# TECHNICAL RESEARCH REPORT

Interactive Exploration of Multidimensional Microarray Data: Scatterplot Ordering, Gene Ontology Browser, and Profile Search (2003)

*by Jinwook Seo, Ben Shneiderman*

**TR 2005-68**

# Interactive Exploration of Multidimensional Microarray Data: Scatterplot Ordering, Gene Ontology Browser, and Profile Search

*Jinwook Seo*[1,2,*], *Ben Shneiderman*[1]

[1]Department of Computer Science & Human-Computer Interaction Laboratory, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA
[2]Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC 20010 USA

## ABSTRACT

**Motivation:** Multidimensional data sets are common in many research areas, including microarray experiment data sets. Genome researchers are using cluster analysis to find meaningful groups in microarray data. However, the high dimensionality of the data sets hinders users from finding interesting patterns, clusters, and outliers. Determining the biological significance of such features remains problematic due to the difficulties of integrating biological knowledge. In addition, it is not efficient to perform a cluster analysis over the whole data set in cases where researchers know the approximate temporal pattern of the gene expression that they are seeking.
**Results:** To address these problems, we add three new features to the Hierarchical Clustering Explorer (HCE): (1) scatterplot ordering methods so that all 2D projections of a high dimensional data set can be ordered according to relevant criteria, (2) a gene ontology browser, coupled with clustering results so that known gene functions within a cluster can be easily studied, (3) a profile search so that genes with a certain temporal pattern can be easily identified.
**Availability:** HCE 2.0 is a PC application written in Microsoft Visual C++. The full application and user's manual of HCE 2.0 with three new features is freely available at http://www.cs.umd.edu/hcil/hce/ for academic or research purposes.
**Contact:** jinwook@cs.umd.edu

## 1. Introduction

Microarray experiments reveal genome-wide profiles of thousands of genes typically under 20-100 experimental treatments or 5-30 time points (Bittner *et al*., 2000; Butte, 2002; Zhao *et al*., 2002). One of the primary goals of gene expression profiling is to understand the functional roles of genes through their associated expression with genes of known function. Even though the entire human genome sequence is known, only a small number of genes/proteins have known functional roles.

To identify genes that have expression profiles similar to the profiles of known genes, researchers often group genes by using clustering algorithms. Numerous clustering and pattern matching strategies have been developed, but hierarchical agglomerative clustering has been the de facto standard for microarray data analyses (Moreau *et al*., 2002; Eisen *et al*., 1998). Our earlier work (Seo and Shneiderman, 2002), presented an interactive visualization tool - Hierarchical Clustering Explorer (HCE) to help researchers understand hierarchical clustering results of microarray experiments. HCE enabled researchers to interactively investigate

* To whom correspondence should be addressed.

clustering results by changing the cluster-tightness threshold (see section 2 for details) to determine natural groupings within the data set. HCE also has 2D scatterplot viewers to examine the 20-100 experimental treatments represented in the dendrogram and color mosaic.

However, the high dimensionality of data sets still makes it difficult to find interesting patterns. To cope with high dimensionality, low dimensional projections of the original data set are generated; human perceptual skills are more effective in 2D or 3D displays (Asimov, 1985; Friedman, 1987). One of the problems of using low dimensional projections for the analysis of high dimensional data sets is that so many projections are possible. Mechanisms have been developed to enable researchers to select low dimensional projections, but most are static and do not allow users to specify what is interesting to them. There is, therefore, a need for a mechanism that allows users to choose the property of projections that they are interested in, rapidly examine projections, and locate interesting patterns, clusters, or outliers.

Even with such improvements, the biological importance of clustering results remains difficult to identify. For most biologists, the names or biological database identities of genes in a cluster do not give sufficient information as to why they cluster together. Gene ontology (GO) annotations can help biologists understand the biological meaning of the clustering result (Gene Ontology Consortium, 2000; www.geneontology.org). With the GO annotation, researchers can easily recognize the biological process, molecular function, or cellular component that genes in a cluster share. Furthermore, it is possible to test a hypothesis that an unknown gene might have the same or similar biological role with the known genes in the same cluster.

Clustering is very useful for microarray experiment data analysis. It is, however, so computationally intensive that clustering over the whole data set disrupts exploration, especially if researchers already know the approximate pattern of gene expression that are seeking (Zhao *et al*., 2002). In such cases, researchers want to quickly find only gene expression profiles similar to the expected pattern of a previously characterized gene. Since they cannot search for the pattern easily, they have to conduct a series of searches for the expression profiles similar to the expected pattern. Therefore, they need an exploratory analysis tool that allows easy modification of the expected pattern and rapid update of the search result.

To address these problems in the analyses of high dimensional microarray experiment data sets, we expanded our Hierarchical Clustering Explorer with three features:

- scatterplot ordering methods so that all 2D projections can be ordered according to user-selected criteria
- gene ontology browser, coupled with clustering results so that known gene functions within a cluster can be easily studied
- profile search so that genes with a specified temporal pattern can be easily identified

## 2. Interactive Exploration of Hierarchical Clustering Results with HCE 2.0

HCE is an interactive visualization tool for hierarchical clustering results with user control of clustering results (dendrograms and color mosaics) (Figure 1). HCE users load a microarray experiment data set from a tab-delimited text file, and apply their desired hierarchical clustering methods to generate a dendrogram and a red-and-green color mosaic (Seo and Shneiderman, 2002). Users can immediately observe the entire clustering result in a single screen that enables identification of high-level patterns, major clusters, and distinct outliers. They can adjust the

color mapping to highlight the separation of groups in the data set. Then they start their exploration of the gene groupings. Instead of using fingers and pencils on a static clustering results, HCE users can use a dynamic query device called "minimum similarity bar" to find meaningful groups. The Y-coordinate of the bar determines the minimum similarity threshold. A cluster (a subtree of the dendrogram) will be shown only if any two items in the cluster are more similar than the minimum similarity threshold specified by the minimum similarity bar. So, users see tighter clusters as they pull the bar lower to increase the minimum similarity threshold.

