International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:6, 2009

Multidimensional Visualization Tools for Analysis of Expression Data

Urska Cvek, Marjan Trutschl, Randolph Stone II, Zanobia Syed, John L. Clifford, Anita L. Sabichi

***Abstract*—**Expression data analysis is based mostly on the statistical approaches that are indispensable for the study of biological systems. Large amounts of multidimensional data resulting from the high-throughput technologies are not completely served by biostatistical techniques and are usually complemented with visual, knowledge discovery and other computational tools. In many cases, in biological systems we only speculate on the processes that are causing the changes, and it is the visual explorative analysis of data during which a hypothesis is formed. We would like to show the usability of multidimensional visualization tools and promote their use in life sciences. We survey and show some of the multidimensional visualization tools in the process of data exploration, such as parallel coordinates and radviz and we extend them by combining them with the self-organizing map algorithm. We use a time course data set of transitional cell carcinoma of the bladder in our examples. Analysis of data with these tools has the potential to uncover additional relationships and non-trivial structures.

***Keywords*—**microarrays, visualization, parallel coordinates, radviz, self-organizing maps.

I. Introduction

MICROARRAYS have become the norm for simultaneous measuring of expression levels of thousands of genes. The "current" and "next generation" platforms have enormous promise in revealing functions of genes, cell populations, tumor classifications [1], understanding cellular pathways, drug target identification, just to name a few [2]-[3]. From the arrays themselves, we have to derive the signal values and then biological conclusions, and the methods that we apply can be roughly divided into preprocessing and data analysis which is further divided into data mining and visualization tools. If we assume that the arrays have been preprocessed and signal data has been obtained, a variety of data mining algorithms are at hand for the next task: from self-organizing maps [4] to hierarchical clustering, one of the most highly utilized data mining methods today. Visualization tools that are heavily used today range from scatter plots to dendrograms (displaying the results of hierarchical clustering) line plots, histograms, box plots and venn diagrams. Statisticians have used majority of these tools for a long time, but modern visualization tools and techniques that are targeted towards high-dimensional data remain underutilized, but not due to the usability or the power of these techniques. Our quick survey of the publications applying multidimensional visualizations in bioinformatics shows that papers using these tools are mostly published in the Information Visualization and Visualization domains rather than life science domains. In 2004 Saraiya et al. evaluated the use of visualization tools by biologists and discovered that there is an overwhelming variety of tools to chose from, and the users are confused about which tool to use [5].

Novel multidimensional visualization techniques enable us to display larger, higher-dimensional data sets in a meaningful, more descriptive manner. They have been shown to have a very high intrinsic dimensionality [6] and uncover non-trivial patterns and relationships in the data. We combine these visualizations with self-organizing maps [7], the topology-preserving neural network and projection technique, which enables us to extend into the third dimension and uncover previously unknown additional relationships.

We are using a data set of the transitional cell carcinoma (TCC) of the bladder generated by Clifford Lab at LSUHSC-S to show these techniques [8]. In Section 2, we first discuss the setup and processing of the TCC data set. We continue with the description of multidimensional visualizations, including parallel coordinates and radviz, and their intrinsic dimensionality in Section 3. We include a list of open source tools that provide these visualizations and their extensions to be utilized by the life scientists for their data analysis tasks. Section 4 is a discussion of self-organizing maps, their extensions and combination with multidimensional visualization techniques: parallel coordinates and radviz. We conclude with Section 5 where we also describe our future goals.

II. Transitional Cell Carcinoma Data Set

Transitional cell carcinoma of the bladder (TCC) ranks 4th in incidence of all cancers in the developed world, yet the mechanisms of its origin and progression remain poorly understood and there are few useful diagnostic or prognostic biomarkers for this disease. In an attempt to generate a mouse

Marjan Trutschl and Urška Cvek are with Department of Computer Science and Laboratory for Advanced Biomedical Informatics, Louisiana State University Shreveport and Center for Molecular and Tumor Virology and Department of Bioinformatics and Computational Biology, LSUHSC-S, Shreveport, Louisiana, USA. (phone: 318-795-4266; e-mail: {mtrutsch, ucvek}@lsus.edu).

Randolph Stone II, Zanobia Syed and John L. Clifford are with Department of Biochemistry and Molecular Biology, Louisiana State University Health Sciences Center-Shreveport, Shreveport, LA, USA. (e-mail: {rstone, zsyed, jcliff}@lsuhsc.edu).

Anita L. Sabichi is with Thoracic/Head and Neck Medical Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA. (e-mail: asabichi@mdanderson.org).

model for bladder cancer progression, investigators in the laboratory of Xue-Ru Wu engineered transgenic mice carrying a low copy number of the SV40 large T (SV40T) oncogene, expressed under the control of the bladder urothelium specific murine uroplakin II promoter [9]. The transgenic mice, UPII-SV40T, develop a condition closely resembling human carcinoma *in situ* (CIS) starting as early as 6 weeks of age, progressing to invasive TCC from 6 months of age onward. We combined this transgenic mouse model for invasive bladder cancer with Affymetrix DNA microarray technology. With the Mouse GeneChip (Mouse Genome 430 2.0) it is possible to determine a relative expression level of over 39,000 mouse transcripts (45,101 probe sets), representing the majority of the transcribed mouse genome, in a given mRNA sample. Duplicate arrays were performed for SV40T and non-transgenic littermates (WT) at each time point, yielding a set of duplicated arrays for two factors: mouse genotype (WT or SV40T) and week (3, 6, 20, 30) creating eight targets. The WT line at the 6 week time point is the exception – due to the degradation of the RNA we only have one quality array. We followed the recommended analysis techniques [10]-[11] using R [12], bioconductor [13], and used the *limma* [14] and *affy* packages [15].

