Dynamic Range-Based Distance Measure for Microarray Expressions and a Fast Gene-Ordering Algorithm

Shubhra Sankar Ray,
Sanghamitra Bandyopadhyay, *Senior Member, IEEE*, and
Sankar K. Pal, *Fellow, IEEE*

***Abstract*—This investigation deals with a new distance measure for genes using their microarray expressions and a new algorithm for fast gene ordering without clustering. This distance measure is called "*Maxrange* distance," where the distance between two genes corresponding to a particular type of experiment is computed using a normalization factor, which is dependent on the dynamic range of the gene expression values of that experiment. The new gene-ordering method called "Minimal Neighbor" is based on the concept of nearest neighbor heuristic involving $O(n^2)$ time complexity. The superiority of this distance measure and the comparability of the ordering algorithm have been extensively established on widely studied microarray data sets by performing statistical tests. An interesting application of this ordering algorithm is also demonstrated for finding useful groups of genes within clusters obtained from a nonhierarchical clustering method like the self-organizing map.**

***Index Terms*—Bioinformatics, clustering, combinatorial optimization, data mining, dynamic range, evolutionary algorithm, gene expression, ordering, self-organizing map (SOM), soft computing.**

I. Introduction

The recent advances in DNA array technologies have resulted in a significant increase in the amount of genomic data [1], [2]. The most powerful and commonly used technique is that involving microarray, which has enabled the monitoring of the expression levels of more than thousands of genes simultaneously. Due to the large quantity of information available from microarray, it is necessary to find an appropriate distance measure for genes and to employ a process of classification of the data in order to obtain initial conclusions about the genes.

This investigation deals with the tasks of measuring the distance between genes, their unidirectional ordering without clustering, and ordering within clusters. The widely used measures for finding the similarity between genes are the Pearson correlation and the Euclidean distance. In computing the similarity, all the aforementioned measures do not assign appropriate weights to gene expressions obtained from different types of experiments, where the expressions differ by orders of magnitude from one type to another. Consequently, gene expression values in the lower dynamic range do get dominated by those with higher dynamic range. A new similarity measure between genes called "*Maxrange* distance" is defined in this correspondence, where local (for a particular type of experiment) similarities between two genes are first normalized with a factor dependent on the dynamic range of gene expression values of that experiment (type) and then summed to find a global distance.

Gene ordering [3] is primarily necessary for identifying groups of highly coregulated genes (discussed in detail in Section II-B). Existing methods using evolutionary algorithms [4], [5], local search [4], [5], and Concorde's linear programming [6] for finding the optimal gene order spend most of the time in repetitive searching for the lowest value of the sum of global similarities within gene groups of the same biological category and result in the same biological score for all possible permutations of genes within the same group. To avoid this situation, a fast gene-ordering algorithm called "Minimal Neighbor" (MN), using nearest neighbor (NN) tour construction heuristic and involving $O(n^2)$ time complexity, is described.

The superiority of the proposed *Maxrange* distance measure over related measures is established by using them on three different ordering algorithms and one hybrid algorithm. Similarly, the comparability of the MN algorithm as compared to two existing algorithms is demonstrated for three different distance measures. An interesting application of the MN for ordering genes in the clusters found by the self-organizing map (SOM) is also demonstrated.

II. Existing Approaches

A. Gene Clustering Methods

Clustering methods can be broadly divided into hierarchical and nonhierarchical clustering approaches. Hierarchical clustering approaches (single, complete, and average linkage) [1]–[3] group gene expressions into trees of clusters. They start with singleton sets and merge all genes until all nodes belong to only one set. Nonhierarchical clustering approaches, such as k means [7], SOM [8], and CLICK [9], separate genes into groups according to the degree of distance among genes. The relationships among the genes in a particular cluster generated by nonhierarchical clustering methods are lost.

B. Gene-Ordering Methods

Hierarchical clustering does not determine unique clusters. So, in the framework of hierarchical clustering, a gene-ordering algorithm helps the user to identify subtrees that are clusters by means of visual display and interpret the data [3]. For nonhierarchical clustering-based approaches as well as for hierarchical clustering approaches, microarray gene ordering within clusters using gene expression information is necessary for the following reasons:

1) Gene ordering helps to identify subclusters in big clusters by means of visual inspection of the clustered gene expression data [3].
2) Genes that are adjacent in linear ordering are often functionally coregulated and involved in the same cellular process [1], [2]. Biological analysis is often done in the context of this linear ordering [3].
3) It provides smooth display of clustered genes, where the functionally related genes are nearer in the ordering [2].
4) The relationships among the genes in a particular cluster generated by nonhierarchical clustering algorithms are lost. This relationship (closer or distant) among genes within clusters can be obtained using gene-ordering approaches.

An optimal gene order can be obtained by minimizing the summation of gene expression distances (or maximizing summation of gene expression similarities) between pairs of adjacent genes in a linear ordering $1, 2, \ldots, n$. This can be formulated as [2]

$$F(n) = \sum_{i=1}^{n-1} C_{i,i+1} \tag{1}$$

Manuscript received April 12, 2006; revised August 29, 2006. This paper was recommended by Associate Editor W. Pedrycz.

S. S. Ray and S. K. Pal are with the Center for Soft Computing Research, Indian Statistical Institute, Calcutta 700108, India (e-mail: shubhra_r@isical.ac.in; sankar@isical.ac.in).

S. Bandyopadhyay is with the Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700108, India (e-mail: sanghami@isical.ac.in).

