A Supervised Learning Approach to the Ensemble Clustering of Genes

Andrew K. Rider[1,3,4], Geoffrey Siwo[2,3], Scott J. Emrich[1,3], Michael T. Ferdig[2,3,4], and Nitesh V. Chawla[1,3,4]
[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*—High-throughput techniques have become a primary approach to gathering biological data. These data can be used to explore relationships between genes and guide development of drugs and other research. However, the deluge of data contains an overwhelming amount of unknown information about the organism under study. Therefore, clustering is a common first step in the exploratory analysis of high-throughput biological data.**

We present a supervised learning approach to clustering 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 any clustering technique and provides results with increased biological significance while not altering the clustering method.

***Keywords*-Clustering; Ensemble; Random subspaces; Classifier; Microarray;**

I. Introduction

A primary goal of systems biology is to uncover the mechanisms underlying the behavior of a cell. 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. An additional complicating factor is that current high-throughput technology used to measure the activity level of genes is notoriously noisy (Daigle, et al. 2010). As there are very few well understood genetic interactions, clustering is a common first step to understanding these data (Hanisch, et al. 2002, Datta & Datta 2003, Eisen, et al. 1998).

We describe a supervised approach to clustering that can increase the biological significance of clustering results by creating an ensemble similarity measure. 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. Our approach uses supervised learning to build an ensemble statistic from any number of descriptive statistics.

We focus on clustering microarray data. Microarrays enable simultaneous high-throughput measurement of the expression level of genes. Our approach leverages the expression data of genes that are known to interact to obtain additional information about relationships between less well understood genes. In contrast to the typical clustering approach, in which a single clustering algorithm uses a single similarity measure, this method has the potential to recognize any relationship that can be described by a statistic.

We apply our approach to the model organisms *Saccharomyces cerevisiae* (yeast) and *Escherichia coli*. Both are ideal organisms to consider given the availability of annotation and experimentally derived gene interaction data (Christie, et al. 2009, Keseler, et al. 2005).

A. Overview

Two general observations about data mining guide our approach. First, noisy data complicates identification of interesting patterns. We approach this problem by using experimentally derived gene interaction data and the random subspaces method. Second, even uninformed ensemble models tend to outperform more straightforward approaches (Chawla, et al. 2004). 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 (Quinlan 1993).
- 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 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 yeast expression data from a line cross experiment composed of a 131 strains and 5979 probes (Brem, et al. 2002). 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 (Jansen 2001). We also considered expression data from an *E. coli* experiment studying the effect of oxygen deprivation (Covert, et al. 2004).

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 (Myers, et al. 2006). 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.

Interacting *E. coli* genes were derived from gene pathway data. Pathway data were gathered from the Ecocyc database (Keseler et al. 2005). Ecocyc is a comprehensive database of the current knowledge about *E. coli*. Genes that occur in the same pathway were considered interacting. Negative interactions were simulated by randomly selecting pairs of genes that did not share a pathway.

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
Cosine similarity	The cosine of the angle between two vectors
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

