Comparative Study on Normalization Procedures for Cluster Analysis of Gene Expression Datasets

Marcilio C. P. de Souto, Daniel A. S. Araujo, Ivan G. Costa
Rodrigo G. F. Soares, Teresa B. Ludermir, Alexander Schliep

***Abstract*—Normalization before clustering is often needed for proximity indices, such as Euclidian distance, which are sensitive to differences in the magnitude or scales of the attributes. The goal is to equalize the size or magnitude and the variability of these features. This can also be seen as a way to adjust the relative weighting of the attributes. In this context, we present a first large scale data driven comparative study of three normalization procedures applied to cancer gene expression data. The results are presented in terms of the recovering of the true cluster structure as found by five different clustering algorithm.**

I. Introduction

As pointed out by [1], cluster analysis techniques of gene expression microarray data is of increasing interest in the field of functional genomics. One of the reasons for this is the need for molecular-based refinement of broadly defined biological classes, with implications in cancer diagnosis, prognosis and treatment [1], [2], [3].

Despite the importance of the choice of the clustering method or data pre-processing in the analysis of cancer data sets, there are in the literature few guidelines or standard procedures about how these data should be analyzed [4]. The choice of methods are mostly driven by the familiarity of biological experts to the methods rather than the method characteristics. For example, the wide use of hierarchical clustering methods is mostly a consequence of its similarity to phylogenetic methods, which biologists are often acquaint to.

In particular, distinct normalization[1] procedures are usually applied to such datasets. These choices are often not justified. Motivated by such issues, in this paper, we present a first large scale data driven comparative study of three normalization procedures applied to cancer gene expression data. The results are presented in terms of the recovering of the true cluster structure as found by five different clustering algorithm.

The remaining of the paper is divided into four sections. In Section II, we introduce the three normalization procedures that will be analyzed. The experimental design, including the description of the datasets and of the algorithms, is given in Section III. Section IV shows our experimental study, as well as a discussion on the results. Some final remarks are presented in Section V.

II. Normalization

Clustering is an important tool for the exploration of datasets with no or very little prior information [5]. In order to cluster a set of patterns, clustering methods need an index of alikeness or association between the data patterns. This can be achieved by the use of proximity (similarity or dissimilarity) indices that calculate the alikeness of two patterns. For the choice of a suitable index, the type of the feature (attributes) and the characteristics of the index should be taken into consideration.

Also, in many practical situations a dataset could present patterns whose attributes or features values lie within different dynamic ranges [5], [6]. In this case, for proximity indices such as Euclidean distance, features with large values will have a larger influence than those with small values. However, this not necessarily will reflect their importance for defining the clusters.

The previous problem is often addressed by normalizing the features values so that they lie within similar ranges. There are several approaches to normalization of attributes values [5], [7]. As the datasets used in our studies have no categoric features, we consider only the case involving numeric values. More precisely, we analyze three different forms of feature normalization.

The first two of them have been widely used in clustering applications [6], [5]: one is based on the z-score formula (standardization) and the other scales the features values to $[0, 1]$. The last normalization presented transforms the values of the attributes in a rank. Such a kind of transformation is more robust to outlier than the other two normalization methods [6].

In order to make the definitions of the normalization procedures clearer, before introducing them, we will give some basic concepts. Following the definitions in [5], the basic unit of data is called a *pattern* (*instance*), denoted by a d-vector, whose components are scalars called *features* (*variables* or *attributes*). The ith pattern is denoted by the column vector $\mathbf{x}_i^*$ and the jth feature value for the ith pattern is denoted by x_{ij}^*. The symbol "*" stands for the unnormalized data. Let n be the number of instances in the analysis. The pattern matrix is the $n \times d$ matrix A^*, where each row is a pattern:

M. C. P. de Souto and D. S. A. Araujo are with the Department of Informatics and Applied Mathematics, Fed. Univ. of Rio Grande do Norte, Natal, Brazil, email: marcilio@dimap.ufrn.br. R. G. F. Soares and T. B. Ludermir are with the Center of Informatics, Federal University of Pernambuco, Recife, Brazil. I. G. Costa and Alexander Schliep are with Max Planck Institute for Molecular Genetics, Berlin, Germany.