Digital Object Identifier 10.1109/TSMCB.2006.889812

TABLE I
SUMMARY FOR DIFFERENT MICROARRAY DATA SETS

Dataset	No. of genes	Category	Experiments performed				Total
Cell Cycle	652	MIPS 16	Cell Cycle (-1.2 to 1.2) 93	sporulation (-3.0 to 3.0) 9	shock (-1.5 to 1.5) 56	diauxic shift (-2.0 to 2.0) 26	184
Yeast Complex	979	MIPS 16	Cell Cycle (-1.2 to 1.2) 18+14+15	sporulation (-3.0 to 3.0) 7+4	shock (-1.5 to 1.5) 6+4+4	diauxic shift (-2.0 to 2.0) 7	79
All Yeast	6221	MIPS 18	Cell Cycle (-1.2 to 1.2) 60	sporulation (-3.0 to 3.0) 13	diauxic shift (-2.0 to 2.0) 7		80
Fibroblast	517	GO 1347	Serum response (-3.0 to 3.0) 12	cycloheximide (-3.0 to 3.0) 6			18
Herpes	106	GeneBank 5	No KSHV (-13.0 to 13.0) 1	-TPA (-13.0 to 13.0) 7	TPA (-13.0 to 13.0) 13		21

where n is the number of genes, and $C_{i,i+1}$ is the distance/similarity between two genes i and $i+1$ obtained from the distance/similarity matrix.

A hybrid method (first clustering then ordering) for ordering genes for a hierarchical clustering solution is proposed in [3]. A method for ordering genes for a nonhierarchical clustering solution is currently missing. Although gene-ordering methods exist (described in the next paragraph), the utility and application of these methods to individual clusters of nonhierarchical solution are not reported. In the current investigation, the summation of gene expression distances for a nonhierarchical solution is defined as

$$F_1(n) = \sum_{j=1}^{k} \sum_{i=1}^{n_{j-1}} C^{j}_{i,i+1} \tag{2}$$

where k is the total number of clusters, n_j is the number of genes in cluster j, and $C^{j}_{i,i+1}$ is the distance/similarity between two genes i and $i+1$ in cluster j obtained from the distance/similarity matrix.

Tsai *et al.* [4] formulated the gene-ordering problem as a travelling salesman problem (TSP). Concorde's TSP solver [6] can obtain the optimal solutions to 107 of the 110 TSPLIB [10] instances; the largest having 15 112 cities. Thus, Concorde appears to be the best TSP solver currently available, and in Section V, comparisons of results for gene ordering are shown with Concorde. Related works on gene ordering are also available in [5] and [11].

III. MATERIALS AND METHODS

A. Preliminary Concepts of Microarray Technology

Fluorescence is currently the predominant method for microarray signal detection [12]. A critical component of a fluorescence scanner is the photomultiplier tube (PMT), in which fluorescent photons produce electrons that are amplified by the PMT gain. For many microarray scanners, the calibration curve (i.e., the curve showing the relationship between dye concentration and fluorescence intensity) depends on the PMT gain setting [12]. This PMT gain is also varied for different types of experiments of different biological origin. DNA microarray measurements normally assume a linear relationship between the detected fluorescent signal and the concentration of the fluorescent dye that is incorporated into the clone DNA or RNA molecules synthesized from the test sample. Each PMT has its own linear dynamic range within which signal intensity increases linearly with the increase of fluorescent dye concentration [12]. This linear dynamic range also fixes the dynamic range of the recorded microarray data (log ratio values) [12] within which the data values are most reliable and used as the normalization factor in the proposed distance measure to remove variations of biological origin. For example, in Cell-Cycle-related experiments, for dye Cy5, the PMT gain at 960 V fixes the intensity range from x1 to x2, and for dye Cy3, the PMT gain at 760 V fixes the intensity range from y1 to y2. So the linear dynamic range of PMT fixes the linear dynamic range of the data from $\log_2 x1/y1$ to $\log_2 x2/y2$. Note that this dynamic range is available either from the supplementary information (website) of the article/data (Yeast data) or upon request to the authors (Herpes data) and not from the data sets, and hence is not sensitive to outliers. However, due to the wide concentration range for genes expressed in a biological sample, the detected fluorescence intensity does not necessarily remain in the linear range for all genes tiled on a microarray. The proposed dynamic range-based normalization (described in Section III-C) belongs to the category of between-slide or multiple-slide normalization [13]. The two other normalization factors in this category, which aim to allow experiment-to-experiment comparisons when different types of experiment have substantially different spreads in log ratios, are median absolute deviation (MAD) and variance regularization. The two normalization methods, viz., MAD and variance regularization, are dynamic range estimators (not the real one) and implemented for the purpose of comparison. However, the results obtained were not very encouraging.

B. Description of Data Sets

For gene ordering, data sets like Cell Cycle [14], Yeast Complex [1], [3], All Yeast [1], [15], Fibroblast [16], and Herpes [17] are chosen. Table I shows the name of the data sets, number of genes in each data set, number of gene categories, name of experiment types and number of experiments performed under each type, and total number of experiments performed for a particular data set. The dynamic range of expression values of each experiment type is shown within parentheses. The dynamic range of available data represents log ratios of -1.2 to 1.2 for the cell-cycle experiments, -3.0 to 3.0 for sporulation, -1.5 to 1.5 for the shock experiments, -2.0 to 2.0 for the diauxic shift, -3.0 to 3.0 for Fibroblast data, and -13.0 to 13.0 for Herpes data. Herpes data are generated using radioactive probes instead of fluorescent probes, and hence, a higher linear dynamic range is observed compared to other data sets. The first three data sets of

Saccharomyces cerevisiae are classified into 16, 16, and 18 groups, respectively, according to the Munich Information for Protein Sequences (MIPS) [18] categorization. The genes in Fibroblast data are classified into 1347 categories according to the Gene Omnibus annotation. In Herpes data, the genes are broadly assigned to five functional groups and available in [17]. For the Cell-Cycle data, first, we downloaded 652 Cell-Cycle-regulated gene names from the MIPS website. These gene names were then uploaded in the Stanford Microarray Database [14], and corresponding gene expression values are downloaded with default parameters by selecting all the cell cycle, sporulation, heat shock, and diauxic shift experiments. Microarray experiments often produce multiple missing expression values, normally due to various experimental problems. In this correspondence, all the genes with more than 50% missing gene expression values are first eliminated from the data set. Thereafter, for the remaining genes, missing gene expression values are estimated using LSimpute [19] software, a statistical java-based package to estimate missing values.