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

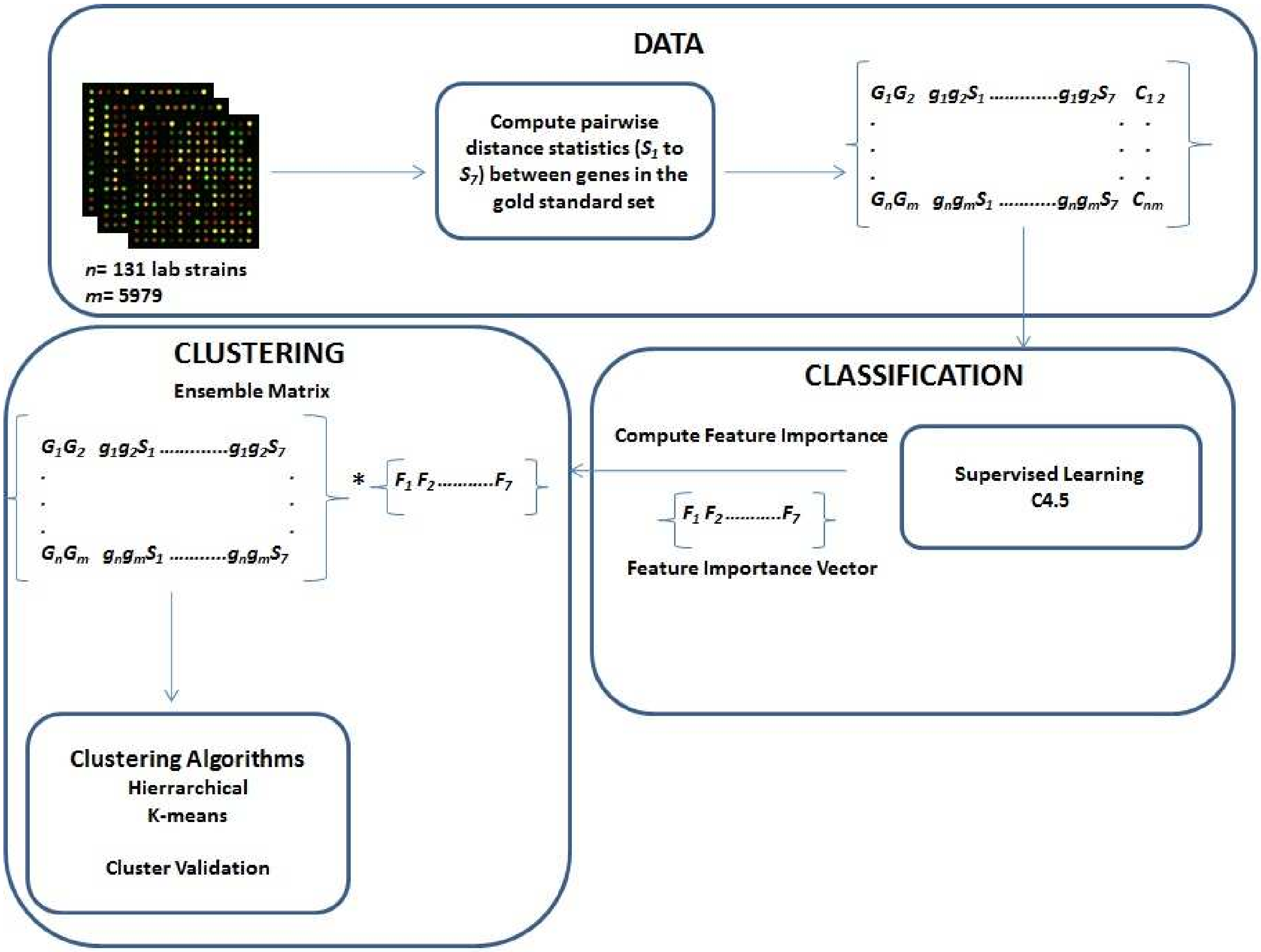
DATA
n= 131 lab strains
m= 5979
Compute pairwise distance statistics (S_1 to S_7) between genes in the gold standard set
G_1G_2 $g_1g_2S_1$ ....$g_1g_2S_7$ $C_{1\,2}$
G_nG_m $g_ng_mS_1$ ....$g_ng_mS_7$ C_{nm}
CLUSTERING
Ensemble Matrix
G_1G_2 $g_1g_2S_1$ ....$g_1g_2S_7$
G_nG_m $g_ng_mS_1$ ....$g_ng_mS_7$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

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 (Ho 1998). 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 (Quinlan 1993). 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.

The original yeast interaction 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 samples composed of equal numbers of positive and negative interactions for the training set. The *E. coli* data was also balanced in terms of positive and negative interactions.

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 was measured as the mean of the feature importance for each similarity measure from all classifiers. Because the classifiers were used exclusively to derive feature importance, to validation was necessary. 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 both data sets. The statistic with the largest feature importance for the yeast data set is correlation closely followed by covariance. In contrast, mutual information had the largest feature importance by far in the *E. coli* data set.

Table II
SIMILARITY MEASURES USED FOR SUPERVISED LEARNING AND THE FEATURE IMPORTANCE ASSIGNED TO THEM.

Similarity measure	Yeast feature importance	*E. coli* feature importance
City block	0.1850	0.1386
Correlation	0.2089	0.1001
Cosine	0.1744	0.1931
Covariance	0.2021	0.1060
Hellinger Distance	0.1073	0.1053
Kolmogorov-smirnov	0.0592	0.0208
Mutual information	0.0627	0.3357

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, k-means clustering, and the Walktrap clustering algorithm were applied to the ensemble similarity measures.

The k-means clustering algorithm attempts to identify the best fit clusters by minimizing the within cluster sum of squared distance from cluster centers (Hartigan & Wong 1979). 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 method. UPGMA groups clusters by the mean distance between elements of each cluster (Sneath & Sokal 1962). This results in a tendency to group clusters with small variance. Ward's method groups clusters explicitly with regard to cluster variance by joining two clusters based on the minimum increase in variance when two groups are merged (Ward 1963). This approach tends to result in equal sized spherical clusters. In contrast to K-means clustering, Ward's method and UPGMA both optimize locally on their clustering criterion (Everitt, et al. 2001).

The walktrap algorithm is designed to capture community structure by simulating random walks in networks (Pons & Latapy 2005). Walktrap creates a similarity measure based on the probability that random walks from each node end at each other node. Communities are merged using Ward's method.

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 (Gibbons & Roth 2002). Additionally, we found that Markov Clustering produced 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. (Myers et al. 2006, Handl, et al. 2005) We use both approaches, using the F-measure to evaluate interactions present in clusters and the Biologiacal Homogeneity Index to measure the validity of cluster results.

The F-measure is a measure of accuracy based on the trade-off between precision, the proportion of gene pairs in a cluster that are known positive interactions (Equation 1), and recall, the proportion of the known interactions that are in the cluster (Equation 2). The F-measure is defined in Equation 3.

$$precision = tp/(tp + fp) \tag{1}$$

$$recall = tp/(tp + fn) \tag{2}$$

$$F - measure = \frac{precision * recall}{(precision + recall)} \tag{3}$$

The Biological Homogeneity Index (BHI) measures cluster validity based on the proportion of genes in each cluster that share at least one GO annotation (Datta & Datta 2006). Each pair of annotated genes x and y in cluster D that share at least one GO term (C(x)=C(y)) in Equation 4 increases the proportion of total genes with shared terms.

$$\frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\varepsilon D} I(C(x)=C(y)) \tag{4}$$

Where k is the number of clusters.

