VizStruct: exploratory visualization for gene expression profiling

Li Zhang\textsuperscript{1,*}, Aidong Zhang\textsuperscript{1} and Murali Ramanathan\textsuperscript{2}

\textsuperscript{1}Department of Computer Science and Engineering and \textsuperscript{2}Department of Pharmaceutical Sciences, State University of New York at Buffalo, Buffalo, NY 14260, USA

Received on December 2, 2002; revised on March 25, 2003; accepted on June 24, 2003

ABSTRACT

Motivation: DNA arrays provide a broad snapshot of the state of the cell by measuring the expression levels of thousands of genes simultaneously. Visualization techniques can enable the exploration and detection of patterns and relationships in a complex data set by presenting the data in a graphical format in which the key characteristics become more apparent. The dimensionality and size of array data sets however present significant challenges to visualization. The purpose of this study is to present an interactive approach for visualizing variations in gene expression profiles and to assess its usefulness for classifying samples.

Results: The first Fourier harmonic projection was used to map multi-dimensional gene expression data to two dimensions in an implementation called VizStruct. The visualization method was tested using the differentially expressed genes identified in eight separate gene expression data sets. The samples were classified using the oblique decision tree (OC1) algorithm to provide a procedure for visualization-driven classification. The classifiers were evaluated by the holdout and the cross-validation techniques. The proposed method was found to achieve high accuracy.

Availability: Detailed mathematical derivation of all mapping properties as well as figures in color can be found as supplementary on the web page http://www.cse.buffalo.edu/DBGROUP/bioinformatics/supplementary/vizstruct. All programs were written in Java and Matlab and software code is available by request from the first author.

Contact: lizhang@cse.buffalo.edu

INTRODUCTION

DNA arrays allow simultaneous measurements of the expression levels of thousands of genes to be made from a single sample. The method has been used to address a range of biomedical research problems including the identification of malignant tissues, the delineation of drug responses, and the discovery of disease subtypes (Alizadeh et al., 2000; Alon et al., 1999; Golub et al., 1999; Hedenfalk et al., 2001; Khan et al., 2001; Virtanerva et al., 2001).

Visualization can facilitate the discovery of structures, features, patterns and relationships in data and exploratory visualization is an important component in hypothesis generation. The most prominent visualization-enhanced analysis tool for gene expression data is TreeView (Eisen et al., 1998), which provides a user-friendly computational and graphical environment for assessing the results from hierarchical clustering. The graphical presentations from TreeView include a dendrogram to reflect the distance relationships between clusters and a heat plot to convey gene expression changes visually between samples. The heat plot can be viewed as a variation of the parallel co-ordinates plot in which color is used to convey dimension values.

Multi-dimensional scaling (MDS) is an important class of visualization techniques that use optimization to preserve distance relationships between points in the multi-dimensional space in the two-dimensional mapping required for effective visualization. Ewing and Cherry (2001) used Sammon’s mapping, a variant of MDS, as a visualization approach for gene expression data sets. More recently, Hautaniemi et al. (2003) have presented a heat map-based strategy for visualizing the U-matrix from self-organizing maps (SOM).

Here, we present an alternative mapping for multi-dimensional data that is based on the first harmonic of the discrete Fourier transform (DFT). The mapping, which is analogous to radial co-ordinate visualization techniques such as Radviz (Hoffman et al., 1997), has interesting properties and preserves certain key characteristics of a variety of data sets, including those from gene expression profiling experiments. An interactive visualization tool, VizStruct, was implemented to perform visualization and sample classification.

THEORY, SYSTEM AND METHODS

Theory

The mapping converts multi-dimensional data into two-dimensions for visualization. An array-derived profile for $N$ genes results in a $N$-dimensional data point containing real valued numerical data. For mapping, we view each data point as a discrete-time real signal on $N$ evenly distributed time points (Cadzow and Landingham, 1985) and each $N$-point
signal (data point) is represented as an indexed sequence of \( N \) real numbers denoted by \( x[n] \). Each term of \( x[n] \) is denoted by \( x[n] \).

The two-dimensional mapping of a \( N \)-dimensional data point is defined by real and imaginary components of:

\[
\mathcal{F}_1(x[n]) = \sum_{n=0}^{N-1} x[n]W_N^n = \sum_{n=0}^{N-1} x[n]e^{-j2\pi n/N},
\]

where \( W_N = e^{-j2\pi/N} \) is called twiddle factor.

