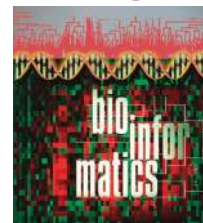


Interactively Exploring Hierarchical Clustering Results



The Hierarchical Clustering Explorer provides a dendrogram and color mosaic linked to two-dimensional scattergrams, a variety of visualization options, and dynamic query controls for use in genomic microarray data analysis.

Jinwook Seo
Ben
Shneiderman
University
of Maryland,
College Park

Molecular biologists and geneticists seek to understand the function of genes, including the more than 6,000 genes in the yeast genome and the estimated 40,000 genes in the human genome. Recently developed for genome analysis, DNA microarrays—also known as gene arrays or gene chips—usually consist of glass or nylon substrates that measure 1 × 3 inches or smaller. These chips contain specific DNA gene samples spotted in an array by a robotic printing device. Researchers spread fluorescently labeled messenger RNA (mRNA) from an experimental condition onto the DNA gene samples in the array. This mRNA binds (hybridizes) strongly with some DNA gene samples and weakly with others. Finally, a laser scans the array and sensors detect the fluorescence levels, indicating the strength with which the sample expresses each gene.

Experimental conditions can include types of cancers, diseased organisms, or normal tissues. Microarray experiments typically have from 100 to 20,000 DNA gene samples and from two to 80 experimental conditions. These experiments produce data sets containing the profiles that include expression levels for each DNA gene sample under each experimental condition.

Researchers often use mathematical clustering methods to discover interesting patterns in these large data sets.¹⁻³ However, several limitations hinder biologists from recognizing important patterns, such as groups of genes with similar profiles, which might indicate similar function. The data volume

makes it impossible to display a large microarray experiment—on one screen.

Researchers also struggle to understand the implications of a specific clustering result. Because the clusters occupy a high-dimensional space and involve so many experimental conditions, researchers find it difficult to view patterns on a 2D or even a 3D display. Further, data can contain hundreds of variously sized clusters, which makes spotting the meaningful clusters a challenge, especially when using a static display. Users need an efficient interactive visualization tool to facilitate pattern extraction from microarray data sets.

Hierarchical clustering has been shown to be effective in microarray data analysis for identifying genes with similar profiles and thus possibly with similar functions. This approach finds the pair of genes with the most similar expression profiles and iteratively builds a hierarchy by pairing genes (or existing clusters of genes) that are most similar.

This hierarchy is usually presented as a dendrogram—a branching diagram that represents the hierarchy of clusters based on degree of similarity. As Figure 1 shows, the binary trees lead down to the leaves, which typically appear at the bottom as a sequence of red and green tiles in a mosaic. Each tile represents an expression level for one of the experimental conditions. The distance of the binary tree's joining points from the root indicates the similarity of subtrees—highly similar nodes or subtrees have joining points farther from the root.

When viewing clusters, researchers use overviews to identify hot spots and understand the distribu-

tion of data. Hot spots are distinctive regions of interest that indicate high or low expression levels. However, screen resolution limitations hamper visualization of large data sets on commonly used displays, which measure only 1,600 pixels across. Even limiting each item to a single pixel means that for data sets larger than 1,600 points, the corresponding dendrogram and color mosaic do not fit onto a single screen.

HIERARCHICAL CLUSTERING EXPLORER

Adding interactive exploration features to hierarchical-clustering visualization tools is needed for the exploration of large data sets. To meet this need, we developed the Hierarchical Clustering Explorer (HCE), which integrates four interactive features:

- overview of the entire data set, coupled with a detail view so that high-level patterns and hot spots can be easily found and examined;
- dynamic query controls^{4,5} that let users eliminate uninteresting clusters and show the interesting clusters more clearly;
- coordinated displays that forge a bidirectional link from the overview mosaic to two-dimensional scattergrams; and
- cluster comparisons to let researchers see how different algorithms cluster the genes.

