VistaClara: An Interactive Visualization for Exploratory Analysis of DNA Microarrays

Robert Kincaid
Agilent Laboratories
3500 Deer Creek Road MS26U-16
Palo Alto, CA 94304
robert_kincaid@agilent.com

ABSTRACT
We have created VistaClara to explore the effectiveness of applying an extended permutation matrix to the task of exploratory data analysis of multi-experiment microarray studies. The permutation matrix is a visualization technique for interactive exploratory analysis of tabular data that permits both row and column rearrangement, and fits well with the tabular forms of data characteristic of gene expression studies. However, this technique has been largely overlooked by current bioinformatics research. Our implementation supports direct incorporation of supplemental data and annotations into the matrix view. This enables visually searching for patterns in gene expression measurements that correlate with other types of relevant data (disease classes, clinical, histological, drug treatments, etc.). The heatmap visualization common in microarray analysis is extended to provide a novel alternative using size as well as color to graphically represent experimental values, thus allowing more effective qualitative comparisons. Methods to sort rows or columns by similarity extend the possible permutation operations, and allow more efficient searching for biologically relevant patterns in very large data sets. Based on overview+detail principles, a dynamic compressed heatmap view of the entire data set provides the user with overall context, including possible correlations not currently visible in the more detailed view. Combined, these techniques make it possible to perform highly interactive ad hoc visual explorations of microarray.

Keywords
microarray, information visualization, user interface, permutation matrix, reorderable matrix, overview+detail, bioinformatics, gene expression.

1. INTRODUCTION
An area of interest in information visualization is interactively manipulating matrix-organized data via spreadsheet-like applications. The goal of such software is to provide interactive mechanisms that enable visual pattern discernment. In analogy to woodworking, Rao refers to this as looking for the “grain” in information [15]. Unlike the more rigorous computational approaches typically used in bioinformatics, this form of visual data mining utilizes the highly developed pattern recognition abilities of human visual perception. The primary aim of this approach is not a static visualization of known results, but rather the highly interactive discovery of previously unknown correlations, outliers, or other areas of interest. The inherent domain knowledge of the user can be leveraged to provide further insight and meaning to the emerging patterns, and direct the path of investigation. Once interesting features of the data are observed, it may be useful or even necessary to verify the result more rigorously by traditional computational techniques.

Multi-experiment microarray data is frequently analyzed in tabular form, thus lending itself to matrix oriented information visualization. This is particularly true of recent experiments that search for insight into cancer and other diseases. Typically many samples are measured using individual microarray experiments, and a resulting matrix of gene vs. experiment is constructed. The problem then becomes one of finding those genes that have high correlation to the disease or disease classes being studied. Considerable efforts have been made to find computational techniques for classification or clustering of such data sets [20]. However, viewing such matrices is generally relegated to generic spreadsheet applications and static visualizations.

Additionally, user studies [13] reveal that molecular biologists often tend to distrust complex computational methods as being opaque and hard to interpret without being an expert in computational biology. The study also found that early-stage ad hoc exploratory analysis of microarray data is important and does take place. Such exploratory analysis often involves examining outliers and statistical anomalies that might be eliminated by more rigorous statistics-based methods.

VistaClara provides a highly interactive environment that allows a typical molecular biologist to intuitively sift through their data to find patterns that are meaningful based on their own domain knowledge, experience and intuition. This provides insight into the underlying correlations between genes, samples and experiments, and may suggest further, more rigorous, confirmatory calculations.

We have created VistaClara principally to explore the effectiveness of this approach to microarray analysis. We have not tried to create a complete microarray analysis environment and have instead concentrated on the research questions around visual exploratory data analysis.

In order to leverage existing user experience and practice, VistaClara takes as a starting point the traditional heatmap visualization commonly used to display gene expression data, and
extends this to a fully interactive permutation matrix supporting both column and row rearrangement. This is an important aspect for analyzing microarray data since correlations are expected to occur between both groups of genes as well as groups of samples. While the permutation matrix has been implemented generically (e.g. VisuLab [18], TableLens [16], Siirtola [19]), no previous implementation has been specifically designed for the unique characteristics of multi-experiment analysis of microarray data. Jacques Bertin initially described the concept of permutation matrices as a visualization [1], and realized that meaningful permutations become difficult with very large data sets. VistaClara implements a novel collection of features specifically designed to facilitate the interactive manipulation of the large data sets typical of microarray experiments. Extended in this way, the permutation matrix can be a useful tool in analyzing such data.