C. New Distance Measure

A number of measures of distance in studying the behavior of two genes can be used, such as Manhattan [20], Euclidean [20], and Pearson correlation distance [2]. Pearson correlation is oversensitive to large threefold changes (peaks) in gene expression profiles due to multiplication of expression vectors in dot product style and therefore leads to false interpretation of distance between genes in certain cases. Moreover, it is observed that often microarray data consist of different sets of expression values corresponding to different experiment types. Existing distance measures usually take the same normalization factor (like standard deviation for Pearson correlation) for a gene. This normalization factor is independent of the type of experiment, varies from gene to gene, and performs global normalization to all the expression values for a particular gene, thus loosing useful local information. But a closer look at the gene expression data reveals that the dynamic range of expression values differs with the type of experiment and remains the same for all the genes in the data set. So, using the same normalization factor is undesirable for all types of experiments, where expression values differ by orders of magnitude from one kind of experiment to another. Consequently, it may be appropriate and better if normalization is performed

- separately for the different types of experiment with different normalizing factors; thereby preserving the local information;
- keeping the same set of normalization factors for all the genes in the data set.

Such an attempt is made in this correspondence, where two new distance measures are developed using Manhattan distance and Euclidean distance, respectively (to avoid oversensitivity to threefold changes), in which normalization is dependent on the type of experiment. This, in turn, results in equal weighting of distance values for different experiment types. The normalization factor is chosen as the linear dynamic range of data values obtained from PMT for a particular type of experiment.

Let

$$X = x_1^{e_1}, \dots, x_{i_1}^{e_1}, x_1^{e_2}, \dots, x_{i_2}^{e_2}, \dots, x_1^{e_m}, \dots, x_{i_m}^{e_m}$$
$$Y = y_1^{e_1}, \dots, y_{i_1}^{e_1}, y_1^{e_2}, \dots, y_{i_2}^{e_2}, \dots, y_1^{e_m}, \dots, y_{i_m}^{e_m}$$

be the expression vectors (levels) of the two genes in terms of log-transformed microarray gene expression data obtained over a series of m different types of experiment $(e_1, e_2, \dots, e_m)$ consisting of $i_1 + i_2 + \cdots + i_m$ experiments in total. Using Manhattan distance, the *Maxrange* distance between X and Y is defined as

$$Maxrange - M_{X,Y} = \frac{1}{m}\sum_{r=1}^{m}\frac{1}{i_r} \times \frac{\sum_{j=1}^{i_r}\left|x_j^{e_r} - y_j^{e_r}\right|}{\text{Max}_{e_r} - \text{Min}_{e_r}} \quad (3)$$

where Max_{e_r} and Min_{e_r} are the maximum and minimum $\log_2(R/G)$ values obtained from the linear dynamic range of the PMT (or radioactive probe) for an experiment of type e_r.

The following can be stated about the measure:

1) $0 \leq Maxrange - M_{X,Y} \leq 1$;
2) $Maxrange - M_{X,Y} = 0$ if and only if $X = Y$;
3) $Maxrange - M_{X,Y} = Maxrange - M_{Y,X}$ (symmetric).

Using the Euclidean distance, the *Maxrange* distance between X and Y is defined as

$$Maxrange - E_{X,Y} = \frac{1}{m}\sum_{r=1}^{m}\frac{1}{i_r} \times \frac{\sqrt{\sum_{j=1}^{i_r}\left(x_j^{e_r} - y_j^{e_r}\right)^2}}{\text{Max}_{e_r} - \text{Min}_{e_r}}. \quad (4)$$

Throughout the literature, we have used *Maxrange-M* and *Maxrange-E* for representing *Maxrange* distance measure using Manhattan and Euclidean distance, respectively.

Let three genes X, Y, and Z with four different types of experiments have the gene expression values $X = 0.02, -0.1, 2.9, 0.1, -0.1, 0.1, -0.15, 0.1,$ $Y = 0.1, -0.05, 0.15, -0.2, -0.3, 0.64, 0.0, 0.3$, and $Z = 0.13, -0.09, 0.1, -0.2, 1.2, 1.2, 1.7, 1.9$.

Assume that the first two expression values for all the genes correspond to cell-cycle experiments with dynamic range between 1.2 and -1.2, the third and fourth values correspond to sporulation experiments with dynamic range between 3.0 and -3.0, the fifth and sixth values correspond to shock experiments with dynamic range between 1.5 and -1.5, and the seventh and eighth values correspond to diauxic shift experiments with dynamic range between 2.0 and -2.0. So, the *Maxrange-M* distance and the Pearson correlation distance between genes X and Y are 0.11208 and 0.85202, respectively.

