Genesis 1.0

Gene Expression Similarity Investigation Suit

Operation Manual

Alexander Sturn

TUG
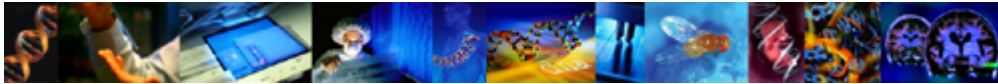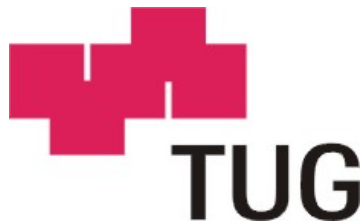
Graz, Austria, 06 January 2001

PROPRIETARY NOTICE

TRADEMARK

CREDITS

The Genesis software was written by Alexander Sturn.
This manual was written by Alexander Sturn.

Manual Version. 1.0

ALEXANDER STURN

Address:
Krenngasse 37
8010 Graz
Austria

Phone: +43 316 876 5345
Fax: +43 316 873 5340
E-mail: alexander.sturn@tugraz.at
Homepage: genome.tugraz.at

Chapter I

The Genesis Software

Purpose

High throughput gene expression analysis is becoming more and more important in many areas of biomedical research. cDNA microarray technology is one very promising approach for high throughput analysis and gives the opportunity to study gene expression patterns on a genomic scale. However, analyzing gene expression data requires sophisticated tools including clustering techniques, motif search, and visualization utilities. We have developed a platform independent Java suite, which integrates these tools. Fluorescence ratios are first imported and can be then normalized in several ways to gain a best possible representation of the data for further statistical analysis. Cluster analysis of fluorescence rations from multiple experiments can be used to identify co-expressed genes, retrieve meaningful patterns of gene expression and point out similarities and/or differences between analyzed conditions. The imported data can be clustered using all common distance similarity measurements and the following methods: hierarchical clustering, k-means, self organizing maps, principal component analysis, and support vector machines. Additionally, motif visualization tools have been implemented, including the drawing of sequence logos and motif diagrams. Moreover it is possible to map gene expression data onto chromosomal sequences. Finally, extensive work has been undertaken to accomplish visualization of the gene expression data and clustering results in a user friendly and intuitive way. The flexibility, the variety of analysis and data visualizations tools, as well as the transparency and portability, provides this software suite with the potential to become a valuable tool in functional genomic studies.

Hardware Requirements PC

Minimum hardware configuration for Genesis:
• Intel P-II (or compatible) CPU, 266 MHz
• 512 KB cache memory
• 128 MB of main memory
• 4 GB hard disk
• 4x CD-ROM
• 17" monitor
• PCI or AGP graphic adapter, 8 MB,
resolution 1024 x 768, 16 bits

Recommended hardware configuration for Genesis:
• Intel P-III (or compatible) CPU, 500 MHz
• 512 KB cache memory
• 384 MB of main memory
• 20 GB hard disk
• 30x CD-ROM
• 19" monitor
• PCI or AGP graphic adapter, 32 MB,
resolution 1280 x 1024, 24 bits

Hardware Requirements Mac

Minimum hardware configuration for Genesis:
• 128 MB of main memory
• 4 GB hard disk
• 4x CD-ROM
• 17" monitor
• resolution 1024 x 768, 16 bits

Recommended hardware configuration for Genesis:
• 384 MB of main memory
• 20 GB hard disk
• 30x CD-ROM
• 19" monitor
• resolution 1280 x 1024, 24 bits

Hardware Requirements Workstation

Minimum hardware configuration for Genesis:
• 128 MB of main memory
• 4 GB hard disk
• 4x CD-ROM
• 17" monitor
• resolution 1024 x 768, 16 bits

Recommended hardware configuration for Genesis:
• 384 MB of main memory
• 20 GB hard disk
• 30x CD-ROM
• 19" monitor
• resolution 1280 x 1024, 24 bits

Software Requirements

• Java 1.3 (with reservations also 1.2.2)
• Java3D 1.2

And one of the following operation systems

• WindowsNT 4.0
• Windows 2000
• Windows 98
• Windows 95
• Windows ME
• Linux
• Unix
• MacOS
• Any other Java supporting platform

Chapter 2

Installation

Open a web browser and go to the URL: genome.tugraz.at. Then got to **Genesis -> Installation** in the subfolder Software.

Web based installation

A certification dialog appears (Figure 2.1). Grant the Browser to access your computer by clicking the "Yes" or "Grant" Button.

A java applet will be started…

Select: I want to search my system for existing Virtual Machine.

The Applet searches your system for a valid java version.

If no valid version is found, please select "I want to specify an existing Virtual Machine" and specify the path to the java program on your system (e.g. java.exe on Windows systems).

If finally a valid java version has been found, the Applet starts to download the software and starts the installation program.

No Java plugin:

If your Browser is not able to execute Java applets, you have two options:

Windows system:

Download the setup.exe program to a temporary location on your system and execute the program.

Other systems:

Download the setup.class file to a temporary location on your system. Open a shell, go to the
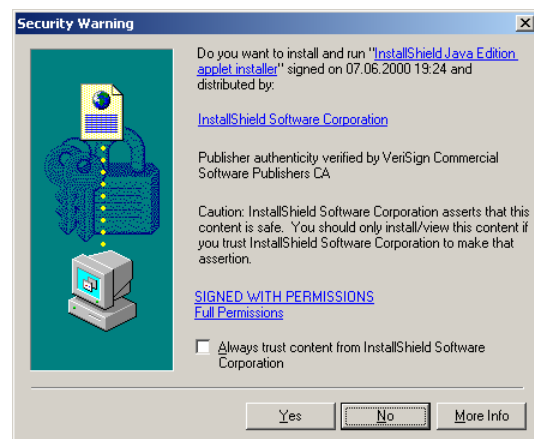
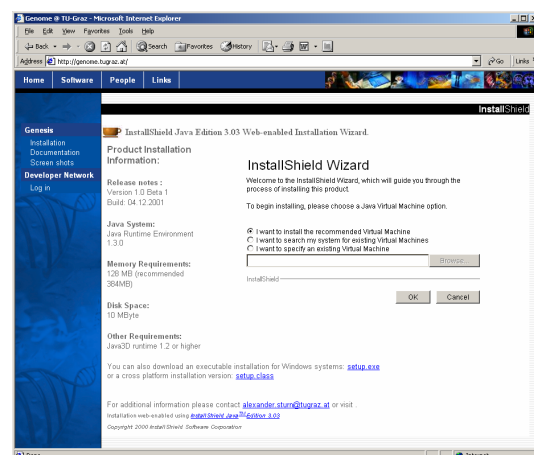*Figure 2.1: Windows example of the certification dialog.*

Figure 2.2: Web based installation with the Java Applet

folder where the setup.class file is located and execute: `java setup`

Maybe you have to specify the full path to the java location like:

```
/usr/java/bin/java setup
```

Installation in Progress:

The InstallShield Wizard will guide you through the installation procedure. Follow the instructions on the screen and enter or select the appropriate information when prompted.

After a welcome screen (figure 2,3) you will be asked to type in the password for the installation (figure 2.4).

A password is no longer necessary for standard versions of Genesis. However passwords are sill used for special releases of Genesis. If you don't have a password, contact the developer.

