Related Documents:


# Table of Contents

Related Documents: .......................................................................................................................... 2
Table of Contents ............................................................................................................................. i

## Chapter 1 Equations for Correlations and other Similarity Measures ................................. 1-1
1.1 Common Correlations .............................................................................................................. 1-2
  1.1.1 Standard Correlation ...................................................................................................... 1-2
  1.1.2 Pearson Correlation ...................................................................................................... 1-2
  1.1.3 Spearman Correlation .................................................................................................. 1-3
  1.1.4 Spearman Confidence ................................................................................................. 1-3
  1.1.5 Two-sided Spearman Confidence .............................................................................. 1-4
  1.1.6 Distance ..................................................................................................................... 1-4
1.2 Special Case Correlations ..................................................................................................... 1-5
  1.2.1 Smooth Correlation .................................................................................................... 1-5
  1.2.2 Change Correlation .................................................................................................. 1-5
  1.2.3 Upregulated Correlation .......................................................................................... 1-6

## Chapter 2 Hierarchical Clustering ......................................................................................... 3-1
2.1 GeneSpring Analysis Methodologies for Making Trees ...................................................... 3-1
2.2 Making a Simple Tree .......................................................................................................... 3-2
  2.2.1 Automatic Annotation of a Gene Tree ....................................................................... 3-4
  2.2.2 Naming a New Gene Tree ......................................................................................... 3-5
2.3 More Complex Ways of Making Trees ................................................................................ 3-8
  2.3.1 Adding or Deleting Experiments from the Correlation ............................................ 3-9
  2.3.2 Restricting the Genes to be correlated ........................................................................ 3-10
  2.3.2.1 Restrictions Over an Entire Experiment or Interpretation .................................. 3-12
  2.3.2.2 Restrictions over a condition or sample ................................................................. 3-13
  2.3.2.3 Data File Restriction ............................................................................................. 3-14
  2.3.2.4 Restricting only certain types of data .................................................................... 3-15
  2.3.3 Similarity Definitions ................................................................................................. 3-17
  2.3.4 Making Trees ............................................................................................................ 3-18
  2.3.5 Creating a Subtree or List from a Node ..................................................................... 3-19
  2.3.6 Minimum Distance and Separation Ratios ............................................................... 3-21
2.4 Tree Commands .................................................................................................................... 3-22
2.5 References ............................................................................................................................ 3-24

## Chapter 3 K-Means Clustering ......................................................................................... 4-1
3.1 Function .................................................................................................................................. 4-1
3.2 Description ............................................................................................................................. 4-1
  3.2.1 Adding or Deleting Experiments from the Correlation ............................................ 4-3
  3.2.2 Restricting the Genes to be correlated ........................................................................ 4-4
  3.2.2.1 Restrictions Over an Entire Experiment or Interpretation .................................. 4-5
  3.2.2.2 Restrictions over a Single Condition or Sample .................................................. 4-7
  3.2.2.3 Restricting only certain types of data .................................................................... 4-9
Chapter 8 Global Error Models ................................................................. 9-1
  8.1 Global Error Model Details ................................................................. 9-1
  8.2 Technical Details ................................................................................. 9-5
  8.3 References ......................................................................................... 9-6
Index ........................................................................................................ 1

Document History for
Advanced Analysis Techniques Manual ...................................................... 3
Chapter 1  Equations for Correlations and other Similarity Measures

Many of the advanced analysis technics are based upon measures of gene similarity. Similarity or “nearness” between genes is usually based on the correlation between the expression profiles of the two genes. GeneSpring offers nine choices of similarity measures. Each is selectable from a drop-down list appearing the Cluster screen, as shown in the “Make New Tree” window presented in Figure 2-15.

Each measure takes two expression patterns and produces a number representing how similar the two genes are. Most of the measures of similarity are correlation measures, and their value will vary from -1 (exactly opposite) to 1 (the same). For a measure of distance, the result will vary from 0 (the same) to infinity (different). For confidences, the result will vary from 0 (no confidence) to 1 (perfect confidence). Both distance and confidence are actually measures of dissimilarity (small means close and large means far away). These are each transformed to measures of similarity by GeneSpring in ways detailed below.

If one expression value for a particular experiment for either gene is missing, that experiment will be not considered in the calculation.

The notation used to describe the formulas:

- **Result**: the result of the calculation for genes A and B.
- **n**: the number of samples being correlated over.
- **a**: the vector \((a_1, a_2, a_3 ... a_n)\) of expression values for gene A.
- **b**: the vector \((b_1, b_2, b_3 ... b_n)\) of expression values for gene B.

Normal mathematical notation for vectors will be used. In particular:

- **a.b = \(a_1b_1 + a_2b_2 + ... + a_nb_n\)**
- **|a| = square root(a.a)**
1.1 Common Correlations

1.1.1 Standard Correlation

Standard correlation measures the angular separation of expression vectors for Genes A and B around zero. As almost all normalized values for genes are positive, you find mostly positive correlations between genes when you use the Standard correlation. This metric is designed to answer the question “do the peaks match up?” or to put it another way, “are the two genes expressed in the same samples?” Since these questions are the most frequent questions a biologist is trying to get answered, GeneSpring calls it “Standard correlation”. It is important to note, what mathematicians and statisticians refer to as “correlation” usually refers to the Pearson correlation. The “Standard correlation” would be called “Pearson correlation around zero” by mathematicians and statisticians.

This is how to compute a Standard correlation:

\[ \text{Standard correlation} = \frac{a \cdot b}{|a| |b|} \]

1.1.2 Pearson Correlation

The Pearson correlation is very similar to the Standard correlation, except it measures the angle of expression vectors for genes A and B around the mean of the expression vectors (for example, the mean of the expression values constituting the profiles for Gene A and Gene B). Generally the mean of the expression vectors will be positive since expression values are based on concentrations of mRNA. Using the Pearson correlation you get more negative correlations than you get from the Standard correlation (for example, you find more genes that behave opposite to each other, because of where you put the baseline—at zero almost all gene values are above it, at 1 there are a fair amount that read below the baseline). It is worth noting that, for data normalized to an overall level of 1 (as with all normalizations that GeneSpring performs) the Pearson correlation gives you almost the same correlations as the Standard correlation when they are both performed on the logarithms of the genes’ expression values.

This is how to compute a Pearson Correlation:

Calculate the mean of all elements in vector a. Then subtract that value from each element in a. Call the resulting vector A. Do the same for b to make a vector B.

\[ \text{Pearson Correlation} = \frac{A \cdot B}{|A||B|} \]
1.1.3 Spearman Correlation

The Spearman correlation is a nonparametric correlation similar to the Pearson correlation except it replaces the data for Gene A and B with the ranks of the data (i.e. the lowest measurement for a gene becomes 1, the second lowest 2, and so forth). Spearman correlation calculates the correlation of the ranks for Genes A and B’s expression data around the mean of the ranks, using the same formula as Pearson correlation. In the Spearman correlation only the order of the data is important, not the level, therefore extreme variations in expression values have less control over the correlation. If there are ties in the data, then all of the tied values are assigned the average of the ranks, e.g. if the 5th, 6th and 7th lowest values are tied, all three datapoints are assigned a rank of 6.

This is how to compute a Spearman correlation:

Order all the elements of vector \( \mathbf{a} \). Use this order to assign a rank to each element of \( \mathbf{a} \). Make a new vector \( \mathbf{a}' \) where the \( i^{\text{th}} \) element in \( \mathbf{a}' \) is the rank of \( a_i \) in \( \mathbf{a} \). Now make a vector \( \mathbf{A} \) from \( \mathbf{a}' \) in the same way as \( \mathbf{A} \) was made from \( \mathbf{a} \) in the Pearson Correlation. Similarly, make a vector \( \mathbf{B} \) from \( \mathbf{b} \).

\[
\text{Spearman correlation} = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}| |\mathbf{B}|}
\]

1.1.4 Spearman Confidence

Spearman confidence is a measure of similarity, not a correlation. Spearman confidence is one minus the p-value for the statistical test when the Spearman correlation is zero versus the alternative when it is larger than zero. There is a high Spearman confidence value if there is a high Spearman correlation and a low p-value, meaning there is a low probability to find a correlation this high. This measure is very similar to looking for large Spearman correlation values, but it takes account of the number of sub-experiments in your experiment set.

This is how to compute a Spearman confidence:

If \( r \) is the value of the Spearman correlation as described in 1.1.3 “Spearman Correlation” on page 1-3, then:

\[
\text{Spearman confidence} = 1 - \text{(probability you would get a value of } r \text{ or higher by chance.)}
\]
1.1.5 Two-sided Spearman Confidence

Two-sided Spearman confidence is again a measure of similarity but not a correlation. It is very similar to the Spearman confidence discussed in 1.1.4 “Spearman Confidence” on page 1-3, except it is based on the two-sided test of whether the Spearman correlation is either significantly greater than zero or significantly lower than zero. There is a high Two-sided Spearman confidence value if the absolute value of the Spearman correlation is large and has a small p-value, meaning there is a low probability to find a correlation with absolute value this large.

This “similarity” measure is really good for answering the question “What genes behave similarly to a specific gene, and at the same time, what genes behave opposite to a specific gene?”. It should probably not be used for the advanced clustering algorithms (such as k-means and hierarchical clustering) because the genes with high two-sided confidence values are really a mixture of similar and dissimilar genes.

This is how to compute a Two-sided Spearman confidence:

If $r$ is the value of the Spearman correlation as described above, then:

\[
\text{Two-sided Spearman confidence} = 1 - \text{(probability you would get a Spearman correlation of } |r| \text{ or higher, or } -|r| \text{ or lower, by chance.)}
\]

1.1.6 Distance

Distance is not a correlation at all, but a measurement of dissimilarity. Distance is the measurement of Euclidian distance between the point for gene A (defined by its expression values for each point in N-dimensional space, where N is the number of experimental points in your experiment) and the point for gene B. This is more formally known as the Euclidian metric.

This is how to compute a Euclidian Distance:

\[
\text{Distance} = |a - b|
\]

Since distance is a measure of dissimilarity, the distance (d) is converted to a similarity measure as $e^{-d}$. 
1.2 Special Case Correlations

The next three metrics should only be used to look at special cases. They are all modified versions of the Standard correlation. Using these three metrics only makes sense when your data is in a sequence, such as “before” and “after”, a time series, or a drug series. The sequence does not have to be continuous, but it must have an order. If your experiment is set up with an experimental point taken at each of “before”, “after”, and “control” then the following correlations will not make sense applied to your data.

1.2.1 Smooth Correlation

This is how to compute a Smooth correlation:

Make a new vector \( \mathbf{A} \) from \( \mathbf{a} \) by interpolating the average of each consecutive pair of elements of \( \mathbf{a} \). Insert his new value between the old values. Do this for each pair of elements that would be connected by a line in the graph screen. Do the same to make a vector \( \mathbf{B} \) from \( \mathbf{b} \).

\[ \text{Smooth correlation} = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| ||\mathbf{B}||} \]

1.2.2 Change Correlation

The Change correlation looks for the opposite of what the Smooth correlation looks for. The change correlation only looks at the change in expression level of adjacent points. However, it is also very similar to the Standard correlation, in that it measures the angular separation of expression vectors for genes A and B around zero (i.e. in comparison to zero), except instead of using the expression values in each experimental point to create the expression vector for gene A, it is based on an arc tangent transformation of the ratio between adjacent pairs of experimental points and uses these to create the expression vector. This correlation looks for when gene A and gene B are changing at the same time. Using the arc tangent makes a measure of change that is less sensitive to outliers than using the ratio directly.

This is how to compute a Change correlation:

Make a new vector \( \mathbf{A} \) from \( \mathbf{a} \) by looking at the change between each pair of elements of \( \mathbf{a} \). Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values \( a_i \) and \( a_{i+1} \) is \( \text{atan}(a_{i+1}/a_i) - \pi/4 \). Do the same to make a vector \( \mathbf{B} \) from \( \mathbf{b} \).

\[ \text{Change correlation} = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| ||\mathbf{B}||} \]
1.2.3 Upregulated Correlation

The Upregulated correlation is very similar to the Change correlation, except that it only considers positive changes. All negative values for the arc tangent transform of the ratio are set to zero. This emphasizes only periods when new RNA is being synthesized.

This is how to compute an Upregulated correlation:

Make a new vector \( A \) from \( a \) by looking at the change between each pair of elements of \( a \). Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values \( a_i \) and \( a_{i+1} \) is \( \max(\arctan(a_{i+1}/a_i)-\pi/4,0) \). Do the same to make a vector \( B \) from \( b \).

\[
Upregulated\ correlation = \frac{A \cdot B}{|A||B|}
\]
Chapter 2  

Hierarchical Clustering

For starter information on Hierarchical Clustering (also known as Gene Trees), please see GeneSpring Basics Instructional Manual Chapter 6 “Trees” on page 6-1. For basic information on viewing Gene Trees, please see GeneSpring User Manual 3.3, “Tree View”. The following examples will use the a genome consisting of 112 rat genes, rather than the yeast genome used in most of the examples in this book.

