Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., and Noll, D.C. (1995), "Improved Assessment of Significant Change in Functional Magnetic Resonance Imaging (fMRI): Use of a Cluster Size Threshold," *Magnetic Resonance in Medicine*, 33, 636-647. *(Our first attempt to increase the power of the t-testing procedures.)*

Genovese, C.R., Noll, D.C., and Eddy, W.F. (1997), "Estimating Test-Retest Reliability in Functional MR Imaging I: Statistical Methodology," *Magnetic Resonance in Medicine* (in press). *(What happens when you repeat an fMRI experiment?)*

Just, M.A., Carpenter, P.A., Keller, T.A., Eddy, W.F., and Thulborn, K.R. (1996), "Brain Activation Modulated by Sentence Comprehension," *Science*, 274, October 4, 1996, 114-116. *(An example application of fMRI.)*

Lange, N. (1996), "Statistical Approaches to Human Brain Mapping by Functional Magnetic Resonance Imaging," *Statistics in Medicine*, 15, 389-324. *(A very long introduction to fMRI.)*

Willian F. Eddy
Carnegie Mellon University
bill@cmu.edu

TOPICS IN INFORMATION VISUALIZATION

Templates for Looking at Gene Expression Clustering

By Daniel B. Carr, Roland Somogyi and George Michaels

1. Introduction

In this paper, we describe the design of graphical displays for investigating clustering evident in gene expression data. The displays include stereo plots, parallel coordinate (time series) plots and conditioned parallel coordinate plots. These basic templates are subject to numerous variations and are potentially useful in many other cluster analysis settings.

As an application of our approach, we will consider gene expression mapping of the developing spinal cord in rats, focusing on only 112 genes. Because tens of thousands of interacting genes control spinal cord development, this study is really addressing just the tip of the iceberg. Some background information is appropriate to place this study in a larger context.

We are in a new era in which biologists can selectively disrupt genes and design genes to perform specific tasks. However, genes do not function in isolation. In gene knock-out experiments, the deletion of a gene can have a disastrous effect in some cases, while in others the working constellation of genes compensates quite well for the gene's absence (see Galli-Taliadoros et al. 1995). This suggests redundancy of gene function and combinatorial regulation of genes: a complex *genetic network*.

The study of genetic networks is one of the topics in recent books and journals addressing complexity (see Kauffman 1993, Somogyi and Sniegoski 1996, and `http://rsb.info.nih.gov/mol-physiol/homepage.html`). The introduction of networks into this area has effectively split the ranks of biologists. The old guard continues to focus on the study of individual genes using an evolving but relatively mature methodology. Those taking on the challenge of genetic network studies are pioneers who must modify and create conceptual models and methodologies to view this new landscape of molecular interactions. This work is much different than studying electronic communications networks where one can at least consult with the engineers who designed the network!

Initial work in studying genetic networks conceptualizes two types of communication paths (Somogyi and Sniegoski 1996). The first consists of proximal paths operating through "cis regions" and "trans elements." Cis are control regions of DNA proximal to gene coding sequences and trans elements are gene products that regulate by interacting with cis regions. The second type of communication pathway is composed of extended paths involving protein-protein and protein-signaling factor interactions governing intra- and extra-cellular communication. Genes encoding the participating proteins control this communication.

Our starting point is very simple; we will look for clusters in gene expression time series data. When the output patterns of different genes are very similar, there is hope that they are a part of a constellation of communicating genes, receiving similar control signals. Again, this is only a starting point. Since some genes turn other genes off, things get complicated quickly. Cluster measures such as mutual information provide clues about additional members of the constellation. (Mutual information is also referred to as the rate of transmission, and is related to conditional entropy.) As this is work in process, we welcome insights into how to proceed in the face of our limited understanding and the apparent com-

plexity of genetic networks. In the next section we provide more details about the nature of our gene expression data.

2. The Observations and Clustering Methods

As mentioned above, we identified 112 genes involved in the development of a rat's spinal cord. This collection includes genes deemed important for the development of the central nervous system: neurotransmitter receptors and metabolizing enzymes, intracellular signaling proteins, peptide factors and their receptors and growth factors. Genes for marker proteins were also selected so that we can associate gene patterns with cell differentiation.

The observations on each gene constitute a time series of length nine. The nine developmental times studied were gestation days 11, 13, 15, 18, 21; and after birth days 0, 7, 14, and 90 (adult). Each point in the series is a value between zero and one. When a gene is not functioning the value is zero, and when it is maximally functioning the value is one. The observations themselves are obtained through a process known as RT-PCR, reverse transcription-polymerase chain reaction (see Somogyi et al1995). The values are determined from digital records of gel images and are actually means of triplicate observations. The variability of the triplicates is very small and not shown in the graphs below. Scaling forces the means to range from zero to one.