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. To avoid this dilemma, 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.

Another troublesome problem related to clustering analysis is that there is no perfect clustering algorithm. Clustering results highly depend on the distance calculation method and linkage method used through clustering process. Therefore, molecular biologists and other researchers need some mechanism to examine and compare two clustering results. HCE users can select two different clustering methods and compare the two clustering results in a single screen. When users double click on a cluster in one clustering result, HCE shows the mapping to the other clustering result by connecting the same items with a line. (see http://www.cs.umd.edu/hcil/hce/ for detail) Through this comparison, users can determine clustering parameters that most faithfully assemble items into the appropriate biological groups according to their known biological function.

In using HCE with biologists and based on requests from other users, we added three features to generate HCE 2.0: scatterplot ordering, gene ontology browser, and profile search.

### 2.1 Scatterplot Ordering

Visualization of high dimensional data has been a hard problem in information visualization (Inselberg and Avidan, 2000; Kandogan, 2001). Dendrogram and color mosaic displays can help users find reasonable groupings, but they do little to identify interesting low-dimensional features. So, 2D and 3D projections have been useful when analyzing a high dimensional data set. Since most users readily understand 2D projections, while 3D projections introduce disorientation from navigation and increase occlusion problems (Cockburn 2002), we decided to use 2D scatterplots. HCE 2.0 users can choose 2 columns (experimental conditions, or samples) for X and Y axes respectively, and see the corresponding scatterplot where each item (a gene) is depicted as a point in (x,y): x is the value of the gene at the sample for X-axis, y is the value of the gene at the sample for Y-axis.

The large number of possible scatterplots for a high dimensional data set can present a problem, so users need efficient mechanisms to investigate the possible scatterplots. The View Tip in Spotfire DecisionSite (www.spotfire.com) provides users with a ranking of 2D scatterplots. This is a good start to browse low dimensional projections of high dimensional data. However, it ranks the 2D scatterplots only by linear correlation between variables. Inspired by the View Tip, we suggest a method to help users browse possible scatterplots and to let them easily find interesting projections. HCE 2.0 provides users with five meaningful criteria to order 2D projections. The first three criteria are useful to reveal statistical relationships between two

experimental conditions (or samples), and the next two are useful to find projections of interesting distributions:

(1) *Pearson's r* orders scatterplots according to the Pearson's correlation coefficient (from +1.0 to −1.0) so that users can easily find the most/least correlated ones.
(2) *Least square error (simple linear regression)* sorts scatterplots in terms of sum of square errors from the optimal line fit so that users can easily isolate ones where all points are closely/loosely arranged along a straight line.
(3) *Least square error (curvilinear regression)* sorts scatterplots in terms of sum of square errors from the optimal quadratic curve fit so that users can easily isolate ones where all points are closely/loosely arranged along a quadratic curve.
(4) *# of items in the region of interest* lists scatterplots in order of the number of items within a user-defined rectangular, elliptical, or free-formed region of interest so that users can easily find ones with most/least genes in the given region.
(5) *Uniformness* orders scatterplots according to the significance level of two-dimensional Kolmogorov-Smirnov test (Chakravarti *et al*., 1967) between a uniform distribution and a scatterplot so that users can easily find the most/least uniform scatterplot.

Users select an ordering criterion, and they see the complete ordering of all possible 2D projections according to the selected ordering criterion (Figure 2). The Score List (C in Figure 2) shows the result of ordering together with scores. Users can easily find the most or least interesting scatterplots by changing the sort order to ascending or descending order of score (or rank). It is also easy to examine the scores of all scatterplots with a certain variable for the X or Y-axis after sorting the list according to the X or Y column by clicking the corresponding column header.

Figure 2

However, users cannot gain an overview of all relationships between variables at a glance in the Score List. Overviews are important because they can show the entire distribution and reveal interesting features. A new visualization component, the Score Table (B), shows a lower triangular matrix where each cell represents a scatterplot. Each cell is color-coded by its score and the color mapping is shown at the top right corner of the Score Table. As users move the mouse over a cell, the scatterplot corresponding to the cell is shown in the Scatterplot Browser (D), and the corresponding item is highlighted in the Score List (C) simultaneously. Users can easily find a variable that is the least (or most) correlated to other variables by just scanning the row or column to find the darkest (or brightest) cell. It is also possible to find an outlying scatterplot whose cell has distinctive color intensity compared to the rest of the same row or column. After locating an interesting cell, users can double click on the cell to select, and enlarge it, and then they can scrutinize it on the Scatterplot Browser (D) and on other tightly coordinated views.

Users can quickly examine scatterplots by using item sliders attached to the Scatterplot Browser (D). Simply by dragging the vertical or horizontal slider bar, users can change the variable for the X or Y-axis. Users can investigate multiple scatterplots at the same time. They select more than one scatterplots in the Score List by clicking them with the Ctrl key pressed. Clicking on the 'Make Views' button above the Score List, produces the selected scatterplot in a separate child window (Figure 1).

Users can select a group of genes by dragging a rubber rectangle over a scatterplot, and the genes within the rubber rectangle will be highlighted with triangles in all other scatterplots

(Figure 1). On some scatterplots they might be neatly clustered, while on other scatterplots they may be widely scattered. Since selected genes are also highlighted in all tightly coupled views, it is possible to see the selected items from various perspectives including the Dendrogram View and the Profile Search (see section 2.3).

