Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Incorporation of biological knowledge into distance for clustering genes

Grzegorz M Boratyn[1*], Susmita Datta[2] and Somnath Datta[2]

[1]Clinical Proteomics Center, University of Louisville, Louisville, KY 40202; [2,3]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY - 40202; Grzegorz M Boratyn* - Email: greg.boratyn@louisville.edu; * Corresponding author

received December 09, 2006; accepted January 20, 2007; published online April 10, 2007

Abstract:

In this paper we propose a data based algorithm to marry existing biological knowledge (e.g., functional annotations of genes) with experimental data (gene expression profiles) in creating an overall dissimilarity that can be used with any clustering algorithm that uses a general dissimilarity matrix. We explore this idea with two publicly available gene expression data sets and functional annotations where the results are compared with the clustering results that uses only the experimental data. Although more elaborate evaluations might be called for, the present paper makes a strong case for utilizing existing biological information in the clustering process.

Availability: Supplement is available at www.somnathdatta.org/Supp/Bioinformation/appendix.pdf

Keywords: knowledge; distance; clustering; genes; expression

Background:

Clustering is routinely used in the exploratory phase of a microarray experiment. Genes are clustered using the pairwise correlation coefficients between two sets of expression profiles as a measure of similarity or closeness. With the growing annotation databases, it is perhaps wise to take advantage of the functional class information of the annotated genes along with the experimental data in grouping genes. Unlike, other approaches (e.g., **[1]**), the purpose of this paper is to outline a general approach of modifying the distance itself. Thus, in a sense, we do not propose a new clustering algorithm - any classical clustering algorithm can be applied based on the new distance (or dissimilarity) matrix. We illustrate this procedure using the agglomerative hierarchical clustering UPGMA and the divisive hierarchical clustering algorithm DIANA applied to two gene expression data sets.

Methodology:

Biological Information

Let $G = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l\}$ be the set of all gene expressions resulting from a microarray experiment, such that $x_i \in R^p$ for some p. Let also $\mathcal{F}_1, \mathcal{F}_2, \ldots\ldots\ldots, \mathcal{F}f$ be not necessarily disjoint sets of labels corresponding to genes with similar biological functions, and $\Im = \bigcup_{j=1}^{f} \mathcal{F}_j$. Note that not all studied genes are functionally annotated, thus $\Im \subset \mathcal{G}$. Let also $\overline{\Im} = \mathcal{G} - \Im$, be the set of unannotated genes. We propose modified distance function that utilizes this prior functional information and promotes clusters of functionally similar genes.

Many clustering algorithms are based on the matrix of distances between each pair of elements (gene expressions). We propose to modify the distance matrix using the information expressed in the sets of biological functions in order to improve clustering results. The presented approach can be used in combination with any distance matrix-based clustering method.

Our new distance (or dissimilarity to be mathematically accurate) combines measurements (gene expressions) and prior information (functional sets). The distance D_{ij} between two genes with expression levels $\mathbf{x}_i$ and $\mathbf{x}_j$ is composed of two parts: 1) the measurement distance d_{ij}^{M} computed with the gene expressions, and 2) functional distance d_{ij}^{F} that is based on the prior biological functional information:

$$\boxed{D_{ij} = d_{ij}^{M} + d_{ij}^{F}} \qquad (1)$$

This new similarity metric (1) corresponds to distances used by semi-supervised clustering techniques in Machine Learning. **[2]** In our case the functional distance plays the role of the similarity-adapting function. Therefore the measurement distance is computed in the same fashion as in the case of standard clustering of gene expressions, and the role of the functional distance is to alter similarities between the gene expressions so that the resulting clusters are in agreement with the functional annotation of genes. Because one of the purposes of cluster analysis of gene expressions is prediction of previously unknown functions of genes, it is desired that: 1) genes with similar functions appear in the same cluster, and 2) genes with unknown functions appear in clusters where majority of genes have known and similar function. In order to satisfy the

presented goals the distance matrix needs to be altered so that:

(1) distance between genes with similar functions is smaller than between genes with different functions,
(2) distance between a pair of annotated and unannotated genes is smaller than between two unannotated genes,
(3) distance between genes with different functions is larger than between two unannotated genes.