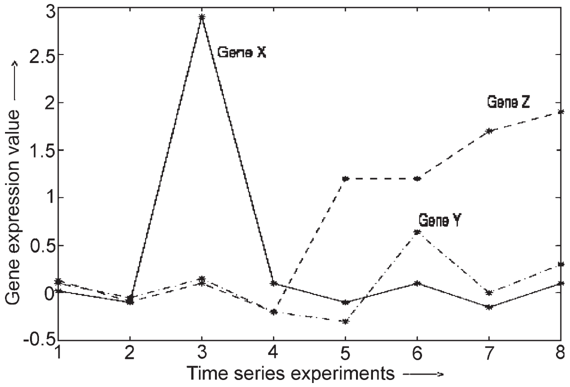

Fig. 1. Expression profile for three genes. According to *Maxrange-M*, the distance between genes X and Y is smaller than Z and Y, which is in opposition with Pearson correlation and Euclidean distance.

To illustrate the difference between *Maxrange-M* and Pearson correlation, consider Gene X and Gene Y in Fig. 1, which shows two profiles (of length 8), which are highly similar according to the *Maxrange-M* but almost dissimilar (uncorrelated) according to Pearson correlation. This is mainly due to the comparatively large value of the threefold change in Gene X. As opposed to this, in *Maxrange-M*, sensitivity to threefold change is avoided using Manhattan distance, and normalization with a dynamic range of experiments correctly reflects the fact that both profiles have similar expressions for three types of experiments, namely cell cycle, shock, and diauxic shift, and differs in only one expression (among two expressions) for sporulation experiments. *Maxrange-E* distance also shows similar performance as *Maxrange-M*. The Euclidean distance between X and Y is 2.8382, and between Y and Z, it is 2.8317. But X differs with Y in only one expression value of high-range experiment type $(\mathrm{Max}_{e_r} - \mathrm{Min}_{e_r} = 6)$, whereas Z differs with Y in three expression values of relatively small-range experiment type. So in the case of Euclidean distances, experiment types with high range dominate experiment types with small-range ones. As opposed to these, the *Maxrange-M* distance between X and Y is 0.11208, which is less than the distance between Y and Z (0.19365). The *Maxrange-E* distance between X and Y is also less than the distance between Y and Z.

D. New Ordering Algorithm

Existing methods, using evolutionary algorithms [4], [5] for finding the optimal gene order, spend most of the time in repetitive searching for the lower value of the sum of gene expression distances in gene groups (genes belonging to same category) and result in the same biological score for all possible permutations of genes within the same group. Under this situation, to avoid repetitive searching, the NN tour construction heuristic can be used to find a near-optimal gene order in terms of gene expression distance. The NN tour has the advantage that it commits only a few severe mistakes in tour construction, while there are long segments connecting nodes with short edges. It has a disadvantage that several genes that are not considered during the course of the algorithm are inserted at high costs in the end. To overcome this to some extent, we propose a new heuristic-based MN algorithm.

Let $1, 2, \ldots, i, \ldots, n$ represent the indices of n genes in the microarray data set, and let the distance between gene i and $i + 1$ be denoted as $C_{i,i+1}$. Given this microarray data set of n genes to be ordered and pairwise distance/similarity (of each gene with all other genes) kept in an $n \times n$ matrix (after calculating), the different steps of applying MN are explained below.

Step 1) Find the closest (most similar) pair of genes and merge them into a single array (string) so that there remains $n - 2$ genes.

Step 2) Consider only the two end genes of the new array and find the two closest genes for each of them from the remaining genes. Out of these two selected genes, find the one closer to one of the end genes of the array and then place it next to that. The other selected gene is not connected and kept with the remaining genes. The index of this gene is stored for use in the next step. (Note that if both the selected genes are the same in this step, then no gene index can be stored and in the next step we have to compute twice for the selection of two genes, else, only one closest gene is needed to be computed.)

Step 3) Repeat Step 2) until all genes are aligned into a single array of size n.

The computational complexity of Step 1) is $O((n/2)^2)$ as the distance matrix is a symmetric one. This step can also be performed during the calculation of $n \times n$ distance matrix. For Steps 2)–3), the worst case complexity is $O(2 * (n - 2) * n)$. So the total complexity of the algorithm is $O(n^2)$.

E. New Hybrid Algorithm for Ordering Genes in Nonhierarchical Clustering

It is mentioned in Section II-B that a method for ordering genes for a nonhierarchical clustering solution is currently missing, and that the utility and application of existing gene-ordering methods to individual clusters of nonhierarchical solution are not reported in literature. Here, we propose a simple hybrid algorithm where MN is applied separately on each of the gene clusters found by SOM to identify subclusters within large clusters and to group functionally correlated genes within clusters. This algorithm is referred to as "SOM + MN." The number of nodes/clusters of SOM is chosen according to MIPS categories for Yeast data and available information in relevant literature for Fibroblast and Herpes data. This hybrid method is proposed to show the efficiency of MN in improving the solution quality of a nonhierarchical solution in a computationally effective way.

IV. Biological Interpretation

A biological score, which is different from the similarity/distance measures, is used to evaluate the final gene ordering. Each gene that has undergone MIPS categorization can belong to one or more categories, while there also are many unclassified genes (no category). A vector $V(g) = (v_1, v_2, \ldots, v_j)$ is used to represent the category status of each gene g, where j is the number of categories. The value of v_j is 1 if gene g is in the jth category and 0 otherwise. Based on information about categorization, the score of a gene order for multiple-class genes is defined as [4]

$$S(n) = \sum_{i=1}^{N-1} G(g_i, g_{i+1}) \tag{5}$$

where N is the number of genes, g_i and g_{i+1} are the adjacent genes, and $G(g_i, g_{i+1})$ is defined as

$$G(g_i, g_{i+1}) = \sum_{k=1}^{j} V(g_i)_k V(g_{i+1})_k \tag{6}$$

where $V(g_i)_k$ represents the kth entry of vector $V(g_i)$. Note that $S(n)$ can also be used as the scoring function for single-class genes. Using scoring function $S(n)$, a gene ordering would have a higher score when more genes within the same group are aligned next to each other.