Figure 1 (page 27) shows the gene expression patterns by functional groups and gene sequence families. The group names appear to the left of the clusters of panels. The groups are members of four general functional categories described later in Figure 4. Since our focus here is on graphics, we will not describe the gene families.

The design of Figure 1 makes heavy use of perceptual grouping and warrants some comment. The scale for the axes appears in the top left panel. Genes in the same functional group appear in a consecutive grouping of panels. Each panel within a gene function group shows the times series for four or fewer genes. The representation is a parallel coordinates plot (see Inselberg 1985 and Wegman 1990) with omitted axes.

Each times series in a panel has its own color. While there is overplotting, the reader can quickly infer values for overplotted points. The color key and corresponding gene label appears at the right of the panel. The color has no meaning other than to serve as a link (see also Carr and Pierson 1996). One can additionally sort the rows of the key by values for the last time period. This positional linking makes more lines run into the rectangles of their own color and linking becomes trivial. However, the current example emphasizes reading labels in order and sorting would scramble the order.

In Figure 1 the plotting order for color is consistent in all panels: cyan, green, orange, and red. We use graphical ordering and plot from the bottom up within each functional group. This might be argued since there is a clash of conventions. Graph reading is bottom up while table reading is top down. Figure 1 is much like a table so there is ambiguity about which convention to apply. Note that with four genes in a panel, the red line, which should appear closest based on wavelength considerations, plots on top. The color selection also makes red the darkest color on a lightness scale and hence it contrasts the most against the light background. The graph convention is slightly advantageous because red appears on top in the panel and at the top of the color key.

The left to right sequence of gene functional group name, panels, and then gene names can also be argued. The task can motivate a different order. If communicating membership in functional groups were much more important than looking at the time series within function groups, then putting the functional group names and gene names together would be the logical design. Putting the text in one place has merit in its own right. However, putting the gene names on the right panels allows the names to be left aligned and to be uniformly close to the key and panel. Since finding the names within the function groups is still easy, we show this variation.

The selection of four genes per panel follows Kosslyn's (1994) advice for creating small perceptual groups. The apparent simplicity of the panels deteriorates quickly as number of time series in each panel increases. Figure 1 appears simple while showing the thousand means in this data set. The four time series per panel design has many applications. In landscape orientation and without the two column format, the design readily accommodates the much longer time series that occur in manufacturing and other applications.

Sorting the time series can make the plot appear simpler (see Carr and Olsen 1996). However, data ordering can serve other purposes. In Figure 1 we ordered the functional groups based on page layout considerations. We kept the provided label ordering within functional groups, because that simplified finding a specific gene within a function group. A time-series sorted version would be interesting. In an interactive setting (see Carr, Valliant and Rope 1996) one might try visual clustering by dragging and dropping time series into different panels. An automated approach can use clustering as a nominal basis for sorting.

In this case my co-authors came to me (Dan) with FITCH (Felsenstein 1993) clustering results. FITCH is an n4 algorithm and produces graphics like that in Figure 2. They had decided that their Euclidean distance clustering was better if they included the differences between consecutive observations in the time series. In other words the input vectors were of length 17, 9 time series values plus 8 differences. This is equivalent to using a weighted distance that emphasizes the seven internal points of the time series. The co- authors note that another option is to add slopes based on the actual spacing in days. They also used mutual information clustering.

Like many statisticians, I am aware of hierarchical clustering algorithms, maximum likelihood clustering, and refinement of clusters using the K-means algorithm. However, I am far from being an expert. I had never heard of FITCH nor mutual information clustering. My co-authors gave me every opportunity to recommend a clustering algorithm that would provide the truth, or one that all scientists would recognize the best of the available choices, but I declined. My early participation was simply to help them look at the data and the results of clustering.

3. Cluster Plots and Stereo Plot Construction

Figure 2 shows a cluster tree produced by FITCH. The follow-the-line distance between points approximates the multivariate interpoint distance. FITCH minimizes a measure of stress that differs somewhat from the measure minimized in traditional 2-D nonmetric multidimensional scaling (MDS). Figure 2 shows the average time series profile for each of the six resulting clusters. The labels for the clusters derive from the profiles and do not necessarily have any deep meaning. Figure 2 is good in that it gets all the labels into the plot and provides a feel for clusters and subclusters. The extra freedom provided by using connecting-line length rather than direct interpoint distance should allow significant reduction in any measure of stress. In this sense Fitch cluster tree views should be better than MDS or first-two principal component views. However, tracking lines for each pair in the cluster to assess interpoint distance is a complicated visual operation. How can we judge the clustering if we can not easily judge interpoint distances? A first reaction is to stick with views that represent interpoint point distances directly even though the distance approximations are not as good.

