Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Matching and Visualization of Multiple Overlapping Clusterings of Microarray Data

Chase Krumpelman and Joydeep Ghosh
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX
Email: chasek@gmail.com

***Abstract*— Algorithms have been recently developed for clustering microarray data that allow elements - usually genes - to belong to more than one cluster. The labellings that these algorithms produce are intuitively closer to the reality of biological processes, but are more difficult to analyze by traditional means. In this paper, we introduce an algorithm for aligning the results of overlapping clusterings and for visualizing the results. We demonstrate the utility of the visualization, and provide an example of the application of the alignment technique to constructing an overlapping clustering ensemble.**

I. Introduction

Clustering is a useful unsupervised way of discovering structure in large datasets. Most clustering algorithms in common use assign each data point to just one cluster. The result of this type of clustering is an assignment of 1 of k labels to each of the n points in the dataset. Such a labelling can be represented as a $k \times n$ membership matrix M where each row has a 1 in the kth position corresponding to the point having label k, and zeros elsewhere. Membership matrices that satisfy $\forall n \sum_{i=1}^{k} M(n,k) = 1$ are called $disjoint$ because they represent a disjoint clustering.

In bioinformatics applications, disjoint clusterings often do not adequately capture the expected behavior. It is known from gene-knockout and other assays that individual genes often play roles in several cellular processes; thus, when finding process clusters by clustering genes' behavior across several experimental conditions, one would expect certain genes to have multiple labels. Several novel approaches to clustering microarray data have been developed to find such labellings, e.g. Gene Shaving [1], Model-Based Overlapping Clustering [2], [3], and Plaid [4]. These approaches allow genes to belong to none, one, or many clusters, and therefore the labellings form membership matrices that do not satisfy $\forall n \sum_{i=1}^{k} M(n,k) = 1$. Such membership matrices are called $nondisjoint$.

The adjusted Rand index (ARI) [5] is commonly used to compare two disjoint clusterings, or to compare a disjoint clustering to a ground truth. The adjusted Rand index evaluates pairs of points; a high ARI indicates that most pairs of points that share a label in one clustering also share a label in the other. The state of a pair of points sharing a label is implicitly binary, and therefore the adjusted Rand index can not be used to compare nondisjoint memberships, where a pair can share 0, 1, or more labels. The Omega index [6] generalizes the adjusted Rand index to compare nondisjoint memberships.

When clustering genes into processes with a nondisjoint clustering approach, the Omega index similarity between clusterings or with a ground truth is useful, but provides no information about the correspondence between individual clusters. This paper describes a method for finding and visualizing the best partial correspondence between two overlapping labellings. Label alignment facilitates comparisons among membership matrices, as well as combining membership matrices into ensembles.

In Sections II and III, we describe the alignment and visualization method. In Sections IV, V, and VI we give examples of the application of the method to to simulated microarray data.

II. Aligning Nondisjoint Membership Matrices

To align two binary membership matrices M_A and M_B, we seek a permutation of the columns of M_B that maximizes the similarity between each column of M_A and its corresponding column in M_B. A simple measure such as Hamming distance between the columns seems like a reasonable choice, but it implicitly assumes each column has the same density (ratio of 1's and 0's), which is not a valid assumption. To compensate for variations in density, we assume the 1's are independently distributed and calculate the probability of seeing the observed match given the densities of the binary vectors, v_1 (a column of M_A) and v_2 (a column of M_B):

