Matrix Visualization

Han-Ming Wu, ShengLi Tzeng, and Chun-houh Chen

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.
hmwu@stat.sinica.edu.tw, h0h1@stat.sinica.edu.tw,
cchen@stat.sinica.edu.tw

1 Introduction

Graphical exploration for quantitative/qualitative data acts as the initial yet essential step in modern statistical data analysis. Matrix visualization (Chen (2002); Chen *et al.* (2004)) is a graphical technique that can simultaneously explore the associations of up to thousands of subjects, variables, and their interactions, without first reducing dimension. Matrix visualization permutes the rows and columns of the raw data matrix by suitable seriation (reordering) algorithms, together with the corresponding proximity matrices. The permuted raw data matrix and two proximity matrices are then displayed as matrix maps through suitable color spectra, and the subject-clusters, variable-groups, and interactions embedded in the data set can be visually extracted.

Since the introduction of Exploratory Data Analysis (EDA, Tukey (1977)), boxplots, along with the scatterplot aided by interactive functionalities, have served the statistical community as major graphical tools. These tools, together with various dimension reduction techniques, are useful for exploring data structure when the number of variables is of moderate size, and when structure is not too complex. Yet, with striking advances in computing, communication, and high-throughput biomedical instruments, the number of variables can easily reach tens of thousands, and the need for practical data analysis remains. Dimension reduction tools often lose effectiveness when it comes to visual exploration of information structure embedded in high dimensional data sets. On the other hand, matrix visualization, integrated with computing, memory, and display, has great potential for visually exploring structure that underlies massive and complex data sets.

We briefly review the literature of related work in the next section. The foundation of matrix visualization under the framework of generalized association plots (GAP, Chen (2002)), with some related issues, is discussed in Sections 3 followed, in Section 4 by some generalization. Section 5 gives an example of matrix visualization with 400 variables (arrays) and 2000 samples (genes). Comparisons of matrix visualization with other popular graphical

tools, for efficiency over size of dimension, are then given in Section 6. Section 7 illustrates matrix visualization for binary data, while Section 8 discusses generalizations and extensions. We conclude this chapter with some perspectives on matrix visualization in Section 9.

2 Related Works

The concept of matrix visualization was introduced in Bertin (1967) as a reorderable matrix for systematically presenting data structures and relationships. Carmichael and Sneath (1969) developed taxometric maps for classifying OUT's (operational taxonomy units) in numerical phenetics analysis. Hartigan (1972) introduced the direct clustering of a data matrix, later known as block clustering (Tibshirani (1999)). Lenstra (1974) and Slagle *et al.* (1975) related the traveling-salesman problem and shortest spanning path to the clustering of data arrays. The colour histogram of Wegman (1990) was the first color matrix visualization in the statistical literature. Minnotte and West (1998) extended the idea of colour histograms to the data image package that was later used for outlier detection (Marchette and Solka (2003)).

Some matrix visualization techniques were developed for exploring proximity matrices only: Ling (1973) looked for factors of variables by examining relationships through a shaded correlation matrix; Murdoch and Chow (1996) used elliptical glyphs to represent large correlation matrices; Friendly (2002) proposed corrgrams, similar to the reorderable matrix method, for analyzing multivariate structure among the variables in correlation and covariance matrices. Chen (1996, 1999, and 2002) integrated visualization for raw data matrix with two proximity matrices (for variables and samples) into the framework of generalized association plots (GAP). The Cluster and TreeView packages by Eisen *et al.* (1998) are probably the most popular matrix visualization packages because of the proliferation of gene expression profiling for microarray experiments.

Permutation (ordering) of columns and rows for a data matrix, and proximity matrices for variables and samples, is an essential step in matrix visualization. Several recent statistical works have touched on the issue of reordering of variables and samples: Chen (2002) proposed the concept of relativity of a statistical graph; Friendly and Kwan (2003) discussed the idea of effect ordering of data displays; Hurley (2004) used scatterplot matrices and parallel coordinate plots as examples to address the problem of placing interesting displays in prominent positions. Different terms (such as the reorderable matrix, the heatmap, color histogram, data image and matrix visualization) have been used in the literature for describing these related techniques. We use matrix visualization (MV) to refer to them all.

3 The Basic Principles of Matrix Visualization

We use the GAP (Chen (2002)) approach to illustrate the basic principles of matrix visualization for continuous data, using the 6400 genes and 851 microarray experiments collected in the published yeast expression data database for visualization and data mining (Marc *et al.* (2001)), and designated henceforth here as Data 0. Detailed descriptions of data pre-processing were given in the yeast Microarray Global Viewer (`http://transcriptome.ens.fr/ymgv/`). For illustration purposes, we selected 15 samples and 30 genes across these samples as Data 1, where rows correspond to genes and columns to microarray experiments (arrays). For various gene expression profile analyses, the roles played by rows and columns are often interchangeable. This interchangeability suits well into the GAP approach of matrix visualization where samples and variables are treated symmetrically and can be interchanged directly.

3.1 Presentation of Raw Data Matrix

The first step of matrix visualization for continuous data is the production of a raw data matrix $X_{30\times15}$, and two corresponding proximity matrices for rows, $R_{30\times30}$, and columns, $C_{15\times15}$, calculated with user-specified similarity (or dissimilarity) measures. The three matrices are then projected through suitable color spectra to construct corresponding matrix maps in which each matrix entry (raw data or proximity measurement) is represented by a color dot. The left panel of Figure 1 shows the raw data matrix of $\log_2$ transformed ratios of expressions coded by a bi-directional green-black-red spectrum for Data 1, with Pearson correlations for between arrays relations coded by a bi-directional blue-white-red spectrum, and Euclidean distances for between genes relations coded by a uni-directional rainbow spectrum.

In the raw data matrix map, a red (green) dot in the ij-th position of the map for $X_{30\times15}$ means the i-th gene at the j-th array is relatively up (down) regulated. A black dot stands for a relatively non-differentially expressed gene/array combination. A red (blue) point in the ij-th position of the $C_{15\times15}$ matrix map represents a positive (negative) correlation between arrays i and j. Darker (lighter) intensities of color stand for stronger absolute correlation coefficients while white dots represent no correlations. A blue (red) point in the ij-th position of the $R_{30\times30}$ matrix map represents a relative small (large) distance between genes i and j while a yellow dot represents a median distance.

Data Transformation

Transformations such as log, standardization (zero mean, unit variance), or normalization (normal score transformation) may have to be applied to raw data before the data map is constructed or proximity matrices calculated in order to have meaningful visual perception of the data structure, or comparable visual effects between displays. The transformation-visualization process

may have to be repeated several times before the embedded information can be fully explored.

Selection of Proximity Measures

Proximity matrices have two major functions: (1) to serve as the direct visual perception of the relationship among variables and between samples; (2) to serve as the media for reordering of variables and samples for better visualization of the three matrix maps. Selection of proximity measures in matrix visualization plays a more important role than it does in numerical or modelling analyses. Pearson correlation often serves as the between-variables proximity measure, Euclidean distance is commonly employed for samples (Figure 1). For potential nonlinear relationships, Spearman's rank correlation and Kendall's tau coefficient can replace Pearson correlation in assessing the between variable relationship while some nonlinear feature extraction methods such as the Isomap (Tenenbaum *et al.* (2000)) distance can be used to measure the nonlinear between-sample distances. More sophisticated kernel methods can also be applied when users see the necessity for them.

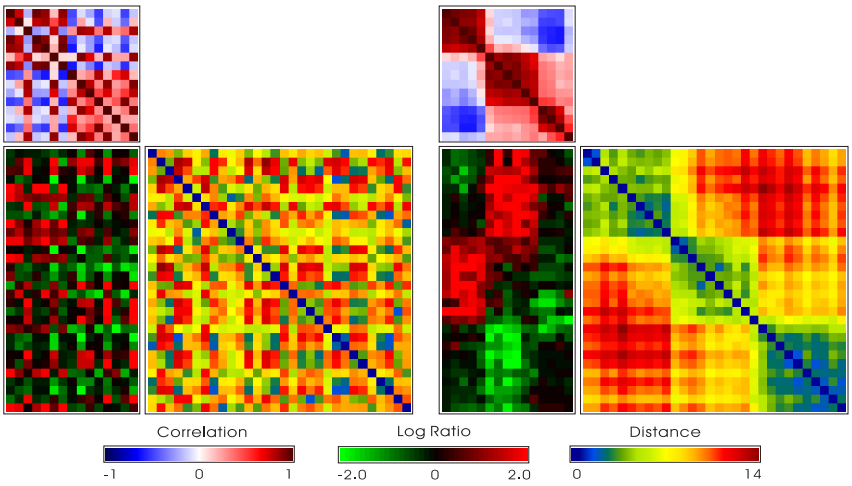

Fig. 1. Left: Unsorted data matrix (log ratio gene expression) map with two proximity matrixes (Pearson correlation for arrays and Euclidean distance for genes) maps for Data Set 1. Right: Elliptical seriations applied to the three matrix maps on the left panel.