We characterized the histologic progression of premalignant, carcinoma in-situ, early invasive TCC and more advanced invasive TCC occurring at 3, 6, 20 and 30 weeks of age, respectively, in the UPII-SV40T mice. We performed a preliminary examination of genes expressed in the urothelium at these time points and revealed approximately 1,900 differentially expressed (≥ 3 fold difference) at one or more of the time points between the urothelium of SV40T mice and their age-matched WT littermates. Preliminary biometric analysis using the Ingenuity Pathways Analysis software package (Ingenuity Systems Inc.) revealed that cell cycle regulatory, DNA replication, and cancer related genes were more strongly expressed in the SV40T bladder urothelium at the highest proportion, even at the 3-week point.

We identified genes that are differentially expressed between the bladders of SV40T mice and their age-matched wild type (WT) littermates at 3, 6, 20 and 30 weeks of age. These are the times, which correspond to premalignant, carcinoma *in-situ*, and early and later stage papillary TCC, respectively. Empirical Bayes method moderated *t*-statistic was used to test each individual contrast equal to zero and compute the moderated F-statistic which combines the *t*-statistic for all the contrasts into an overall test of significance for that gene. *p*-values were adjusted for multiple testing using the method of Benjamini and Hochberg to control the false discovery rate. Tests were considered to be significant if the adjusted *p* value did not exceed 0.05. We eliminated the control probes from our analysis using the cutoff *p*-value and required at least a one-fold change between the arrays to consider them as differentially expressed.

Figure 1 shows the counts of genes that were up or down regulated for each of the lines, when compared to the first time point for the WT. There is virtually no regulation present in the WT, while we observe an exponential growth of the number of differentially expressed genes for the SV40T line from approximately 1,300 to 2,100 and 4,400 at time points 6, 20 and 30, respectively. We proceeded by identifying the difference in expression between the WT and SV40T lines for each of the time points. Figure 2 shows the number of genes that are up or down regulated at each time point using the F-statistic with an additional requirement of a log fold change of 1.5 or greater. The increase of the number of differentially expressed genes is still apparent, and there are 17 genes that are exponentially increasing the regulation of expression (either up or down) at every point from time 3 to time 30.

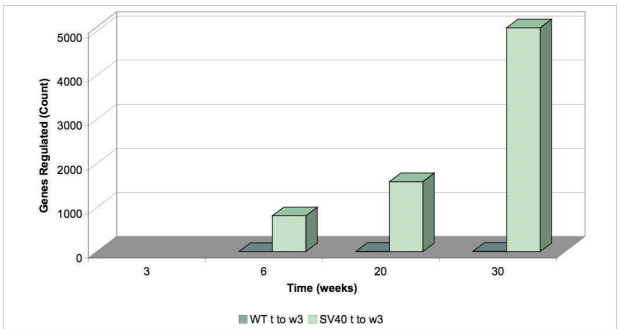

Fig. 1 Gene regulation at time points 6, 20 and 30 for the WT and SV40T lines. Differential expression is present only in the SV40T line (green).

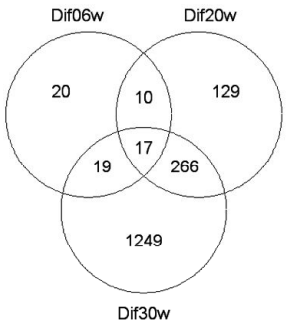

Fig. 2 Two-way analysis confirms the exponential increase of regulation of genes in SV40T line.

We focused first on the set of 17 probe sets in the intersection. We clustered them using two-way hierarchical clustering (implemented in R/bioconductor) followed by the analysis in Expression Profiler [16]. One of our surprising findings was CLDN3 gene that has 5 probe sets on 430 2.0 array and appears with 4 sets in our small list of 17. We further analyzed the set of 585 genes that are differentially expressed at the early stages, namely weeks 6 and 20 (Fig. 2). The goal of the project is the early identification of bladder cancers, thus the choice of time points (week 20 is classified as early stage papillary TCC). We tested several of the genes upregulated in SV40T urothelium, including hyaluronan mediated motility receptor (RHAMM), RacGAP1, PCNA and others as biomarkers for premalignancy, in urine samples from a completed chemoprevention trial.

III. Multidimensional Visualization

Data sets of four or more dimensions are harder to display on a two-dimensional (2D) screen, or a piece of paper. We can handle three or four dimensions by projecting them onto the 2D surface. We could also use two of the dimensions as *x* and *y* coordinates and the other dimensions as the color, texture, size, shape of each record. Our perceptual ability limits us to five to seven dimensions that we can normally track on a visual display. In recent years several research efforts resulted in new techniques to display large multivariate data sets but due to the exponential data growth, we have to continue with the process of tool development. The *first set* of novel techniques is focused on the reduction of the data size and preservation of its significant features. Pixel level visualization schemes were proposed that permit the display of a large number of records and are not scalable but rather dependent on the size of the display area [17]. The *second* set of techniques is based on matrices [18]-[20] as one of the techniques first utilized to address this problem. A scatter plot matrix, for example, shows all pairwise scatterplots, mirrored across the diagonal. The *third* set of techniques is the group of line-based (parallel coordinates) and point projections (radviz). Parallel coordinates use parallel axes instead of perpendicular axes [21]-[22]. Radviz [23] positions dimensional anchors around the perimeter of a circle using Hooke's law. We can measure the visual effectiveness of these visualizations to determine their benefits. However, no matter how effective a visualization, we can always identify a data set with a larger number of dimensions or records than can be displayed. Nevertheless, it is our goal to steer the reader towards multidimensional tools that can accommodate medium and larger data set exploration and aid in the visual knowledge discovery. These are not to replace the statistical and computational tools but rather complement them. We continue to enhance current techniques and devise new ones that will help us identify non-trivial patterns in the data.