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. (Demsar 2006)

$$W_+ = \sum_{d_i>0} rank(d_i) + 1/2 \sum_{d_i=0} rank(d_i) \tag{5}$$

$$W_- = \sum_{d_i<0} rank(d_i) + 1/2 \sum_{d_i=0} rank(d_i) \tag{6}$$

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

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \tag{7}$$

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 III-G then comparing the mean performance of algorithms in in Equation III-G.

$$R_j = \frac{1}{N}\sum_i r_i^j \tag{8}$$

Where R_j is the rank of algorithm j.

$$\chi_F = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}] \tag{9}$$

IV. Results

In previous work we trained 50 C4.5 trees on random subspaces of every size, from subspaces using single similarity measures to subspaces using every similarity measure (Rider, et al. 2010). Observations made during that analysis lead us to extend the approach to take all random subspace sizes into account simultaneously. Therefore we trained classifiers on 100 random subspaces of random sizes, utilizing anywhere from a single feature to all features. A set of ensemble similarity measures was created for each data set. In the following sections we compare the effects of using these ensembles as the basis of clustering to the effect of using the correlation alone.

A. Yeast interaction based validation

We performed Friedman's tests to determine if any of the ensembles significantly affected the cluster results. TableIII shows the Friedman's test results comparing the effect of different similarity measures on each clustering algorithm. The table shows that there are significant differences in results depending on the similarity measure used for all algorithms except k-means.

Table III

Friedman's test results comparing the effect of similarity measures on each clustering algorithm. Each p-value represents a comparison of clustering results gathered using each ensemble and correlation with the algorithm named in the row. P-values significant at $\alpha = 0.05$ appear in bold.

Algorithm	p-value
UPGMA	**1.38e-06**
Ward	**0.00129**
K-means	0.07717
Walktrap	**0.00825**

We used signed rank tests to determine which ensembles have the greatest effect on the clustering results in Table IV. We tested the hypothesis that the median F-measure of ensemble-based results is greater than the median F-measure of correlation-based results. The min ensemble is the only similarity measure with a statistically significant effect on cluster results at $\alpha = 0.05$. We also tested the opposite hypothesis, that the median F-measure of correlation-based results is greater than the corresponding ensemble-based results, and found no significant effects. In absolute terms, the best BHI results were achieved by the min ensemble in combination with either the Walktrap algorithm or UPGMA.

Figure 2 shows the F-measure versus the number of clusters for all similarity measures and all clustering algorithms. As indicated by Table IV, the min ensemble performs very well in combination with UPGMA and Walktrap. Regardless of clustering algorithm, all ensembles appear to always do at least as well as correlation alone and noticeably better in UPGMA and Walktrap results.

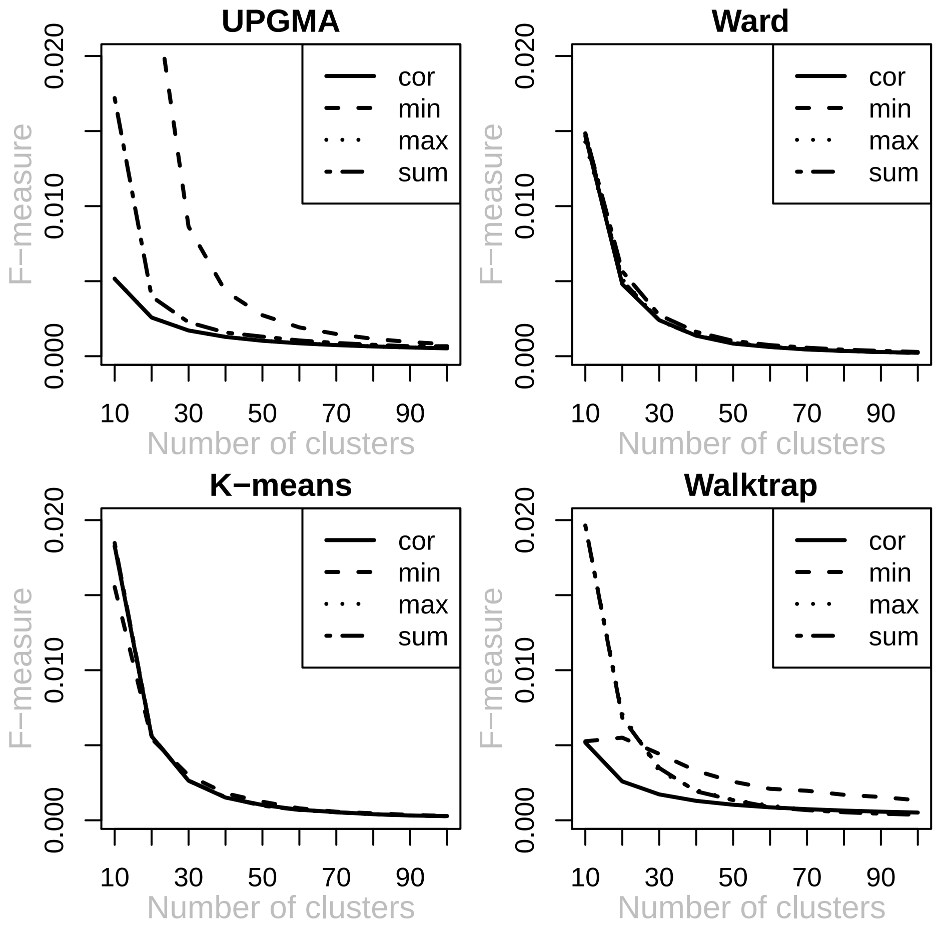

Figure 2. F-measure versus number of clusters.