On closer inspection, the mapping of Equation (1) is seen to be identical to the first harmonic of the DFT (Morrison, 1994). The DFT is defined by:

\[
\mathcal{F}_k(x[n]) = \sum_{n=0}^{N-1} x[n]W_N^{nk}, \quad k = 0, 1, \ldots, N - 1.
\]

Each harmonic \( \mathcal{F}_k \) in the DFT is a measure of the \( k \)-th sinusoidal frequency component in the signal. Because the Fourier harmonics, in general, are complex numbers, they provide the two-dimensional point estimate for mapping a multi-dimensional signal. For this reason, we refer to the mapping as the first Fourier harmonic projection (FFHP). The computational complexity of the mapping is \( O(N \log N) \) using the fast Fourier transform (FFT) algorithm in Cooley and Tukey (1965). In this paper, all figures of FFHP visualization are two-dimensional scatter plot. The \( x \)-axis is labeled \( \text{Re}[F1] \) and represents the real part of the first Fourier harmonic and the \( y \)-axis is labeled \( \text{Im}[F1] \) because it represents the imaginary part of the first Fourier harmonic. The units for both axes are those of the input gene expression values.

Equivalently, the complex numbers representing the individual terms of FFHP in Equation (1) can be expressed in terms of magnitude \( r \) and phase \( \theta \) to provide a useful geometric interpretation of the mapping. For a data point with \( N \) dimensions, the complex exponential divides a unit circle centered at the origin of the complex plane into \( N \) equally spaced angles. The value of the first dimension is projected on the radial line corresponding to \( \theta = 0 \) and, similarly, the value of the \( k \)-th dimension is projected on to the radial line corresponding to \( \theta = 2\pi (1 - k)/N \) radians. The overall two-dimensional FFHP mapping is the complex sum of all \( N \) projections from a data point. Figure 1 illustrates the geometric interpretation for a point containing eight dimensions.

The FFHP has several useful properties that preserve the correlation between dimensions in the multi-dimensional data point. A detailed mathematical description of the properties is available as supplementary material.

**System**

The FFHP approach is implemented as a visualization tool called VizStruct, which is written in Java and Matlab. VizStruct is capable of mapping multi-dimensional data and can perform binary classification and multi-class classification.

**Methods**

**Data sets for visualization**

The approach was tested using Fisher’s iris data set and several gene-array data sets. The iris data set is widely used as a benchmark for classification and clustering (Fisher, 1936). The multiple sclerosis (MS) data set consists of array-derived gene expression profiles that were provided by our collaborators in the Department of
Fig. 2. (A)–(C) presents three snapshots from a dimension tour through a synthetic three-dimensional data set containing 25,000 points. The parameter settings for (A)–(C) were $\langle 0.5, 0.5, 0.5 \rangle$, $\langle 0.25, 0.5, 0.5 \rangle$, and $\langle 0.25, 0.25, 0.75 \rangle$, respectively.

Table 1. Sources of the published data sets used

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Size</th>
<th>Reference</th>
<th>URL from which data set was obtained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia-A</td>
<td>$7129 \times 72$</td>
<td>Golub et al. (1999)</td>
<td><a href="http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi">http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi</a></td>
</tr>
<tr>
<td>Lymphoma</td>
<td>$2984 \times 42$</td>
<td>Alizadeh et al. (2000)</td>
<td><a href="http://llmpp.nih.gov/lymphoma">http://llmpp.nih.gov/lymphoma</a></td>
</tr>
<tr>
<td>Leukemia-B</td>
<td>$7129 \times 27$</td>
<td>Virtaneva et al. (2001)</td>
<td><a href="http://thinker.med.ohio-state.edu/projects/aml8/index.html">http://thinker.med.ohio-state.edu/projects/aml8/index.html</a></td>
</tr>
<tr>
<td>BRCA</td>
<td>$3226 \times 22$</td>
<td>Hedenfalk et al. (2001)</td>
<td><a href="http://www.nhgri.nih.gov/DIR/Microarray/NEJM_Supplement/">http://www.nhgri.nih.gov/DIR/Microarray/NEJM_Supplement/</a></td>
</tr>
<tr>
<td>Iris</td>
<td>$4 \times 150$</td>
<td>Fisher (1936)</td>
<td><a href="http://www.ics.uci.edu/~mlearn/MLRepository.html">http://www.ics.uci.edu/~mlearn/MLRepository.html</a></td>
</tr>
</tbody>
</table>

Pharmaceutical Sciences and in the Department of Neurology. The data set contains two pair-wise group comparisons of interest. The first data subset, ‘MS versus Controls’, contains array data from 15 MS patients and 15 age- and sex-matched controls while the second subset is referred to as ‘MS–IFN’ because it contains array data from 14 MS patients prior to and 24 h after interferon ($\text{IFN}_\beta$) treatment. The remaining gene-array data sets were obtained from the publicly available sources listed in Table 1.