Conceptually, higher-dimensional plots reduce the measure of stress and hence provide a better representation of interpoint distances than low dimensional plots. In practice the analyst must translate the differences between encoded multivariate points into interpoint distances. The merits of a higher-dimensional representations can be more than counterbalanced by the difficulty and inaccuracy of the decoding process. In fact a definitive test for multivariate representations should be how well the user can assess the distance between two points and the ratio of two such distances. The position of Carr et al 1986 is that 3-D stereo plots (and possibly 4-D stereo ray glyph plots) allow quick distance judgements that are good enough to be worthwhile. In the principal components context, if a third or fourth component adds little to the percent of variability explained, then one might get by focusing on a 2-D plot. However, a 3-D or 4-D plot is often a better starting point. Those busy interpreting a 2-D plot can seem naive when important structure is obvious in a 3-D view. Of course naiveté is relative. Those that can incorporate and understand even more information in the graphics have an advantage.

Figure 3 (page 28) is a side-by side stereo plot that distinguishes the six groups using color and symbol. Carr (1990) discusses stereo projections. The slightly rotated view in Figure 3 seems to help image fusion over a directly facing view. Many people can learn to fuse side-by-side images without the aid of a view device. This learned skill involves the decoupling of eye-convergence and lens focusing that normally work together in a process called accommodation. Proper fusion results in the square dot appearing in the back left corner of the plot frame.

The plot axes in Figure 3 are the first three principal components of the 17 variables. The three coordinates capture 65 percent of the variability. The figure uses global scaling for the three coordinates and the plot frame reflects the range of the principal components. That is, the x-axis represents the first principal component and the frame is largest in the x direction. The y-axis represents the second principal component. Representing the third principal component with stereo depth reduces overplotting and the complications of looking through many layers of data. The analyst should view the stereo plot from the correct distance to perceive interpoint distances properly. The assignment of variables to the axes gives an important clue. If the frame appears deeper than the frame is tall, then the analyst is too far away.

The cluster and color pairing are: Constant = red, Wave–1 = orange, Wave–2 = green, Wave–3 = cyan, Wave–4 = magenta, and Other = black. A minimal spacing tree based on the three axes connects the points in each cluster. This helps to constrain visual traversal paths in

Figure 2. A cluster tree using follow-the-line distance.

repeated viewing, and the perceptual grouping makes the plot look simpler (Carr et al 1986). Given fusion, we see plausible clusters (red triangles and green octagons) at a glance. The scale for the red and green clusters raises serious doubts about some of the other clusters such as the orange squares and magenta x's.

Another way to look at cluster results is to use local averages of the times series rather than principal components. The averaging of adjacent times series values is a standard dimension reduction technique. The advantage is that when patterns appear, interpretation often less complicated than for a principal components view. (A disadvantage is that researchers don't like loosing temporal resolution especially when the experiments were laborious.) After grouping the nine values into sets of three and averaging, we used the shuttering glass stereo in ExplorN (Carr, Wegman, and Luo 1997) to look at the resulting three coordinates. (ExplorN also supports touring in parallel coordinate and scatterplot matrix views of up to 20 dimensions.) Color and ray angle represented the cluster membership. The clusters were plausibly coherent just as they are in Figure 3. However there was one bad exception in our initial look at the clustering. The data for that gene had a transcription error. The principal component view in Figure 3 shows the corrected data. Before assessing clusters further, we pause for more comments on stereo views and plot construction.

4. Stereo Viewing and Plot Production

Small side-by-side stereo plots are less than optimal. A good stereo viewer with appropriate mirrors and lenses allows use of much larger left-eye and right-eye plots. Stereo workstations using shuttering eye-glasses work quite well, although they sacrifice a bit in terms of spatial and brightness resolution. Rotation of points provides a good depth cue, motion parallax. However, rapidly rotating plots are hard to study. In our use of ExplorN we found very slowly rotating stereo views to be a desirable compromise.