InstallShield will check if you have already installed the Java-VM – if not or an older version is already present, you will be asked if you would like to install the latest Java-VM on your hard disk (figure 2.5).

Warning: The bundled Java version is only for Windows systems. If you are installing Genesis on an other platform do under no circumstances install the bundled Java-VM. Install the correct Java-VM for your platform. It has to be release 1.2.2 or higher for proper function of Genesis.

InstallShield will proceed installing Genesis into the selected destination directory (figure 2.6).
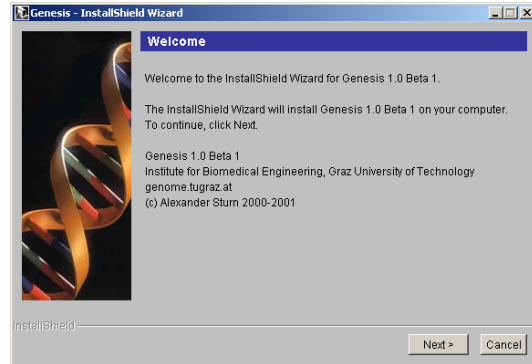

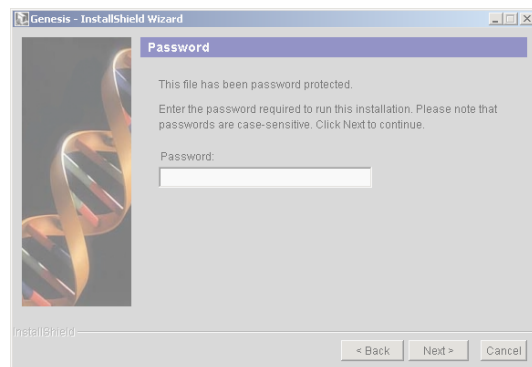*Figure 2.3: The Installation Welcome Screen*
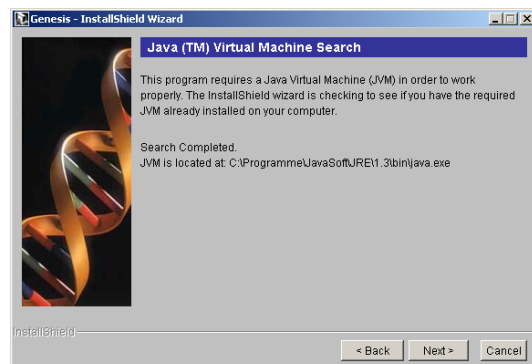

Figure 2.4: The Password checking dialog.

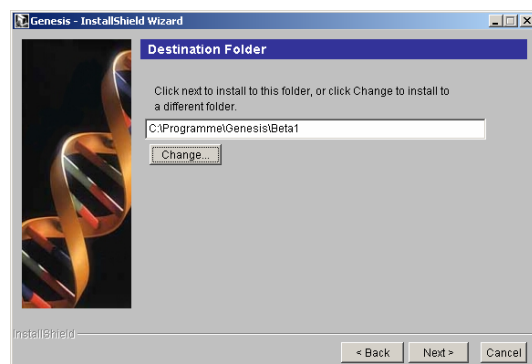*Figure 2.5: Checking if Java-VM is installed already*

Figure 2.6: Selecting the destination folder for the Genesis software

During the installation:

Next you can specify if you would like to install an icon on your desktop (only windows systems, figure 2.7)
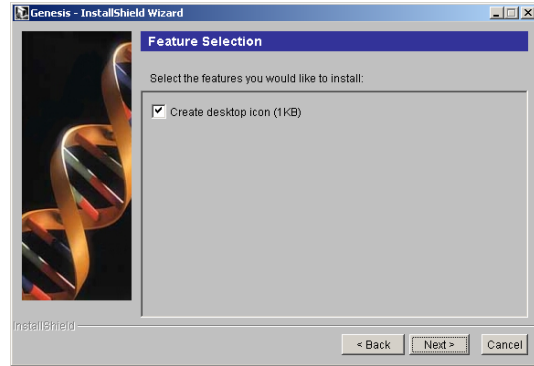

Figure 2.7: Desktop icon installation feature

Before the installation is started, the Wizard will show a summary of the selected install options figure 2.8).
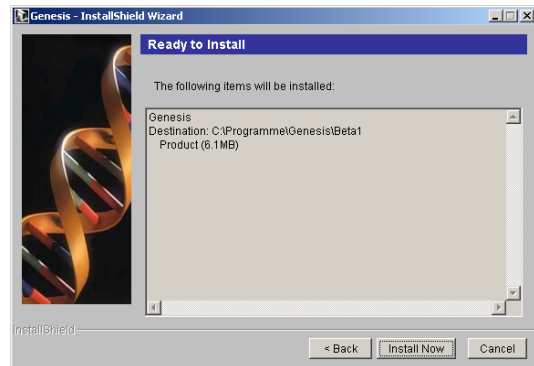
Figure 2.8: The summary of the selected options

The Wizard displays a progress bar during the installation process (figure 2.9).
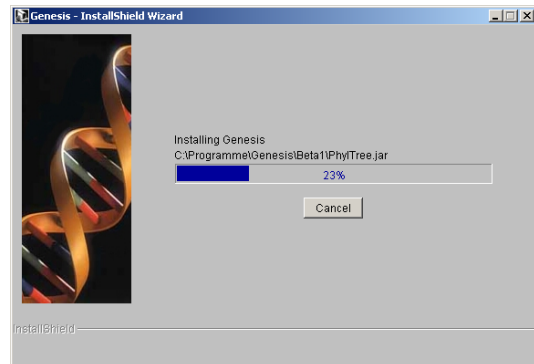
Figure 2.9: The installation in progress

Genesis requires the Java3D runtime environment release 1.2 or higher for the 2D view of PCA results. You can retrieve Java3D form the Java homepage with the URL:

java.sun.com/products/java-media/3D/

Do not use Beta releases of Java3D since they can interfere with the correct function of Genesis!

Genesis is able to run without Java3D, although the 3D view of PCA will not work in this case.

Figure 2.10: The Java3D information dialog

Warning: This is a Beta version of Genesis. Therefore it is possible that some features are not finished and errors can occur. If you discover an error, please report it to the developer, so that it can be fixed. We appreciate your help.

After the installation is finished, the Wizard will show a summary of the installed items and their current path (figure 2.12).
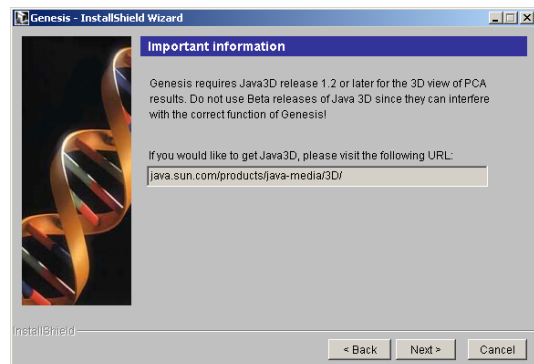
Program start setup

Windows systems:

The installation program creates an icon on the desktop and Program-menu. If the creation fails execute the Genesis.bat file in the Genesis directory.