With

1-4244-0710-9/07/$20.00 ©2007 IEEE

- N as the vector length (total number of data points)
- d_1 as the number of 1's appearing in v_1, i.e. $d_1 = \sum_{i=1}^{N} v_1[i]$
- d_2 as the number of 1's appearing in v_2, i.e. $d_2 = \sum_{i=1}^{N} v_2[i]$
- S is the number of "overlapping" 1's, i.e. $S = \sum_{i=1}^{N} v_1[i] \times v_2[i]$
- the 1's in v_1 and v_2 uniformly distributed

the probability of seeing an observed overlap of S 1's is given by:

$$P(s = S) = \frac{\begin{pmatrix} N \\ d_1 \end{pmatrix} \begin{pmatrix} d_1 \\ S \end{pmatrix} \begin{pmatrix} N - d_1 \\ d_2 - S \end{pmatrix}}{\begin{pmatrix} N \\ d_1 \end{pmatrix} \begin{pmatrix} N \\ d_2 \end{pmatrix}} \qquad (1)$$

where the denominator is the total number of permutations of the two vectors, and the numerator is the number of those permutations that have the observed overlap S:

- $\begin{pmatrix} N \\ d_1 \end{pmatrix}$ counts the number of ways d_1 1's can be placed in a vector of length N;
- $\begin{pmatrix} d_1 \\ S \end{pmatrix}$ counts the number of ways to choose the S overlapping points from the d_1 1's in v_1; and
- $\begin{pmatrix} N - d_1 \\ d_2 - S \end{pmatrix}$ counts the number of ways to place the remaining 1's in vector 1 such that they do not overlap with 1's in v_2.

While the denominator is obviously symmetric in assignment of v_1 and v_2, the numerator is not so clearly symmetric. Expansion of the binomial in Equation 1 coefficients makes it clear that the numerator is symmetric:

$$\frac{N!}{(d_1 - S)!S!(N - d_1 - d_2 + S)!(d_2 - S)!} \qquad (2)$$

Algebraic simplification of Equation 1 yields:

$$P(s = S) = \frac{\begin{pmatrix} d_1 \\ S \end{pmatrix} \begin{pmatrix} N - d_1 \\ d_2 - S \end{pmatrix}}{\begin{pmatrix} N \\ d_2 \end{pmatrix}} \qquad (3)$$

Equation 3 calculates the probability that two binary vectors v_1 and v_2 with densities d_1 and d_2 respectively will have S matched 1's (Equation 3 is the hypergeometric distribution evaluated at $s = S$). If we observe two N-length binary vectors, then

$$p - \text{value} = \sum_{s=S}^{s=\min\{d_1, d_2\}} P(s) \qquad (4)$$

The p-value defined in Equation 4 gives the total probability of seeing the observed overlap (S) or a greater overlap. This value essentially measures the likelihood of the observed overlap being a random event, hence a small p-value indicates a small probability of seeing the observation at random. [1]

Returning to our original goal of aligning the columns of two membership matrices M_1 and M_2, it is clear that if we find that the p-value of matching between column i of M_1 and column j of M_2 is very small, we can surmise that those two columns represent the same cluster. Finding the best possibly global matching of columns is an instance of the *Stable Marriage* problem [9], which is known to be NP-complete. The well-known Gale-Shapley algorithm [10] gives asymmetric solutions, so we opt to use a simple greedy matching algorithm.

1) Find the pairwise alignment p-value for every column of M_1 matched with every column of M_2. (For M_1 and M_2 each having k columns, this operation will take $(\frac{k^2}{2} - k)$ p-value calculations.)
2) Match the pair of columns $M_1[:, i]$ and $M_2[:, j]$ with the lowest pairwise p-value.
3) Repeat 1-2 until all columns have been assigned.

Note that this algorithm does not allow a single cluster in one clustering to be represented by a combination of clusters from the other clusterer. Future work will explore allowing such combinations.

III. Visualization of p-value based alignment

Along with the overlapping correspondence matching algorithm, we have developed a visualization tool which clearly presents correspondence alignments and other information. The visualization, as shown in Fig. 1, presents three frames. The first two show cluster "signatures", and the last a bar chart of overlap p-values.

