

New Distance Measure for Microarray Gene Expressions using Linear Dynamic Range of Photo Multiplier Tube

Shubhra Sankar Ray
Center for
Soft Computing Research
Indian Statistical Institute
Kolkata, India
shubhra_r@isical.ac.in

Sanghamitra Bandyopadhyay
Machine Intelligence Unit
Indian Statistical Institute
Kolkata, India
sanghami@isical.ac.in

Sankar K. Pal
Center for
Soft Computing Research
Indian Statistical Institute
Kolkata, India
sankar@isical.ac.in

Abstract

This paper deals with a new distance measure for genes using their microarray expressions. The distance measure is called, "Maxrange distance", where an experiment specific normalization factor is incorporated in the computation of the distance. The normalization factor is dependent on the linear dynamic range of the photo multiplier tube (PMT) for scanning fluorescence intensities of the gene expression values. Superiority of this distance measure in the microarray gene ordering problem has been extensively established on widely studied microarray data sets by performing statistical tests.

1 Introduction

The recent advances in DNA array technologies have resulted in a significant increase in the amount of genomic data [3, 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.

The present article deals with the tasks of measuring the distance between genes and evaluating their biological ordering in clustering framework. The widely used measures for finding the global similarity (where all the gene expression values present in the gene are taken into consideration) between genes are the Pearson correlation [3, 2] and the Euclidean distance [8]. In computing the similarity, all the above mentioned 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 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 article, where gene expression (for a particular type of experiment) distance between two genes are first normalized with a factor dependent on the linear dynamic range of photo multiplier tube (used for scanning fluorescence intensities of that experiment), and then summed to find a global distance.

Superiority of the proposed *Maxrange* distance measure over the related measures is established by using them on four different algorithms.

2 Gene Ordering Methods

Cluster analysis, ordering, and display of gene expression patterns are considered to be useful tools to detect genes that are co-expressed or implicated in similar cellular functions [3, 2]. Hierarchical clustering approaches (single, complete and average linkage) [3, 1] group gene expressions into trees of clusters. They start with singleton sets and merge all genes until all nodes belong to only one set. Hierarchical clustering does not determine unique clusters. Thus the user has to determine which of the subtrees are clusters and which subtrees are only a part of a bigger cluster. So in the framework of hierarchical clustering a gene ordering algorithm helps the user to identify clusters, and subclusters in big clusters, by means of visual inspection of the clustered gene expression data [1]. Moreover, genes that are adjacent in a linear ordering are often functionally co-regulated and involved in the same cellular process [2, 3] and biological analysis is often done in the context of this linear ordering [1].

Ideally, one would like to obtain a linear order of all

genes that puts similar genes close to each other; such that for any two consecutive genes the distance between them is small. 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 $1, 2, \dots, n$. This can be formulated as [2]

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

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 distance/similarity matrix.

Though hierarchical clustering provides good gene order [3] by grouping co-regulated genes, there is still much room in improving gene order. A hybrid method (first clustering then ordering) for ordering genes for a hierarchical clustering solution is proposed in [1] where dynamic programming is applied to flip internal nodes to reorder the leaves in a hierarchical solution.

3 Materials and Methods

3.1 Preliminaries of Microarray Technology

In general, microarray data can be represented by a real valued matrix; each row represents a gene and each column (or a set of columns) represents a condition, or experiment. In cDNA (clone DNA) microarray-based investigations, RNA from experimental samples (taken at selected times during the process) is labeled during reverse transcription with the red-fluorescent dye Cy5 and is mixed with a reference sample labeled in parallel with the green-fluorescent dye Cy3 [3]. After hybridization and appropriate washing steps, separate images/spots are acquired for each fluor, and fluorescence intensity ratios are obtained for all target elements. If R (red) and G (green) are the spot-specific, quantitated, fluorescent intensities of the target and reference expression signals respectively, relative gene expression is defined as the log ratio $M = \log_2 \frac{R}{G}$. For microarray data table each cell represents the M value at the corresponding target element [3] obtained from the gene under that experimental condition.

