Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Biologically Supervised Hierarchical Clustering Algorithms for Gene Expression Data

Grzegorz M Boratyn, *Member IEEE*, Susmita Datta, and Somnath Datta

***Abstract*— Cluster analysis has become a standard part of gene expression analysis. In this paper, we propose a novel semi-supervised approach that offers the same flexibility as that of a hierarchical clustering. Yet it utilizes, along with the experimental gene expression data, common biological information about different genes that is being complied at various public, web accessible databases. We argue that such an approach is inherently superior than the standard unsupervised approach of grouping genes based on expression data alone.**

It is shown that our biologically supervised methods produce better clustering results than the corresponding unsupervised methods as judged by the distance from the model temporal profiles.

R-codes of the clustering algorithm are available from the authors upon request.

I. INTRODUCTION

We introduce a novel hierarchical clustering algorithm for grouping genes based on gene expression data. Unlike other approaches to clustering gene expressions, it is not an unsupervised method. We characterize it as "semi-supervised" since it uses a "training set" of annotated genes whose functional information is known. This approach produces a hierarchy of clusters in either an agglomerative or a divisive manner. However, at each stage of cluster formation, the algorithm uses a distance measure based on the gene expression profiles plus the biological information obtainable from the public GO databases. It is argued that such an approach is inherently superior than the standard unsupervised approach in producing biologically meaningful clusters.

A. Motivation

Cluster analysis is routinely used in analyzing gene expression data. Typically, a standard hierarchical clustering algorithm such as the UPGMA with correlation similarity measure is used. A post hoc analysis is done to identify each cluster by associated biological functions often using a handful of genes with known biological behavior. As pointed out in earlier works [1], [2], [3], the result of such an approach is sensitive to the choice of the clustering algorithm used.

B. Related work

Numerous approaches of cluster validations are available in the literature. As for example, Figure of Merit (FOM) measures [4] such as the silhouette width [5] or the homogeneity index [6] can be used to evaluate the external (visual) characteristics of the results of a clustering algorithm. Indices to measure the stability of a clustering algorithm were introduced in [3], [7]. Analysis of biological validity of clustering results through computation of distance from a model profile was presented in [3]. A resampling based validity scheme was proposed in [8]. Use of GO databases in validating the results of an unsupervised method are available in a number of recent papers including [9], [10], [11], [12]. Utilization of GO terms to improve biological relevance of clustering results is considered in [13]. Semi-supervised clustering methodologies for general applications are presented in [14], [15].

The next section introduces the presented clustering algorithms. Experimental validation of the proposed method along with description of utilized data sets of gene expression, and experimental results are described in Section 3. Conclusions and discussion are presented in Section 4.

II. BIOLOGICALLY SUPERVISED CLUSTERING

Let $G = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ be the set of all gene expressions resulting from a microarray experiment, such that $\mathbf{x}_g \in R^p$, for some p. Let also $F_1, F_2, \ldots, F_m$ be not necessarily disjoint sets of labels corresponding to genes with similar functions. We propose such semi-supervised clustering algorithms that utilize this prior functional information and promote clusters of functionally similar genes.

A. BSC algorithms

The Biologically Supervised Clustering (BSC) algorithms are novel clustering techniques that take set of gene expressions G and set of functional classes $\mathcal{F} = \bigcup_{k=1}^{m} F_k$ as the input and produce a hierarchy of clusters as a result. The crucial part of the presented method is the distance metric that combines measurements (gene expressions) and prior information (functional sets). The distance $D(A, B)$ between two clusters A and B is composed of two parts: 1) the mathematical distance $d_M(A, B)$ computed with the gene expressions, 2) and biological distance $d_B(A, B)$ that is based on the prior biological functional information:

$$D(A, B) = (1 - \lambda) d_M(A, B) + \lambda d_B(A, B), \qquad (1)$$

where $\lambda \in [0, 1]$ is a user-specified coefficient, representing the relative importance of the components. Consider two genes with expression levels $\mathbf{x}_g$ and $\mathbf{x}_{g'}$, $g \neq g'$ belonging to two different clusters. The mathematical distance is the distance between each pair of gene expressions, that belong