By integrating this additional data directly into the visualization, when rows or columns of the matrix are reordered, meaningful correlations between the supplemental and gene expression data can be easily observed. For example, expression patterns may correlate with invasive ability, or chromosome location, etc. To further aid the visual discovery of such correlations, binary data types (yes/no, +/-, male/female, etc.) are encoded with either pale red for “positive” values or pale green for “negative” values. In this way, the user can visually correlate red/green expression encodings with their supplemental counterparts.

2.2 Permutation Matrix
As first described by Bertin [1, 2], the key feature of a Permutation Matrix is that both rows and columns are reorderable. The intent is to allow manipulating rows and/or columns in appropriate ways to visually reveal correlated patterns in the data. Siirtola provides a recent concise discussion of permutation matrices [19] along with some usability data.

VistaClara starts with a basic permutation matrix by allowing both row and column sorting. However, in order to support including supplemental gene and sample annotations, this sorting is restricted as follows:

1. If a column is being sorted, only those rows corresponding to microarray data are reordered.
2. If a row is being sorted, only those columns corresponding to microarray data are reordered.

A simple way to understand this process is to consider the supplemental data as extensions of the corresponding row or column headers. This permits maintaining alignment between the microarray data and corresponding annotations, while still allowing sorting by relevant annotations. For example, in the melanoma data shown in Figure 1, one can sort gene expression data by chromosome location, and/or one can sort by invasive ability.

2.3 Ink blobs
When looking for correlations in both the horizontal and vertical directions, it is useful to graphically represent data in a way that is not directionally biased. Heatmap views have this property (e.g. Figure 1) and can be useful in some situations. However, there is good evidence that intensity of color alone is difficult to resolve visually, and not effective for discriminating small numerical differences. Bertin and others have advocated size as a preferred, more visually comparable representation. Following similar representations found in Bertin’s work, we optionally replace the typical heatmap with filled circles, whose diameters are in proportion to the represented values. Further, if the diameter goes beyond a critical threshold, the square in which the circle appears is filled. This change in shape is an additional highly visible cue that the data has exceeded this threshold.

Standard ink blob representations are generally monochrome with no way of indicating the sign of the data (positive or negative). To overcome this we simply compute the diameter to correspond to the absolute value of the underlying log-transformed ratio data, and then apply the same red/green color gradients as used in the heatmap to indicate sign. An adjustable scale factor is applied to the data when calculating ink blob diameters so that the diameter and threshold can be optimally scaled. Typically a user would set...
this threshold to cause all cells beyond some specific fold increase/decrease to “saturate”. This makes significant fold increases/decreases readily apparent.

Following a previous method [11], we calculate the color intensity \( i \) for a value \( x \) using the sigmoid function:

\[
i = \frac{2}{1 + e^{-sx}} - 1
\]

where \( s \) is used to specify the steepness of the color gradient and can be adjusted by the user to optimize the gradient for the range and distribution of data being viewed. This function ranges from \(-1\) to \(1\) and is multiplied by 255 to obtain an appropriate RGB color value (sign is used to distinguish red versus green).

This combined ink blob heatmap (Figures 3-6) provides better value comparisons, while retaining the signed nature of the heatmap and the color-coding familiar to researchers working with microarrays. Figure 2 shows specific examples of how ink blobs provide superior visual discrimination of the underlying values. In the leftmost heatmap view it is difficult to determine if the absolute magnitude of the green cell C6 is less or greater than red cells like A3. In the rightmost ink blob view the difference is much more apparent. Since the ink blob threshold is set to fill cells at +/- four-fold ratios, it is immediately obvious that A3, A5, B3, C1 and E1 are all up-regulated four-fold or greater, and there are no four-fold down-regulated values. Also, B4 and B3 are distinguishable in the ink blob view, but not in the heatmap view. Finally, while B6, C3 and D3 are all clearly shown down-regulated in the heatmap view, the ink blob view reveals that these ratios are significantly smaller than might appear from just the color alone.