A. Parallel Coordinates

The parallel coordinates [21]-[22] algorithm has been applied to a wide range of data analysis tasks. Parallel coordinates (PC) are a geometric projection visualization technique that represents dimensions on parallel common axes, arranged horizontally. Each record corresponds to a polyline that intersects the axes at the record's dimensional values. They can display a large number of dimensions of a data set, but suffer from the disadvantage that the number of records that can be displayed is limited. Visualizing a medium or larger size data set usually results in over-plotting or clutter, a featureless blob, which hides the underlying data structure. Several techniques have extended the original parallel coordinates algorithm to overcome this problem by dimension reordering or summarization, such as clustering, sampling and filtering. Regardless of their application, we can group existing techniques into 2D and 3D based on their topology.

A number of PC-based techniques cluster or organize the records first and then visualize the cluster centroids. Peng et al. [24] first measure the clutter and then minimize it by dimension ordering not only for parallel coordinates but also for other multidimensional visual techniques. This is implemented in tools such as XmdvTool [25], that allow manual or automatic reordering of (hierarchical dimension ordering), increasing and decreasing spacing between axes, filtering and focus+context technique DOSFA [26]. Clustering is applied to dimensions or to records or both. Fua et al. [27] propose a multiresolutional view of the data using hierarchical clustering to records and displaying aggregate clusters as bands faded from an opaque centre to a transparent edge. Users can navigate and filter the data and select a desired focus region and level of detail [28]. Clustered parallel coordinates are also developed by Johansson et al. [29], where clusters and transfer functions are combined with textures to highlight different aspects of cluster characteristics. Siirtola [30] visualizes the correlation coefficients between polyline subsets and replaces the polylines with their average, introducing quick manipulation with the plots. Correlation between variables is plotted as a bar between the ranges, pointing upwards or downwards, indicating positive or negative correlation. Lesh and Mitzenmacher [31] propose an interactive data summarization where the results of an exhaustive search for manipulations that would change the summary are identified. Frequency plots [32] calculate and highlight the frequency regions for each dimension, a very useful technique for exploration of clusters but less efficient for the whole data structure. Using fuzzy rules Berthold and Hall [33] apply a similar approach and Andrienko and Andrienko [34] utilize envelopes and ellipse plots. These approaches divide each dimension's range of values into frequency intervals and convey the information on independent variables. Envelopes or polygons are representing clusters in the parallel coordinate approach by Novotny [35] combined with clusters displayed as striped textures.

Parallel coordinates are also augmented by additional information displays or interactive tools, such as histograms, frequency or density information, glyphs and coupled visualizations. Average shifted histograms visualize density plots with parallel coordinates [36]-[37] aiming to remove the problem of dimension's bins or frequency intervals. This approach is extended by painting pixels of polylines with intensity proportional to the pixel's record overlap [38]. Artero et al. [39] create an interactive parallel coordinates frequency and density plot by computing frequency and density information from the data and mapping data density to the intensity of parallel coordinate polylines. Together with the constraint that denser lines cannot be drawn on top of less dense lines, this creates a visualization in which the clusters can be detected. Visual data mining displays [40] use cluster centroids placed on top of the parallel coordinates and tracked statistical measures displayed as static or animated glyphs in a separate coordinate system. SpringView [41] explores coupled multiple views of radviz and parallel coordinates and applies brushing, color and data clustering to facilitate data exploration and reduce clutter. Parallel coordinates have also been interactively linked with star glyphs, scatter plots and dimensional stacking [42]-[45], among others.

In Figure 3 we show the over-plotting or occlusion that occurs when we visualize the set of 585 unique TCC genes that are differentially expressed in SV40T line at week 6 and/or week 20. We display all of the 15 dimensions of this data set and order them by first listing the SV40T lines followed by the WT lines. We cannot notice any significant differences across the dimensions, although we can detect that majority of the records have low to mid values. There are only a few records whose signal values were high. In the center of the display we notice a dip in the expressions, which is due to the switch from the SV40T dimensions to the WT dimensions and is not a true pattern in the data set.

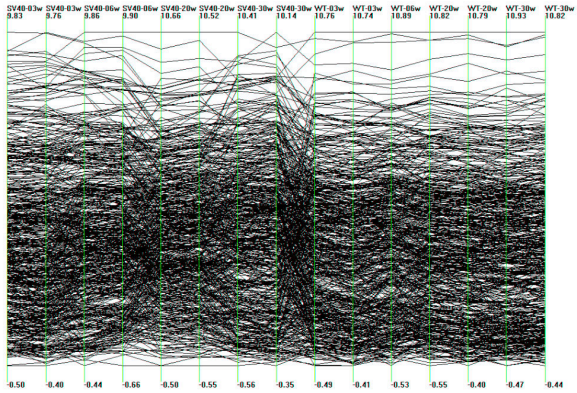

Fig. 3 Parallel coordinates of the 585 early changing genes of the TCC data

B. Radviz

Radial display technique places dimensional anchors (dimensions) around the perimeter of a circle and utilizes spring constants to represent relational values among points is the technique known as radviz (radial visualization). As shown in Figure 4, radviz utilizes spring constants to represent relational values among points. A record is represented as a vector $(x_{i1},...,x_{im})$ on m dimensions and its position is determined by the pull of the position vectors (dimensions) $(\bar{S}_{1,...,}\bar{S}_m)$. The record in Figure 4 has m=8 dimensions ordered on the circle in counter-clockwise equidistant fashion. Each position vector points from (0,0) to the corresponding fixed point on the perimeter of a unit circle. The values of each dimension are usually normalized to a 0 to 1 range to eliminate any effects of the variable minimum and maximum values in the range. Each data point is displayed at the point where the sum of all spring forces equals zero and the stiffness of each spring is proportional to the value of the corresponding dimension. The point ends up at the position where the spring forces are in equilibrium:

$$\sum_{j=1,m}(S_j - u_i)x_{ij} = 0 \qquad (1)$$

The position of the data point depends largely on the arrangement of dimensions around the circle; however, data items with similar values are always placed close together. The technique has been complemented by dimension ordering approaches, where the dimension order is determined by the structure of the data or the inherent class separation [46]-[47],[41].