Hierarchical Clustering allows you to visualize a set of samples or genes by organizing them into a mock-phylogenetic tree, often referred to as a dendrogram. In these trees, samples or genes having similar effects on the gene expression patterns are clustered together. How far across the tree you can go and still find a subtree containing both samples or genes can be considered a measure of how closely correlated those samples or genes are. This can be an exceedingly powerful application. For example, it allows you to find incidents in which environmental stresses cause similar effects to the expression levels of mutant organisms or if the ligands/drugs applied have similar effects on the tested samples.

Many of the commands used for Experiment Trees are the same as the ones used for Gene Trees.

To view an existing tree, simply go to the Navigator, open the folder of “Gene Trees” or “Experiment Trees” and select by clicking on the Tree you would like to view.

2.1  GeneSpring Analysis Methodologies for Making Trees

A tree of genes is constructed by successively finding the two closest genes or subtrees at each stage of the process, and merging them to form a larger subtree. Then branches where the distance between subbranches are less than a user-specified separation ratio are merged into a single branch, in order to simplify the tree. Finally, labels are assigned to branches by examining all known gene lists to see if they are statistically significantly similar to the genes in the branch. Again, a wide selection of distance measures is available, including Euclidean distance and distance measures based on correlation (distance = 1 – correlation).

For example, the following six-step algorithm is used to build gene or experiment trees in GeneSpring:

1. Determine if there is only one gene or subtree left. If yes, go to step five.
2. Find the two closest genes/subtrees.
3. Merge these two into one subtree.
4. Return to step one.
   To prune the tree:
5. Merge together branches where the distance between sub-branches is less than the separation ratio, subject to considering genes with less than the minimum distance apart.
To annotate the tree:

6. Determine if there is a label (for each branch) that could fit on that branch by looking at all gene lists and looking for statistically significant correlations. (This function can be turned off by deselecting “Do Automatic Annotation” in the “Clustering” window.)

The following four formulas are used:

- The distance formulas depend on the correlation used. If “Euclidean Distance” is chosen, then the definition of distance is the Euclidean Distance. See Chapter 1, “Equations for Correlations and other Similarity Measures” for more information. Otherwise, distance equals 1 minus correlation.
- The ratio of branch distances equals distance1/max (distance2, min distance).
- The statistical significance formula is the same as for similar gene lists, taking into account the total number of computations.
- The formula to merge together branches is the average of the expression values of the two branches to get a new artificial gene.

The implementation is more complex to improve efficiency, but produces results equivalent to the ones described.

## 2.2 Making a Simple Tree

In Figure 2-1, The default “Clustering” window, is displayed. This window is the result of selecting Tools > “Cluster...” in the main Genome Browser. The kind of tree being referred to by this command is a mock-phylogenetic tree. The “Clustering” window allows you to make a gene tree or an experiment tree. In these trees, genes or experiments with similar expression patterns are clustered together. For a further description of complex trees, see either 2.3 “More Complex Ways of Making Trees” on page 2-8 for a description of experiment trees or GeneSpring User Manual 3.3, “Tree View” for a description of gene trees. You can create simple trees by correlating one experiment. You can make complex trees by correlating a number of different experiments together and/or placing restrictions upon the genes to be looked at, before those correlations are done.

The “Clustering” window will appear showing the selected gene list from the main Genome Browser window already in the “Experiments to be correlated together” box in the middle of the window. Its weight is the default value (1.0).
The default “Clustering” window for Gene Trees

Directly below the title bar is the menu bar, containing three drop-down menus: “Experiments”, “Restrictions”, and “Options”. In the Navigator at the left are folders named “Gene List” and “Experiments”.

There are three boxes.

1. The topmost gray box is the “Genes to Use” box. It containing information directly relating to the list of genes to be correlated to create the gene tree or experiment tree. In Figure 2-1 the gene list specified after the “Choose from Genes” is “all genes”. The “all genes” list is the default list to correlate, but you may choose any previously defined list from the Navigator and limit the tree to only those genes.

In the middle is a white box labeled “Additional Restrictions on genes for clustering”. It contains all of the restrictions placed on the gene lists prior to their inclusion in the tree. It will be discussed in greater detail later, as it applies to the more complex trees. Please see 2.3.2 “Restricting the Genes to be correlated” on page 2-11.

At the bottom of the “Genes to Use” box you can see how many of the genes out of the total
number pass the restrictions listed in the white “Restrictions” box at the middle of the “Genes to Use” box.

2. The middle “Experiments to Use” box currently contains the title of and the weight given to the current experiment, NIH Spinal Cord Study. For a simple tree, you will not need to change this default setting. To remove this experiment, select it to highlight, then click the “Remove” button. To add another experiment, select it in the Navigator and click the “Add” button. You can alter the weight of an experiment when you add it to the clustering.

3. The bottom “Gene Trees” box contains a drop-down menu of similarity definitions. See Chapter 1, “Equations for Correlations and other Similarity Measures” for more information on them.

Below that are the editable boxes for the minimum distance and separation ratios. See 2.3.6 “Minimum Distance and Separation Ratios” on page 2-21 for more information. There are two other checkboxes, “Do Automatic Annotation” and “Only Annotate With Standard Lists”. These two checkboxes allow you to turn on or off the automatic annotation. Please refer to 2.2.1 “Automatic Annotation of a Gene Tree” on page 2-4 for more information.

To make a Gene Tree of “all genes” in the current experiment, accept the defaults and click the “Start” button at the bottom center of the window. After a few moments of processing, a new window will come up asking you to name your new tree. The “Close” button will not cancel the new tree. It will close the original “Clustering” window.

2.2.1 Automatic Annotation of a Gene Tree

After clustering the genes according to their expression patterns, GeneSpring checks all known lists against all subtrees of the new gene tree, to assign names to the tree nodes where possible. All of the branches intersecting to form a node constitute the subtree defined by that node. In other words, GeneSpring attempts to name the groupings of genes the clustering process has revealed. If you do not want GeneSpring to do this, you will need to de-select the “Do Automatic Annotation” and “Only Annotate With Standard Lists” checkboxes in the “Clustering” window. See GeneSpring User Manual 3.3, “Tree View” for a complete description of gene trees and illustrations of what branches, subtrees and nodes are. If you deselect only “Only Annotate With Standard Lists” GeneSpring will attempt to annotate your tree with every list associated with that genome. This may cause a great deal of clutter in your tree. However, many of the labels are so small they will only be visible in high magnification. Please see Figure 2-2 for an example of visible labels. More examples and general tree information is available in GeneSpring Basics Instructional Manual Chapter 6 “Trees” on page 6-1.

GeneSpring also assigns a measure of statistical significance to each name it associates with a node, or cluster of genes. The number it lists is not the P-value (probability that this name is a false positive) for its associated cluster, but the negative of the log base 10 of its P-value. Therefore the higher the number, the more significant the comparison is. The comparisons between the lists and the cluster of genes forming a subtree are not exact matches. Instead GeneSpring looks for statistically significant overlaps, which may mean the cluster of genes forming the subtree is a subset or a superset of the list named. A particular subtree may be statistically similar to more than one list and its name will be displayed as a label along the top when there is enough room in the Genome Browser. While this naming is happening a window labeled “Title Search” appears.
The “Title Search” window, like the “Clustering Progress” window, contains a blue bar indicating how far along in the naming process GeneSpring is. When the blue bar approaches the right edge of the window the naming is almost done. The naming is done when the entire “Title Search” window is filled with blue. You may stop the naming process at any point, and thereby go straight to the “Name New Gene Tree” window, which allows you to name the new tree. This tree will not have all of the names associated with the nodes displayed, but only the ones labeled before you canceled the “Title Search” window. To close the “Title Search” window, click the “X” button at the far right of the title bar of that window.

2.2.2 Naming a New Gene Tree

After the “Title Search” is complete the “Name New Gene Tree” window will appear. This window is illustrated in Figure 2-2. This window allows you to name and accept the tree you have just made or to cancel it. To name your tree move the cursor into the white box next to the “Name” label, click in the box, and type in a new name. In Figure 2-2 the name of the tree is the default name, “NIH Spinal Cord Study Default Interpretation”. After renaming the tree click the “OK” button to accept the tree and the name you have given it. If you give your new tree a redundant name, you will get an warning message. Click the “OK” button of the error message to return to the “Name New Gene Tree” window. Once properly named your new tree will be saved under the folder Gene Trees in the main Navigator.
Figure 2-2 A simple gene tree, made from the “all genes” list

Highlight over the default name and type in a new name. After you name it, your new tree it will appear in the main Genome Browser and be listed under the Gene Trees folder in the Navigator. In Figure 2-3, the tree has been named “simple NIH Spinal Cord Study”.

Hierarchical Clustering

Advanced Analysis Techniques Manual

Making a Simple Tree

Figure 2-3 A Gene Tree in the Genome Browser

This is a very simple tree made from one experiment. There are many more complex and revealing ways to make trees as described in the next section, 2.3 “More Complex Ways of Making Trees” on page 2-8.
2.3 More Complex Ways of Making Trees

Complex trees can be made from multiple experiments or by tightly defining the types of genes to use. You can open the “Gene Lists” folder in the Navigator and select a list to reduce the number of genes to be made into a tree.

To begin an Experiment Tree, select “Make New Tree...” from the “Cluster” drop down menu. Use the drop down menu in the middle of the Clustering Panel to select “Experiment Tree”.

![The default Clustering window for experiment trees](image)

The name of the experiment you were previously viewing in the main window is listed in the “Experiments to Use” box and the current list is in the “Genes to Use” box.
2.3.1 Adding or Deleting Experiments from the Correlation

To add an experiment, look in the “Experiments to be correlated together” box in the middle of the window. Your default experiment should already be there. Open the “Experiments” folder in the Navigator of the “Clustering” window. Click over an experiment set or an interpretation, highlighting it. Click the “Add>>” button in the “Experiments to Use” section of the “Clustering” window. This will bring up the “New Experiment” window.

![Figure 2-5 The “New Experiment” window](image)

You can change the weight of the experiment in this box. Clicking the “OK” button will add this experiment to the list in the white box under “Experiments to Use”. If you want to change the weight of the experiment after it is added to the list, double click its name in the “Experiments to Use” box and change the weight in the “New Experiment” pop-up menu.

Correlations of multiple experiments are done through a weighted correlation, in which you specify the weight of each experiment. You may make one experiment or experiment set more important than another. If all of the experiments, or experiment sets, are given the same weight, they will be averaged equally. The name of the experiment is noted directly after its relative weight. In Figure 2-6 the “test” experiment is given a weight of 1.0 verses the “NIH Spinal Cord Study” experiment, which is given a weight of 2.0. Therefore, in this example, the correlations found in the “NIH Spinal Cord Study” will be twice as influential in creating the tree as the correlations between the genes in “test” study.

The equation used to determine the overall correlation is:

\[
X = \frac{(Aa + Bb + Cc + \ldots)}{(a + b + c + \ldots)}
\]

- \(A\) is the correlation coefficient between the gene in question in experiment 1 and the gene named in the “Experiments to Use” box, also from experiment 1.
- \(a\) is the weight specified for experiment 1.
- \(B\) is the correlation coefficient of the gene in question in experiment 2, to the gene named in the title bar, also from experiment 2.
- \(b\) is the weight associated with experiment 2.
- \(C\) is the correlation coefficient of the gene in question in experiment 3 to the gene named in the title-bar, also from experiment 3.
- \(c\) is the weight associated with experiment 3.

and so on.
Experiments 1, 2, 3, and so forth, are all of the experiments selected in the white “Correlations” box. If $X$ is between the minimum and maximum correlations specified in the “Clustering” window, then the gene in question passes the correlations.

To delete an experiment, click the name of the experiment in the white “Experiments to Use” window, highlighting it. Select “Remove Selected” from the “Experiments” drop-down menu.

![Figure 2-6 The “Clustering” window with experiment/interpretation pop-up menu](image)

To see the pop-up menu in Figure 2-6, right click over an experiment.
2.3.2 Restricting the Genes to be correlated

The top white box is labeled “Additional restrictions on genes for clustering”. This indicates the restrictions placed upon the set of genes to be correlated, before they ever reach the correlation stage. To add restrictions, right click any experiment in the Clustering window Navigator and select one of the significant restriction options in the pop-up menu presented in Figure 2-6. These restrictions are the same as in the “Complex Correlations” window, please see GeneSpring User Manual 8.6, “Making Lists with the Complex Correlation Command” or GeneSpring User Manual 7.2, “Placing Restrictions on a List” for details.
2.3.2.1 Restrictions Over an Entire Experiment or Interpretation

- **Expression Percentage Restriction**: finds genes with certain values in some proportion of the experiments in this set.

