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

Shubhra Sankar Ray[1], Sanghamitra Bandyopadhyay[2], and Sankar K. Pal[1]

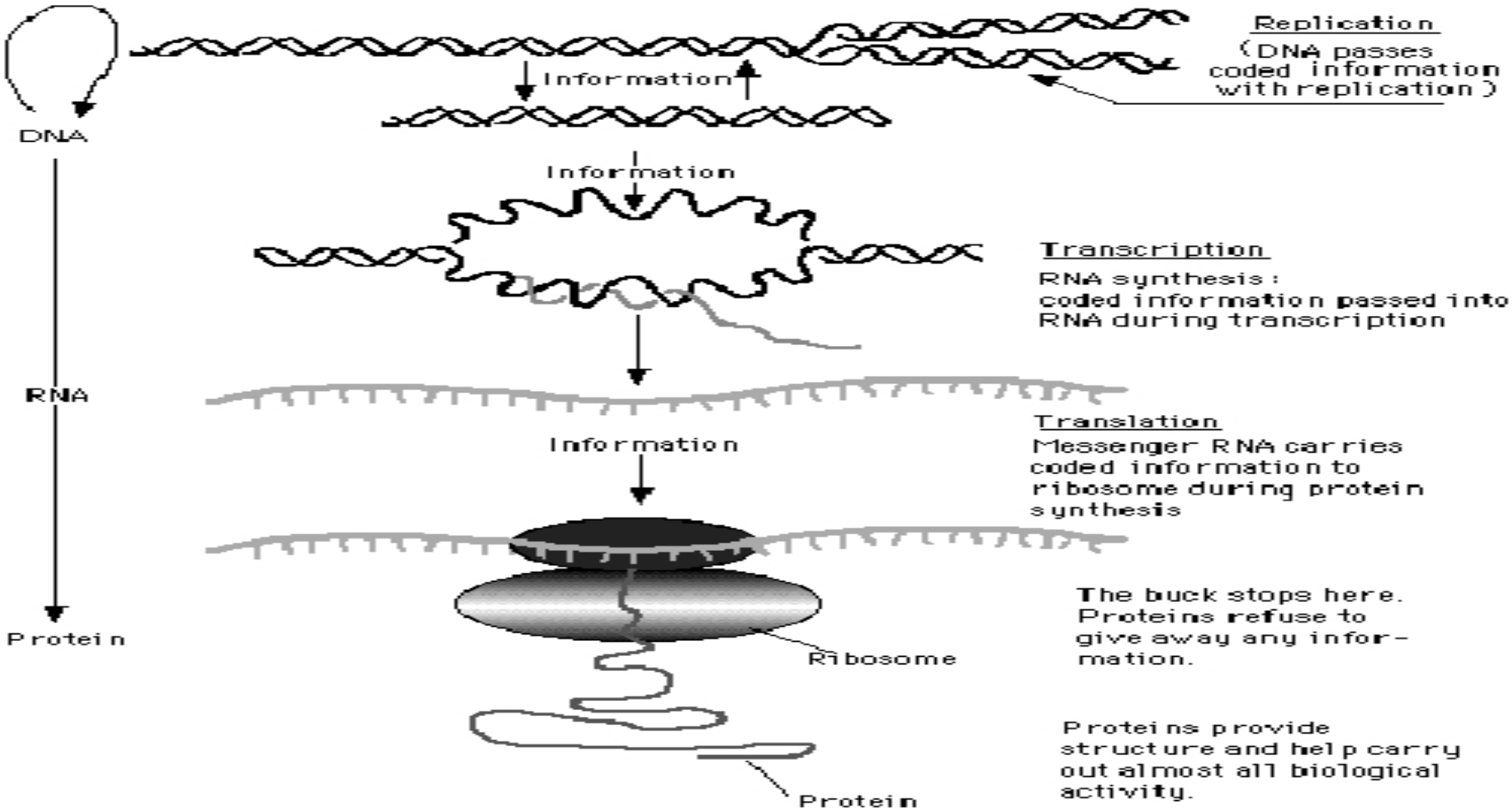
[1]Center for Soft Computing Research, [2]Machine Intelligence Unit

Indian Statistical Institute, Kolkata, India

Content:

- How gene is expressed ?
- Basics of DNA Microarray
- Tasks involving Microarray: Gene Ordering
- Problem Definition
- Proposed Dynamic Range Based Distance Measure
- Application of Dynamic Range Based Distance Measure and Gene Ordering in Partitive Clustering
- Results
- Summary
- References

How Gene is Expressed ?