Table IV
Signed rank test results testing the hypothesis that the median F-measure of ensemble-based results is greater than the median F-measure of correlation-based results. For each clustering algorithm and ensemble, the vector of F-measures corresponding to the number of clusters were compared to the corresponding set of F-measures from correlation-based clustering results. P-values less than 0.05 appear in bold.

	Algorithm			
Ensemble	UPGMA	Ward	K-means	Walktrap
min	**0.0177**	0.5147	0.4267	**0.0057**
max	0.3696	0.5147	0.5147	0.4267
sum	0.1965	0.3152	0.4852	0.4267

B. Yeast BHI validation

We took the same approach to validation with the BHI. Friedman's test results in Table V showed that there are significant differences ($\alpha = 0.001$) between cluster results for Ward's method, and Walktrap depending on which similarity measure is used.

Table V
Friedman's test results comparing the effect of similarity measures on each clustering algorithm. Each p-value represents a comparison of clustering results gathered using each ensemble and correlation with the algorithm named in the row. P-values significant at $\alpha = 0.05$ appear in bold.

Algorithm	p-value
UPGMA	0.0771
Ward	**0.0001**
K-means	**1.32e-5**
Walktrap	**0.0001**

We further investigated the inconsistencies by performing signed rank tests comparing the clustering results given by one clustering algorithm and each ensemble to the same clustering algorithm and correlation. Table VI shows that ensemble-based results were always significantly better than correlation-based results with the exception of Ward's method and the combination of UPGMA and the min ensemble. Correlation and Ward's method was significantly better than all ensembles at $\alpha = 0.0001$. The best BHI was observed for the Walktrap algorithm with the min ensemble. It was marginally greater than the second best combination, the sum ensemble and UPGMA with a p-value of 0.0525.

Table VI
SIGNED RANK TEST RESULTS TESTING THE HYPOTHESIS THAT THE MEDIAN BHI OF ENSEMBLE-BASED RESULTS IS GREATER THAN THE MEDIAN BHI OF CORRELATION-BASED RESULTS. FOR EACH CLUSTERING ALGORITHM AND ENSEMBLE THE VECTOR OF BHIS CORRESPONDING TO THE NUMBER OF CLUSTERS WERE COMPARED TO THE CORRESPONDING SET OF BHI FROM CORRELATION-BASED CLUSTERING RESULTS. P-VALUES LESS THAN 0.05 APPEAR IN BOLD.

	Algorithm			
Ensemble	UPGMA	Ward	K-means	Walktrap
min	0.1237	1	**5.41e-6**	**5.41e-6**
max	**0.0376**	1	**5.41e-6**	**0.0002**
sum	**0.0177**	1	**5.41e-6**	**0.0007**

Figure 3 shows the BHI across numbers of clusters produced. The figure further supports the signed rank test results. The BHI appears much more erratic in UPGMA results than in any other algorithm. Most of the variation in all algorithms appears to occur in the smaller numbers of clusters and level off as the number increases.

Comparing Figure 2 and Figure 3 we see that both the min and sum ensembles provide the best clustering results overall and provide the best clustering results according to both annotation-based and interaction-based validation methods in UPGMA and Walktrap clusters.

Viable clusters are those that contained enough genes with GO terms and interactions for analysis. With the exception of UPGMA and Walktrap, all clustering experiments resulted in precisely the number of desired viable clusters. Table VII shows the number of viable clusters produced by UPGMA and walktrap with all similarity measures across all numbers of clusters. Clustering experiments that resulted in few viable clusters may be finding interesting clusters of genes that simply do not have the necessary annotation or studied interaction data for validation. Results with very small numbers of viable clusters should be considered suspect because of the lack of validation data. In such cases, validation measures may appear artificially high because all known data about genes occurs in the same few clusters. In this light, the combination of min ensemble and UPGMA should be considered suspect as well as the combination of correlation and Walktrap. However, the results from combination of the min ensemble and Walktrap are still greater than all other results except for the min ensemble and UPGMA.

Table VII
THE NUMBER OF VIABLE CLUSTERS FOR CLUSTERING EXPERIMENTS USING UPGMA AND WALKTRAP AND ALL ENSEMBLE SIMILARITY MEASURES AND CORRELATION IN THE YEAST DATA SET. VIABLE CLUSTERS ARE THOSE THAT CONTAINED ENOUGH GENES WITH GO TERMS AND INTERACTIONS FOR ANALYSIS. ALL CLUSTERS PRODUCED BY WARD'S METHOD AND K-MEANS CLUSTERING WERE VIABLE.