Cluster signatures are a visual representation of a nondisjoint label assignment designed to facilitate quick inspection and comparison of labellings. Given an $n \times m$ membership matrix M, a cluster signature is constructed as follows:

```
1) For each row $i$,
    a) construct a vector $t$ containing the indices of the 1's and set $\text{score}[i] = 0$
    b) If $\text{length}(t) = 0$, $\text{score}[i] = n + 1$
    c) if $t[2] - t[1] \neq 1$, $\text{score}[i] = t[1] + \sum_{k=2}^{\text{length}(t)} \frac{t[k]}{m^k}$
    d) if $t[2] - t[1] = 1$, $\text{score}[i] = t[1] + (1 - \frac{1}{m^{m+1}}) + \sum_{k=2}^{\text{length}(t)} \frac{t[k]}{m^{m+k}}$
```

[1] A similar measure called the S-measure has been used previously in [7], [8].

2) Sort rows by increasing score

Ordering the membership matrix as described above puts the data points with no memberships at the bottom (step 1b) and sorts data points with memberships into blocks according to their minimum cluster label (steps 1c and 1d). Additionally, step 1d creates a visual overlap between consecutive overlapping clusters. Both steps 1c and 1d place points with additional cluster memberships into a unique order. This algorithm leads to a unique ordering for a given membership matrix, and is independent of the input order of the points.

IV. Comparison of Clustering Methods

We present the alignment and visualization with a comparison of two algorithms, Model-based Overlapping Clustering (MOC) [3] and thresholded soft k-means [2] . MOC takes as input an observed $n \times m$ data matrix E and factors it into an $n \times k$ binary matrix M and an $m \times k$ (real) activation matrix A. Soft k-means is very similar to the standard k-means algorithm, except that points are given partial ("soft") assignment to centers. The resulting membership matrix is real, with the property that for any row j, $\sum_{i=1}^{m} x[i,j] = 1$. A soft clustering can be converted into a hard clustering by thresholding the soft membership matrix.

For ease of explanation and analysis, we demonstrate the application of our alignment and visualization on synthetic data. We generated a 10 cluster synthetic dataset using the MOC generative model, which is a conceptual representation of the biological and experimental processes that produce collections of microarray experiments. The MOC model assumes that an observed $n \times m$ data matrix E can be expressed as the product an $n \times k$ (binary) nondisjoint membership matrix M and an $m \times k$ (real) activation matrix A. For this example, we have used $n = 1000, m = 30, k = 10$. (For a full discussion of the MOC model, please refer to [3].)

Fig. 1 illustrates the p-value based alignment of Model Based Overlapping Clustering with the ground truth for a synthetic data set. The uppermost box shows the signatures of each of the 10 clusters of the ground truth labels. The points have been sorted as described in Section III. The next box shows the signatures of each of the 10 clusters of the MOC labelling, with the points sorted as above and the clusters aligned. The final box shows the p-values of each of the alignments.

By visually comparing the first and second frames (the cluster signatures of the ground truth and of MOC's labelling, respectively) in Fig. 1, one can observe that this clusterer, MOC, has found a cluster labelling that corresponds well to the actual clusters in the data. Columns 1,3,5, and 6 show log_{10} p-values of less than -110, indicating infinitesimal odds of those matches occurring by chance. The other columns show good alignment, although the imperfections in the result are evident. Overall, this alignment is very good, which is to be expected since the data was generated using the clusterer's generative model.

Fig. 2 shows the alignment of a soft k-means clusterer run on the same artificial data set and thresholded at 0.3. Both the signature visualization and the p-value chart show that the clustering does not match the truth as well as the MOC clustering. While several clusters show reasonable correspondence with the ground truth, two clusters - columns 2 and 4 - completely fail to match.

V. Alignment for Cluster Ensembles

When the underlying generative model is unknown, combining the results of several diverse clusterers often improves the overall clustering result. One method of aligning clusters ensemble techniques is matching label assignments from each clusterer in the ensemble. Effective methods exist for disjoint membership matrices [13]; however, such methods are not applicable to nondisjoint membership matrices.

The p-value alignment method described in Section II provides a means of combining overlapping clusterings where each constituent clusterer uses the same k. In this section, we present results from combining three overlapping clusterers - MOC, thresholded soft k-means, and gene-shaving - on the previously described synthetic microarray dataset.

