A Supervised Learning Approach to the Unsupervised Clustering of Genes

Andrew Rider[1,3,4], Geoffrey Siwo[2,3], Scott Emrich[1,3], Michael Ferdig[2,3,4], and Nitesh V. Chawla[1,4]
[1]*Department of Computer Science and Engineering,*
[2]*Department of Biological Sciences,*
[3]*Eck Institute for Global Health,*
[4]*Interdisciplinary Center for Network Science and Applications,*
Notre Dame IN 46556, USA,
**Corresponding author: nchawla@nd.edu*

***Abstract*—Clustering is a common step in the analysis of microarray data. Microarrays enable simultaneous high-throughput measurement of the expression level of genes. These data can be used to explore relationships between genes and can guide development of drugs and further research. A typical first step in the analysis of these data is to use an agglomerative hierarchical clustering algorithm on the correlation between all gene pairs. While this simple approach has been successful it fails to identify many genetic interactions that may be important for drug design and other important applications.**

We present an approach to the clustering of expression data that utilizes known gene-gene interaction data to improve results for already commonly used clustering techniques. The approach creates an ensemble similarity measure that can be used as input to common clustering techniques and provides results with increased biological significance while not altering the clustering approach at all.

***Keywords*-clustering; ensemble; random subspaces; classifier; microarray;**

I. Introduction

An application of systems biology is to uncover the mechanisms underlying the behavior of a cell. The systems biology approach uses multiple sources of high-throughput data to build models of processes essential to life. Relationships between genes encode most of this information and are often discovered and represented as pathways that lead to essential products. Understanding these relationships is a very challenging problem as even the simplest organisms contain a multitude of genes that interact in complex combinations to deal with environmental conditions. Another complicating factor is that current high-throughput technology used to measure the activity level of genes is notoriously noisy [1]. As there are very few well understood genetic interactions, unsupervised clustering is a common first step to understanding these data [2], [3].

We describe a procedure for creating an ensemble statistic based on a number of distance measures that can increase the biological significance of clustering results. We posit that an intelligent combination of multiple statistics can describe the extent to which two genes are similar more precisely than any single statistic. We propose a supervised learning approach for building an ensemble statistic from any number of descriptive statistics. The approach leverages the expression data of genes that are known to interact to obtain additional information about relationships between less well understood genes and identify previously overlooked relationships between genes. It has the additional benefit of recognizing any relationship that can be described by a statistic. We apply our approach to the model organism *Saccharomyces cerevisiae* (yeast). Yeast is an ideal organism to consider given the availability of annotation and experimentally derived gene interaction data.

A. Overview

Our approach is inspired by two general principles in data mining. First, noisy data complicates identification of interesting patterns. This principle guides our approach in two ways: we use experimentally derived gene interaction data and we use the random subspaces method. A second guiding principle is that ensemble models tend to outperform more straightforward approaches. This principle supports our assessment that even weakly descriptive statistics contain information that is missed by stronger predictors.

The approach can be described roughly in four steps.

- Calculate descriptive statistics on the microarray data for each gene pair.
- Train C4.5 decision trees on random subspaces of the features using experimentally derived positive and negative interacting gene classes[8].
- Calculate a measure of feature importance based on the structure of the trees in our model.
- Weight each statistic by its feature importance and create ensemble similarity measures.
- Cluster the ensemble data.

First we calculate descriptive statistics on the microarray data for each gene pair. Each statistic describes a different type of relationship between a pair of genes. In order to demonstrate the success of our approach we use a set of statistics that, with the exception of correlation, we believe will result in poor clustering results. Second we train C4.5 trees on random subspaces of the features using experimentally derived positive and negative interacting gene classes from gold standard data sets. The random subspaces