The move from the workstation stereo graphics to printed side-by-side views raises the issue of color overplotting inconsistencies. In non-translucent stereo mode, our SGI graphics workstation uses a z-buffer methods to make sure that whatever is closest to the viewer plots on top. One could utilize the workstation graphics by accessing the separate eye views and copying the low resolution bit maps to a high resolution printer. For small side-by-side views the size reduction ameliorates the problem of limited resolution in workstation views. For the graphics here we sought to use more conventional software. Production of Figure 3 is straightforward using high resolution black lines. However, color inconsistencies arise using conventional vector graphics. The wrong color overplots when drawing a distant line of one color after drawing a close line of a different color.

The partial solution used to construct Figure 3, broke the line segments into a sequence of short line segments based on a large number of depth planes. The algorithm used the closest of the short segment endpoints as the measure of the segment's depth. The algorithm then sorted both points and lines back to front before plotting. This procedure, while computationally tedious, corrects the problem except for overplotting of segments and points at almost identical depth.

5. Cluster Interpretation and Assessment

In an unsupervised clustering problem, one hopes to use corroborating scientific information as well as cluster tightness to assess the clustering. Here the genes tyrosine hydroxylase (Th), insulin 1 (Ins1) and insulin-like growth factor II (IFGII), appear as a tight subgroup in the Wave–1 cluster. They turn out to be located on the same human cytogenetic band (11p15.5) and are close together on mouse chromosome 7 (see Mouse Genome Database). This suggests the genes are regulated in parallel due to their close proximity on the chromosomes.

Figure 4 shows the residuals from the cluster means by gene functional group and cluster. (Using a different scale, one can also show the panel means in row and column margins.) Both Waves–2 and 3 are notably confined to neurotransmitter signaling. Wave–4 cluster genes primarily belong to several functional families. Genes showing largely constant expression (the Constant group) originate from diverse families but strictly exclude the neurotransmitter signaling and neuroglial markers. The variability in Figure 4 raises concern about the adequacy of the clustering and the stability of the clusters variations when using more data or other algorithms. It seems doubtful that all of Wave–1 is just one constellation of genes. With more data, Wave–1 may break into several defensible clusters. Currently the subcluster indicated above could begin to define a constellation. The co-location of genes on a chromosome is reasonable confirmation.

6. Cluster Comparison

As indicated above my co-authors also brought results from a mutual information clustering algorithm. Again they selected six classes. A natural step is to compare the clustering.