V. Experimental Results

The algorithms of gene ordering and clustering are implemented using mex files in Matlab 7 on Sun Fire V 890 (1.2 GHz and 8 GB RAM). The codes for single, average, and complete linkage and the method of Bar-Joseph *et al.* [3] are downloaded from [21]. The performances of the proposed *Maxrange-M* and *Maxrange-E* distance are compared with Pearson correlation, Euclidean distance, and Manhattan distance, whereas the MN algorithm for gene ordering is compared mainly with Concorde's linear programming [6] algorithm. SOM is used with 16, 16, 18, 6, and 5 nodes (clusters) for clustering Cell Cycle, Yeast Complex, All Yeast, Fibroblast, and Herpes data, respectively. Finally, MN is applied separately on the gene clusters obtained by SOM in the new hybrid algorithm (SOM + MN).

A. Comparative Performance of Algorithms and Distance Measures

Table II shows the summation of gene expression distances in terms of $F(n)$ (computed using (1) for Concorde, MN, and Bar-Joseph)

TABLE II
SUMMATION OF GENE EXPRESSION DISTANCES COMPUTED IN TERMS OF $F(n)$ [(1) FOR CONCORDE, MN, AND BAR-JOSEPH] AND $F_1(n)$ [(2) FOR SOM + MN] VALUE FOR DIFFERENT ORDERING ALGORITHMS (ALGO.) AND DISTANCE MEASURES (DIST.)

		Data Sets				
Dist.	Algo.	Cell cycle	Yeast comp.	All Yeast	Fibro-blast	Herpes
1	Concorde	69.00	80.87	463.85	26.21	2.73
	MN	71.50	84.31	481.69	27.87	2.89
	B-joseph	71.67	84.75	493.51	27.87	2.80
	SOM+MN	73.91	88.27	505.12	29.27	3.12
2	Concorde	286.51	306.14	1773.15	71.82	11.69
	MN	298.25	327.92	1874.23	81.53	12.37
	B-joseph	300.51	330.17	1920.82	81.71	12.12
	SOM+MN	323.67	361.17	1970.16	99.96	14.34
3	Concorde	3913.9	3244.7	20302.2	851.9	419.5
	MN	4039.4	3386.1	21101.7	902.2	431.7
	B-joseph	4051.4	3388.9	21530.4	897.3	431.4
	SOM+MN	4197.9	3518.3	21525.8	943.3	475.2

and $F_1(n)$ value (computed using (2) for SOM + MN) with (1) *Maxrange-M*, (2) Pearson correlation, and (3) Euclidean distance for all the data sets and four ordering algorithms. Hereafter, the serial numbers of these distances are used to denote them in the tables. In this comparative study among ordering algorithms, Concorde provides the lowest sum of gene expression distances in terms of $F(n)$ (1) value for all the distance measures and data sets, although it has the highest computational complexity ($O(2^n)$). MN and Bar-Joseph's algorithm provide comparable results in terms of $F(n)$ value.

The ultimate goal of an ordering algorithm is to order the genes in a way that is biologically meaningful. In this regard, Table III compares the performance of our proposed approach with those of the other ordering methods in terms of the S value (5). Three distance measures are considered, namely: 1) *Maxrange-M*; 2) Pearson; and 3) Euclidean. The biological scores corresponding to Manhattan distance are found to be comparable to those for Pearson correlation distance and hence omitted here. The percentages of improvement over the lowest biological score (in terms of S value) in a particular data set are shown within parentheses and defined as

$$PI_{i,j} = \frac{d_{i,j} - \min_i(d_{i,j})}{\min_i(d_{i,j})} \times 100 \qquad (7)$$

where $d_{i,j}$ indicates the biological score (S value) in the ith row and jth column of the result matrix in the concerned tables (Tables III and IV), and $\min_i(d_{i,j})$ indicates the minimum biological score in column j for all i.

Table IV shows the performance of our proposed approach "SOM + MN" with respect to SOM alone for the same set of parameters. These PI values in Tables III and IV are used in the next section for conducting t-tests.

TABLE III
BIOLOGICAL SCORE AND PERCENTAGE OF IMPROVEMENT (PI) VALUE (WITHIN PARENTHESES) FOR DIFFERENT GENE-ORDERING ALGORITHMS (ALGO.) AND DISTANCE (DIST.) MEASURES

		Data Sets			
Algo. & complexity	Dist.	Cell cycle	Yeast comp.	All Yeast	Herpes
Concorde	1	420 (10.24)	1089 (11.69)	2383 (6.05)	44 (29.41)
	2	400 (4.99)	1039 (6.56)	2350 (4.58)	34 (0.00)
$O(2^n)$	3	400 (4.99)	1051 (7.79)	2415 (7.48)	40 (17.65)
MN	1	425 (11.55)	1075 (10.26)	2388 (6.28)	43 (26.47)
	2	403 (5.77)	1031 (5.74)	2349 (4.54)	37 (8.82)
$O(n^2)$	3	406 (6.56)	1010 (3.59)	2382 (6.01)	40 (17.65)
B-Joseph	1	423 (11.02)	1074 (10.15)	2371 (5.52)	43 (26.47)
et.al.	2	381 (0.00)	1024 (5.03)	2350 (4.58)	38 (11.76)
$O(n^4)$	3	421 (10.50)	1013 (3.90)	2346 (4.41)	40 (17.65)
SOM+MN	1	409 (7.35)	1039 (6.56)	2335 (3.92)	42 (23.53)
	2	386 (1.31)	1002 (2.77)	2302 (2.45)	38 (11.76)
$O(n^2)$	3	381 (0.00)	975 (0.00)	2247 (0.00)	37 (8.82)