978-1-4244-8305-1/10/$26.00 ©2010 IEEE

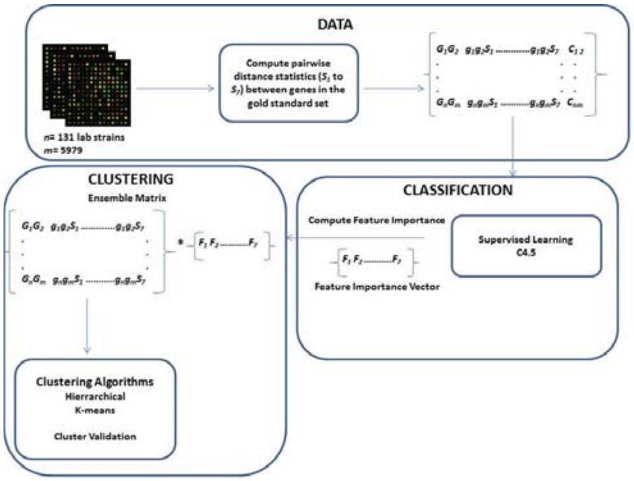
DATA
Compute pairwise distance statistics (S_1 to S_7) between genes in the gold standard set
G_1G_2 $g_1g_2S_1$ .... $g_1g_2S_7$ $C_{1\,2}$
G_nG_m $g_ng_mS_1$ .... $g_ng_mS_7$ C_{nm}
n= 131 lab strains
m= 5979
CLUSTERING
Ensemble Matrix
G_1G_2 $g_1g_2S_1$ .... $g_1g_2S_7$
G_nG_m $g_ng_mS_1$ .... $g_ng_mS_7$
* F_1 F_2 .... F_7
CLASSIFICATION
Compute Feature Importance
Supervised Learning C4.5
F_1 F_2 .... F_7
Feature Importance Vector
Clustering Algorithms
Hierrarchical
K-means
Cluster Validation

Figure 1. Overview of the approach. The method begins by computation of pairwise distance statistics. We calculated seven statistics (S1 to S7) between each interacting gene pair (e.g G1G2) in both the positive and negative gold standards data set. Many gene pairs have a class label C which can either be positive or negative as defined in the gold standards data set. C4.5 was then used to estimate feature importance (F1 to F7) as the sum of information again across all splits of a given feature.

approach builds classifiers using subsets of the available statistics. The use of random subspaces allows the classifiers to investigate how different combinations of statistics work together to predict gene interactions. Some statistics may act in combination to improve classification whereas others may interfere with classification or obscure the positive effects of less reliable predictor variables. We overcome this issue by evaluating our approach across all possible subspace sizes. Next we calculate a measure of feature importance based on the structure of the trees in our model. C4.5 trees were chosen because of the conceptual ease of determining feature importance as a function of tree structure. We use the feature importance to weight the individual statistics and combine them into an ensemble statistic. The use of feature importance as a weighting mechanism in combination with the random subspaces method has the effect of increasing the weight of statistics that were good predictors of gene interactions. Finally, we cluster the ensemble data. Figure 1 depicts the full experimental design. The individual steps are explained in detail in the methods section.

II. Data

A. Gene Expression Data

We considered a yeast expression data set from a line cross experiment composed of a 131 strains and 5979 probes[4]. Genetic line cross experiments have great potential to elucidate the causative agents of drug resistance and can shed light on the intricate relationships between genes and ultimately targets for drug design[5]. Additionally, yeast has the advantage of having a thoroughly annotated genome and experimentally derived gene interaction data.

B. Positive and Negative Gold Standards

Positive and negative gold standard sets of gene interaction data for yeast were obtained from a manually curated set of GO terms, which was balanced in terms of functional classes of genes[6]. Interacting genes were selected by voting results from a team of six expert biologists. Gene pairs were said to be interacting if each gene shared a GO term specific enough to imply functional association. Positive and negative sets consist of pairs of genes that have been confirmed or refuted as interacting through laboratory experiments rather than computational approaches. Six expert biologists voted on whether each of a large set of GO terms should be considered interacting. Terms with many votes were considered interacting while terms with one or less vote were considered non-interacting. The use of these data allowed us to create an ensemble statistic while minimizing functional bias due to an unbalanced hierarchy of GO terms.

III. Methods

A. Similarity measures

Our approach is flexible in that the number of statistics that can be combined into an ensemble is only limited by the computational burden in the classification step. Individual statistics are weighted by the amount they contribute to predicting interactions. This approach reduces the effects of statistics that do not contribute to the identification of interacting gene pairs. We demonstrate this property by using a collection of statistics, most of which we do not expect to discriminate well between interacting and non-interacting genes. These weakly predictive statistics may have regions in which they are locally good predictors or they may be good predictors in combination with other statistics. Our approach is designed to take advantage of these effects.

We calculated seven similarity measures on the expression data for each gene pair, shown in Table I.

Table I
FEATURES USED FOR SUPERVISED LEARNING.

Feature	description
City block distance	Distance along a grid
Correlation	The extent to which two variables are linearly related
Covariance	Amount a pair of variables change together
Hellinger distance	The similarity in shape of marginal distributions
Kolmogorov-smirnov	Similarity in shape and magnitude of two distributions
Mutual information	Mutual dependence of two variables
Spearman rank correlation	Extent to which two variables are monotonically related

B. Classification

Each statistic describes a different relationship in the data. Our goal is to combine the strengths of all the statistics into a single ensemble. We do this by leveraging patterns in the expression of gene pairs that are known to be interacting.