G. M. Boratyn is with the Kidney Disease Program and Clinical Proteomics Center, University of Louisville, Louisville, KY (corresponding author, e-mail: greg.boratyn@louisville.edu)

S. Datta and S. Datta are with Faculty of Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY

1-4244-0033-3/06/$20.00 ©2006 IEEE.

to different clusters, normalized by the number of elements in the clusters:

$$d_M(A,B) = \frac{1}{\mathrm{n}(A)\mathrm{n}(B)} \sum_{\mathbf{x}_g \in A,\ \mathbf{x}_{g'} \in B} d(\mathbf{x}_g, \mathbf{x}_{g'}), \quad (2)$$

where $d(\cdot,\cdot)$ is a distance (or dissimilarity) measure, and $\mathrm{n}(\cdot)$ is the cardinality of a set. On the other hand, the biological distance is found by counting all such pairs of genes whose expressions belong to different statistical clusters and labels do not belong to the same functional set, normalized by the number of genes in each statistical cluster for whose functional information is known. Thus for $\mathrm{n}(A \cap \mathcal{F})\mathrm{n}(B \cap \mathcal{F}) > 0$:

$$d_B(A,B) = \frac{1}{\mathrm{n}(A \cap \mathcal{F})\mathrm{n}(B \cap \mathcal{F})} \cdot \cdot \sum_{g \in A \cap \mathcal{F},\ g' \in B \cap \mathcal{F}} (1 - \mathrm{I}(g, g' \in F_k, \text{ for some } k)), \quad (3)$$

where $\mathrm{I}(\cdot)$ is an indicator of the logical value. We assume that for $\mathrm{n}(A \cap \mathcal{F})\mathrm{n}(B \cap \mathcal{F}) = 0$, $d_B(A,B) = 0$.

We consider two versions of the BSC algorithm. It could be either agglomerative (BSAC) or divisive (BSDC) hierarchical clustering. In the first case, at the initial level all genes form their own clusters. At each subsequent levels two least distant clusters are combined to form one bigger cluster. A hierarchy of clusters is produced as the result, as in case of hierarchical clustering (UPGMA).

On the other hand, the BSDC algorithm starts with one large cluster that contains all genes. At each subsequent stage the cluster that contains two most distant observations is selected for division. The observation that has the largest average dissimilarity from all other observations in the cluster forms so-called *splinter group*. All genes with smaller average distance to the splitter group than to the old cluster are assigned to the new cluster, as in the case of the Divisive Analysis method (DIANA).

B. Selection of parameters

The performance of a BSC algorithm depends on the value of λ. This parameter controls the degree of influence of the prior functional information in the clustering process. Note that for $\lambda = 0$, the functional information is not considered for constructing clusters and the BSC algorithm behaves as UPGMA or DIANA. With $\lambda = 1$ only prior biological information is utilized for cluster construction, and λ between 0 and 1 causes combination of gene expressions and functional information with various weights. The optimal value of λ should reflect the degree of confidence in the gene expression measurements, prior biological information and how well the annotated genes represent all genes in the study. Because the this confidence is difficult to assess, we assume that the optimal λ corresponds to the optimal performance of a BSC algorithm.

We propose to apply the BSC algorithm with several values of λ and select the value that yields the optimal effectiveness in terms of some performance measure. An example of such measure along with experimental results are provided in the next section.

III. Experimental Results

We illustrate our clustering methods on two very different data sets, described below.

A. Data Sets

1) Yeast time course cDNA data: As an illustrative data set, we use the classical data set collected by Chu et al. and presented in [16]. This data set records expression profiles during sporulation of *Saccharomyces cerevisiae* at seven time points. The original data set was filtered using the same criterion as in [16]. For our illustration, we look at a further subset of 513 genes (ORF's to be correct) that were overall positively expressed (i.e., $\sum_{time} \log$ expression ratio > 0).