The idea behind cluster ensembles is that each constituent clusterer will return a noisy representation of the actual underlying clustering. Combining several clusterers in an ensemble averages out the noise and often provides a better estimate of the underlying clustering. Clustering algorithms in general return clustering information in arbitrary order, necessitating cluster matching prior to ensemble operations.

We applied MOC, soft k-means thresholded at 0.3, and gene shaving to our synthetic microarray dataset. We aligned the results to each other, as shown in Fig. 3. We then performed a majority-vote combination; that is, if a gene is marked as belonging to a cluster m by 2 out of the 3 clusters, we assign that gene to cluster m in the final result. Fig. 4 shows the final consensus result aligned to the ground truth.

[2] Soft k-means is the application of the expectation maximization (EM) algorithm to a mixture of k spherical Gaussians. Soft k-means minimizes an objective function equivalent to the fuzzy c-means [11] objective with "fuzziness" parameter m set to 1 and with dimensional scaling matrix A_k as identity. These are reasonable - but not necessarily optimal - parameters for this algorithm on our dataset. For a study on choosing m and A_k for a given dataset, see [12].

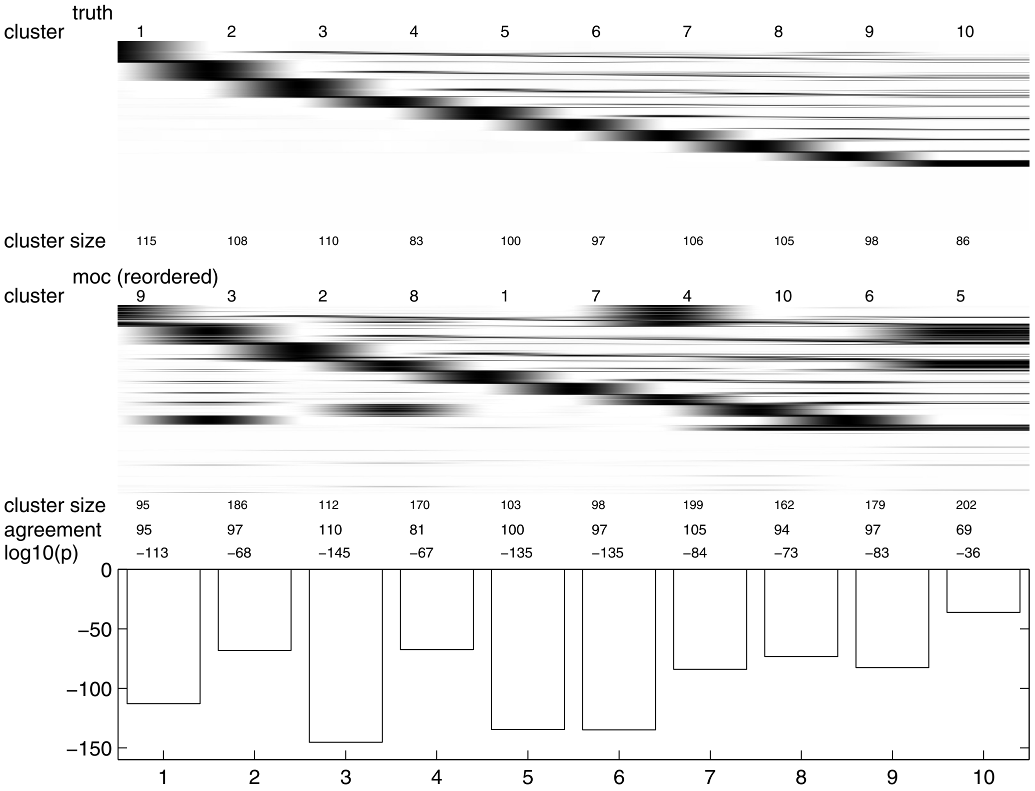

Fig. 1. Visualization of correspondence between ground truth and MOC overlapping clustering for synthetic microarray data. The top frame is the "signature" of the ground truth, with the rows sorted as described in Section III. The second frame is the "signature" of the MOC clustering, with the rows in the same order as in the top frame, and the columns matched using the algorithm described in Section II. The final frame is a bar chart of the alignment p-values. The more negative the alignment p-values, the less likely the alignment happened due to random chance.

