Identification of Breast Cancer Subtypes Using Multiple Gene Expression Microarray Datasets

Alexandre Mendes

Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine
School of Electrical Engineering and Computer Science
Faculty of Engineering and Built Environment
The University of Newcastle, Callaghan, NSW, 2308, Australia
Alexandre.Mendes@newcastle.edu.au

Abstract. This work is motivated by the need for consensus clustering methods using multiple datasets, applicable to microarray data. It introduces a new method for clustering samples with similar genetic profiles, in an unsupervised fashion, using information from two or more datasets. The method was tested using two breast cancer gene expression microarray datasets, with 295 and 249 samples; and 12,325 common genes. Four subtypes with similar genetic profiles were identified in both datasets. Clinical information was analysed for the subtypes found and they confirmed different levels of tumour aggressiveness, measured by the time of metastasis, thus indicating a connection between different genetic profiles and prognosis. Finally, the subtypes identified were compared to already established subtypes of breast cancer. That indicates that the new approach managed to detect similar gene expression profile patterns across the two datasets without any *a priori* knowledge. The two datasets used in this work, as well as all the figures, are available for download from the website http://www.cs.newcastle.edu.au/∼mendes/BreastCancer.html.

Keywords: Bioinformatics, breast cancer, data mining, genetic algorithms.

1 Introduction

The introduction of the microarray technology imposed a series of new challenges in terms of producing relevant and statistically sound results. Current research indicates that with the amount of data publicly available, the use of a single dataset is no longer acceptable to justify new medical discoveries. Comparisons with previous, similar studies need to be carried out. A problem that arises in this situation is that microarray data is highly heterogeneous, noisy, and in general, different unsupervised techniques will find different configurations of clusters for the same dataset. In addition, clusters found using a specific dataset sometimes are not observed in other datasets. Consensus clustering techniques try to overcome these problems, with two main types being found in the literature.

D. Wang and M. Reynolds (Eds.): AI 2011, LNAI 7106, pp. 92–101, 2011.

The first deals with single datasets and proposes the concurrent use of several unsupervised clustering techniques, which will likely produce different partitions of the samples. A consensus clustering is then determined using information from all clusters found, usually based on some similarity measure among elements [15,10,4]. The second type of consensus clustering involves finding clusters which have similar profiles across multiple datasets. This is the goal of the method introduced in this paper, and two previous works should be cited. First, Filkov and Skiena (2003) [2] modeled the consensus clustering of multiple datasets as a median partition problem and use three types of heuristics (local search, greedy and simulated annealing) to address it. Then, in Hoshida et al. (2007) [5], the authors use a statistical test to find the consensus clusters. The literature on consensus clustering and microarrays is extensive and even though several methods are available, no single approach dominates the scientific literature.

This paper offers a new consensus clustering technique, which differs from the previous ones mainly because it optimizes three criteria at once. Those are the number of biomarkers that characterize the clusters; consistency of the clusters across datasets; and statistical relevance of the clusters, measured by a classification test.

The method introduced in this work extends the study in Mendes (2008) [8]. It uses a Genetic Algorithm as the search engine and was tested with two well-known datasets from previous breast cancer studies. The first contains 24,158 probes, 295 samples and was introduced in Vijver et al. (2002) [16]. The second dataset has 44,928 probes, 249 samples and was introduced in Miller et al. (2005) [9].

The results presented in this work show the clustering of breast cancer samples into four subtypes. These subtypes were then compared to subtypes of breast cancer already established in the medical literature, using well-known markers. Finally, the subtypes are justified from a clinical standpoint as well, by performing an analysis of the *time of metastasis* associated to the samples in each subtype. Even though such clinical information was not directly used in the determination of the subtypes by our method, the Kaplan-Meier curves of the time of metastasis are consistently distinct in both datasets. In other words, the subtypes found share similar genotypical and phenotypical profiles in the two datasets, even though the method only uses genotypical information.