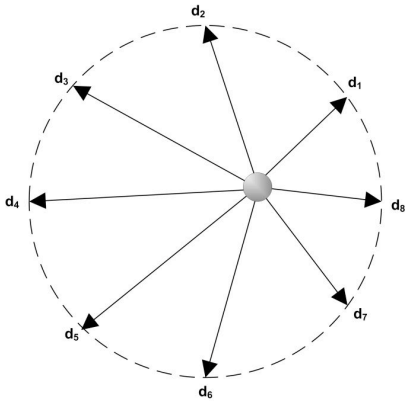

Fig. 4 Example of radviz spring forces of an 8-dimensional record

One of the major disadvantages of radviz is the overlap of points that occurs not only when the records have identical values on the displayed dimensions, but also when the records are scaled. For example, records (1,1,1,1,1,1) and (10,10,10,10,10,10) would appear at the same point in the center of the circle (they are pulled by all dimensions equally). Dimension ordering and placement of dimensions away from the radial layout minimizes this problem, but does not completely solve it. We developed an approach that utilizes the third dimension to organize the data when overlap occurs.

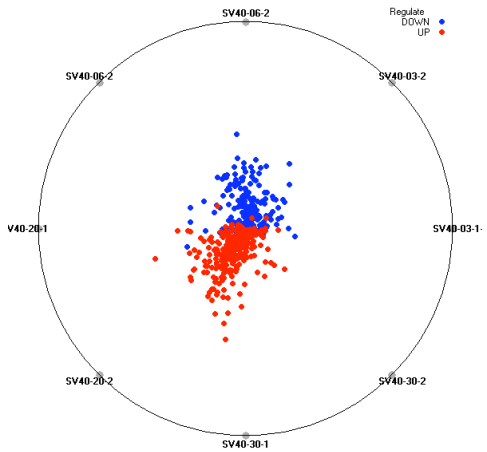

Fig. 5 Radviz display of the 585 early changing genes of the TCC data's SV40T line

When we examine the radviz display of the SV40T line on the 585 early differentially expressed genes of the TCC data set, we noticed that majority of them are either in the center (pulled equally by all dimensions) or positioned towards the bottom of the visualization. We can conclude that the signal values at the 30-week time point are relatively large and that our genes are more likely upregulated. When we show the UP or DOWN regulation for the genes as the record color, we can confirm this.

C. Intrinsic Dimensionality

The goal of intrinsic dimensionality [6] is to measure how visualizations deal with *n* dimensions when displayed on the screen. This information guides our decision when selecting a visualization appropriate for a data set at hand.

One of the main issues is the point overlap, also called occlusion or over-plotting, and the loss of data in projection. Given an *n*-dimensional space, the *intrinsic dimension* (ID) of a visualization is the largest *k* for which a set of *k* unit vectors can be uniquely identified (perceived) in the visualization. We consider 10 and 100 unit vectors (its dimensional values are 0 or 1) in 2D and 3D to produce scatter plots in 10- and 100-dimensional space. The intrinsic dimension of a 2D scatter plot is 2 and 3D is 3. The *intrinsic coordinate dimension* (ICD) of a visualization is the largest *k* for which the coordinates of any vector in that space can be uniquely identified in the visualization. The intrinsic coordinate dimension for the 2D scatter plot is 2, and for the 3D scatter plot it is 2 as well (the projected point may come from several points projecting to a line in 3D). If we drew the parallel coordinates plot using the 10- and 100-dimensional unit vector datasets, we would be able to identify 10 and 100 intrinsic dimensions. Each of the coordinate values can be uniquely identified and thus we also have a 10- and 100- dimensional intrinsic coordinate dimension in the parallel coordinate plot. The intrinsic dimension for the radviz display is 10 and 100, respectively. The intrinsic coordinate dimension is not identifiable in general, if the point is not on the boundary of the circle.

D. Multidimensional Visualizations Tools

Our focus is to present a few of the most powerful multidimensional tools from the information visualization domain that are available in the public, non-commercial domain and must include at least a parallel coordinates and a projection display (such as radviz). We do not include the techniques that are static, for example, tools that work through a browser and do not allow the user to select or in some other way interact with the data. This step also excludes Matlab, R project and others. Selection enables the user to select and highlight items in the visualization(s) and we find this to be an important feature of a visualization tool. Navigation of the visual space includes focus and context techniques, allowing scrolling, panning, zoom, rotation, etc. We prefer to utilize tools that provide multiple coordinated plots, where the selection, for example, is propagated or linked from the scatter plot to the parallel coordinates and other displays as this provides for a richer exploration environment. We focus on general and bioinformatics data analysis tools and exclude the tools that are geared specifically towards hierarchical or time-series data.

XmdvTool [25] is an exploratory tool that provides a brushing-and-linking concept where the data can be brushed (in *m* dimensions) in one plot and it is propagated to the other visualizations displaying the same data set. Xmdv includes the implementation of hierarchical PC to overcome the difficulty of interpretation in large data sets in addition to scatter plots, star glyphs, dimensional stacking and pixel-oriented display. Xmdv also allows the axes to be moved to different positions, providing an insight into additional data relationships.

Hierarchical Clustering Explorer [48] is a tool specifically geared towards microarray experiment data sets. It includes coordinated plots of parallel coordinate-like space, histograms and scatter plots (or projections onto the 2D space). The users can order records in histograms or in scatterplots, depending on the number of dimensions.