### 2.4 Similarity Sorting

Navigation of large data spaces is problematic and can be greatly improved via computationally assisted navigation [6]. A very simple form of this is row or column sorting, which raises highly regulated genes to the top or bottom of the matrix display. However, these correlations are usually confounded with various noise contributions to the data, which often make the standard simple row and column sorts ineffective as permutation operations. With this in mind, VistaClara implements an intuitive extension of simple column and row sorting. We allow sorting rows (or columns) using measures of similarity between the entire rows (or columns) of microarray data. This is particularly advantageous for microarray data, since we typically expect to find such correlations between samples (columns) as well as between genes (rows). Since we do not rely on a single experiment or gene to determine the ordering, the result is more robust to experimental error, and can also collect patterns of expression rather than just those rows or columns of data that exhibit similar ordering.

Which genes exhibit similar expression patterns across many experiments, or which experiments have similar behavior, are common and fundamental questions asked by typical microarray users. Similarity sorting provides a natural and intuitive mechanism to explore such questions in a rapid ad hoc manner.

Similarity sorting of rows is performed as follows:

1. A row of interest is chosen. For the case of gene expression data, we might choose a row because the pattern of expression across the set of microarrays can be visually seen to correlate with the supplemental sample information.
2. This row then becomes the first row of microarray data.
3. A distance measure is computed between the chosen row of interest and all other rows.
4. The remaining rows are then sorted in ascending order by distance (increasing dissimilarity).

VistaClara currently implements two methods of calculating similarity. For two vectors \( X \) and \( Y \), we can compute the Euclidian distance \( D \) as:

\[
D(X, Y) = (\sum (X_i - Y_i)^2)^{1/2}
\]

Since rows are simply being ordered, computing the square root of the formula above is unnecessary, and is omitted to improve performance slightly.

We can compute an alternative distance based on the Pearson correlation coefficient using the following computational formula:

\[
r = \frac{\sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{\sqrt{\left( \sum_{i=1}^{N} X_i^2 - \frac{(\sum_{i=1}^{N} X_i)^2}{N} \right) \left( \sum_{i=1}^{N} Y_i^2 - \frac{(\sum_{i=1}^{N} Y_i)^2}{N} \right)}}
\]

Distance is measured as 1-r. Pearson coefficients are often used in similar manner for cluster analysis of microarray data [8].

Similarity is only computed between vectors of microarray data (supplemental data is not included). Currently, only ratio measurements are handled (vs. single channel microarray experiments). Ratios are always log-transformed so that distance measurements will not be unduly biased toward high ratios.

The Euclidian distance is often better at finding rows closely similar in overall amplitude, but does a poor job at separating correlated and anti-correlated rows to opposite ends of the matrix. Calculating the Pearson coefficient does a better job of sorting and separating correlated and anti-correlated rows, but similarity is weighted more toward the overall pattern or shape of the expression profile rather than amplitude. Which similarity measure to use depends on what type of correlation is being sought, and can be selected by the user. For microarray data, this process often corresponds to choosing a particular gene of interest and ordering the rows based on their similarity of expression across the multiple samples.
The distance calculations and subsequent sort compute rapidly enough that this manipulation can be included as a component of the data browsing in a manner similar to, but more powerful than, simple row sorting by a single column of expression ratios. A similarity sort merely requires the additional n-1 calculations to compute the array of distance measures provided to the underlying mergesort algorithm. Since the number of rows (genes) will generally far exceed the number of columns (experiments), the time complexity of similarity sorting rows is dominated by the O(n log n) complexity of the underlying mergesort for large n.

Table 1. Typical execution times for various methods of row sorting

<table>
<thead>
<tr>
<th>DATA</th>
<th>ROWS</th>
<th>COLS</th>
<th>EXECUTION TIME (SECS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>SIMPLE</td>
</tr>
<tr>
<td>Melanoma</td>
<td>8067</td>
<td>31</td>
<td>0.06</td>
</tr>
<tr>
<td>Prostate</td>
<td>6500</td>
<td>26</td>
<td>0.05</td>
</tr>
<tr>
<td>Multi-Array</td>
<td>23353</td>
<td>11</td>
<td>1.93</td>
</tr>
</tbody>
</table>

Table 1 presents some typical execution times for row sorting, including results from an extremely large data set. The melanoma and prostate data are the same shown in Figures 3, 4 and Figures 5, 6 respectively. The Multi-Array data is a proprietary set of candidate microarray probes printed on a set of arrays and measured against a number of different validation samples. In all cases, rows represent genes and columns represent samples. All timings were taken on a system with a 1GHz processor, 512Mb RAM, and minimal extraneous processes.