2 The Consensus Clustering Problem

The consensus clustering problem addressed in this work can be described as follows. Given k input datasets (D_1, D_2,..., D_k), identify partitions of the samples in D_1, D_2,..., D_k into two clusters, which:

- Are supported by the same set of biomarkers; and preferably by a large number of them (higher statistical significance of genetic signatures);
- Have a high accuracy classification of the samples in each dataset (higher intra-cluster similarity and inter-cluster dissimilarity);

- Present a similar proportion of samples in both clusters for all datasets (both clusters should be observed in all datasets to indicate consistency).

The large number of biomarkers indicate that the clusters are not product of a statistical artifact. Although in practice biologists will use just a small number of biomarkers for classification purposes, or when designing a diagnostic kit, a relatively large number of biomarkers is generally recommended for the determination of subtypes. That follows the ‘data-driven’ approach to biomarker discovery, which is discussed in reference [17] (i.e. analyzing the entire genome rather that working from a hypothesis about one or few candidate genes).

The second characteristic is the classification accuracy obtained with a cross-validation procedure, associated to a classification model. High classification accuracy can be associated to high intra-cluster similarity and inter-cluster dissimilarity, and will reflect on the accuracy of future classifiers for prognosis.

Finally, the third characteristic is the proportion of the samples in each subtype and in each dataset, which reflects the consistency of these subtypes across datasets.

These three characteristics are combined into a single objective function used to assess the quality of putative partitions of the samples. Next, we formalize the objective function, but before doing so, consider the following notation:

- $D = \{D_1, D_2,...,D_k\}$: Set of k datasets;
- $C = \{c_1, c_2\}$: Set of classes. In every iteration, the samples are partitioned into two classes: c_1/c_2;
- S_{D_i}: Set of samples in D_i; $|S_{D_i}| = m_{D_i}$;
- $S_{D_i(c_j)}$: set of samples in D_i that belong to class c_j; $|S_{D_i(c_j)}| = m_{D_i(c_j)}$.

The identification of breast cancer subtypes is done iteratively. Initially, the samples are divided into two clusters. Then, those two clusters are further divided into four, and so on, resulting in a binary tree structure. The criterion to stop the division was based on the clinical analysis of the time of metastasis for the samples in each cluster. When no significant difference is observed between two new clusters, in terms of the time of metastasis, we consider that they actually represent the same subtype of the disease, and stop the division.

2.1 Objective Function

The objective function takes into account three characteristics that should be observed in high quality partitions.

- Partitions should be supported by a large number of biomarkers: In each division, the partitions should be supported by the same set of biomarkers in all datasets; and preferably be composed of a large number of them. The method implementation played an important role in this aspect. If we considered all k datasets separately and tried putative partitions for each of them, the search space would be prohibitively large and the sets of biomarkers would be considerably different for each partition in each dataset; i.e. there would be no consistency between biomarkers for any given subtype across datasets.

To overcome this, first we force all datasets D_1, D_2,...,D_k to contain the same genes; i.e. *any gene that is not present in all datasets is removed from the analysis.* Then, one of the datasets is selected as the *main dataset.* This main dataset will have its samples partitioned first, and this partition will induce the partitions in the other $k-1$ datasets.

Let the main dataset chosen be D_1. Given a putative partition for the samples in D_1 into classes c_1 and c_2, a t-Student statistical test is used to determine the $n_{markers}$ associated biomarkers ($p < 0.01$). The biomarkers for the partition in D_1 are then used to induce partitions in the other datasets D_2,...,D_k. A Nearest Neighbor classification model [19] is created with the biomarkers and samples in D_1 and then used to assign the samples in D_2,...,D_k either to class c_1 or c_2.