Orange is an open-source component-based data mining and visualization suite [49]. It provides several visualization and data mining techniques, including radviz, polyviz, parallel coordinates and data-driven arrangement of dimensions in both (based on class data). The interface provides objects that identify the workflow and processes the data go through (the user can drag and connect them on the display). Clicking on the object provides an interface to the settings and parameters and execution of the underlying functionality. The suite does not provide coordinated plots, but does allow linking of dependent activities (different visualizations, data mining algorithms, etc.).

IV. Self-Organizing Map

A. Algorithm and Extensions

Self-organizing map (SOM) is an unsupervised neural network that facilitates mapping of a set of *n*-dimensional vectors to a two-dimensional topographic map [7]. Training of the unsupervised neural network is data-driven, without a target condition, therefore, the output of a SOM algorithm represents the relationships among the input vectors. It is a summarization technique that attempts to reduce the complexity of the data set by displaying clusters of the data in a grid. Its widespread use is attributed to its simplicity.

The learning of the SOM is the process in which we form a nonlinear projection of the records onto a map. The self-organizing grid or map consists of an array of output nodes (neurons), each of them associated with an *m*-dimensional weight vector m_i (corresponding to the *m* dimensions of the input data set). Initial values of m_i may be randomly selected, preferably from the data set. Each record is positioned on the map, one by one, until the data set is exhausted. The assignment of weight vectors is formed in an unsupervised learning process, and the records are randomly drawn from the input distribution and presented to the network one at a time.

A record is mapped onto the SOM by calculating the similarity between the input vector and node *i*'s weight vector m_i. Each node *i* receives the same input vector and produces a single similarity value. The input record maps onto the best-matching (winning) node *c*, based on the largest similarity or the smallest distance (depending on the implementation). The weight vectors of nodes topologically close to the winning node (up to a certain geometric distance) adjust their weight vectors, "learning" something about the input. The adjustment depends on the size of the neighborhood, the value of the neighboring function and the learning function. This results in a local relaxation or smoothing of the neighborhood, which

with continued learning leads to global ordering. This process is repeated until the output map converges to a stable or organized state when the average error falls below a pre-specified value or a certain number of iterations have been reached. The self-organizing process works by repeated refinement and progressively smaller values of the learning function.

Most SOMs are visualized on a rectangular display of output nodes, although hexagonal and irregular grids are also used. Numerous SOM algorithms and extensions have been developed in a multitude of fields, which include biomedical applications. Investigations include self-adaptive and incremental learning neural networks (SANN) that would replace the static topology networks [50]-[51], tree-structured SOM network architecture [52], alternate neural-network based projections [53]-[56]. Some of these approaches aim to determine the shape and size of the self-organizing structure during the learning process and are targeted towards specific domains. One of the seminal applications of SOMs in the biomedical arena was the work by Tamayo et al. [4] in which SOMs were used to find the classes in 828 genes of the Yeast cell cycle.

B. *Combining SOM with Multidimensional Visualizations*

1) *3D PC and 3D Radviz Algorithms*

The dimensionally capable nature of multidimensional visualization plots results in displays of large, multi-dimensional data sets. These techniques have their drawbacks (for example, a parallel coordinates plot results in unwanted inter-dimensional blobs of edges) and we overcome them by combining them with SOMs.

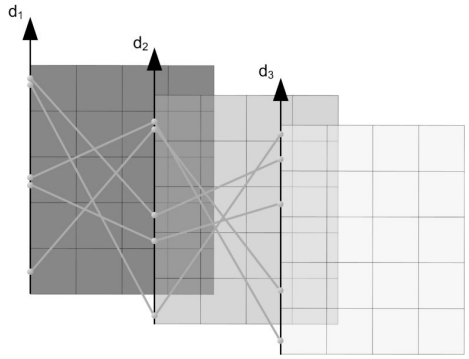

Fig. 6 Primary and secondary mapping steps of our 3D PC algorithm

Original SOM algorithm only has one mapping step at which the output node for the record is determined. Our algorithms consist of the primary and secondary mapping steps that help us merge the SOM and multidimensional visualization. For the 3D *parallel coordinates* we first start by replacing the original dimensional axes by grids of output nodes and then proceed with a two-step mapping process (Fig. 6). *Primary mapping* determines the location for a record on the dimensional axis, just as in the original step. The main difference is that instead of mapping to the value's position on the axis, we map to a primary bin that has been created by the grid in the *y* dimension. The second step is *secondary mapping* where the single grid cell is replaced by the single-dimensional set of SOM output nodes to which the record can map. The output node is chosen based on the Euclidean distance between the record and the representative weight vectors of the output grid (Fig 6.). The neighborhood, learning function, repetition of this process and adjustment of neighboring nodes for all of the records adjusts these grids in three dimensions and places the records with similar dimensional values closer together. The neighborhood function works in three dimensions: *y* and *z* (grid output nodes) and *x* (horizontally) across the grids. We stop the process after the learning stops or after we pass a learning threshold.

The organization of records into the three PC dimensions provides a larger surface on which the records' polylines can be organized, which helps remove some of the overlap, but does not completely eliminate it. There are cases where additional approaches have to be taken to deduce the structure of the data displayed. One of the options we provide is the option of *bundling* records. We represent the records plotted in a 3D PC plot mathematically in the form of Bezier curves (replacing polylines). This breed of curve interpolation utilizes a series of appropriately placed control points to guide the direction of the curve. When the edges are bundled, we are actually grouping the control points together. To set the number of bundles in the *y* and *z* directions, we implemented partitioning of the plane into a number of partitions. The second option we provide is *force-directed record placement*. This approach utilizes weight vectors used for self-organization along the grids. Using these weights, nodes are either pushed apart or pulled together, depending upon their similarity.

