

## RESEARCH BRIEF

# HIERARCHICAL CLUSTER EXPLORER FOR MULTIDIMENSIONAL CLUSTERING AND FEATURE DETECTION

### The potential

Cluster analysis of multidimensional data is widely used in many research areas such as financial, economic, sociological and biological analyses, including microarray experiment data sets. In particular, genome researchers are using cluster analysis to find meaningful groups in microarray data, also known as gene arrays or gene chips.

Cluster analysis reveals the underlying structure of an input data set, natural subclasses, interesting unusual patterns, and potential outliers. It serves as a basis for further analyses.

### The challenge

There are many challenges in visualizing such data. For example, the huge volume of the data makes it impossible to show the dendrogram of a large microarray experiment in one screen.

Researchers also struggle to understand the implications of a clustering result for their research. Since the clusters are in a high dimensional space (typical results have from 2–40 conditions), it is difficult to see patterns on a 2D (or even 3D) display. Another problem is that there may be hundreds of clusters of various sizes; spotting the meaningful clusters is a challenge, especially when a static display is used.

Users need an efficient visualization tool extract patterns from microarray datasets. They also need to be able to quickly apply knowledge to interpret a cluster by visual display in coordinated views.

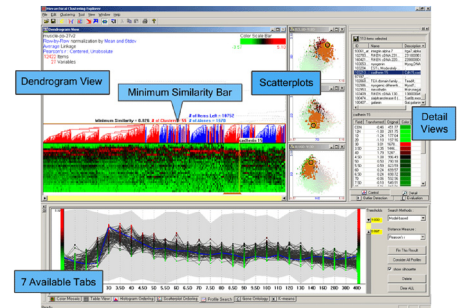
Software developers have been working on these problems. Early tools produced only printed results, while newer ones enable some online exploration.

Some clustering algorithms, such as k-means, require users to specify the number of clusters as an input, but users rarely know the right number beforehand. Other clustering algorithms automatically determine the right number of clusters, but

users may not be convinced of the result since they had little or no control over the clustering process.

### The research

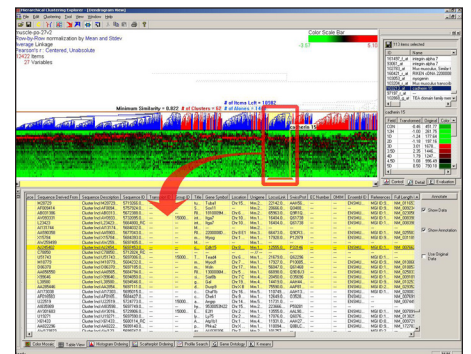
Hierarchical Clustering Explorer (HCE) software, currently in version 3.0, is being developed to give users more control over data analysis processes and to enable more interactions with analysis results through interactive visual techniques. With such control, users will be able to perform exploratory data analysis, establish meaningful hypotheses and verify results.



Features of Hierarchical Clustering Explorer

HCE applies the hierarchical clustering algorithm without a predetermined number of clusters, and then enables users to determine the natural grouping with interactive visual feedback (dendrogram and color mosaic) and dynamic query controls.

HCE includes four tools that allow users to interactively explore outcomes and gain a stronger understanding of the significance of the clusters.



A view incorporating a dendrogram (top) and a table (bottom). Here, each row has annotations for a gene. Each column represents an annotation from an external database. All 12422 genes in the dendrogram are in the tabular view, and there are 28 annotation columns. When users select a cluster of 113 genes in the dendrogram view, the annotation information for those genes is highlighted in the table.

- An overview lets users see the entire dataset, which helps them easily spot high-level patterns and hot spots. A detail view allows these areas to be examined.
- Dynamic query slider bar controls allow users to view clusters of varying size more clearly and reduce clutter from too much detail.
- Users can see how the hierarchical clusters are presented in a familiar and easy-to-understand 2-dimensional scatterplots. The coordination between the overview color mosaic and the scattergram is bi-directional; users can select a group of items in either view and see where they fall in the other view. The scatterplots can be ordered according to relevant criteria.
- Cluster comparisons allow researchers to see how different clustering algorithms group the data.

HCE also features:

- A gene ontology browser, coupled with clustering results so that known gene functions within a cluster can be easily studied
- A profile search so that genes with a certain temporal pattern can be easily identified.
- Histogram ordering and table view
- Continuous zooming in the dendrogram view

### Availability

HCE 3.0 is a stand-alone Windows® application that runs in a general PC environment. It is freely downloadable for academic and/or research purposes. A user manual is also available. Download at [www.cs.umd.edu/hcil/multi-cluster/hce\\_3.html](http://www.cs.umd.edu/hcil/multi-cluster/hce_3.html)

### Support

HCE research has been partially supported by National Institutes of Health grant N01 NS-1-2339.

### Contact

#### Ben Shneiderman

Professor, Computer Science Department and the Institute for Systems Research  
3177 A.V. Williams Bldg.  
University of Maryland  
College Park, MD 20742

Phone: 301.405.2680

Email: [ben@cs.umd.edu](mailto:ben@cs.umd.edu)

Web: [www.cs.umd.edu/~ben/](http://www.cs.umd.edu/~ben/)

#### Jinwook Seo

Graduate Research Assistant  
Computer Science  
3174 A.V. Williams Bldg.  
University of Maryland  
College Park, MD 20742

Phone: 301.405.2725

Email: [jinwook@cs.umd.edu](mailto:jinwook@cs.umd.edu)

Web: [www.cs.umd.edu/~jinwook/](http://www.cs.umd.edu/~jinwook/)

### Links

Hierarchical Clustering Explorer site at HCIL, including papers  
[www.cs.umd.edu/hcil/hce/](http://www.cs.umd.edu/hcil/hce/)

Bioinformatics Visualization at HCIL  
[www.cs.umd.edu/hcil/bioinfovis/](http://www.cs.umd.edu/hcil/bioinfovis/)

Human-Computer Interaction Laboratory  
[www.cs.umd.edu/hcil/](http://www.cs.umd.edu/hcil/)

“Software for the 4th Dimension: visual interface lets users easily query time-ordered data,” ComputerWorld, June 4, 2001 [www.computerworld.com/industrytopics/financial/story/0,10801,60995,00.html](http://www.computerworld.com/industrytopics/financial/story/0,10801,60995,00.html)