Our work meshes with the current trend to take the substantial progress in data mining algorithms and give users more than just a printout. With novel information visualization techniques,⁶ users can now control the processes and interact with the results. For example, recent decision tree packages let users manipulate incoming data and the rules generated, then examine the results with color- and size-coded visualizations. The capacity to interact and explore lets domain experts apply their knowledge by quickly testing hypotheses and performing exploratory data analysis. We hope to provide similar features to users of hierarchical clustering algorithms.

OVERVIEW IN A LIMITED SCREEN SPACE

To accommodate large data sets, HCE provides a compressed overview based on replacing dendrogram leaves with the average values of adjacent leaves. This view shows the entire hierarchy, at the cost of some lost detail at the leaves, as Figure 2 shows. A second overview allocates two pixels per item, but requires scrolling to view all items. In this scrolling overview, users can adjust the level of detail shown in the overview by moving the slider

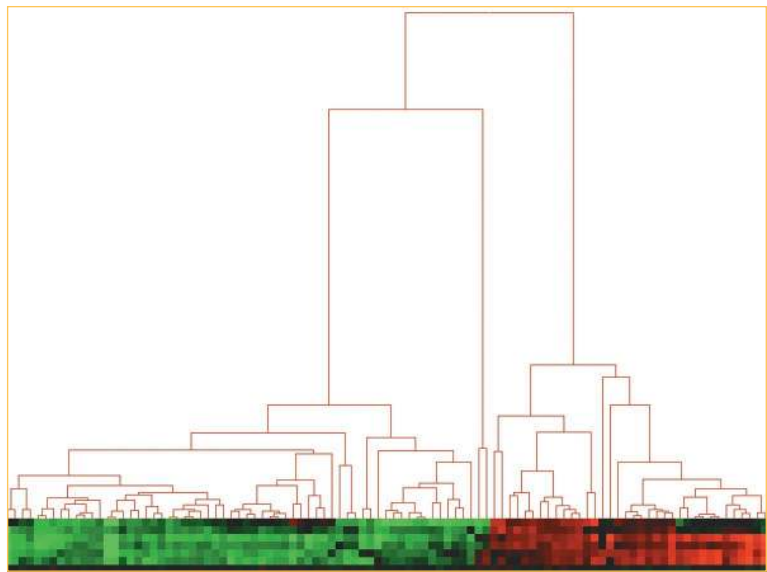


Figure 1. Sample dendrogram—a binary tree in which subtrees are each a cluster and the leaves are individual genes. The distance from the root to a subtree indicates the similarity of subtrees—highly similar nodes or subtrees have joining points farther from the root.