[1] In this paper, we use the term *normalization* in a generic sense: it could mean the transformation of the values of a feature to have zero mean and unity variance or some kind of linear scaling of them.

$$A^* = \begin{bmatrix} \mathbf{x}_1^* & \mathbf{x}_2^* & \dots & \mathbf{x}_n^* \end{bmatrix}^T = \begin{bmatrix} x_{11}^* & x_{12}^* & \dots & x_{1d}^* \\ x_{21}^* & x_{22}^* & \dots & x_{2d}^* \\ \dots & \dots & \dots & \dots \\ x_{n1}^* & x_{n2}^* & \dots & x_{nd}^* \end{bmatrix}$$

Based on the previous matrix, we can define the following type of normalization:

$$x_{ij} = \frac{x_{ij}^* - m_j}{s_j}, \qquad (1)$$

where m_j and s_j are, respectively, the sample mean and standard deviation of attribute j. This type of normalization, which uses the z-score formula, translates and scales the axes so that transformed feature x_{ij} will have zero mean and unit variance. Hereafter, for short, we will refer to this normalization as Z_1.

The second procedure involves normalization with the use of the maximum and minimum values on the attribute:

$$x_{ij} = \frac{x_{ij}^* - Min(j)}{Max(j) - Min(j)}, \qquad (2)$$

where $Min(j)$ and $Max(j)$ are, respectively, the smallest and largest values that unnormalized feature j takes in the data. Hereafter, for short, we will refer to this normalization as Z_2. Assuming nonnegative values, an attribute normalized with Z_2 is bounded by 0.0 and 1.0, with at least one observed value at each of these end points [6]. Its standard deviation will be $s_j/(Max(i) - Min(j))$. Also, differently from Z_1, the transformed mean and variance resulting from Z_2 will not be the constant across all features.

Both procedures could be adversely affected by the presence of outliers on the features, mainly Z_2 that depends on the minimum and maximum values. A different approach to normalization, which is more robust to outliers, converts the values of the attributes to ranks:

$$x_{ij} = Rank(x_{ij}^*) \qquad (3)$$

Equation 3 yields a normalized feature with mean of $(n+1)/2$, range $n-1$, and variance $(n+1)[((2n+1)/6) - ((n+1)/4)]$ for all features [6]. For short, hereafter, we will refer to this normalization as Z_3.

III. Experimental Design

We present the study of the effect of normalization procedures Z_1, Z_2 and Z_3 on the recovery of cluster structure in 20 Affymetrix gene expression datasets. In order to provide a basis for comparison, we also include the analysis based on the non-normalized data. We refer to the results with the untransformed data as Z_0.

A. Datasets

Twenty microarray datasets are included in this analysis. They were built from the data available in [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21]. As can be observed in Table I, these datasets present different values for characteristics such as number of patterns (second column), number of classes (third column), distribution of patterns within the classes (fourth column), dimensionality (fifth column), and dimensionality after feature selection (last column).

To be more precise, we focus our study on microarray data from cancer. In such a context, one of the main goals is to identify previously unknown cancer subtypes for which gene expression profiles are homogeneous within a subtype but different between subtypes [22], [23], [12], [24], [21]. As pointed out before, the discovering of new subtypes of a disease could aid the decision-making process related to the choice of existing treatments, as well as in the development of new target-specific therapeutics [1], [2], [3].

Microarray technology is usually available in two different platforms, cDNA and Affymetrix [1], [2], [3]. Measurements of Affymetrix arrays are estimates on the number of RNA copies found in the cell sample, while cDNA microarrays values are ratios of the number of copies in relation to a control cell sample. Data from these platforms have distinct distributions, thus normalization procedures would have distinct impacts on each platform. In order to make our analysis less complex, we will approach only data produced by Affymetrix microarrays.

One characteristic of the data produced via Affymetrix microarray technology, which is interesting for our analysis of the normalization procedures, is the variability of the magnitude of the expression level of a gene. For example, in a given array, there could be a gene whose expression level is around 10 and other whose level is around 10,000.