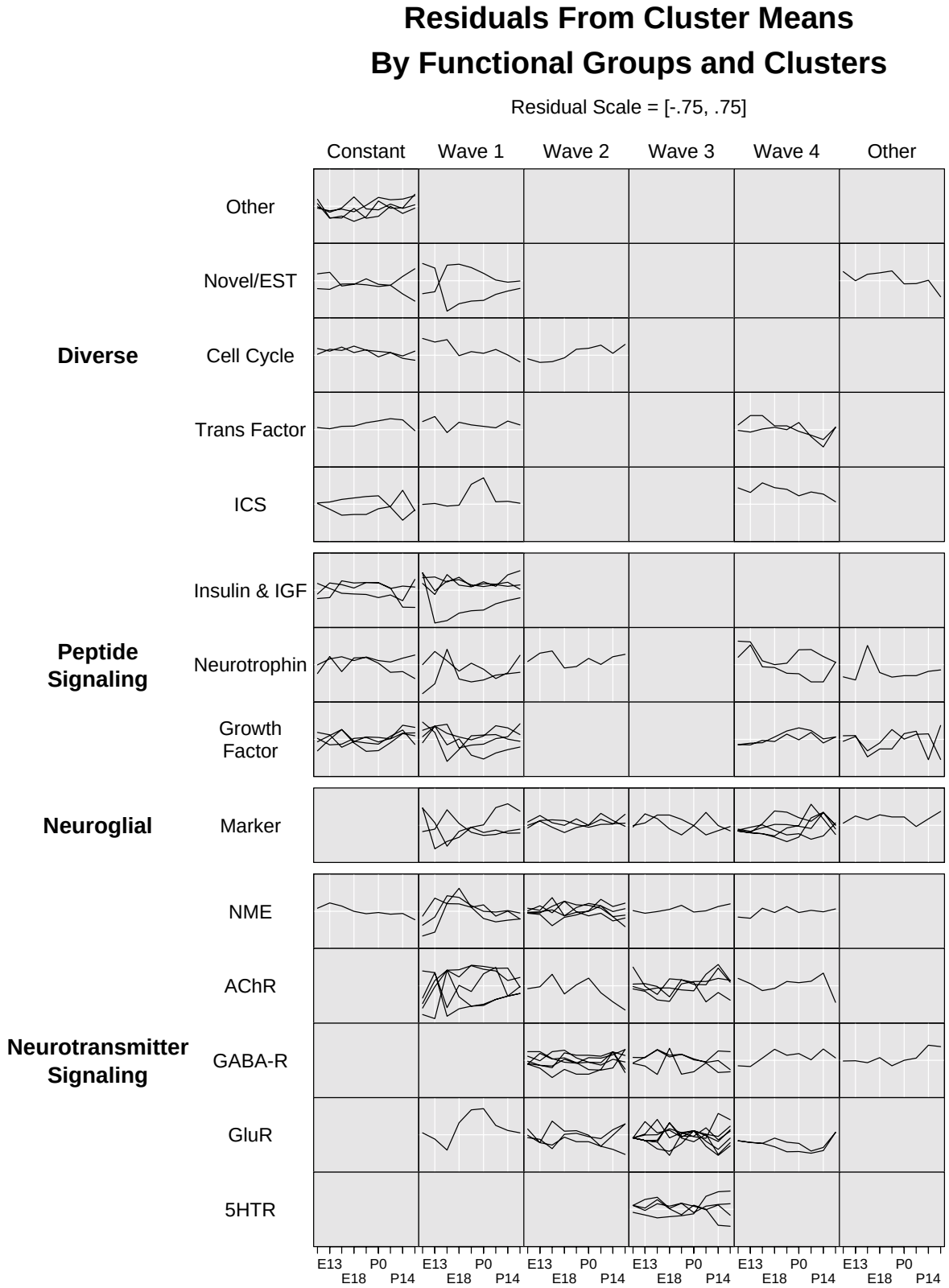

Figure 4. Multivariate residuals in a two-way layout.

Comparison graphics can take many forms. One template is a two-way panel layout. The rows are class membership for one algorithm and the columns are the class membership for other. Each panel contains the corresponding times series (if any). After suitable rearrangement of rows and columns, a strong diagonal pattern would indicate effective similarity of the clustering algorithms. The shape of the series in off diagonal panels provides insight into algorithm differences. One can augment such graphics. In the current cases, there would be a 6 x 6 panel layout and color could indicate membership in one of the four major functional groups diverse, peptide signaling, neuroglial and neurotransmitter signaling. Another variation incorporates the information using a (4 x 6) x 6 layout. The 24 rows of panels result from crossing the major functional groups with one of the classifications. The potential variations using conditioning and color to incorporate additional information are numerous.

Figure 5 (page 28) shows a cluster comparison approach based on parallel coordinates. Here the time series are omitted and the plot emphasizes three classifications: gene function, Euclidean distance clusters and mutual information clusters. The gene function axis appears at both the top and bottom of the plot. The regular spacing between a small number of crossing points distinguishes the classification axes.

The Figure 5 design introduces two unique-case axes between each classification axis. Every case (gene) has its own unique plotting position on these axes. With only a hundred or so lines, all lines are visibly distinct. George's idea behind using two such axes was to confine messy line crossing to the region between the unique-id axes. This creates regular patterns of lines reaching the classification axes.

The choice of color in the figure serves two purposes. First it collapses the 13 gene function groups in the four functional families. Second the color selection purposely calls attention to the peptide signaling (high contrast yellow) and down plays the distinction between neuroglial and neurotransmitter signaling.

The precursors to Figure 5 raised an interesting sorting issue. The crossing lines made the plots look complicated. The challenge then is to order classifications, order classes within each classification, subclasses within nested classifications, and genes within (sub)classes to minimize line crossings. An all permutations approach works for small problems. Unfortunately the combinatorics become overwhelming in a general table setting.

For two classifications there is a convenient approximate approach. Wegman (1990) observes that few crossings correspond to high correlations. Kendall and Stuart (1979) describe an eigenvector scoring approach for categorical data that will maximize the correlation for two classification. This provides a basis for ordering the classes within each classification. Unfortunately, we have not been able to generalize this approach to three variables and fear that the general case may be np-complete.

Some find the string art in Figure 5 appealing but for others the plot is still too complicated. The advantage of the small panel approach described previous is that it shows the times series in addition to the classification. We present Figure 5 because it illustrates one way of converting classification tables into graphs and because it raises a sorting challenge.