- High accuracy classification of samples in all datasets: The high accuracy classification of the samples in all datasets acts as a proxy for high intra-cluster similarity and inter-cluster dissimilarity. Given the nearest neighbor-based classification model from D_1 and the partitions of the samples in D_1, D_2,...,D_k, we perform a 10-fold cross-validation [19] in all datasets D_i, calculating the accuracy of each classification acc_{D_i}. The overall accuracy $\overline{acc}_D$ is:

$$\overline{acc}_D = \frac{1}{k}\sum_{i=1}^{k} acc_{D_i} \tag{1}$$

- Similar proportion of samples in clusters across all datasets: It is arguably recommended to have a similar proportion of samples in each cluster, across all datasets. First, this would indicate that subtypes of diseases identified are present in all datasets. Moreover, the proportion of the number of samples in each class indicates that a subtype of the disease, more/less common in a dataset, should be more/less common in all other datasets as well. This is a strong assumption, which only holds if different cohorts share similar sampling characteristics. The balance of the partition of the samples is denoted as B, and is calculated as follows. First, let:

$$\overline{m}_{c_j} = \frac{1}{k}\sum_{i=1}^{k} \frac{m_{D_i(c_j)}}{m_{D_i}} \tag{2}$$

be the average proportion of samples in class c_j in all datasets. The balance should be optimum when $m_{D_i(c_j)}/m_{D_i}$, i.e. the proportion of samples clustered in c_j is the same in every dataset D_i. The equation for the balance is:

$$B = \sum_{i=1}^{k} \left|\overline{m}_{c_1} - \frac{m_{D_i(c_1)}}{m_{D_i}}\right| + \sum_{i=1}^{k} \left|\overline{m}_{c_2} - \frac{m_{D_i(c_2)}}{m_{D_i}}\right| \tag{3}$$

Finally, the objective function used in this work is stated as:

$$obj = n_{markers} * \overline{acc}_D * \frac{1}{B + \epsilon} \tag{4}$$

The objective function aims at a trade-off between large number of biomarkers for D_1 ($n_{markers}$); high average accuracy of the classification across datasets ($\overline{acc}_D$); and good balance of classes across datasets ($B \approx 0$).

2.2 The Genetic Algorithm

The problem of finding the partition of the samples that maximizes Eq. 4 was addressed using a Genetic Algorithm (GA). GAs are population-based search methods [3] where a population of solutions evolves through the application of special operators (recombination and mutation), and selection pressure.

- Representation: The search space of the consensus clustering problem consists of all the possible partitions of the samples in the dataset D_1 into two classes. In terms of genetic algorithm implementation, a partition P is represented as a binary array $P = [p_1, p_2, ..., p_{m_{D_1}}]$, with $p_i \in \{0, 1\}$.

- Population structure: The GA employs a population structure that follows a complete ternary tree with three levels, i.e. 13 individuals. This structure was object of study in the past, and genetic/memetic algorithms using it performed better compared to non-structured approaches in several combinatorial optimization problems [1,11]. Also, the use of fewer individuals is critical because, in this problem, the objective function calculation is very time-consuming, as it involves several, complex steps.

- Mutation: The mutation operator implemented was the bit-swap. A sample is chosen uniformly at random and moves from a class to another, i.e. either $c_1 \rightarrow c_2$ or $c_2 \rightarrow c_1$. This 1-bit mutation is applied to 10% of the offspring created, also chosen uniformly at random.

- Recombination and acceptance policy: The recombination operator chosen was the uniform crossover (UX) [12]. In every generation, a number of individuals equal to the size of the population is created and evaluated. Offspring that are better than at least one of their parents survive to the next generation, directly replacing their worst parent. Even though this scheme creates a strong evolutionary pressure, premature convergence is controlled by checking population diversity and applying restart procedures.

- Population diversity and restart: The diversity check procedure verifies at every generation whether any offspring created was better than at least one of its parents. If none was better, a population restart follows, which keeps the current best solution within the population (*elitist restart*), and replaces all others by randomly-generated solutions. Indeed, if no solution created within a generation was better than one of its parents, that indicates that the current population has evolved enough generations to be consisted of high-quality individuals only, which are also likely to be very similar.