In fact, following other works, for our datasets, all genes with expression level below to 10 are set to the minimum threshold of 10. The maximum threshold is set at 16,000. Values below or above these thresholds are often not reliable [1], [17], [25]. That is, our analysis is performed on the scaled data to which the ceiling and threshold values have been applied.

Furthermore, in order to remove uninformative genes, we apply the following procedure. For each gene j (attribute), we compute the mean m_j. But before doing so, in order to get rid of extreme values, we discard the 10% largest and smallest values. Based on this mean, we transform every value x_{ij}^* to:

$$y_{ij} = \log_2(x_{ij}^*/m_j)$$

After the previous transformation, we select for further analysis genes whose expression level differed by at least l-fold, in at least c samples, from their mean expression level across samples. With few exceptions, the parameters l and c were chosen in such a way as to yield a filtered dataset with around at least 10% of the original number of genes (features).

Finally, it is important to point out that the data transformed with the previous equation is only used in the filtering step.

B. Clustering Methods and Recovery Measure

Five clustering algorithms are used to generate partitions solutions and formed one factor in the overall experiment design. These are the single linkage, complete linkage, average linkage, k-means and Shared Nearest Neighbors (SNN). These algorithms have been chosen to provide a wide range of recovery effectiveness, as well as to give some generality to the results. In our analysis, all of them are implemented with Euclidean distance. The Euclidean distance between two patterns $\mathbf{x}_i$ and $\mathbf{x}_k$ is given by the following equation.

$$d(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_{j=1}^{d} (x_{ij} - x_{kj})^2} \qquad (4)$$

Hierarchical clustering methods, more specifically the agglomerative ones, are procedures for transforming a distance matrix into a dendrogram or tree [5]. These algorithms start with each pattern representing a cluster, then the methods gradually merge these clusters into larger ones. Intuitively, agglomerative methods yield a sequence of nested partitions starting with the trivial clustering in which each item is in a unique cluster, and ending with the trivial clustering in which all patterns are in the same cluster.

Among the different agglomerative methods, there are three broader used variations, which are used in this paper: Complete Linkage (CL), Average Linkage (AL), and Single Linkage (SL). These variations differ in the way the distance between two clusters is calculated. For the single linkage algorithm, the distance between two clusters is determined by the two closest patterns in different clusters. In contrast, the complete linkage method employs the farthest distance of a pair of patterns to define the inter-cluster distance. In the case of the average method, the distance between two clusters is calculated by the average distance between the patterns in one group and the patterns in the other group. Such a method has been extensively used in the literature of gene expression analysis [2], [3], [26], [27], although experimental results have shown that in many cases the complete linkage outperforms it [4].

Another method popular in the literature of gene expression analysis is the k-means [2], [3]. k-means is a partitional iterative algorithm that optimizes the best fitting between clusters and their representation, using a predefined number of clusters [5]. Starting with prototypes values from randomly selected patterns, the method works on two alternates steps: (1) an allocation step, where all patterns are allocated to the cluster with the prototype with lower dissimilarity; (2) and a representation step, where a prototype is constructed for each cluster. A major problem of this algorithm is its sensitivity to the selection of the initial partition. As a consequence, the algorithm could converge to a local minimum [5]. In order to prevent the local minimum problem, a number of runs with different initializations are executed. Then, the best run, based on some cohesion measure, is taken as the result. Another characteristic of this method is its robustness to noisy data.

The Shared Nearest Neighbor algorithm (SNN) is a recent technique and was selected because it can robustly deal with high dimensionality, noise and outliers [28]. Such an algorithm searches for the nearest neighbors of each pattern and uses the number of neighbors that two points share as the proximity index between them. With this index, SNN employs an approach based on density to find representatives patterns and build clusters around them. Besides the number of nearest neighbors (NN), two other kinds of parameters are considered: the ones regarding the weights of the shared nearest neighbor graph (strong, merge and label) and others related to the number of strong links (topic and noise). These parameters are thresholds on which each step of the algorithm is based.

In terms of the index to measure the success of the algorithm in recovering the true partition of the dataset, as in [6], we use the corrected Rand [5], [6]. The corrected Rand index can take values from -1 to 1, with 1 indicating a perfect agreement between the partitions, and the values near 0 or negatives corresponding to cluster agreement found by chance.