Color Spectrum

The selection of an appropriate color spectrum can be critical and is user dependent in visualization and information extraction of data and proximity matrices. The selection of a suitable color spectrum should focus on the

capacity for expressing numerical nature individually and globally in the matrices. Our above mentioned choices for gene expression profiles might well give way to others in different circumstances. Thus, illustrated in Figure 2 is a correlation matrix map of fifty psychosis disorder variables (Chen (2002)) coded with four different bi-directional color spectra. While displays (a) and (b) appear more agreeable to human perception, displays (c) and (d) actually provide better resolution for distinguishing different levels of correlation intensities. The relative triplet color codes (red, green, blue) in the RGB cube for these four color spectra are shown in Figure 3.

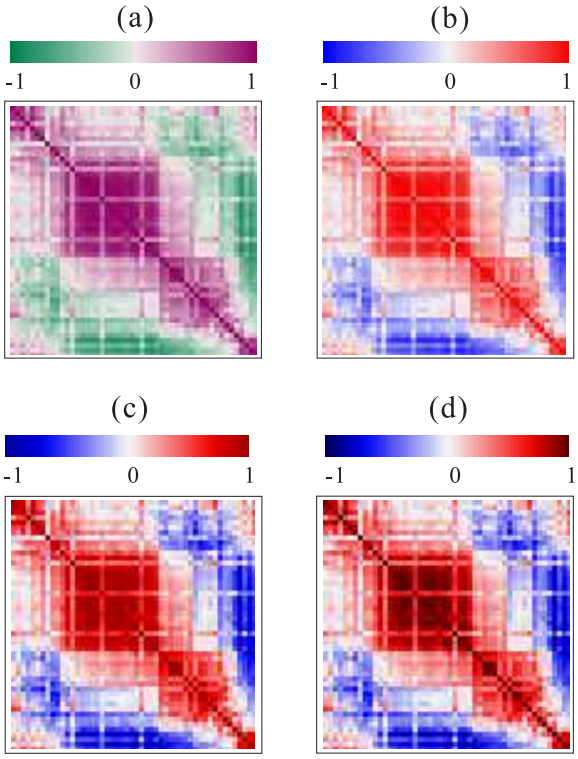

Fig. 2. Four color spectra applied to the same correlation matrix map for fifty psychosis disorder variables (Chen (2002)).

Display Conditions

Display condition is analogous to data transformation for colors. Usually, the whole color spectrum is used to represent the complete range of values in the data matrix (range matrix condition). The matrix condition can be switched to row or column conditions for emphasizing individual variable distributions

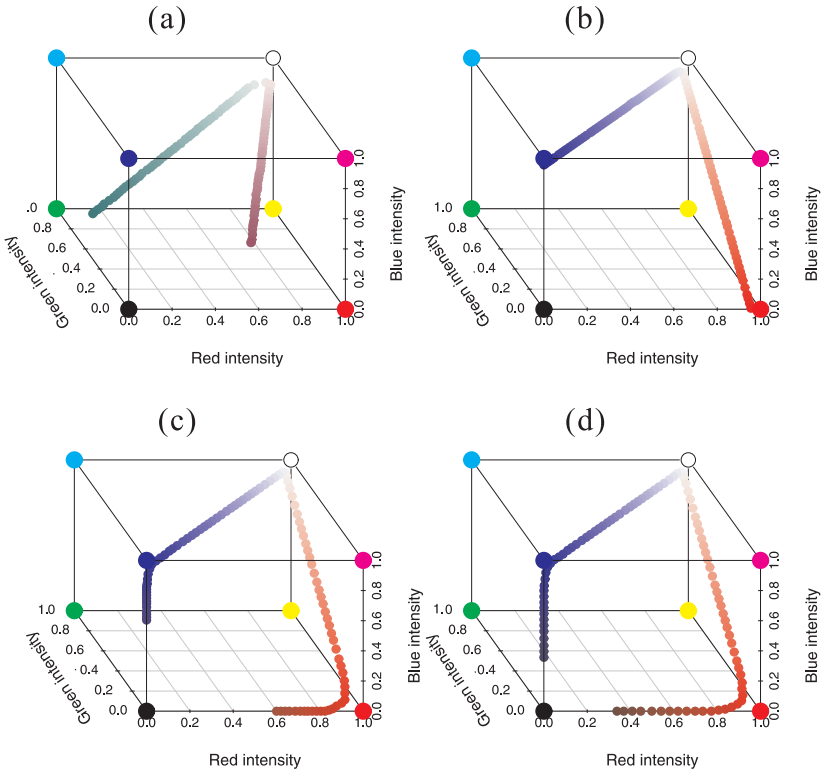

Fig. 3. Relative (red, green, blue) hues in the RGB cubes for the four color spectra in Figure 2.

or subject profiles. For a bi-directional color spectrum (green-black-red for differential gene expressions, blue-white-red for correlation coefficients), the center matrix condition symmetrizes the color spectrum around the baseline numeric value (1:1 for log2 ratio gene expression, zero for correlation coefficient). On occasion, we might like to downweight the effects of extreme values in the data set, and the use of ranks as a replacement for numerical values is a possibility. This is termed the rank matrix condition.

Resolution of a Statistical Graph

If the data matrix or proximity matrices contain potential extreme values, the relative structure of the extreme values to the main data cloud will dominate the overall visual perception of the raw data map and the proximity matrix maps. The problem can be handled by using rank conditions or by compressing the color spectrum to a suitable range. Variously, we can apply a logarithm or similar transformation to reduce the outlier effect or to simply remove the outlier.

3.2 Seriation of Proximity Matrices and Raw Data Matrix

Without suitable permutations (orderings) of the variables and samples, matrix visualization is of no practical use in visually extracting information (Figure 1, Left Panel). It is necessary to compute meaningful proximity measures for variables and samples, and to apply suitable permutations to these matrices before matrix visualization can reveal information structure of the given data set. We discuss below some concepts and criteria for evaluating the performances of different seriation algorithms in reordering related matrices.

Relativity of a Statistical Graph

Chen (2002) proposed a concept, the relativity of a statistical graph, for evaluation of general statistical graphic displays. The idea is that of placing similar (different) objects at closer (more distant) positions in a statistical graph. In a continuous display, such as the histogram or a scatterplot, relativity always holds automatically. An illustration is the histogram, in Figure 4, of the Petal Width variable and a scatterplot of Petal Width and Petal Length variables for 150 Iris flowers (Fisher (1936)). Two flowers coded in $\times$ and $\circ$ are placed next to each other on these two displays automatically, because they share similar petal widths and lengths. Friendly and Kwan (2003) proposed a similar concept for ordering information in general visual displays which they called the effect-ordered data display. Hurley (2004) also studied related issues with examples in scatterplot matrices and parallel coordinate plots.

The relativity concept does not usually hold for a matrix visualization or parallel coordinate plot type of display since one can easily destroy the property with a random permutation. It is a common practice to apply various permutation algorithms to sort the columns and rows of the designated matrix so that similar (different) samples/variables are permuted at closer (distant) rows/columns.

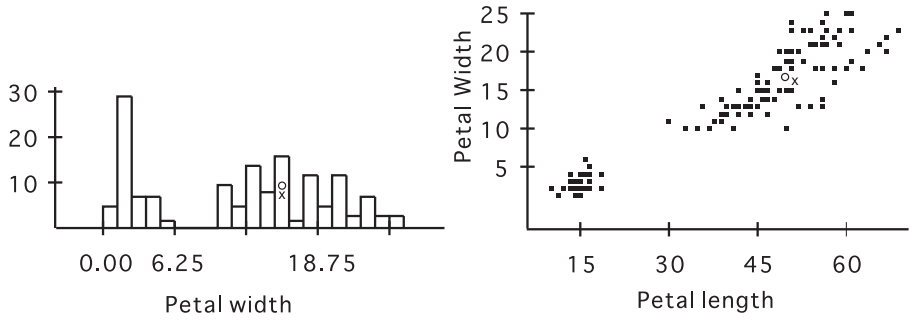

Fig. 4. Concept of Relativity of a Statistical Graph for a continuous data set (the Iris data).