![Image of Expression Percentage Restriction window](image)

**Figure 2-8 The Expression Percentage Restriction window for experiments**

Highlight the default number in the Strength box (in this instance 0) and type in your minimum control strength required for the list you are making. Click the “OK” button and the restriction will be added to the list in the big white box.

You can change the number in the “In at least [ ] out of total Experiments”. This line can refer to the whole experiment set or it can be for a single experiment.
• **Data File Restriction:** looks at the column names and allows some to be chosen or restricted.

![Data File Restrictions window](image)

**Figure 2-9 The Data File Restriction window for experiments**

2.3.2.2 **Restrictions over a condition or sample**

• **Expression Restriction:** finds genes with certain values at a particular condition.

![Expression Level Restrictions window](image)

**Figure 2-10 The Expression Restriction window for conditions**
• **Condition to Condition Comparison Restriction**: finds genes based on two conditions.

![Figure 2-11 The Condition to Condition Comparison Restriction window](image)

**2.3.2.3 Data File Restriction**

The Data File Restriction looks at the column names and allows some to be used or restricted. The Data File Restriction finds genes based on any column from the original data files for this experiment.

![Figure 2-12 The Data File Restriction window for conditions](image)

You can select any column name from your experiment (GeneSpring automatically scans all of the columns in your files and allows you to select both columns of interest and restriction values) from the Column drop-down menu. If not you will need to enter the column number in the Num-
ber box. If you have access to the original data files entered in GeneSpring, you can check them for column numbers.

For example, if you had entered an Affymetrix file as your experiment, you could use the drop-down menu to select the Abs/call column and select for all entries equal to “M” if you wanted to make a list of just the marginal data.

2.3.2.4 Restricting only certain types of data

You can also change the type of data the restriction applies to. Currently the types are as listed under a drop-down list in the applicable windows.

- **Normalized Data:** see *GeneSpring Loading Data Manual* Chapter 1, “Normalizing Options” for more information on normalizing options during the data loading process. See 5.4 “Experiment Normalizations” on page 5-17 for more information on changing the normalization methods from within GeneSpring.

- **Raw Data:** The unnormalized experimental data. (In the original data file as put into GeneSpring, you must use the correct type of decimal markers for your computers’ settings. If your computer is set for a non-English language that typically uses commas for decimal markers, GeneSpring will recognize this. If, for example, your computer is set for French, the comma will be recognized as a decimal marker. You cannot use comma and periods interchangeably.)

- **Control Signal:** The normalization factor. Please refer to *GeneSpring User Manual* 5.4, “Experiment Normalizations” for information on the various positive and negative controls used in GeneSpring’s normalization features.

- **Number of Replicates:** the number of replicates can vary depending on your original data entered into GeneSpring and your current interpretation.

- **Range of Normalized Data:** The difference between the minimum and maximum of the normalized data. You can use the Range of normalized data feature if you want genes with (for example) a compact range of data. This range is how much variability is on a single measurement, not in the mean expression level.

  If your original data did not include measurement flags, you can use this feature to filter out “absent” genes by specifying a value 0 or above because Absent genes are not assigned any value.

- **Standard Error of Normalized Data:** Standard Error is calculated to +/- the standard error of the mean. It is calculated by taking the mean of several measurements. This represents the
precision of the mean of the replicates at that experimental condition. It is calculated as the standard deviation of a single measurement divided by the square root of the number of replicates.

- **Standard Deviation of Normalized Data:** Standard Deviation error bars are drawn to +/- one standard deviation, and represent the variability in an individual measurement. You can use Standard Deviation of normalized data if you want genes within a certain range of data. The range may still be quite large; one standard deviation encompasses about 68% of the data. Standard deviation is only applicable to a single measurement; not the mean expression level.

- **T-test probability:** A low value (for example, less than 0.05) means good replication away from normality (usually 1).

Normalized, Control and Raw data are also displayed in the upper right corner of the Gene Inspector window. In the fictitious Yeast Extraterrestrial Study, if you set the Kryptonite concentration parameter as continuous and the variety of yeast as non-continuous, you will see something like Figure 2-14 in the Gene Inspector.

<table>
<thead>
<tr>
<th>Description</th>
<th>Normalized</th>
<th>Control</th>
<th>Raw</th>
<th>T-test p-value</th>
<th>Flags</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Kryptonite concentration 0 ppm, Variety of yeast A, Andromeda strain infection Andromeda</td>
<td>6.215 (6.341 to 7.09)</td>
<td>0.15</td>
<td>1.007 (0.951 to 1.063)</td>
<td>0.04167/0.0436</td>
<td>P</td>
</tr>
<tr>
<td>2. Kryptonite concentration 0 ppm, Variety of yeast A, Andromeda strain infection Healthy</td>
<td>6.169 (6.071 to 6.267)</td>
<td>0.15</td>
<td>0.974 (0.920 to 1.031)</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>3. Kryptonite concentration 10 ppm, Variety of yeast A, Andromeda strain infection Andromeda</td>
<td>6.529 (6.033 to 7.026)</td>
<td>0.15</td>
<td>0.979 (0.905 to 1.054)</td>
<td>0.06701/0.039</td>
<td>P</td>
</tr>
<tr>
<td>4. Kryptonite concentration 10 ppm, Variety of yeast A, Andromeda strain infection Healthy</td>
<td>6.928 (6.839 to 7.013)</td>
<td>0.15</td>
<td>1.024 (0.965 to 1.083)</td>
<td>0.02553/0.089</td>
<td>P</td>
</tr>
<tr>
<td>5. Kryptonite concentration 20 ppm, Variety of yeast A, Andromeda strain infection Andromeda</td>
<td>6.913 (6.892 to 7.134)</td>
<td>0.15</td>
<td>0.982 (0.914 to 1.07)</td>
<td>0.05923/0.237</td>
<td></td>
</tr>
<tr>
<td>6. Kryptonite concentration 20 ppm, Variety of yeast A, Andromeda strain infection Healthy</td>
<td>8.114 (7.289 to 9.502)</td>
<td>0.15</td>
<td>1.017 (1.001 to 1.335)</td>
<td>0.06144/0.994</td>
<td>P</td>
</tr>
<tr>
<td>7. Kryptonite concentration 30 ppm, Variety of yeast A, Andromeda strain infection Andromeda</td>
<td>8.845 (8.886 to 7.022)</td>
<td>0.15</td>
<td>1.027 (1.019 to 1.063)</td>
<td>0.02926/0.836</td>
<td>P</td>
</tr>
<tr>
<td>8. Kryptonite concentration 30 ppm, Variety of yeast A, Andromeda strain infection Healthy</td>
<td>9.960 (9.818 to 8.750)</td>
<td>0.15</td>
<td>1.044 (0.774 to 1.114)</td>
<td>0.18659/0.085</td>
<td>P</td>
</tr>
<tr>
<td>9. Kryptonite concentration 40 ppm, Variety of yeast A, Andromeda strain infection Andromeda</td>
<td>6.625 (6.416 to 6.836)</td>
<td>0.15</td>
<td>0.894 (0.962 to 1.025)</td>
<td>0.02722/0.299</td>
<td>P</td>
</tr>
<tr>
<td>10. Kryptonite concentration 40 ppm, Variety of yeast A, Andromeda strain infection Healthy</td>
<td>5.408 (4.121 to 6.698)</td>
<td>0.15</td>
<td>0.818 (0.681 to 1.039)</td>
<td>0.18103/0.373</td>
<td>P</td>
</tr>
<tr>
<td>11. Kryptonite concentration 50 ppm, Variety of yeast B, Andromeda strain infection Andromeda</td>
<td>6.768 (6.888 to 6.896)</td>
<td>0.15</td>
<td>1.015 (1.011 to 1.028)</td>
<td>0.01108/0.493</td>
<td>P</td>
</tr>
<tr>
<td>12. Kryptonite concentration 50 ppm, Variety of yeast B, Andromeda strain infection Healthy</td>
<td>5.418 (5.034 to 5.799)</td>
<td>0.15</td>
<td>0.812 (0.505 to 1.028)</td>
<td>0.19313/0.538</td>
<td>P</td>
</tr>
<tr>
<td>13. Kryptonite concentration 10 ppm, Variety of yeast B, Andromeda strain infection Andromeda</td>
<td>6.203 (5.751 to 6.636)</td>
<td>0.15</td>
<td>0.846 (0.962 to 1.025)</td>
<td>0.06096/0.942</td>
<td>P</td>
</tr>
<tr>
<td>14. Kryptonite concentration 10 ppm, Variety of yeast B, Andromeda strain infection Healthy</td>
<td>6.529 (6.812 to 6.247)</td>
<td>0.15</td>
<td>0.829 (0.722 to 0.937)</td>
<td>0.08958/0.797</td>
<td>P</td>
</tr>
<tr>
<td>15. Kryptonite concentration 20 ppm, Variety of yeast B, Andromeda strain infection Andromeda</td>
<td>6.317 (6.718 to 6.868)</td>
<td>0.15</td>
<td>0.847 (0.985 to 1.013)</td>
<td>0.06971/0.965</td>
<td>P</td>
</tr>
<tr>
<td>16. Kryptonite concentration 20 ppm, Variety of yeast B, Andromeda strain infection Healthy</td>
<td>5.639 (5.812 to 7.445)</td>
<td>0.15</td>
<td>0.848 (0.665 to 1.07)</td>
<td>0.03942/0.492</td>
<td>P</td>
</tr>
<tr>
<td>17. Kryptonite concentration 30 ppm, Variety of yeast B, Andromeda strain infection Andromeda</td>
<td>6.946 (6.751 to 6.314)</td>
<td>0.15</td>
<td>0.807 (0.962 to 0.951)</td>
<td>0.0371/0.921</td>
<td>P</td>
</tr>
<tr>
<td>18. Kryptonite concentration 30 ppm, Variety of yeast B, Andromeda strain infection Healthy</td>
<td>4.519 (2.168 to 8.892)</td>
<td>0.15</td>
<td>0.877 (0.335 to 1.028)</td>
<td>0.37481/0.287</td>
<td>P</td>
</tr>
<tr>
<td>19. Kryptonite concentration 40 ppm, Variety of yeast B, Andromeda strain infection Andromeda</td>
<td>5.747</td>
<td>0.15</td>
<td>1.065</td>
<td>0.0053/0.053</td>
<td>P</td>
</tr>
<tr>
<td>20. Kryptonite concentration 40 ppm, Variety of yeast B, Andromeda strain infection Healthy</td>
<td>4.952 (4.333 to 12.55)</td>
<td>0.15</td>
<td>1.424 (0.965 to 1.882)</td>
<td>0.23903/0.373</td>
<td>P</td>
</tr>
</tbody>
</table>

Figure 2-14 Data display in the Gene Inspector
2.3.3 Similarity Definitions

The equations used to determine the nine types of correlations are described in detail in Chapter 1, “Equations for Correlations and other Similarity Measures”.

The default correlation is the Standard Correlation, \( \text{Standard correlation} = \frac{a \cdot b}{|a||b|} \).
2.3.4 Making Trees

The “Start” button creates a tree fitting all the criteria given in the “Clustering” window. Clicking the “Start” button will not close the “Clustering” window, so you can begin planning another tree immediately. Changing the information given in the “Clustering” window after you have started clustering a tree (i.e., after you have clicked the “Start” command) does not change the parameters of the tree in the process of being made. Changing the parameters displayed changes the parameters required for the next tree you make from this window. The “Close” button, at the bottom of the window, closes the “Clustering” window. This will not halt the making of a tree currently in the process of clustering. You cannot start clustering a new tree while there is already one in the process of being computed.

Select a folder and name your tree and it will be shown in the Navigator.

![Figure 2-16 Completed gene tree for “Test” and “NIH Spinal Cord Study”](image)
2.3.5 Creating a Subtree or List from a Node

Once you have a tree, you can right click over the nodes to make a new list or a mini-tree of the genes in that branch to get the pop-up menu displayed in Figure 2-16.

Select the “Make Subtree” option and GeneSpring will bring up just the genes that node connects to in a “Name New Gene Tree” window.

Select the “Make list” option and GeneSpring will bring up just the genes that node connects to in a “New Gene List” window.

Figure 2-17 A new Tree from a Node

Select the “Make list” option and GeneSpring will bring up just the genes that node connects to in a “New Gene List” window.
Figure 2-18 A new List from a Node
2.3.6 Minimum Distance and Separation Ratios

To make a tree, GeneSpring calculates the correlation for each gene with every other gene in the set. Then it takes the highest correlation and pairs those two genes, averaging their expression profiles. GeneSpring then compares this new composite gene with all of the other unpaired genes. This is repeated until all of the genes have been paired. At this point the minimum distance and the separation ratio come in to play. Both of these commands affect the branching behavior of the tree. The minimum distance deals with how far down the tree discrete branches are depicted. A value smaller than 0.001 has very little effect, because most genes are not correlated more closely than that. A higher number will tend to lump more genes into a group, making the groups less specific. The separation ratio determines how large the correlation difference between groups of clustered genes has to be for them to be considered discrete groups, and not be lumped together. This number should be between 0 and 1.