Thus the functional distance is composed of the three parts that correspond to the above tasks:

$$\mathbf{d^F} = \mathbf{d^{F1}} + \mathbf{d^{F2}} + \mathbf{d^{F3}} - d_{\min}(\mathbf{J} - \mathbf{I}), \quad (2)$$

Where J is an arbitrary size matrix of ones, I is the identity matrix, and $d_{\min} = \min_{i,j}\{d_{ij}^{M} + d_{ij}^{F1} + d_{ij}^{F3} + d_{ij}^{F3}\}$, added in order to ensure that $D_{ij} = d_{ij}^{M} + d_{ij}^{F} \geq 0, \forall i, j$.

Each element on the right hand side of (2) modifies distances between a group of genes. Computation of d^{F1}, d^{F2}, and d^{F3} is presented below.

Decreasing the distance between genes with similar functions

The often overlooked difficulty in assessing gene similarity is the lack of clear structure in the public gene ontology data bases. Some functions are more general than others and some functions are sub-functions of others. As a solution to this problem we assume that genes are more similar if they share more functions. Let F be a binary matrix that represents gene membership in the functional sets or lack of functional annotation:

$$F_{ik} = \begin{cases} 1 \text{ if } i \in \mathcal{F}_k \text{ or } i \in \bar{\mathfrak{I}} \\ 0 \text{ otherwise,} \end{cases} \quad (3)$$

Then the distance between genes with similar biological functions is modified by the first element of (2) that is given by:

$$d_{ij}^{F1} = \begin{cases} -\mathbf{F}_i \alpha \mathbf{F}_j^T \quad for \quad i \neq j \\ 0 \text{ otherwise,} \end{cases} \quad (4)$$

where α is a diagonal matrix of scaling coefficients for each functional set. The scaling coefficients α are introduced because distance variance may vary across functional sets.

Increasing distance between unannotated genes

Because the purpose of gene expression clustering is to predict functions of genes not studied previously, it is desired that the unannotated genes are placed into clusters composed primarily of genes with known functional annotations. Increasing distances between unannotated genes will result in change of the behavior of a clustering algorithm. The annotated genes will be clustered earlier and will thus form basis of functional clusters to which unannotated genes will later be assigned. Because the only accessible information about unannotated genes are their expressions, $\mathbf{d}^{F2}$ should not alter the relative distances between expressions of unannotated genes. Adding a constant to the distance between each pair of unannotated genes will satisfy the goal without changing the properties of the data. The element modifying the distance matrix that accomplishes this task is equal to:

$$d_{ij}^{F2} = \begin{cases} \beta u_i u_j \quad for \quad i \neq j \\ 0 \quad \text{otherwise,} \end{cases} \quad (5)$$

where u is a binary vector that denotes genes whose biological functions are not known *a priori*:

$$u_i = \begin{cases} 1 \quad if \quad i \in \bar{\mathfrak{I}} \\ 0 \quad \text{otherwise,} \end{cases} \quad (6)$$

and β is a scaling factor that controls the magnitude of the increase of the distance between a pair of unannotated genes.

Increasing the distance between genes with different functions

The previous paragraph presents the modification of distance matrix that increases distances between all pairs of unannotated genes so that clusters composed primarily of genes with unknown functions are avoided. However the set of unannotated genes may contain genes with similar functions that should be placed in the same clusters. The modified distance matrix counteracts assigning twounannotated genes to one cluster. In the absence of functional information of unannotated genes, it is not known how the distances between unannotated genes should be altered. Gene expressions and functional information of annotated genes provide the only accessible knowledge about the unannotated genes. Thus the distances between annotated genes that have different functions will be increased so that unannotated genes with similar gene expressions can be placed in the same cluster. The distance matrix is updated by:

$$\mathbf{d^{F3}} = \gamma(\mathbf{1} - \text{sign}(\mathbf{F}\mathbf{F}^{\mathrm{T}})), \quad (7)$$

where γ is a scaling factor.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net

Selection of parameters

The new updated distance matrix depends on the values of the parameters α, β, and γ. These are found by considering the constraints imposed on the distances between genes that lead to formulation of the functional distance. The mathematical details are provided in the supplement.