Global Criterion: Robinson Matrix

It is usually desired to permute a matrix to resemble as closely as possible a Robinson matrix (Robinson (1951)) because of the smooth and pleasant visual effect on examining permuted matrix maps. A symmetric matrix is called a Robinson matrix if its elements satisfy $r_{ij} \leq r_{ik}$ if $j < k < i$ and $r_{ij} \geq r_{ik}$ if $i < j < k$. If the rows and columns of a symmetric matrix can be permuted to those of a Robinson matrix, we call it pre-Robinson. For a numerical comparison, three anti-Robinson loss functions (Streng, (1978)) are calculated for each permuted matrix, $D = \{d_{ij}\}$, for the amount of deviation from a Robinson form with distance-type proximity:

$$AR(i) = \sum_{i=1}^{p}[\sum_{j<k<i} I(d_{ij} < d_{ik}) + \sum_{i<j<k} I(d_{ij} > d_{ik})],$$

$$AR(s) = \sum_{i=1}^{p}[\sum_{j<k<i} I(d_{ij} < d_{ik}) \cdot |d_{ij} - d_{ik}| + \sum_{i<j<k} I(d_{ij} > d_{ik}) \cdot |d_{ij} - d_{ik}|],$$

$$AR(w) = \sum_{i=1}^{p}[\sum_{j<k<i} I(d_{ij} < d_{ik})|j-k||d_{ij}-d_{ik}| + \sum_{i<j<k} I(d_{ij} > d_{ik})|j-k||d_{ij}-d_{ik}|].$$

$AR(i)$ counts only the number of anti-Robinson events in the permuted matrix; $AR(s)$ sums the absolute value of anti-Robinson deviations; $AR(w)$ is a weighted version of $AR(s)$ penalized by the difference of column indices of the two entries.

Elliptical Seriation

Chen (2002) introduced a permutation algorithm called rank-two elliptical seriation that extracts the elliptical structure of the converging sequence of iteratively formed correlation matrices using eigenvalue decomposition. Given a p-dimensional proximity matrix D, a sequence of correlation matrices $R = (R^{(1)}, R^{(1)}, \cdots)$ is iteratively formed from it. Here $R^{(1)}$ is the correlation matrix of the original proximity matrix D, and $R^{(n)}$ is the correlation matrix of $R^{(n-1)}$ for $n > 1$. The iteratively formed sequence of correlation matrices gradually cumulates the variation information to the leading eigenvectors. At the iteration with rank two, there are only two eigenvectors left with non-zero eigenvalues, and all information is reduced to the ellipse spanned by the two eigenvectors. Every object has its relative position on this two-dimensional ellipse, and a unique permutation is obtained. The elliptical seriation usually identifies very good global permutations, and is useful for identifying global clustering patterns and smooth temporal gene expression profiles (Tien *et al.* (2006)) by optimizing the Robinson criterion.

Local Criterion: Minimal Span Loss Function

The minimal span loss function $MS = \sum_{i=1}^{n-1} d_{i,i+1}$ for a permuted matrix $D = \{d_{ij}\}$ focuses on the optimization of local structures. The idea is to find a shortest path through all data elements as in the travelling salesman problem. The local seriation method produces tighter blocks than the global method does around the main diagonal of the proximity matrix. In addition, we can combine the anti-Robinson measure and minimal span loss into a measure in which a band along the diagonal of a proximity matrix is selected with width w $(0 < w < n)$ and the anti-Robinson measurement is computed within that band.

Tree Seriation

The hierarchical clustering tree with a dendrogram (Eisen *et al.* (1998)) is the most popular method for two-way sorting the gene-by-array matrix map employed in gene expression profiling. The ordering of terminal nodes generated by an agglomerative hierarchical clustering tree automatically keeps good local grouping structure, since the tree dendrogram is constructed through a sequential bottom-up merging of "most similar" sub-nodes. On the other hand, a divisive hierarchical clustering tree usually keeps better global patterns through a top-down splitting of "most heterogeneous" substructures. Divisive hierarchical clustering trees are rarely used due to their computational complexity.

Flipping of Intermediate Nodes

One critical issue in applying the leaves of the dendrogram in sorting the rows/columns of an expression profile matrix is the flipping of the intermediate nodes. As illustrated in Figure 5 with a schematic dendrogram (Figure 5a), the $n-1$ intermediate nodes (red points) for a dendrogram of n objects can be flipped independently (Figure 5b) resulting in $2^{(n-1)}$ different dendrogram layouts (Figure 5c, for example) and corresponding permutations for the n objects with identical proximity matrices (Pearson correlation or Euclidean distance) and the same tree linkage method (single, complete, average or centroid). The flipping mechanism of intermediate nodes can be guided either by an external or internal reference list. For example, the Cluster software developed by Eisen's lab (1998) guides the tree flips based on the average expression level. He also suggests one can use the results of a one-dimensional SOM to guide the tree seriation. This makes the tree seriation as close to the external references as possible. In Alon *et al.* (1999), it is suggested that one order the leaf nodes according to the similarity between a node and its parent's siblings. Bar-Joseph (2001) proposed the fast optimal leaf ordering for hierarchical clustering that maximizes the sum of the similarities of adjacent leaves in the ordering. These are two examples of internal references.

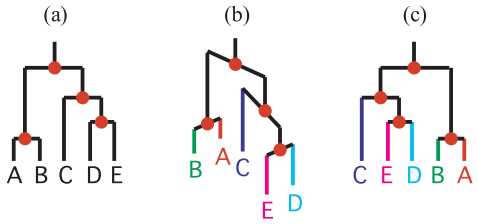

Fig. 5. Flipping mechanism for intermediate notes of a dendrogram.

4 Generalization and Flexibility

4.1 Summarizing Matrix Visualization

Sorted matrix maps are capable of displaying the raw expression pattern and the association structures among genes and arrays. One can go one step further to identify clusters in the permuted matrix maps using the dendrogram branching structure or other partitioning methods, such as the converging sequence of Pearson's correlation matrices (Chen (2002)) and block searching (Hartigan (1972)). Once the partitioned matrix maps are obtained (Figure 6, Left Panel), a summarizing matrix visualization which Chen (2002) coined sufficient matrix vsiualization can be constructed by representing individual data points and proximity measures in each identified subject-subject, variable-variable and subject-variable block by the summary statistic (means, medians or standard deviations) for that particular block.

The three maps in Figure 6, Right Panel summarize the sufficient information of the data matrix and the corresponding proximity matrices for the gene expression profiles in the Left Panel. In the sufficient MV of Figure 6, Right Panel, users can easily extract the within and between correlation structure for the three array-groups, the relative clustering pattern of the four gene-clusters, and the interaction behavior of the four gene-clusters on the three array-groups. Three essential steps are necessary to ensure the effectiveness of a sufficient MV in extracting the overall information structure embedded in the original data matrix and two proximity matrices: (1) appropriate permuted variables and samples; (2) carefully derived partitions for variables and samples; and (3) representative summary statistics.

4.2 Sediment Display

The sediment display of a row data matrix for rows (columns) is constructed by sorting the column (row) profiles for each row (column) independently according to its magnitude. This display expresses the distribution structure for all rows (columns) simultaneously. The middle panel of Figure 7 has the sediment display for all 30 gene expression profiles, while the right panel has

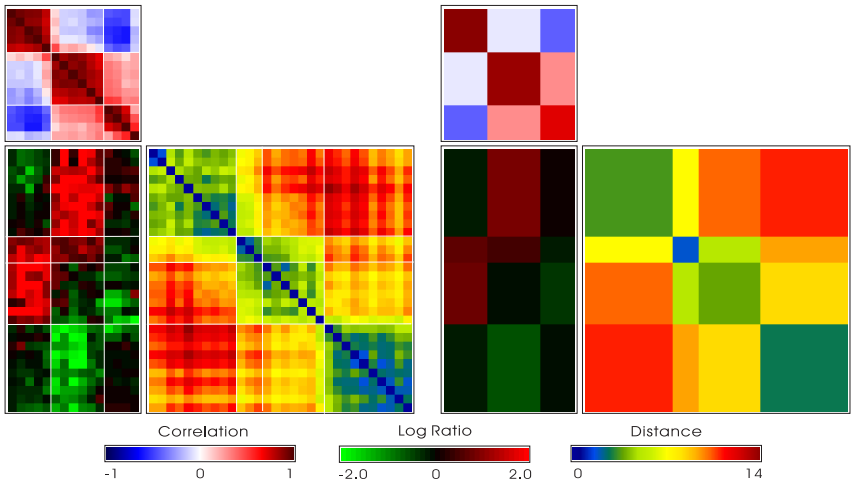

Fig. 6. Left: Partitioned data and proximity matrix maps for Data Set 1. Right: Sufficient data and proximity matrix maps.

the expression distributions for each of the 15 selected arrays. The sediment displays for genes and arrays convey similar information to that given by a boxplot when the color strips at the quartile positions are extracted.