Formally, let $U = \{u_1, \ldots, u_r, \ldots, u_R\}$ be the partition given by the clustering solution, and $V = \{v_1, \ldots, v_c, \ldots, v_C\}$ be the partition formed by an a priori information independent of partition U (the gold standard). The corrected Rand is defined as

$$cR = \frac{\sum_i^R \sum_j^C \binom{n_{ij}}{2} - \binom{n}{2}^{-1} \sum_i^R \binom{n_{i\cdot}}{2} \sum_j^C \binom{n_{\cdot j}}{2}}{\frac{1}{2}[\sum_i^R \binom{n_{i\cdot}}{2} + \sum_j^C \binom{n_{\cdot j}}{2}] - \binom{n}{2}^{-1} \sum_i^R \binom{n_{i\cdot}}{2} \sum_j^C \binom{n_{\cdot j}}{2}}$$

where (1) n_{ij} represents the number of objects in clusters u_i and v_j; (2) $n_{i\cdot}$ indicates the number of objects in cluster u_i; (3) $n_{\cdot j}$ indicates the number of objects in cluster v_j; (4) n is the total number of objects; and (5) $\binom{a}{b}$ is the binomial coefficient $\frac{a!}{b!(a-b)!}$.

IV. Empirical study and discussion

All the five clustering algorithm and the three normalization procedures were used to produce the respective partition of the datasets. We also run the algorithms with the unnormalized version of the datasets. The number of cluster is set to be equal to the true number of the classes in the data. The known class labels were not used in any way during the clustering.

In order to build the partition from the hierarchical methods, the trees were run from root to the leaves, then the k first sub-trees were taken as the clusters, with k equal to the exact number of classes in the dataset. In the case of the k-means, as it is nondeterministic, for each configuration pair (dataset, normalization procedure), we run the algorithm 30

TABLE I
DATASET DESCRIPTION

Dataset	*n*	Nr. Classes	Dist. Classes	*d*	Filtered *d*
Armstrong-V1 [8]	72	2	24,48	12582	1081
Armstrong-V2 [8]	72	3	24,20,28	12582	2194
Bhattacharjee [9]	203	5	139,17,6,21,20	12600	1543
Chowdary [10]	104	2	62,42	22283	182
Dyrskjot [11]	40	3	9,20,11	7129	1203
Golub-V1 [12]	72	2	47,25	7129	1877
Golub-V2 [12]	72	3	38,9,25	7129	1877
Gordon [13]	181	2	31,150	12533	1626
Laiho [14]	37	2	8,29	22883	2202
Nutt-V1 [15]	50	4	14,7,14,15	12625	1377
Nutt-V2 [15]	28	2	14,14	12625	1070
Nutt-V3 [15]	22	2	7,15	12625	1152
Pomeroy-V1 [16]	34	2	25,9	7129	857
Pomeroy-V2 [16]	42	3	10,10,10,4,8	7129	1379
Ramaswamy [17]	190	14	11,10,11,11,22,10,11,10,30,11,11,11,11,20	16063	1363
Shipp [18]	77	2	58,19	7129	798
Su [19]	174	10	26,8,26,23,12,11,7,27,6,28	12533	1571
West [20]	49	2	25,24	7129	1198
Yeoh-V1 [21]	248	2	43,205	12625	2526
Yeoh-V2 [21]	248	6	15,27,64,20,79,43	12625	2526

times. Then, for further analysis, we pick the partition with the best corrected Rand (cR).

For the SNN, we execute the algorithm with several values for its parameters (2%, 5%, 10%, 20%, 30% and 40% of NN), topic (0, 0.2, 0.4, 0.6, 0.8 and 1) and merge (0, 0.2, 0.4, 0.6, 0.8 and 1). Preliminary experiments showed that variations of the other parameters did not produce very different results. Thus, the default value were used for the parameter strong, and the value 0 was used for the parameters noise and label (to have all points assigned to a cluster). From the partitions created with such parameters values, we pick for further analysis the partition with best cR and with k in the interval of interest.

In terms of results, Table II illustrates, for each dataset, the configuration pair (algorithm, normalization procedure) that showed the best corrected Rand value (third column). In order to put this result into perspective, on the fourth column of the table we present the best corrected Rand value achieved with the non-normalized data (Z_0). As one could expect, the majority of the best corrected Rand values were obtained with some kind of normalization procedure. This happens for 14 out of the 20 datasets.