3 The Breast Cancer Datasets

Two breast cancer microarray datasets were used in this work. The first one (dataset D_1) is from a study with 295 patients diagnosed with primary breast carcinomas presented in Vijver et al. (2002) [16]. A 25,000-gene cDNA array consisting of 24,479 probes was used for each patient. The second dataset (D_2) comes from a study comprising 259 primary breast cancer patients presented in Miller et al. (2005) [9]. Each patient was sampled using an Affymetrix genechip with 38,061 probes.

A first pre-processing procedure removed duplicate genes from both datasets, resulting in D_1 keeping 14,547 unique genes; and D_2 keeping 18,342 unique genes. A second step involved forcing the two datasets to contain exactly the same genes (to enforce consistency of classifiers' attributes). Using the gene symbols as identifiers, there was a total of 12,325 common genes.

4 Results

After applying the clustering algorithm to the two datasets, a binary tree with the partition of the breast cancer samples was produced. It is shown in Figure 1 and depicts the types found, the biomarkers for the partitions found in both datasets, and the Kaplan-Meier (K-M) curves for the time of metastasis associated to the types identified.

Samples were first divided into two subtypes (Types 1 and 2) and then into four others (Types 3 to 6). Note that the biomarkers in each specific division are the same for the two datasets D_1 and D_2, and the types have a similar clinical profile in terms of prognosis. Type 2 is more aggressive than Type 1; and Types 3 and 5 are more aggressive as well, compared to Types 4 and 6.

Additional divisions of Types 3 to 6 into more subtypes were tested, but the clinical profiles obtained were not consistent across the two datasets. The classification shown in Figure 1 contains only those subtypes that present consistent clinical profiles.

4.1 Comparison with Existing Subtypes

There are five subtypes of breast cancer broadly accepted by the medical community: *normal breast-like*, *basal*, *luminal A*, *luminal B*, and *HER2+/ER-*. In order to compare the four subtypes identified in this work with them, we analyzed a number of genetic markers associated to breast cancer, collected from the following studies:

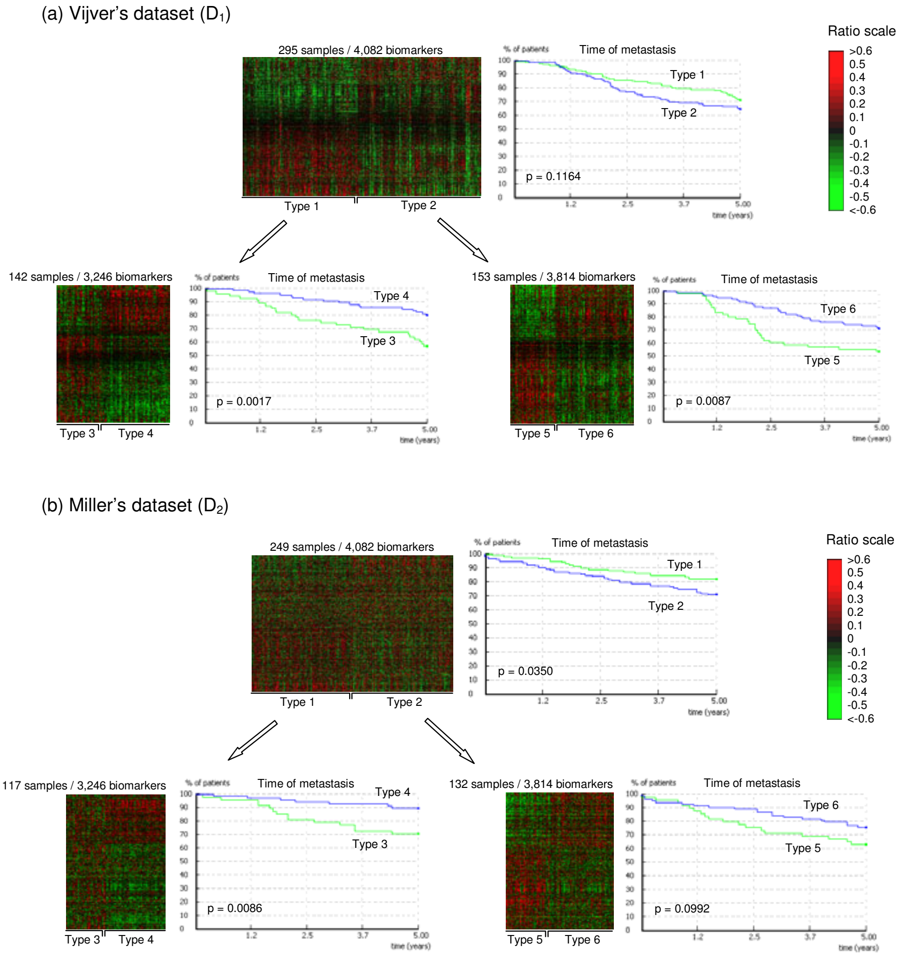


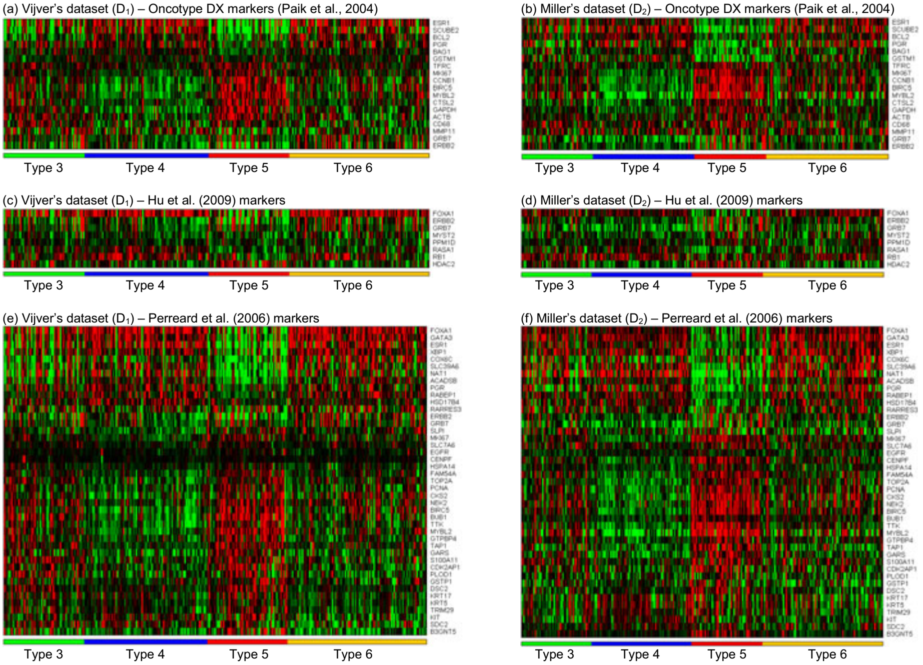
Fig. 1. (a) Classification of breast cancer subtypes for Vijver's dataset. Samples were initially divided into two subtypes – Types 1 and 2 – which were further divided into the final four subtypes – Types 3 to 6. For each division we present a genetic signature with the biomarkers obtained by a t-student statistical test ($p < 0.01$). Next to each signature we present the associated Kaplan-Meier curves for the time of metastasis. (b) Classification of breast cancer subtypes for Miller's dataset. The subtypes are analogous to the ones identified in (a).

- *Perreard et al. (2006)* [14]: 53 biomarkers for different subtypes of breast cancer – 37 so-called 'intrinsic' genes to classify the subtypes, plus PGR, EGFR and 14 proliferation-related genes.
- *Hu et al. (2009)* [6]: 9 oncogenes and tumor suppressor genes.
- *Paik et al. (2004)* [13] – *Oncogene DX*: a breast cancer prognosis kit based on 21 genes for ER+, lymph node-negative patients.