Unix systems:

Open a Unix shell

Change to the Genesis directory

Change the execution attribute of the GenesisUnix file by typing for instance:

```
chmod u+x GenesisUnix
```

Start the program:

```
./GenesisUnix
```

If the starting fails: Open GenesisUnix file in a text editor and change the path in front of the java command to the Java-VM on your computer, like

```
/usr/java/bin/java –xm512m –classpath...
```

If you have problems with memory, change the value in the –xm512m statement to a lower or higher value (memory in Mbytes).

Compaq Alpha systems:

If you have a FastVM installed use the AlphaFast file in the same way as described above. Probably it is necessary to change the stack limit by typing:

```
limit stack 32768
```
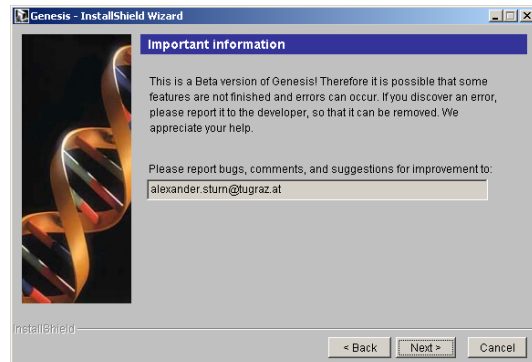

Figure 2.11: The Beta release information dialog
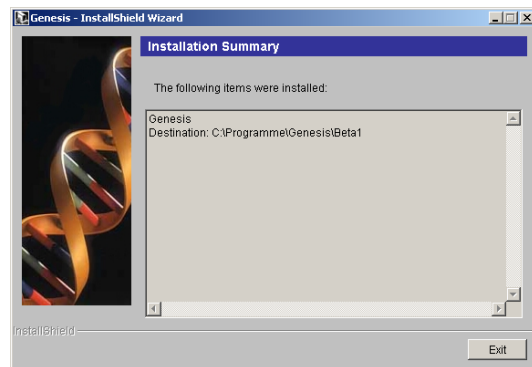
Figure 2.12: The summary after finished installation

Important notice:

Although this software is able to run on various platforms it has been developed and tested for and under Windows 2000/NT. Due to the different amount of support of, and varieties between other virtual machines, complicacies executing Genesis may occur using non commonly used VMs. Therefore we recommend using a sun supported platform (Windows NT/2000, Linux or Solaris).

Reinstallation

For re-installing Genesis, you have to uninstall it first.

Uninstalling Genesis

Windows systems:

Open the Start/Settings/Control Panel folder.

For Windows NT:

Open Add/Remove Programs.

For Windows 95/98/2000:

Open Software.

Continue with:

Select Genesis from the list and press the Add/Remove button.

Follow the uninstall instructions (figure 2.12).

Unix systems:

Simply remove the complete Genesis directory and all its files.
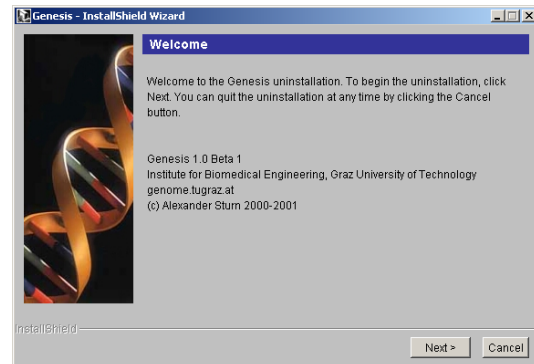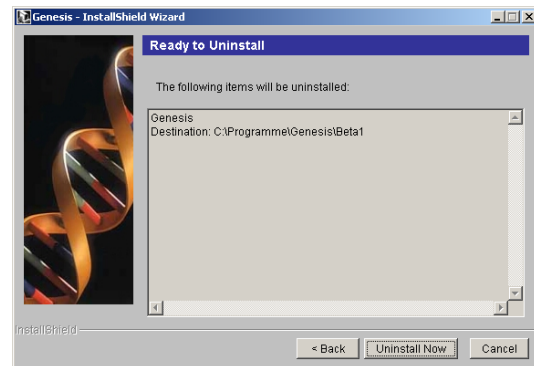

Figure 2.13: The Uninstallation Welcome Screen

Figure 2.14: Starting the uninstallation
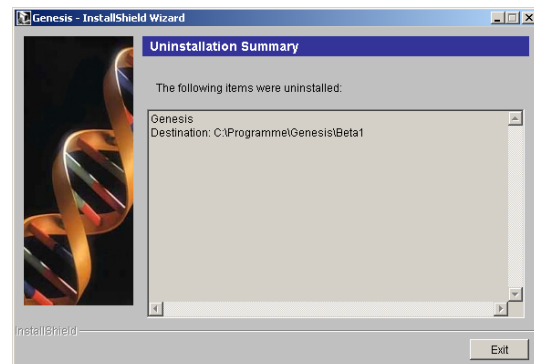
Figure 2.15: The summary after finished uninstallati

Chapter 3

First Steps

Starting the Main Program

Windows systems: To start Genesis double-click on the shortcut icon (on your desktop) or press the Start button on the taskbar. Then open the Programs menu and select Genesis -> Genesis 1.0 Beta 1. You can also open a Dos shell and execute the Genesis.bat file in the Genesis home directory.

Unix systems: Start the program with the command: `./GenesisUnix` in the Genesis home directory.

See also: **Program start setup** in the previous chapter.

After a short time of waiting the Genesis screen appears, showing your system properties on the right panel. Be sure that the Java VM and the Java3D is running properly (figure 3.1). The memory information specifies the memory within the Java VM, not the total memory of your system.

Figure 3.1: Genesis' user-interface after the startup.