- Transcription factors control gene expression by binding the gene's promoter and either activating (switching on) the gene's transcription, or repressing it (switching it off).

DNA MICROARRAY

- Typically a glass slide, onto which cDNAs are attached and colored with the green-fluorescent dye Cy3 .
- By performing biological experiments RNA from experimental samples are colored during reverse transcription with the red-fluorescent dye Cy5 and mixed with a reference sample labeled in parallel with green-fluorescent dye Cy3.
- After hybridization and appropriate washing steps, separate images were acquired for each fluor.
- Cy5/Cy3 fluorescence ratio (gene expression) are obtained by measuring the spot intensities with fluorescence scanner.

Tasks involving Microarray Gene Expression

- Grouping genes with functional similarity
  - Clustering of Genes
  - Ordering of Genes

Why gene ordering?

•Genes that are adjacent in an ordering are often functionally co-regulated and involved in the same cellular process

•Gene ordering helps to identify subclusters in clusters by means of visual inspection of the clustered gene expression data

•Every gene is influenced by a set of eight to ten other genes

- it can be achieved by gene ordering

Definition of Gene Ordering

An optimal gene order, a minimum sum of distances between pairs of adjacent genes in an ordering *{1, 2, ... n}*, can be formulated as

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

where n is the number of genes and $C_{i,\ i+1}$ is the gene expression distance between two genes *i* and *i+1*.

- Smaller the distance in terms of gene expression values greater the functional similarity between genes.
- ***F(n)*** can be used as the fitness function.

Biological Evaluation of Gene Ordering

A biological score, that is different from the fitness function, is used to evaluate the final gene ordering. The biological score is defined as

$$S(n)=\sum_{i=1}^{n-1} C_{i,\,i+1}$$ where $C_{i,\,i+1}$ = 1, if gene i and $i+1$ are in the same group
= 0, if gene i and $i+1$ are not in the same group

- Ordering would have a higher score when more genes within the same group are aligned next to each other.

Genes	YML120C	YJR048W	YMR002W	YDR432W
groups	Metabolism	Transcription	Transcription	Metabolism
Score		0+1+0=1		

Genes	YJR048W	YMR002W	YDR432W	YML120C
score		1+0+1=2.		

Problems in Gene Expression related Applications

1) Dynamic Range of gene expressions differ by orders of magnitude from one kind of experiment to another
2) Experiments with higher dynamic range dominate the lower ones.
3) Existing normalization methods adjust median (Ex: Median absolute deviation).
4) Normalization with dynamic range obtained from data is sensitive to outliers.

Solving the Dynamic Range Problem

- Normalization is performed with the linear dynamic range of Photo Multiplier Tube (PMT), a component of a Fluorescence scanner.
- **Linear dynamic range is obtained from the characteristics of PMT.**
- Dynamic range is not obtained from the datasets and hence is **not sensitive to outliers**.