**2.2 Gene Ontology Browser**

Hierarchical clustering algorithms combined with interactive visualization techniques can help researchers discover meaningful groups in the data set. Adding biological evidence as to why a set of genes is in a cluster is the next step in understanding functional aspects of biological pathways.

In recent decades, biological knowledge has become available in many genomic databases and will increase rapidly in the future (Baxevanis, 2003). However, the lack of a shared controlled vocabulary is one of the major reasons that biologists cannot efficiently utilize the abundant knowledge in the databases. The databases are so diverse that researchers have difficulties in identifying relevant information from the databases and combining them (Karp, 2000).

In an effort to increase the utility of the knowledge in the biological databases, the Gene Ontology project provides the biology community with a set of structured vocabularies (the Gene Ontology) to describe domains of molecular biology (Gene Ontology Consortium, 2000; Hill *et al*., 2002). There have been many attempts to annotate gene expression profiles with the Gene Ontology (GO), such as the MAPPFinder (Doniger *et al*., 2003). Users define a criterion for a significant gene-expression change, MAPPFinder calculates the percentage of genes meeting the criterion for each GO term, and the results are shown in a GO browser in order of the percentage. In this way, users can rapidly find GO terms associated with genes of significant expression changes.

HCE 2.0 combines GO annotation data with clustering results of microarray experiment data sets to present the biological significance of the results in a unified and structured manner. Since most microarray experiment stations don't produce GO annotation in the output by default, scripts or relational database queries are necessary to add GO annotations to the microarray experiment data. We join biological databases to get gene ontology identifiers of genes. For example, we used *UniGene* and *LocusLink* to add GO annotation to the melanoma microarray data set (Bittner *et al*., 2000). Genes can be compared in terms of up-to-date GO annotations available at the Gene Ontology consortium website.

The gene ontology hierarchy is a directed acyclic graph (dag), but we use a tree structure to show the hierarchy since the tree structure is easier for biologists to understand and easier for developers to implement than a dag. In our tree representation of the gene ontology, many gene ontology terms may appear several times in different branches, but all paths from the root to the nodes are guaranteed to be unique.

HCE 2.0 users can click on a cluster in the Dendrogram View, and see the list of genes in the selected cluster together with their GO identifiers (Figure 3). The oval in the Dendrogram View located at the upper left corner of Figure 3 indicates the current selected cluster. 25 genes in the cluster and their GO identifiers are listed at the Gene List Control in the bottom right corner of Figure 3. Gene names are preceded by the 'G'-shape icon and GO identifiers are preceded by a flag-shape icon. GO identifiers are listed below the gene name with an indentation. For example, in the Figure 3 the gene '*transcription factor Dp-2*' has three GO identifiers (GO:0003677, GO:0003700 and GO:0003712). If users select a *GO identifier* in the Gene List Control, all possible paths from the root to the selected GO identifier in the entire GO hierarchy are shown at

the Ontology Tree Control (in the bottom left corner of Figure 3). To reduce clutter, all non-relevant paths are hidden. If users select a *gene* in the Gene List Control as in the Figure 3, paths for all GO identifiers of the gene are shown in the Ontology Tree Control. By taking a look at GO term names shown in the Ontology Tree Control, users can see the detail biological functions described using a shared controlled vocabulary. If users want more information about a GO identifier, they can double click on it and HCE2 will launch a web browser and open up a web page for the identifier at *godatabase.org* where users can also find all associated genes across many data sources (FlyBase, MGI, SRS, etc.). Clicking on the 'Show All' button shows all paths from the root to GO identifiers of all genes in the selected cluster. By carefully investigating the shared paths in the Ontology Tree Control, users can learn which biological functions are shared among the genes in the cluster.

We examined the melanoma microarray data (Bittner *et al*., 2000) with HCE 2.0 and set the minimum similarity threshold to 0.55. For more than half the clusters, there are genes in the same cluster that share at least one edge in the molecular function ontology dag. Since there are still many unknown genes in the human genome and there are many genes whose molecular functions are unknown, this successful experience might not always be repeated. However, the combination of clustering result and ontology may produce more meaningful insights as the gene ontology becomes more comprehensive.

**2.3 Profile Search**

Many microarray experiments measure gene expression over time (Butte, 2002; DiGiovanni et al.; Zhao *et al*., 2002). Researchers would like to group genes with similar expression profiles or find interesting time-varying patterns in the data set. Often times, they roughly know the time varying patterns that they want to find. For example, they might be interested in the genes that are up-regulated in a certain time period and down-regulated in remaining periods (Zhao *et al*., 2002). In such cases, researchers might benefit from a query environment where they can easily specify queries, instantly see the result of the queries, and easily modify their queries.

'Profile Search' in the Spotfire DecisionSite calculates the similarity to a search pattern (so called 'master profile') for all genes in the data set and adds the result as a new column to the data set. The built-in profile editor makes it possible to edit the search pattern, but the editor window is separate from the main information window. The modification of master profile in the profile editor is interactive, but search results are not updated dynamically.

TimeSearcher (Hochheiser and Shneiderman, 2001) supports interactive querying and exploration of time-series data. Users can specify interactive timeboxes over the time-varying patterns, and get back the profiles that pass though all the timeboxes. Users can drag and drop an item from the data set into the query window to create a query with a separate timebox for each time point over the item in the data set. Each timebox at each time point can be modified to change the query.

HCE 2.0 reproduces TimeSearcher's basic functions with a novel interface, Profile Search, that allows for rapid creation and modification of desired profiles. Key design concepts are:

- interactive specification of a search pattern on the Information Space : Users can submit their queries simply by mouse drags over the Information Space, rather than using a separate query specification window.
- dynamic query control : Users get the query results instantaneously as they change the search pattern, similarity function, or similarity threshold.