While more computationally intensive than simple sorts, similarity sorting is still fast and scalable enough to be effective as an interactive manipulation. Further, the complexity analysis and empirical timings confirm that the difference between simple sorting and similarity sorting becomes less a factor as n becomes larger, implying that there are no significant scalability issues inherent in similarity sorting procedure.

Similarity sorting is also permitted between columns of microarray data using an analogous computation. This results in ordering the experiments by similarity to the chosen experiment of interest.

GeneSpring [9], BioConductor [3] and dChip [7], allow creating lists of genes with similar expression patterns and use computational measures similar to those described here. However, they do not provide this function as a highly interactive sorting mechanism useful for exploratory data analysis of entire microarray data collections. Further, the typical use case for these systems is to create a separate and relatively small subset of such similar genes for further study, vs. interactively manipulating an entire microarray study as in VistaClara.

2.5 Overview+Detail

To provide global context for the more detailed table view, we employ a style of context management referred to as Overview+detail [5, 14], which provides coordinated views of data. An overview is typically used for orientation and navigation and shows the entire set of data, while the detail view displays only a portion of the data, but at a higher level of detail. In VistaClara, the overview takes on the additional role of displaying trends and correlations that are not visible in the detail view, but are the result of column and row rearrangements of the permutation matrix.

As indicated in Figure 1, VistaClara includes an overview display of the entire data set in the form of a dynamic heatmap. As rows and columns are rearranged, the overview is updated to reflect the change and any emerging correlations that might be visible beyond the range of the detail view. A transparent cursor rectangle in the overview outlines the position and range of the detail view as the user scrolls the display, and provides further context and navigation orientation for the user. This also exposes the striking observation of how small a slice of the total microarray data is generally viewed in standard tabular spreadsheet visualizations. The cursor is difficult to make out in most of the printed figures, but Figure 1 shows an enlarged view of this feature. The cursor can be seen more clearly in the operating software.

3. RESULTS

To test and demonstrate VistaClara we examined the gene expression data from two previous studies with known results [4, 12] in order to validate that VistaClara can find similar correlations. Our intent is not to reproduce exactly the more rigorous results. Instead, we wish to show that making reasonable and simple assumptions about what genes should be relevant, VistaClara manipulations reveal a qualitatively similar result.

In this way we mimic the thought process of a typical molecular biologist. Also, a user will typically explore the data interactively in search of previously unknown correlations or relationships in the data. The unstructured nature of a typical session of this type is difficult to convey in printed form, but these simple examples should at least demonstrate the potential such operations and visual pattern finding can have. It is important to realize that the final display of results shown in the figures is only the end result. It is the interactive process, driven by the domain knowledge of the user that leads to these results and is the core of VistaClara.

We first examined Bittner’s microarray data [4] consisting of 8067 cDNA measurements for each of 31 patient samples. Using computational techniques Bittner singled out 22 cDNA clones as being highly discriminating for one class of melanoma. We chose Melan-A as a gene of interest, as it is associated with melanoma and might be reasonably chosen in the absence of Bittner’s results by a knowledgeable investigator. Rows are sorted by similarity using Pearson coefficients as a distance measure.

Within the first set of rows similar to Melan-A (Figure 3) we find instances of all 11 of the 22 discriminating genes reported by Bittner which have expression profiles similar to Melan-A. Based on our distance measure, the most distant rows (Figure 4) consist of the most anti-correlated patterns relative to Melan-A. This list contains instances of 7 of the previously reported discriminating genes. Further, we visually find good overall correlation to the partitioning of melanoma patients found computationally by Bittner (row 5 in Figure 3). While no single gene perfectly separates the classes, the overall trend is quite apparent in the ink blob representation, and qualitatively similar to the static heatmap presented by Bittner.
Using only user interface manipulations, and visual pattern finding, we are able to reproduce qualitatively similar results to more exact computational methods.