TABLE IV
BIOLOGICAL SCORE AND PERCENTAGE OF IMPROVEMENT (PI) VALUE (WITHIN PARENTHESES) FOR "SOM + MN" AND SOM

		Data Sets			
Algo. & complexity	Dist.	Cell cycle	Yeast complexes	All Yeast	Herpes
SOM+MN	1	409 (13.30)	1039 (15.19)	2335 (14.40)	42 (20.00)
$O(n^2)$	2	386 (6.93)	1002 (11.09)	2302 (12.79)	38 (8.57)
	3	381 (5.54)	975 (8.09)	2247 (10.09)	37 (5.71)
SOM	1	389 (7.76)	973 (7.87)	2100 (2.89)	40 (14.29)
	2	369 (2.22)	944 (4.66)	2073 (1.57)	36 (2.86)
$O(n^2)$	3	361 (0.00)	902 (0.00)	2041 (0.00)	35 (0.00)

For Fibroblast data, no biological score can be provided as genes in the same biological group for these data are rare. For each of the distance measure and any algorithm, the biological scores (in terms of S value) obtained using MAD (or variance regularization factor) normalization are found to be inferior to the biological scores with *Maxrange* normalization and hence are not provided here. Although in most cases, *Maxrange-E* distance is found to be superior to Euclidean distance and inferior to *Maxrange-M*; for All Yeast data, it performs better ($S(n) = 2431$) than *Maxrange-M* ($S(n) = 2388$) for the MN algorithm. However, the superiority of *Maxrange-M* is evident when different types of experiments are present in a particular microarray data. For example, superior results are obtained with *Maxrange-M* for most of the available algorithms for the Cell Cycle, Yeast Complex, and All Yeast data sets (shown in first row for each algorithm in Table III). The available measures for gene distance, like Manhattan distance, Euclidean distance, and Pearson correlations, are suitable for

TABLE V
RESULTS OF t-TEST FOR DIFFERENT PAIRS OF DISTANCE MEASURES

	Pairs of distance measure	
	Maxrange-M & Pearson	*Maxrange-M* & Euclidean
t	3.4247	2.1563
p	$0.001 > p$	$0.02 > p$

the same type of experiments in microarray data, but they are unable to assign different weights of distance for different types of experiments. In contrast, the *Maxrange-M* and *Maxrange-E* distance provides this flexibility, and hence, better results are obtained for multiple types of experiments.

B. Statistical Analysis of Maxrange-M Distance Measure and MN Ordering Algorithm

To statistically compare the performance of *Maxrange-M* distance with Pearson correlation in the case of ordering algorithms, t-tests are performed with the PI (7) values shown within parentheses in Table III using

$$t = \frac{\overline{PI_1} - \overline{PI_2}}{\sqrt{\frac{VariancePI_1}{n_1} + \frac{VariancePI_2}{n_2}}} \tag{8}$$

where $\overline{PI_1}$ and $VariancePI_1$ are the mean and the variance of all the available PI values for *Maxrange-M* distance in Table III. PI_2 is used for Pearson correlation and $n_1 = n_2 = 16$, as there are 16 PI values available in total from Table III for each of the distance measures with four data sets and four algorithms. So, the degrees of freedom for t-test are $16 \times 2 - 2 = 30$. Similarly, t-test is also performed for *Maxrange-M* distance and Euclidean distance. The two t values and related p values are shown in Table V. The alternative hypothesis (H_1) that the average of "percentages of improvement over the lowest biological score" for the *Maxrange-M* distance is better than the related one (Pearson or Euclidean) is used in the calculation of t-statistics. After finding the p values (from t-table) for corresponding t values, we reject the null hypothesis for both cases with significance level of 0.001 and 0.02, respectively, which suggests that there is strong evidence against the null hypothesis in favor of the alternative.

Similar types of t-tests for the MN and related algorithm (Concorde or Bar-Joseph) are also performed with the percentages of improvement shown in Table III. The results are shown in Table VI. For each algorithm, there are 12 PI values (for four data sets and three distance measures), and hence, $12 \times 2 - 2 = 22$ degrees of freedom are available for each t-test. From the results of t-test and p values, the null hypothesis that "there is no difference between the averages of "percentages of improvement over the lowest biological score" for the two algorithms" is accepted for the pairs MN–Concorde and MN–Bar-Joseph. The alternative hypothesis that the average of "percentages of improvement over the lowest biological score" for "SOM + MN" is better than SOM is favored in t-test with the PI values shown in Table IV.

From the biological scores (Table III) and t-test results (Table VI), it is evident that MN provides biologically comparable gene order with respect to Concorde for all data sets and distance measure. Note that the time complexity of MN is $O(n^2)$, whereas the time complexity of Concorde is $O(2^n)$, where n is the number of genes. Therefore, it is preferable to use the MN algorithm since it has the minimum complexity. For example, MN took 0.008 s to order Yeast Complex data (979 genes) as compared to Concorde and Bar-Joseph's method that took 272 and 3.328 s, respectively.