7. Closing Remarks

Alternatives and extensions to the above templates for viewing clustering results are numerous. Hierarchical cluster tree views are common. Visually connecting the cluster trees to the multivariate data can help provide insights about the clustering. Buja, Cook, and Swayne (1996) used color brushing in XGobi to link cluster tree branches to other scatterplot views of data. Such interactivity facilitates following the visual clues provided by graphics. The idea of joining multiple window brushing capability with adaptable thoughtfully-designed multiple panel plots has occurred to many but implementations have been slow to appear.

Many multiple panel-designs scale to much large sample sizes. Density methods apply to parallel coordinate plots (Wegman and Luo 1997). The data need not necessarily be time series. Plots like Figure 4 have many extensions.

As usual, S-PlusT functions and scripts for producing the graphics are available via anonymous ftp to `galaxy.gmu.edu`. Change directory to `pub/dcarr/newsletter/gene`. In contrast to the past, the data provided is artificial. The real data will be substituted when my co-authors have published in a refereed journal.

Splus users may find the matrix layout functions of particular interest. Currently there is a Bureau of Labor Statistics technical report describing the functions and selected connections to TrellisT graphics. Contact Dan for a copy.

Those with Silicon Graphics workstations may be interested in ExplorN. A tar file containing an executable and sample data sets is available. Conversion to OpenGL and available on other OpenGl compatible platforms may happen later in the year.