- sequential query refinement : Users can keep the current query results as a new narrowed information space for subsequent queries. This enables users to refine their query results, which follows the process of general problem solving.

The Profile Search consists of three parts (Figure 4): the Information Space where input profiles are drawn and queries are specified, the range slider to specify similarity thresholds, and a set of controls to specify query parameters. Users specify a search pattern by simple mouse drags. As they drag the mouse over the Information space, the intersection points of mouse cursor and vertical time lines define control points. A search pattern is a set of line segments connecting the contiguous control points specified. Users choose a search method and a similarity measure on the control panel. They can change the current search pattern by moving a control point (a rectangular point on the search pattern), by moving a line segment vertically or horizontally, or by adding or removing control points. All of these modifications are done by mouse clicks or drags, and the results are updated instantaneously. This integration of the space where the data is shown and the space where the search pattern is composed reduces users' cognitive load by removing the overhead of context switching between two different spaces.

Figure 4

Incremental query processing enables instantaneous updates (within 100 ms) so that dynamic query control is possible for most microarray data sets. The easy and fast search for interesting patterns enables researchers to attempt multiple queries in a short period of time to get important insights into the underlying data set.

In the Profile Search, users can submit a new query over the current query result. If users click "Pin This Result" button after submitting a query, the query result becomes a new narrowed search space (Figure 4). We define this process as "pinning." Pinning enables sequential query refinement, which makes it easy to find target patterns without losing the focus of the current analysis process. If users click on a cluster in the Dendrogram View, all items in the cluster are shown in the Profile Search. By pinning this result, users can limit the search to the cluster to isolate more specific patterns in the cluster.

Genes included in the search result are highlighted in the Dendrogram View. Conversely, if users click on a cluster in the Dendrogram View, profiles of the genes in the cluster are shown in the Profile Search so that users can see the patterns of genes in a different view other than color mosaic. Through the coordination between the Profile Search and the Dendrogram View, users can easily see the representative patterns of clusters and compare patterns between clusters. Since queries done in the Profile Search identify genes with a similar profile, the search results should be consistent with clustering results if the same similarity function is used. In this regard, the Profile Search is also useful to validate the clustering results.

Two different types of queries are possible in the Profile Search: Model-based queries and Ceiling-and-Floor queries.

**Model-based queries:** Users can specify a model pattern (or a search pattern) simply by mouse drags as shown in Figure 4, and then select a distance/similarity measure among 3 different ones and assign threshold values. All profiles satisfying the threshold range will be instantaneously shown in the Information Space. The three different measures are 'Pearson correlation coefficient', 'Euclidean distance', and 'absolute distance from each control point'. The first measure is useful when the up-down trends of profiles are more important than the magnitudes, while the second and the third measures are useful when the actual magnitudes are more important. Assume users select the third measure and the threshold values are 0 and 5. A profile

will be selected if the distance between each point of the profile and its corresponding control point of the search pattern is within the distance between 0 and 5. It's like selecting profiles that flow through an equi-width pipe of the diameter of 5 whose centerline is the search pattern.

**Ceiling-and-Floor queries:** It is possible to define ceilings and floors on the Information Space so that only the profiles between ceilings and floors are shown as a result (Figure 5). Users can specify a ceiling by dragging with the left mouse button depressed, and a floor by dragging with the right mouse button depressed. They can change ceilings and floors with mouse actions in the same way as they did for changing search patterns in model-based queries. A ceiling imposes upper bounds and a floor imposes lower bounds on the corresponding time points. This type of query is useful when users know the up-down patterns of the target profiles and the range of values at the corresponding time points. Compared to the queries using the similarity measure for model-based queries, ceil-and-floor queries allow users to specify separate bounds for each control point.

Figure 5

## 3. Implementation

The Hierarchical Clustering Explorer 2.0 (HCE 2.0) was implemented as a stand-alone application using Microsoft Visual C++ 6.0. The Microsoft Foundation Class (MFC) library was statically linked. HCE 2.0 runs on personal computers running Windows (at least Window 95) without special hardware or external library support. HCE is freely available at http://www.cs.umd.edu/hcil/hce/ for academic or research purposes.

Figure 6

Figure 6 shows four tightly coupled components of HCE (newly added three components and the Dendrogram View) and linkages between them. All linkages except ones with Ontology Browser are bi-directional. Outgoing interactions from the Ontology Browser are not useful in microarray data analysis, so we don't activate them to prevent distractions due to too many interactions. Updates by each linkage in Figure 6 are instantaneous (<100ms) for most microarray data sets.

To achieve rapid responses to user's actions, hash and map data structures were used because they enable constant time lookup of items, with only a modest storage overhead. Incremental data structures were used to support rapid query update in the Profile Search by maintaining active index sets for intermediate query results.

Microarray experiment data set can be imported to HCE 2.0 from tab-delimited text files. The latest gene ontology annotation data is automatically downloaded from the Gene Ontology Consortium's ftp server, but the current implementation requires some effort to add gene ontology annotation to input data sets.

## 4. Discussion and Conclusion

Recent advances in microarray technology have enabled highly parallel measurements (~500000) of nucleic acid molecules in biological samples. Since up to about 100 microarrays are used in a project, large high dimensional data sets are common in microarray data analyses. Members of the statistical and algorithmic communities have made major efforts to develop analyses of these data sets. Those efforts contributed significantly to advances in functional bioinformatics, but most of the software tools lack interactive exploration by human experts. In this paper, we deal with three problems by offering interactive tools for the analyses of

microarray data sets. These are presented as additions to our previously described interactive visualization tool, the Hierarchical Clustering Explorer (HCE).