TABLE VI
RESULTS OF t-TEST FOR DIFFERENT PAIRS OF ALGORITHMS

	Algorithm pairs		
	MN & Concorde	MN & B-Joseph	SOM+MN & SOM
t	0.051	0.067	4.103
p	$p > 0.5$	$p > 0.5$	$0.0001 > p$

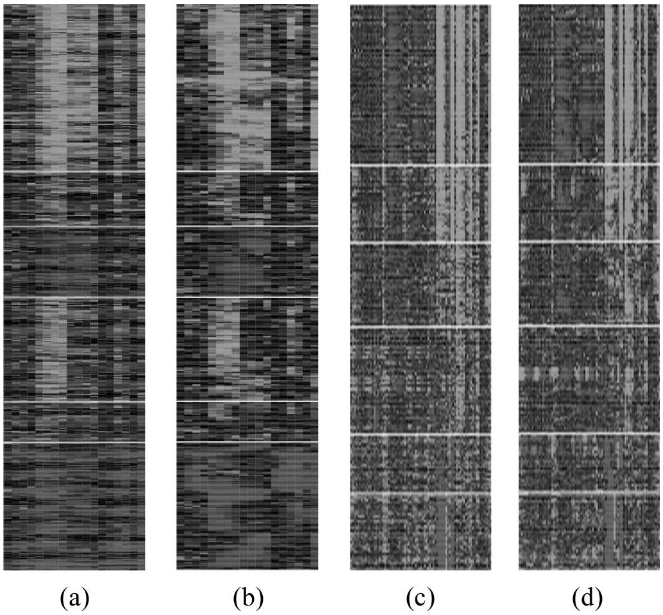

Fig. 2. Comparing SOM with "SOM + MN" for (a) and (b) Fibroblast data and (c) and (d) Yeast Complex data using *Maxrange-M* distance. The expression profiles are represented as lines of colored boxes using treeview software [1]. Some grouped genes obtained by MN [(b) and (d)] have similar expression patterns.

C. Subcluster Identification and Grouping of Correlated Genes by MN with SOM

To show how MN helps to identify subclusters within large clusters and groups functionally correlated genes within clusters to improve the solution quality of a nonhierarchical solution, MN is applied separately on the gene clusters found by SOM. The results/improvements found by combining these two algorithms are shown in Tables II, III, and VI. Here, the visual displays are presented for Fibroblast [Fig. 2(a) and (b)] and Yeast Complex [Fig. 2(c) and (d)] data. Fibroblast genes are first clustered using SOM with six nodes. A visual display of these six clusters is shown in Fig. 2(a). Observing this visual pattern, no subcluster can be identified in each cluster. After applying MN on each cluster, closely related genes with similar expressions are aligned next to each other, as shown in Fig. 2(b). Gene ordering here suggests that two or more subclusters exist at least in Clusters 1, 4, and 6, and it will be useful to increase the number of nodes of SOM to at least nine for Fibroblast data. Note that Iyer *et al.* [16] identified ten clusters of genes for these data.

The Yeast Complex data set is first clustered in 16 groups using SOM with 16 nodes. A visual display of the first six clusters/groups is shown in Fig. 2(c). When the genes are ordered in each cluster with

TABLE VII
GENE SUBCLUSTERS IN THIRD AND FOURTH CLUSTER AND THEIR FUNCTIONAL CATEGORY INDEXES FOR YEAST COMPLEX DATA. THESE SUBCLUSTERS ARE IDENTIFIED USING SOM + MN

Cluster	Subcluster	Genes	Functional index
3	1	YMR260C, YDR429C, YPL237W, YLR406C, YJR007W, YER025W, YPR041W, YDR172W, YDR211W	5
	2	YDR212W, YIL142W, YPL210C, YKL057C, YPL243W	6 and 7
	3	YLR060W, YOR260W, YDL040C, YKR026C, YLR291C, YBR142W, YBL087C, YHL001W, YDR450W, YHL033C, YBR191W, YBR189W, YBR048W, YBR118W	5
	4	YBR142W, YHR062C, YHR065C, YNR003C, YMR043W, YIL021W, YOR210W, YDR194C, YHR069C	4
4	1	YLR093C, YNL121C, YLR170C, YML112W, YBR160W, YBR171W, YLR378C, YML019W, YPL234C, YOR039W	6
	2	YKR068C, YLL050C, YGL200C, YML012W, YPL218W, YKL080W, YDR086C, YNL153C, YKL122C, YLR292C, YGL112C, YLR268W YLR447C	6 and 9
	3	YBR010W, YNL031C, YBL003C, YDR225W, YDR224C, YNL030W, YBR009C, YBL002W, YPL256C,	3, 4, and 7
	4	YJL025W, YPR101W, YMR061W, YGR195W, YOR244W, YLR105C, YDL043C, YPR056W, YPR057W	4

MN, four, four, five, and two distinct subclusters are identified using visual display in clusters 2, 3, 4, and 5, respectively. Gene names along with their functional category (indexes) for each subcluster within the third and fourth cluster are shown in Table VII. The name of the functional categories corresponding to their index is shown in Table VIII. For example, all the nine genes in the third subcluster of cluster 4 (YBR010W, YNL031C, YBL003C, YDR225W, YDR224C, YNL030W, YBR009C, YBL002W, and YPL256C) are involved in Cell Cycle and DNA processing, Transcription, and Protein with Binding Function or Cofactor Requirement. While using SOM, these genes are distributed in the cluster 4 and no subcluster can be identified. After ordering with MN, they are tightly grouped and identified easily using visual display.

VI. CONCLUSION

A new measure called *Maxrange*, for evaluating the distance between genes, and a new MN gene-ordering algorithm are described in this correspondence. These are used for efficiently ordering the genes in terms of their expression values for complete microarray data sets as well as in individual clusters found by SOM for those data sets. In *Maxrange-M* and *Maxrange-E* distance, normalization is performed separately with different normalizing factors for different types of experiment. This makes it suitable for both single type and multiple types of experiments. As a basic distance measure, Manhattan/Euclidean distance is used in *Maxrange* for their insensitiveness to large threefold changes in the gene expression profiles.