The expression profiles of the genetic markers mentioned in the three studies above are shown in Figure 2. They were divided according to the study and the dataset. Samples are ordered from Type 3 to Type 6, in all figures.

Basal tumors are characterized by being ESR1, PGR and ERBB2 negative, i.e. these three markers are under-expressed. This subtype is also referred to as triple receptor negative [7]. Type 5 is the cluster where those genes are the least expressed. This is an aggressive subtype and that behaviour agrees with K-M curves in Figure 1a-b. Therefore, we can associate Type 5 to the basal breast cancer subtype.

Two other types are also identifiable: *Luminal A* and *luminal B*. These types are molecularly similar, being characterized by the over-expression of ESR1, PGR, GATA3 and FOXA1. That occurs in both Types 4 and 6. The main



(a) Vijver's dataset (D_1) – Oncotype DX markers (Paik et al., 2004)

(b) Miller's dataset (D_2) – Oncotype DX markers (Paik et al., 2004)

(c) Vijver's dataset (D_1) – Hu et al. (2009) markers

(d) Miller's dataset (D_2) – Hu et al. (2009) markers

(e) Vijver's dataset (D_1) – Perreard et al. (2006) markers

(f) Miller's dataset (D_2) – Perreard et al. (2006) markers

Fig. 2. Gene expression of breast cancer markers found in the literature, considering the four subtypes identified in Figure 1a-b. Three sets of genetic markers are compared: *(a-b)* Oncotype DX [13], *(c-d)* Hu et al. (2009) [6] and *(e-f)* Perreard et al. (2006) [14]. Based on the markers' expression, we can make the following mapping: Type 5 corresponds to *basal* samples. Basal is one of the most aggressive subtypes of breast cancer, which is in agreement with the Kaplan-Meier curves in Figure 1a-b. Types 4 and 6 correspond to *Luminal A* and *Luminal B* samples, respectively. They are similar with respect to the biomarkers, but proliferation-related genes are under-regulated in Type 4 and over-regulated in Type 3. Finally, Type 3 corresponds to HER2+/ER- tumors, which is also a very aggressive subtype – again showing agreement with the K-M curves in Figure 1a-b.

difference between luminal A and B is that *proliferation-related genes* are under-expressed and over-expressed in those subtypes, respectively [18]. The proliferation genes that we refer to are listed in Perreard et al. (2006) [14] (HSPA14, GTPBP4, PCNA, CKS2, NEK2, TOP2A, BUB1, TTK, FAM54A, MKI67, MYBL2, BIRC5 and CENPF). This difference indicates that Type 4 corresponds to *luminal A* and Type 6 to *luminal B*.

Finally, Type 3 appears to correspond to HER2+/ER- tumors. This type is characterized by the under-expression of ESR1 and PGR; and over-expression of ERBB2. In addition, proliferation-related genes are over-expressed. From the clinical standpoint, HER2+/ER- is, together with basal, one of the most aggressive breast cancer tumor subtype. That would be in agreement with the K-M curves in Figure 2. These findings illustrate how the method managed to identify, across two distinct datasets, four subtypes broadly accepted by the scientific community. Moreover, the clinical aspects have also shown consistency across datasets and agreed with the scientific literature for the subtypes.

5 Conclusion

In this paper we introduce a new method to perform classification of microarray samples using multiple datasets, and test the approach using two publicly available breast cancer datasets. Four subtypes were identified and presented similar gene expression profiles across both datasets, as well as similar clinical profiles (based on time of metastasis). A subsequent analysis comparing those four subtypes with the currently accepted subtypes of breast cancer in the scientific community provided a mapping between them. The types basal, luminal A, luminal B and HER2+/ER- were mapped into the four subtypes identified by our algorithm by analyzing the expression profile of several markers reported in the literature. That result was also corroborated by the analysis of the time of metastasis, which shows that the types mapped into basal and HER2+/ER- subtypes have a more aggressive behavior.