First, high dimensionality of the data sets hinders users from recognizing important patterns in the data sets. We add a scatterplot browsing mechanism that helps users select interesting 2D projections of the high dimensional data set. Second, an even more difficult problem is to understand the biological significance of the patterns found in the data sets. We add a gene ontology browser coupled with clustering results to help users study biological functions of genes through the gene ontology annotation. Third, a lightweight interactive profile search tool is necessary when researchers roughly know the pattern that they want to find. HCE 2.0's interactive direct profile search mechanism allows users to easily specify queries, instantly see the result of the queries, and easily modify the queries.

HCE 2.0 was successfully used in a couple of case studies. We proposed a general method of using HCE 2.0 to identify the optimal signal/noise balance in Affymetrix gene chip data analyses. HCE 2.0's interactive features helped researchers find the optimal combination of three variables (probe set interpretation methods, present call filters, and clustering linkage methods) to maximize the effect of the desired biological variable on data interpretation (Seo *et al*., 2003). HCE 2.0 was also used to analyze *in vivo* murine muscle regeneration expression profiling data using Affymetrix U74Av2 (12,488 probe sets) chips measured in 27 time points. HCE 2.0's dynamic query controls and profile search played an important role in finding 12 "novel" downstream targets that are biologically relevant during myoblast differentiation (Zhao *et al*., in press).

This work is a part of our continuing effort to give users more controls over data mining processes and to enable more interactions with analysis results through interactive information visualization techniques. These efforts are designed to help users perform exploratory data analysis, establish meaningful hypotheses, and verify results. In this paper, we show how those visualization methods can help molecular biologists analyze and understand multidimensional gene expression profile data. Empirical validation on standard tasks, more case studies with biological researchers, and feedback from users will help refine this and similar software tools.

**Acknowledgements**

We appreciate thoughtful comments from Dr. Harry Hochheiser and Dr. Eric Baehrecke. We thank Dr. Eric Hoffman for making this work possible by providing financial support for the first author through grant N01 NS-1-2339 from the National Institutes for Health.

**References**


Asimov, D. (1985) The grand tour: a tool for viewing multidimensional data, *SIAM J. Sci. Statist. Computing* 6, 1, 128-143.

Baxevanis, A.D. (2003) The Molecular Biology Database Collection: 2003 update, *Nucleic Acids Research*, 31, 1-12.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., and Trent, J. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature*, 406, 536-540.

Butte, A. (2002) The use and analysis of microarray data, *Nature Reviews Drug Discovery,* Vol. 1 No. 12, 951-960.

Chakravarti, I.M., Laha, R.G., and Roy, J. (1967) *Handbook of Methods of Applied Statistics, Volume I*, Wiley, New York, 392-394.

Cockburn, A. and McKenzie, B. (2002) Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments, *Proceedings of ACM CHI'2002 Conference on Human Factors in Computing Systems*, 203-210.

DiGiovanni, S., Knoblach, S.M., Brandoli, C., Aden, S.A., Hoffman, E.P., and Faden, A.I., (2003) Temporal gene profiling after experimental spinal cord injury identifies cell cycle genes associated with neuronal damage and cell death, *Annals. Neurol.*, in press.

Doniger, S., Salomonis, N., Dahlquist, K., Vranizan, K., Lawlor1, S. and Conklin, B. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, *Genome Biology*, 4, 1, R7.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, Vol. 95, 14863-14868.

Friedman, J.H. (1987) Exploratory projection pursuit, *Journal of the American Statistical Association*, Vol. 82, No. 397, 249-266.

Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology, *Nature Genet*, 25, 25-29.

Hill, D.P., Blake, J.A., Richardson, J.E. and Ringwald, M. (2002) Extension and Integration of the Gene Ontology (GO): Combining GO vocabularies with external vocabularies. *Genome Research* 12, 1982-1991.

Hochheiser, H. and Shneiderman, B. (2001) Visual specification of queries for finding patterns in time-series data, *Proceedings of Discovery Science*, Springer, Berlin, 441-446.

Inselberg, A. and Avidan, T. (2000) Classification and visualization for high-dimensional data, *Proc. 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 370-374.

Kandogan, E. (2001) Visualizing multi-dimensional clusters, trends, and outliers using star coordinates, *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 107-116.

Karp, P.D. (2000) An ontology for biological function based on molecular interactions, *Bioinformatics*, Vol. 16, No. 3, 269-285.

Moreau, Y., De Smet, F., Thijs, G., Marchal, K., De Moor, B. (2002) Functional bioinformatics of microarray data: from expression to regulation, *Proceedings of the IEEE* , Vol. 90, Issue 11, 1722-1743.

Seo, J. and Shneiderman, B. (2002) Interactively exploring hierarchical clustering results, *IEEE Computer*, Vol. 35, No. 7, 80-86.

Seo, J., Bakay, M., Zhao, P., Chen, Y., Clarkson, P., Shneiderman, B., Hoffman, E.P. (2003) Interactive Color Mosaic and Dendrogram Displays for Signal/Noise Optimization in Microarray Data Analysis, to appear in the *Proc. IEEE International Conference on Multimedia and Expo*.

Zhao, P., Iezzi, S., Carver, E., Dressman, D., Gridley T., Sartorelli, V. and Hoffman, E.P. (2002) Slug is a novel downstream target of MyoD. Temporal profiling in muscle regeneration, *J. Biol. Chem*, 277, 30091-30101.

Zhao, P., Seo, J., Wang, Z., Wang, Y., Shneiderman, B., and Hoffman E.P., "In vivo filtering of in vitro MyoD target data: An approach for identification of biologically relevant novel downstream targets of transcription factors," *Comptes Rendus Biologies*, in press.