One surprising result is the good performance achieved with Z_3. Such a normalization procedure was the best one for eight datasets, whereas the second best procedure, Z_1, which is one of most traditionally used procedure in cluster analysis (including gene expression data) [5], [6], [2], worked best for seven datasets. Thus, contrary to conventional expectation, Z_3 outperformed Z_1 and Z_2.

One reason for the behavior previously described could be the presence of outliers or noise in our dataset sets, as Z_3 is more robust to deal with this sort of problem [6]. In fact, it is well known that microarrays suffers from several sources of noise; either by manufacture failures, problems in the reading procedure; unspecific probes, variability in biological samples, or variations in the environment conditions in which experiments were performed [29].

In another kind of analysis, we investigate the impact of each normalization procedure on the performance of the algorithms. For instance, Table III illustrates the mean and standard deviation, across all datasets, of the corrected Rand (cR) for the pair (algorithm, normalization procedure). Based on this table, as it occurred in the study in [6], we can observe that, for some cases, there is a clear interaction between algorithm and normalization procedures.

This impact was more significative on the hierarchical methods. For instance, the mean of the cR for the AL algorithm with Z_3 was of 0.22 against a value of 0.05 (second best value) for the case of Z_2. Again, as stated before, one reason for this could be the presence of outliers in the dataset. A similar behavior happens for the case of the CL algorithm. For the k-means and SNN the the magnitude of the mean difference among the different normalization procedures was not as large as in the previous cases. That is, in general, for each of these algorithm there was no single normalization procedure that could be stated to be the best (this includes also the unnormalized data - Z_0).

V. FINAL REMARKS

In the context of cluster analysis, for the case of proximity indices, such as Euclidean distance, that are sensitive to differences in the magnitude or scales of the attributes, before clustering, the normalization of them could be needed [5].

In order to address the previous issue, in this paper, we presented a first large scale data driven comparative study of three normalization procedures applied to 20 cancer gene expression datasets. The results were presented in terms of the recovering of the true cluster structure as found by five different clustering algorithm.

Two of the normalization procedures investigated have been widely used in clustering applications [6], [5]: one

TABLE II
RESULTS

Dataset	**Alg.**	**Norm.**	**cR**	Z_0
Armstrong-V1	KM	Z_1	0.64	0.47
Armstrong-V2	KM	Z_3	0.81	0.70
Bhattacharjee	SNN	Z_3	0.62	0.47
Chowdary	KM	Z_2	0.92	0.32
Dyrskjot	SNN	Z_0	0.63	0.63
Golub-V1	KM	Z_1,Z_2	0.94	0.59
Golub-V2	KM	Z_0	0.70	0.70
Gordon	KM	Z_1,Z_2,Z_3	0.97	0.16
Laiho	KM	Z_1	0.28	0.24
Nutt-V1	SNN	Z_0,Z_1,Z_2	0.44	0.44
Nutt-V2	SNN	Z_0,Z_1,Z_2	0.72	0.72
Nutt-V3	AL,CL,KM	Z_1,Z_2,Z_3	1.00	0.82
Pomeroy-V1	SNN	Z_1	0.24	0.16
Pomeroy-V2	KM	Z_3	0.56	0.54
Ramaswamy	KM	Z_3	0.49	0.30
Shipp	KM	Z_3	0.12	0.10
Su	KM	Z_1	0.66	0.56
West	KM	Z_3	0.50	0.39
Yeoh-V1	KM	Z_0	0.92	0.92
Yeoh-V2	KM	Z_0	0.25	0.25

TABLE III
CLUSTERING ALGORITHM X NORMALIZATION: MEAN AND STANDARD DEVIATION OF cR FOR ALL DATASET

Alg.	$\mathbf{Z_0}$	$\mathbf{Z_1}$	$\mathbf{Z_2}$	$\mathbf{Z_3}$
SL	0.02 ± 0.04	0.01 ± 0.03	0.01 ± 0.03	0.00 ± 0.03
AL	0.05 ± 0.09	0.03 ± 0.08	0.05 ± 0.10	0.22 ± 0.28
CL	0.13 ± 0.20	0.05 ± 0.09	0.14 ± 0.25	0.22 ± 0.23
KM	0.36 ± 0.29	0.37 ± 0.34	0.44 ± 0.32	0.42 ± 0.29
SNN	0.29 ± 0.22	0.27 ± 0.24	0.26 ± 0.23	0.28 ± 0.19

is based on the z-score formula (Z_1) and the other scales the features values to $[0, 1]$ (Z_2). The other normalization presented transforms the values of the attributes in a rank (Z_3).