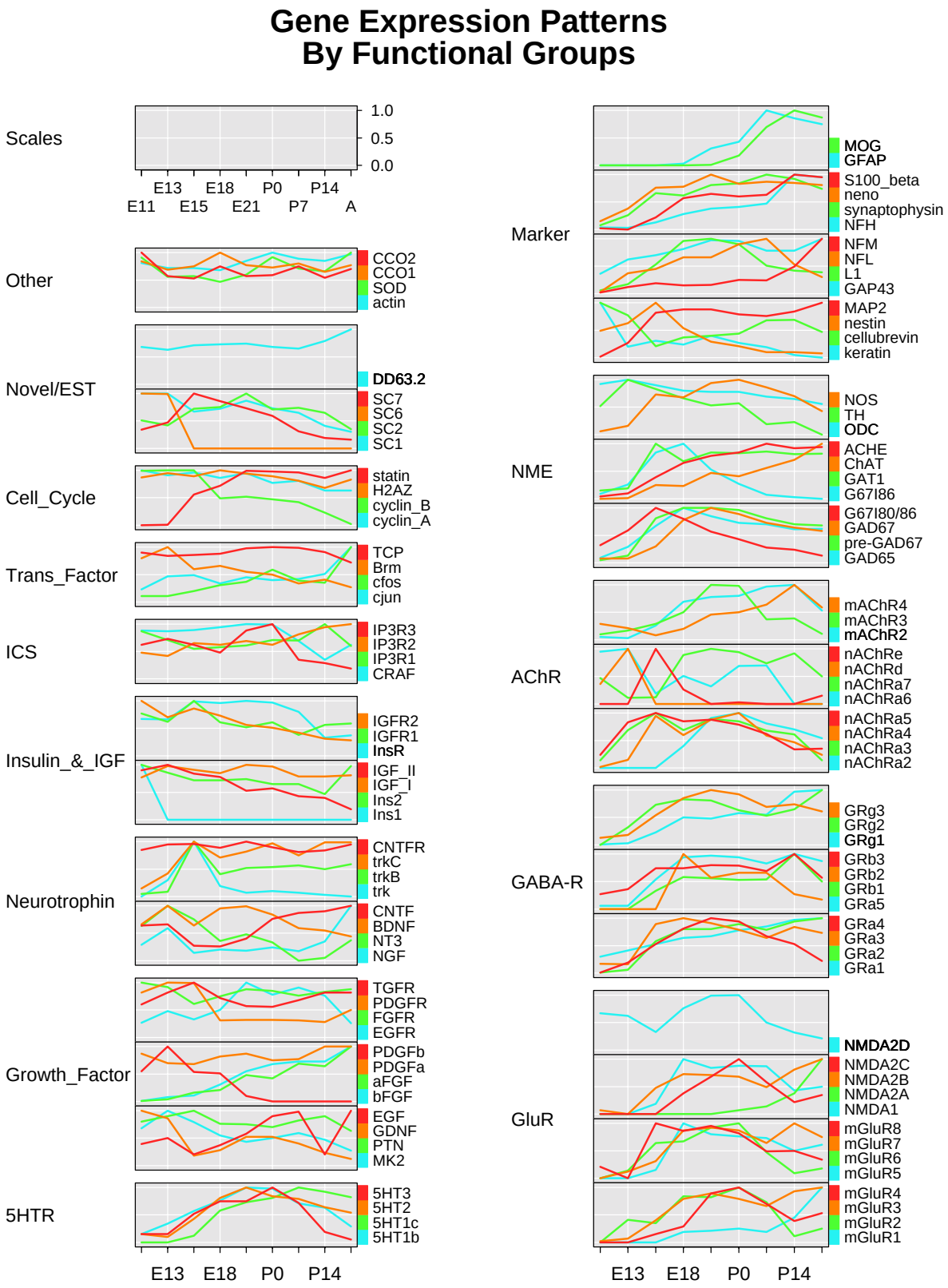

Figure 1. Labeled time series with controlled overplotting. Here, the colors have no meaning, but serve only as a linking device within each panel.

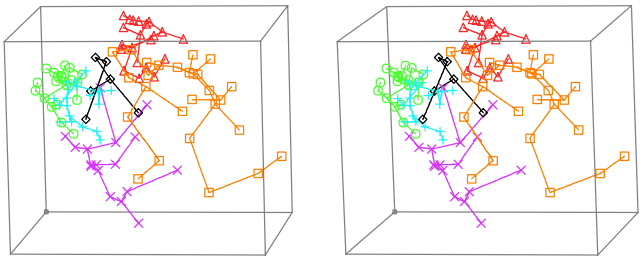

Figure 3. Stereo pairs with careful color overplotting. In this case, colors correspond to clusters (see the text for a detailed description).