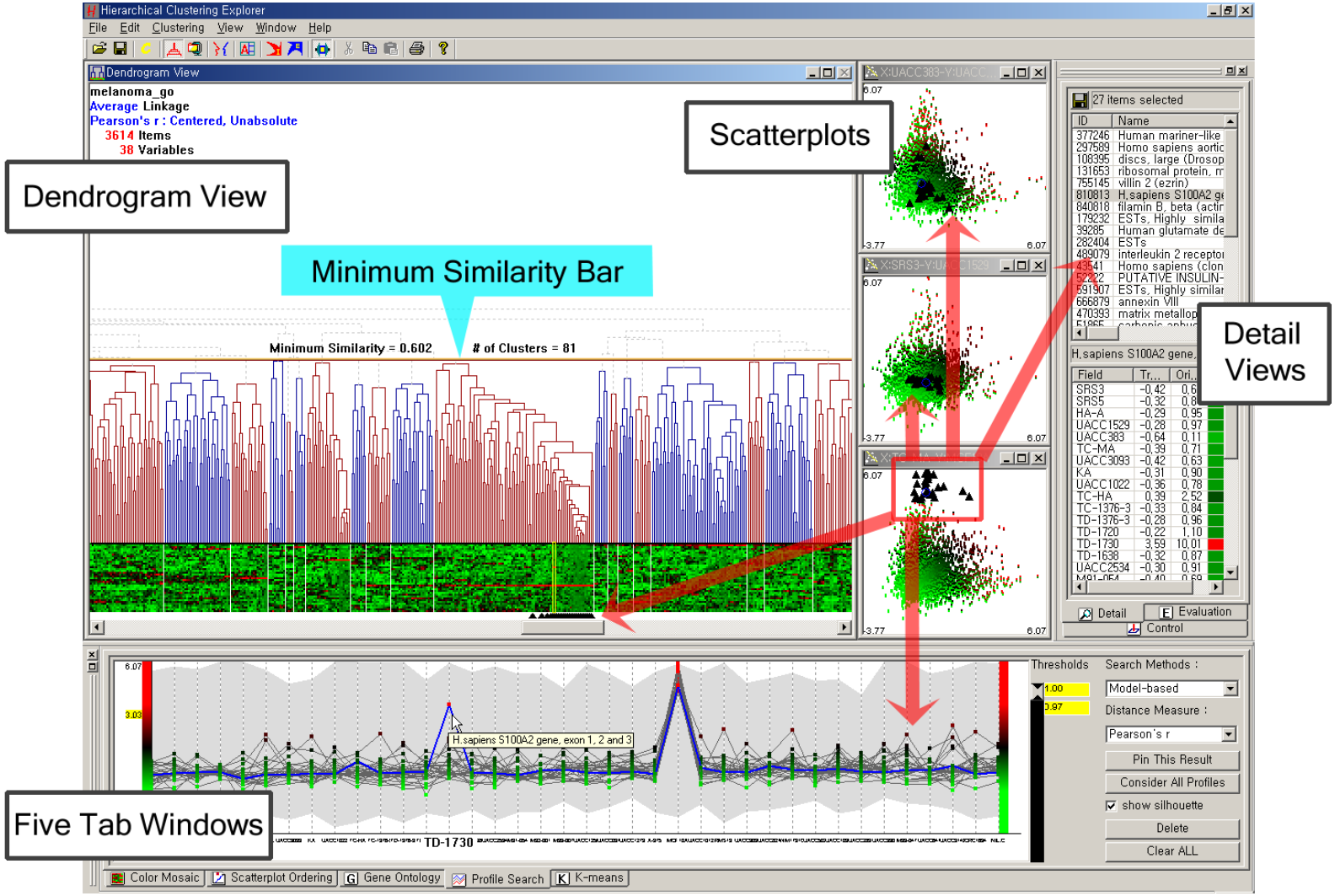


**Figure 1.** Overall layout of HCE 2.0. Minimum similarity bar was pulled down to get 81 clusters in the Dendrogram View. 27 genes are selected in the third scatterplot and they are highlighted with triangles in the other scatterplots, the Dendrogram View, detail view, and Profile Search tab window (see section 2.3). Users can select a tab window among the five tab windows at the bottom pane to investigate genes in the data set. Users can see the names of the selected genes and the actual profile values in the detail views. "Scatterplot Ordering" and "Gene Ontology" tabs are explained in sections 2.1 and 2.2 respectively.