📂 Loading a dataset

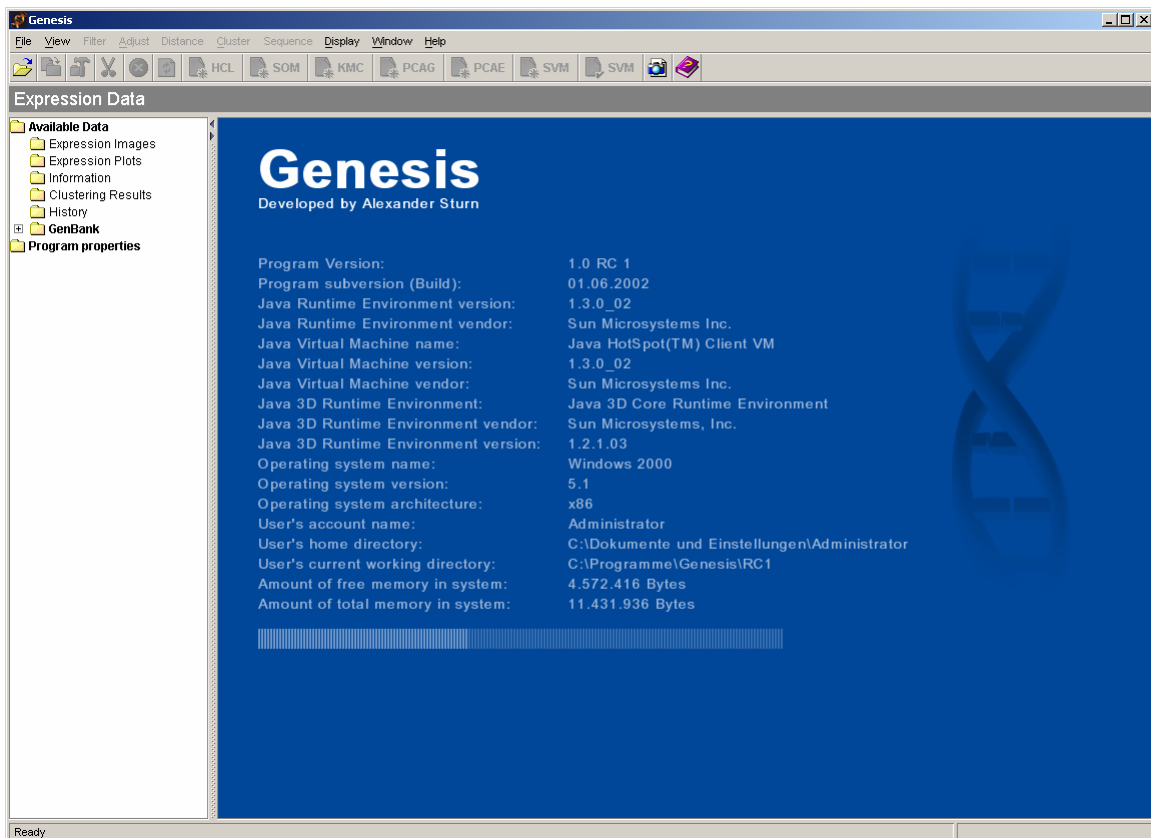To load a sample dataset, either select Files-> Open from the menu, or click on the "Open" button. Genesis supports tab delimited text files in the Stanford Format. Genesis can handle just one file at a time!

It is also possible to open one of the 10 most resent files by clicking on Files->Filename.

To demonstrate the features of Genesis the fibroblast dataset (Fibroblasts.txt in the Sample directory) from Iyer et. al. is used.

After loading the dataset, an expression image is calculated and displayed. If the dataset is very big, the calculation of the expression image can take quite long. Remember: If you have 5000 genes and 20 experiments, 100.000 rectangles have to be drawn. It is therefore possible to divide the image and to specify the number of genes per page : View-> Set genes per page… The default value is 1000 genes per page. If you have problems opening a file because of its size, change this value before opening the dataset!

You can flip between the images by clicking on the specific Leaf in the Expression image folder.

To display all genes in one page simply enter a number greater than the number of genes in the dataset.

You can also view the dataset as expression plot or centroid plot by opening the Expression Plots folder in the tree on the left side of the program window (later referred as data tree) (figure 3.4).

If you open the Information folder you can get information about the current dataset (number of genes, number of experiments and so on).

If you move with the mouse over one of the colored squares the square is selected and the Gene ID, the experiment name, and the value of the cell are displayed in the status bar.

If you click on a cell, a dialog pops up with some information about the selected cell. You can uses this information to search in the Entrez database of the NCBI. The data are retrieved form the UniqueID and Name field in the opened flat file (figure 3.5).
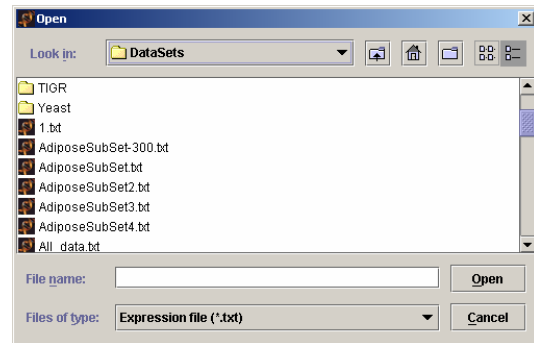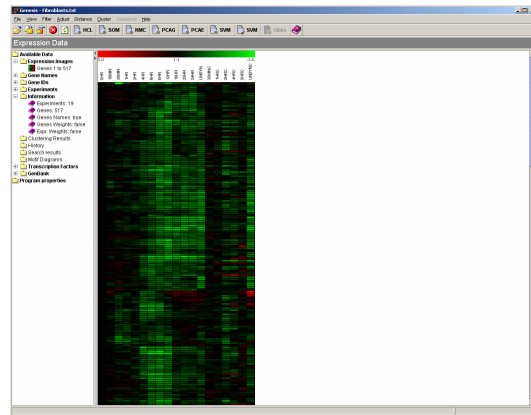

Figure 3.2: The file open dialog box

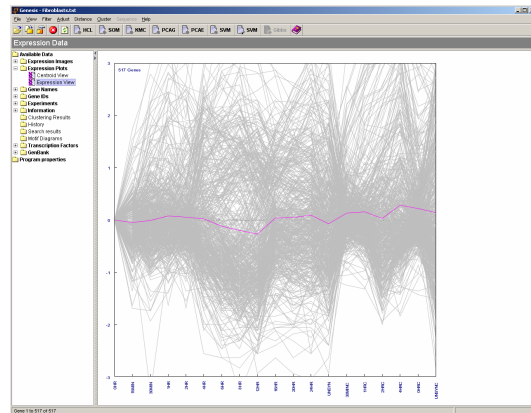*Figure 3.3: Display of the expression data*

Figure 3.4: Data Tree showing information about the Fibroblasts.txt dataset.
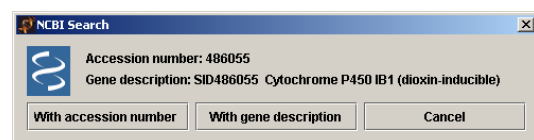

Figure 3.5: NCBI Search dialog.

Changing the image properties

Genesis is able to render the expression image in a variety of ways. You can uses predefined schemas or specify the cell height and with of a cell yourself. Unit is pixel.

Important: If the height of a cell is grater or equal 10 pixels the describing text (i.e. accession number, figure 3.6 right) is displayed besides the expression image. If the height is smaller than 10 pixels this would be unreadable and therefore the text is not displayed.

If the text is visible, you can switch on or off whether the gene description (can be very long) should be displayed or not with the menu entry Show gene names.

You can switch the drawing of the borders between the cell on and of by clicking on the Show borders menu item.

Additionally you can specify the max. value for saturated colors. The default value is 3.0, that means that a ratio grater or equal to 3.0 is rendered with saturated red and a ratio less of equal to –3.0 is rendered with saturated blue for instance (figure 3.6).

With Adjust to maximum the ratio value for the saturated pos. and neg. color is set to the maximum absolute ratio in the dataset.

If you have integer data (e.g. number of ESTs in a specific library) you can use the Adjust to integer command. This increases the color "distance" between two values (e.g. 1 and 2) so that it is easier to distinguish between different values. The color range (e.g. a rainbow) is in this case divided into 25 different colors so that 0 is for example color red and 25 is color magenta (see also figure 3.9).

With Reset maximum to 3.0 you can reset the ratio for saturated colors to the default value of 3.0.

You can specify this value of course to any value you want. Lower values increase contrast, but values greater than the specified maximum are rendered with the saturated color.

Important: This settings also influence the axis scale of the centroid view and expression view.

