E-CAST: A Data Mining Algorithm For Gene Expression Data

Abdelghani Bellaachia and David Portnoy*
The George Washington University
Department of Computer Science
801 22nd St NW
Washington, DC 20052

Yidong Chen and Abdel. G. Elkahloun
NIH/NHGRI/CGB
National Institute of Health
Bethesda, MD 20892-4470

All correspondences should be sent to Dr. Abdelghani Bellaachia at:

E-mail: bella@seas.gwu.edu
Phone: (202) 994-8166

*** This work was supported by National Institute of Health (NIH).**

Abstract
Data clustering methods have been proven to be a successful data mining technique in the analysis of gene expression data. The Cluster affinity search technique (CAST) developed by Ben-Dor, et. al., 1999, which has been shown to cluster gene expression data well, has two drawbacks. First, the algorithm uses a fixed initial threshold value to start the clustering. As stated in the original paper, this parameter directly affects the size and number of clusters produced. Second, the algorithm requires a final cleaning step, which takes $O(n^2)$, to relocate n data points among the existing clusters.

In this paper, we have developed and enhanced CAST algorithm, called E-CAST, that uses a dynamic threshold. The threshold value is computed at the beginning of each new cluster.

We have implemented both CAST and E-CAST algorithms and tested their performance using three different data sets. The datasets are real gene expression data from melanoma, pheochromocytoma and brain cell tissue samples generated using micro-arrays technology. The results of both implementations were compared to the output from the hierarchical clustering program, written by Michael Eisen, with very comparable results. Not only did the final results compare favorably with the hierarchical approach, but they also indicate that the cleaning step of the original CAST algorithm may be unnecessary.

Keywords: Clustering, Data mining, Bio-informatics, Gene expression, Graph theory, Micro-array.

1. Introduction

Several data mining solutions have been presented for Bioinformatics [ALTM02], [ZHAN02], [KANK02], [THIM95], and [WWW01]. Clustering analysis has received significant attention in the area of gene expression. It allows the identification of the structure of a data set, i.e. the identification of groups of similar objects in multidimensional space. Clustering procedures yield a data description in terms of clusters or groups of data points that possess strong internal similarities. Some possible methods of clustering are:

Divisive or Partitional Clustering: these methods start with each point as a part of a random or guessed cluster and iteratively move points between clusters until some local minimum is found with respect to some distance metric between each point and the center of the cluster it belongs to.

Hierarchical Clustering: These methods start with each point being considered a cluster and recursively combine pairs of clusters (subsequently updating the inter-cluster distances) until all points are part of one hierarchically constructed cluster.

Graph Theoretic Methods: These methods are partitioning methods that partition the space into sub-graphs with respect to some geometric properties.

In this paper we study a graph-based clustering algorithm that uses a divisive approach. This approach uses top-down analysis. It starts with a large cluster and split into smaller ones until each sub-cluster contains only one data point. Examples of this approach are the Two-way clustering binary tree [ALON99], the Coupled Two-way clustering [GETZ00], the Cluster affinity search technique (CAST) [Ben99], and the Gene shaving [HAST00].

In bio-informatics, these techniques are used to analyze data expression generated using micro-array technologies. They try to identify groups of genes that have similar patterns. For instance, the genes that govern chromosome function or meiosis may be more tightly linked to each other than to the genes involved with another function, such as apoptosis [WWW01].

Formally, a set of genes can be viewed as a set of vectors $V = \{v_1, v_2, v_3, \ldots, v_n\}$ with each expression level of a given experiment, x_i, being the components in the vector $v_i = (x_1, x_2, x_3, \ldots, x_n)$, where m is the number of genes in the experiments and n is the number of experiments. (This works equally well when the experiments form the vectors.) These vectors can then be viewed as points in n dimensional space and a similarity measurement between points can be calculated and stored in a m-by-m similarity matrix M. Where M_{ij} is the

distance (similarity) measure between gene i and gene j. There are several similarity measures, e.g., Euclidean distance and Pearson correlation. Then one of many algorithms used for clustering is run on the similarity matrix to group the members of V into clusters, which attempts to maximize the intra-cluster similarity and minimize the inter-cluster similarity.