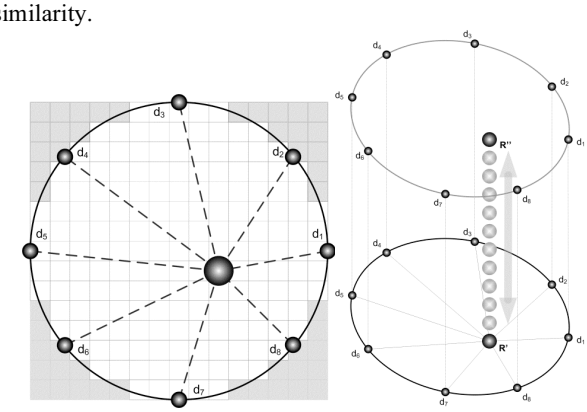

Fig. 7 Binned radviz surface and secondary mapping into the third dimension

For the 3D *radviz* we first start by placing the records based on the original radviz algorithm. The main difference is that instead of positioning it in the exact position, we grid the radviz surface and position the record into its primary bin (Fig. 7). The position is then augmented in the *secondary mapping* step, where we build a single-dimensional SOM at each radviz grid cell into the third dimension. The same process is repeated for all of the records, adjusting the neighborhood based on the neighborhood function and learning functions and stopping after no learning takes place (or it falls below a

certain threshold). The neighborhood function affects the three dimensions (x, y, z) as all of the neighboring output nodes are adjusted.

2) Application to the TCC Data Set

We show the results of our techniques on the TCC data set of 585 records. Figure 8 shows the binned parallel coordinates plot of the 585 records that are differentially expressed early (weeks 6 and 20). We are displaying one replicate at the 3, 6, 20 and 30 week points (in order from left to right) of the SV40T line. The data is colored by the 30 week time point (the dimension furthest to the right). If we compare it to the original parallel coordinates plot, it is less cluttered, as the records have been collapsed into the bins (10 bins were create for each dimension). The points appear to distribute evenly on each dimension. There is a lot of occlusion and the only way that structure like this can be explored further is by utilizing brushing, selection, filtering and similar interaction techniques to explore the properties of the data set.

Our next step was to explore the projection using the 3D PC, which position a grid of 10x10 output nodes at each of the dimensional axes (Fig. 9). We can start discovering more structure – not all the records that have the same dimensional continue to stay together in the third dimension. We can observe that while some of the records have higher dimensional values, most of them do not stay at the high value across all dimensions but rather project onto the low or medium dimensional values in majority of the dimensions.

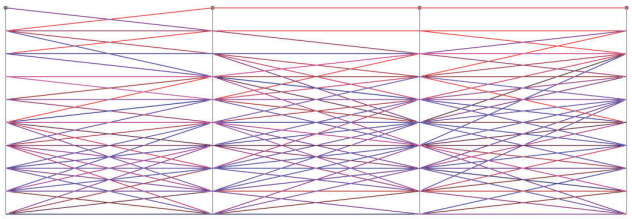

Fig. 8 Example of binned classic parallel coordinates – 585 records of the SV40T set, 4 time points. Colored by week 30 values.

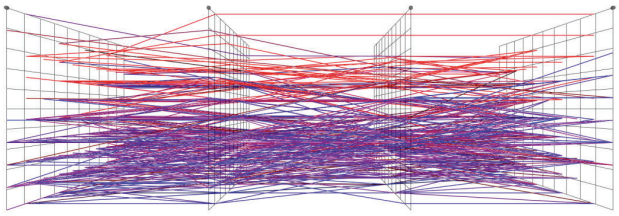

Fig. 9 The records shown with the SOM grids replacing dimensional axes. Colored by week 30.

Our next step is to apply the bundling approach to the data (Figure 10). This process separates the data into two distinct groups; one at the top and one at the bottom of the dimensional values. When we reexamine that these 585 records belong to the two groups of records that are differentially expressed at the 6 and/or 20 week time point, we could expect that this type of organization is going to occur. This is especially due to the small number of overlapping records (27) that are differentially expressed at both time points. We explore this visualization by using the selection, brushing and rotational tools.

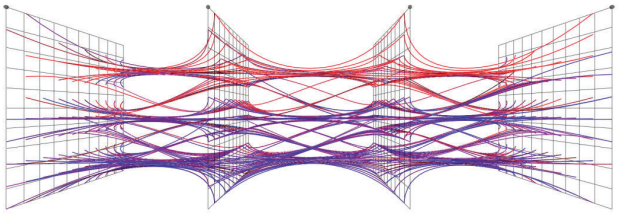

Fig. 10 Creating bundles of records that represent record clusters

The next visualization displays the bundling and force-directed placement of records (Figure 11). Record edges are pulled together whenever the output nodes contain similar weight vectors. This helps to further reduce the clutter and provide a cleaner set of data patterns to be interactively explored.

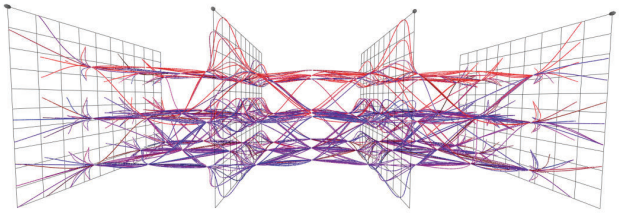

Fig 11 Force-directed record placement and bundling of the records.