	UPGMA				Walktrap			
Num. clusters	cor	min	max	sum	cor	min	max	sum
10	6	1	2	2	1	6	10	10
20	14	1	9	6	1	13	20	20
30	21	2	18	9	1	13	30	30
40	31	3	28	12	1	18	40	40
50	39	3	37	14	1	21	50	50
60	49	5	47	19	1	28	60	60
70	59	8	57	20	1	29	70	70
80	69	8	66	20	2	33	80	80
90	79	11	76	22	1	34	90	90
100	89	15	86	29	1	38	100	100

C. E. coli *interaction based validation*

Friedman's test results comparing the effect of different similarity measures on each clustering algorithm showed significant ($\alpha = 0.001$) differences in results depending on the similarity measure used for all algorithms.

We tested the hypothesis that the median F-measure of ensemble-based results is greater than the median F-measure of correlation-based results in Table VIII. The max and sum ensembles performed significantly better for UPGMA than correlation at $\alpha = 0.05$. All ensembles appear to not quite significantly outperform correlation for Walktrap clustering. Tests on the opposite hypothesis, that the median F-measure of correlation-based results is greater than the corresponding ensemble-based results, found no significant effects.

Table VIII
SIGNED RANK TEST RESULTS TESTING THE HYPOTHESIS THAT THE MEDIAN F-MEASURE OF ENSEMBLE-BASED RESULTS IS GREATER THAN THE MEDIAN F-MEASURE OF CORRELATION-BASED RESULTS. FOR EACH CLUSTERING ALGORITHM AND ENSEMBLE THE VECTOR OF F-MEASURES CORRESPONDING TO THE NUMBER OF CLUSTERS WERE COMPARED TO THE CORRESPONDING SET OF F-MEASURES FROM CORRELATION-BASED CLUSTERING RESULTS. P-VALUES LESS THAN 0.05 APPEAR IN BOLD.

	Algorithm			
Ensemble	UPGMA	Ward	K-means	Walktrap
min	0.6303	0.4267	0.3979	0.0827
max	**5.41e-6**	0.1763	0.0525	0.0715
sum	**0.0262**	0.3696	0.3696	0.0715

Figure 4 shows the F-measure versus the number of clusters for all similarity measures and all clustering algorithms. The max ensemble does not appear in the UPGMA figure because it is nearly ten times better than the next best results. A large part of UPGMA's success is due to the tendency to produce a single large cluster containing the vast majority

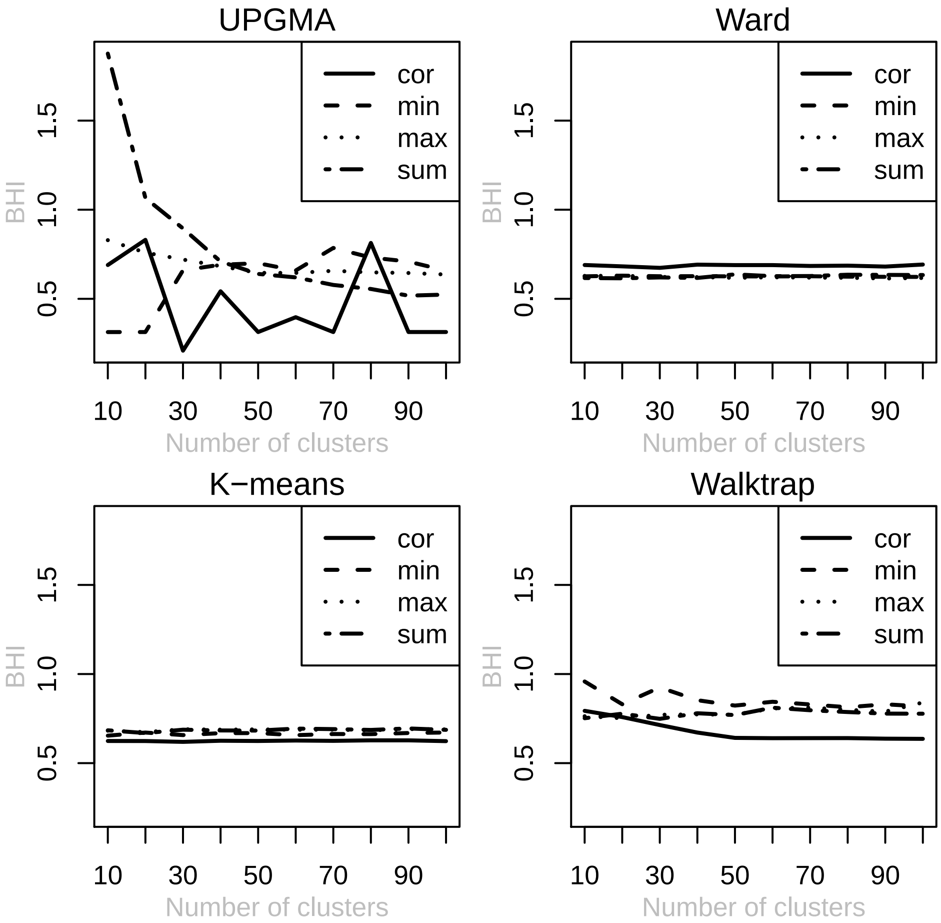

Figure 3. BHI versus number of clusters.

of observations. As a result, the number of viable clusters (those containing known interactions and at least two genes) was less than the number produced with other algorithms. As indicated by Table VIII, the min, max, and sum ensembles perform very well in combination with both UPGMA and Walktrap. The max and sum ensembles also performed well compared to correlation with Walktrap clustering.