![Figure 2-19 The effects of minimum distances and separation ratios](image)

- md=0.001 and sr=0.1
- md=0.1 and sr=0.5
- md=0.01 and sr=0.5
- md=.001 and sr=0.5 (Default)
- md=.001 and sr=1.0
- md=.25 and sr=0.5
• **Separation Ratio**

The separation ratio determines how large the correlation difference between groups of clustered genes has to be for the groups to be considered discrete groups and not be joined together.

- Increasing separation increases the ‘branchiness’ of the tree.
- Default separation ratio is 0.5. Separation ratio can range from 0.0 to 1.0.
- At a separation ratio of 0, all gene expression profiles can be regarded as identical.

• **Minimum Distance**

The minimum distance deals with how far down the tree discrete branches are depicted. A higher number will tend to lump more genes into a group, making the groups less specific.

- Decreasing minimum distance increases the ‘branchiness’ of the tree.
- Default minimum distance is 0.001. A value smaller than .001 has very little effect, because most genes are not correlated more closely.
- In the last example presented in Figure 2-19, a minimum distance of 0.25 (at separation ratio=0.5) treats all expression profiles as equally different (i.e. all diverge once from a single branch).

It is not normally appropriate to change separation ratio or minimum distance.

### 2.4 Tree Commands

Please see *GeneSpring User Manual* 8.6, “Making Lists with the Complex Correlation Command”, and ignore the commands with “Correlation” in the name.

- **Add/Remove Experiment to/from the Correlation:** The “Add>>” and “<<Remove” buttons are the first step in adding or deleting a single experiment or set of experiments to the correlation. To add an experiment (or set of experiments) to be correlated, go to the “Experiments” folder in the Navigator. Open the folder by clicking the plus icon next to the folder. Then, select the experiment you want to add. Click the “Add>>” button. See “New Experiment Window” command, below, for details. Clicking the “OK” button in the window adds the experiment to the “Experiments to Use” list.

- **New Experiment Window:** (See Figure 2-5.) This window is the result of either selecting one of the experiments for correlation by right clicking or by double clicking one of the experiments in the white “Experiments to Use” box. In this window you specify a weight for the selected experiment. The default weight of an experiment is 1. To change this default weight, click in the weight box, highlight the existing weight and type in a new one. The “OK” button at the bottom of the window accepts the value indicated and adds the experiment and its associated weight to the correlation list. The experiment name and weight will then appear in the white “Experiments to Use” box in the “Clustering” window. If you reached the “New Experiment” window by double clicking the experiment name displayed in the white “Experiments to Use” box in the “Clustering” window, the “OK” button accepts any changes you have made to the weight. It then correspondingly changes the line for the experiment in the white “Experiments to Use” box. The “Close” button at the bottom of the window dismisses the window.
• **Remove Selected Experiment:** This command deletes an experiment from the list of experiments being correlated. To do this, select the experiment you wish to remove in the white “Experiments to Use” box, highlighting it. Go to the “Experiments” menu in the menu-bar and click “Remove Selected”. If an experiment is not currently selected (highlighted) this command is disabled.

• **Minimum distance:** The number specified in the “Minimum distance” box determines the minimum separation considered significant between genes. This reduces meaningless structure at the base of the tree. Decreasing minimum distance increases the ‘branchiness’ of the tree. The default minimum distance is 0.001. A value smaller than .001 has very little effect, because most genes are not correlated more closely. In most experiments a minimum distance of 0.25 (at separation ratio=0.5) treats all expression profiles as equally different (i.e. all diverge once from a single branch). In Figure 2-1 the minimum distance is set to 0.001, which is the default value. To change this number move the cursor into the white box next to the “Minimum distance” label, and click in the box, then use the keyboard to alter the text, just like using a word processing program. You will not normally want to modify this.

• **Separation Ratio:** The number specified in the “Separation Ration” box determines the extent to which the tree is ‘branchy’. The separation ratio determines how large the correlation difference between groups of clustered genes has to be for them to be considered discrete groups, and not be lumped together. The increasing separation increases the ‘branchiness’ of the tree. The default separation ratio is 0.5. The separation ratio can range from 0 to 1.0. At a separation ratio of 0, all gene expression profiles can be regarded as identical. In Figure 2-1 the maximum correlation is set to 0.5, which is the default value. To change this number click in the white box next to the “Separation Ration” label, and type in a new value. You will not normally want to modify this.

• **Start:** Clicking the “Start” button initiates the process of building a gene tree according to the parameters specified in the “Clustering” window. Restrictions on the genes’ expression levels and control strengths limit which genes in the experiments being correlated are considered, and may change which experiments are clustered together as similar, and how similar they are considered. To do this, GeneSpring clusters the genes according to their expression profiles. When this is occurring, a window, labeled “Clustering Progress” appears. The “Clustering Progress” window contains a blue bar indicating how far along in the clustering process GeneSpring is. You may stop the clustering process at any point, and thereby stop making the gene tree, by closing the “Clustering Progress” window. Do this by clicking the “X” button at the far right of the title bar for that window. Closing the “Clustering” window at this point does not cancel making the gene tree.

• **Close:** Clicking the “Close” button at the bottom of the “Clustering” window does not cancel a tree in the process of clustering, but simply closes the Cluster window. Restrictions and other selected items in the “Clustering” window will not be saved after closing.
2.5 References


Chapter 3  K-Means Clustering

3.1 Function

K-means clustering divides the genes into distinct groups based on expression pattern. Genes are initially divided into some number (k) of user-defined and equal-sized groups, in gene order. Centroids are calculated for each group as averages of the expression profiles. Genes are reassigned to the group with the highest similarity between the expression profile for the gene and the group centroid. Group centroids are then recalculated, and the process is iterated until the group compositions converge. A wide selection of similarity measures (parametric and non-parametric correlations, Euclidean distance, etc.) is available.

3.2 Description

You can cluster from any gene list. Smaller gene lists will take less time. To begin the K-means clustering, select Tools > “Cluster...”. You will get a pop-up like Figure 3-1.

Figure 3-1 The “K-means initialization” window
In the bottom “K-Means” box, enter the number of classes you would like by highlighting over the default number (5, in the example above) and typing in the number you want. Select a similarity measure by clicking the black arrow at the end of the box and scrolling through the options.

![K-Means Initialization Window](image)

**Figure 3-2** The “K-means initialization” window, showing the similarity measures

- **Number of Clusters**: This is the number of clusters your genes will be divided into. You can choose as many as you like, but larger numbers may lead to difficulties in interpretation. There is no upper limit for this number.

- **Maximum Iterations**: From the “Cluster” window, you can change the number of iterations of the k-means process GeneSpring will put your data through. If your k-means do not converge in this number of iterations, you can re-start the k-means from here by selecting the “Start from Current Classifications” and specifying the number of additional iterations.

- **Measure similarity by**: Please refer to Chapter 1, “Equations for Correlations and other Similarity Measures” for details comments on the nine possible correlations. There are many instances when you will want to define “nearness” for gene expression patterns. There are many ways to define nearness, and GeneSpring allows you to select one of them from the list shown in Figure 3-2.

  The default setting is the **Standard correlation** = $a.b/(|a||b|)$

- **Start from Current Classification**: This instructs GeneSpring to use the current classification. This option will only be available if you have a classification selected.

- **Animate Display While Clustering**: Selecting this option will show colors in the main Genome Browser reflecting the cluster assignments as the algorithm proceeds. This may slow your analysis slightly.
3.2.1 Adding or Deleting Experiments from the Correlation

To delete an experiment, click the name of the experiment in the white “Experiments to Use” window, highlighting it. Select Experiments > “Remove Selected”.

![Figure 3-3 The “Experiments to Use” Box](image)

To add an experiment, the “Clustering” window will start off showing the experiment from the main browser window already in the “Experiments to be correlated together” box in the middle of the window. Open the “Experiments” folder in the Navigator of the “Clustering” window. Click over an experiment set or an interpretation, highlighting it. Click the “Add>>” button in the “Experiments to Use” section of the “Clustering” window. This will bring up the “New Experiment” window.

![Figure 3-4 The “New Experiment” window](image)

You can change the weight of the experiment in this box. Clicking the “OK” button will add this experiment to the list in the white box under “Experiments to Use”. If you want to change the weight of the experiment after it is added to the list, double click its name in the “Experiments to Use” box and change the weight in the “New Experiment” pop-up menu.

Correlations of multiple experiments are done through a weighted correlation, in which you specify the weight of each experiment. Weight will make one experiment or experiment set more important than another. If all of the experiments, or experiment sets, are given the same weight, they will be averaged equally. The name of the experiment is noted directly after its relative weight. In Figure 3-5 the “test” experiment is given a weight of 1.0 versus the “NIH Spinal Cord Study” experiment, which is given a weight of 2.0. Therefore, in this example, the correlations found in the “test” will be twice as influential in creating the tree as the correlations between the genes in NIH Spinal Cord Study.
3.2.2 Restricting the Genes to be correlated

The top white box is labeled “Additional restrictions for clustering”. This indicates the restrictions placed upon the set of genes to be correlated, before they ever reach the correlation stage. To add restrictions, right click the experiments in the Navigator and select one of the four significant restriction options in the pop-up menu presented in Figure 3-3. Of the four types of gene restriction you may choose; two are for entire experiments and two are for a single sample. These restrictions are the same as in the “Filter Genes” window, see GeneSpring User Manual 7.1, “Making Lists with the Filter Genes Window” for details.
3.2.2.1 Restrictions Over an Entire Experiment or Interpretation

- **Expression Percentage Restriction**: finds genes with certain values in some proportion of the experiments in this set.
Figure 3-7 The Expression Percentage Restriction window for experiments

Highlight the default number in the Strength box (in this instance 0) and type in your minimum signal strength required for the list you are making. Click the “OK” button and the restriction will be added to the list in the big white box.

You can change the number in the “In at least [ ] out of total Experiments”. This line can refer to the whole experiment set or it can be for a single experiment.

- **Data File Restriction**: looks at the column names and allows some to be choose or restricted.

Figure 3-8 The Data File Restriction window for experiments

- **Statistical Group Comparison Restriction**: finds genes having significantly different mean expression values in two groups. You will need to specify which parameter is to be used for the comparison, and which two values of the parameter are to be compared.
This filter function statistically compares mean expression levels between two or more groups of samples. This comparison is performed for each gene, and the genes with the most significant differences (smallest p-value) are returned. The comparisons can be done with parametric or non-parametric methods. The parametric comparison for 2 groups is known as Student’s two-sample t-test. For multiple groups, this is known as one-way analysis of variance (ANOVA). The user can specify whether to assume within-group variances are equal across all groups. Calculations without the assumption of equality of variances are done using Welch’s approximate t-test and ANOVA. Non-parametric comparisons are also available, corresponding to the Wilcoxon two-sample text (also known as the Mann-Whitney U test) for two groups, and the Kruskal-Wallis test for multiple groups.

3.2.2.2 Restrictions over a Single Condition or Sample

- **Expression Restriction**: finds genes with certain values at a particular experimental point.
**Condition to Condition Comparison Restriction:** finds genes based on two conditions.

---

Figure 3-11 The Expression Restriction window for conditions

Figure 3-12 The Condition to Condition Comparison Restriction window
• **Data File Restriction**: looks at the column names and allows some to be used or restricted.

![Data File Restrictions](image)

Figure 3-13 The Data File Restriction window for conditions

### 3.2.2.3 Restricting only certain types of data

You can also change the type of data the restriction applies to. Currently the types are as listed under a drop-down list in the applicable windows.

![Application options for restricted data](image)

Figure 3-14 Application options for restricted data

• **Normalized Data**: see *GeneSpring Loading Data Manual* Chapter 1, “Normalizing Options” for more information on normalizing options during the data loading process. See 5.4 “Experiment Normalizations” on page 5-17 for more information on changing the normalization methods from within GeneSpring.

• **Raw Data**: The unnormalized experimental data. (In the original data file as put into GeneSpring, you must use the correct type of decimal markers for your computers’ settings. If your computer is set for a non-English language that typically uses commas for decimal markers, GeneSpring will recognize this. If, for example, your computer is set for French, the comma will be recognized as a decimal marker. You cannot use comma and periods interchangeably.)

• **Control Signal**: The normalization factor. Please refer to *GeneSpring User Manual* 5.4, “Experiment Normalizations” for information on the various positive and negative controls used in GeneSpring’s normalization features.

• **Number of Replicates**: the number of replicates can vary depending on your original data entered into GeneSpring and your current interpretation.
Advanced Analysis Techniques Manual

K-Means Clustering Description

Chapter 3-10 Copyright 2000-2001 Silicon Genetics

- **Range of Normalized Data**: The difference between the minimum and maximum of the normalized data. You can use the Range of normalized data feature if you want genes with (for example) a compact range of data. This range is how much variability is on a single measurement, not in the mean expression level.

  If your original data did not include measurement flags, you can use this feature to filter out “absent” genes by specifying a value 0 or above because Absent genes are not assigned any value.