Identification of informative genes Informative genes are genes whose expression levels are significantly different across different sample classes. Neighborhood analysis (Golub et al., 1999) was used to create subsets of informative genes for the MS versus Controls, MS–IFN and Leukemia-A data sets. The Significance Analysis of Microarrays (SAM) algorithm (Tusher et al., 2001) was used for identifying subsets of informative genes for the remaining data sets.

Classifier construction and evaluation The oblique decision tree approach, which constructs straight lines of arbitrary slope to separate classes of data, was used to construct classifiers in the visualizations (Murthy et al., 1993). The Oblique Classifier 1 (OC1) algorithm (Murthy et al., 1994), which combines hill climbing with two forms of randomization, was used to identify good oblique splits at each node.

The accuracy of the classifier was evaluated using the holdout and leave-one-out cross-validation techniques. The holdout technique was used for the larger data sets, and involves constructing the classifiers using a subset of training samples and assessing the accuracy using a mutually exclusive subset of the test samples. The leave-one-out method was used for the smaller data sets. The classifiers were constructed using all the data points except one, which was withheld and used for testing. The step of classifier construction and testing were repeated for all the data points in the data set.

The accuracy of the classifier predictions was assessed using overall error rate (the ratio of the number of errors to the total number of test samples) and the $\kappa$-coefficient. The $\kappa$-coefficient (Cohen, 1960) is defined as:

$$
\kappa = \frac{p_0 - p_c}{1 - p_c},
$$

where $p_0$ is the observed agreement ratio and $p_c$ is the chance agreement ratio.
where $p_0$ is the proportion of cases in which the raters agree and $p_c$ is the proportion of cases for which agreement is expected only by chance on the assumption that the marginal distributions of ratings are independent.

RESULTS

Feasibility assessment with the benchmark iris data set

The iris data set, which has been widely employed to assess discriminant and cluster analysis algorithms, was used as a benchmark data set to obtain a preliminary assessment of the usefulness of the FFHP approach prior to analysis of the more challenging, high-dimensional array data sets. The iris data set contains 150 observations of four attributes (sepal length and width, petal length and width) from three iris species (50 observations each of iris setosa, iris versicolor, iris virginica). Figure 3A and B show the data set for two-dimensional parameter settings. The three clusters are clearly apparent from the visualization and segregate by species. The iris setosa cluster (plus symbols) is clearly separated from the cluster formed by iris virginica (circles) and iris versicolor (stars). The separation between virginica and versicolor improves in Figure 3B compared with Figure 3A.

These findings demonstrated the feasibility of using the FFHP mapping for array-derived gene expression profiles.

Visualization of two-class array-derived data sets

Figure 4A presents the visualization results from the MS–IFN data set, which contains gene expression profiles from 14 MS patients before and 24 h after IFN treatment while Figure 4B presents the results from the MS versus Controls data set which compares 15 MS patients to age- and sex-matched controls. Subsets of 88 informative genes were identified from the 4132 genes in each data set. The informative genes were mapped using VizStruct; oblique classifiers then were constructed and evaluated using the leave-one-out method. The pre-treatment group (plus symbols) in Figure 4A segregated to upper half of the visualization field while the post-treatment group (circle) occupied the lower half. The results for MS patients (plus symbols) and controls (circles) in Figure 4B were generally similar. The pre-treatment and post-treatment groups in Figure 4A were clearly separated while the separation between MS patients and controls in Figure 4B less distinct. Table 2 shows that the accuracy of the prediction was 100% (the $\kappa$-coefficient was 1.00) for the MS–IFN data set and 90% (the $\kappa$-coefficient was 0.80) for the MS versus Controls data set.

The leukemia-A data set (Golub et al., 1999) consists of expression values of 7129 genes from 72 patients. Fifty informative genes were identified using the neighborhood analysis approach of the original report. The training set consists of 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML) patients’ samples (plus symbols and circles in Fig. 4C). An oblique classifier was constructed based on visualization of the training set, and the holdout method was used for classifier evaluation. The testing set contained 20 ALL and 14 AML samples (triangles and stars in Fig. 4C). The overall accuracy of the classifier was 88% and the $\kappa$-coefficient was 0.75.