D. E. coli *GO annotation based validation*

Friedman's test results for the BHI were similar to the F-measure results in that the effect of different similarity measures on each clustering algorithm showed significant differences in results depending on the similarity measure used for all algorithms at $\alpha = 0.01$.

Table IX shows results for tests of the hypothesis that the median BHI of ensemble-based results is greater than the median BHI of correlation-based results. All ensembles significantly outperformed correlation in UPGMA results at $\alpha = 0.001$. The sum ensemble with Ward's method significantly outperformed correlation and all other ensembles. The min ensemble with the K-means algorithm significantly outperformed correlation and all other ensembles. The best overall BHI was a result of the min ensemble and Walktrap. Signed rank tests of the hypothesis that the median BHI of ensemble-based results is less than the median BHI of correlation-based results showed that all other ensembles performed significantly worse ($\alpha = 0.05$) than correlation except Ward's method with the min ensemble.

Figure 5 shows the BHI versus the number of clusters for all similarity measures and all clustering algorithms. As indicated by Table IX, the min, max, and sum ensembles perform very well in combination with UPGMA.

Table X shows the number of viable clusters produced by UPGMA and walktrap with all similarity measures across all numbers of clusters. In UPGMA results the max ensemble produced a single viable cluster. Therefore the overwhelmingly positive F-measure results for this combination should be considered highly suspect. Walktrap and the sum and max ensembles produced only single viable clusters for most numbers of desired clusters. This data explains why their

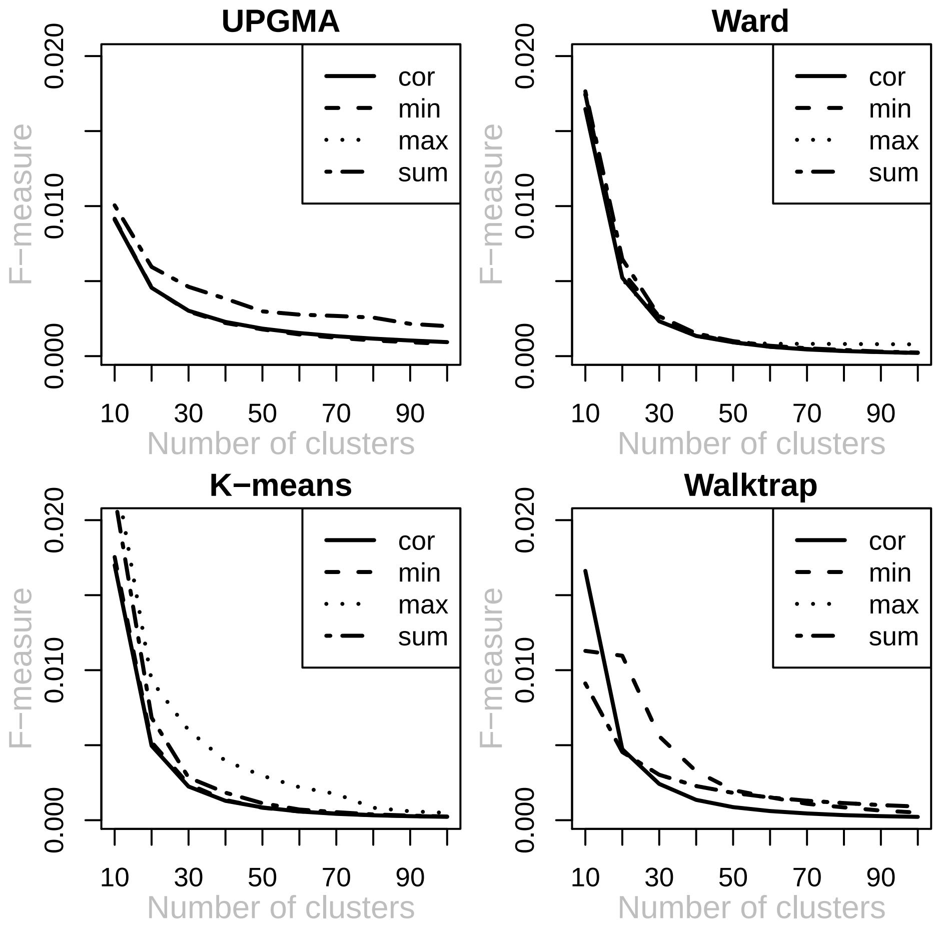

Figure 4. F-measure versus number of clusters.

Table IX

SIGNED RANK TEST RESULTS TESTING THE HYPOTHESIS THAT THE MEDIAN BHI OF ENSEMBLE-BASED RESULTS IS GREATER THAN THE MEDIAN BHI OF CORRELATION-BASED RESULTS. FOR EACH CLUSTERING ALGORITHM AND ENSEMBLE THE VECTOR OF BHIS CORRESPONDING TO THE NUMBER OF CLUSTERS WERE COMPARED TO THE CORRESPONDING SET OF BHIS FROM CORRELATION-BASED CLUSTERING RESULTS. P-VALUES LESS THAN 0.001 APPEAR IN BOLD.