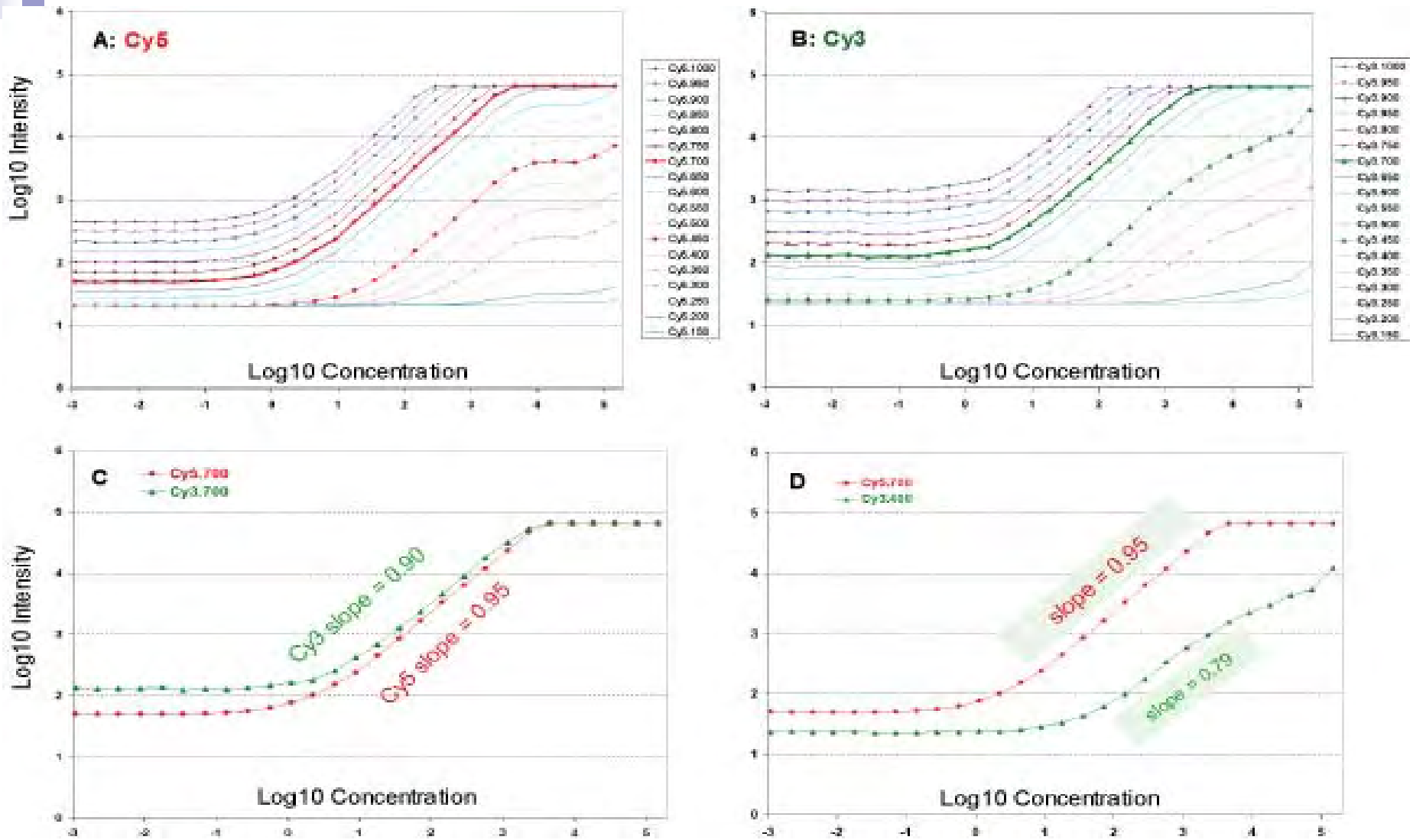
Properties of Photo Multiplier Tube (PMT)

- PMT has its own linear dynamic range within which signal intensity increases linearly with the increase of fluorescent dye (Cy5 and Cy3) concentration.

- Linear dynamic range is varied for experiments of different biological origin, by varying the PMT voltge.

  - For example, in Cell Cycle related experiments, for dye Cy5, PMT gain at 960 volts fixes the intensity range from x1 to x2, and for dye Cy3, PMT gain at 760 volts fixes the intensity range from y1 to y2. So the linear dynamic range of PMT fixes the linear dynamic range of the data from

$$\log_2 \frac{x1}{y1} \quad \text{to} \quad \log_2 \frac{x2}{y2}$$



Calibration curves under different PMT gains. X-axis: log10 concentration, Y-axis: log10 fluorescence intensity. A: Cy5 dye; B: Cy3 dye. Representative calibration curves are presented in C (Cy5 and Cy3 channels are scanned under the same PMT gain of 700 V) and D (the Cy5 and Cy3 channels are scanned at 700 V and 400 V, respectively). BMC Genomics. 2004; 5: 10.
Published online 2004 February 3. doi: 10.1186/1471-2164-5-10.

Microarray Gene Expression values

gene	Cell Cycle 1	Cell Cycle 2	Sporulation 1	Sporulation 2	Shock 1	Shock 2	Diauxic Shift 1
YDR029W	0.05	-0.21	0.16	0.19	0	2.19	-2.0
YBL052C	-0.38	-0.71	0.14	0.33	0.01	0.09	1.7
YOR337W	-0.42	1.78	0	0.62	0.64	0.45	2.92
YMR183C	-0.86	-0.67	-3.23	0.09	-0.09	0.48	0.49
YKR021W	-0.85	1.31	-0.46	3.42	-2.38	0.19	0.64
YHR023W	-1.77	-0.86	-1.07	0.1	0.28	0.91	0.97
YHR029C	-0.58	-0.91	-0.62	-0.13	0.06	-0.08	2.03