The visualization of colon cancer data set (Alon et al., 1999) consists of gene expression profiles for 2000 genes.
Fig. 4. Two-class classifications. (A) shows the results from mapping the MS–IFN data set. The pre-treatment samples are shown in plus symbols, the post-treatment samples are in circles and a test sample is shown in a filled circle highlighted with arrow. (B) shows the results from mapping the MS versus Control data set. The MS patient samples are shown in plus symbols, the post-treatment samples are in circles, and a test sample is shown in a filled circle highlighted with arrow. On all panels, the solid dark line is the oblique classifier used. There were two misclassified samples in (B), which occurred during training and were not counted toward classifier evaluation. (C) shows the results from mapping the leukemia-A data set. Test samples are overlaid with training samples. The plus symbols are ALL samples and the circles are AML during the training. The triangles and the stars are test ALL and test AML samples, respectively. (D) shows the results from mapping the colon cancer data set. The plus symbols are normal colon tissues and the circles are cancerous colon tissues. The test cancerous sample is shown in the filled circle highlighted by the arrow; it was misclassified. (E) shows the results from mapping the lymphoma data set. The germinal center (GC) lymphoma samples are in plus symbols and the activated B cell-like lymphomas are in circles. The test activated B cell-like lymphoma sample is shown in the filled circle highlighted by the arrow; it was correctly classified.

Table 2. Summary of results from sample classifications

<table>
<thead>
<tr>
<th>Data set</th>
<th>Classes</th>
<th>Genes in classifier</th>
<th>Evaluation</th>
<th>Training set size</th>
<th>Testing set size</th>
<th>Errors count</th>
<th>Accuracy (%)</th>
<th>( \kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS–IFN</td>
<td>2</td>
<td>88</td>
<td>Leave-one-out</td>
<td>28</td>
<td>28</td>
<td>0</td>
<td>100</td>
<td>1.00</td>
</tr>
<tr>
<td>MS versus Control</td>
<td>2</td>
<td>88</td>
<td>Leave-one-out</td>
<td>30</td>
<td>30</td>
<td>3</td>
<td>90</td>
<td>0.80</td>
</tr>
<tr>
<td>Leukemia-A</td>
<td>2</td>
<td>50</td>
<td>Holdout</td>
<td>38</td>
<td>34</td>
<td>4</td>
<td>88</td>
<td>0.75</td>
</tr>
<tr>
<td>Colon</td>
<td>2</td>
<td>50</td>
<td>Leave-one-out</td>
<td>62</td>
<td>62</td>
<td>7</td>
<td>89</td>
<td>0.75</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>2</td>
<td>100</td>
<td>Leave-one-out</td>
<td>42</td>
<td>42</td>
<td>0</td>
<td>100</td>
<td>1.00</td>
</tr>
<tr>
<td>Leukemia-B</td>
<td>3</td>
<td>50</td>
<td>Leave-one-out</td>
<td>27</td>
<td>27</td>
<td>8</td>
<td>70</td>
<td>0.55</td>
</tr>
<tr>
<td>BRCA</td>
<td>3</td>
<td>50</td>
<td>Leave-one-out</td>
<td>22</td>
<td>22</td>
<td>0</td>
<td>100</td>
<td>1.00</td>
</tr>
<tr>
<td>SRBCT</td>
<td>4</td>
<td>100</td>
<td>Holdout</td>
<td>63</td>
<td>19</td>
<td>1</td>
<td>95</td>
<td>0.93</td>
</tr>
</tbody>
</table>

from 22 normal samples and 39 cancerous colon tissues. The result from mapping 50 informative genes is summarized in Figure 4D. The upper part of the visualization field was dominated by normal colon tissue while cancerous colon tissues occupied the lower half. The overall accuracy was 89% and the \( \kappa \)-coefficient was 0.75.