- **Standard Error of Normalized Data**: Standard Error is calculated to +/- the standard error of the mean. It is calculated by taking the mean of several measurements. This represents the precision of the mean of the replicates at that experimental condition. It is calculated as the standard deviation of a single measurement divided by the square root of the number of replicates.

- **Standard Deviation of Normalized Data**: Standard Deviation error bars are drawn to +/- one standard deviation, and represent the variability in an individual measurement. You can use Standard Deviation of normalized data if you want genes within a certain range of data. The range may still be quite large; one standard deviation encompasses about 68% of the data. Standard deviation is only applicable to a single measurement; not the mean expression level.

- **T-test probability**: A low value (for example, less than 0.05) means good replication away from normality (usually 1).

Normalized, Control and Raw data are also displayed in the upper right corner of the Gene Inspector window. In the fictitious Yeast Extraterrestrial Study, if you set the Kryptonite concentration parameter as continuous and the variety of yeast as non-continuous, you will see something like Figure 3-15 in the Gene Inspector.

<table>
<thead>
<tr>
<th>Description</th>
<th>Normalized</th>
<th>Control</th>
<th>Raw</th>
<th>T-test p-value</th>
<th>Flanks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Kryptonite concentration 0 ppm, Variety of yeast A, Andromeda strain infected Andromeda</td>
<td>6.71±6.34</td>
<td>6.15±6.93</td>
<td>6.09±6.93</td>
<td>0.007±0.95</td>
<td>Healthy</td>
</tr>
<tr>
<td>2. Kryptonite concentration 0 ppm, Variety of yeast A, Andromeda strain infected Healthy</td>
<td>6.15±6.93</td>
<td>6.15±6.93</td>
<td>6.15±6.93</td>
<td>0.95±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>3. Kryptonite concentration 10 ppm, Variety of yeast A, Andromeda strain infected Andromeda</td>
<td>6.53±6.34</td>
<td>6.03±6.93</td>
<td>6.05±6.93</td>
<td>0.95±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>4. Kryptonite concentration 10 ppm, Variety of yeast A, Andromeda strain infected Healthy</td>
<td>6.03±6.34</td>
<td>6.05±6.93</td>
<td>6.05±6.93</td>
<td>0.95±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>7. Kryptonite concentration 30 ppm, Variety of yeast A, Andromeda strain infected Andromeda</td>
<td>6.94±6.34</td>
<td>6.05±6.93</td>
<td>6.05±6.93</td>
<td>0.95±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>8. Kryptonite concentration 30 ppm, Variety of yeast A, Andromeda strain infected Healthy</td>
<td>6.05±6.34</td>
<td>6.05±6.93</td>
<td>6.05±6.93</td>
<td>0.95±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>11. Kryptonite concentration 0 ppm, Variety of yeast B, Andromeda strain infected Andromeda</td>
<td>6.68±6.34</td>
<td>6.68±6.93</td>
<td>6.68±6.93</td>
<td>0.68±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>12. Kryptonite concentration 0 ppm, Variety of yeast B, Andromeda strain infected Healthy</td>
<td>6.68±6.34</td>
<td>6.68±6.93</td>
<td>6.68±6.93</td>
<td>0.68±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>15. Kryptonite concentration 20 ppm, Variety of yeast B, Andromeda strain infected Andromeda</td>
<td>6.94±6.34</td>
<td>6.94±6.93</td>
<td>6.94±6.93</td>
<td>0.94±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>16. Kryptonite concentration 20 ppm, Variety of yeast B, Andromeda strain infected Healthy</td>
<td>6.94±6.34</td>
<td>6.94±6.93</td>
<td>6.94±6.93</td>
<td>0.94±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>17. Kryptonite concentration 30 ppm, Variety of yeast B, Andromeda strain infected Andromeda</td>
<td>6.94±6.34</td>
<td>6.94±6.93</td>
<td>6.94±6.93</td>
<td>0.94±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>18. Kryptonite concentration 30 ppm, Variety of yeast B, Andromeda strain infected Healthy</td>
<td>6.94±6.34</td>
<td>6.94±6.93</td>
<td>6.94±6.93</td>
<td>0.94±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>19. Kryptonite concentration 40 ppm, Variety of yeast B, Andromeda strain infected Andromeda</td>
<td>6.94±6.34</td>
<td>6.94±6.93</td>
<td>6.94±6.93</td>
<td>0.94±6.93</td>
<td>Healthy</td>
</tr>
<tr>
<td>20. Kryptonite concentration 40 ppm, Variety of yeast B, Andromeda strain infected Healthy</td>
<td>6.94±6.34</td>
<td>6.94±6.93</td>
<td>6.94±6.93</td>
<td>0.94±6.93</td>
<td>Healthy</td>
</tr>
</tbody>
</table>

**Figure 3-15 Data display in the Gene Inspector**
### 3.2.2.4 Starting the K-Means analysis

Once you have entered your estimated number of classes and selected a similarity measure, click the “Start” button. Rather than the clustering bar you see with the Trees, in this screen you will get to watch the colored lines shuffle around for a minute (or several minutes, depending on how many genes are being correlated) while the screen updates.

![Figure 3-16 The graph display of a K-means clustered set](image)

Figure 3-16 The graph display of a K-means clustered set
3.3 Saving Clustering Results

Once the K-means clustering is done the graph view will be filled with different colors and the color legend on the right side of the screen defines the sets.

When the K-means clustering is done GeneSpring will automatically bring up a screen asking you where you would like to save the current clustering.

Name and save your new classification in the “Choose Classification Name” window.

If you wanted to save only one cluster, you can right click over one of the colors in the color bar and select “save” in the pop-up menu. Choose a folder and give your new cluster a sensible name. Click “OK” when you are done.

Open the “Classification” folder in the Navigator and look for the new name, to be sure it was saved properly.
3.4 Viewing Classifications

The best and least visually complicated way to view classifications is by using GeneSpring’s “Split Window” feature. You can view any classification by opening the “Classifications folder and selecting the classifications you would like to view. Then, right click over the name of the chosen classification in the Navigator and select Split Window > Both. The screen will change to several small subscreens. (You can also choose vertically or horizontally.) Please refer to Figure 3-25 and Figure 3-26 for examples of vertically and horizontally split windows.

You can select any view and each of the little screens will change to Graph, Ordered List, Classification or Physical Position, and so on. You can also change the list being viewed to get a more sensible display.
3.5 Making Lists from Classifications

Once it is saved as a classification, there are various things you can do with it. You can look at the groups individually by right clicking the classification you just created. Select “Make gene lists” and Genespring will create one list for each grouping.

Figure 3-19 Classification View
Open up the Gene Lists folder in the Navigator. There should be a folder with the name of the classification (default, “Calculated”). Click this list to open it. It will contain one entry for each group (default 5). Click one of these gene lists to view those genes. If you are not seeing them as graphs, you probably want to do this. Click View > “Graph” or Ctrl+4.
3.6 Clustering Commands

Zoom, select, panning and the “Gene Inspector” work as you would expect.

3.6.1 Colorbar Commands

Right clicking over the color bar on the far right of the screen produces a new menu of options.

- **Save**: Selecting the “Save” option from the pop-up menu will bring you to the “Choose Classification Name” window. Highlight over the default name (in the example below, calculated) and type in a sensible name. You will be able to view these lists again from the Classification folder in the Navigator.
If you want to make list of all the classifications at once, right click over the classification name in the Navigator and select “Make Gene Lists”.

- **Modify**: The Modify option is not currently enabled.
- **About**: Selecting the “About” option bring up the following menu in the default system.

![Figure 3-22 The “Choose Classification Name” panel in K-means clustering](image)
3.6.2 Pop-Up Menu Commands

By right clicking over the lists presented in the classification folder you can access the following commands:

- **Set as classification**: This will arrange all of the previously sorted classification into an order on the screen. You can then look at the animate feature (while in color by primary experiment) and watch as all the colors change at the same time points in the cycle.

- **Set as coloring scheme**: This will return the display to the colors set in the initial K-means clustering.
  You can also use these lists to define the colors in a Venn Diagram.
- **Bookmarks**: If you have not saved your clusters you may get a warning dialog box telling you that GeneSpring cannot bookmark unsaved data. Save the data, then bookmark.

- **Split Window**: You can view all of your classification by right clicking over the name in the Navigator and selecting “Split Window”. Figure 3-25 is split horizontally and Figure 3-26 is vertically.
• **Make gene lists:** Select this command and all of the classification are made into gene lists and stored in their own sub-folder under Gene Lists. Double clicking the name of the list in the Navigator will bring up the List Inspector window.

• **Properties:** This brings up the Properties/About box. By clicking the edit button you can change the data presented to suit your work. The name of your list is in the banner bar. You
can close the box without changes by clicking the “X” in the upper right corner. See Figure 3-23.

**Figure 3-26 Vertically Split Windows screen**

- **Attachments:** You can add a text or picture attachment to your k-means clusters. Select the “Attachment” command after right clicking over the name of the cluster.
- **Delete:** Selecting this will delete your list. You will get a warning dialog box before the list is deleted.

```plaintext
Attachments: You can add a text or picture attachment to your k-means clusters. Select the “Attachment” command after right clicking over the name of the cluster.
Delete: Selecting this will delete your list. You will get a warning dialog box before the list is deleted.
```
3.7 Classification Inspector for k-means

The Classification Inspector provides details on classifications produced by k-means clustering. The Classification Inspector can be viewed by double-clicking on a classification, or by right clicking over any Classification in the Navigator and selecting “Properties” from the resultant pop-up menu. The Inspector window shows a table of classes, including the number of genes in each class, the average class radius from its class centroid (in Euclidean distance), and the overall percentage of variance explained by the classification. The latter can be used to compare the effectiveness of different numbers of groups used in classification.

Right click over any classification and select the “Properties” option from the pop-up menu.

Figure 3-27 shows the Classification Inspector results for a classification from a k-means clustering with five groups. The name of the classification is shown at the top with other details and the (automatically generated) notes about the source of the classification. You can change the name of your classification in the Name box or add notes of your own in the Notes text box. These changes will be saved when you click the OK button. If you do not want to save changes, close the window by clicking the small black x in the upper right corner.
Next is a table with one row for each class in the classification (set1…set5), plus the unclassified genes, and genes together (All classes). The number of genes in the class is shown (Genes), plus the root mean square Euclidean distance of the gene expression profiles for the genes in the class around the expression profile (or centroid) of the class (Average Radius). Classes with a small Average Radius are tightly grouped, classes with a large Average Radius are more spread out.

At the bottom is a measure of the percent of variability explained by this classification. Effective classifications explain much of the variability. The measure GeneSpring calculates is based on the Calinski and Harabasz index of quality, expressed as:

\[ G = \frac{B}{(c-1)} \frac{1}{W/(n-c)} \]

Where \( B \) = sum of the squared distances between the cluster centroids and the overall data mean, \( W \) = sum of the squared distances of the individual genes from their cluster centroids, \( n \) = number of genes, and \( c \) = number of groups. GeneSpring reports the quantity

\[ E = 100 \left( \frac{G}{1 + G} \right) \]

which can thought of as the percentage of explained variability, adjusted for the number of clusters. This can be compared across different classifications with different numbers of groups as a measure of effectiveness. The correction for the number of classes in the formula means that the
percentage variance explained will not increase without limit as the number of classes increases, so you can try different numbers of classes and see which is most effective. For purposes of this calculation, genes that are unclassified are treated as a separate class, but are counted in the index. As an example, in the data example shown in Figure 3-27, the six group clustering explained 60.6% of the variability. A 10-group clustering explains 88.5% of the variability. But when you do 20 classes, the variability explained is only 87.7%, so the data suggests that 10 clusters is a more effective classification of the genes than 20. Further examination shows that the percentage explained variation rises slightly to a peak near 15, with 89% variation explained, though the small gain in explained variation over 10 groups may not justify the complexity of reporting 15 groups.

3.8 References


Chapter 4  Pathways

Visual Pathways allow you to potentially find genes on a pathway related to your experiment. You can draw pathway pictures (JPEG or GIF images) or use publicly available pathways and place the genes in them.

4.1 Description

You can look at expression data from an experiment overlaid onto a previously made pathway image. The Pathways can be used to confirm and display hypotheses on such things as regulatory and metabolic pathways. With GeneSpring, you can import your own pathway images and specify the position of genes on them. In the default installation of the yeast genome, a picture of a cell cycle overlaid with some genes believed to be involved in the cell cycle is provided.

The following example is from the Yeast genome. Select the “all genes” list in the “Gene List” folder of the Navigator so you are not restricting which genes are displayed. Open the “Pathways” folder, and the “Cell growth & division” subfolder. Select the “mitosis” option.