With the Adjust to maximum command you can scale the axis so that all values are within the drawing area.
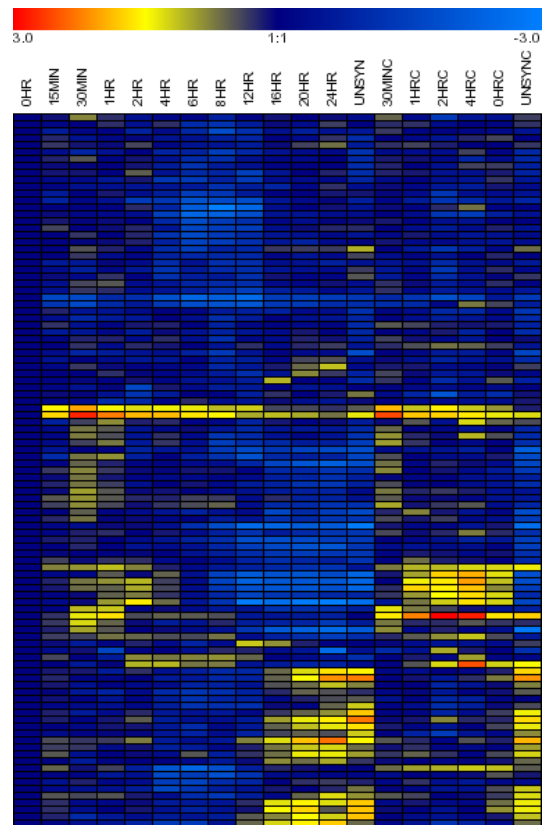

Figure 3.6: Standard cell rendering (20x5, borders) of an expression image.
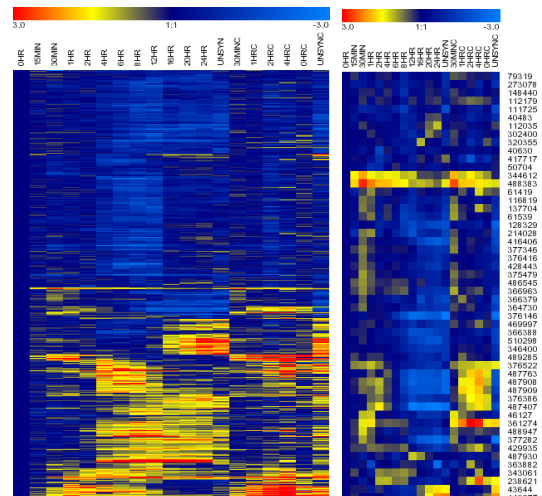

Figure 3.7: Compact (20x1, no borders) and Detail view (10x10, no borders) of an expression image.

Changing the color scheme

To change the color scheme of the expression image, either select View -> Adjust colors… or click on the Color properties button. The Color Configuration dialog will be displayed, where you can either select specific colors by clicking on the Choose button or use one of the predefined color schemes on the right side (figure 3.8).

standard color scheme used in the literature

color scheme for color blind persons

this one gives a better contrast for weak ratios

for better distinction between pos. and neg.

standard color scheme used in the literature

false color scheme for the pos ratios
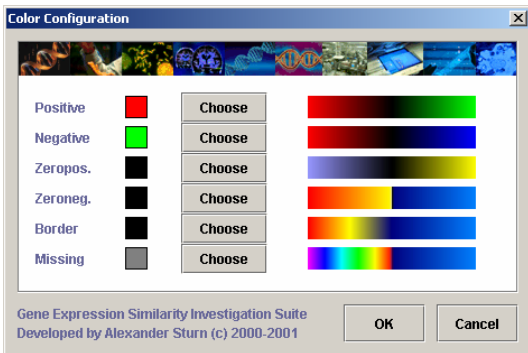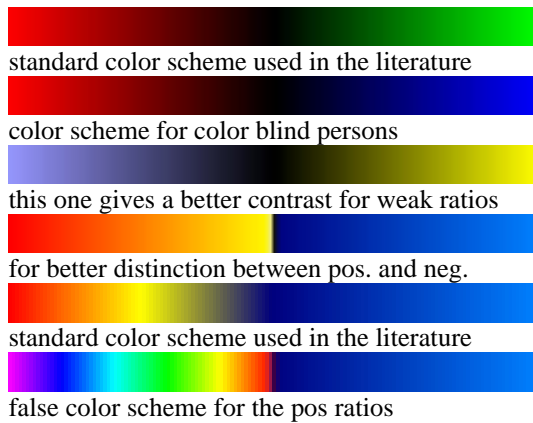
Figure 3.8: Color Configuration dialog.

You can also mirror the color scheme with the command Flip colors.

Changing the Look&Feel

It is possible to change the Look&Feel of Genesis at any time to one of the supported Look and Feels (see figures 3.10-3.13). Display -> …

The default look and feel is the Genesis Look&Feel.

Figure 3.10: Genesis Look&Feel.

Figure 3.11: Genesis Blue Look&Feel.

Figure 3.12: Java Metal Look&Feel.

Figure 3.13: CDE/Motif Look&Feel.

Anti aliasing effects

One word to anti aliasing: Anti aliasing smoothens the edges of lines, curves and text and therefore improves the look of a drawing tremendously. However, unfortunately it require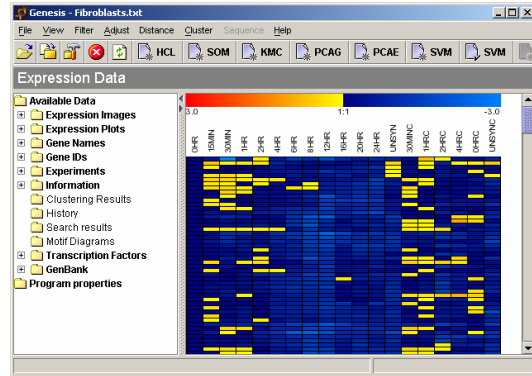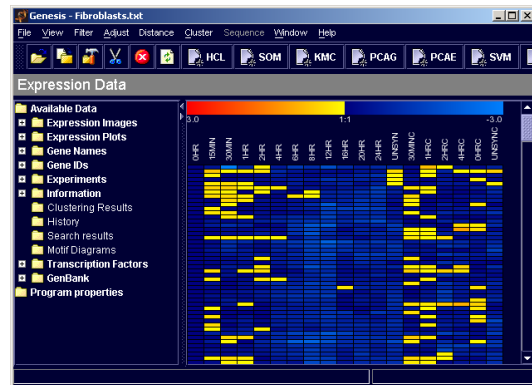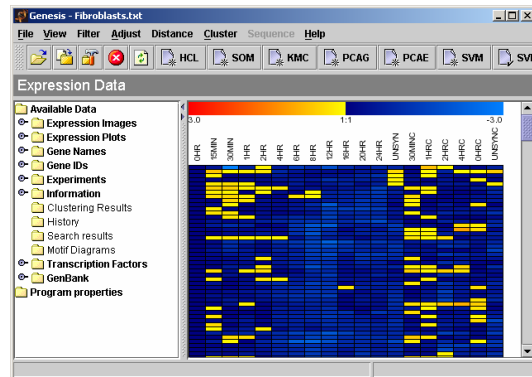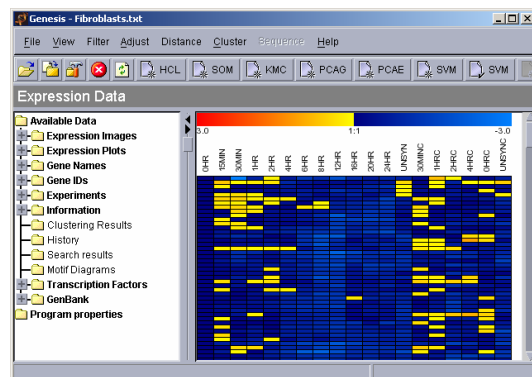s a lot of computational power. If you have for instance a dataset with 5000 genes and 21 experiments, Genesis has to draw 100.000 squares in the expression view and that can take a while. The abandonment of the advantages of anti aliasing improves rendering time significantly. You can switch anti aliasing on and of with the Use anti aliasing command.