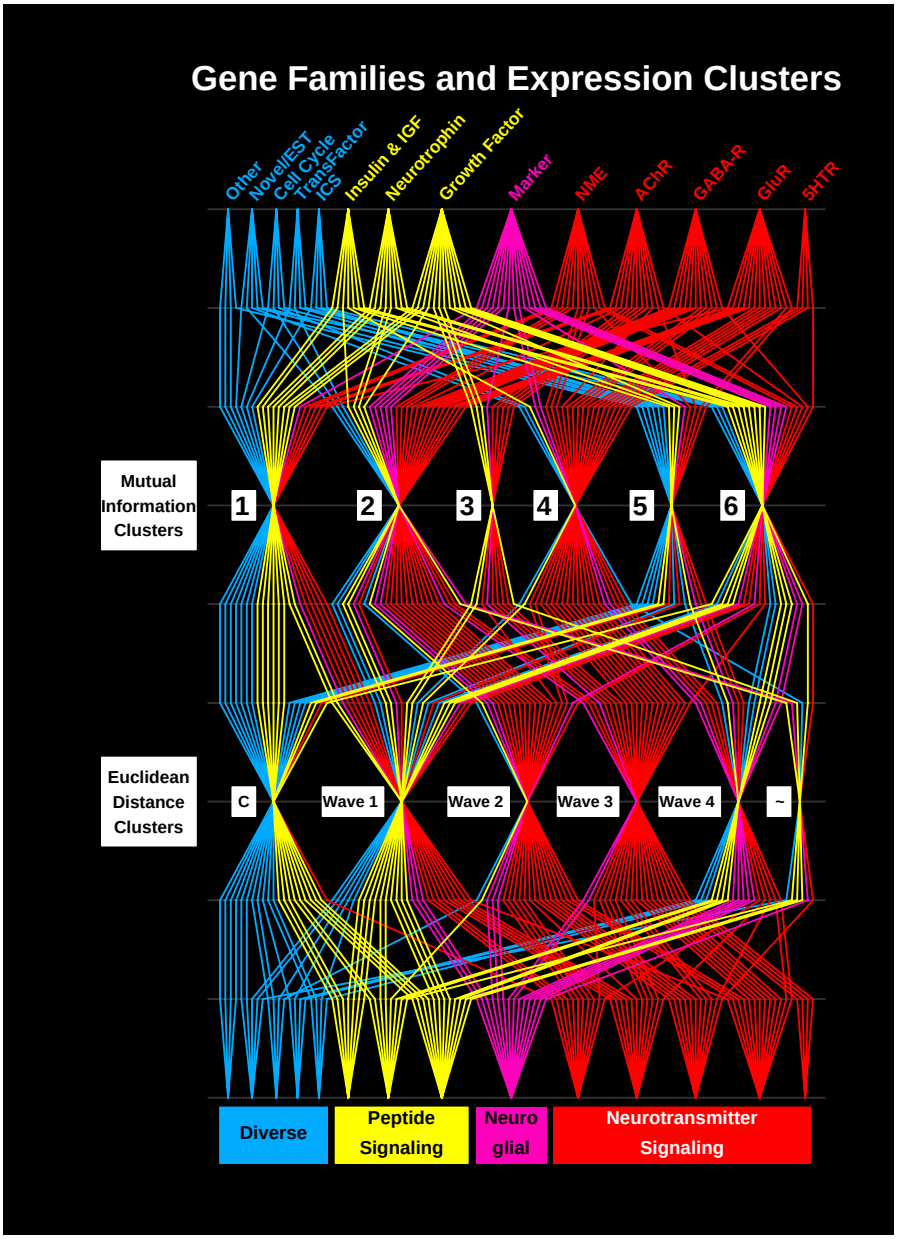

Figure 5. A sorted parallel coordinates view of multiway classified cases.

As always, the authors are open to gentle constructive suggestions. Comments on graphics are best addressed to Dan and comments on the study or on genetic networks are best addressed to Roland and George.

Acknowledgements

The authors thank Drs. Xiling Wen and Stefanie Fuhrman, Laboratory of Neurophysiology, NINDS, for their outstanding work in the experimental data acquisition and analysis. Thanks also go to Andreas Buja for a discussion on ordering classes.

References

Buja, A., Cook, D. and Swayne, D. F. (1996), "Interactive High-Dimensional Data Visualization," *Journal of Computational and Graphical Statistics*, 5(1), 78–99.

Carr, D. B. (1993), "Production of Stereoscopic Displays for Data Analysis," *Statistical Computing & Graphics Newsletter*, 4(1), 2–7.

Carr, D. B. and Olsen, A. R. (1996), "Simplifying Visual Appearance By Sorting: An Example Using 159 AVHRR Classes," *Statistical Computing & Graphics Newsletter*, 7(1), 10–16.

Carr, D. B., Nicholson, W. L., Littlefield, R. J. and D. L. Hall (1986), "Interactive Color Display Methods for Multivariate Data," *Statistical Image Processing and Graphics*, eds. E. J. Wegman and D. J. DePriest, Marcel Dekker, New York, pp. 215–250.

Carr, D. B. and Pierson, S. M. (1996), "Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps," *Statistical Computing & Graphics Newsletter*, 7(3), 16–23.

Carr, D. B., Valliant, R. and Rope, D. (1996), "Plot Interpretation and Information Webs: A Time-Series Example From the Bureau of Labor Statistics," *Statistical Computing & Graphics Newsletter*, 7(2), 19–26.

Carr, D. B., Wegman, E. J., and Luo, Q. (1997), "ExplorN: Design Considerations Past and Present," Center for Computation Statistics Technical Report No. 137, George Mason University, Fairfax, Va. 22030.

Felsenstein, J. (1993), *PHYLIP (Phylogeny Inference Package), version 3.5c*, distributed by the author, Department of Genetics, University of Washington, Seattle.

Galli-Taliadoros, L.A., Sedgwick, J. D., Wood, S. A., and Korner, H. (1995), *J. Immunol. Methods*, 181, 1–15.

Inselberg, A. (1985), "The Plane with Parallel Coordinates," *The Visual Computer*, 1, 69–96.

Kauffman S. A. (1993), *The Origins of Order, Self-Organization and Selection in Evolution*, Oxford University Press, London.

Kendall,M. and Stuart, A. (1979), *The Advanced Theory of Statistics, Vol. 2, Inference and Relationship*, Charles Griffin & Company, London.

Kosslyn, S. M. (1994), *Elements of Graph Design*, W. H. Freeman and Company, New York.

Mouse Genome Database, Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine.
`http://www.informatics.jax.org/`

Somogyi R. and C. A. Sniegoski (1996), "Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation," *Complexity* 1(6), 45–63.

Somogyi R, Wen, X., Ma, W., and Barker, J. L. (1995), "Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord," *J. Neurosci*, 15, 2575–2591.

Wegman, E. J. (1990), "Hyperdimensional Analysis Using Parallel Coordinates," *Journal of the American Statistical Association*, 85, 664–675.

Wegman, E. J. and Luo, Q. (1997), "High-Dimensional Clustering Using Parallel Coordinates and the Grand Tour," *Computing Science and Statistics*, 28, 361–368.

Dan Carr
Institute for Computational Sciences and Informatics
George Mason University
`dcarr@voxel.galaxy.gmu.edu`

Roland Somogyi
National Institute of Neurological Disorders and Stroke
National Institutes of Health
`rolands@helix.nih.gov`

George Michaels
Institute for Computational Sciences and Informatics
George Mason University
`gmichael@gmu.edu`