Dynamic Range of PMT	-1.2 to 1.2	-3.0 to 3.0	-1.5 to 1.5	-2.0 to 2.0

Dynamic Range Based Distance Measure

Gene $X = x_1^{e_1}, x_2^{e_1}, \cdots, x_{i_1}^{e_1}, x_1^{e_2}, x_2^{e_2}, \cdots, x_{i_2}^{e_2}, \cdots\cdots, x_1^{e_m}, x_2^{e_m}, \cdots, x_{i_m}^{e_m}$

Gene $Y = y_1^{e_1}, y_2^{e_1}, \cdots, y_{i_1}^{e_1}, y_1^{e_2}, y_2^{e_2}, \cdots, y_{i_2}^{e_2}, \cdots\cdots, y_1^{e_m}, y_2^{e_m}, \cdots, y_{i_m}^{e_m}$

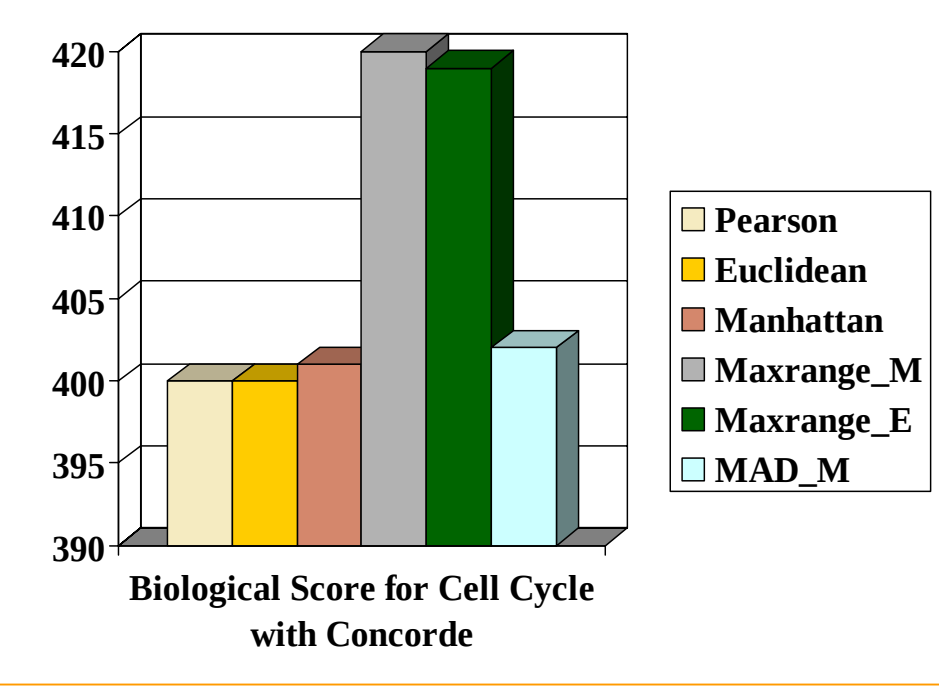
Series of *m* different types of experiment (e_1+e_2….e_m) consisting of i_1+i_2+…+i_m experiments in total.

Using Manhattan distance and Dynamic Range Normalization, the distance between X and Y is defined as