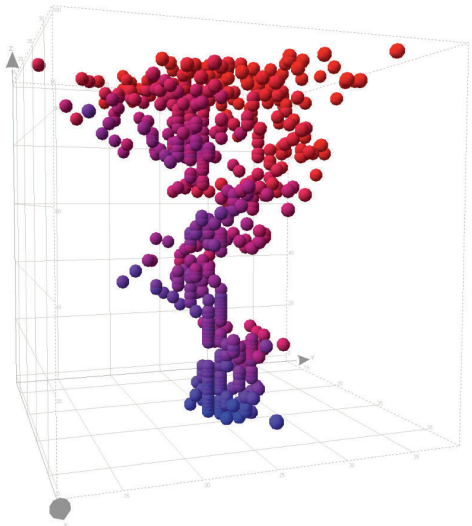

Fig. 12 3D Radviz projection of the eight dimensions of the SV40T line, extending from the primary positioning of the records in radviz (Fig. 5)

We proceed by exploring the data in the 3D radviz visualization. We already showed the classic radviz on the 585 record data set (Fig. 5), and the 3D radviz approach moves the records into the third dimension (Fig. 12). The pull is driven by the self-organization of the selected data - all of the eight dimensions of the SV40T set (a set of duplicates for each of the four time points). We color the records by one of the 30 week time points, creating a color map from red (low) to blue (high) values. The distribution of signal values of the data is not uniform, and there are more values at the lower end of the range. The records are positioned with regards to all eight of their SV40T dimensions, thus placing the records with low

values on all of the dimensions at the "ceiling" and higher end of the range at the "floor." We use the selection (brushing), filtering, rotation and zoom tools to interact with the data.

V. Conclusions and Future Work

We emphasize the importance of multidimensional visualization tools in the knowledge discovery process of microarray and other life science data. We list today's most powerful free public tools that provide coordinated multiple-view visualizations. We discuss the details of two visualization techniques: parallel coordinates and radviz, which we combine with the self-organizing map projection to showcase novel data exploration methods. We are currently extending the capabilities of our tools for larger data sets, implementing them in parallelized versions. As with all software packages, their thorough evaluation is an ongoing process and we continue our designs based on the feedback from life scientists that are using these techniques in their everyday work.

Appendix

Appendixes, if needed, appear before the acknowledgment.

Acknowledgment

The project was supported by NIH Grant Number P20RR016456 and P20RR018724 from the National Center for Research Resources. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health.

References

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439), pp. 531-537, 1999.

[2] P.T. Spelman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Fucher. Comprehensive identification of cell-cycle regulated genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9(12), pp. 3273-3297, 1998.

[3] T. Zhang, R. Ramakrishnan, M. Livny. Birch: an efficient data clustering method for very large databases. *Proc.Int. Conf. Management of Data,* pp. 103-114, 1996.

[4] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Atl. Acad. Sci.,* 96(6), pp. 2907-2912, 1999.

[5] P. Saraiya, C. North, K. Duca. An evaluation of microarray visualization tools for biological insight. *Proc. Information Visualization 2004*, pp. 1-8, 2004.

[6] G. Grinstein, M. Trutschl, U. Cvek, High-dimensional visualizations. *7th ACM/SIGKDD Data mining Conference (KDD), 2001*.

[7] T. Kohonen, Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, vol. 43, pp. 59-69, 1982.

[8] R. Stone II, A.L. Sabichi, J. Gill, I.Lee, R. Loganatharaj, M. Trutschl, U. Cvek, J.L. Clifford. Identification of genes involved in early stage bladder cancer progression. *Unpublished.*

[9] Z.T. Zhang, J. Pak, E. Shapiro, T.T. Sun, X.R. Wu. Urothelium-specific expression of an oncogene in transgenic mice induced the formation of carcinoma in situ and invasive transitional cell carcinoma. *Cancer Res*., 59(14), pp. 3512-7, 1999.

[10] R. Gentleman, V. Carey, et al. (editors) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, 2005.

[11] R. Gentleman, W. Huber. Working with Affymetrix data: estrogen, a 2x2 factorial design example. Practical Microarray Course, Heidelberg, 2003.

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna Austria, 2008.

[13] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), R80, 2004.

[14] G.K. Smyth. Limma: Linear models for microarray data. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. R. Genleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (editors) Springer pp. 397-420, 2005.

[15] L. Gautier, L. Cope, B.M. Bolstad, R.A. Irizarry. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics,* 12(3), pp. 307-315, 2004.

[16] A. Torrente, M. Kapushesky, A. Brazma. A new algorithm for comparing and visualizing relationships between hierarchical and flat gene expression data clusterings. *Bioinformatics* 21(21), pp. 3993-3999, 2005.

[17] D. Keim, H. Kriegel, M. Ankerst. Recursive pattern: a technique for visualizing very large amounts of data. *Proc. Visualization 1995*, pp. 279-286, 1995.

[18] D.F. Andrews. Plots of high-dimensional data. *Biometrics*, 29, pp. 125-136, 1972.

[19] J.M. Chambers, W.S. Cleveland, B. Kleiner, P.A. Tukey. *Graphical Methods for Data Analysis*, Chapman and Hall, 1976.

[20] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin, Madison, WI, 1983.

[21] A. Inselberg, The plane with parallel coordinates. *The Visual Computer,* pp. 69-92, 1985.

[22] A. Inselberg, B. Dimsdale, Parallel coordinates: A tool for visualizing multidimensional geometry. *Proc. IEEE Visualization*, pp. 361-378, 1990.

[23] P. Hoffman, G. Grinstein. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. Presented at NPIV 99 (Workshop on New Paradigms in Information Visualization and Manipulation), 1999.

[24] W. Peng, M.O. Ward, E.A. Rundensteiner, Clutter reduction in multi-dimensional data visualization using dimension reordering. Proc. IEEE Symposium on Information Visualization, pp. 89-96, 2004.

[25] M.O. Ward, XmdvTool: Integrating multiple methods for visualizing multivariate data. Proc. IEEE Visualization 1994, pp. 326-333, 1994. URL: http://davis.wpi.edu/~xmdv/.

[26] J. Yang, W. Peng, M.O. Ward, E.A. Rudensteiner, Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. Proc. IEEE Symposium on Information Visualization, pp. 14-21, 2003.

[27] Y.-H. Fua, M.O. Ward, E.A. Rundensteiner, Hierarchical parallel coordinates for exploration of large datasets. Proc. IEEE 5th International Conference on Information Visualization, pp. 425-432, 2001.