Functional classes were obtained using the web-based GO mining tool at http://mips.gsf.de/proj/funcatDB/search_main_frame.html. Overall, 503 of the 513 genes were annotated into the following seventeen functional classes: metabolism (138), energy (27), cell cycle and DNA processing (152), transcription (50), protein synthesis (10), protein fate (72), protein with binding function or cofactor requirement (81), protein activity regulation (16), transport (63), cell communication (12), defense (36), interaction with environment (33), cell fate (17), development (41), biogenesis (77), cell differentiation (82).

2) Normal versus breast carcinoma, SAGE data: This data set comes from the study presented in [17]. We illustrate our methods using the expression profiles of 258 genes (SAGE tags) that were judged to be significantly differentially expressed at 5% significance level between four normal and seven ductal carcinoma in situ (DCIS) samples.

For constructing the functional classes, we have used a publicly available web-tool called Amigo (http://www.godatabase.org/cgi-bin/amigo/go.cgi) . We were able to annotate 113 SAGE tags into the following eleven functional classes based on their primary biological functions. They were as follows: cell organization and biogenesis (24), transport (7), cell communication (15), cellular metabolism (48), cell cycle (6), cell motility (7), immune response (7), cell death(7), development (5), cell differentiation (5), cell proliferation (5) where the numbers in parentheses were the numbers of SAGE tags in a class. There were 23 genes that belonged to more than one functional class.

B. Performance Measures

The BSC algorithms were implemented with the R programming language. The performance of the proposed clustering method was assessed with the yeast and SAGE data sets described in Sec. III-A. The distance from model profiles, validating biological relevance of resulting clusters, was utilized as performance measure. The following dissimilarity measure:

$$d(\mathbf{x}_g, \mathbf{x}_{g'}) = \frac{1 - r(\mathbf{x}_g, \mathbf{x}_{g'})}{2} \quad (4)$$

was utilized in (2), where $r(\cdot,\cdot)$ is the correlation coefficient.

The distance from model profiles, proposed in [3], measures biological validity of statistical clusters. Model profiles are created from a small group of hand-selected genes that were available from the original studies and classified into biological classes as deemed appropriate by the biologists for that particular experiment. The gene expressions averaged over each class create the model profiles. The averaged gene expressions are calculated for each resulting cluster, and the distance between so created profile and the model profile is computed:

$$dist = \min_{\pi} \sum_{i=1}^{K} d(\bar{\mathbf{x}}_i^m, \bar{\mathbf{x}}_{\pi(i)}), \tag{5}$$

where K is the number of clusters and the minimum is taken over all permutations π of integers $\{1, 2, \ldots, K\}$, and $\bar{\mathbf{x}}_i^m$ is the (average) model profile for the i-th cluster. The rational behind using a permutation was to match the levels of the clusters with those for the model profiles. The expression (4) was also used here as distance metric. Smaller $dist$ indicates that studied clusters are more similar to the model profiles thus more biologically valid.

The distance from model profiles (5) was computed for the yeast data set. In the original paper Chu et al., [16] determined on the basis of first induction of expression that seven is the right number of clusters to be used for grouping genes for this data set. In addition they created a model expression profile by using certain handpicked genes in each class. We use the same number of clusters ($K = 7$) and the benchmark model profile.

It may be worth pointing out that the biological information used in [16] is different from the functional information from GO used in BSC. Thus, (5) serves as a true validation measure.

The resulting distances from the model profiles computed for the yeast data set, clustered with BSAC and BSDC algorithms are presented in Figs. 1 and 2, respectively.

In the case of the BSAC algorithm (Fig. 1), the minima (and hence the optimal values) within the grid of selected λ, occur at 0.5, 0.7 and 0.8. The distance from model profiles produced by the BSDC algorithm (Fig.2) shows greater improvement of the performance measure for $\lambda = 0.4$.