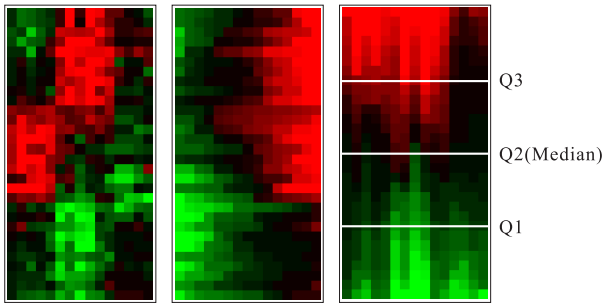

Fig. 7. Sediment displays for genes (middle panel) and arrays (right panel) for the permuted data matrix (left panel) of Data Set 1.

4.3 Sectional Display

The purpose of a sectional display is to exhibit only those numerical values that satisfy certain conditions in the data or the associated proximity maps. For example, one can choose to ignore the values below some threshold by not displaying the corresponding color dots. For a permuted distance map, one can emphasize more coherent neighboring structure by displaying only the

corresponding neighbors dynamically. Figure 8 has a series of such sectional displays for the distance matrix for genes in Figure 1, Right Panel (and Figure 6, Left Panel).

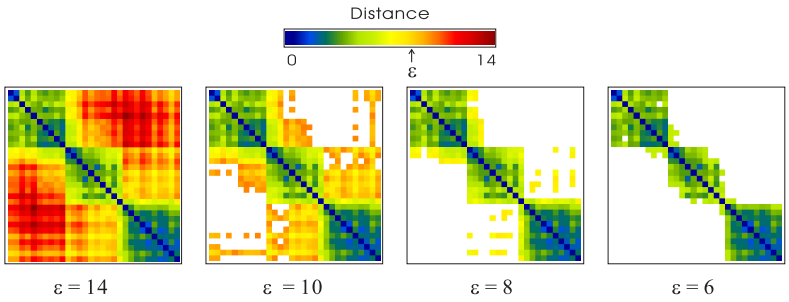

Fig. 8. Sectional displays for the permuted gene distance map. Only distances smaller than the threshold, ϵ, are displayed.

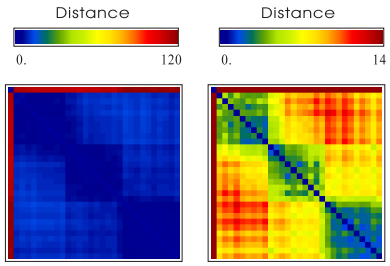

Fig. 9. Original (left) and restricted (right) displays for the permuted gene distance map in Figure 8 with an outlying gene added.

4.4 Restricted Display

Outlying data points or proximity measures may mask detailed color resolutions. The situation can be improved by displaying only rank conditions instead of original magnitudes, or by compressing the color spectrum to represent only the main body of the data values, i.e., one displays data values that fall in some range of the data using the whole color spectrum. Figure 9 Left Panel shows the restricted display of Figure 8 with an artificial outlier observation added. The relative large distance of this outlier to other observations exhausts the color spectrum and masks the main feature embedded in the distance matrix. The right panel of Figure 9 uses the whole rainbow spectrum to represent only the distance range between 0 and 14 and thus reveals the main

three-group structure. A nonlinear color mapping (for the distances) like the one implemented in MANET (Unwin (1998)) can also resolve the problem.

5 An Example

Construction of An MV Display

Many microarrays in Data Set 0 have many missing values because of technical issues and because different experiments studied different sets of genes in the yeast genome. Two thousand genes with four hundred arrays with relatively fewer missing values were then selected from the original Data Set 0, resulting in Data Set 2. Illustrated in Figure 10 is the MV display of Data Set 2. Pearson's correlation coefficient is used for measuring both the between genes and between arrays association, as commonly practiced in gene expression profile analysis. Average linkage clustering trees are then grown on the two correlation matrices for genes and arrays. Relative positions of the terminal nodes of the two dendrograms are then applied to sort the corresponding correlation matrix maps and the data matrix map (gene expression profile). The basic gene clustering structure and array (experiments) grouping patterns can be identified using these tree sorted matrix maps.

The enlarged permuted data matrix map for gene expression profiling is displayed in Figure 11. Red dots represent relatively high expression of message RNA of gene/experiment combinations, green dots display relatively low expression ones, with black dots designating relatively little differential expression. Missing values are coded in white so one can see that many arrays (experiments) still contain some missing observations. Such an MV display presents each gene expression profile as a horizontal strip of color dots across all arrays (experiments), and the important visual information is carried by the relative variation of hues of colors.

Without suitable permutations to sort the similar genes at closer rows and identical arrays next to each other so that the relativity property holds, an MV display is basically useless. From this two-way permuted display, one looks for horizontal strips of genes that share similar expression profiles, and vertical strips of arrays that exhibit close experimental results. The blocks of the two directions illustrate the interaction patterns of gene-clusters and experiment-groups. All of the numerical information is displayed in this raw expression profile map (with proximity maps for genes and arrays and corresponding dendrograms). Careful and patient examination of these color maps can lead to valuable insights on embedded information structure.

Examination of An MV Display

As with other visualization tools, proper training and experience is need to get the most information out of these complex matrix visualization displays.

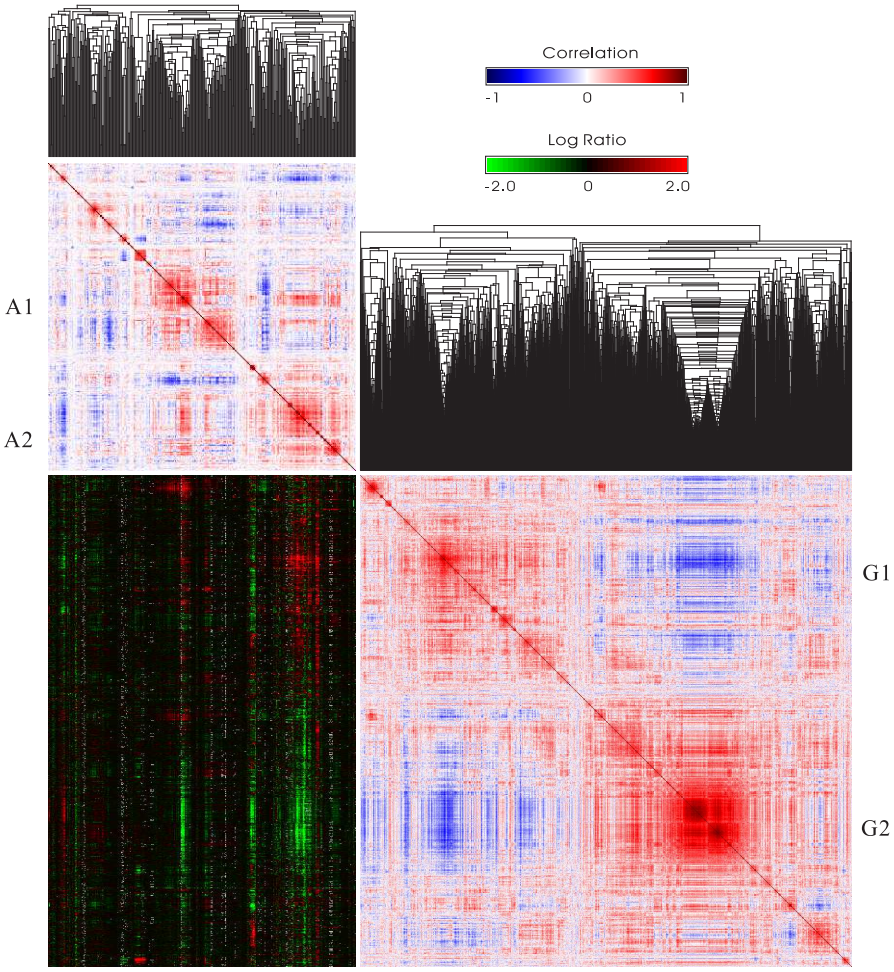

Fig. 10. Average linkage trees (for both genes (rows) and arrays (columns)) permuted data matrix ($\log_2$ ratio gene expression) with two proximity matrix maps (Pearson correlation for both genes and arrays) for Data Set 2.