In the first step, α is found using the set of annotated genes $\Im$. The functional distance should alter $\mathbf{d}^M$ so that for each function *k*, the expected distance between genes $i \in \mathcal{F}_k$ and $j \in \mathcal{F}_k$ is smaller or equal to distance between $i \in \mathcal{F}_k$ and $j \in \Im - \mathcal{F}_k$ (that does not have function *k*). Thus α is found as a solution of the following linear programming problem:

$$\begin{aligned} &\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \sum_{k=1}^{f} \alpha_k \\ &\text{subject to:} \\ &\sum_{l=1}^{f} \alpha_l \left(\sum_{i \in \mathcal{F}_k, j \in \Im - \mathcal{F}_k} \frac{F_{il} F_{jl}}{N_0^k} - \sum_{i,j \in \mathcal{F}_k, i \neq j} \frac{F_{il} F_{jl}}{N_1^k} \right) \\ &\leq \frac{1}{N_0^k} \sum_{i \notin \mathcal{F}_k, j \in \Im - \mathcal{F}_k} d_{ij}^M - \frac{1}{N_1^k} \sum_{i,j \in \mathcal{F}_k, i \neq j} d_{ij}^M + C \\ &\alpha_k \geq 0 \\ &\text{for } k = 1,2,...,f \end{aligned} \tag{8}$$

where $N_0^k = \left| \{(i,j) : i \in \mathcal{F}_k, j \in \Im - \mathcal{F}_k\} \right|$, $N_1^k = \left| \{(i,j) : i,j \in \mathcal{F}_k, i \neq j\} \right|$, $|\cdot|$ indicates cardinality of a set, and $C \geq 0$ is a user-specified constant that relaxes the constraints in (8) in the case when a data set does not satisfy them with $C = 0$. Arbitrarily large C will result in $\alpha_k = 0$ for each k. In a typical case $C = 0$ and should be increased only if the constraints are not satisfied for $C = 0$.

Let us now consider the remaining two goals presented in the previous section. The maximal distance between genes $i \in \Im$ and $j \notin \Im$ is smaller or equal than the minimal distance between $i \notin \Im$ and $j \notin \Im$. Also, the maximal distance between genes $i \notin \Im$ and $j \notin \Im$ is smaller or equal to the minimal distance between genes that do not share any functions. The above constraints lead to the following expressions for β and γ:

$$\begin{aligned} \beta &= \max\{\beta', 0\} \\ \gamma &= \max\{\gamma', 0\} \end{aligned} \tag{9}$$

where β' and γ' are given by:

$$\beta' = \max_{i \in \mathcal{F}_k, j \in \Im - \mathcal{F}_k, k = 1,2...,f} \left(d_{ij}^M - d_{ij}^{F1} \right) - \min_{i,j \in \Im, i \neq j} \left(d_{ij}^M + d_{ij}^{F1} \right) \tag{10}$$

and

$$\gamma' = \beta + \max_{i,j \in \Im, i \neq j} \left(d_{ij}^M + d_{ij}^{F1} \right) - \min_{i,j \in \Im, \mathbf{F}\mathbf{F}_j^T = 0} \left(d_{ij}^M + d_{ij}^{F1} \right) \tag{11}$$

where $\mathbf{d}^{F1}$ is computed with α resulting from (8). The detailed derivation of the above presented equations is given in the supplemental document.

Note that this constrained distance matrix controls the behavior of a clustering algorithm. The annotated genes that belong to the same functional sets are the closest to one another and thus create basis of clusters. Then genes with unknown functions are assigned to clusters created by annotated genes. Because (5) and (7) increase the distances between unannotated genes and genes that do not share any function by a constant, cluster assignment is delayed but performed on the basis of gene expressions.

Results:

Two illustrative examples of clustering gene expression data are included here. We compare the results of two distance based clustering algorithms – UPGMA and DIANA that utilize the proposed distance matrix $\mathbf{d}^M + \mathbf{d}^F$ with prior functional information with those of the respective clustering algorithms based on distance matrix computed only with gene expressions $\mathbf{d}^M$. We measure biological validity of clustering results and the distribution of functions in gene clusters. Two publicly available sets of gene expressions and functional annotations obtained from public databases are used. The detailed description of data sets and performance measures is presented below.

Data

Two publicly available sets of gene expressions: 1) Yeast time course cDNA, and 2) Normal versus breast carcinoma, SAGE data, are utilized in our illustration.

Yeast time course cDNA data

This set of gene expressions was collected by Chu et al. and presented in **[5].** This data set records expression profiles during sporulation of *Saccharomyces cerevisiae* at seven time points. The original data set was filtered using the

same criterion as in **[5]**. We consider a subset of 513 genes (ORF's to be correct) that were overall positively expressed (i.e., $\sum_{\text{time}}$ log expression ratio > 0).

As in **[4]**, the sets of functional classes were obtained using the web-based GO mining tool at http://mips.gsf.de/proj/funcatDB/search main frame.html. Overall, 342 of the 513 genes were annotated into the following sixteen functional classes: metabolism (121 genes), energy (25), cell cycle and DNA processing (140), transcription (45), protein synthesis (9), protein fate (66), protein with binding function or cofactor requirement (73), protein activity regulation (15), transport (57), cell communication (11), defense (36), interaction with environment (27), cell fate (11), development (11), biogenesis (69), cell differentiation (72).