Next, a similar performance measure was computed for the SAGE data set. The model profiles were composed of a small collection of genes reported in [17] (Fig. 5 in reference [17]), whose deregulation is altered in the ductal carcinoma in situ stage of breast cancer. Three model clusters were created from the following functional classes: Cell cycle (3 genes), Apoptosis (3), and Cytokines (4). Due to the small number of model clusters and genes in them, we took one more cluster, e.g., $K = 4$ with the hope of representing genes that may be involved with "other" types of cellular activities. The distance measure (5) was modified so that the smallest distance between model profiles and $K - 1$ cluster profiles

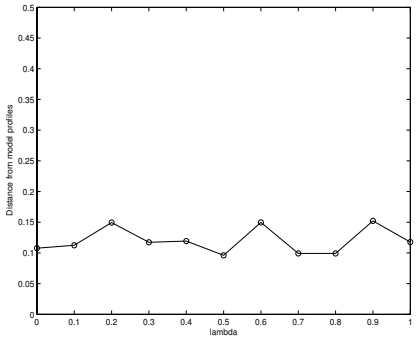

Fig. 1. Distance from model profiles, computed for the yeast data set with set II of functional classes, for various values of λ, with the BSAC algorithm

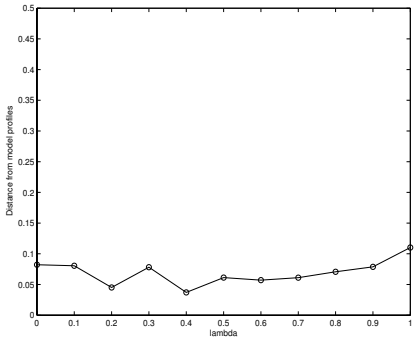

Fig. 2. Distance from model profiles, computed for the yeast data set with set II of functional classes, for various values of λ, with the BSDC algorithm

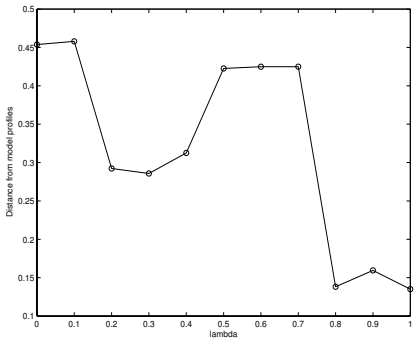

Fig. 3. Distance from model profiles, computed for the SAGE data set, for various values of λ, with the BSAC algorithm

is measured:

$$dist^* = \min_{\pi} \sum_{i=1}^{K-1} d(\bar{\mathbf{x}}_i^m, \bar{\mathbf{x}}_{\pi(i)}), \tag{6}$$

where π denotes the same permutations as in (5). The values of $dist^*$, computed for the SAGE data with BSAC and BSDC methods, for several values of λ are presented in Figs. 3 and 4. For this data set, we notice substantial gain in using our BSC algorithms compared to the standard UPGMA and Diana (which correspond to $\lambda = 0$). Once again, BSDC (Fig. 4) seems to produce slightly better results than BSAC (Fig. 3) for the optimal λ of 0.8 for both clustering methods.

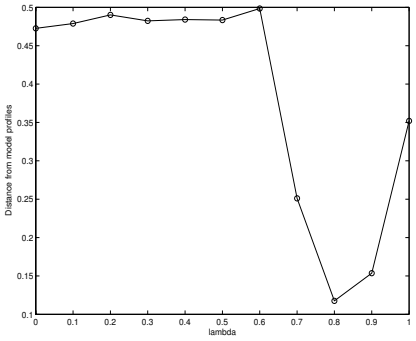

Fig. 4. Distance from model profiles, computed for the SAGE data set, for various values of λ, with the BSDC algorithm

IV. Discussion

The BSC algorithms presented here are novel clustering methods which are based on biological as well as the experimental information. These are also novel additions to the collection of tools that attempt to utilize the growing GO databases. As a clustering algorithm, it maintains the full generality of hierarchical clustering producing a dendrogram or hierarchy of clusters which can be utilized at any specified level (height). It can use a general dissimilarity measure just like the standard hierarchical clustering for the expression component of the "distance". Even though we have used the average distance in computing the mathematical distance between sets, clearly, other choices such as the minimum (single linkage) or the maximum (complete linkage) are possible.