While examining a complex MV display such as that in Figures 10 and 11, several general steps are to be taken:

1. For column (array) proximity matrix:
   a) Search for coherent clusters of arrays along the main diagonal of the correlation (maybe distance for other circumstances) matrix with darker red points. Two dominant array groups of arrays can be identified around the middle and at the lower-right corner of the correlation matrix with several small but coherent clusters scattered along the main diagonal. Let's denote these two major groups of arrays as A1

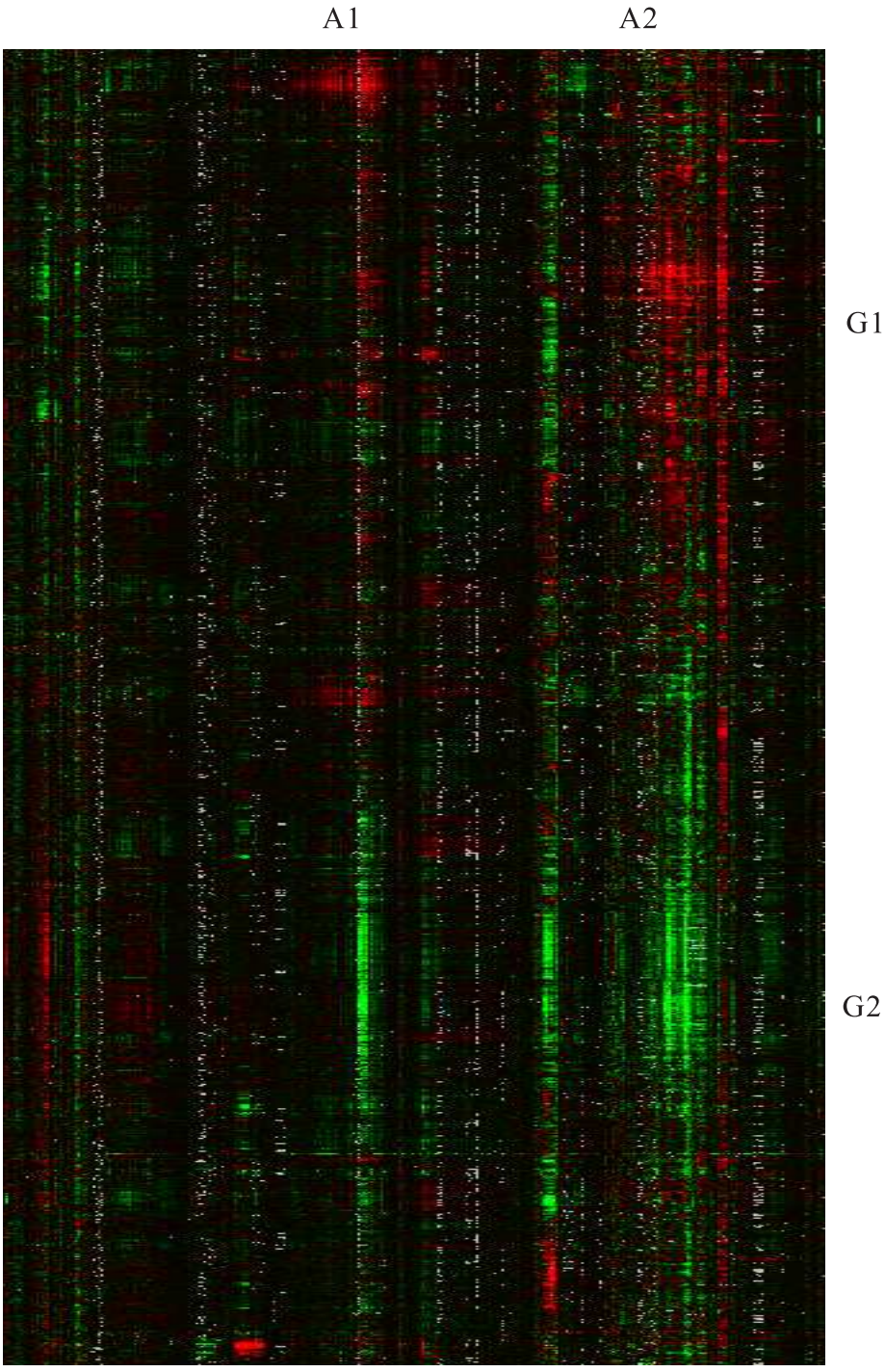

Fig. 11. Enlarged expression profile matrix map of Figure 10 (missing observations are coded in white).

and A2. The arrays grouped into these clusters must have similar expression patterns across all the 2000 genes (to be examined at later steps).

b) Look for interactions between the array-clusters at off-diagonal locations. Various types of between-cluster correlation patterns with sub-structures can also be easily pined down.

c) The arrays represent many different biological assays for various functions of Saccharomyces cerevisiae yeast such as cell-cycle control, stress (environmental changes, relevant drug-affected), metabolic/genetic control, transcriptional control and DNA-binding (`http://transcriptome.ens.fr/ymgv/`). Different biomedical assays activate and suppress expression patterns of certain functional groups of genes. We need to integrate these biological/medical knowledge with the numerical/graphical findings in 1.a) and 1.b) for validating known information and more importantly for exploring and interpreting novel interesting patterns.

d) Both hierarchical clustering trees for arrays and genes also provide partial visual exploration of the data and proximity structure but not as comprehensive as direct visualization of the two proximity matrix maps since the dendrograms only keep partial information of the proximity matrices from which they are constructed.

2. For row (gene) proximity matrix:

a) Similar procedures as in 1.a) and 1.b) for arrays (columns) have to be repeated here for the genes (rows) proximity matrix. Of particular interest is the dichotomous pattern of these 2000 genes. The up-regulated (red) genes at the upper half and the down-regulated (green) genes at the bottom half of the A2 arrays are responsible for this dichotomous structure. We shall denote these two clusters of genes as G1 and G2. Several small sub-clusters of genes within G1 and G2 can also be identified along the main diagonal.

b) It is necessary to go one step further to consult various annotation databases for more detailed interpretations and explanations of the potential clusters of genes identified this way. Some of the genes are not annotated yet. Their potential functions can be roughly determined through the positively correlated (up-regulated) gene-clusters and negatively correlated (down-regulated) patterns.

3. For raw data (gene expression profile) matrix:

a) Many major and minor array-groups and gene-clusters have been found in steps 1 and 2. In step 3 we use the raw data (gene expression profile) matrix map to search for the interaction patterns of the gene-cluster on each array-group. It is also necessary to use vertical-strips of expression profiles to contrast between array-group structure variations and horizontal-strips to identify between gene-cluster distribution differences. With careful examination, one can associate certain pieces of expression block in raw data matrix to the formation of

each array-group and gene-cluster and the between groups (clusters) differentiation.

b) There are about 10000 (∼1.25%) missing observations in this data matrix of 400 arrays with 2000 genes. One sees that the missing pattern is not of random manner. Different array-group and gene-cluster combinations are associated with various proportions of missing observations. The visualization of the missing structure greatly assists users in choosing more appropriate missing value estimation or imputation machanism for further analyses.

c) These visual information provide valuable insight into more advanced studies such as the confirmation of existing metabolite pathways (see Section 7 for an example) and exploration of novel pathways.

This paragraph only discusses some general issues in the examination of such an MV display. There are actually many more interesting patterns to be explored with the input of expert knowledge and interaction with the biologists who are familiar with related experiments of Saccharomyces cerevisiae yeast. In Figure (10) and (11) we demonstrated an MV can easily handle thousands of samples. An MV display can also handle thousands of variables since samples and variables are treated symmetrically in the MV framework.

6 Comparison with Other Graphical Techniques

In this section, we discuss the visualization efficiency of the scatterplot (SP), the parallel coordinate plot (PCP)(Inselberg (1985), Wegman (1990)), and matrix visualization (MV), based on the dimensionality of the data at small, moderate, and large sample sizes.

Low Dimension

For one-dimensional data, a scatterplot and the PCP display amount to a dotplot, while one-dimensional MV yields a colored bar chart. In any event, it is unlikely that an alternative to the histogram can prove more popular in the display of one-dimensional data. For two-dimensional data, a scatterplot is the most efficient graphical display. While a PCP display relies on the n connecting line segments between two vertical dotplot to represent the association of the two variables, an MV displays each sample as a single row with two color dots. The efficiency of scatterplots decreases with increasing dimension. For three-dimensional data, a rotational scatterplot is commonly practiced for extracting geometric structure through a sequence of two-dimensions scatterplots with changing angles controlled by the user. The usefulness of PCP and MV displays of three-dimensional data is a subtle point, and the appropriate permutation of variables is surely needed to enforce relativity for both types of displays.