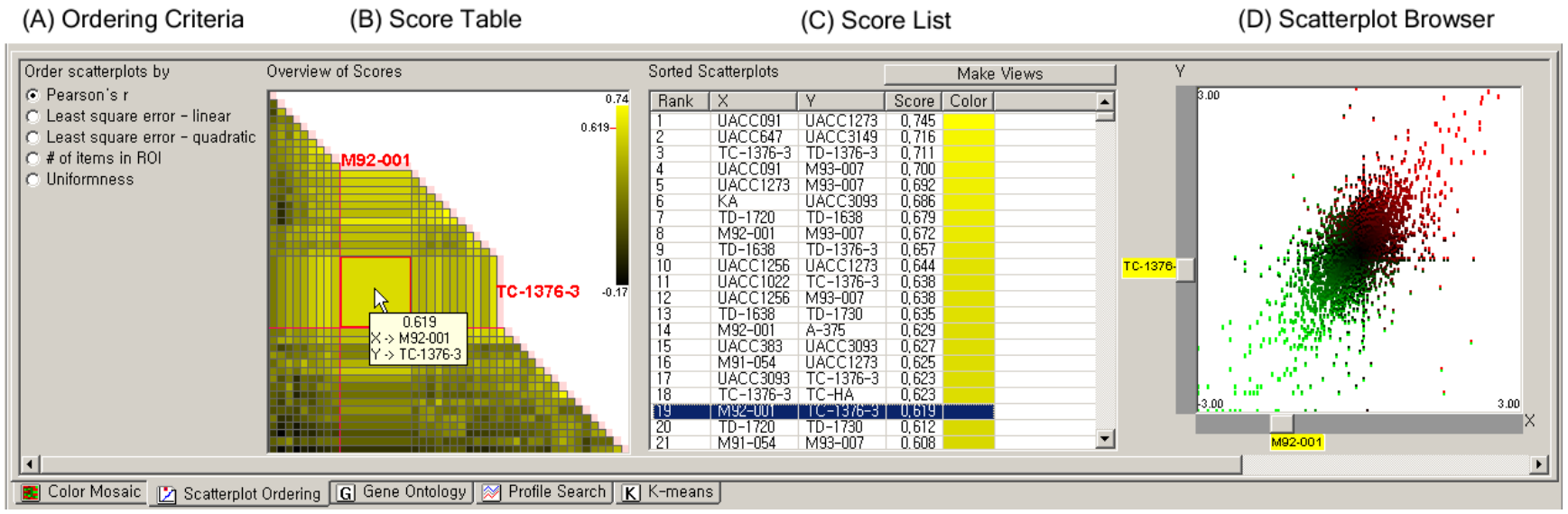


**Figure 2.** Ordering 2D scatterplots : All 2D scatterplots are ordered according to the current ordering criterion (A) in the Score List (C). Users can select multiple scatterplots at the same time and generate separate scatterplot windows for them to compare them in a screen. The Score Table (B) shows an overview of scores of all scatterplots with each cell color-coded by its score value. Mouseover event activates a cell in the Score Table, highlights the corresponding item in the Score List (C) and shows the corresponding scatterplot in the Scatterplot Browser (D) simultaneously. A double click on a cell enlarges the cell and activates it until another double click event. A selected scatterplot is shown in the Scatterplot Browser (D), where it is also easy to traverse scatterplot space by changing X or Y-axis using item sliders on the horizontal or vertical axis. Data displayed is from a cDNA microarray experiment data set (31 melanoma + 7 controls) by Bittner *et al*., 2000.

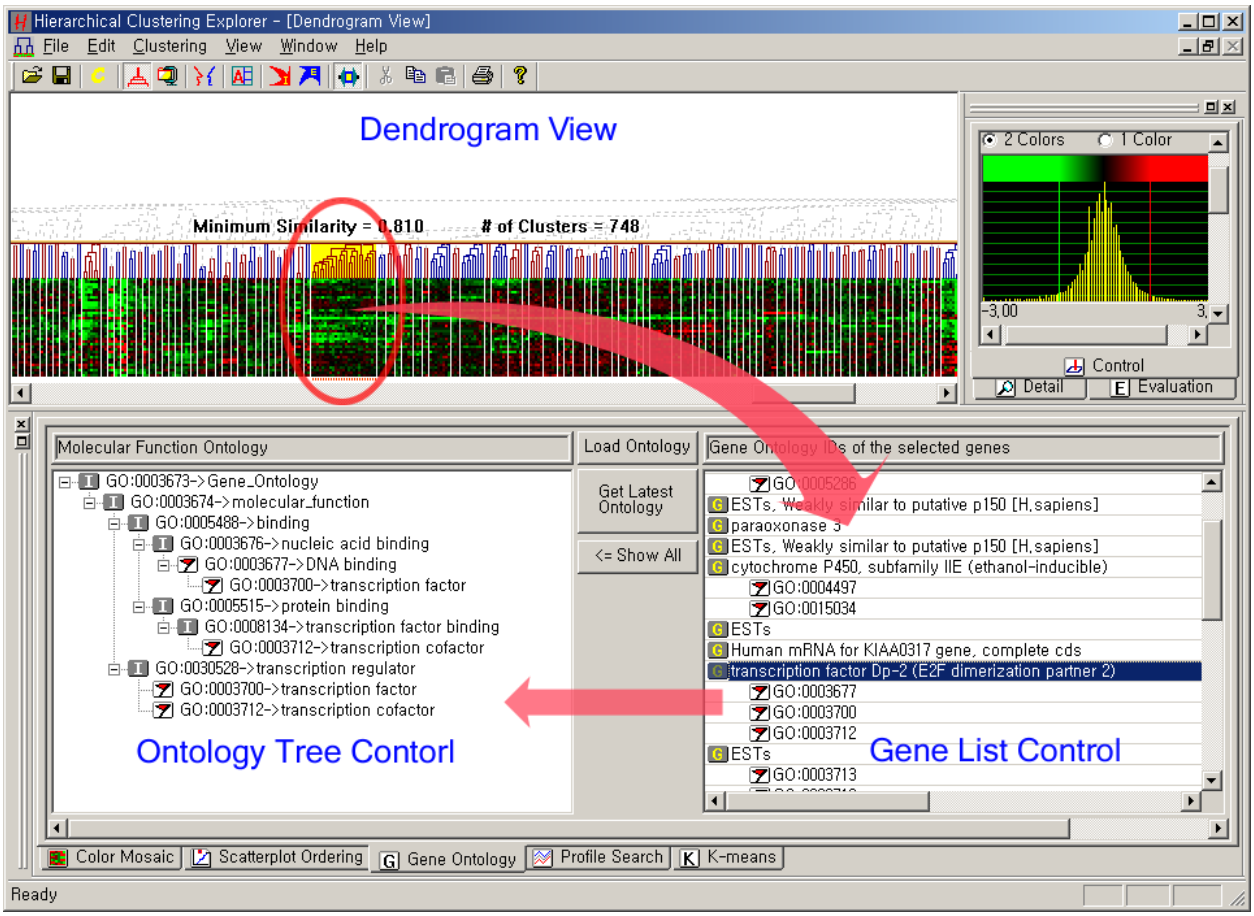


**Figure 3.** Hierarchical Clustering Explorer 2.0 with gene ontology browser on. Users can select a cluster in the Dendrogram View (at the top left corner), which is highlighted with an oval. 25 genes in the selected cluster are shown in the Gene List Control at the bottom right corner. All paths to the selected GO terms are shown with a flag-shape icon in the Ontology Tree Control at the bottom left corner. 'I' represents 'IS-A' relationship and 'P' represents 'PART-OF' relationship. The data set shown is a cDNA microarray experiment data set (31 melanoma + 7 controls) from Bittner *et al*., 2000.