The algorithm can be implemented in either an agglomerative or a divisive fashion. Thus the standard clustering methods UPGMA and Diana can be obtained as special cases of this method.

Although we were motivated by gene expression data, the basic idea is more general. It can be easily adapted to other biological (e.g., proteomics) and non-biological applications where one is interested in clustering but there is additional relevant information. This information could be a like a training set in a classification problem with the following important distinctions. The informative groups could be overlapping and the cluster levels do not have to correspond to or be restricted to the levels in that information set.

References

[1] J. Quackenbush, Computational analysis of microarray data, *Nat. Rev. Genet*, vol. 2, pp. 418-427, 2001.

[2] S. Datta and J. Arnold, Some comparisons of clustering and classification techniques applied to transcriptional profiling data, in *Advances in Statistics, Combinatorics and Related Areas*, C. Gulati, Y-X. Lin, S. Mishra, and J. Rayner (editors), World Scientific, 2002, pp. 63 – 74.

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

[4] K. Yeung, D. R. Haynor, and W. L. Ruzzo, Validating clustering for gene expression data, *Bioinformatics*, vol. 17, pp. 309 – 318, 2001

[5] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Computat. Appl. Math*., vol. 20, pp. 53 – 65, 1987.

[6] R. Shamir and R. Sharan, Algorithmic approaches to clustering gene expression data, *Current Topics in Computational Molecular Biology*, MIT Press, 2002, pp. 269 – 300.

[7] S. Dudoit and J. Fridlyand, A prediction-based resampling method to estimate the number of clusters in a dataset, *Genome Biology*, vol. 3, pp. 0036.1 – 0036.21, 2002.

[8] K. M. Kerr, and G. A. Churchill, Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments, *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 8961 – 8965, 2001.

[9] F. D. Gibbons, and F. P. Roth, Judging the quality of gene expression-based clustering methods using gene annotation, *Genome Research*, vol. 12, pp. 1574 – 1581, 2002.

[10] M. D. Robinson, J. Grigull, N. Mohammad, and T. R. Hughes, FunSpec: a web-based cluster interpreter for yeast, *BMC Bioinformatics*, vol. 3, pp. 1471 – 2105, 2002.

[11] S. G. Lee, J. U. Hur, and Y. S. Kim, A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, vol. 20, pp. 381 – 388, 2004.

[12] S. Datta, and S. Datta, Combining functional information in selecting clustering algorithms, In *Proceedings of Interface 2005*, CD-ROM, 2006.

[13] D. Hunag and W. Pan, Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data, *Bioinfromatics*, vol. 22, pp. 1259 – 1268, 2006.

[14] N. Grira, M. Crucianu, and N. Boujemaa, Unsupervised and semi-supervised clustering: a brief survey, in *A Review of Ma chine Learning Techniques for Processing Multimedia Content*, Report of the MUSCLE European Network of Excellence (FP6), 2004.

[15] S. Basu, Semi-supervised clustering: probabilistic models, algorithms and experiments, Ph.D. Thesis, Department of Computer Sciences, University of Texas at Austin, 2005.

[16] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, I. Herskowitz, The Transcriptional Program of Sporulation in Budding Yeast, *Science*, vol. 282, pp. 699 – 705, 1998.

[17] M. C. Abba, J. A. Drake, K. A. Hawkins, Y. Hu, H. Sun, C. Notcovich, S. Gaddis, A. Sahin, K. Baggerly, and C. M. Aldaz, Transcriptomic changes in human breast cancer progression as determined by serial analysis of gene expression. *Breast Cancer Res*, vol. 6, pp. R499 – R513, 2004.