The consensus clustering shown in Fig. 4 is superior to any of the constituent clusterings shown in Fig. 3. It should also be noted that while the individual alignments for MOC and soft k-means shown in Fig. 1 and Fig. 2 have some lower alignment p-values, the overall ensemble result appears to be much less noisy.

VI. Conclusion

Overlapping clustering techniques provides a means of clustering microarray data in a way that matches nicely with biologic intuition about the participation of genes in biological processes. The results of overlapping clusterings, though, can be difficult to interpret. In this paper, we presented a cluster alignment method and a visualization tool which facilitates comparison, evaluation, and combination of overlapping clustering results.

Notes

1) The model-based clustering algorithm was implemented by Bannerjee, Basu, and Krumpelman [3].
2) **geneclust** [14] was used for gene shaving.
3) **netlab** [15] was used for soft k-means.

Acknowledgement

This research was supported by NSF grant IIS-0325116.

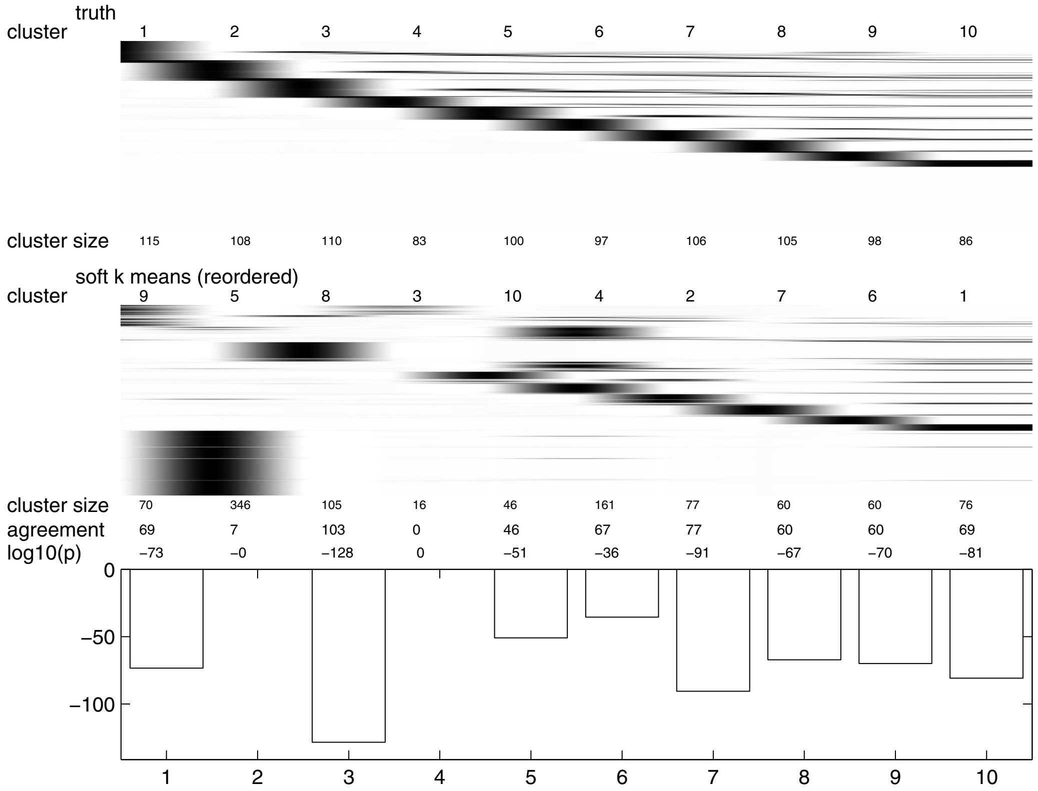

Fig. 2. Visualization of correspondence between ground truth and Soft k-means overlapping clustering for synthetic microarray data.