Figure 4E shows the results from mapping the 100 informative gene subset of the lymphoma data set (Alizadeh et al., 2000). The germinal center lymphomas were relatively well separated from the activated B cell-like lymphomas. The robustness of the separation contributed to 100% (the \( \kappa \)-coefficient was 1.00) accuracy upon classification.
Multi-class classifications. (A) shows the results from mapping the leukemia-B data set. The plus symbols are samples from AML patients with trisomy 8 patients (AML+8), the circles are samples of AML patients with normal cytogenetics (AML–CN) and the stars are CD34+ cells from controls. There were three training errors (1 AML–CN and 2 AML+8) that were not counted toward classifier evaluation. (B) shows the mapping of the BRCA data set (plus symbols: patients with BRCA1 mutations; circles: patients with BRCA2 mutations; stars: sporadic cases). The filled circle highlighted with the arrows in (A) and (B) are a representative test sample. (C) shows the results of mapping the informative genes from the SRBCT data set. The triangles represent neuroblastomas (NB), the stars are rhabdomyosarcomas (RMS), the plus symbols are Ewing’s family of tumors (EWS) and the cross symbols are Burkitt’s lymphomas (BL). The symbols represent the training set and for the corresponding test samples four sets of symbols were used: filled circle, diamond, square and unfilled circle. In all panels, the solid dark lines are the classifiers.

Visualization of multi-class array-derived data sets
The performance of VizStruct for multi-class array-derived data sets was examined in the next series of experiments. The leukemia-B data set (Virtaneva et al., 2001) contains three classes and a 50-gene informative subset was mapped using VizStruct (Fig. 5A). The normal CD34+ samples segregated to the northwestern region of the visualization and were well separated from the AML samples. The samples from cytogenetically normal AML (AML–CN) patients and trisomy 8 AML (AML+8) patients segregated, albeit less distinctly. This was reflected in the lower accuracy of 70% and the \( \kappa \)-coefficient was 0.55.

The FFHP mapping of 50-gene informative subset of the BRCA data set (Hedenfalk et al., 2001) resulted in a clear separation of the BRCA1, BRCA2 and sporadic classes (Fig. 5B) and the accuracy was 100% (the \( \kappa \)-coefficient was 1.00).

For the four-class SRBCT data set (Khan et al., 2001), a 100-gene informative subset was mapped (Fig. 5C). The four classes of small round blue cell tumors segregated to different regions of the visualization field. The accuracy of the oblique classifiers was of 95% and the \( \kappa \)-coefficient was 0.93 (Table 2).

As a rule of thumb, \( \kappa \) values of 0.41–0.60 indicate moderate agreement, values in the range 0.61–0.80 indicate substantial agreement and values in the range 0.81–1.0 indicate almost perfect agreement (Landis and Koch, 1977). The summary in Table 2 indicates that for MS–IFN, MS–Control, lymphoma, BRCA and SRBCT data sets, almost perfect agreement was obtained using VizStruct. For the leukemia-A and colon cancer data sets substantial agreement was obtained. However, only moderate agreement was achieved in for the Leukemia-B data set. Taken together, these experimental results highlight the effectiveness of the FFHP mapping for visualization-driven sample space classification of both binary and multi-class gene-array data sets.

Comparison of FFHP to Sammon’s mapping
Multi-dimensional scaling is a competing approach for visualizing multi-dimensional data. We compared FFHP to Sammon’s mapping (Sammon, 1969), variant of MDS that optimizes the following stress function \( \mathcal{E} \):

\[
\mathcal{E} = \frac{1}{\sum_i \sum_{j<i} d_{ij}^*} \sum_i \sum_{j<i} \left( \frac{d_{ij}^* - d_{ij}}{d_{ij}^*} \right)^2.
\]

where \( d_{ij}^* \) is the distance between points \( i \) and \( j \) in the \( N \)-dimensional space and \( d_{ij} \) is the distance between \( i \) and \( j \) in the visualization.

Figure 6A shows the Sammon’s mapping of the iris data set. A comparison of Figure 6A with Figure 3A reveals the extensive similarities between FFHP and Sammon’s mapping. For example, the iris setosa cluster is clearly separated from the virginica and the versicolor cluster and the versicolor cluster is closer to setosa than the virginia. The relative locations of individual samples are also remarkably similar (data not shown).

Figure 6B is the Sammon’s mapping of the SRBCT data set. A comparison of Figure 6B with Figure 5C reveals extensive similarities; e.g. the relative locations of various tumor clusters are quite similar and the Burkitt’s lymphoma cluster is distal from the rhabdomyosarcomas.
Fig. 6. Sammon’s mapping. (A) shows the results from Sammon’s mapping of the iris data set (iris setosa: plus symbols; iris versicolor: circles; and iris virginica: stars). (B) shows results from Sammon’s mapping of the informative genes from the SRBCT data set. The triangles represent neuroblastomas (NB), the stars are rhabdomyosarcomas (RMS), the plus symbols are Ewing’s family of tumors (EWS) and the cross symbols are Burkitt’s lymphomas (BL). The symbols represent both the training set and the test data set.