Filtering the data

It is very often necessary to filter data in order to get a better basis for later clustering and data adjustment procedures. Each filter dialog has a Try button, where you can calculate how many genes pass the criteria. This operation does not change the data. The Apply button opens a Save File dialog, where you can specify the name and location of the new file, which contains only the genes that have passed the specified criteria. Currently only the filtering of genes is supported Genesis supports at the moment two filter:

Figure 3.14: Missing values filter dialog.

Missing values filter:

This is a filter to construct a new dataset out of the currently loaded by excluding all genes with less or equal *n* values present (=not missing values), where *n* has to be specified by the user (figure 3.14). Missing values are difficult to handle during the usage distance measuring procedures. A few are usually no problem, but if there are too many in comparison to the number of vector-elements, an arbitrary result can be expected. Therefore it is important to filter out genes with to many missing values.

Standard deviation filter:

Figure 3.15: standard deviation filter dialog.

This is a filter to exclude all genes that have a standard deviation of less than the stated one (figure 3.15). This filter can be used to get a set of genes that change significantly between conditions. It can be used to retrieve for instance the 500 genes with the "biggest" changes out of a whole genome dataset.
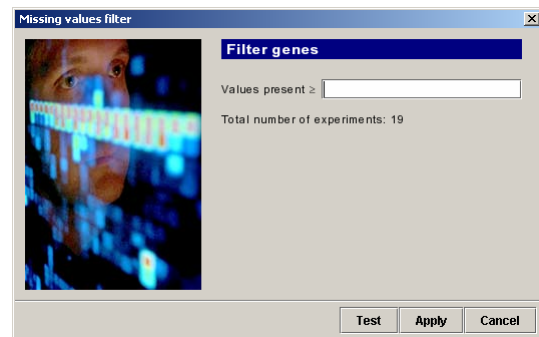
Chapter 4

Hierarchical Clustering

HCL Starting the calculation

To start the Hierarchical Clustering calculation simply select the menu item **Clustering -> Calculate Hierarchical Clustering** or click onto the HCL button on the toolbar. Dialog 4.1 will appear to select the agglomeration rule (average, complete or single linkage) and whether genes or experiments (samples) or both should be clustered. Preset is average linkage clustering of only the genes, since this is probably the most commonly used procedure.

To start the calculation process simply click OK. The calculation will now be started in its own thread and a progress bar will be displayed. As soon as the calculation has been completed, a new entry is added to the **Clustering Results** folder.

Handling the result

If you click onto the result leaf the reordered expression matrix and the tree representing the relations between genes are displayed on the right panel (Figure 4.3). In the tree horizontal branch lengths represent the similarity between genes or experiments.

This result representation is very similar to the representation discussed earlier in this manual. You can change the color schema, cell sizes and so on in the same way we have discussed before.

To change the tree properties click with the right mouse button anywhere on the right panel containing the expression matrix and the tree. A popup panel will appear with many menu items manipulate the result.

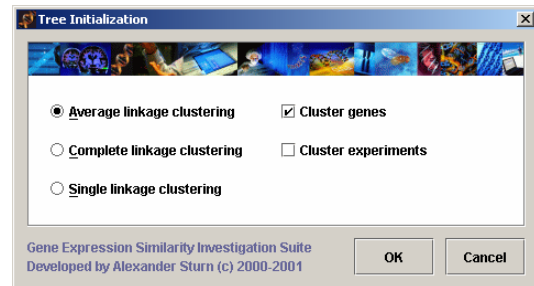To change branches of the gene tree for instance select **GeneTree properties…**

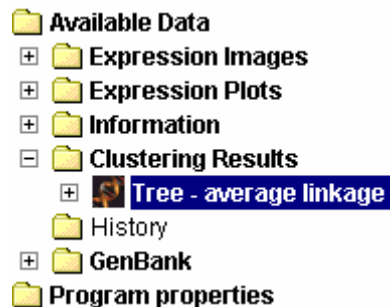

Figure 4.1: Hierarchical Clustering Dialog

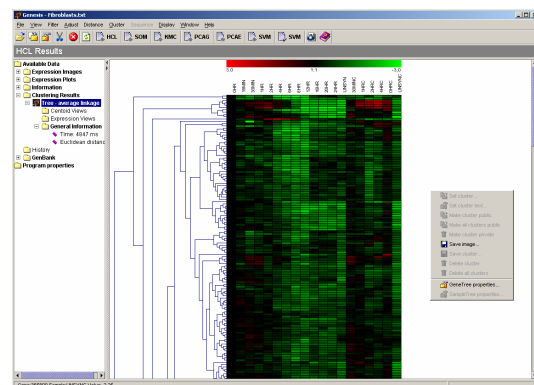*Figure 4.2: New clustering result*

Figure 4.3: Hierarchical Tree Result

A dialog with tree parameters will appear (Figure 4.4). Minimum pixel distance represents the minimum horizontal branch length in pixel. The maximum pixel distance represents the maximum horizontal branch length in pixel. If the distance would be greater than this value the branch will be branch length will be truncated to this length.

Distance threshold represents the distance value between two genes that will be considered as zero. Any distance smaller than this threshold will be considered to be zero and the branch length between this genes will be the minimum distance threshold. Generally we can say that increasing this value will lead to shorter branch length. If you would like to have equidistant branch length just enter the same pixel value (for instance 10) in the minimum and maximum distance fields (Figure 4.5).

Selecting clusters

To select a subtree simply click with the mouse on the note which should be the top of the subtree. The subtree will be selected and the genes in this branch will be highlighted (Figure 4.5).

The current version supports just selecting subtrees of the gene tree, not the experiment tree.

Specifying gene clusters

After selecting a subtree, click once with the right mouse button. A popup dialog will appear. Select Set gene cluster… .

A new dialog will appear and you have to select a color for the cluster (Figure 4.6).

After pressing OK, the selected subtree will appear in the specified color with a colored bar right beside the selected genes (Figure 4.7).

To specify a text for the cluster click again using the right mouse button and select Set cluster text …

Enter the cluster name and press OK.

Important notice: For all cluster manipulations the specific subtree has to be selected!