Fluorescence is currently the predominant method for microarray signal detection [5]. A critical component of a fluorescence scanner is the photomultiplier tube (PMT), in which fluorescent photons produce electrons that are amplified by the PMT voltage, also referred to as the PMT gain. For many microarray scanners, the PMT gain is an easily adjustable parameter, and the calibration curve (i.e., the curve showing the relationship between dye concentration and fluorescence intensity) depends on the gain setting

[5]. This PMT gain is also varied for different types of experiments of different biological origin. DNA microarray measurements normally assume a linear relationship between detected fluorescent signal and the concentration of the fluorescent dye. Each PMT has its own linear dynamic range within which signal intensity increases linearly with the increase of fluorescent dye concentration [5]. This linear dynamic range also fixes the dynamic range of the recorded microarray data (log ratio values) 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, PMT gain at 960 volts fixes the intensity range from x_1 to x_2 , and for dye Cy3, PMT gain at 760 volts fixes the intensity range from y_1 to y_2 . So the linear dynamic range of PMT fixes the linear dynamic range of the data from $\log_2 \frac{x_1}{y_1}$ to $\log_2 \frac{x_2}{y_2}$. Note that, this dynamic range is available either from the supplementary information (website) of the article/data (Yeast datas), or upon request to the authors (Herpes data) and not from the datasets, and hence is not sensitive to outliers. The proposed dynamic range based normalization belongs to the category of between-slide or multiple-slide normalization with two other members median absolute deviation (MAD) and variance regularization. The MAD and variance regularization are dynamic range estimators (not the real one) and are also implemented for the purpose of comparison. However, the results obtained were similar to without any normalization.

3.2 Description of Data Sets

For gene ordering, data sets like Cell Cycle [4], Yeast Complex [3, 1], All Yeast [3], and Herpes [7] are chosen. Table 1 shows the name of the data sets, number of genes in each dataset, number of gene categories, name of experiment types and number of experiments performed under each type, and finally the total number of experiments performed for a particular dataset. The dynamic range of expression values of each experiment is shown within parenthesis. 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, and -13.0 to 13.0 for Herpes data. The first three data sets of *Saccharomyces cerevisiae* consists of about 652, 979 and 6221 genes, and 184, 79 and 80 microarray experiments respectively. The genes in the three data sets are classified according to MIPS [6] categorization into 16, 16, and 18 groups respectively. Herpes virus genes are broadly assigned to five functional groups and available in [7].

Table 1. 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
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

3.3 New Distance Measure

A natural basis for organizing gene expression data is to group together genes with similar patterns of expression. The first step to this end is to adopt a mathematical description of distance. A number of measures of distance in the behavior of two genes can be used, such as the Manhattan distance [8], Euclidean distance [8], Pearson Correlation distance [2]. These 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 and performs global normalization to all the expression values for a particular gene; thus losing 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 dataset. 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 the 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 dataset.

Such an attempt is made in this article where two new distance measures are developed using Manhattan distance and Euclidean distance respectively (to avoid over sensitivity to three fold changes), in which the 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 photo multiplier tube, for a particular type of experiment.

Let

$$X = x_1^{e_1}, \dots, x_{i_1}^{e_1}, x_1^{e_2}, \dots, x_{i_2}^{e_2}, \dots, x_1^{e_m}, \dots, x_{i_m}^{e_m} \text{ and}$$

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

be the expression 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 $i_1 + i_2 + \dots + i_m$ experiments in total. Using Manhattan distance the *Maxrange* distance between X and Y is defined as

$$Maxrange-M_{X,Y} = \frac{1}{m} \sum_{r=1}^m \frac{1}{i_r} \times \frac{\sum_{j=1}^{i_r} |x_j^{e_r} - y_j^{e_r}|}{Max_{e_r} - Min_{e_r}} \quad (2)$$

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 photo multiplier tube (or radioactive probe) for an experiment of type e_r .