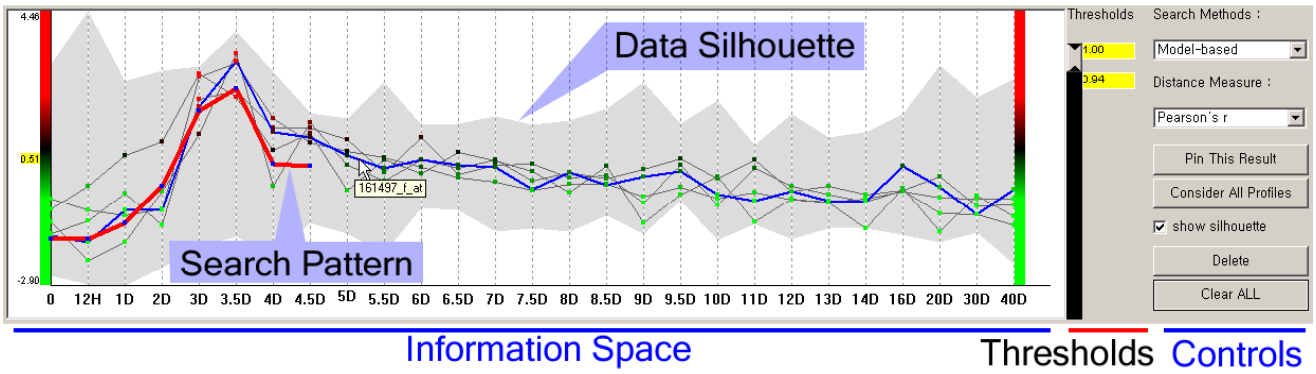


**Figure 4.** Profile Search: Layout of profile search view and an example of model-based query on the mouse muscle regeneration data. The data silhouette (the gray shadow) represents the coverage of all expression profiles. The bold line is a search pattern specified by users' mouse drags. Thin solid lines are the result of the current query that satisfies the given similarity threshold (more than 94% similar to the search pattern). The data set shown is a temporal gene expression profile on the mouse muscle regeneration (Zhao *et al*., 2002).

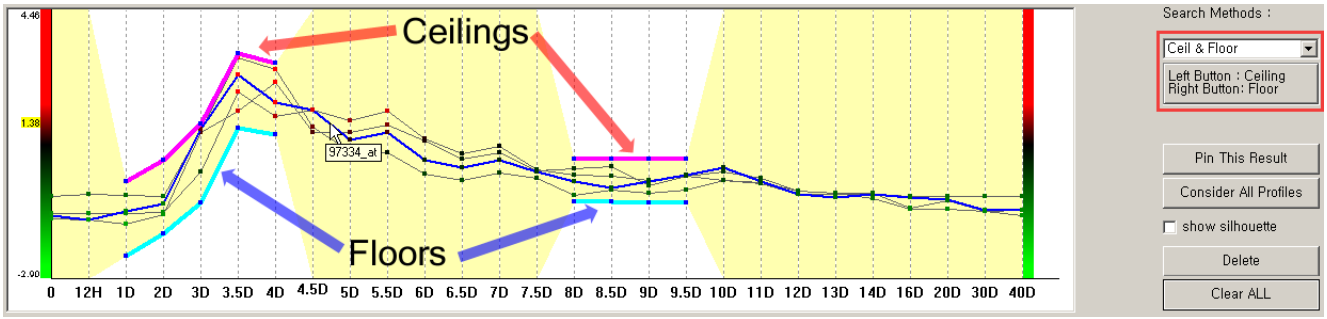


**Figure 5.** An example of the Ceiling-and-Floor query. Bold line segments above profiles define ceilings, and bold line segments below profiles define floors. Profiles only both below ceilings and above floors are shown as a result. Users can move a line segment or a control point of ceilings or floors to modify current query. Satisfactory region is indicated by a shadowed region to give users informative visual feedbacks of current query. The data set shown is a temporal gene expression profile on the mouse muscle regeneration (Zhao *et al*., 2002).

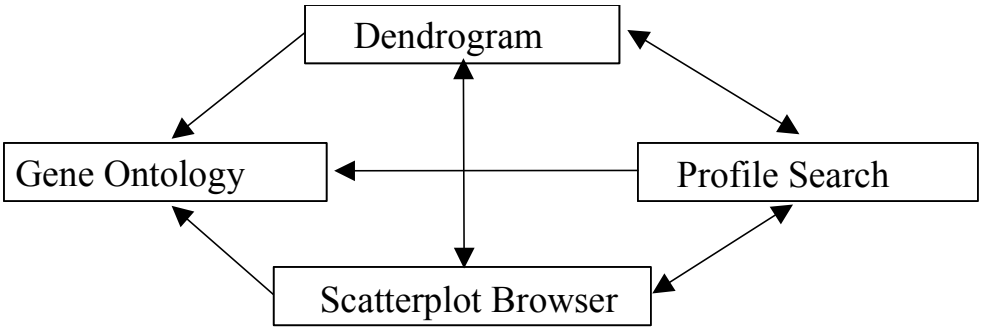


**Figure 6.** Diagram of interactions between components of HCE 2.0. All interactions except those with the Gene Ontology browser are bidirectional.