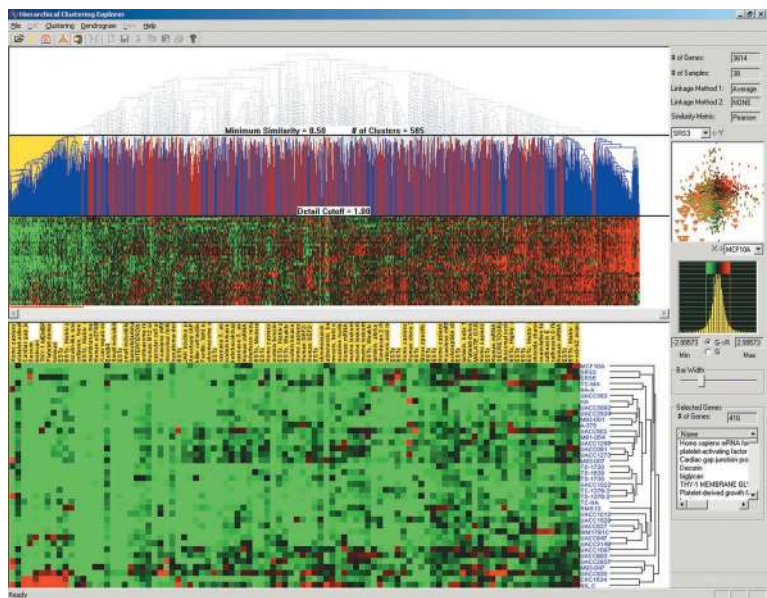


Figure 2. Hierarchical Clustering Explorer's compressed overview. The melanoma gene expression profile contains 3,614 genes and 38 experimental conditions. This view shows the entire hierarchy in one screen by replacing leaves with the average values of adjacent leaves. The detail information of a selected cluster, shown as a yellow highlight in the upper left, appears below the overview, together with the gene names and the other dendrogram at the lower right, by clustering the 38 experimental conditions.

to change item widths from two to 10 pixels. With either overview, users can click on a cluster and view the detailed information at the bottom of the display, which also includes the item names.

Researchers can examine the color mosaic to identify hot spots and understand the data distribution. In general, HCE displays a dendrogram with

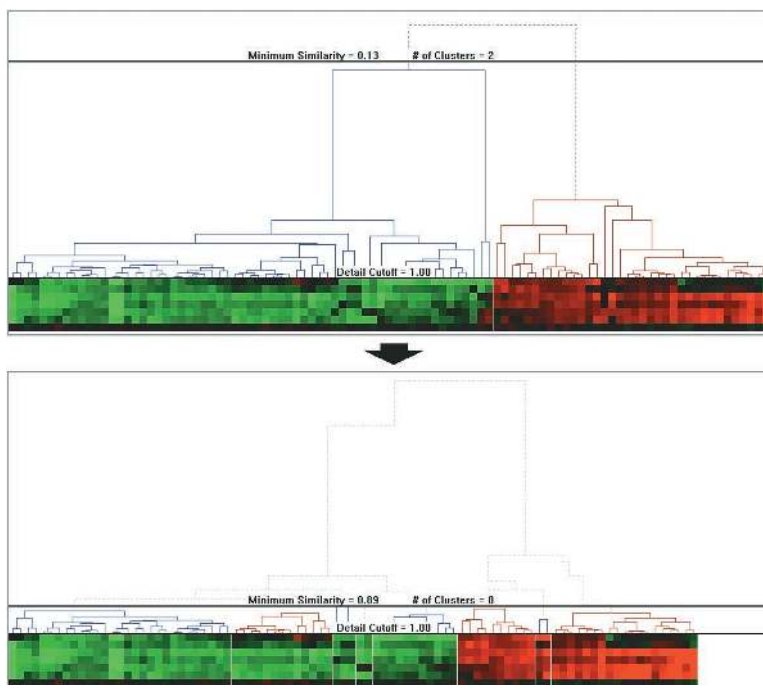


Figure 3. Using the minimum similarity bar. The bar's y coordinate determines the minimum similarity value. Users can drag the bar down to filter out items distant from a particular cluster. In this example, the minimum similarity values changed from 0.13 to 0.89 as the user separated two large clusters into eight small clusters.

a color mosaic at the leaves to show the underlying graphical pattern. The mosaic colors each tile on the basis of measured fluorescence ratio, which denotes the gene's expression level.⁷

The gene expression profile data consists of the ratio or relative amount of each specific gene in the two mRNA or DNA samples—corresponding to the normal and test conditions. Some data sets, including melanoma and yeast mutants, are more complicated. These data show expression levels for several mutants or cancer cell lines relative to a control condition. Researchers commonly use the log of ratio values and display the result using a 2D colored mosaic.

The HCE control panel, to the right of the dendrogram visualization in Figure 2, shows the data histogram by expression level. User controls help viewers see subtle differences in the ranges of interest. For skewed data distributions, this feature helps prevent large areas of all green or red, which would indicate all low or high gene expressions, respectively. Users can change the color mapping by adjusting the range of the color stripe displayed over the histogram. Because they can instantly see the result of a new color mapping on the display, users can identify the proper mapping for the data set.