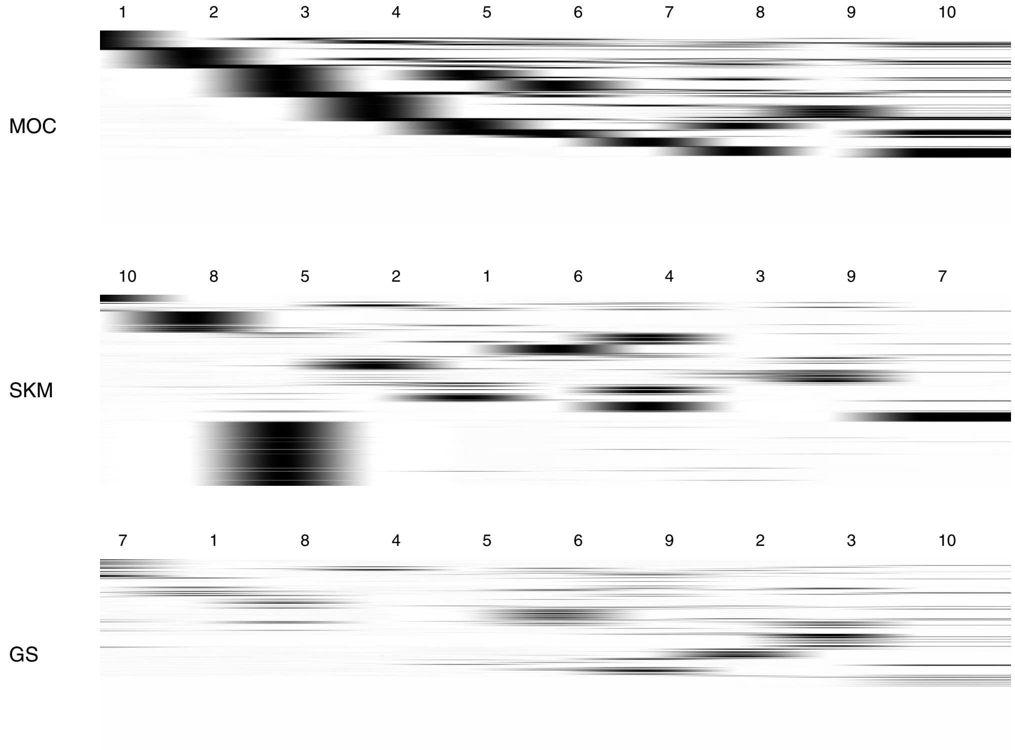

Fig. 3. Visualization of aligned cluster signatures of MOC, soft k-means, and gene shaving clustering results on the synthetic data.