Several clustering techniques have been applied to the analysis of gene expression data. But each method has shortcomings. These shortcomings include problems of cluster boundaries, as for Hierarchical techniques, where the output is a tree depicting the relation of each object to every other object in the dataset. And, the requirement for knowing the number of expected clusters, as for Self-Organizing Maps.

The Cluster affinity search technique (CAST) developed by Ben-Dor, et. al., 1999, appears to address both of these issues [Ben99]. Its output is of discrete clusters, as apposed to the dendrogram produced by hierarchical methods, and the number of clusters produced does not have to be predetermined. But, it requires an input parameter, which ultimately determines the size and number of clusters produced. This parameter, the so-called, *affinity threshold*, is used to determent what the minimum similarity between an object and a cluster is required for that object to be a member. Leading us back the to the problem of requiring information about the clusters before we can cluster them.

In this paper we discuss a threshold assignment function that can be used to determine the threshold parameter based solely on the data to be clustered. Further, we show that using the threshold function before each new cluster is formed obviates the need for a step in the CAST algorithm with has a time complexity on the order of $O(n^2)$.

The paper is organized as follows. The next section reviews related work. Section 3 presents the CAST algorithm and our enhanced version of CAST, E-CAST. Our clustering results and the data sets we have used are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related Work

Several clustering techniques have been studied for analyzing expression data [HEDE01], [GETZ00], and [EISE98]. Hedenfalk *et al.*, 2001, used an agglomerative hierarchical clustering algorithm to investigate any relation among discriminator genes in hereditary breast cancer [HEDE01]. The discriminator genes were found using three methods, the modified F tests and t-tests, a weighted gene analysis, and mutual-information scoring (InfoScore). They concluded that gene-expression technology can increase the specificity of the molecular classification of breast cancer.

Bittner *et al.*, 2000, used an average linkage hierarchical clustering algorithm with a dissimilarity measure of one minus the Pearson correlation in combination with multidimensional scaling (MDS) plots to successfully cluster and predict the spreading and migration of closely related malignant melanomas [BITT00]. A non-hierarchical clustering algorithm, CAST , was used to define experimental clusters [Ben99]. They found that the classification of melanoma on the basis of gene expression pattern is possible and believe that their identification of genes 'weighted' for their ability to discriminate a subset of melanomas should provide a sound basis for the dissection of other subsets of the melanoma tumor.

Ben-Dor *et al.*, 2000, compared three classification techniques, Nearest Neighbor, Support Vector Machines (SVM) (Cortes *et al.*, 1995), AdaBoost (Freund et al., 1997), and Cluster Affinity Search Technique (CAST) [Ben99], in the classification of two sets of tumor and normal clinical samples, which consisted of 62 colon samples and 32 ovarian samples [Ben00]. Their results indicated that clustering using CAST can be very useful in the classification of cancers. It also highlights the complications in classifications due to the fact that clinical samples are likely to contain a mixture of cells and that there is an inherent genomic instability in tumor samples which may lead to random fluctuations in the gene expression patterns.

Alon *et al.*, 1999, used a two-way deterministic-annealing algorithm to separate cancerous from non-cancerous tissues [ALON99]. The data set was composed of 40 colon tumor samples and 22 normal colon tissue samples and was analyzed with an oligonucleotide array. The data-clustering algorithm was used to build a dendrogram. Each gene, k, was represented by a vector, $v_k = (x_1, x_2, x_3,\ldots, x_n)$, whose components, x_n, corresponded to expression levels in each sample. The vectors were then normalized such that the sum over their components equaled zero, $\sum_m x_m = 0$, and the magnitude equaled one, $|v_k| = 1$. The clusters were split into two groups by first defining two cluster centroids, C_j, where j= 1, 2. A probability of belonging to each cluster was then determined for each gene:

$$P_j(v_k) = \frac{e^{(-b|v_k - C_j|^2)}}{\sum_j e^{(-b|v_k - C_j|^2)}}$$

The cluster centroids were determined by the equation:

$$C_j = \frac{\sum_k v_k P_j(v_k)}{\sum_k P_j(v_k)}$$

which was solved by iterations. For $\beta = 0$ there is only one cluster $C_1 = C_2$. β was increased in small steps until two distinct, converged centroids were formed. Each gene was then assigned to a cluster with the larger $P_j(v_k)$. The process was then repeated to split each one of the new clusters. The algorithm was then run against the tissue samples, where each tissue sample, k, was represented by the vector, v_k. They found that gene grouping can be achieved on the basis of variation between tissue samples from different individuals and that displaying the data with both samples and gene clustered revealed wide-scale patterns that hint at an extensive underling organization of gene expression.