DYNAMIC QUERY CONTROLS

Once users find a closely related group of genes, they can infer that an unknown gene clustered with a known gene may have a similar biological function to the known gene.¹ HCE users select a data set,

apply their desired clustering algorithm, then begin the process of understanding the output. First, they adjust the color mapping to get a clearer presentation of similarities and differences in expression levels. They can then study the main groupings—the two high-level clusters—which may not themselves be interesting but may combine interesting subclusters. For example, a set of 800 genes may be composed of 10 or 20 interesting subgroups, so looking at a simple two-group clustering does not reveal the relevant subgroups. Currently, static dendrogram users rely on their eyes and fingers to traverse the hierarchy and identify interesting clusters.

HCE provides a dynamic query on the dendrogram in the form of a filtering bar on which the y coordinate determines the minimum similarity value, as Figure 3 shows. As users pull down the minimum similarity bar, the mosaic display splits into two, three, four, or more groups. As the bar moves further down, the system removes the items far from a cluster, but users can still see the overall dendrogram structure. As the system removes more items, the tighter clusters can be seen more easily. Users' domain knowledge guides them in determining how far to go and how many clusters to examine.

To prevent users from losing global context during dynamic filtering, HCE maintains the entire dendrogram in the background. Users can see the position of a cluster in the original data set simply by clicking on the cluster, which causes it to highlight in yellow, as Figure 4 shows. Users can easily identify each cluster by the alternating blue and red lines in the dendrogram and by the one-pixel gaps placed between clusters. HCE highlights the selected cluster with a yellow rectangle, displays the corresponding gene names in the detailed color mosaic in the lower pane, and shows the other dendrogram that the clustering produces. When executed in a Windows 2000 environment run on at least a Pentium II, HCE processes all interactions and updates almost instantaneously for a moderate-sized data set of up to 4,000 genes in 40 experimental conditions.

Researchers need a sufficiently comprehensive overview as much as they need adequate detail. The overview reveals patterns across the entire data set, which guides users to the next step in their search. One generally accepted visualization scheme starts with an overview, then lets users dynamically access detail information.⁶ Providing a data set overview while allowing detailed analysis of a selected part helps users place that part in a general context during their search.

However, too much detail can be a problem. If the set of 800 genes fits neatly into between six and

16 clusters, then seeing details of 800 genes below the cluster level can be confusing. To avoid this effect, HCE lets users represent highly similar items with the same coloring. Users can reduce the detail level by dragging the detail cutoff bar higher. The system renders all subtrees below the bar using the average of the leaf node values belonging to the subtree. Users can thus hide the detail below the bar to help them concentrate on more global structures.

Especially for a large dendrogram, the detail cutoff bar helps visually present cluster structures that satisfy the current minimum similarity level. Once users find an interesting cluster in the dendrogram, they can restore detail by dragging the detail cutoff bar downward.

COORDINATED DISPLAYS

The hierarchy shown in the dendrogram, and the linear presentation in the color mosaic, help reveal clusters that represent important patterns. However, they can hide some aspects of the data's high-dimensional nature. High-dimensional displays such as parallel coordinates⁸ and other novel techniques⁹ could be useful, but many users have difficulty comprehending these visualizations. Even three-dimensional displays can be problematic because of the disorientation brought on by the cognitive burden of navigation.⁶

The x and y axes of two-dimensional scattergrams limit them to two variables at a time, but most users can readily understand them. Further, without the distraction of operating 3D navigation controls, users can concentrate on finding patterns in the data.

HCE thus contains a scattergram view in which users first select any two dimensions for the x and y axes, representing two of the conditions. Then they use a rubber rectangle to sweep out an area on the scattergram, producing orange triangles that highlight the items in the scattergram and the related items in the overview color mosaic, as Figure 5 shows. Marking the selected items with similar shapes and orange color facilitates perceptual integration. Often, the neighboring items in the scattergram appear in the same cluster in the dendrogram, but items contained in other clusters can serve as important indicators to domain experts.