Although, the FFHP uses an approach to mapping multi-dimensional data that is distinctively different from Sammon’s mapping, it yields results that are consistently comparable.

DISCUSSION

In this paper, we have explored the use of the FFHP mapping for visualizing and classifying multi-dimensional array data sets. The accuracy, which ranges from 70 to 100% with a $\kappa$-coefficient range of 0.75–1.0, demonstrated its promise for sample classification in gene-array data sets. The FFHP mapping results in two-dimensional visualizations that are identical to those of radial co-ordinate visualization techniques, e.g. Radviz (Hoffman et al., 1997). However, in place of the vector notation and the spring paradigm of Radviz, we have used a complex number notation. This substantive reformulation of the mapping provides valuable theoretical insights and allows several important properties of the mapping, including its relationship to the DFT, to be easily derived.

We assessed the robustness of the visualisation approach by assessing its sample classification accuracy in the face of large perturbations in input data. The leukemia-A and SR–BCT data sets were perturbed by switching the training and test inputs. The accuracy for the leukemia-A data set with switching was 88% and it remained the same as that in Table 2. For the SRBCT data set, the accuracy after switching was 73%; we attribute the reduced accuracy to the small training sample size ($n = 3$) in the Burkitt’s lymphomas (BL) group. Using this operational approach, it became apparent that the robustness of the approach is dependent on several factors, including the intrinsic separation of the classes in the input array data, the sample size of the training set and the classifier algorithm. Generally, data sets with large relative separation between classes—large inter-class effect sizes—are least sensitive to perturbations of various types, e.g. changes in the position and angles of classifier lines, or small decreases in the training set size. Another factor is the training sample size: the statistical uncertainty with small samples sizes is large and accuracy diminishes once the classifiers overlap the prediction interval of the data. When the training sets become larger, the lines and their splits can be more accurately positioned to separate the classes. Classifiers of arbitrary shape may be more robust compared with the oblique classifiers when the distribution of points in the various classes in the two-dimensional mapping is not defined by a polygonal space partitioning. In a separate report (Zhang et al., 2003), we compared oblique classifiers to SOM (Kohonen, 1995): In colon cancer data set, the oblique classifier misclassified seven samples (Table 2) while a $4 \times 3$ unit SOM misclassified eight samples. Furthermore, the same seven samples misclassified by the oblique classifier were also misclassified by the SOM. The VizStruct approach readily admits more complex clustering and classification strategies but oblique classifiers appear effective and parsimonious and their simplicity may appeal to end-users in the life sciences.

Parallel co-ordinates and MDS represent competing approaches to FFHP for visualizing multi-dimensional data sets. Parallel co-ordinates visualization has obvious drawbacks: it becomes increasingly unreadable when the data
size gets larger. The FFHP mapping was shown to yield results that were similar to those from Sammon’s mapping, a variant of MDS. Despite providing results that are mathematically optimal in some sense, MDS is not ideal for gene-array data because: (1) it provides a single final result and the user cannot intervene interactively during visualization and (2) the incremental addition of any single point requires a complete repetition of the optimization procedure and possibly extensive reorganization of all the previously mapped points to new locations. Secondarily, from a computational complexity standpoint, the parallel co-ordinates method requires relatively little computational effort because each dimension is plotted directly, while MDS requires time-consuming optimization procedures of time complexity $O(N^2)$ or greater. The computational complexity of FFHP ($N \log N$) is intermediate.

The FFHP is sensitive to dimension order and VizStruct has the capability to reorder dimensions so that class separation is enhanced. To determine whether the canonical reordering could cause misleading pseudo-classes to appear during visualization, a random data set containing 50 data points each with 100 dimensions was simulated; 20 points were arbitrarily assigned to one class and the remainder to another class followed by canonical dimension reordering. The visualization of the reordered data set did not suggest the presence of pseudo-classes.

The performance of VizStruct suggests that visualization may be capable of a larger and richer role than is currently appreciated. However, with multi-dimensional gene expression data, one also has to be always mindful of the curse of dimensionality (Bellman, 1961) and rigorously confirm any experimental findings in the two-dimensional mapping with appropriate quantitative techniques.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Science Foundation and the grants RG3258A2 from the National Multiple Sclerosis Society.

REFERENCES