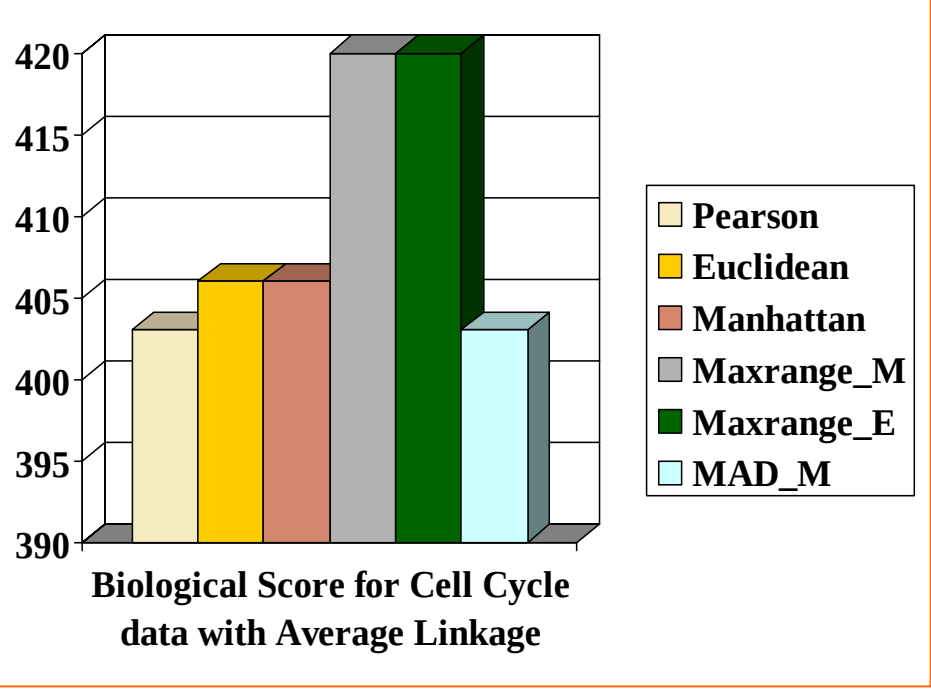
$$Maxrange\text{-}M_{X,Y} = \frac{1}{m}\sum_{r=1}^{m}\frac{1}{i_r} \times \frac{\sum_{j=1}^{i_r}|x_j^{e_r} - y_j^{e_r}|}{Max_{e_r} - Min_{e_r}}$$

Using Euclidean distance

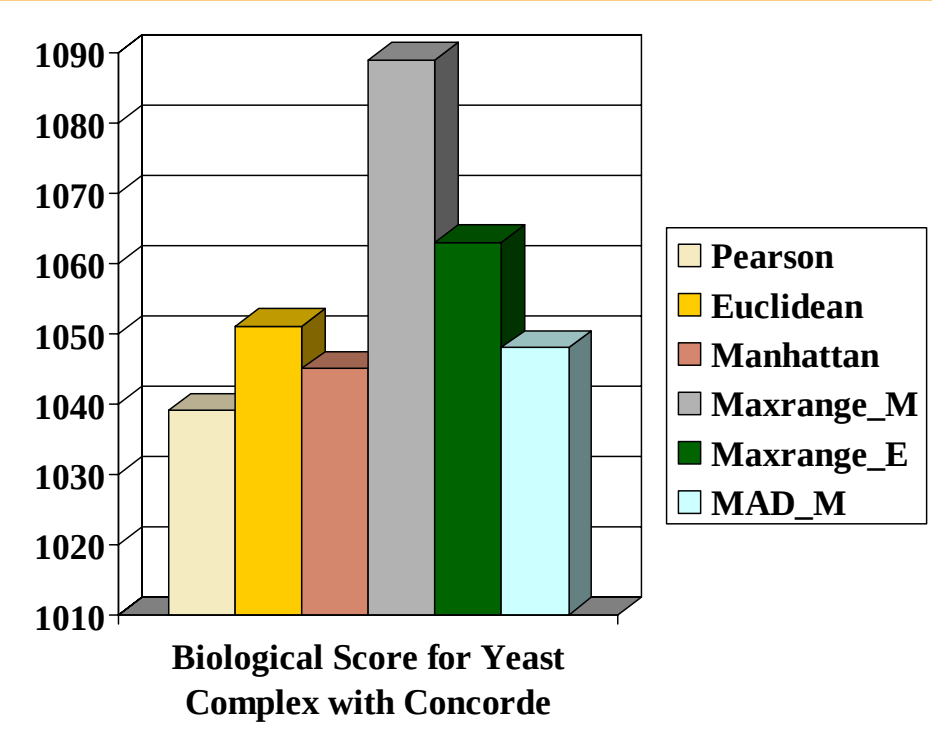
$$Maxrange\text{-}E_{X,Y} = \frac{1}{m}\sum_{r=1}^{m}\frac{1}{i_r} \times \frac{\sqrt{\sum_{j=1}^{i_r}(x_j^{e_r} - y_j^{e_r})^2}}{Max_{e_r} - Min_{e_r}}$$