Getz *et al.*, 2000, used a coupled two-way clustering (CTWC) algorithm on colon cancer and leukemia [GETZ00]. And, were able to discover partitions and correlation that were masked and hidden when the full data set was used in the analysis, by identifying relevant gene and sample subsets and focusing on them. The method stems from the idea that only a limited number of genes in each experiment have any useful information. And, the same may be true for the samples. So one should look for stable submatrices within the expression matrix to find the genes and or samples to be studied. Because the time complexity of a brute force implementation of this analysis would be too great, they developed a iterative heuristic. The iterative process is initialized with the full

matrix – i.e. the set of all genes (g_0) and of all samples (s_0) are used as (both) features and objects, to perform standard two-way clustering. The stable clusters of the genes and samples found in the first step are denoted by $g_{1,i}$ and $s_{1,j}$. The process is repeated until some criteria (such as stability or critical size) are reached. Any clustering algorithm can be used in combination with CTWC to find stable clusters. The study used a hierarchical algorithm, SPC (Blatt *et al*., 1996).

3. Enhanced CAST Algorithm

The Cluster Affinity Search Technique, or CAST, clustering method takes a graph theoretic approach that relies on the concept of a clique graph and uses a divisive clustering approach. A clique graph is an undirected graph that is the union of disjoint complete graphs. Thus, the model assumes that there is a "true biological partition of the genes into disjoint clusters bases on the functionality of the genes" [Ben99]. The clique graph would then be composed of clusters (cliques) of genes (vertices) whose interconnections (edges) are present or not present corresponding to their respective similarity measures (i.e. if two genes are similar there is an edge between them). So, ideally, the genes would form sub-graphs (cliques) where every gene would be completely similar to every other gene in the clique and completely dissimilar to every gene not in the clique. Thereby, producing a clique graph G of $U = \{u_1, u_2, \ldots, u_h\}$ vertices partitioned such that every clique S_i contains edges connecting every vertex $u \in S_i$ to every other $u \in S_i$ and no edges connecting any $u \in S_i$ to any $u \in U \backslash S_i$. This, model can be applied just as easily to experiments instead of genes. Where, the experiments become the vertices and one experiment is linked to another based on the similarity of their respective patterns.

Given that it is very probable that a set of gene (or experiment) vectors will tend to have a similarity gradient across other vectors and the high incidence rate of errors in micro-array technology, the ideal clique graph would be impossible to generate, or, at the very lease, would create very small clusters. So small, in fact, that many would contain single data points, and therefore defeat the purpose of the algorithm. Thus, an approximation of the preceding model is called for.

Accepting the extremely remote possibility of disjoint partitions CAST tries to approximate the model by just striving to maximizing the intra-cluster edges and minimizing the inter-cluster edges.

The CAST and E_CAST algorithms take as input an *n*-by-*n* similarity matrix S where $(S(i, j) \in [1, 0])$ and an affinity threshold T is defined. T is used to determine node membership to a cluster.

Definition 1: The affinity of a node x to a cluster C is defined as follows:

$$a(x) = \sum_{k \in C} S(x,k)$$

Definition 2: The connectivity threshold, χ, of a cluster C is:

$\chi = T|C|$ where $|C|$ is the cardinality of C.

Definition 3: A high connectivity node is a node that will be included in a cluster. Its affinity satisfies the following:

$a(i) \geq \chi$ where a(i) is the affinity of i.

Definition 4: A low connectivity node is a node that will be removed from a cluster. Its affinity satisfies the following:

$a(i) < \chi$ where a(i) is the affinity of i.

Each cluster is formed by alternating between adding and removing nodes from the current cluster until such time that changes no longer occur or a maximum of iterations has been executed:

Node Addition:Add nodes with high connectivity to the nodes in the open cluster.

Node Removal: Remove any nodes in the open cluster with low connectivity to the other nodes in the cluster.

Cluster Cleaning: Make sure all nodes are in clusters with highest affinity.