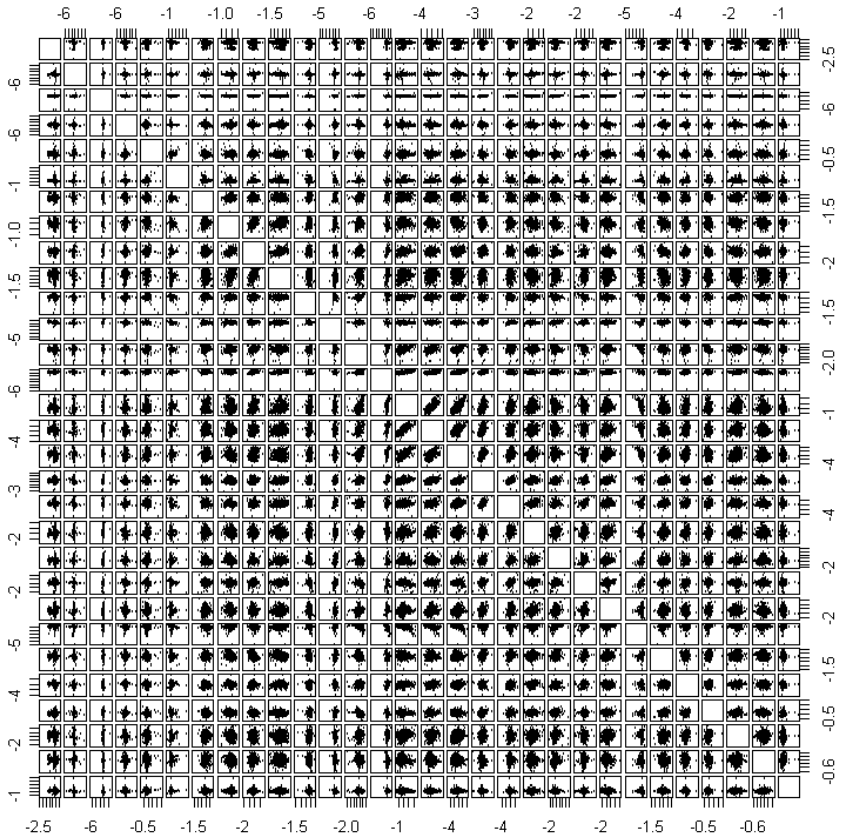

Fig. 12. Scatter-plot matrix for the first thirty arrays of Data Set 2.

High Dimension

A scatterplot matrix (SM) is used to simultaneously visualize information structure embedded in all $C(p, 2)$ pairs of variables for data dimension larger than three. Grand Tours (Asimov, (1985)) are sometimes undertaken in the hope of extracting high- dimensional data structure through rotation of randomly projected three-dimensional plots. Dimension reduction techniques, such as principal component analysis, are also useful for displaying structural information from high-dimensional data to low-dimensional displays. Figure 12 shows the scatterplot matrix display of the first 30 variables (arrays) in Data Set 2, while Figure 13 gives the corresponding PCP for these data. We note that a PCP display of high dimensional data with a large sample size can simultaneously display all the samples, but it is usually necessary to use some interactive mechanism for selecting subsets of samples in order to study

the relative structure across all variables, as in Figure 13. Moreover, for these plots, more than one pixel width is needed to display each variable.

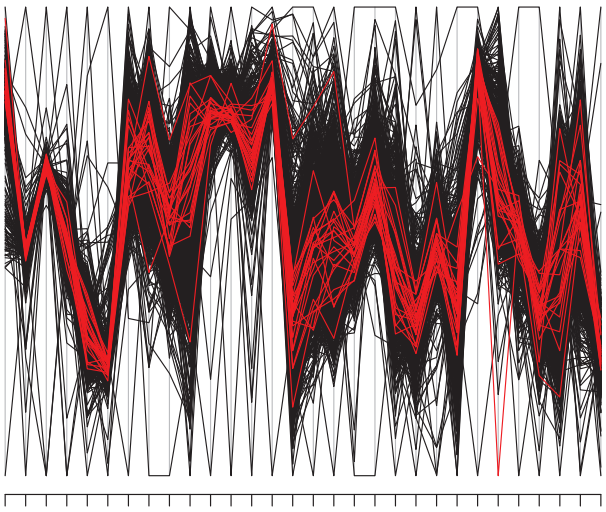

Fig. 13. Parallel coordinate plot for the first thirty arrays of Data Set 2.

Generally, a scatterplot-matrix needs $C(p, 2) \times n$ dots to display a data set with n samples measured on p variables, a PCP display needs p vertical lines plus $(p - 1) \times n$ line segments, and an MV plots requires $n \times p$ dots. When p becomes large, larger than 15 say, a scatterplot-matrix is basically useless. A PCP display does well with up to a few hundred variables, but founder at higher levels due to the space required for displaying the line segments connecting sample points. on the other hand, a scatterplot-matrix wasted a high proportion of display space. An MV display, on the other hand, utilizes every column pixel for displaying a variable on a computer screen. PCP has an advantage over MV on the sample side, but MV plots provide better resolution.

Overall Efficiency

Schematically illustrated in Figure 14 is a diagram of efficiency against dimensionality for conventional scatterplot (matrix) and dimension-free visualization tools such as the parallel coordinate plot (PCP) and matrix visualization (MV). While direct visual perception of geometric pattern makes scatterplots most efficient for visualization of low-dimensional data, MV and PCP definitely have the advantages for visualization of data sets with fifteen or more variables.

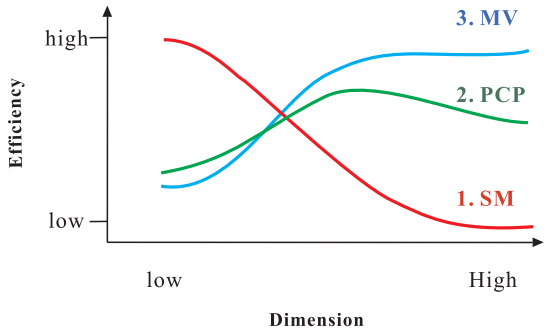

Fig. 14. Schematic illustration for the relative efficiency of the Scatter-plot matrix, the Parallel coordinates plot, and Matrix visualization, with varying numbers of dimensions.

Missing Values

It is very difficult to display missing values in a scatterplot while one can always displays missing values above or below the regular data range of each variable for a PCP display. The MANET system by Unwin *et al.* (1996) can be used to display missing information interactively. In an MV plot, a missing value can be simply displayed at the corresponding position (row and column) with a color that can be easily distinguished from the color spectrum of the numerical values. The missing values of the gene expression profiles of Figures 10 and 11 are coded in white. With appropriate permutations for rows and columns, the corresponding variable/sample combinations of missing structure can be visually accessed. MV users can have a simple visual perception of the missing mechanism (random or not, ignorable or nonignorable) of the data (variables) before formal statistical modeling of missing values is implemented.

7 Matrix Visualization for Binary Data

While scatterplots, PCP, and MV displays have their own advantages and disadvantages over varying dimension size for continuous data structure, an MV display is the only statistical graph that can meaningfully display binary data sets over all dimensions. We use the KEGG (Kyoto Encyclopedia of Genes and Genomes) metabolism pathways (`http://www.genome.jp/kegg/pathway.html`) for Saccharomyces cerevisiae yeast to illustrate how an MV display can be generalized to visually extract all important information embedded in multivariate binary data. The KEGG web site provides detailed information on the 1177 related genes involved in 100 metabolism pathways of Saccharomyces cerevisiae yeast. We simplified the complex information structure down to a two-way binary data matrix of 1177 genes by 100 pathways. This binary data matrix is called Data Set 3 in our study. A one (zero) encoded at the i-th row

and j-th column of the matrix means the i-th gene is (not) involved in j-th pathway activities.

7.1 Similarity Measure for Binary Data

The usual measurements for evaluating associations between samples and variables for continuous data, Euclidean distance and correlation coefficients, cannot be applied directly to binary data sets. Two issues are noted here in the selection of similarity measures for binary data in an MV display.

Symmetric and Asymmetric Binary Variables

A binary variable is considered symmetric if both of its states are equally valuable, that is, there is no preference on which outcome should be coded as 0 or 1. A binary variable is asymmetric if the outcomes of the states are not equally important, such as the positive and negative outcomes of a disease diagnosis. Conventionally the most important outcome, the rarer one, is coded as 1, the other as 0. Thus, asymmetric binary variables are often considered "monary" (as if there is only one state).

Sparseness and Dimensionality

Asymmetric binary variables are usually sparse in nature and it is difficult to identify appropriate association measurement for assessing the relationships among samples and between variables. Dimension reduction techniques also fail in attempts at summarizing high-dimensional data structure in low-dimensional fashion. Listed in Figure 15 are some commonly used similarity measurements for binary data. For sparse data, it is common practice to use the Jaccard coefficient instead of the simple matching coefficient.