To give some generality to our results, these procedures were studied in conjunction with five different clustering algorithms (all of them implemented with Euclidean distance): the single linkage, complete linkage, average linkage, k-means and Shared Nearest Neighbors (SNN) algorithms.

In terms of results, as it was expected, most of the best corrected Rand values were obtained with some of sort normalization in the datasets. More precisely, this happens for 14 out of the 20 datasets.

Nevertheless, one surprising result was the excellent performance achieved with Z_3: it was the best normalization procedure for eight datasets. The second best procedure, Z_1, which is one of most traditionally used procedure in cluster analysis, worked best for seven datasets. Thus, in contrast to conventional expectation, Z_3 outperformed Z_1 and Z_2.

Our experimental results also showed that there is a clear interaction between algorithm and normalization procedures. The impact of the interaction was stronger for the case of the hierarchical clustering methods: the mean of the cR obtained with Z_3 was much higher than with the other procedures. As previously discussed, one reason for this could be the presence of outliers and noise in the dataset.

Finally, based on Table III, we can see that in all datasets, either SNN and or k-means are the methods with best recovery of class labels. The average linkage and complete linkage methods had cR values as high as k-means for only one dataset, "Nutt-V3". While the main objective of our study is not the comparison of the clustering methods themselves, the shortcoming of hierarchical methods is noticeable in other comparative studies on gene expressions data [26], [27], [4]. However, hierarchical methods are still widely used in clustering gene expression datasets.

REFERENCES

[1] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, pp. 91–118, 2003.

[2] J. Quackenbush, "Computational analysis of cDNA microarray data," *Nature Reviews*, vol. 6, no. 2, pp. 418–428, 2001.

[3] D. Slonim, "From patterns to pathways: gene expression data analysis comes of age," *Nature Genetics*, vol. 32, pp. 502–508, 2002.

[4] P. D'haeseleer, "How does gene expression clustering work?" *Nat Biotech*, vol. 23, no. 12, pp. 1499–1501, Dec. 2005. [Online]. Available: http://dx.doi.org/10.1038/nbt1205-1499

[5] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice Hall, 1988.

[6] G. W. Milligan and M. C. Cooper, "A study of standardization of variables in cluster analysis," *Journal of Classification*, vol. 5, pp. 181–204, 1988.

[7] ——, "A study of the comparability of external criteria for hierarchical cluster analysis," *Multivariate Behavorial Research*, vol. 21, pp. 441–458, 1986.

[8] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia." *Nat Genet*, vol. 30, no. 1, pp. 41–47, Jan 2002. [Online]. Available: http://dx.doi.org/10.1038/ng765

[9] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson, "Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses." *Proc Natl Acad Sci U S A*, vol. 98, no. 24, pp. 13 790–13 795, Nov 2001. [Online]. Available: http://dx.doi.org/10.1073/pnas.191502998

[10] D. Chowdary, J. Lathrop, J. Skelton, K. Curtin, T. Briggs, Y. Zhang, J. Yu, Y. Wang, and A. Mazumder, "Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative." *J Mol Diagn*, vol. 8, no. 1, pp. 31–39, Feb 2006.

[11] L. Dyrskjt, T. Thykjaer, M. Kruhffer, J. L. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. F. Orntoft, "Identifying distinct classes of bladder carcinoma using microarrays." *Nat Genet*, vol. 33, no. 1, pp. 90–96, Jan 2003. [Online]. Available: http://dx.doi.org/10.1038/ng1061

[12] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science*, vol. 286, no. 5439, pp. 531–537, Oct 1999.