CAST algorithm relies on the *affinity* threshold, T, being an input variable defined by the user before initiating the clustering process. This is a problem because the size and quantity of the clusters produced by the algorithm is directly affected by this parameter [Ben99]. Implying that some knowledge of the data set is required before the clustering can be performed. We have enhanced the algorithm to calculate this threshold. Further, the threshold can be calculated dynamically based only on the objects in that have yet to be assigned a cluster, $U' = U \backslash (C_0 \cup C_1 \cup \ldots \cup C_n)$, before each cluster is created. Thus, providing a means of fine-tuning while clusters are formed. The threshold parameter, T, is calculated based on the similarity values of the nodes left to be clustered. This dynamic threshold is computed as follows:

$$T = \left(\frac{\sum_{i,j \in U' \text{ and } S(i,j) \geq 0.5} S(i,j) - 0.5}{\left| \{ u : u \in U' \text{ and } a(u) \geq 0.5 \} \right|} \right) + 0.5$$

Currently, we have shown very good performance using the average scaled (0 to 1) similarity values above 0.5, on a number of data sets. Varying methods can be used to determine T and is one area we would like to further research.

The following provides pseudo-code for both CAST and E_CAST algorithms:

Threshold:
// T is an input parameter
CAST:
T= fixed value (for example 0.76)

// executed before each new C_{open} is created
E-CAST:
a = 0;
count = 0;
for all $u \in U$ such that $a(u) \geq 0.5$ {
  a += a(u)-0.5
  count++
}
T = (a/count) + 0.5

Cluster Formation:
while (U <> ∅){
  E-CAST: Calculate Threshold, T
  for all $u \in U$ set a(u) = 0
  create empty cluster C_{open}

```
  Pick an element u ∈ U such that S(u,x)=max{S(w,x)|w and
    x ∈ U}
  C_open = C_open ∪ u
  U = U \ u
  For all x ∈ U set a(x) = a(x) + S(x,u)
  while (changes in C_open occur) or (iterations < max
    iterations){
    //Addition Step
    while max{a(w)|w ∈ U} ≥ χ {
        Pick an element u ∈ U such that a(u)=max{a(w)|w ∈
          U}
        C_open ← C_open ∪ {u}
        U ← U \ {u}
        // Update affinity of all nodes
        For all x ∈ U ∪ C_open set a(x) = a(x) + S(x,u)
    }
    //Removal Step
    while min{a(w)|w ∈ C_open} < χ {
        Pick an element u ∈ C_open such that a(u)=min{a(w)|w
          ∈ C_open }
        C_open ← C_open \ {u}
        U ← U ∪ {u}
        // Update affinity of all nodes
        For all x ∈ U ∪ C_open set a(x) = a(x) - S(x,u)
    }
  }
}
```

Cleaning Step:

```
while (changes in any C_i occur) or (iterations < max
  iterations){ // cleaning step may not converge
  for each c ∈ C_i and C_i ∈ C and C_j ∈ C{
    Compute a normalized affinity of c to each cluster C_j such
      that a_j(c)= (∑_{k∈Cj} S(c,k))/(|C_j|)
  }

  if max{ a_j(c) } > a_i , for all C_j ∈ C and i ≠ j {
    C_i = C_i \ c
    C_j = C_j ∪ c
    }
}
```

The dynamic threshold assignment has been shown by our results to obviate the need for the "cleaning" step as proposed in the original algorithm. The cleaning step is used to move any vector from its current cluster to one that it may have a higher affinity for and has a time complexity on the order of $O(n^2)$.

4. Experiments

Three real gene expression data sets were used to compare the performance of E-CAST, CAST, and an average linkage hierarchical clusterer. The dataset will be referred to as melanoma, Thanhall and brain. The melanoma dataset [Bitt00] consists of expression for 38 samples, including 31 melanomas and 7 controls. Each profile consists of the expression levels of 3,614 cDNA clones. Their conclusions relied on, in part, a dendrogram generated using an average linkage hierarchical clustering algorithm. They found a major cluster consisting of 19 samples and, what was considered, a non-clustered group of 12. This information was then used to aid the prediction that the tumors represented by the samples in the major cluster would have reduced motility and reduced invasive ability as compared to the melanomas outside the this cluster. The prediction was verified using a series of cellular assays.