This coordination of displays is bidirectional. Users can click on the overview color mosaic to select a cluster, which highlights the related items in the scattergram. Seeing the distribution of items in the scattergram confirms the clustering and often produces intriguing patterns that invite further

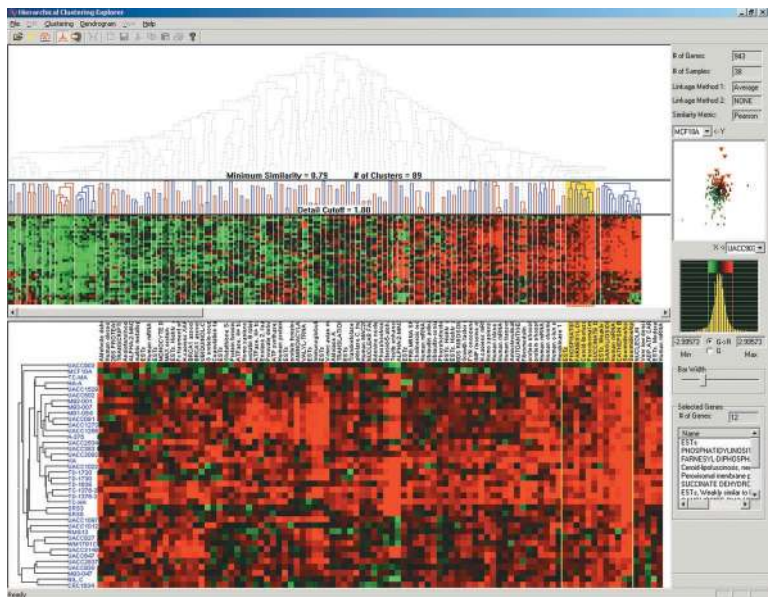


Figure 4. Highlighting a cluster. The alternating blue and red lines just below the minimum similarity bar and the one-pixel white gaps placed between clusters make it easy to identify each cluster. Users can select a cluster simply by clicking on it, which highlights it with a yellow rectangle. The system also highlights the corresponding gene names in the detailed color mosaic and displays the other dendrogram that clustering the data in the transposed dimension produces, as shown on the lower left side.

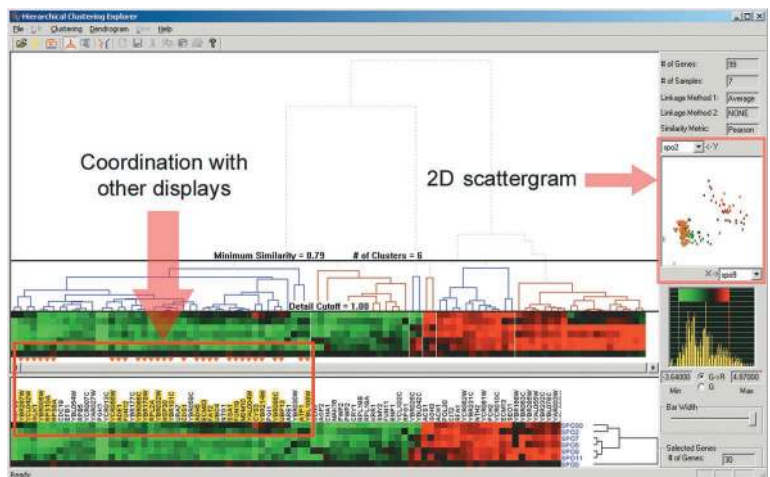


Figure 5. Two-dimensional scattergram and its coordination with other visualizations. Users can select a group of items by sweeping out a rectangular area on the scattergram. The Hierarchical Clustering Explorer simultaneously highlights the selected items in the scattergram, and the related items just below the overview color mosaic, with orange triangles.

investigation. Clusters are not always neatly circular—they often show up as odd-shaped bunches with some points embedded in nearby groups. These surprising patterns can be important to domain experts.