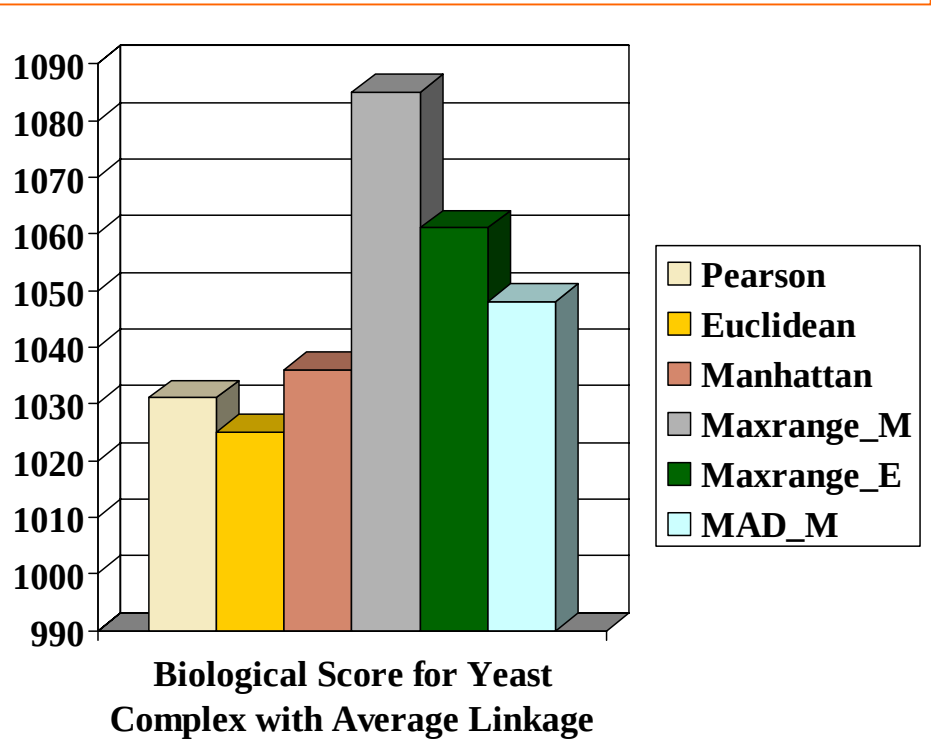
Biological Score for Cell Cycle with Concorde