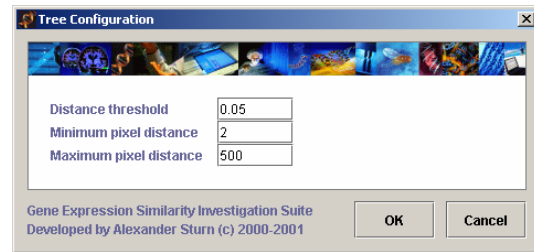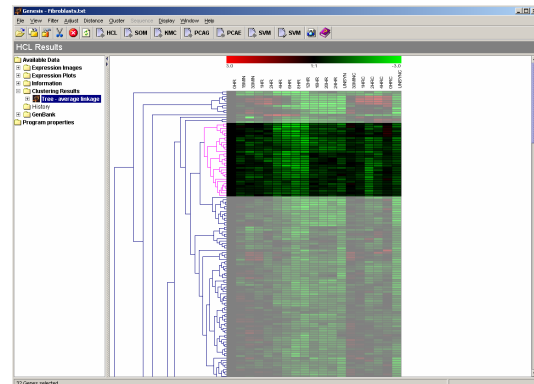*Figure 4.4: Gene tree properties dialog.*
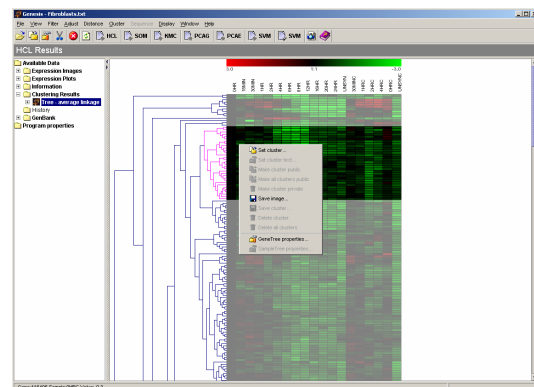

Figure 4.5: Selected subtree
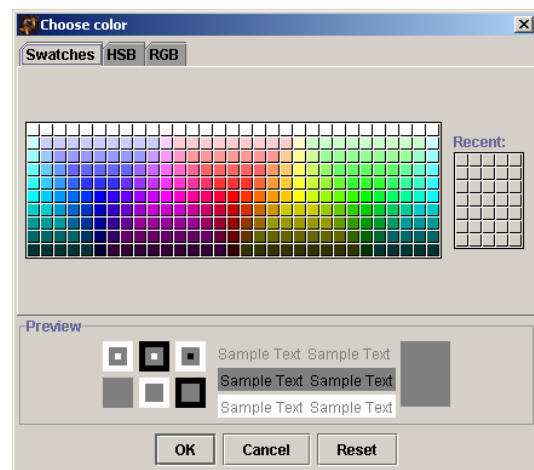

Figure 4.6: Creating a cluster

Figure 4.7: Dialog to choose a cluster color

Each time a cluster is created, a node in the Cetroid Views and Expression views folder is added which allows taking a look on this kind of representation of a gene cluster (Figure 4.9).

Time needed for calculation and used distance function can be determined in the general information folder.

It is also possible to save a cluster into a file, save the images, delete clusters or delete all clusters via the popup menu described above.
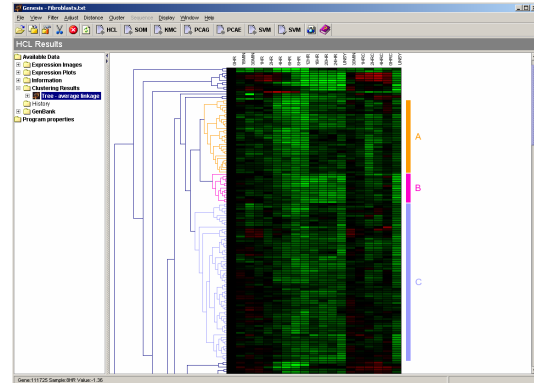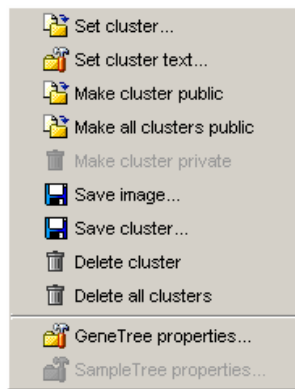
Figure 4.8: Colored clusters with description

Making clusters public

A very useful feature of Genesis is the ability to make clusters public. This means that all genes of a specific cluster are marked and visible in other panels and result views, e.g. PCA or SOM result panels in the same color. This gives you the ability to compare clusters in a easy way.

You can define specific clusters public by selecting a cluster and clicking onto the Make cluster public menu item in the popup menu or make all clusters public via the Make all clusters public entry.

Genes in a public cluster are marked with a little colored square on the right side of the expression image and are rendered with the cluster color in expression views (Figure 4.10 and Figure 4.11).

Figure 4.9: Expression view of cluster A
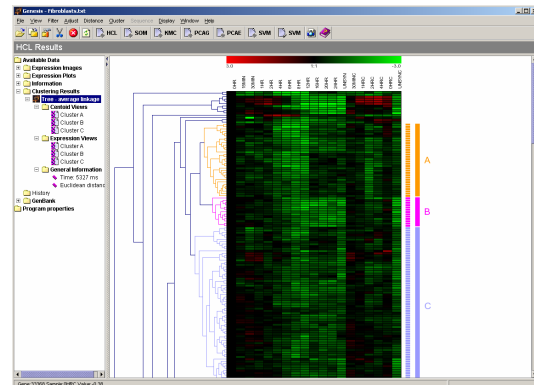
Figure 4.10: Public clusters in the result view.
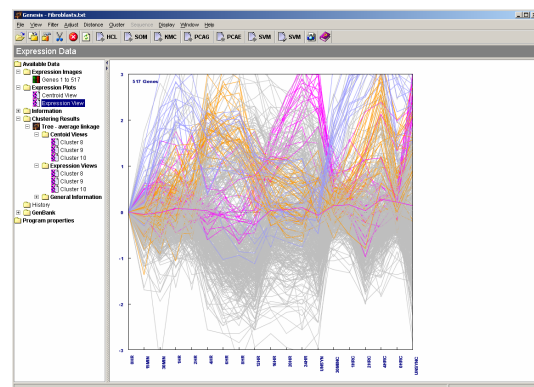
Figure 4.11: Public clusters in expression view

Important information

Many features of Genesis can be controlled via context menus which can be activated using the right mouse button.

Hierarchical Clustering

Clusters can be labeled by clicking on any node of the tree and then using the right mouse button (context menu) to specify the details.

PCA

The context menu can be reached by clicking with the right mouse button on the 3D-View leaf of the PCA-Result in the navigation tree located at the left panel of the program window.

Chapter 5

Chromosomal Mapping

It is possible to import GenBank flat files into the program by just adding another line to the XML-file that represents the program tree on the left part of the program window (Genesis.xml) or by using the simple wizards. To start a wizard simply select the leaf where you would like to add some sequence information and use the right mouse button. A popup menu will occur and you will be asked what to add and where the sequence file is located. If you would like to add a new genome for instance simply click on the GenBank leaf, press the right mouse button and select Add Genome.

Important notice:

Be very careful with the Genesis.xml file! This file is crucial for the operation of Genesis and if it gets corrupted, Genesis will not be able to start any more. Save the file before you make any changes!

Genesis is able to read exclusively GenBank flatfiles from NCBI (uncompressed text files with the extension *.gbk).

The sequence and annotation information contained in these files can be used for further analysis. One promising application is the mapping of gene expression studies onto a chromosome of the studied organism and the calculation of so called Chromosome correlation maps as it is described in the paper:

Cohen BA, Mitra RD, Hughes JD, Church GM. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet. 2000 Oct;26(2):183-6.