Normal versus breast carcinoma, SAGE data

The second data set comes from the study presented in **[6]**. We illustrate our methods using the expression profiles of 258 genes (SAGE tags) that were judged to be significantly differentially expressed at 5% significance level between four normal and seven ductal carcinoma *in situ* (DCIS) samples. Abba et al. **[6]** combined various normal and tumor SAGE libraries in the public domain with their own SAGE libraries and used a modified form of *t*-statistics to compute *p*-values. Further details can be obtained from their paper and its supplementary web-site. The functional classes were constructed using a publicly available web-tool called Amigo (http://www. godatabase.org/cgibin/ amigo/go.cgi). As in **[4]**, a total of 113 SAGE tags were annotated into the following eleven classes of molecular function based on their primary biological functions. They were as follows: cell organization and biogenesis (24), transport (7), cell communication (15), cellular metabolism (48), cell cycle (6), cell motility (7), immune response (7), cell death(7), development (5), cell differentiation (5), cell proliferation (5).

Clustering algorithms and distance metric

The Unweighted Pair Group Method with Arithmetic mean (UPGMA) and Divisive Analysis (DIANA) were applied to the two illustrative data sets. The following expression

$$d_{ij}^{M} = \left(1 - r\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)/2, \qquad (12)$$

where $r(\mathbf{x}_i, \mathbf{x}_j)$ is the correlation coefficient of two gene expressions $\mathbf{x}_i$ and $\mathbf{x}_j$, was used as the dissimilarity measure. The values of parameters α, β, and γ utilized for computation of the functional distance (2) for yeast and SAGE data are presented in Tables 1 and 2, respectively. These were calculated using our algorithm stated in the Methods section (c.f., (8)-(11)).

Parameter	Value
α_1	0
α_2	0.0683
α_3	0.0005
α_4	0.0185
α_5	0
α_6	0
α_7	0
α_8	0.0302
α_9	0
α_{10}	0
α_{11}	0
α_{12}	0.0176
α_{13}	0
α_{14}	0
α_{15}	0.1406
α_{16}	0
β	0.9542
γ	1.8764
c	0

Table 1: The values of the parameters α, β, and γ used for computation of functional distance for the yeast gene expressions

In order to demonstrate how the distances between genes change we selected six genes: YBL043W, YBR168W that belong to the biogenesis functional set, YGL210W that belongs to both biogenesis and transport functional groups, YAL067C from transport group, and YAL018C, YBL010C both unannotated. The measurement distances between the listed genes are presented in Table 3. Note that YAL067C that belongs to the transport group is, according to gene expressions, more similar to YBL043W and YBR168W that belong to biogenesis group than to YGL210W, which

also belongs to the transport group. Note also that the two unannotated genes YAL018C and YBL010C are more similar to each other than the genes from the biogenesis or the transport functional group.

The functional dissimilarities between the selected six genes are shown in Table 4. The complete functional distance (2) is not listed in Table 4 so that the change with respect to d^M can be observed. The parameter α_k is equal to 0 for the biogenesis group and 0.0302 for the transport group. The similarities between the genes from the biogenesis group are not changed. The distance between YGL210W and YAL067C (that belong to the transport group) is decreased and that between the unannotated genes YAL018C and YBL010C is increased. The distances between annotated genes that do not share any functions YAL067C and YBL043W, as well as YAL067C and YBR168W are also increased.

The resulting new distances between the selected genes are given in Table 5, where the distance between the unannotated genes YAL018C and YBL010C is larger than between genes that share a function: YBL043W, YBR168W, YGL210W, but smaller than between annotated genes that do not share any function, such as YBL043W and YAL067C.