7.2 Matrix Visualization of the KEGG Metabolism Pathway Data

The $1-$Jaccard distance coefficient is used to compute the proximity matrices for both genes and pathways in Figure 16. Elliptical seriations (Chen (2002)) are employed to permute the two $1-$Jaccard distance matrices and the binary pathway data matrix. One can easily see, from the binary data matrix map and the proximity matrix map for genes, that there are many genes that are only involved in the activities of a single pathway. We then exclude those genes from further analysis, since they provide no association information. This reduces the original 1177 genes by 100 pathways binary matrix to a 432 genes by 88 pathways matrix (some pathways are also excluded after the exclusions of genes). When not enough horizontal or vertical pixels are available for MV display, users can either use the scroll bars to visualize certain portion of the display or to zoom out the display to visualize the overall structure with averaging effect as in a typical computer graph.

Binary Data		Object B		
		1	0	
Object A	1	a	b	(a+b)
	0	c	d	(c+d)
		(a+c)	(b+d)	(a+b+c+d)

Similarity for Binary Data	Formula
Kulczynskl	$\frac{a}{b+c}$
Rao	$\frac{a}{a+b+c+d}$
Jaccard	$\frac{a}{a+b+c}$
simple match	$\frac{a+d}{a+b+c+d}$
Sneath	$\frac{a}{a+2b+2c}$
Rogers	$\frac{a+d}{a+2b+2c+d}$
Hamman	$\frac{a+d-(b+c)}{a+b+c+d}$
Phi	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
Yule	$\frac{ad-bc}{ad+bc}$

Fig. 15. Some similarity measures for binary data.

Average linkage clustering trees are then employed to sort the resulting 1−Jaccard distance matrices for genes and pathways and the corresponding binary data matrix, see Figure 17. The association structure between genes and among pathways can now be comprehended using the three corresponding permuted matrix maps. In the upper left corner of the data matrix and the upper-left corner of the proximity maps for genes and pathways, we can identify several groups of genes involved in activities of only a few pathways, and several small groups of pathways that share highly similar groups of genes. The rest of the genes and pathways have more complicated interactions of activities. It is of course possible to further exclude pathways and genes with simpler behavior, and to focus on the details of interactions of those more active genes and pathways.

8 Other Modules and Extensions of MV

We have so far introduced the fundamental framework for matrix visualization in the GAP (Chen (2002)) approach to visualization of continuous and

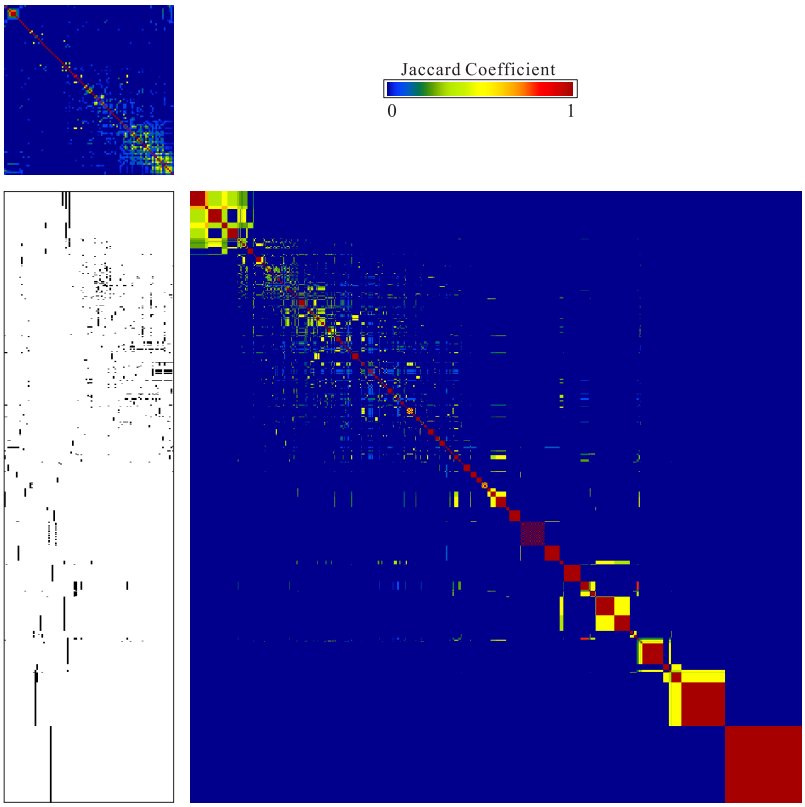

Fig. 16. Binary data matrix map for Data Set 3 (KEGG metabolite pathway database with 1177 genes (rows) for 100 pathways (columns)) with two Jaccard proximity matrices for genes and for pathways sorted by elliptical seriations on both directions.

binary data matrices, with corresponding derived proximity matrices. We have also presented some generalizations, such as the sufficient MV, the sediment, sectional, and restricted displays. In practice, observed data can be highly complex, to the degree that the basic matrix visualization procedures are not rich enough to comprehend the data structure. In some situations one may not be able to apply MV directly to the given data or proximity matrices. This section discusses ongoing projects and future directions that will make matrix visualization a more promising statistical graphical environment. One important feature of the GAP (Chen (2002)) approach to matrix visualization is that it usually contains four basic procedures: (1) color projection of the raw data matrix; (2) computation of two proximity matrices for variables and sample; (3) color projection of the two proximity matrices; and (4) permutations of variables and sample. Most of the extensions of MV have something

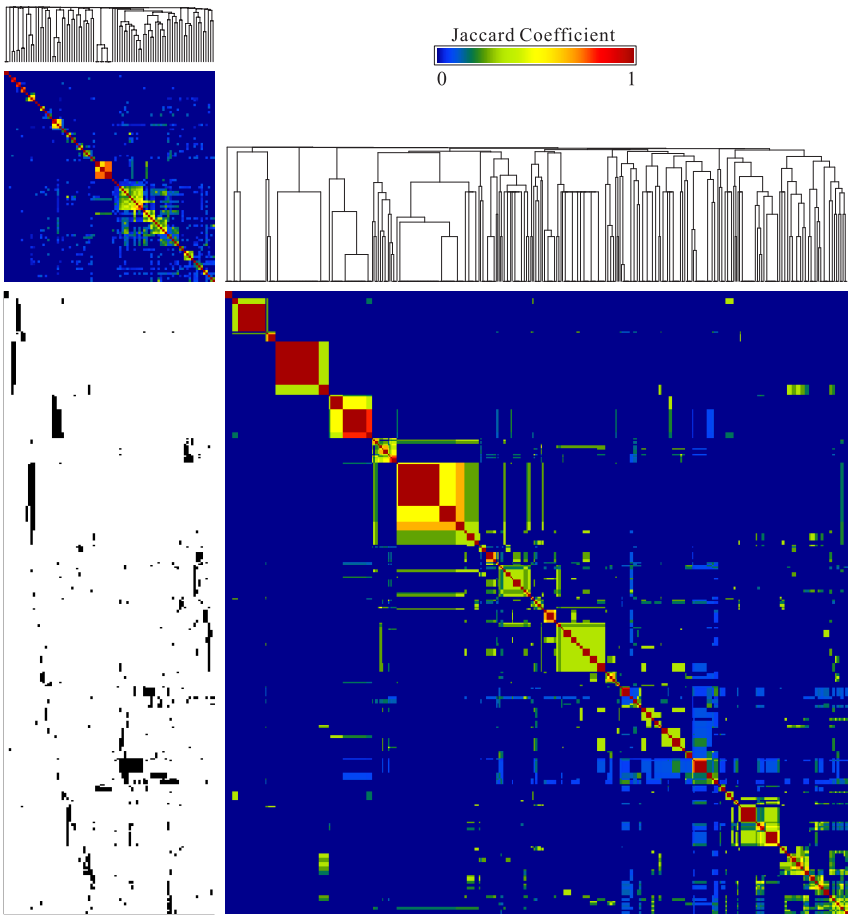

Fig. 17. Binary data matrix for reduced Data Set 4 (432 genes (rows) for 88 pathways (columns)) with two Jaccard proximity matrices for genes and for pathways?sorted by?average linkage trees on both directions.

to do with the first two procedures. The aforementioned algorithms for the other two steps can be simply adapted once the first two procedures are fixed.

8.1 MV for Nominal Data

It is much more difficult to create MV for nominal data than it is for binary data, since one can use black/white to code 1/0 if the binary data is of asymmetric nature and can use the Jaccard and other coefficient for binary data to derive the between variables and between sample relationships. There is no natural way to guide the color-coding for multivariate nominal data in such a way that the color version of relativity of a statistical graph still holds (Chang

et al. (2002)). Derivation of meaningful between-variable and between-sample proximity measures for nominal data is another challenging issue. Chen (1999) and Chang *et al.* (2002) utilized the Homals (de Leeuw (1998)) algorithm and developed a categorical version of matrix visualization that naturally resolved the two critical problems.