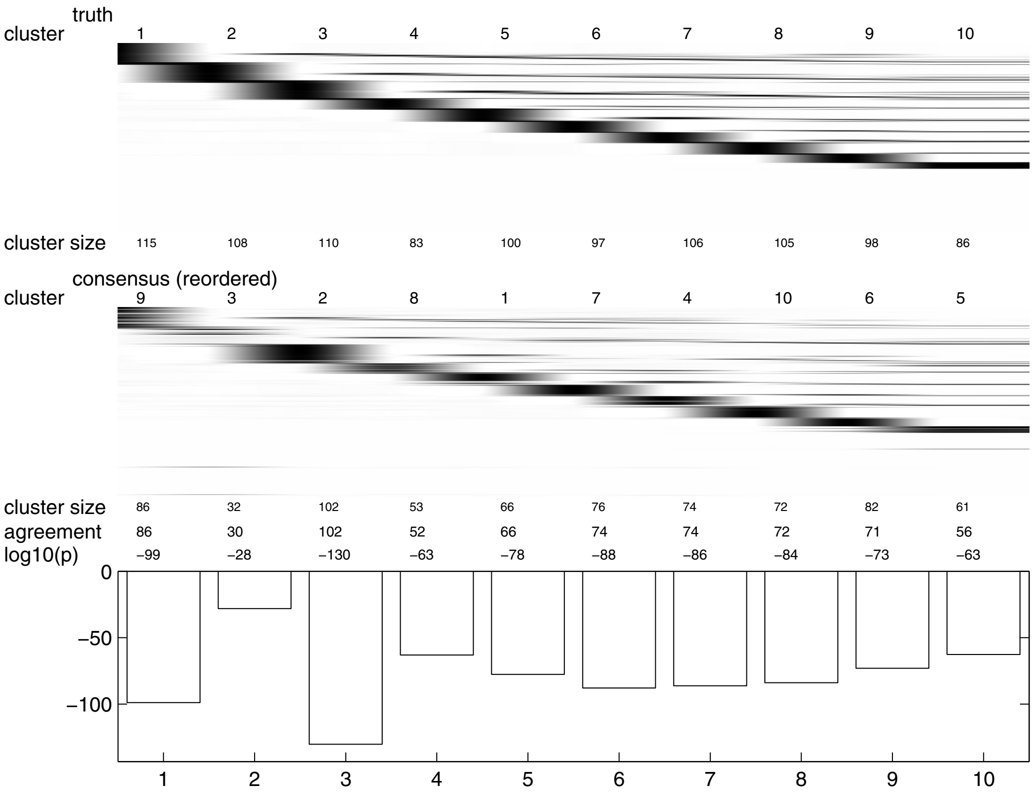

Fig. 4. Visualization of ground truth cluster labels and aligned majority-vote consensus of MOC, soft k-means and gene shaving. The consensus recovers the actual labelling better than any of the individual clusterers.

REFERENCES

[1] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown, "'gene shaving' as a method for identifying distinct sets of genes with similar expression paterns," *Genome Biology*, vol. 2, pp. 0003.1–0003.21, 2000. [Online]. Available: http://genomebiology.com/2000/1/2/research/0003

[2] E. Segal, A. Battle, and D. Koller, "Decomposing gene expression into cellular processes," in *Proc. 8th Pacific Symposium on Biocomputing (PSB), Kaua'i*, Jan 2003.

[3] A. Banerjee, S. Basu, C. Krumpelman, J. Ghosh, and R. Mooney, "Model based overlapping clustering," in *Proceedings of KDD2005*, 2005, pp. 100–106.

[4] L. Lazzeroni and A. B. Owen, "Plaid models for gene expression data," *Statistica Sinica*, vol. 12, no. 1, pp. 61–86, 2002.

[5] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[6] L. M. Collins and C. W. Dent, "Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions," *Multivariate Behavioral Research*, vol. 23, no. 2, pp. 203–230, 1988.

[7] X. Li and R. C. Dubes, "The selection of significant dichotomous features," in *Proceedings of the Seventh International Conference on Pattern Matching*, 1984, pp. 260–264.

[8] ——, "The first stage in two-stage template matching," in *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI 7*, 1985, pp. 700–707.

[9] D. Gusfield and R. W. Irving, *The Stable Marriage Problem: Structure and Algorithms*. Cambridge, MA: MIT Press, 1989.

[10] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *American Mathematics Monthly*, vol. 69, pp. 9–14, 1962.

[11] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

[12] D. Dembélé and P. Kastner, "Fuzzy c-means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 973–980, 2003.

[13] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining partitionings." in *Proc. Conference on Artificial Intelligence (AAAI 2002)*, July 2002, pp. 93–98.

[14] K.-A. Do, R. Nikolova, P. Roebuck, and B. Broom, "GeneClust." [Online]. Available: http://odin.mdacc.tmc.edu/ kim/geneclust/

[15] I. Nabney and C. Bishop, "netlab." [Online]. Available: http://www.ncrg.aston.ac.uk/netlab/index.php