Parameter	Yeast
α_1	0.0377
α_2	0
α_3	0.1166
α_4	0.0089
α_5	0
α_6	0
α_7	0.01
α_8	0.0322
α_9	0
α_{10}	0
α_{11}	0
β	0.9816
γ	1.9478
c	0

Table 2: The values of the parameters α, β, and γ used for computation of functional distance for the SAGE gene expressions

	YBL043	YBR168	YGL210	YAL067	YAL018	YBL010
YBL043	0.00	0.55	0.45	0.17	0.55	0.46
YBR168	0.55	0.00	0.11	0.26	0.10	0.14
YGL210	0.45	0.11	0.00	0.29	0.08	0.02
YAL067	0.17	0.26	0.29	0.00	0.25	0.29
YAL018	0.55	0.10	0.08	0.25	0.00	0.05
YBL010	0.46	0.14	0.02	0.29	0.05	0.00

Table 3: Measurement dissimilarities d_{ij}^M, between 6 selected genes from the yeast data set that belong to the following functional groups: biogenesis (2), biogenesis and transport (1) , transport (1), unannotated (2)

Performance measures

We compare the performance of the resulting clusterings with the following two measures: 1) distance from model profiles and 2) average proportion of functions in clusters. These quantities are described below. A more extensive comparison along the lines of **[7]** or **[8]** might be possible but is deemed to be beyond the scope of this paper.

Distance from model profiles

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

$$dist = \min_{\pi} \sum_{i=1}^{K} d\left(\bar{\mathbf{x}}_i^m, \bar{\mathbf{x}}_{\pi(i)}\right), \qquad (13)$$

where $d(.,.)$ is a dissimilarity measure, K is the number of clusters and the minimum is taken over all permutations π of

integers {1,2,…..,*K*}, and $\bar{\mathbf{x}}_i^m$ is the (average) model profile for the i-th cluster. The expression (12) for a dissimilarity was also used here. Smaller *dist* indicates that resulting clusters are more similar to the model profiles thus more biologically valid. Datta and Datta **[3]** proposed to use the model profiles as a benchmark for result produced by a clustering algorithm. In the original paper, Chu et al. **[5]**, determined on the basis of first induction of expression that seven is the right number of clusters to be used for grouping genes for this data set. In addition they created a model expression profile by using certain handpicked genes in each class. We use the same number of clusters (K = 7) and the benchmark model profile. The genes used for construction of model profiles have no functional information assigned. The distance from model profiles (13) was computed for the yeast data clustered with UPGMA and DIANA using $\mathbf{d}^{\mathrm{M}}$ and $\mathbf{d}^{\mathrm{M}} + \mathbf{d}^{\mathrm{F}}$ as distance matrices. The resulting values of *dist* are presented in Table 6. The same performance measure was computed for the SAGE data set. The model profiles were composed of genes reported in **[6]**, whose deregulation is altered in the ductal carinoma *in situ* stage of breast cancer. Three model clusters were created from the following functional classes: Cell cycle (3 genes), Apoptosis (3), and Cytokines (4). The values of the distance from model profiles, computed for the SAGE data set clustered with UPGMA and DIANA are presented in Table 7.

Incorporation of the functional information into the distance (dissimilarity) matrix decreased the distance from model profiles in all but one cases (Table 6) indicating a closer agreement with the selected profiles.

	YBL043'	YBR168'	YGL210'	YAL067	YAL018	YBL010
YBL043	0.00	0.00	0.00	1.88	0.00	0.46
YBR168	0.00	0.00	0.00	1.88	0.00	0.14
YGL210	0.00	0.00	0.00	-0.03	-0.03	-0.03
YAL067	1.88	1.88	-0.03	0.00	-0.03	-0.03
YAL018	0.00	0.00	-0.03	-0.03	0.00	0.68
YBL010	0.00	0.00	-0.03	-0.03	0.68	0.00

Table 4: Functional distances not corrected for negative $d_{ij}^{F1} + d_{ij}^{F2} + d_{ij}^{F3}$ between 6 selected genes, from the yeast data set, that belong to the following functional groups: biogenesis(2), biogenesis and transport (1), transport (1), unannotated (2)

	YBL043W	YBR168W	YGL210W	YAL067C	YAL018C	YBL010C
YBL043W	0.00	0.76	0.67	2.26	0.77	0.67
YBR168W	0.76	0.00	0.32	2.35	0.31	0.36
YGL210W	0.67	0.32	0.00	0.47	0.27	0.21
YAL067C	2.26	2.35	0.47	0.00	0.44	0.48
YAL018C	0.77	0.31	0.27	0.44	0.00	0.95
YBL010C	0.67	0.36	0.21	0.48	0.95	0.00

Table 5: Biologically motivated distance $d_{ij}^{M} + d_{ij}^{F}$ between 6 selected genes in Table 3 from the yeast data set that belong to the following functional groups: biogenesis (2), biogenesis and transport (1), transport (1), unannotated (2)

Clustering algorithm	Distance from model profiles for distance matrix	
	$\mathbf{d}^{\mathrm{M}}$	$\mathbf{d}^{\mathrm{M}} + \mathbf{d}^{\mathrm{F}}$
UPGMA	0.1077	0.1218
DIANA	0.0822	0.0604