The Thanhalldataset is used for the first time for clustering analysis and has not been published yet. Expression profiles consist of 12,024 cDNA clones, for pheochromocytoma tumors. Pheochromocytomas are rare but clinically important tumors of chromaffin cells that arise typically in the adrenal gland and constitute a surgically correctable cause of hypertension. Acute catecholamine release by a pheochromocytoma not only can lead to malignant hypertension but also lethal arrhythmias, heart failure, myocardial infarction and sudden death. Most pheochromocytomas are sporadic, but some are familial, associated with von Hippel-Lindau(VHL) disease, multiple endocrine neoplasia type 2 (MEN2), or neurofibramatosis 1 (NF1). Genes for these hereditary conditions have been identified. Somatic mutations of the same genes, or different, as yet unidentified genes may underlie sporadic pheochromocytoma. Pheochromocytomas that develop sporadically or in patients with hereditary predispositions differ in terms of their rate of growth, likelihood of recurrence, malignant potential, and catecholamine phenotype. The molecular mechanisms by which genotypic changes predispose to development of pheochromocytoma and the bases for the variable clinical presentation and course of the tumor remain unknown. About 10-20% of pheochromocytomas are malignant, with life expectancy 3-5 years. There is no known effective treatment for malignant pheochromocytoma. Out of the 22 samples included in this data set, 17 were predicted to be placed into 3 clusters.

The third data set (not yet published), includes samples that consisted of brain cells and bone marrow T-cells treated as to become brain cells. Expression profiles consist of 12,024 cDNA clones. Out of the 22 samples, 20 were predicted to be placed into 6 clusters.

All samples for each of the datasets where compared to the same universal RNA source. The calibrated ratios represent the mean sample intensity over the mean intensity of the reference RNA.

Each data set was clustered using E-CAST and CAST (with the threshold, T, set to the value calculated by E-CAST before the first cluster was formed). NIH/NHGRI's GeneCluster tool was used to generate the dendrograms for the Thanhall and brain datasets, while the dendrogram published in Bittner, M., *et. al.*, 2000 was used for the melanoma dataset.

The pheochromocytoma samples were taken from tumors with know mutations, therefore it was possible to predict the cluster formations. Predictions of cluster formation for the brain cell data also could be made based on treatment type and cell line. For these two datasets, comparisons were made between the predicted clusters and the results generated by the three clustering algorithms. The performance was measured in misplaced experiments and misjoined clusters**.** Misplaced experiments refer to the number of experiments that were placed into a cluster for which they were not predicted to belong. While, misjoined clusters refer to the number of merges of clusters which were predicted to be disjoint.

Because no cluster formation was anticipated for the melanoma dataset, misplaced experiments and misjoined clusters were measured from the major cluster of 19 samples as described. This appears to make the hierarchical approach 100% accurate, but the validity of the major cluster's components has not, as yet, been proven.

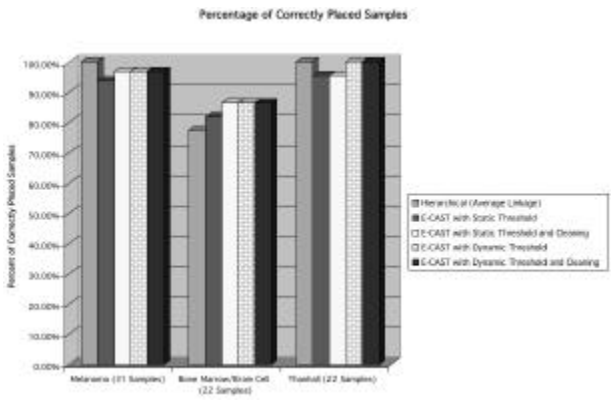

Figure 1 "Percentage of Correctly Placed Samples"

Figure 1 shows the percentage of samples that were placed into their predicated clusters. As can be seen, E-CAST with dynamic threshold assignment without the cleaning step consistently outperforms static assignment. Cluster membership under E-CAST is not changed using the additional cleaning step. Figure 1 also shows that E-CAST equals or out performs the hieratical method for the BoneMarrow/Brain Cell and Thanhall datasets. The hieratical method performs slightly better than E-CAST for the Melanoma since the clusters were defined using the hierarchical method.