![Figure 4-1 A Pathway display of the cell cycle](image-url)
Figure 4-1 and Figure 4-2 represent a picture of the cell cycle displayed as you might see it in a textbook. In Figure 4-2 the Genome Browser has been enlarged so that the names of the genes are visible. Some cyclin and other genes are shown, as well as pictures of what the cells are expected to be doing when the cell is in that phase. In Figure 4-2, at 21.1 minutes, you can see the genes believed to be involved in S phase are active (colored in red). If you adjust the time with the scroll bar on the bottom of the window, you can see that the phase changes and the active genes change. You can correlate these activities with the pictures in the lower right corner, to see that the observed expressions match the predicted phase based upon the real cell state.

![Pathway display of the cell cycle in Metaphase](image)
Figure 4-3 shows the same image during metaphase. The active (red) genes are now further around the cycle.

**Figure 4-3 A Pathway display of the cell cycle advanced in time**
4.2 New Pathways

For analysis, you can try to extend an existing pathway by finding more genes that could conceivably fit on the pathway. To do this, you load a pathway, and an experiment that is likely to involve changes in that pathway. You then right click on a place in the pathway that you consider to be interesting, and choose the option “Find genes that could fit here”. GeneSpring will then try to find genes that it thinks could fit there based upon the pathway and the expression values in the current experiment. The purpose of this is to allow a powerful question to be asked in a simple way. This can be asked of any pathway and any experiment, but of course there is no guarantee that it will actually produce results. If the experiment has nothing to do with the pathway, then the genes in the pathway will probably not be affected, and there is no information in the data that could lead to an extension of that pathway.

The way that this extension is done is by investigating each gene, and seeing if it reasonably fits on that point in the pathway. A gene is considered to be a candidate fit if the physical distances on the picture between the position of the place clicked and the other genes placed on the picture have a similar relationship to each other as the abstract distance in expression space between that gene and the other genes in the pathway. The abstract distance in expression space is like a similarity measure between expression patterns.

4.2.1 Adding a gene to a pathway

To make a new Pathway you must have an accessible picture (in .jpg or .gif format). You could place the file in the data folder of the relevant genome, but if you do not, GeneSpring will copy it into the correct folder. In this example the file is in default directory, C:\Program Files\SiliconGenetics\GeneSpring\data\yeast\Pathways.

From the File drop-down menu, select the “New Pathway...” command. It will bring up the familiar “Find Image File” dialog box. Browse for the file you want to use and select it. The picture will be brought into GeneSpring. You can rename your picture, if you like.
By holding down the Control key [CTRL] and using the cursor, you can draw a small rectangle. Then the New Genes box will appear.

- **Mac Users:** Use Option-click to draw a new element.

You can then type in the a word or part of a word expected to be in the descriptor of the gene you would like to place there and GeneSpring will bring up a list of all the genes including that word or partial word. For example, in the sample genome. The Keyword “metab” generated the following list:
Figure 4-6 The “Found Multiple Results” box for “metab”

Clicking any of these genes will cause GeneSpring to place that gene into the shape you just drew. The list will vanish, if you want another gene with that same keyword, you must repeat the process.

You can also search by accession number.

If you want to display another name for that gene in your pathway you will need to rename that gene within the genome, reload the genome and begin the process again. GeneSpring can only display the “true” name it knows.
4.2.2 Deleting a Pathway Element

You can also delete the genes you have placed in the picture. GeneSpring will ask you for confirmation before deleting.

4.2.3 The “Find Genes which could fit here” command

If you select the “Find Genes which could fit here” menu option, GeneSpring will offer you a list of genes it considers to be statistically likely matches.

“Find Genes which could fit here” menu option finds genes more similar to genes on your diagram (close to the point you selected) than to other genes on your diagram. When you select a point and choose this option from the popup menu, GeneSpring makes two lists of genes from those currently displayed on your diagram.
1. List A is the two genes (or pseudo-genes, i.e. the average gene from a search) closest to your selected point on the diagram (in terms of screen distance - it doesn’t matter what picture you are showing).

2. List B is all other genes on the Pathway.

GeneSpring then examines all the genes on your currently selected gene list, and finds all genes whose minimum similarity (correlation) with genes on list A is higher than their maximum similarity with genes on list B. These genes are reported in a new gene list window for examination by you. You can then make a note of the gene you would like to place on your Pathway and go through the normal process of placing it as demonstrated in 4.2 “New Pathways” on page 4-4.

### 4.3 KEGG Pathways

To enter a KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway into GeneSpring you must first get the KEGG pathways from KEGG. These pictures and associated data are copyright, so Silicon Genetics cannot redistribute them. The represented pathways are not organism specific. The folders which correspond to organism-specific pathways (such as hsa for human) are not always recognized by GeneSpring because the annotation for some genes is in a modified format. Put the ‘maps’ folder in the “Pathways” subdirectory for your genome. This will include the .gif, .html, .ec, .conf, and .gene files. The .gif and .html files are the ones that will be utilized. When you next start up GeneSpring, the pathways should be present in the Pathway folder.

KEGG uses EC numbers to refer to genes rather than gene names. This has several consequences:

- If your organism is not one of the ones officially supported by KEGG you may still get something useful.
- If you do not have any EC numbers loaded for your genes (please refer to *GeneSpring Loading Data Manual* Chapter 3, “Installing a New Genome”), you will not get any genes on your pathways
- If the set of genes you are studying does not include genes known to have the EC numbers being shown on the diagram, there will be blank spaces on the diagrams.
4.4 Pathway Commands

4.4.1 Commands found in the Genome Browser

By right clicking the Pathway itself you will be given the following options:

- **Find Genes which could fit here**: This command will find genes that could fit on the pathway near where you right-clicked, based on expression patterns in the current experiment. See 4.2.3 “The “Find Genes which could fit here” command” on page 4-7 for details.

- **Delete Pathway Element**: This command will delete the gene where you right-clicked. You will see a confirmation of the delete before it happens.

- **Select**: By holding down the Shift key and left-clicking in the Pathway you can select several genes and make a list. Make a list by right clicking in the Genome Browser and selecting “Make list from selected genes” from the pop-up menu.

All of the usual zooming, panning, Gene Inspector and color options work in this view.

4.4.2 Commands found in the Navigator

By right clicking over the Pathway name in the Navigator you will be given the following options:

- **Display Pathway**: This will immediately take your screen to the Pathway view and that particular Pathway.

- **Properties**: This option will bring up the editable “Properties” box.

![Figure 4-8 The editable Properties box](image-url)
• **Attachments:** You can add a text or picture attachment to your Pathway. Select the “Attachment” command after right clicking over the name of the Pathway.

![Figure 4-9 The New Overlay Attachment box](image)

• **Make Gene List:** Selecting this option will bring up a “New Gene List” window with all the genes present on this pathway listed. For example, the Cell Cycle pathway currently results in the following list.

![Figure 4-10 The “New Gene List” window](image)
• **Publish to GeNet:** Selecting this option will upload your information and the pathway picture to GeNet, please see 11.1.1 “Upload to GeNet” on page 11-1 for details.

• **Delete Pathway:** You will get a confirmation of deletion dialog box before the Pathway is completely deleted.

By right clicking over the Pathway folder in the Navigator you will be given the following options:

• **Publish to GeNet:** Selecting this option will upload your information and the pathway picture to GeNet, please see 11.1.1 “Upload to GeNet” on page 11-1 for details.

• **Make Gene Lists:** A new folder will appear in the “Gene Lists” folder, with all the Pathways in it as lists. For example, the Cell Cycle pathway currently results in the following Graph view.

![Figure 4-11 Viewing the Pathway mitosis as a graph](image)

Copyright 2000-2001 Silicon Genetics
• **Delete Folder:** You will get a confirmation of deletion dialog box before the Folder is completely deleted.
Chapter 5  Principal Components Analysis

Principal components analysis (PCA) explores the variability in gene expression patterns and finds a small number of themes, or building blocks. These themes can be combined to make all the different gene expression patterns in a data set. The first principal component is obtained by finding the linear combination of expression patterns explaining the greatest amount of variability in the data. The second principal component is obtained by finding another linear combination of expression patterns that is at “right angles” to (i.e. orthogonal and uncorrelated with) the first principal component. The second principal component must explain the greatest amount of the remaining variability in the data after accounting for the first principal component. Each succeeding principal component is similarly obtained. There will never be more principal components than there are variables (experimental points) in the data. Any individual gene expression pattern can be recreated as a linear combination of the principal component expression patterns.

In mathematical terms, the principal components of a data set are the eigenvectors obtained from an eigenvector-eigenvalue decomposition of the covariance matrix of the data. The eigenvalue corresponding to an eigenvector represents the amount of variability explained by that eigenvector, so the eigenvector of the largest eigenvalue is the first principal component, and so on.

An eigenvalue-eigenvector decomposition is performed on the covariance matrix of the gene expression values around zero. The eigenvector corresponding to the largest eigenvalue is called the first principal component. The first principal component is the dimension explaining the largest amount of variability. Successive principal components are eigenvectors corresponding to each smaller eigenvalue, and indicate successively orthogonal and uncorrelated dimensions explaining the largest proportions of the remaining variability. Principal components which explain significant variability are displayed by GeneSpring. The component loadings which represent the composition of each original gene as a linear combination of principal components are stored in gene lists for user examination.

5.1 Starting the PCA Analysis

In order to perform a principal components analysis, select Tools > “Principal Components Analysis...”. How long the analysis takes depends on the speed of your computer and the hugeness of your gene list. A word of warning: if you try to use PCA on list generated by the “Find Similar” command, you may get an error message. There may not be enough significant patterns to generate a sensible display.

To get a really good PCA, you will want to re-interpret the data using the log scale because PCA looks at the data in an additive way. (Please refer to GeneSpring User Manual 5.5 “Change Experiment Interpretation” on page 5-30 for details on how to change the current interpretation of your data.) The distance between 1 and 2 on a log scale is the same as the distance between 0.5 to 1. PCA was built for data in a Gaussian distribution. The log takes the data and puts it on a more balanced scale. If you don’t use the log interpretation, the first principal component will spend all of its energy on the fact that data is centred on 1.

In the following example, PCA is performed on the fictitious Yeast Extraterrestrial Studies. Of course, you can get different results by changing the initial interpretation. Here, the parameters
are set with Kryptonite concentration as a continuous element, Andromeda strain infection as non-continuous element and all other parameters are set as replicates.

GeneSpring will display a new window (similar to Figure 5-2) showing the most significant principal components. Components not explaining a significant proportion of the variability (at least twice the average variability explained by any one variable) are not displayed. The components are keyed by color to the color bar on the right of the window. Figure 5-1 shows this analysis for the fictitious Yeast Extraterrestrial Study, using a “log of ratio” interpretation\(^1\). For more information on how to change the experiment interpretation in GeneSpring, please refer to *GeneSpring User Manual 5.5 “Change Experiment Interpretation”* on page 5-30.

---

1. GeneSpring performs principal components analysis on the covariance matrix of the interpreted gene expressions (i.e. the normalized ratio, log ratio, or fold, according to the experimental interpretation) around zero. A consequence of this is that if you perform principal components with under the ratio interpretation, in most cases the first principal component will be fairly flat and basically represent the mean of the data. In most cases, it will be more informative if you change the experiment interpretation to a log ratio before performing PCA, this makes the data less skewed and centers the data around zero. All the examples shown in this chapter have been done with a logarithmic experiment interpretation.
5.1.0.1 The PCA Window

Figure 5-2 Principal Components Analysis display

As noted in the display, a folder of Gene Lists is automatically created and named “PCA” followed by the experiment name, with one Gene List for each significant principal component.

If your ModalGenes are difficult to see because they are too high or too low on the vertical axis, right click in the display and select Options > Change Vertical Axis Range.
5.1.1 Examining Components with the Gene Inspector

You can double-click any of the components in this display to bring up a Gene Inspector window. The eigenvalue associated with this principal component and the proportion of the total variance explained by this component are shown in the Description field. Figure 5-3 shows the result of double clicking on the first principal component (red) – the first principal component explains 42.31% of the overall variability. You can perform the usual operations in the Gene Inspector window, such as finding similar genes, etc.

![Gene Inspector display of the first principal component](image)

Figure 5-3 The Gene Inspector display of the first principal component
5.1.2 PCA and the Scatter Plot View

After performing principal components analysis, the main data display is set to a scatter plot of genes plotted on the first two principal components, as shown in Figure 5-4. By the way principal components are defined, this shows the 2-dimensional linear transformation of the gene expression pattern that displays the largest amount of the overall variability in the genes. You can animate this display and watch the expression colors to get some feel for which areas of the plot represent different expression levels. (It is also interesting to initiate a k-means clustering with this display showing, you will be able to watch the clusters form as the clustering proceeds.)

Meanwhile, back in GeneSpring’s main screen, a new folder and lists have been added to the Navigator. The new lists can be examined with the List Inspector, please refer to GeneSpring User Manual 6.2 “List Inspector” on page 6-10.