Table 6: Distance from model profiles computed for the yeast data set clustered with UPGMA and DIANA using measurement and functional distances

Clustering algorithm	Distance from model profiles for distance matrix	
	$\mathbf{d}^{\mathbf{M}}$	$\mathbf{d}^{\mathbf{M}} + \mathbf{d}^{\mathbf{F}}$
UPGMA	0.4753	0.2307
DIANA	0.4887	0.2433

Table 7: Distance from model profiles computed for the SAGE data set clustered with UPGMA and DIANA using measurement and functional distances

Average proportion of functions in clusters

The average proportion of functions in clusters assesses the ability of a clustering algorithm to group genes with similar biological functions into the same clusters. For a given number of clusters K, the proportion of the largest group of genes with common biological function is found in each cluster. The performance measure is given by the average proportions weighted by the number of elements in a cluster:

$$E(K) = \frac{1}{l}\sum_{m=1}^{K}\left|D_m\right| \max_{k=1,2,\ldots,f} \frac{\left|D_m \cap \mathcal{F}_k\right|}{\left|D_m\right|}, \qquad (14)$$

where D_m denotes the m-th cluster of genes. The value of E(K) closer to 1 indicates that a majority of genes in the clusters belongs to one functional set, therefore denotes better clustering performance. Only the genes with known biological functions are used for computation of (14). Note however that all genes are clustered, but only the annotated ones are used for performance assessment. If a gene belongs to more than one functional set, it is considered in finding proportion of all those sets.

Note that E(1) is the proportion of the largest functional group in the entire set of genes under consideration and is independent of a clustering algorithm. A plot of E(K) vs. K can be used to compare the effectiveness of clustering algorithm. Generally speaking, a rapidly increasing curve reaching values close to 1 would indicate better clustering results.

The average proportion of functions in clusters (14), computed for the yeast gene expressions clustered with the UPGMA and DIANA algorithms are presented in Figs. 1 and 2, respectively. The performance measure was computed for K = 1…,10. The average proportion of functions in Fig. 1 for the proposed distance matrix $\mathbf{d}^{\mathrm{M}} + \mathbf{d}^{\mathrm{F}}$ is larger than for the distance $\mathbf{d}^{\mathrm{M}}$ computed only with gene expressions. Note also that the performance of UPGMA with $\mathbf{d}^{\mathrm{M}}$ as a distance matrix is stable for K = 2,…,10. Thus consecutive divisions of clusters into smaller parts do not improve the distribution of functions in clusters. The clustering with $\mathbf{d}^{\mathrm{M}} + \mathbf{d}^{\mathrm{F}}$ as distance matrix, on the other hand, yields monotonically increasing E(K). Therefore as K increases the distribution of functions in clusters improves. Similarly, the performance of the DIANA clustering with the proposed distance matrix is superior to distances computed only with gene expressions (Fig. 2) for K > 1.

The average proportion of functions in clusters was also computed for the SAGE data set. The resulting E(K) for several number of clusters produced with UPGMA and DIANA are presented in Figs. 3 and 4. The biologically motivated distance matrix $\mathbf{d}^{\mathrm{M}} + \mathbf{d}^{\mathrm{F}}$ provides larger E(K), for K > 2, for clusters constructed with UPGMA (Fig. 3) and DIANA (Fig.4) than the gene expression-based distances $\mathbf{d}^{\mathrm{M}}$. Therefore inclusion of prior functional information improves the distribution of functions in clusters.

Discussion:

Although somewhat limited in nature, our studies make astrong case for using semi-supervised clustering whenever possible - one that merges existing biological knowledge with experimental data in grouping genes. A penultimate stage of this approach is available in. **[9]** The present approach has at least two distinct advantages over previous approaches **[1, 9]**: it offers a one step algorithm that determines the appropriate modifications for various categories of genes in an automatic and data based fashion. In addition, since it just modifies the distance (or dissimilarity) matrix, it can be used in conjunction with any dissimilarity based clustering techniques. Furthermore, unlike the approaches presented in **[1]** and **[9]**, we provide analytical as well as computationally inexpensive procedure for parameters selection.