The greatest challenge in evaluating the usefulness of each statistic is that they can interact in complex ways. A classifier built on a pair of statistics may have additional predictive power over a single statistic classifier. However, a different pair of statistics may be less useful if they both contain similar information or are poor predictors. An additional challenge is that statistics can be locally strong predictors. Locally strong predictors may be overshadowed by generally better predictors. This is a loss because each statistic, even a generally poor predictor, contains some information about a relationship in the data. Our approach seeks to weight statistics in proportion to their overall usefulness in identifying interacting genes.

We trained classifiers on random subsets of similarity measures. Using a single random subset of two similarity measures, we might train a classifier on only correlation and mutual information data. This approach is known as the random subspaces method[7]. It allows us to investigate the effects of various combinations of similarity measures on prediction of gene interaction.

We used C4.5 decision tree classifiers on random subspaces of similarity measures. The C4.5 algorithm splits data into subsets by an information gain criterion[8]. Because of this, a similarity measure that is a locally good predictor of interactions may not be split on in a C4.5 tree that has access to similarity measures with more predictive power. Random subspaces allow even poor predictors to be built into a tree. This allows our approach to recognize similarity measures that are good predictors of specific small subsets of interactions even when they are poor predictors in general.

We trained fifty C4.5 trees on random subspaces of every size, from single similarity measure subspaces to seven similarity measure subspaces. The largest combination of similarity measures for any subspace of seven is thirty-five. Creating fifty trees for each subspace size ensures that each similarity measure has a chance to be chosen with every combination of other similarity measures as the basis for splitting in a tree.

The original data contained a large imbalance towards non-interacting gene pairs. In order to create classifiers with an emphasis on identifying positive interactions, we took a random sample of 140,000, composed of equal numbers of positive and negative interactions for the training set.

C. Feature importance

We calculated feature importance as the sum of information gain across all splits in decision trees for each similarity measure. We believe that this is an informative metric because information gain depends on the amount of data split as well as the usefulness of the split for prediction. Splits further down the tree typically affect less data and have lower information gain. This trend agrees with the intuition that splits lower in the tree are less important to overall tree structure.

The feature importance for each subspace size was the mean of the feature importance for each similarity measure from all fifty trees. Classifiers were validated using 10-fold cross validation. The total feature importance was the mean of the feature importance for all classifiers across the ten folds. Finally, we transformed each feature importance measure into the proportion of total feature importance across all similarity measures. Table II contains the similarity measures used in classifiers and the scaled feature importance for yeast in subspaces of size three.

Table II
SIMILARITY MEASURES USED FOR SUPERVISED LEARNING AND THE FEATURE IMPORTANCE ASSIGNED TO THEM AS DETERMINED FOR RANDOM SUBSPACES OF THREE SIMILARITY MEASURES.

Similarity measure	Feature importance
Kolmogorov-smirnov	0.1185
Covariance	0.2119
Correlation	0.1827
Spearman	0.0780
City block	0.0815
Mutual information	0.2228
Hellinger Distance	0.1043

D. Ensemble similarity measure

We used three approaches to build ensemble similarity measures. All component similarity measures were range standardized such that all elements fell between zero and one. Each similarity measure was weighted by multiplying all values in the similarity matrix with the corresponding feature importance. A weighted sum ensemble was created by computing the sum of each corresponding element from all similarity measure matrices. Similarly, weighted min and max ensembles were created by taking the min and max respectively for each element of the matrix.

E. Clustering Algorithms

Hierarchical clustering and k-means clustering were performed on the ensemble matrices.

The k-means clustering algorithm attempts to identify the best fit clusters by minimizing the within cluster sum of squared distance from cluster centers [9]. Given unlimited iterations, the k-means algorithm attempts to optimize globally on its clustering criterion and tends to result in clusters with spherical shape and size. We used k-means clustering with five random restarts and a voting process for cluster membership to reduce the possibility of the algorithm converging to locally optimized clusters.

We report results for two agglomerative hierarchical clustering criterion: UPGMA and Ward's criterion. UPGMA groups clusters by the mean distance between elements of each cluster [10]. This results in a tendency to group clusters with small variance. Ward's criterion groups clusters explicitly with regard to cluster variance by joining two clusters based on the minimum increase in variance when two groups are merged [11]. This approach tends to result in equal sized spherical clusters. In contrast to K-means clustering, Ward's criterion and UPGMA both optimize locally on their clustering criterion[12].