The default view of PCA clustered data is displayed in a scatterplot because the scatterplot shows loadings on PCA 1 versus PCA 2. PCA demonstrates that variation is made up of patterns. It tries to pick up the most important single pattern.
For more information about the Scatter Plot display in GeneSpring, please refer to *GeneSpring User Manual* 3.9 “Scatter Plot View” on page 3-33. Right clicking over the vertical or horizontal axis lines will give you several options for changing the display.

The linear coefficients for each gene that could be used to reconstruct the original gene expression pattern from the principal components are called the component loadings. These are automatically stored as a set of gene lists in a folder titled PCA plus the data set name (in the example it is PCA Yeast Extraterrestrial Studies, although you can’t see the whole name in the graphics). If you select one of these lists, and then choose View > “Ordered List”, you get a display like Figure 5-5. Genes with the largest loading on the selected component (and hence a large positive correlation) will be at the left end of the plot, those with a low loading (often negative loading and hence negative correlation) with the selected component will be at the right of the display. The expression patterns for each gene are shown vertically. The usual tools of selection are available from this display if you wish to create gene lists, for instance of genes with very high or very low component loadings.

For more information on working with the Ordered List View, please refer to *GeneSpring User Manual* 3.10 “Ordered List View” on page 3-41.

![Figure 5-5 Ordered list of the component loadings for the third principal component](image)
5.2 PCA Commands

In the PCA window you have many option to change the current display.

You can also access the “Save”, “Modify”, and “About” by right clicking over the Colorbar on the right side of the screen.

Back in the main screen all of the usual commands work in the Scatter Plot and Graph views. Please refer to *GeneSpring User Manual* Chapter 3 “Genome Browser Views” on page 3-1.
5.3 References


Hotelling, H. *Analysis of a Complex of Statistical Variables into Principal Components*. Journal of Educational Psychology 24, 417-441, 498-520 (1933).


Chapter 6  Self-Organizing Maps

Self-Organizing Maps (SOMs) are a way of exploring and mapping the variations in expression patterns within an experiment. The method is similar to k-means clustering, but adds an additional feature where the resulting groups of genes can be displayed in a rectangular pattern, with adjacent groups being more similar than groups further away. Self-Organizing Maps were invented by Tuevo Kohonen (1991, 2000) and are used to analyze many kinds of data. Applications to gene expression analysis were described by Tamayo, et. al. (1999).

A set of “nodes” are placed in the space of gene expression patterns, initially arranged in a 2-dimensional rectangular grid. Then the data for each gene is successively examined, and the nodes closest to the gene pattern are adjusted a small distance towards it. This is iterated over all the genes a number of times, with each node migrating towards the genes it is closest to, so the nodes spread throughout the space of gene expression patterns and represent the important features. As the iteration proceeds, the neighborhood of nodes that all are moving together is made smaller and smaller, so eventually each node moves towards different genes. In this process the grid is stretched and twisted like chicken-wire so the nodes follow the contours and clusters of the data set. When the iteration process is finished, each node is assigned the set of genes closest to the node, and a rectangular array of expression graphs is produced, each showing the gene patterns closest to the node. So this array of graphs is a “map” of the different gene patterns in the data.

6.1 GeneSpring Analysis Methodologies for Self-Organizing Maps

A rectangular grid of nodes (of user-specifiable size) is constructed in the space of gene expressions. In a large number of iterations, data for each gene are successively examined, and nodes close to that gene (in Euclidean distance of expression profile) are moved closer to that gene profile. In this way, the grid of nodes is stretched and wrapped to best represent the variability in the data, while still maintaining similarity between adjacent grid nodes. As the iteration proceeds, nodes are moved by smaller and smaller amounts to produce convergence. After the iteration, genes are assigned to the nearest grid node, and a display grid of gene expression graphs is shown corresponding to the final state of the grid. Node assignments can be saved as classifications for further study.

6.2 Function

The Self-Organizing Map (SOM) is one of the ways of exploring variability in a data set. SOMs are similar to clustering in the algorithm divides genes so their patterns are more similar within a group than between groups.
6.3 Description

In order to generate a Self-Organizing Map, choose Tools > “Cluster...” > SOM.

Figure 6-1 The Tools drop-down menu

A new dialog box will appear, as in Figure 6-2.
You can choose the genes and experiments to use in the analysis by selecting them in the Navigator. You can also restrict the information used by right clicking over any experiment or sample and selecting any of the restriction options. Please refer to 2.3.2 “Restricting the Genes to be correlated” on page 2-11 for details on all these options. For instance, you may well wish to select only genes with high control strength (signal strength) for analysis.

You can also specify the size and shape of the grid of nodes to be used to make the map (Rows and Columns), plus the neighborhood radius and the number of iterations.

The Neighborhood Radius parameter controls how many points move towards the data point in a particular iteration, and therefore how similar the profiles will be for each node. The neighborhood radius is in terms of Euclidean distance in grid units relative to the abstract grid of the SOM points, so for instance point 1,2 is one unit away from 1,3. (This is different from how far apart the nodes may be in the gene expression space.) As the iteration proceeds, the neighborhood radius is decreased smoothly from it’s starting value, so points move more independently later in the process. If you make the starting neighborhood radius very small (less than 1) each point will always move independently, and adjacent clusters will not be related. If you specify a very large
neighborhood radius, initially all the nodes will move towards every data point, and the grid will act as if it is very “stiff”, with more similarity between node results, but less flexibility to explore the variations in the data.

The Iterations parameter controls how many gene profiles are examined in the process of producing the SOM. If there are 10,000 genes and 60,000 iterations are specified, then each gene will be examined 6 times during the process. The step and neighborhood sizes are automatically adjusted as the SOM proceeds so the process finishes at the specified number of iterations.

The first time you request a SOM for a particular experiment during a session, GeneSpring fills in some suggested values for the parameters, based on the number of genes and the number of experimental points in the experiment. These are just suggestions, the user is free to modify them. GS will remember whatever parameters you are using until the end of the session.

A SOM analysis produces a classification, named by the (row, column) index of the node, as shown in Figure 2. You can name and save the results as either a Classification or a folder of Gene Lists.

When you have set your parameters, click “Start”. In a few moments a new box will appear, as in Figure 6-3.
You can save these lists under the Classification folder or the Gene List folder. Either can be used as coloring later. If you save as a list you can make a new SOM with that list later.

Choose a folder and give your new SOMs an appropriate name.

The results of the SOM analysis are shown as an array of gene expression graphs, as shown in Figure 3. Each graph corresponds to an SOM node, and each gene is assigned to the node it is closest to. Nodes are identified in the upper right corner of each plot.

![Figure 6-4 The Split Screen SOM display](image)

All of the genes have been selected for easier viewing.

You can Zoom In on any section of the grid for more data.
There is now more data visible in the split screen, including the section (2,4) and the number of genes (41) in each region. One gene (NOS) has been selected.
6.4 SOM Commands

All of the commands for selection, zooming etc. will work in each little screen as they do in normal graph screens. Each selection or Zoom will only effect the small screen in which you zoomed or selected. Animate will effect each of the small screens equally.

- **Undo Split Screen**: The View > “Unsplit Window” command can be used to return to viewing just one graph of all genes, when you are done with your analysis. But you can recreate the SOM graph at any time by right-clicking on the saved SOM classification or folder of gene lists, and choosing “Split Window/Both”.

![Figure 6-6 The Unsplit Screen, colored by classification](image)

If you would like to look at all the genes at once, the most sensible way to do this as a classification is to color by the classification. After you selected the unsplit screen command, posi-
tion your cursor over the name of your classification in the Navigator. Right click and select “Set as coloring scheme”.

- **Show just Averages**: You may also wish to choose View > “Show just Averages” or adjust the features shown from the View > Visible, in order to best display the results, as shown in Figure 4. As is seen here, if space is available on the graphs, the number of genes assigned to each node will be displayed for each graph.

![Figure 6-7 SOM graphs showing just averages](Image)

**Figure 6-7 SOM graphs showing just averages**

- **Show/hide vertical label**: Right click in the Genome Browser and go to Options > Hide Vertical Labels. In Figure 6-7 the vertical labels has been hidden. This allows the number of genes represented in each screen to be shown. If you have a larger screen this data will be visible in the normal view.
6.4.1 Classification Inspector for SOM

The Classification Inspector provides details on classifications produced by SOMs. The Classification Inspector can be viewed by double-clicking on a classification, or by right clicking over any Classification in the Navigator and selecting “Properties” from the resultant pop-up menu. The Inspector window shows a table of classes, including the number of genes in each class, the average class radius from its class centroid (in Euclidean distance), and the overall percentage of variance explained by the classification. The latter can be used to compare the effectiveness of different numbers of groups used in classification.

Right click over any classification and select the “Properties” option from the pop-up menu.

Figure 6-8 shows the Classification Inspector results for a classification from a SOM clustering with 12 groups. The name of the classification is shown at the top with other details and the (automatically generated) notes about the source of the classification. You can change the name of your classification in the Name box or add notes of your own in the Notes text box. These changes will be saved when you click the OK button. If you do not want to save changes, close the window by clicking the small black x in the upper right corner.
Figure 6-8 Classification Inspector for a SOM clustering

6.4.2 SOM Tips and Notes

If you have selected many panels, you may want to hide the horizontal and vertical labels for easier viewing. Right click in the main Genome Browser and look under the Options submenu for the various hide and show commands. You can also increase the amount of space available to the Genome Browser by choosing the View > Visible > Hide All command.

To change the interpretation of the graphs use Experiments > “Change Experiment Interpretation...”.

Occasionally with small datasets, you may see nodes with no genes in your new SOM. There is nothing in the theory behind SOM to prevent geneless nodes. This should not happen very often, if it does, you might try a smaller grid size (rows and columns) in the Clustering window.

![Figure 6-9 A SOM with empty nodes](image)

If your graphs are looking “too full” or appear to have “eraser” marks in the middle of the data, consider making an Ordered List (see *GeneSpring User Manual* 3.10, “Ordered List View” for more details) with the Interesting Genes function (see *GeneSpring User Manual* 8.8, “Find Interesting Genes” for more details) of GeneSpring. Once you have an ordered list by “interesting-
ness” graphically select the top end of the graph and re-run the SOM. This will (visually) get rid of the genes with little trust.

6.5 References


Chapter 7  The Predictor

The objective of the class predictor is to find a list of genes and a classification rule based on the gene expression level of those genes that will classify of samples into two or more classes of interest (such as cancer type, etc.). In order to do this, the algorithm starts with a training set of data where the classification of the samples is known. First, genes are rated and selected based on their ability to individually distinguish between the classes, then their predictive power is combined in order to predict the classifications of the samples in the independent test set using the k-nearest-neighbors method. A user-specified threshold can be set in the method such that if there is not a sufficient preponderance of the evidence in favor of a particular class, then no prediction will be made.

7.1 Using the Class Predictor

For illustration of this analysis, we will use the cancer classification data set presented in Golub, et. al (1999). The data consists of samples from patients with one of two forms of leukemia, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), and is divided into two parts, a training set of 38 samples (27 ALL, 11 AML) on which the predictor will be based and an independent set of 34 samples (20 ALL, 14 AML) on which the effectiveness of the predictor will be tested. These data are set up in GeneSpring as two experiments, called “AML ALL training set” and “ALL AML independent set”. In order for the class predictor to work well, you should have data sets with a number of samples at least comparable to those in this example.

To use the class predictor, choose “Predict Parameter Values…” from the Tools menu. Open the Experiments folder in the navigator, select the your training data set (“AML ALL training set” here), and then click the first “Set” button to assign it as the training set for the analysis, as shown in Figure 1. Then select your test data set (“AML ALL independent set”) and click the second “Set” button to assign it as the test data set.

![Figure 7-10 Selecting the Training Set](image)
Next you select the parameter to be predicted, as shown in Figure 2. Here we are interested in predicting Leukemia Type, which takes on the values ALL or AML for each of the samples in both the training and test sets. You are not restricted to parameters taking on only two values, the class predictor will work with any small number of groups. However the more classes to be distinguished, the more samples are needed in the training set to work well.

![Figure 7-11 Selecting the parameter to predict, and other options](image)

Here we have chosen to use 50 genes (out of the 6800 genes on the chip) as the basis of prediction, to use neighborhoods consisting of the 10 nearest training samples as the basis for classifying new samples in the test set, and to only make a prediction if the preponderance of evidence in favor of one class over the next best is 5-fold, i.e. the p-value for the best class is less than 20% of the p-value for the second best class.

Default values are suggested by the program for these values, but you can change them as appropriate. Experience suggests that the number of neighbors should not be too large a fraction (perhaps between one quarter and one third) of the number of training samples or the algorithm will not discriminate well.

If you know the classifications for the test data set, they should be specified in a parameter in the test set with the same name as in the training set, and those classifications will be used to evaluate the effectiveness of the predictor. This is a fair trial of the predictor, since the classifications in the test set are not used as part of the prediction process. If you do not know the real classifications for the test data set, you can proceed with the analysis by assigning your training set to be the test set as well, however the results of the prediction will not constitute a fair independent trial of the prediction. It will, however, allow you to examine which genes are strong predictors, etc.