It is worth emphasizing that the method introduced in this study successfully discovered subtypes in an unsupervised, unbiased (data-driven) fashion, using data from a genetically heterogeneous disease. It has the potential to impact the discovery of subtypes of other heterogeneous diseases for which microarray data is available.

References

1. Buriol, L., Franca, P., Moscato, P.: A new memetic algorithm for the asymmetric traveling salesman problem. Journal of Heuristics 10, 483–506 (2004)
2. Filkov, V., Skiena, S.: Integrating microarray data by consensus clustering. In: Proceeding of the 15th IEEE International Conference on Tools with Artificial Intelligence, pp. 418–426. IEEE Computer Society (2003)
3. Glover, F., Kochenberger, G.: Handbook of Metaheuristics. Springer, USA (2003)

4. Grotkjaer, T., Winther, O., Regenberg, B., Nielsen, J., Hansen, L.: Robust multiscale clustering of large dna microarray datasets with the consensus algorithm. Bioinformatics 22, 58–67 (2006)
5. Hoshida, Y., Brunet, J., Tamayo, P., Golub, T., Mesirov, J.: Subclass mapping: Identifying common subtypes in independent disease data sets. PLoS ONE 2, e1195 (2007)
6. Hu, X., Stern, H.M., Ge, L., O'Brien, C., Haydu, L., Honchell, C.D., Haverty, P.M., Wu, B.P.T., Amler, L.C., Chant, J., Stokoe, D., Lackner, M.R., Cavet, G.: Genetic alterations and oncogenic pathways associated with breast cancer subtypes. Molecular Cancer Research 7, 511–522 (2009)
7. Irvin Jr., W., Carey, L.: What is triple-negative breast cancer? European Journal of Cancer 44, 2799–2805 (2008)
8. Mendes, A.: Consensus clustering of gene expression microarray data using genetic algorithms. In: Proceedings of PRIB 2008 - Third IAPR International Conference on Pattern Recognition in Bioinformatics (Supp. volume), pp. 181–192 (2008)
9. Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T., Bergh, J.: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the National Academy of Sciences 102, 13550–13555 (2005)
10. Monti, S., Mesirov, P.T.J., Golub, T.: Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 52, 91–118 (2003)
11. Moscato, P., Mendes, A., Berretta, R.: Benchmarking a memetic algorithm for ordering microarray data. Biosystems 88, 56–75 (2007)
12. Olariu, S., Zomaya, A.: Handbook of Bioinspired Algorithms and Applications. Chapman & Hall/CRC, USA (2005)
13. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., Hiller, W., Fisher, E.R., Wickerham, L., Bryant, J., Wolmark, N.: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England Journal of Medicine 351, 2817–2826 (2004)
14. Perreard, L., Fan, C., Quackenbush, J., Mullins, M., Gauthier, N., Nelson, E., Mone, M., Hansen, H., Buys, S., Rasmussen, K., Orrico, A., Dreher, D., Walters, R., Parker, J., Hu, Z., He, X., Palazzo, J., Olopade, O., Szabo, A., Perou, C.M., Bernard, P.: Classification and risk stratification of invasive breast carcinomas using a real-time quantitative rt-pcr assay. Breast Cancer Research 8, R23 (2006)
15. Swift, S., Tucker, A., Vinciotti, V., Martin, N., Orengo, C., Liu, X., Kellam, P.: Consensus clustering and functional interpretation of gene-expression data. Genome Biology 5, R94 (2004)
16. van de Vijver, M., He, Y., van't Veer, L., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E., Friend, S., Bernards, R.: A gene expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine 347, 1999–2009 (2002)
17. van't Veer, L., Bernards, R.: Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 452, 564–570 (2008)
18. Weigelt, B., Baehner, F., Reis-Filho, J.: The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. Journal of Pathology 220, 263–280 (2010)
19. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, USA (2005)