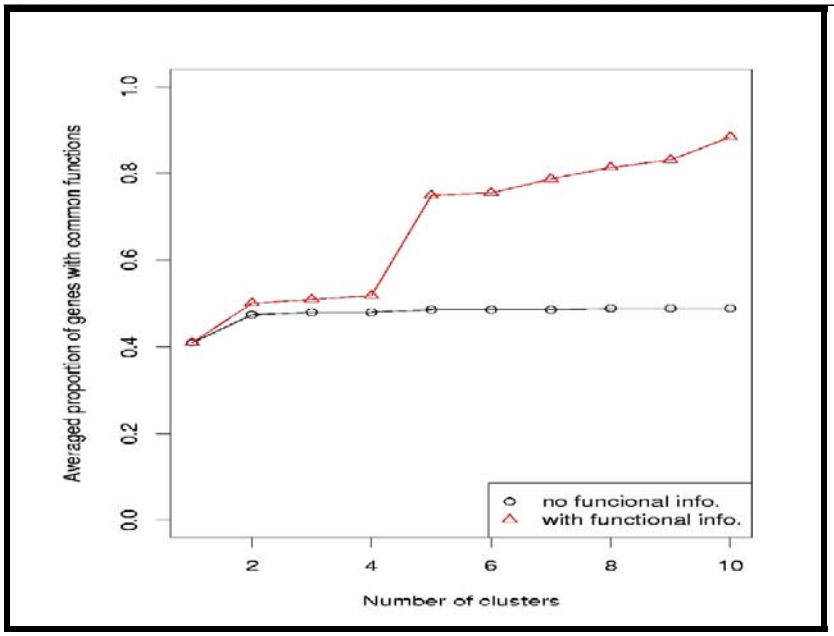

Figure 1: Average proportion of functions in clusters computed for the yeast data clustered with the UPGMA method with (triangles) and without (circles) functional information

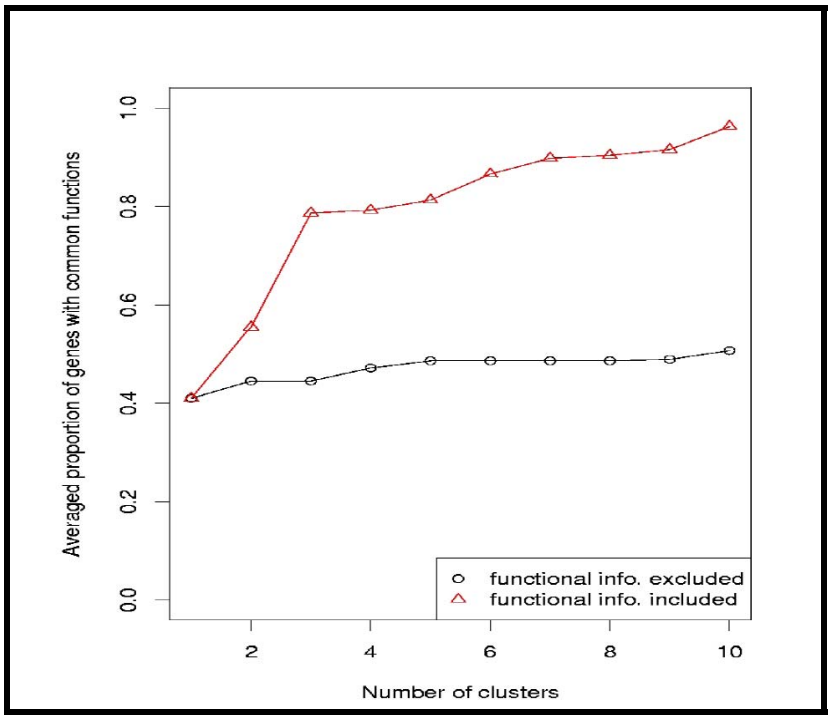

Figure 2: Average proportion of functions in clusters computed for the yeast data clustered with DIANA with (triangles) and without (circles) functional information