Coordinated windows have their problems, too. Scarce screen space must be allocated to two visualizations rather than one, and users must shift their attention back and forth rapidly. Minimizing the

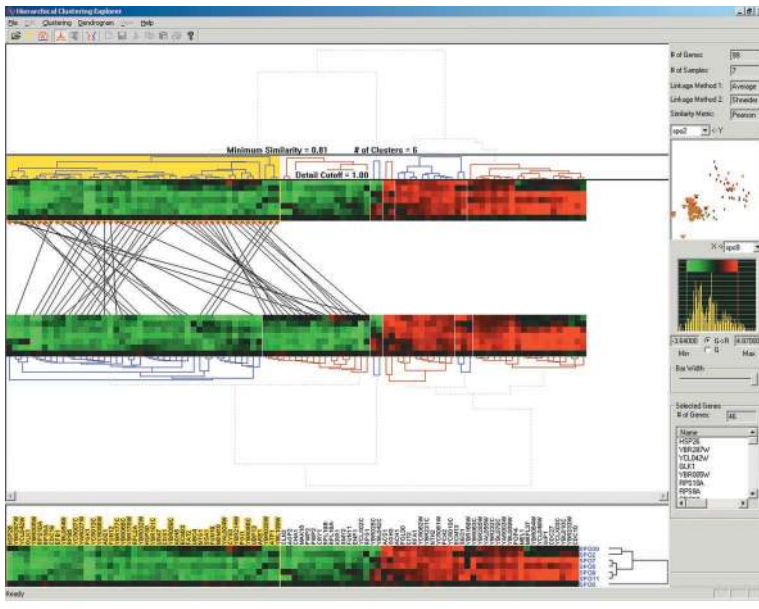


Figure 6. Cluster comparisons. Users can double-click on a specific cluster to see the mapping of each gene between the two different clustering results. The selected cluster will highlight in yellow, and lines from each item in that cluster will extend to their position in the second clustering result.

distance between visualizations, avoiding overlaps, and making rapid updates all contribute to improved human performance. After some exploration with a given scattergram, users often explore another pair of axes. Showing multiple scattergrams concurrently can help, but doing so exacerbates the screen-space management problem.

CLUSTER COMPARISONS

Often when performing a microarray data analysis, researchers know that groups of genes have similar functions. Similar genes should reside in the same cluster, and the researchers want to confirm that the hierarchical clustering algorithm finds these similar groups. They also want to see which unknown genes cluster together with known genes because this information may lead them to understand the unknown gene’s function. Reviewing the clusters in the detail color mosaic helps uncover these relationships.

The lack of a perfect clustering algorithm complicates this process, however. Because molecular biologists and other researchers have different ways of computing distances between items in a multi-dimensional data set and the similarity values between groups of items, they need a mechanism to examine and compare two clustering results.

HCE users can view the results of two hierarchical clustering algorithms on the screen at once, as Figure 6 shows. Double-clicking on a specific cluster reveals the mapping of each gene between the two different clustering results. The selected cluster will highlight in yellow, and lines from each item in that cluster will extend to their position in the second clustering result. If some genes map to dif-

ferent clusters, users can examine the genes more carefully to understand what caused the difference.

Although crisscrossing lines can cause confusion, they enable users to compare competing clustering algorithms. Showing relationships between non-proximal items is a basic problem in information visualization research. Each of the three basic methods—color coding, blinking, and drawing lines—has problems. HCE already uses color coding heavily, and blinking would add distraction to an already complex display, so drawing lines provided our best alternative.

Our biology users, excited to have this capability, spent hours using an alternate clustering algorithm to probe the clusters to see which genes had switched into other clusters. Metrics for measuring similarity and tools to highlight important changes would further improve HCE.

Another possible verification method selects a subset of the experimental conditions and does the clustering on the reduced set. It is easier to verify the correctness of a clustering method in a low dimension that involves between two and four conditions than to do so in higher dimensions that involve between five and 80 conditions. HCE users can use a dialog box to select a subset of the conditions to take part in the clustering. The resulting color mosaic has a white space between the selected conditions and the others. Users can concentrate their inspection on the selected conditions and see the clusters more clearly in the scattergram. Using different conditions to redo the clustering helps users understand the relationships among conditions and identify which of them strongly affect the outcomes.