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

$$Maxrange-E_{X,Y} = \frac{1}{m} \sum_{r=1}^m \frac{1}{i_r} \times \frac{\sqrt{\sum_{j=1}^{i_r} (x_j^{e_r} - y_j^{e_r})^2}}{Max_{e_r} - Min_{e_r}} \quad (3)$$

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

4 Biological Interpretation

A biological score, that 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 category, while there are many unclassified genes also (no category). A vector $V(g) = (v_1, v_2, \dots, v_j)$ 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 j th category; otherwise is zero. Based on the information about categorization, the score of a gene order for multiple class genes is defined as [9]

$$S(n) = \sum_{i=1}^{N-1} G(g_i, g_{i+1}), \quad (4)$$

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

$$G(g_i, g_{i+1}) = \sum_{k=1}^j V(g_i)_k V(g_{i+1})_k, \quad (5)$$

where $V(g_i)_k$ represents the k^{th} entry of vector $V(g_i)$. Note that, $S(n)$ can also be used as scoring function for single class genes like Herpes 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. So higher values of $S(n)$ are better and can be used to evaluate the goodness of a particular gene order. Note that, although these scoring functions provide a good quantitative index for gene ordering, using $S(n)$ as the similarity measure in ordering is not practical, since the information about gene categories is unknown for most of the genes in the real world.

5 Experimental Results

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 Bar-Joseph et al.'s [1] method are downloaded from [10]. Performance of the proposed *Maxrange-M* and *Maxrange-E* distance are compared with Pearson correlation, Euclidean distance, and Manhattan distance.

5.1 Comparative Performance of Distance Measures

Table 2 compares the performance of our proposed measure with those of the other measures in terms of the $S1$ value (Eq. 4). Three distance measures are considered, namely, *Maxrange-M*, Pearson and Euclidean. The biological scores corresponding to Manhattan Distance are found to be comparable to those for Pearson Correlation, and hence omitted here. The percentages of improvement over

the lowest biological score (in terms of $S1$ value) in a particular data set are shown within parenthesis, and defined as:

$$PI_{i,j} = \frac{d_{i,j} - \min_i(d_{i,j})}{\min_i(d_{i,j})} \times 100 \quad (6)$$

where, $d_{i,j}$ indicates the biological score ($S1$ value) in i th row and j th column of the result matrix in Table 2, and $\min_i(d_{i,j})$ indicates the minimum biological score in column j for all i . These PI values in Table 2 are used in the next section for conducting t-tests.

Though in most of the cases *Maxrange-E* distance is found to be superior to Euclidean distance and inferior to *Maxrange-M*, for All Yeast data, it performs better ($S(n)=2441$) than *Maxrange-M* ($S(n)=2341$) for average linkage algorithm. When the microarray data sets contain experiments with data value of same dynamic range, like Herpes, then *Maxrange-M* provides identical results with Manhattan distance for all widely used ordering algorithms. 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 2).

5.2 Statistical Analysis of *Maxrange-M* Distance Measure

To statistically compare the performance of *Maxrange-M* distance with Pearson Correlation in case of ordering algorithms, t-tests are performed with the PI (Eq. 6) values shown within parenthesis in Table 2, using the equation

$$t = \frac{\overline{PI}_1 - \overline{PI}_2}{\sqrt{\frac{\text{Variance}PI_1}{n_1} + \frac{\text{Variance}PI_2}{n_2}}}. \quad (7)$$

where, \overline{PI}_1 and $\text{Variance}PI_1$ are the mean and the variance of all the available PI values for *Maxrange-M* distance in Table 2. PI_2 is used for Pearson Correlation and $n_1 = n_2 = 16$, as there are 16 PI values available in total from Table 2 for each of the distance measures with 4 datasets and 4 algorithm. 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 3. 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. The final conclusion, once the test has been carried out, is always given in terms of the null hypothesis (H_0), that there is no difference between the averages of 'percentages of improvement over the lowest biological score' for the two distance measures. After finding the p values (from t-table)

Table 2. Biological Score ($S(n)$) and Percentage of Improvement (PI) value (within parenthesis) for different distance measures and algorithms