	Algorithm			
Ensemble	UPGMA	Ward	K-means	Walktrap
min	**0.0007**	0.3979	**5.41e-6**	**1.08e-05**
max	**0.0005**	0.9855	1	0.9384
sum	**5.41e-06**	**5.41e-06**	0.9999	0.9384

Table X

THE NUMBER OF VIABLE CLUSTERS FOR CLUSTERING EXPERIMENTS USING UPGMA AND WALKTRAP AND ALL ENSEMBLE SIMILARITY MEASURES AND CORRELATION IN THE *E. coli* DATA SET. VIABLE CLUSTERS ARE THOSE THAT CONTAINED ENOUGH GENES WITH GO TERMS AND INTERACTIONS FOR ANALYSIS. ALL CLUSTERS PRODUCED BY WARD'S METHOD AND K-MEANS CLUSTERING WERE VIABLE.

	UPGMA				Walktrap			
Num. clusters	cor	min	max	sum	cor	min	max	sum
10	1	7	1	2	10	7	2	2
20	1	17	1	8	20	16	2	2
30	2	27	1	11	30	26	1	1
40	5	37	1	12	40	35	1	1
50	7	47	1	16	50	41	1	1
60	8	57	1	17	60	48	1	1
70	10	67	1	17	70	55	1	1
80	13	77	1	20	80	63	1	1
90	13	87	1	25	90	70	1	1
100	16	97	1	27	100	78	1	1

results according to both validation methods are the same. Regardless of these considerations, clustering with ensemble similarity measures outperform clustering with correlation in all cases.

E. Comparing validation measures across organisms

Overall, the interaction-based validation in Figures 2 and 3 show similar trends. In both figures the F-measure decreases

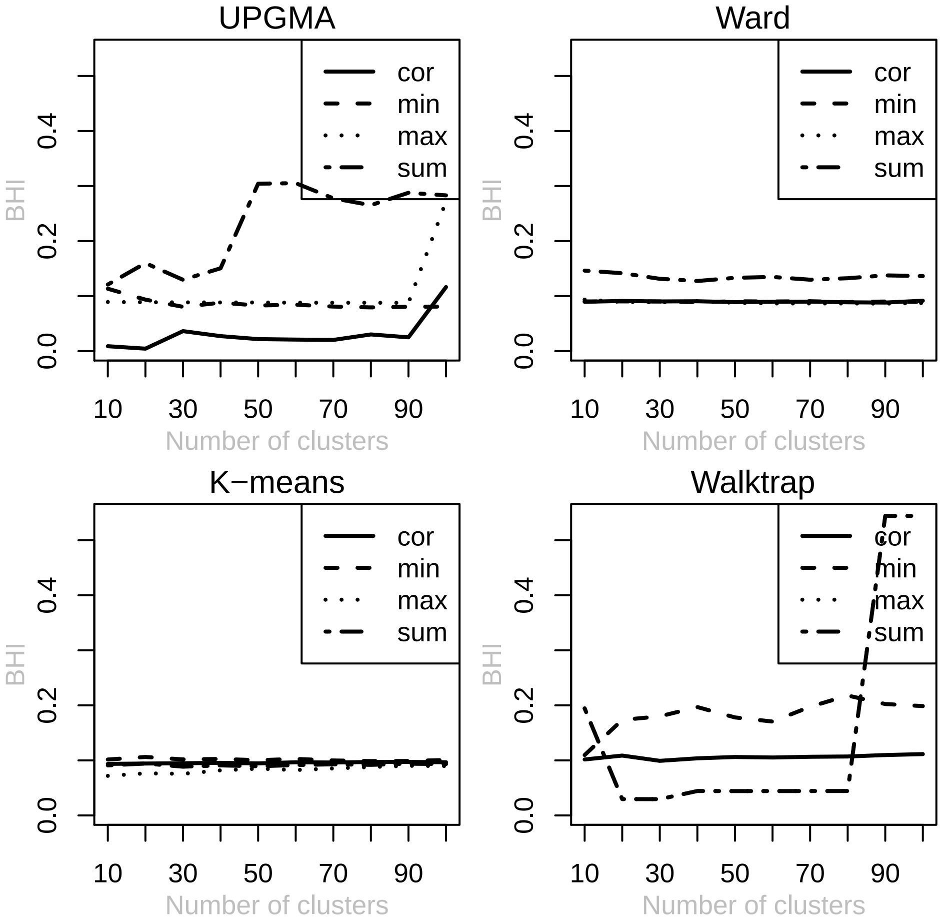

Figure 5. BHI versus number of clusters.

as the number of clusters increases, indicating that the precision and recall grow more unbalanced. In both organisms ensembles perform significantly better than correlation in UPGMA and Walktrap results. The best F-measure in each organism was achieved by using an ensemble similarity measure.

The annotation-based validation in Figures 4 and 5 was similarly uniform across organisms. BHI results for Yeast and *E. coli* agree about the success of ensembles in UPGMA clustering results and the min ensemble in K-means clustering results. They also agree about the poor performance of ensembles in Ward's method. The best BHI was achieved with Walktrap in both organisms.