We tested our method with two additional hierarchical clustering criterion, including single linkage and median linkage. We found that Single linkage and median linkage produced very poor clustering results. Our findings with respect to Single linkage agree with results reported in [13]. Additionally, we used Markov Clustering with results similar to single and median linkage. We focus here on the results that best demonstrate the differences between clustering with a single similarity measure and clustering with an ensemble statistic.

F. Cluster validation

There are two general approaches to validation of microarray cluster results: validation based on internal measures and validation based on additional biological knowledge. [6], [14] We focus on biological validation and use the GO similarity to measure the validity of cluster results.

GO similarity measures the similarity of pairs of terms by the distance between them in a tree describing the hierarchy of GO terms. The similarity of GO terms was computed by the Lin method [15] via the GOSim package [16]. The mean GO similarity across all clusters for each combination of clustering algorithm and similarity measure was used to measure the utility of cluster results.

G. Statistical comparison of results

We utilized the Wilcoxon signed-rank test to compare pairs of cluster results. A signed-rank test is a non-parametric analog of a t-test. It compares the difference between tied pairs of items by ranking the differences into positive and negative sets of ranks. [17]

$$W_+ = \sum_{d_i>0} rank(d_i) + 1/2 \sum_{d_i=0} rank(d_i) \tag{1}$$

$$W_- = \sum_{d_i<0} rank(d_i) + 1/2 \sum_{d_i=0} rank(d_i) \tag{2}$$

Where d_i is the distance between tied pair i. The smaller of the two values, T, is given a z-score as follows:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \tag{3}$$

Where N is the number of observations.

We used the Friedman test to rank the performance of ensembles across different clustering algorithms. The Friedman test is a non-parametric equivalent of ANOVA. In contrast to ANOVA, the Friedman test does not make the assumption that the sample means being tested have related means or that the underlying variables have equal variance. Instead, the Friedman test assumes that the data come from populations with the same continuous distributions and that all observations are mutually independent. These assumptions are desirable for our data because clustering results from separate algorithms may be extremely variable.

The Friedman test compares multiple treatments across multiple data sets under the hypothesis that all treatments are equivalent and should have the same rank. The test compares the mean rank of all combinations of sample i (of N total data sets) and algorithm j (of k total algorithms) by first calculating the mean performance of each algorithm across samples in Equation 5 then comparing the mean performance of algorithms in in Equation 6.

$$R_j = \frac{1}{N} \sum_i r_i^j \tag{4}$$

Where R_j is the rank of algorithm j.

$$\chi_F = \frac{12N}{k(k+1)} [\sum_j R_j^2 - \frac{k(k+1)^2}{4}] \tag{5}$$

IV. Results

The random subspaces resulted in classifiers with a very low classification accuracy. The classifiers had a range of classification accuracy between 55.97% to 60.89% for the positive class. These are poor classifiers but the following results show that the ensemble statistics built from even a poor classifier can significantly improve cluster results.

A set of ensemble similarity measures was created for each subspace size. In the following sections we compare the effects of using these ensembles for clustering to the effect of using the correlation alone.

A. GO similarity based validation

A separate Friedman's test was performed for each clustering algorithm in combination with each similarity measure for all subspace sizes. Table III shows Friedman rank results for twenty-one Friedman tests, each comparing a clustering algorithm to itself with different similarity measures for a single subspace size. Higher rank indicates more significant effects. The ranks in Table III indicate that every combination we tried outperformed UPGMA and correlation. For all algorithms the sum ensemble appears to perform well for smaller subspaces and the max seems to perform well for larger subspaces.

The mean GO similarity refers to the mean across clusters from an individual clustering. For example, each cluster in a clustering experiment with ten cluster results has a GO similarity. The mean of this quantity across clusters describes the overall performance of the algorithm on the similarity measure. The mean of that quantity across all clustering experiments for a combination of clustering algorithm and

Table III
RANKED RESULTS FOR THE MEAN GO SIMILARITY FOR EACH COMBINATION OF SIMILARITY MEASURE AND CLUSTERING ALGORITHM. CLUSTERING RESULTS WERE RANKED AGAINST EACH OTHER USING THE SAME SUBSPACE SIZE. GREATER RANK IN THE LAST ROW CORRESPONDS TO GREATER EFFECTS ACROSS SUBSPACE SIZES.