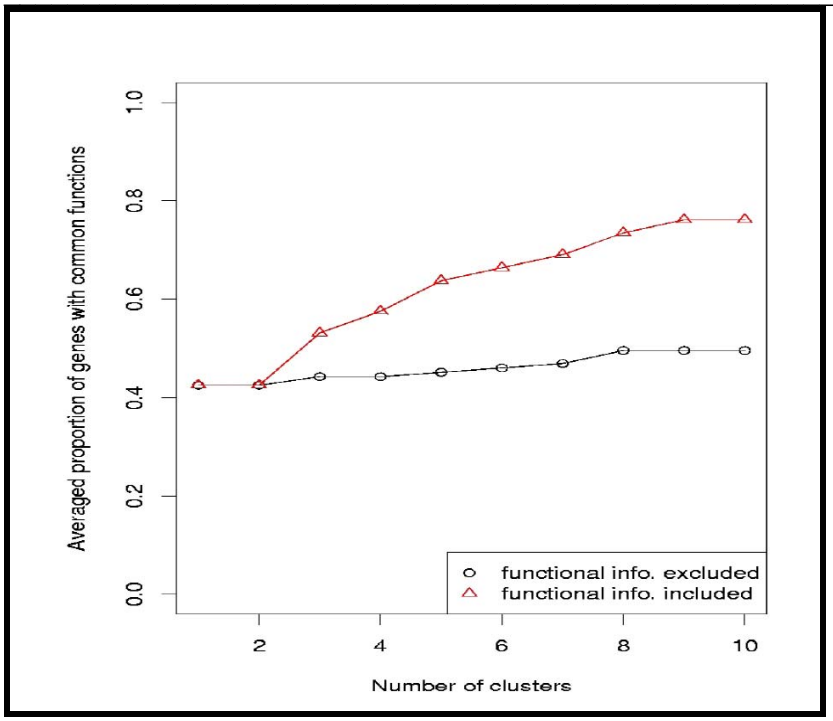

Figure 3: Average proportion of functions in clusters computed for the SAGE data clustered with UPGMA with (triangles) and without (circles) functional information

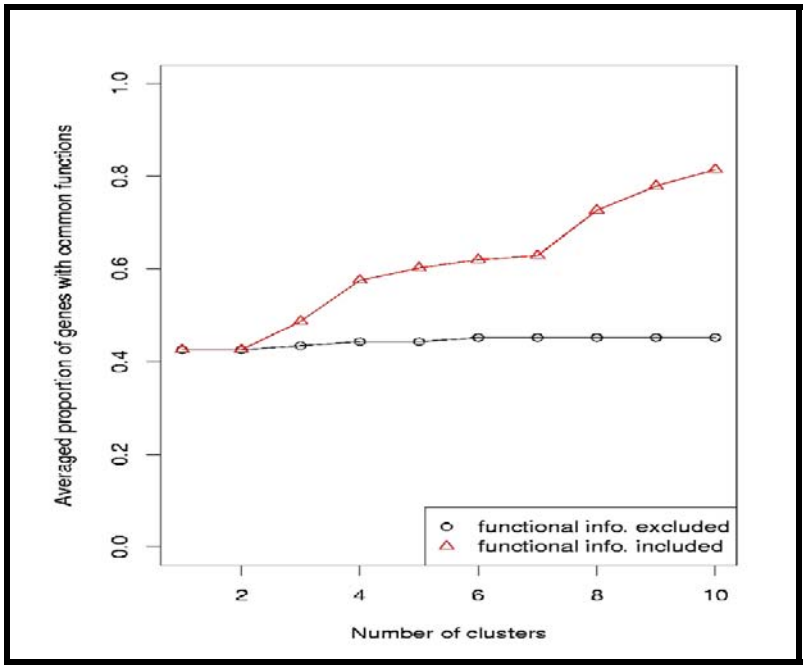

Figure 4: Average proportion of functions in clusters computed for the SAGE data clustered with DIANA with (triangles) and without (circles) functional information

Acknowledgment:
This research was supported by grants from the National Science Foundation (MCB-0517135) and the National Security Agency (H98230-06-1-0062).

References:

[01] D. Huang & W. Pan, *Bioinformatics*, 22:1259 (2006) [PMID: 16500932]
[02] N. Grira, *et al.*, *A Review of Machine Learning Techniques for Processing Multimedia Content*, Report of the MUSCLE European Network of Excellence (2005)
[03] S. Datta & S. Datta, *Bioinformatics*, 19:459 (2003) [PMID: 12611800]
[04] S. Datta & S. Datta, *BMC Bioinformatics*, 7:397 (2006) [PMID:16945146]
[05] S. Chu, *et al.*, *Science*, 282:699 (1998) [PMID: 9784122]
[06] M. C. Abba, *et al.*, *Breast Cancer Res*, 6:R499 (2004) [PMID: 15318932]
[07] J. Handl, *et al.*, *Bioinformatics*, 21:3201 (2005) [PMID: 15914541]
[08] V. Pihur, *et al.*, Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach, preprint (2006)
[09] G. M. Boratyn, *et al.*, *Proc of the 28th IEEE EMBS Annual International Conference*, 1:5515 (2006)

Edited by Susmita Datta
Citation: Boratyn *et al.,* Bioinformation 1(10): 396-405 (2007)