Biological Score for Cell Cycle data with Average Linkage



Biological Score for Yeast Complex with Concorde



Biological Score for Yeast Complex with Average Linkage

Gene Ordering in Partitive Clustering

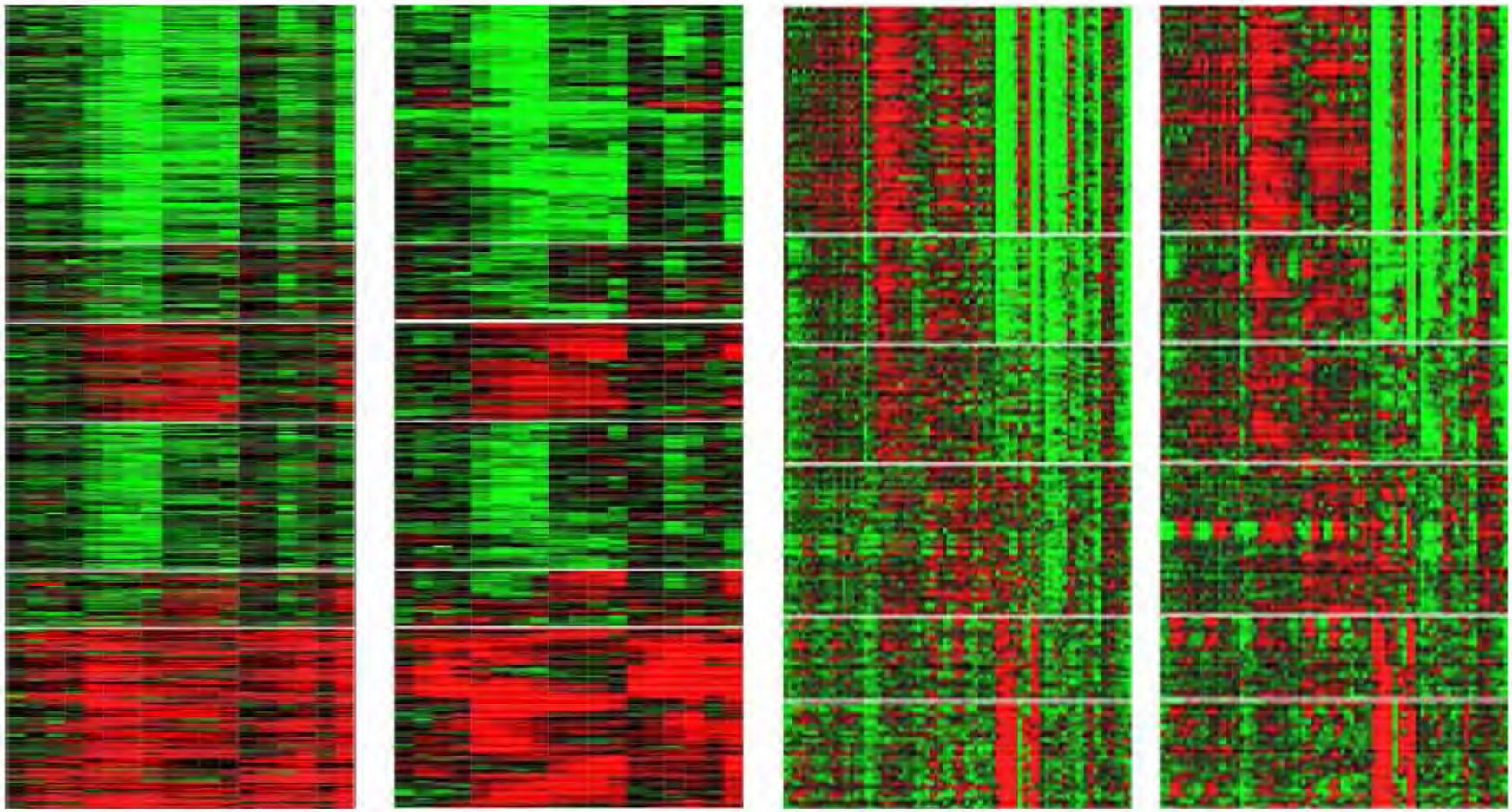


Figure 1: Comparing SOM with 'SOM+MN' for Fibroblast data (Fig. a and Fig. b respectively) and Yeast Complex data (Fig. c and Fig. d respectively). The expression profiles are represented as lines of coloured boxes using Expander [8], each of which corresponds to a single experiment.

Subclusters found by Minimal Neighbor Algorithm in Cluster 4

Cluster	Sub-cluster	Genes	Functional index
4	1	YLR093C, YNL121C, YLR170C, YML112W, YBR160W, YBR171W, YLR378C, YML019W, YPL234C, YOR039W	6
	2	YKR068C, YLL050C, YGL200C, YML012W, YPL218W, YKL080W, YLR447C, YDR086C, YNL153C, YKL122C, YLR292C, YGL112C, YLR268W	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
	5	YGL100W, YNL261W, YKL144C, YNL151C, YJL008C, YER148W	7

3 → Cell Cycle and DNA Processing
4 → Transcription
6 → Protein Fate (folding, modification, destination)
7 → Protein with Binding Function or Cofactor Requirement
9 → Cellular Transport, Transport Facilitation and Transport Routes

Summary

•A new normalization method using linear dynamic range of photo multiplier tube is described.

• Biological relevance of the dynamic range based normalization and gene ordering in partitive clustering are established.

Selected References

1. L. Shi et al., “Microarray scanner calibration curves: characteristics and Implications,” BMC Bioinformatics, vol. 6, no. Suppl2:S11, pp. 1-14, 2005.
2. S. Pickett, “Understanding and evaluating fluorescent microarray imaging instruments,” IVT Technology, vol. 9, no. 4, pp. 1-6, 2003.
3. http://www.psrg.lcs.mit.edu/clustering/ismb01/optimal.html
4. 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. Computer Sci., Univ. Waterloo, Tech. Rep. 2001-14, 2001.
5. Munich Information for Protein Sequences, http://www.mips.com
6. 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, vol. 95, pp. 14863-14867, 1998.
7. 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. National Academy of Sciences, pp. 2907-2912, 1999.
8. D. Applegate, R. Bixby,V. Chvátal, and William Cook, “Concorde Package. [Online],” www.tsp.gatech.edu/concorde/downloads/codes/src/ co031219.tgz, 2003.