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 ensemble approach consistently produces better clustering results than correlation alone across organisms and validation types. An important observation to come out of this study is that the most commonly used clustering algorithms in microarray analysis were consistently outperformed by Walktrap. Although UPGMA and K-means clustering are the most commonly used algorithms for exploratory analysis of microarray data, we found that the absolute best results with either validation approach and across organisms were most often achieved by use of the Walktrap algorithm.

The few apparent discrepancies 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 in the case of Yeast or a shared pathway in the case of *E. coli*. 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. The GO analysis of *E. coli* was particularly affected by a lack of annotation as there were only slightly more annotated gene products than there were genes. 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 unreliable (Myers et al. 2006). Although each gold standard or validation method has its own bias it can still be informative.

The importance of the data set used in this procedure cannot be overstated. We attempted the same experiment with Biogrid Yeast and *E. coli* interaction data (Stark, et al. 2006) with less definitive but still encouraging results. The clustering results using ensembles built on the Biogrid interaction set tended to be statistically indistinguishable from clustering results using correlation alone. Although the Biogrid database contains a larger collection of curated interaction data, we feel that the data sets we used were important contributing factors to the success of this approach.

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) 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. iii) It complements and improves upon existing common and successful methods of analyzing high-throughput biological data. iv) Because it creates an ensemble similarity measure rather than altering a clustering approach, it could be used with clustering methods beyond those discussed here.

Our analysis also leads us to the recommendation that Walktrap replace UPGMA and K-means clustering as the *de facto* clustering algorithm of choice for microarray analysis.

References

R. B. Brem, et al. (2002). ‘Genetic Dissection of Transcriptional Regulation in Budding Yeast’. *Science* **296**(5568):752–755.

N. V. Chawla, et al. (2004). ‘Learning Ensembles from Bites: A Scalable and Accurate Approach’. *Journal of Machine Learning Research* **5**:421–451.

K. R. Christie, et al. (2009). ‘Functional annotations for the Saccharomyces cerevisiae genome: the knowns and the known unknowns’. *Trends in Microbiology* **17**(7):286–294.

M. W. Covert, et al. (2004). ‘Integrating high-throughput and computational data elucidates bacterial networks’. *Nature* **429**(6987):92–96.

B. J. Daigle, et al. (2010). ‘Using pre-existing microarray datasets to increase experimental power: application to insulin resistance.’. *PLoS computational biology* **6**(3):e1000718+.

S. Datta & S. Datta (2003). ‘Comparisons and validation of statistical clustering techniques for microarray gene expression data’. *Bioinformatics* **19**(4):459–466.

S. Datta & S. Datta (2006). ‘Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes’. *BMC Bioinformatics* **7**(1):397+.

J. Demsar (2006). ‘Statistical Comparisons of Classifiers over Multiple Data Sets’. *Journal of Machine Learning Research* **7**:1–30.

M. B. Eisen, et al. (1998). ‘Cluster analysis and display of genome-wide expression patterns’. *Proceedings of the National Academy of Sciences* **95**(25):14863–14868.

B. S. Everitt, et al. (2001). *Cluster Analysis*. Wiley, 4th edn.

F. D. Gibbons & F. P. Roth (2002). ‘Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation’. *Genome Research* **12**(10):1574–1581.

J. Handl, et al. (2005). ‘Computational cluster validation in post-genomic data analysis.’. *Bioinformatics* **21**(15):3201–3212.

D. Hanisch, et al. (2002). ‘Co-clustering of biological networks and gene expression data’. *Bioinformatics* **18**(suppl 1):S145–S154.

J. A. Hartigan & M. A. Wong (1979). ‘Algorithm AS 136: A K-Means Clustering Algorithm’. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1):100–108.

T. K. Ho (1998). ‘The random subspace method for constructing decision forests’. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8):832–844.

R. Jansen (2001). ‘Genetical genomics: the added value from segregation’. *Trends in Genetics* **17**(7):388–391.

I. M. Keseler, et al. (2005). ‘EcoCyc: a comprehensive database resource for Escherichia coli.’. *Nucleic Acids Res* **33**(Database issue).

C. Myers, et al. (2006). ‘Finding function: evaluation methods for functional genomic data’. *BMC Genomics* **7**(1):187+.

P. Pons & M. Latapy (2005). ‘Computing Communities in Large Networks Using Random Walks’. In p. Yolum, T. Güngör, F. Gürgen, & C. Özturan (eds.), *Computer and Information Sciences - ISCIS 2005*, vol. 3733, chap. 31, pp. 284–293. Springer Berlin Heidelberg, Berlin, Heidelberg.

J. R. Quinlan (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

A. K. Rider, et al. (2010). ‘A supervised learning approach to the unsupervised clustering of genes’. In *2010 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 323–328, Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA. nchawla@nd.edu. IEEE Computer society.

P. H. Sneath & R. R. Sokal (1962). ‘Numerical taxonomy.’. *Nature* **193**:855–860.

C. Stark, et al. (2006). ‘BioGRID: a general repository for interaction datasets’. *Nucleic Acids Research* **34**(suppl 1):D535–D539.

J. H. Ward (1963). ‘Hierarchical Grouping to Optimize an Objective Function’. *Journal of the American Statistical Association* **58**(301):236–244.