Distance	Algorithm	Data Sets			
		Cell cycle	Yeast complexes	All Yeast	Herpes
<i>Maxrange-M</i>	Bar-Joseph	423 (17.83)	1074 (26.50)	2371 (22.85)	43 (19.44)
	Average Linkage	415 (15.60)	1040 (22.50)	2341 (21.30)	39 (8.33)
	Complete Linkage	407 (13.37)	1043 (22.85)	2305 (19.43)	38 (5.56)
	Single Linkage	382 (6.41)	903 (6.36)	1970 (2.07)	41 (13.89)
Pearson	Bar-Joseph	381 (6.13)	1024 (20.61)	2350 (21.76)	38 (5.56)
	Average Linkage	385 (7.24)	987 (16.25)	2292 (18.76)	38 (5.56)
	Complete Linkage	393 (9.47)	955 (12.49)	2301 (19.22)	36 (0.00)
	Single Linkage	359 (0.00)	902 (6.24)	1973 (2.23)	39 (8.33)
Euclidean	Bar-Joseph	421 (17.27)	1013 (19.32)	2346 (21.55)	40 (11.11)
	Average Linkage	403 (12.26)	1011 (19.08)	2431 (25.96)	39 (8.33)
	Complete Linkage	403 (12.26)	999 (17.67)	2269 (17.56)	37 (2.78)
	Single Linkage	361 (0.56)	849 (0.00)	1930 (0.00)	36 (0.00)

Table 3. Results of t-test for different pairs of distance measures

	Pairs of distance measure	
	<i>Maxrange-M</i> & Pearson	<i>Maxrange-M</i> & Euclidean
t	2.0134	1.2709
p	0.027 > p	0.107 > p

for corresponding t values, we reject the null hypothesis for both the cases with significance level 0.027 and 0.107 respectively, which suggests that there is strong evidence against the null hypothesis in favor of the alternative.

6 Conclusion

A new measure called *Maxrange*, for evaluating the distance between genes, is used for efficiently ordering the genes in terms of their expression values for microarray datasets. The available measures for gene distance, like Manhattan Distance, Euclidean distance, and Pearson correlation, use only one normalization factor (1, 1, and standard deviation respectively) for all types of experiments, although the expression values may differ by orders of magnitude from one kind of experiment to another. As a consequence, the distance between genes may not be properly reflected in these measures for microarray data having different types of experiments. In contrast, normalization is performed separately with different normalizing factors for the different types of experiment in our *Maxrange-M* and *Maxrange-E* distance. This makes it, suitable for both sin-

gle type and multiple type of experiments and, promising for microarray gene expression related experiments.

References

- [1] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17:22–29, 2001.
- [2] T. Biedl, B. Brejov, E. D. Demaine, A. M. Hamel, and T. Vinar. Optimal arrangement of leaves in the tree representing hierarchical clustering of gene expression data. Technical Report 2001-2014, Dept. Computer Sci., Univ. Waterloo, 2001.
- [3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. National Academy of Sciences*, 95:14863–14867, 1998.
- [4] G. S. et al. The stanford microarray database. *Nucleic Acids Research*, 29(1):152–155, 2001.
- [5] L. S. et al. Microarray scanner calibration curves: characteristics and implications. *BMC Bioinformatics*, 6((Suppl2):S11):1–14, 2005.
- [6] M. I. for Protein Sequences. <http://www.mips.com>.
- [7] R. G. Jenner, M. M. Alb, C. Boshoff, and P. Kellam. Kaposi's sarcoma-associated herpesvirus latent and lytic gene expression as revealed by dna arrays. *Journal of Virology*, 75(2):891–902, 2001.
- [8] E. F. Krause. *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*. 1986.
- [9] H. K. Tsai, J. M. Yang, Y. F. Tsai, and C. Y. Kao. An Evolutionary Approach for Gene Expression Patterns. *IEEE Trans. on Info. Tech. in Biomedicine*, 8(2):69–78, 2004.
- [10] D. Venet. MatArray: a Matlab toolbox for microarray data. *Bioinformatics*, 19(5):659–660, 2003.