After you have chosen all the parameters, click on the “Ok” button, and the analysis will be performed, and the results of the predictions will be shown in the Prediction Results window, as shown in Figure 3. Each sample in the test data set is shown, with the true class as given in the data set, and the predicted class from the prediction algorithm. In cases where the prediction algorithm decides not to make a prediction, the entry in the Predict column is left blank. Incorrect pre-
dictions (there aren’t any in this example) would be highlighted in red. In this example, out of the 34 samples, 31 were correctly predicted, and the algorithm chose not to make predictions in the other 3 cases. The value given under “P-value” is compared to the “P-value cutoff” specified on the first panel, a prediction is made if the value is less than the cutoff, and no prediction is made if the value is above the cutoff.

Several options are available at the bottom of the Prediction Results window. “Save Minimal Experiment” will allow you to save a subset of the training set containing just the information on the selected predictor genes. This can save time in loading the experiment if you want to be able to apply the prediction to new test data sets in the future. “Save Predictor Genes” will let you save a Gene List with the genes chosen as predictors. “Show Details” will show the individual class counts (“votes”) among the k neighbors, and the p-value for each class, as described under “Decision Threshold” below.
### Figure 7-12 Class Predictor Results

<table>
<thead>
<tr>
<th>Sample</th>
<th>True Value</th>
<th>Prediction</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 2</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.016</td>
</tr>
<tr>
<td>Sample 3</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 4</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 5</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 6</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 7</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 8</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 9</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 10</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 11</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 12</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0</td>
</tr>
<tr>
<td>Sample 13</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0.019</td>
</tr>
<tr>
<td>Sample 14</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0</td>
</tr>
<tr>
<td>Sample 15</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0</td>
</tr>
<tr>
<td>Sample 16</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0</td>
</tr>
<tr>
<td>Sample 17</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.016</td>
</tr>
<tr>
<td>Sample 18</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 19</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0.002</td>
</tr>
<tr>
<td>Sample 20</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0</td>
</tr>
<tr>
<td>Sample 21</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 22</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0</td>
</tr>
<tr>
<td>Sample 23</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0.884</td>
</tr>
<tr>
<td>Sample 24</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0</td>
</tr>
<tr>
<td>Sample 25</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0</td>
</tr>
<tr>
<td>Sample 26</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0.884</td>
</tr>
<tr>
<td>Sample 27</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0</td>
</tr>
<tr>
<td>Sample 28</td>
<td>AML ALL indepen</td>
<td>AML</td>
<td>0.441</td>
</tr>
<tr>
<td>Sample 29</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.129</td>
</tr>
<tr>
<td>Sample 30</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 31</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 32</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 33</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
<tr>
<td>Sample 34</td>
<td>AML ALL indepen</td>
<td>ALL</td>
<td>0.018</td>
</tr>
</tbody>
</table>

31 correct predictions, 0 incorrect predictions, 3 not predicted
7.2 Technical Details

7.2.1 Gene Selection

In order to select genes for use in the predictor, all genes are examined individually and ranked on their power to discriminate each class from all others, using the information on that gene alone. For each gene, and each class, all possible cutoff points on gene expression level for that gene are considered to predict class membership either above or below that cutoff. Genes are scored on the basis of the best prediction point for that class. The score function is the negative natural logarithm of the p-value for a hypergeometric test (Fisher’s exact test) of predicted versus actual class membership for this class versus all others.

A combined list containing the most discriminating genes for each class is produced as the predictor list. Each class is examined in turn, and the gene with the highest score for that class is added to the list, if it is not already on the list. Then genes with the next highest scores for each class are added. This is continued in rotation among the classes until the specified number of predictor genes is obtained. If you save the list of predictor genes as a Gene List, the best prediction score of the gene among the classes for which it would have been added to the list is saved as the attached number on the list.

7.2.2 Classifying the Test Samples

Based on the selected genes, classifications are then predicted for the independent test data, using the k-nearest-neighbors rule. A sample in the independent set is classified by finding the (user specified) k nearest neighbors of the sample among the training set samples, based on Euclidean distance between the normalized expression ratio profiles of the samples. The class memberships of the neighbors are examined, and the new sample is assigned to the class showing the largest relative proportion among the neighbors after adjusting for the proportion of each class in the training set.

7.2.3 Decision Threshold

P-values are computed for testing the likelihood of seeing at least the observed number of neighborhood members from each class based on the proportion in the whole training set. The class with the smallest p-value is given as the predicted class. The column labeled “P-value” is the ratio of the p-value for the best class to that of the second-best class. The predictor will make a prediction if this ratio is less than the “P-value Cutoff” specified on the initial panel, and will not make a prediction if the ratio is above this cutoff. Setting the p-value cutoff to 1 will force the algorithm to always make a prediction but may result in more actual prediction errors.
7.2.4 References


The global error model combines information from all genes in order to estimate the curve relating control strength (overall measurement level) with precision. Two components of precision are fitted, an absolute error component that dominates at low measurement levels, and a relative error component that dominates at high measurement levels. If replicate measurement data are available, the error model can be fitted using the replicate variability. If replicate information is not available, an approximate and conservative error model is fitted considering the variation of gene measurements around the normalization center. Robust fitting methods are used in order to separate the measurement variability from the biological variability. Error model results may be employed to evaluate the statistical precision of all measurements in an experiment, and as the basis for finding differentially expressed genes and making comparisons.

8.1 Global Error Model Details

Error models are ways of estimating the precision (standard error) corresponding to a measured expression level. The global error model in GeneSpring combines information from all genes in order to estimate the curve relating precision to control strength. The precision values calculated from the error model can be used as the basis of other statistical calculations throughout GeneSpring, including criteria for finding differentially expressed genes, tests of statistical significance, error bars, and other uses.

The error model makes precise the relationship between control strength and trust that has always been a part of the GeneSpring data presentation. Expression measurements made at high control strength are more precise and reliable than those made at low control strengths. The error model combines the effect of two types of error, base or background variability that provides a constant component of error in all measurements, and a proportional component of error that with measurement level.

The global error model is not used in GeneSpring unless you request it. You turn on the use of the error model as an option in the current interpretation, by checking “Use Global Error Model” in the Change Interpretation dialog (Experiments > Change Experiment Interpretation…), as shown in Figure 8-1.
Details about how the error model is fitted are specified in the Error Models dialog (Select Experiments > Error Models…), as shown in Figure 8-2.
At the top of the dialog, you choose whether to use replicate variability (if replicates are available) or the absolute deviation of the normalized gene intensity from 1 as the basis for the error model. Below that is a table showing the fitted error model coefficients for each sample, Base (a) and Proportional (b), as discussed the 8.2 “Technical Details” on page 8-5. The standard error of normalized expression levels will be approximately equal to the second coefficient (Proportional) for high control strengths, and is larger than that at low control strengths by an amount that relates to the first coefficient (Base). Please see 8.2 “Technical Details” on page 8-5 for more information.

Ideally, replicate measurements are available and their variability can be used as the basis for the error model. If replicate information is not available, an approximate and conservative error model is fitted considering the variation of gene measurements around their normalization center (1.0). This “Deviation from 1.0” option is chosen by default. Under this option, the error model is fitted to the expression data for all genes for each sample separately.

If you have replicate measurements you should choose the “Replicates” option, which then presents a detail panel where you can specify which of your parameters represent replicates, as shown in Figure 8-3. This is similar to assigning parameters as replicates in the Change Experiment Interpretation dialog, but the designation in the Interpretation dialog controls how the data is
grouped for display, and the designation in the Error Models dialog controls how the data is grouped for the analysis of the error model.

![Error Models](image)

**Figure 8-3 Error Model with Replicates**

This example combines four samples with the same disease state as replicates, and fits the error model to their standard deviation. Thus the fitted error model coefficients are the same for all of the samples in each replicate group. Using replicates will in general produce smaller error curves, and allow better discrimination of differential expression, etc.

If the global error model is turned on then the standard errors from the model will be used in drawing error bars in the Gene Inspector and as the basis for the Color by Significance display. It is also used as the basis for testing for whether the normalized gene expression level is different than 1 in the Gene Inspector (t-test p-value).
8.2 Technical Details

The form of the error model is a two-components error model known as the Rocke-Lorenzato model. The two components are an absolute error component that dominates at low control strengths, and a relative error component that dominates at high control strengths. The formula for the error model for raw (pre-normalization, except the background is subtracted) expression levels can be written as:

$$\sigma_{\text{raw}} = \sqrt{a^2 + b^2 C^2}$$

where $\sigma_{\text{raw}}$ is the measurement standard error of the raw expression data, $C$ is the control strength, and $a$ and $b$ are the fitted coefficients of the model.

Expressed in terms of the normalized expression levels, which are the result of dividing raw expression levels by control strength, the standard errors can be written as:

$$\sigma_{\text{norm}} = \sqrt{a^2 / C^2 + b^2}$$

The basis of fitting the error model is the set of squared control strengths and either the variance of the replicates or the squared deviation from 1.0. Before fitting the model, a robust smoothing is performed on each series, in order to remove the effect of biologically active genes on the deviation measure, and to produce a robust fit. GeneSpring uses a set of non-overlapping windows, and each window (consisting of the data from 11 genes) contributes a single point to the curve fit. Then an iteratively reweighted linear regression of variation or squared deviation versus squared control strength is fitted to estimate the coefficients.

If you have replicate measurements you should choose the “Replicates” option, which then presents a detail panel where you can specify which of your parameters represent replicates, as shown in Figure 3. This is similar to assigning parameters as replicates in the Change Experiment Interpretation dialog, but the designation in the Interpretation dialog controls how the data is grouped for display, and the designation in the Error Models dialog controls how the data is grouped for the analysis of the error model.

Figure 8-4 shows the results of the error model to the deviation data for one chip. The graph plots absolute deviation of the raw expression level from the control strength versus control strength. The blue points are individual genes. The purple points are the results of the running median filter on the data. The yellow curve shows the fitted error model.
Figure 8-4 GeneSpring Error Model

8.3 References

Index

A
Automatic Annotation 3-2
C
Change correlation 1-5
Classification Inspector 4-22, 7-9
   Average Radius 4-23
   class 4-23
   notes 4-22, 7-9
Cluster
   commands 4-16
   results 4-12
Clustering window
   similarity definitions 3-17
   correlation 1-2
Correlation Equations
   Change correlation 1-5
   Distance 1-4
   Pearson correlation 1-2
   Smooth correlation 1-5
   Spearman Confidence 1-3
   Spearman correlation 1-3
   Standard correlation 1-2
   Two-sided Spearman Confidence 1-4
   Upregulated correlation 1-6
   correlation, weighted 3-9, 4-3
D
decimal markers 3-15, 4-9
dendrogram 3-1
Distance 1-4
E
Euclidian metric 1-4
F
formula notation 1-1
G
gene similarity 1-1
Gene Trees 3-1
H
Hierarchical Clustering 3-1
   commands 3-22
   Complex 3-8
   Experiments 3-1
   simple 3-2
K
KEGG 5-8
K-means
   Maximum Iterations 4-2
   Number of Clusters 4-2
   k-means clustering 4-1
Kyoto Encyclopedia of Genes and Genomes 5-8
M
mathematical notation 1-1
Minimum Distance 3-21, 3-22, 3-23
missing expression values 1-1
mock-phylogenetetic tree 3-1, 3-2
N
new Pathway 5-4
nodes 7-1
P
Pathway view 5-1
   commands 5-9
PCA see Principal Component Analysis
   Pearson correlation 1-2
   Principal Component Analysis 6-1
R
restrict data types
   Control Signal 3-15, 4-9
   Normalized Data 3-15
   Number of Replicates 3-15, 4-9
   Range of Normalized Data 3-15, 4-10
   Raw Data 3-15, 4-9
   Standard Deviation of Normalized Data 3-16, 4-10
   Standard Error of Normalized Data 3-15, 4-10
   T-test probability 3-16, 4-10
S
Separation ratio 3-21, 3-22, 3-23
similarity definitions see also correlations
   Smooth correlation 1-5
SOM
   Euclidean distance 7-1
   Set as coloring scheme 7-8
   Show just Averages 7-8
   Spearman Confidence 1-3
   Spearman correlation 1-3
   Split Window 4-19
   Standard correlation 1-2

Copyright 2000-2001 Silicon Genetics
T
Trees 3-1
    commands 3-22
    Complex 3-8
    Do Automatic Annotation 3-4
    Experiments 3-1
    Minimum Distance 3-21
    Only Annotate With Standard Lists 3-4
    Separation ratio 3-21
    simple 3-2
Two-sided Spearman Confidence 1-4
U
Upregulated correlation 1-6
Document History for
Advanced Analysis Techniques Manual

Update from 2.0 to 2.1
Improved correlation chapter
Updates to PAC
Added reference sections

Update from 1.1 to 2.0
Updates to SOM

Update from 1.0 to 1.1
New Cluster window