TABLE VIII
FUNCTIONAL INDEXES AND CORRESPONDING FUNCTIONAL CATEGORIES

Functional index	Functional Category
1	Metabolism
2	Energy
3	Cell Cycle and DNA Processing
4	Transcription
5	Protein Synthesis
6	Protein Fate (folding, modification, destination)
7	Protein with Binding Function or Cofactor Requirement
8	Protein Activity Regulation
9	Cellular Transport, Transport Facilitation and Transport Routes

In MN, the repetitive searching for optimal gene order in gene groups (closely related genes) is avoided. While this results in reduced time complexity ($O(n^2)$) for MN, in terms of biological score, it is comparable with Concorde ($O(2^n)$), the best TSP solver currently available. Also, it will be computationally expensive to apply Concorde or similar local search-based evolutionary algorithms to order genes in individual clusters of a nonhierarchical clustering solution. A novel hybrid method of gene ordering in SOM and its utility in finding useful subgroups of genes within clusters is also demonstrated. Experiments for each data set are also conducted with $\sqrt{n}$ nodes for SOM. In all these cases, the cluster number increased marginally, many nodes are found with no genes associated with them, and some clusters are found where genes belong to different biological categories and cannot be identified without gene ordering.

A huge number of different types of experiment by different research groups all over the world are conducted over genes to find the functional correlation between them. In the future, more experiments are likely to be appended in the same existing microarray. This demands a distance measure like *Maxrange-M*, and a growing number of genes for the same microarray data sets require fast ordering algorithm like MN. It is evident from the experimental results that *Maxrange-M* with MN performs the best in such situations. As such, this combination seems to be a promising tool for microarray- and gene-expression-related experiments.

ACKNOWLEDGMENT

The authors would like to thank B. L. Narayan for his help in statistical analysis, anonymous reviewers for their suggestions in improving the quality of research, and Dr. P. Kellam for providing the linear dynamic range of Herpes data.

REFERENCES

[1] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” in *Proc. Nat. Acad. Sci.*, Dec. 1998, vol. 95, no. 25, pp. 14 863–14 868.

[2] T. Biedl, B. Brejová, E. D. Demaine, A. M. Hamel, and T. Vinar, “Optimal arrangement of leaves in the tree representing hierarchical clustering of gene expression data,” Dept. Comput. Sci., Univ. Waterloo, Waterloo, ON, Canada, Tech. Rep. 2001-2014, 2001.

[3] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, “Fast optimal leaf ordering for hierarchical clustering,” *Bioinformatics*, vol. 17, no. 90 001, pp. 22–29, 2001.

[4] H. K. Tsai, J. M. Yang, Y. F. Tsai, and C. Y. Kao, “An evolutionary approach for gene expression patterns,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 2, pp. 69–78, Jun. 2004.

[5] C. Cotta, A. Mendes, V. Garcia, P. Franca, and P. Moscato, “Applying memetic algorithms to the analysis of microarray data,” in *Proc. Evo Workshops*, 2003, pp. 22–32.

[6] D. Applegate, R. Bixby, V. Chvátal, and W. Cook, (2003), *Concorde Package*. [Online]. Available: www.tsp.gatech.edu/concorde/downloads/codes/src/co031219.tgz

[7] R. Herwig, A. J. Poustka, C. Muller, C. Bull, H. Lehrach, and J. O’Brien, “Large-scale clustering of cDNA-fingerprinting data,” *Genome Res.*, vol. 9, no. 11, pp. 1093–1105, Nov. 1999.

[8] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, “Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation,” *Proc. Nat. Acad. Sci.*, vol. 96, no. 6, pp. 2907–2912, Mar. 1999.

[9] R. Sharan and R. Shamir, “CLICK: A clustering algorithm with applications to gene expression analysis,” in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2000, pp. 307–316.

[10] TSPLIB. [Online]. Available: http://www.iwr.uniheidelberg.de/groups/comopt/software/TSPLIB95/

[11] J. S. de Sousa, L. de C. T. Gomes, G. B. Bezerra, L. N. de Castro, and F. J. V. Zuben, “An immune-evolutionary algorithm for multiple rearrangements of gene expression data,” *Genet. Program. Evol. Mach.*, vol. 5, no. 2, pp. 157–179, 2004.

[12] L. Shi *et al.*, “Microarray scanner calibration curves: Characteristics and implications,” *BMC Bioinformatics*, vol. 6, no. (Suppl2):S11, pp. 1–14, 2005.

[13] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed, “Normalization of cdna microarray data,” in *Proc. SPIE—Microarrays: Optical Technologies and Informatics,* M. L. Bittner, Y. Chen, A. N. Dorsel, E. R. Dougherty, Eds., 2001, vol. 4266, pp. 141–152.

[14] G. Sherlock *et al.*, “The Stanford microarray database,” *Nucleic Acids Res.*, vol. 29, no. 1, pp. 152–155, 2001.

[15] Website. [Online]. Available: http://rana.lbl.gov/EisenData.htm

[16] V. R. Iyer *et al.*, “The transcriptional program in the response of human fibroblasts to serum,” *Science*, vol. 283, no. 5398, pp. 83–87, 1999.

[17] R. G. Jenner, M. M. Albà, C. Boshoff, and P. Kellam, “Kaposi’s sarcoma-associated herpesvirus latent and lytic gene expression as revealed by dna arrays,” *J. Virol.*, vol. 75, no. 2, pp. 891–902, 2001.

[18] *Munich Information for Protein Sequences*. [Online]. Available: http://www.mips.com

[19] T. H. Bo, B. Dysvik, and I. Jonassen, “Lsimpute: Accurate estimation of missing values in microarray data with least squares methods,” *Nucleic Acids Res.*, vol. 32, no. 3, 2004. e34, pp. online.

[20] E. F. Krause, *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*. New York: Dover, 1986.

[21] D. Venet, “MatArray: A Matlab toolbox for microarray data,” *Bioinformatics*, vol. 19, no. 5, pp. 659–660, 2003.