Thank You

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
Maxrange-M	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)

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

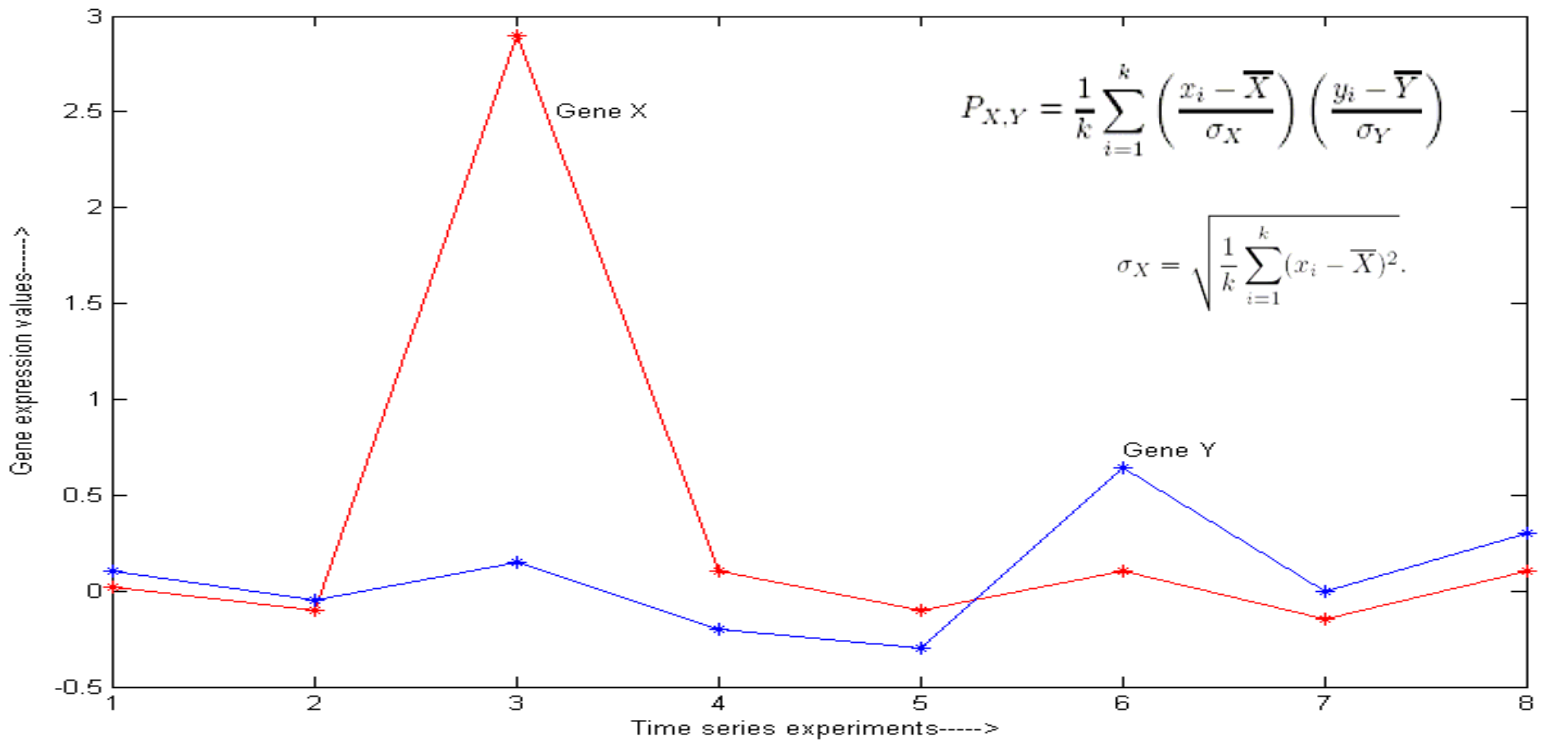
$$t = \frac{\overline{PI_1} - \overline{PI_2}}{\sqrt{\frac{VariancePI_1}{n_1} + \frac{VariancePI_2}{n_2}}}.$$

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

	Pairs of distance measure	
	Maxrange-M & Pearson	*Maxrange-M* & Euclidean
t	2.0134	1.2709
p	$0.027 > p$	$0.107 > p$

Limitation of Pearson Correlation

- Over sensitive to large three fold changes
  - One spike is sufficient for misleading results



- Sensitiveness to large three fold changes are minimized with Manhattan or Euclidean distance

Median Absolute Deviation

Let us assume that there are m different types of experiments, experiment type i (eg., sporulation) has a total of n_i (where, i = 1, 2, …. m) no. of M values (log ratios), and Set_i denotes the total set of M values for experiment i. The scaling factor S_MAD_i for experiment of type i, in terms of median absolute deviation (MAD), is defined as

$$S\_MAD_i = \frac{MAD_i}{\sqrt[m]{\prod_{i=1}^{m} MAD_i}},$$

where, MAD is defined as

$$MAD_i = median\{|Set_i - median(Set_i)|\}$$