[28] Y.-H. Fua, M.O. Ward, E.A. Rundensteiner, Navigating hierarchies with structure-based brushes. Proc. IEEE 5th International Conference on Information Visualization, pp. 58-64, 1999.

[29] J. Johansson, P. Ljung, M. Jern, M. Cooper, Revealing structure within clustered parallel coordinates displays. Proc. IEEE Symposium on Information Visualization, pp. 125-132, 2005.

[30] H. Siirtola, Direct manipulation of parallel coordinates, Proc. IEEE 4th International Conference on Information Visualization, pp. 373-378, 2000.

[31] N. Lesh, M. Mitzenmacher, Interactive data summarization: an example application. Proc. Working Conference on Advanced Visual Interfaces, pp. 183-187, 2004.

[32] J.F. Rodrigues, Jr., A.J. Traina, C. Traina, Jr., Frequency plot and relevance plot to enhance visual data exploration. Proc. XVI Brazilian Symposium on Computer Graphics and Image Processing, pp. 117-134, 2003.

[33] M. Berthold, L.O. Hall, Visualizing fuzzy points in parallel coordinates. IEEE Transactions on Fuzzy Systems, pp. 369-374, 2003.

[34] G. Andrienko, N. Andrienko, Parallel coordinates for exploring properties of subsets. Proc. 2nd IEEE Conference on Coordinated and Multiple Views in Exploratory Visualization, pp. 93-104, 2004.
[35] M. Novotny, Visually effective information visualization of large data. Proc. 8th Central European Seminar on Computer Graphics, pp. 41-48, 2004.
[36] J.J. Miller, E.J. Wegman, Construction of line densities for parallel coordinate plots. Computational Statistics and Graphics, eds. A. Buja, P. Tukey, Springer-Verlag, pp. 107-123, 1990.
[37] E.J. Wegman, Hyperdimensional data analysis using parallel coordinates. Journal of American Statistical Association, 85 (411), pp. 664-675, 1990.
[38] E.J. Wegman, Q. Luo, High dimensional clustering using parallel coordinates and the grand tour. Proc. Conf. German Classification Society, Freiburg, Germany, 1996.
[39] A.O. Artero, M.C. Ferreira de Oliveira, H. Levkowitz, Uncovering Clusters in Crowded Parallel Coordinates Visualizations. Proc. IEEE Symposium on Information Visualization, pp. 81-88, 2004.
[40] D. Ericson, J. Johansson, M. Cooper, Visual data analysis using tracked statistical measures within parallel coordinate representations. Proc. 3rd IEEE Conference on Coordinated and Multiple Views in Exploratory Visualization, pp. 42-53, 2005.
[41] E. Bertini, L. Dell' Aquila, G. Santucci, Springview: cooperation of radviz and parallel coordinates or view optimization and clutter reduction. Proc. 3rd IEEE International Conference on Coordinated & Multiple Views in Exploratory Visualization, pp. 22-29, 2005.
[42] P.C. Wong, R.D. Bergeron, Multivariate visualization using metric scaling. Proc. IEEE Visualization 1997, pp. 111-118, 1997.
[43] Y.-H. Fua, M.O. Ward, E.A. Rundensteiner, Hierarchical parallel coordinates for exploration of large datasets. *Proc. IEEE 5th International Conference on Information Visualization*, pp. 425-432, 2001.
[44] M.O. Ward, XmdvTool: Integrating multiple methods for visualizing multivariate data. *Proc. IEEE Visualization 1994,* pp. 326-333, 1994.
[45] J. Yang, A. Patro, S. Huang, N. Mehta, M.O. Ward, E.A. Rundensteiner, Value and relation display for interactive exploration of high dimensional datasets. Proc. IEEE Symposium on Information Visualization 2004, pp. 73-80, 2004
[46] G. Leban, I. Bratko, U. Petrovic, T. Curk, B. Zupan. VizRank: finding informative data projections in functional genomics by machine learning. *Bioinformatics*, 21, 2005.
[47] P. Au, M. Carey, S. Sewraz, Y. Guo, S. Ruger. New paradigms in information visualization. *Proc. 23rd International ACM SIGIR Conference*, Athens, Greece, 2000.
[48] J. Seo, B. Shneiderman. A Rank-by-Feature framework for unsupervised multidimensional data exploration using low dimensional projections. *Proc. IEEE InfoVis2004*, pp. 65-72, 2004.
[49] URL: http://www.cs.umd.edu/hcil/hce/
[50] J. Demsar, B. Zupan, G. Leban. Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper. Faculty of Computer and Information Science, University of Ljubljana.
[51] URL:www.ailab.si/orange
[52] M.A. Nour, G.R. Madey. Heuristic and optimization approaches to extending the Kohonen self-organizing algorithm. *European Journal of Operational Research*, 93(2), pp. 428-448, 1996.
[53] B. Fritzke. Growing cell structures - a self-organizing network for unsupervised and supervised learning. *Neural Networks* 7, 9, pp. 1441-1460, 1994.
[54] P. Koikkalainen, E. Oja. Self-organizing hierarchical feature maps, *International Joint Conference on Neural Networks* IJCNN'90, pp. 279-284, 1990.
[55] E. Oja. A simplified neuron model as a principle component analyzer. *Journal of Mathematical Biology* ,15, pp. 267-273, 1982.
[56] M. A. Kraaijveld, J. Mao, A.K. Jain. A nonlinear projection method based on Kohonen's topology preserving maps. *IEEE Transactions on Neural Networks*, 6(3), pp. 548-559, 1995.
[57] D. Merkl, A. Rauber. Alternative ways for cluster visualization in self-organizing maps, *Proc. Workshop on Self-Organizing Maps*, pp. 106-111, 1997.
[58] M.-C. Su, H.-T. Chang. Fast self-organizing feature map algorithm, *IEEE Transaction on Neural Networks*, 11(3), pp.721-727, 2000.