Microarrays, sequenced genomes, and the explosion of bioinformatics research have led to astonishing leaps in our understanding of molecular biology. To date, work in these fields has focused largely on algorithmic methods for processing and manipulating vast biological data sets. These efforts have made impressive gains, but additional help may be needed.

Future improvements will likely provide users with guidance in selecting the most appropriate algorithms and metrics for identifying meaningful clusters. Hybrid approaches that combine powerful algorithms with interactive visualization tools—such as those described in the “Visualization Software for Clustering in Bioinformatics” sidebar—will join the strengths of fast processors with the detailed understanding of domain experts. We

Visualization Software for Clustering in Bioinformatics

As computing became widespread, statistical analysts quickly developed the hierarchical clustering technique.¹ Extensions included alternative ways to compute

- distances between items in a multi-dimensional data set, such as Euclidean, correlation coefficient, and Manhattan distance; and
- the similarity values between groups of items—or linkage—such as average, complete, and single.

Meanwhile, presentation refinements focused on producing effective color printouts for publication.

TreeView

Software tools for hierarchical clustering have been developed in many disciplines and incorporated into a variety of software products. The widely used TreeView (<http://www.pnas.org/cgi/content/full/95/25/14863>), developed especially for genetic research, generates a dendrogram and color mosaic. Users can get an overview and detail view by selecting a contiguous region of the mosaic, which is magnified in a second view. Because TreeView's main purpose is to produce a good image in many formats for publications, the current version does not allow direct manipulation of the visualization.

GeneMaths

Developed by Applied Maths (<http://www.applied-maths.com/ge/ge.htm>), GeneMaths displays dendrograms for gene samples and experimental condi-

tions on a single screen. Users can select a cluster by clicking a subtree's root. GeneMaths offers one of the fastest clustering algorithms and a visually appealing design. It shows only a few genes at a time, however, which makes it difficult to get an overview of the entire data set.

Spotfire DecisionSite

The Spotfire Array Explorer, now included in the DecisionSite product (<http://www.spotfire.com/>), does the hierarchical clustering, and users can view the entire green-black-red color mosaic or selected components. Users select a subtree in the dendrogram by clicking on the root of the subtree, or they can select a group of subtrees by selecting a similarity threshold. Users can coordinate scattergrams and bar charts with the dendrogram display to help understand the clustering results.

Expression Profiler

The European Bioinformatics Institute's Expression Profiler tool set does clustering, analysis, and visualization of gene expression and other genomic data (<http://ep.ebi.ac.uk/>). Among these tools, Epclust lets users do a hierarchical clustering with many different distance measures and linkage methods. When users select a dendrogram node, it shows detailed information about that node in a new window. Users can load their own data and try many kinds of hierarchical clustering algorithms. The institute's Web site also shows the great diversity in outcomes for different correlation-related distance metrics.

Clustering methods

In recent years, many clustering methods have been developed and implemented in software products. Popular methods include k -means clustering, which identifies starting points for a fixed number of clusters and then grows the region around the clusters. Recent work seeks to get beyond the limitation of spherical clusters² by using graph representations and developing clusters of arbitrary shapes, including interlocking geometries.³ All clustering methods face validity challenges.⁴ Does the clustering reflect known classifications? How many clusters are best? What should be done about outliers or intruders to clusters? What metrics could confirm or reject a perceived cluster?