	UPGMA				Ward				K-means			
subspace size	Corr.	min	max	sum	Corr.	min	max	sum	Corr.	min	max	sum
1	1	2	3	4	1	2	4	3	2	3	1	4
2	1	3	4	2	1	3	4	2	1	2	4	3
3	2	1	4	3	1	3	4	2	1	2	4	3
4	2	1	4	3	1	3	4	2	1	3	4	2
5	1	3	2	4	1	3	2	4	2	3	1	4
6	1	3	2	4	2	1	3	4	2	3	1	4
7	1	3	2	4	1	2	3	4	3	2	1	4
GOSim rank	1	2	3	4	1	2	4	3	1	3	2	4

similarity measure is the mean of the mean GO similarity. It describes how well the combination performs across various numbers of clusters.

We used the mean of the mean GO similarity to measure the performance of a combination of clustering algorithm and similarity measure across different clustering results and varying numbers of resulting clusters. Figure 2 depicts the mean of the mean GO similarity across cluster numbers for each subspace size. The combinations of UPGMA and the max ensemble and UPGMA and the sum ensemble out-perform all three clustering algorithms and the correlation. These observations agree with the ranks given in Table III.

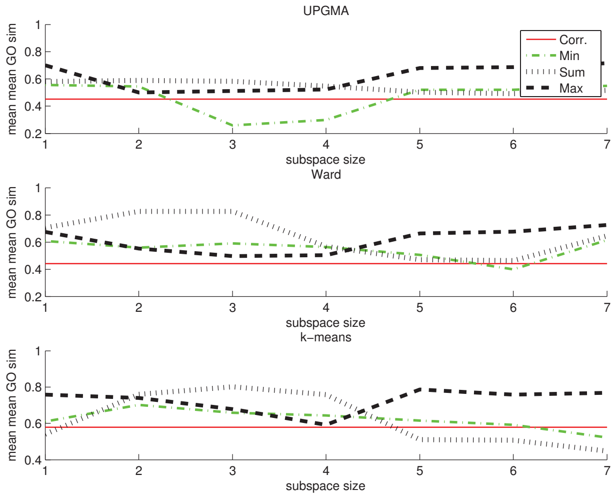

Figure 2. The subspace size versus the mean mean GO similarity.

We used signed-rank tests to compare the mean GO similarity between all combinations of clustering algorithm and similarity measure. We compared the mean GO similarity across the range of cluster numbers tested for each given combination of clustering algorithm and similarity measure. UPGMA in combination with the max and the sum ensembles significantly outperformed ($\alpha = 0.05$) UPGMA in combination with correlation for all subspace sizes. The best performance overall in terms of GO similarity was from the combination of k-means clustering and the sum ensemble followed in performance by the max ensemble and k-means clustering.

V. DISCUSSION

We have presented strong evidence that an ensemble can out-perform correlation across different clustering algorithms already in use for analysis of microarray data. Furthermore, we have shown that our ensembles consistently produce better clustering results than correlation alone across subspace sizes and these common clustering algorithms.

Figure 2 shows that the sum ensemble performs well where the max ensemble performs most poorly and vice versa for all clustering algorithms. This variation in performance shows that subspace size is an important factor in our approach. These trends can be explained by the way the ensembles are created. The sum ensemble performs well for small subspace sizes because it factors in effects of all statistics even when they were not all used in combination as much as they would have been at larger subspace sizes. The max ensemble filters out the noise of having more statistics built into each classifier by simply choosing the largest effect for each gene pair. A second interesting feature of Figure 2 is that the sum and max ensembles have a sharp change in performance at subspace size four across all three clustering algorithms. This pattern further supports the idea that the max ensemble filters out noise while the sum ensemble strengthens models built with less data. Subspaces of size four are exactly between the minimum and maximum size and are a low point for both ensembles.

Although we report results with GO similarity based validation, we also used interaction-based validation. We calculated the F-measure based on the number of positive interactions in clusters. We found that The sum and max ensembles perform well for UPGMA and Ward's criterion, respectively, but we did not find significant trends on the order of those found in the GO similarity validation data.

The apparent discrepancy between the interaction based validation results and the GO similarity validation results may be attributed to the underlying differences in what is measured by the two approaches. The interaction based validation considered only gene pairs which shared a GO term. The GO similarity validation on the other hand used all available GO terms and measured the distance between all pairs on the GO hierarchy. Not only did the GO similarity method have more data available but it considered less specific relationships. In light of the differences between the results, the combinations of clustering algorithm and similarity measure that performed well by both measures are particularly interesting. Poor agreement between gold standards and validation methods is a pervasive problem in biological validation but it should not imply that the approaches are useless[6]. Each gold standard or validation method may have its own bias but is still informative.