Dynamic and Static Threshold Assignment Performance

		Before Cleaning Step		After Cleaning Step	
Data Set	Threshold	Misplaced Exp.	Misjoined Clusters	Misplaced Exp.	Misjoined Clusters
Brain	Dynamic	3	2	No Cluster Change	No Cluster Change
20 exp. 6 clusters	T = 0.67	4	2	4	2
Melanoma	Dynamic	1	0	No Cluster Change	No Cluster Change
25 exp. 3 clusters	T = 0.61	2	0	1	0
Thanhall	Dynamic	0	0	No Cluster Change	No Cluster Change
17 exp. 3 clusters	T = 0.66	1	1	No Cluster Change	No Cluster Change

Table 1 "Dynamic and Static Threshold Assignment Performance"

Table 1 shows the number of misplaced experiment and misjoind clusters for the three datasets using static and dynamic threshold assignment with, and without the cleaning step. As can be seen the cleaning step never changes the cluster formation while using dynamic assignment, but does for static. In addition, it can be clearly seen that the performance of dynamic threshold assignment is consistently better than static. In only one case did the cleaning step improve the clusters created by the static method. And, these modifications did not improve performance over the dynamic method without cleaning.

5. Conclusion

In this paper we presented a data mining algorithm for the analysis of gene expression data. Cluster affinity search technique (CAST) has been successfully used in clustering gene expression data. However, CAST has two main drawbacks: (1) the threshold for clustering is fixed and assigned a value before clustering starts and (2) a very expensive cleaning step. We have introduced an enhanced CAST (E-CAST) algorithm that uses a dynamic threshold. This threshold is computed at the creation of each cluster. Our experimental results show that E-CAST does not necessarily requires the cleaning step.

Three real datasets were used to evaluate our algorithm against the original CAST algorithm and the Eisen's hierarchical algorithm. The first dataset is a melanoma set that has 38 samples, including 31 melanomas and 7 controls. This set has been previously clustered by other authors.

The second set is a Thanhall dataset that we are used for the first time and it has not been published yet. The set has 22 samples out of which 17 were predicted to be placed into 3 clusters.

The third data set (not yet published), includes samples that consist of brain cells and bone marrow T-cells treated as to become brain cells. This set has 22 samples out of which 20 were predicted to be placed into 6 clusters.

Our results show that E-CAST performs better than the original algorithm. The results also confirm that the cleaning step may not be required and thus an improvement in the overall clustering performance. The dynamic computation of the threshold indicates great promise for using this technique to glean information from gene expression profiles. We have also clustered these data sets using Eisen's hierarchical algorithm. Overall, E-CAST has shown better performance than the hierarchical algorithm.

Future work includes theoretical analysis of the determination of the threshold parameter. We are also investigating the performance of our E-CAST algorithm on labeled clusters.

Acknowledgement:
We would like to thank Dr. Graeme Eisenhofer at NIH for sharing with us the Thanhall dataset.

References

[ALON99] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. & Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucle otide arrays, *Proceedings of the National Academy of Sciences*

[Ben99] Ben-Dor, A., Shamir, R. & Yakhini, Z. (1999). Clustering gene expression patterns, *Journal of Computational Biology*.

[HAST00] Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D. & Brown, P. (2000). "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biology*

[GETZ00] Getz, G., Levine, E. & Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data, *Proceedings of the National Academy of Sciences*

[WWW01] http://www.wiley.co.uk/wileychi/eob/genetics/Pab004.pdf(visited on May 10, 2002).

[ALTM02] R. Altman et al., "Text Mining for Bioinformatics," Second SIAM International Conference on Data Mining, Arlington Virginia, 2002.

[ZHAN02] Li Zhang, Chun Tang, Yong Shi, Yuqing Song, Aidong Zhang, Murali Ramanathan, "IVADA: An Interactive Visualization Approach to Data Analysis and Its Application on Microarray Data," Second SIAM International Conference on Data Mining, Arlington Virginia, 2002.

[KANK02] P. Kankar et al. “MedMeSH Summarizer: Text Mining for Gene Clusters
Second SIAM International Conference on Data Mining, Arlington Virginia, 2002.

[THIM95] Thimothy J. Ross, *Fuzzy Logic with Engineering Applications*, Mc Graw Hill, Inc, 1995.

[Ben00] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. & Yakhini, Z., Tissue classification with gene expression profiles. Journal of Computational Biology 2000, vol 7, 559-584.

[BITT00] Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N. & Trent, J. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature August 2000, vol 406, 536-540.

[EISE98] Eisen, M., Spellman, P., Brown, P. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA December 1998, vol 95, 14863-14868.