8.2 MV for Covariate Adjustment

Quite often covariate data, such as gender and age, are collected in a study in addition to the variables of primary interest. When effects of covariates are at issue, covariate adjustment has to be taken into consideration much as in a formal statistical modelling process. Wu and Chen (2006) introduced a unified regression model approach which partitions the raw data matrix into model and residual matrices, and ordinary MV can be applied on these two derived matrices. The covariate adjustment process is accomplished through the estimation of conditional correlations. For a discrete covariate, a correlation matrix for variables is decomposed into within- and between-component matrices. When the covariate is continuous, the conditional correlation is equivalent to the partial correlation under the assumption of joint normality.

8.3 Data with Missing Values

The relativity of a statistical graph (Chen (2002)) is the main concept in seriation algorithms for constructing meaningful clustered matrices. This property can be used for developing a weighted pattern method to impute the missing values. The initial proximity measurements for rows and columns with missing values can be computed with pair-wise complete observations first, then imputed values are estimated and updated iteratively for the subsequent proximity calculations and imputation.

8.4 Modelling of Proximity Matrices

Many statistical modelling procedures try to visually explore the high- dimensional pattern embedded in a proximity matrix that records pair-wise similarity or dissimilarity for a set of objects through a low-dimensional projection. Multidimensional scaling, hierarchical clustering analysis, and factor analysis are three such statistical techniques. Four types of matrices are usually involved in the modelling process of these statistical procedures. The input proximity matrix is transformed into a disparity matrix prior to fixing the statistical model that summarizes the information into the output distance matrix. A stress (residual) matrix is calculated for assessing the badness-of-fit of the modelling. Such a study aims at creating a comprehensive diagnosis environment for statistical methods through various types of matrix visualization for the numerical matrices involved in the modelling process.

9 Conclusion

Matrix visualization is the color order-based representation of data matrices. It is of benefit to employ human vision for exploration of the structure in a matrix in the pursuit of further appropriate mathematical operations and statistical modelling. A good matrix visualization environment helps us gain comprehensive insights into the underlying process. Rather than rely solely on numerical characteristics, matrix visualization is suggested as a preliminary step in modern exploratory data analysis and is a continuing and active topic of research and application.

A matrix visualization displays provide five levels of information: (1) raw scores for every sample/variable combination; (2) an individual sample score vector across all variables, and an individual variable vector across all samples; (3) an association score for every sample-sample and variable-variable relationship; (4) a grouping structure for variables and a clustering effect for sample; and (5) an interaction pattern of sample-clusters on variable-groups.

With the capacity for displaying thousands of variables in a single picture, the flexibility for working with all types of data, and the ability for handling the various manifestations of extraordinary data patterns (missing value, covariate adjustment), we believe matrix visualization has the opportunity to become one of the major graphical tools for the new generation of exploratory data analysis (EDA).

References

1. Alon, U., Barkai,N., Notterman, D.A. Gish, K. Ybarra, S. Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc Natl Acad Sci U S A.*, 96(12):6745-V 6750.
2. Asimov, D. (1985) The grand tour: a tool for viewing multidimensional data, *SIAM Journal of Scientific and Statistical Computing*, 6(1):128 – 143.
3. Bar-Joseph, Z., Gifford, D.K. and Jaakkola, T.S. (2001). Fast optimal leaf ordering for hierarchical clustering, *Bioinformatics*, 17:S22-V S29.
4. Bertin, J. (1967). *Semiologie Graphique*, Paris: Editions gauthier-Villars. English translation by William J. Berg. as Semiology of Graphics: : Diagrams, Networks, Maps. TheUniversity of Wisconsin Press, Madison, WI, 1983.
5. Carmichael, J.W. and Sneath, P.H.A. (1969). Taxometric maps, *Systematic Zoology*, 18:402 – 415.
6. Chang, S.C., Chen, C.H., Chi, Y.Y., and Ouyoung, C.W. (2002). Relativity and resolution for high dimensional information visualization with generalized association plots (GAP), *Section for Invited Papers, Proceedings in Computational Statistics 2002 (Compstat 2002)*, Berlin, Germany, 55 – 66.
7. Chen, C.H. (1996). The properties and applications of the convergence of correlation matrices, in *1996 Proceedings of the Section on Statistical Graphics of the American Statistical Association*, 49 – 54.

8. Chen, C.H. (1999). Extensions of generalized association plots, *1999 Proceedings of the Section on Statistical Graphics of the American Statistical Association*, 111 – 116.
9. Chen, C.H. (2002). Generalized association plots: information visualization via iteratively generated correlation matrices, *Statistica Sinica,* 12:7 – 29.
10. Chen, C.H., Hwu, H.G., Jang, W.J., Kao, C.H., Tien, Y.J., Tzeng, S. and Wu, H.M. (2004). Matrix visualization and information mining, *Proceedings in Computational Statistics 2004 (Compstat 2004)*, 85 – 100, Physika Verlag, Heidelberg.
11. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A,* 95:14863 – 14868.
12. R.A. Fisher (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7: 179 – 188.
13. Friendly M. (2002). Corrgrams: exploratory displays for correlation matrices. *The American Statistician*, 56(4):316-324.
14. Friendly, M. and Kwan, E. (2003). Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4): 509 – 539.
15. Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, 78:123 - V 129.
16. Hurley, C.B. (2004). Clustering visualization of multidimensional data, *Journal of Computational and Graphics Statistics*, 13: 788 – 806.
17. Inselberg, A. (1985), The plane with parallel coordinates, *The Visual Computer*, 1:69 – 91.
18. Lenstra, J. (1974). Clustering a data array and the traveling salesman problem, *Operations Research*, 22:413 – 414.
19. Ling, R.L. (1973). A computer generated aid for cluster analysis, *Communications of the ACM*, 16(6):355 – 361.
20. Marc, P., Devaux, F., and Jacq, C. (2001). yMGV: a database for visualization and data mining of published genome-wide yeast expression data, *Nucleic Acids Research*, 29(13):e63.
21. Marchette, D. J. and Solka, J. L. (2003). Using data images for outlier detection, *Computational Statistics and Data Analysis,* 43, 541 – 552.
22. Michailidis, G. and de Leeuw, J. (1998). The Gifi system for descriptive multivariate analysis, *Statistical Science*, 13:307 – 336.
23. Minnotte, M. and West, W. (1998). The data image: a tool for exploring high dimensional data sets, in *Proceedings of the ASA Section on Statistical Graphics,* Dallas, Texas, 25 – 33.
24. Murdoch, D.J. and Chow, E.D. (1996). A graphical display of large correlation matrices, *The American Statistician*, 50:178 – 180.
25. Robinson, W.S. (1951). A method for chronologically ordering archaeological deposits, *American Antiquity* 16: 293 – 301.
26. Slagle, J.R., Chang, C.L. and Heller, S.R. (1975). A clustering and data-reorganizing algorithm, *IEEE Trans. Syst. Man Cybern*, 5:125-128.
27. Streng, R. (1991). Classification and seriation by iterative reordering of a data matrix. In Classification, Data Analysis, and Knowledge Organization: Models and Methods with Applications (EditedbyH.H.Bock and P. Ihm), 121-130. Springer-Verlag, NewYork.

28. Tenenbaum, J.B., de Silva, V. and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science*, 290(5500):2319 – 2323.
29. Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., Brown, P. (1999). Clustering methods for the analysis of DNA microarray data. Technical Report, Stanford University, Oct. 1999.
30. Tien, Y.J., Lee, Y.S, Wu, H.M. and Chen, C.H. (2006). Integration of clustering and visualization methods for simultaneously identifying coherent local clusters with smooth global patterns in gene expression profiles. Technical Report 2006-11, Institute of Statistical Science, Academia, Taiwan.
31. Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.
32. Unwin, A. R, Hawkins, G., Hofmann, H., and Siegl, B. (1996). Interactive graphics for data sets with missing values - MANET, *Journal of Computational and Graphical Statistics* **5**, 113 – 122.
33. Unwin, A. R, Hofmann, H. (1998). New interactive graphics tools for exploratory analysis of spatial data. *In Innovations in GIS 5, ed. S Carver,* pp. 46 – 55. London: Taylor & Francis.
34. Wegman, E.J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association.* 85:664 – 675.
35. Wu, H.M. and Chen, C.H. (2005). Covariate adjusted matrix visualization. Technical Report. Institute of Statistical Science, Academia, Taiwan.