Luo [12] studied gene expression differences between tissues representing Human prostate cancer and benign prostatic hyperplasia (BPH). We also analyzed this data set (25 patient samples, 6500 cDNA clones), including the provided supplemental data. Figures 5 and 6 show the result of the following operations:

1. Rows are sorted by similarity using AMACR as the gene of interest. AMACR is a well-known marker gene for prostate cancer [17] and so would make a suitable candidate to consider in this way. It also visually correlates well with the supplemental data identifying cancer and BPH samples.

2. Columns are sorted by similarity to sample 13. We chose this column since it had the highest percentage of cancer cells (see row 3 of Figure 5) and hence might be most representative of the general expression profile of prostate cancer.

The effectiveness of the similarity sorting procedure is again demonstrated, as well as the clarity of the visualization. Luo provided a list of 210 cDNA clones with statistically significant differentiation between cancer and BPH. Many of these are in fact found in the vicinity of AMACR in the sorted VistaClara list as shown in Figure 5. The anti-correlated end of the list (Figure 6) shows additional genes from the discriminating set. Similarity sorting with respect to hepsin (not shown) further improves agreement with Luo et. al., extracting more relevant genes from the total body of expression data.

Figure 3. Ink blob visualization of melanoma data showing genes ranked by similarity to Melan-A. Highlighted gene annotations indicate qualitative agreement with Bittner [7]. Red (darker colored ink blobs) indicates up-regulation.
It is important to stress that such visual manipulations are not meant to substitute for more formal methods. We merely wish to point out from these examples that data can be visually manipulated to reveal real, biologically meaningful correlations. Identifying such correlations in VistaClara is not meant to be definitive evidence of a biological finding, but to be suggestive of phenomena that warrant further investigation using more rigorous methods.

In both examples, a relatively small set of highly relevant correlations are extracted from thousands of individual expression profiles. This can be accomplished repeatedly, quickly and efficiently as part of an interactive data browsing activity, guided by the domain expertise and insight of the investigator. Interesting gene expression patterns can be easily compared visually to supplemental data, since all data is aligned and integrated into the same view.

We have also successfully used VistaClara for analysis of proprietary data of our own and that of research collaborators. A variety of experiment types yielded striking correlations via similarity sorting, including cancer studies, time courses, drug treatments, etc. All the features of VistaClara taken together (similarity sorting, ink blob views, overview+detail, etc.) provide an effective and powerful environment for exploratory analysis.

Two key features of VistaClara (similarity sorting and the ink blob representation) have been transferred to the commercial product Synapsia[21].

4. CONCLUSIONS AND FUTURE WORK

VistaClara can efficiently perform novel exploratory analysis of multi-experiment microarray data. It is possible to manipulate in their entirety, large heterogeneous data sets consisting of multiple microarray experiments and relevant supplemental annotations and data. These manipulations can be guided by the investigator's own insights and domain knowledge. This enables interactive visual searching for biologically meaningful patterns in the data. Testing with melanoma and prostate data show that it is possible to obtain qualitative insights via these interactive matrix permutations, and that these results are similar to more rigorous computational methods. Large data sets are effectively handled by showing a dynamic overview of the entire set. Similarity sorts provide computationally assisted methods to find relevant gene correlations in such large sets, using an intuitive and simple user interface.

Figure 4. Ink blob visualization of melanoma data, showing genes most anti-correlated to Melan-A. Highlighted gene annotations indicate qualitative agreement with Bittner [7]. Green (lighter colored ink blobs) indicates down-regulation.
interface. We have further verified these results through our own internal use, as well as that of research collaborators.

We are currently investigating various extensions, such as improved methods of similarity sorting and incorporating the supplemental data as well as error estimates in similarity measures.

While we have concentrated primarily on building user interfaces for manipulating gene expression data, data from other microarray systems such as comparative genomic hybridization as well as protein arrays, should be amenable to this kind of visual exploration. Further, extensions to handle non-ratio data, or data from experiments other than microarrays are also possible.

**REFERENCES**


Figure. 5. Prostate data ranked by similarity to AMACR and Sample 13. Highlighted gene annotations indicate qualitative agreement with Luo [16]. Green (lighter colored ink blobs) indicates up-regulation.

Figure. 6. Prostate data showing rows anti-correlated with AMACR. Highlighted gene annotations indicate qualitative agreement with Luo [16]. Red (darker colored ink blobs) indicates down-regulation.