The importance of the data set used in this procedure cannot be overstated. We attempted the same experiment with Biogrid yeast interaction data [18] with less definitive but still encouraging results. The clustering results using ensembles built on the Biogrid interaction set were statistically indistinguishable from clustering results using correlation alone. Although the Biogrid database contains a larger collection of curated interaction data, we feel that the carefully balanced data set we used was an important contributing factor to the success of this approach.

In addition to the data from [6] and the Biogrid data, we applied this approach to the malaria parasite *Plasmodium falciparum*. We obtained expression data from a line cross experiment between drug sensitive and drug resistant strains[19]. Unfortunately, *Plasmodium falciparum* is not as thoroughly annotated as yeast and we were unable to obtain reliable GO similarity validation results. It was very interesting to note that the same combinations of ensemble similarity measure and clustering algorithms that were promising in the yeast data performed well on the malaria data. Namely, the combination of K-means with the max and sum ensembles performed well on both data sets according to interaction-based based validation using the F-measure.

VI. Conclusions

We have described a method that provides a number of advantages over typical approaches to gene clustering: i) it intelligently weights similarity measures by their predictive power, allowing a number of statistics to be utilized regardless of their individual usefulness. ii) It can result in an increase in cluster validity even when a very poor classifier is built on noisy data. iii) The method employs prior biological knowledge in the form of known gene to gene interactions represented by positive and negative gold standards and integrates this into the similarity measure. iv) It complements and improves upon existing common and successful methods of analyzing high-throughput biological data. v) Because it creates an ensemble similarity measure rather than altering a clustering approach, it could be used with clustering methods beyond those discussed here.

References

[1] B. J. Daigle, A. Deng, T. McLaughlin, S. W. Cushman, M. C. Cam, G. Reaven, P. S. Tsao, and R. B. Altman, "Using pre-existing microarray datasets to increase experimental power: application to insulin resistance." *PLoS computational biology*, vol. 6, no. 3, pp. e1 000 718+, March 2010.

[2] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, "Co-clustering of biological networks and gene expression data," *Bioinformatics*, vol. 18, no. suppl 1, pp. S145–154, July 2002.

[3] S. Datta and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data," *Bioinformatics*, vol. 19, no. 4, pp. 459–466, March 2003.

[4] Quinlan, J. Ross, *C4.5: Programs for Machine Learning*, 1st ed. Morgan Kaufmann, January 1993.

[5] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak, "Genetic dissection of transcriptional regulation in budding yeast," *Science*, vol. 296, no. 5568, pp. 752–755, April 2002.

[6] R. Jansen, "Genetical genomics: the added value from segregation," *Trends in Genetics*, vol. 17, no. 7, pp. 388–391, July 2001.

[7] C. Myers, D. Barrett, M. Hibbs, C. Huttenhower, and O. Troyanskaya, "Finding function: evaluation methods for functional genomic data," *BMC Genomics*, vol. 7, no. 1, pp. 187+, July 2006.

[8] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, Aug 1998.

[9] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[10] P. H. SNEATH and R. R. SOKAL, "Numerical taxonomy." *Nature*, vol. 193, pp. 855–860, March 1962.

[11] Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[12] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, 4th ed. Wiley, January 2009.

[13] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Research*, vol. 12, no. 10, pp. 1574–1581, October 2002.

[14] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis." *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, August 2005.

[15] D. Lin, "An information-theoretic definition of similarity," in *In Proceedings of the 15th International Conference on Machine Learning*, 1998, pp. 296–304.

[16] H. Frohlich, N. Speer, A. Poustka, and T. BeiSZbarth, "Gosim - an r-package for computation of information theoretic go similarities between terms and gene products," *BMC Bioinformatics*, vol. 8, no. 1, pp. 166+, May 2007.

[17] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[18] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucl. Acids Res.*, vol. 34, no. suppl 1, pp. D535–539, January 2006.

[19] J. M. Gonzales, J. J. Patel, N. Ponmee, L. Jiang, A. Tan, S. P. Maher, S. Wuchty, P. K. Rathod, and M. T. Ferdig, "Regulatory hotspots in the malaria parasite genome dictate transcriptional variation." *PLoS biology*, vol. 6, no. 9, pp. e238+, September 2008.