[13] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma." *Cancer Res*, vol. 62, no. 17, pp. 4963–4967, Sep 2002.

[14] P. Laiho, A. Kokko, S. Vanharanta, R. Salovaara, H. Sammalkorpi, H. Jrvinen, J.-P. Mecklin, T. J. Karttunen, K. Tuppurainen, V. Davalos, S. Schwartz, D. Arango, M. J. Mkinen, and L. A. Aaltonen, "Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis." *Oncogene*, vol. 26, no. 2, pp. 312–320, Jan 2007. [Online]. Available: http://dx.doi.org/10.1038/sj.onc.1209778

[15] C. L. Nutt, D. R. Mani, R. A. Betensky, P. Tamayo, J. G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, T. T. Batchelor, P. M. Black, A. von Deimling, S. L. Pomeroy, T. R. Golub, and D. N. Louis, "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification." *Cancer Res*, vol. 63, no. 7, pp. 1602–1607, Apr 2003.

[16] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression." *Nature*, vol. 415, no. 6870, pp. 436–442, Jan 2002. [Online]. Available: http://dx.doi.org/10.1038/415436a

[17] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures." *Proc Natl Acad Sci U S A*, vol. 98, no. 26, pp. 15 149–15 154, Dec 2001. [Online]. Available: http://dx.doi.org/10.1073/pnas.211566398

[18] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub, "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." *Nat Med*, vol. 8, no. 1, pp. 68–74, Jan 2002. [Online]. Available: http://dx.doi.org/10.1038/nm0102-68

[19] A. I. Su, J. B. Welsh, L. M. Sapinoso, S. G. Kern, P. Dimitrov, H. Lapp, P. G. Schultz, S. M. Powell, C. A. Moskaluk, H. F. Frierson, and G. M. Hampton, "Molecular classification of human carcinomas by use of gene expression signatures." *Cancer Res*, vol. 61, no. 20, pp. 7388–7393, Oct 2001.

[20] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles." *Proc Natl Acad Sci U S A*, vol. 98, no. 20, pp. 11 462–11 467, Sep 2001. [Online]. Available: http://dx.doi.org/10.1073/pnas.201162998

[21] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling." *Cancer Cell*, vol. 1, no. 2, pp. 133–143, Mar 2002.

[22] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.

[23] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak, "Molecular classification of cutaneous malignant melanoma by gene expression profiling." *Nature*, vol. 406, no. 6795, pp. 536–540, Aug 2000. [Online]. Available: http://dx.doi.org/10.1038/35020115

[24] J. Lapointe, C. Li, J. P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, P. Ekman, A. M. DeMarzo, R. Tibshirani, D. Botstein, P. O. Brown, J. D. Brooks, and J. R. Pollack, "Gene expression profiling identifies clinically relevant subtypes of prostate cancer." *Proc Natl Acad Sci U S A*, vol. 101, no. 3, pp. 811–816, Jan 2004. [Online]. Available: http://dx.doi.org/10.1073/pnas.0304146101

[25] K. Stegmaier, K. N. Ross, S. A. Colavito, S. OMalley, B. R. Stockwell, and T. R. Golub, "Gene expression-based high-throughput screening(ge-hts) and application to leukemia differentiation," *Nature Genetics*, vol. 36, no. 3, pp. 257–263, 2004.

[26] I. G. Costa, F. A. D. Carvalho, and M. C. P. D. Souto, "Comparative analysis of clustering methods for gene expression time course data," *Genetics and Molecular Biology*, vol. 27, no. 4, pp. 623 – 631, 2004. [Online]. Available: http://www.scielo.br/pdf/gmb/v27n4/22434.pdf

[27] S. Datta and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data," *Bioinformatics*, vol. 19, pp. 459–466, 2003.

[28] L. Ertoz, M. Steinbach, and V. Kumar, "A new shared nearest neighbor clustering algorithm and its applications," in *Workshop on Clustering High Dimensional Data and its Applications*, 2002, pp. 105–115.

[29] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martnez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu, "Multiple-laboratory comparison of microarray platforms." *Nat Methods*, vol. 2, no. 5, pp. 345–350, May 2005. [Online]. Available: http://dx.doi.org/10.1038/nmeth756