References

1. S.C. Johnson, "Hierarchical Clustering Schemes," *Psychometrika* 32, 1967, pp. 241-254.
2. G. Karypis, E-H. Han, and V. Kumar, "Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling," *Computer*, Aug. 1999, pp. 68-75.
3. D. Harel and Y. Koren, "Clustering Spatial Data Using Random Walks," *Proc. 7th Int'l Conf. Knowledge Discovery and Data Mining (KDD-2001)*, ACM Press, New York, 2001, pp. 281-286.
4. G.S. Davidson, B.N. Wylie, and K.W. Boyack, "Cluster Stability and the Use of Noise in Interpretation of Clustering," *Proc. IEEE Symp. Information Visualization (InfoVis 01)*, IEEE CS Press, Los Alamitos, Calif., 2001, pp. 23-30.

need further research into bioinformatics visualization to develop the tools that will meet the upcoming genomic and proteomic challenges. ■

Acknowledgments

We thank Eric Baehrecke, Harry Hochheiser, Eser Kandogan, Yehuda Koren, Bill Ladd, and the anonymous reviewers for their thoughtful comments. Partial support for this project came from the University of Maryland Institute for Advanced Computer Studies. Bongshin Lee served as an early

contributor to the software's development. Additional information about the HCE, more screen images, a user manual, and the software are available at <http://www.cs.umd.edu/hcil/multi-cluster>.

References

1. M.B. Eisen et al., "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of Sciences USA*, vol. 95, no. 25, 1998, pp. 14,863-14,868; <http://www.pnas.org/cgi/content/full/95/25/14863>.

2. M. Bitter et al., "Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling," *Nature* vol. 406, 2000, pp. 536-540; http://www.nhgri.nih.gov/DIR/Microarray/selected_publications.html.
3. I. Hedenfalk et al., "Gene-Expression Profiles in Hereditary Breast Cancer," *The New Journal of Medicine*, vol. 344, no. 8, 2001, pp. 539-548; http://www.nhgri.nih.gov/DIR/Microarray/selected_publications.html.
4. B. Shneiderman, "Dynamic Queries for Visual Information Seeking," *IEEE Software*, Nov./Dec. 1994, pp. 70-77.
5. C. Williamson and B. Shneiderman, "The Dynamic HomeFinder: Evaluating Dynamic Queries in a Real-Estate Information Exploration System," *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR 92)*, ACM Press, New York, 1992, pp. 338-346.
6. S.K. Card, J.D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization*, Morgan Kaufmann, San Francisco, 1999.
7. P.O. Brown and D. Botstein, "Exploring the New World of the Genome with DNA Microarrays," *Nature Genetics Supplement*, vol. 21, 1999, pp. 33-37; <http://www.nature.com/ng/>.
8. A. Inselberg and T. Avidan, "Classification and Visualization for High-Dimensional Data," *Proc. 6th Int'l Conf. Knowledge Discovery and Data Mining (KDD 00)*, ACM Press, New York, 2000, pp. 370-374.
9. E. Kandogan, "Visualizing Multi-Dimensional Clusters, Trends, and Outliers Using Star Coordinates," *Proc. 7th Int'l Conf. Knowledge Discovery and Data Mining (KDD 01)*, ACM Press, New York, 2001, pp. 107-116.

Jinwook Seo is a graduate research assistant in the Department of Computer Science & Human-Computer Interaction Laboratory, Institute for Advanced Computer Studies, University of Maryland, College Park. His research interests include information visualization and human-computer interaction. Seo received an MS in computer science from Seoul National University. Contact him at jinwook@cs.umd.edu.

Ben Shneiderman is a professor in the Department of Computer Science & Human-Computer Interaction Laboratory, Institute for Advanced Computer Studies, University of Maryland, College Park. His research interests include human-computer interaction and information visualization. Shneiderman received a PhD in computer science from the State University of New York, Stony Brook. Contact him at ben@cs.umd.edu.

The IEEE Computer Society thanks these sponsors for their contributions to the Computer Society International Design Competition.



FINANCIAL SPONSORS

ABB
AMD
EMC²
Hewlett-Packard
Intel
Lucent Technologies
Microsoft
Motorola
Science Applications International Corp.
Sun Microsystems

TECHNOLOGY SPONSORS

Ericsson
Intel
Toshiba
Microsoft

computer.org/csids/