Consecutive genes are often similarly expressed and can be easily identified by this method (Figure 4.1). This can have various reasons. One is that the genes are so similar that cross-hybridization effects occur, a second could be that the consecutive genes are regulated by the same regulatory network. The latter information can be used for further analysis of the transcriptional program of these genes.

To demonstrate this program feature chromosome III of the model organism *saccaromyces cerevisieae* (yeast) is used, since its sequence is fully annotated. As microarray experiment a sporulation dataset form Stanford University (genome-www.stanford.edu) is chosen, because it contains almost all genes of the genome. Unfortunately the annotation is not very consistent, therefore not all genes from the experiment can be mapped. For that reason it is important that the official gene names from GenBank are used in the microarray data files (for the mapping the UniqueID from a gene in the expression file is compared with the gene name field [/gene="name"] in the annotation header of the GenBank flatfiles)!

Procedure:

1.) Open an expression data set (i.g. yeast sporulation dataset)
2.) Select the chromosome or contig you would like the expression data to be mapped on (i.g. yeast Ch. III)
3.) Select the menu item: Sequence->Map expression data

The mapping and cromosome correlation map will be calculated and displayed.
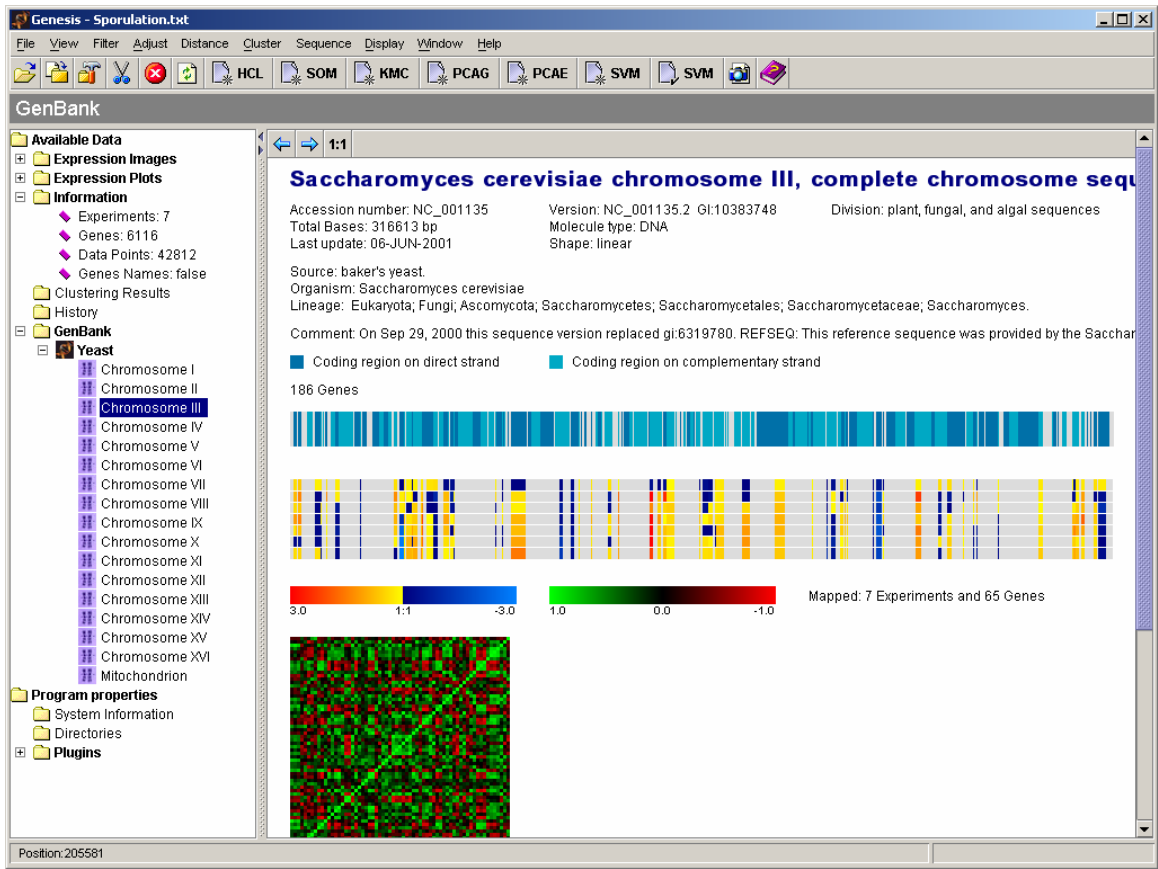
Figure 4.1: Program window showing the chromosomal mapping and chromosome correlation map of a yeast sporulation gene expression study onto chromosome III of saccharomyces cerevisiae. First 3 genes are consecutive and similarly expressed.

Appendix A

File format

Currently, Genesis only reads tab-delimited text files in a format, described and used by Michael Eisen from Stanford University. Such tab-delimited text files can be created and exported in any standard spreadsheet program, such as Microsoft Excel.

The relative expression levels of n genes, which may constitute almost the entire genome of an organism, are probed simultaneously by a single microarray. A series of m arrays, which are almost identical physically, probe the genome wide expression levels in m different samples - i.e., under m different experimental conditions. Let the matrix $\mathbf{X}$, of size (n-genes x m-arrays), tabulate the full expression data, where x_{ij} is the for instance $\log_2$ of the expression ratio of the ith gene in the jth sample as measured by the array j. The vector in the ith row of the matrix $\mathbf{X}$ lists the fluorescence ratio of the ith gene across the different samples, which correspond to the different arrays. The vector in the jth column of the matrix $\mathbf{X}$, lists the genome wide fluorescence ratios measured by the jth array.

$\mathbf{X}$ is presented in the following way (minimum information required):

UNIQID	Sample1	Sample 2	Sample 3	Sample 4	Sample 5
Gene 1	-0.275505795	-0.614654186	-1.444680246	-0.08685839	0.686924912
Gene 2	-0.863129991	0.217531843	-0.670604845	-0.941195232	-1.246143743
Gene 3	-0.091081723	-0.200490419	-0.124179411	0.357309311	-0.009984089
Gene 4	-0.182828898	0.278716028	0.067933676	-0.024940107	0.448381269
Gene 5	-0.353557047	1.606237652	0.510961919	0.686924912	0.193543785

Optional it is possible to specify a description for each gene. It has to be the 2nd column and the term in the 1st row has to be "NAME". The columns GWEIGHT and GORDER as well as the rows EWEIGHT and EORDER of the Stanford format are currently not taken into account by Genesis and are therefore ignored.

UNIQID	NAME	GWEIGHT	GORDER	Sample1	Sample 2
EWEIGHT				1	1
EORDER				1	2
Gene 1	Name 1	1	1	-1.444680246	-0.08685839
Gene 2	Name 2	1	2	0.217531843	-0.670604845
Gene 3	Name 3	1	3	-0.124179411	0.357309311
Gene 4	Name 4	1	4	0.067933676	-0.024940107
Gene 5	Name 5	1	5	0.510961919	0.686924912

There are 2 sample datasets in the Sample directory, one with gene descriptions (Fibroblast.txt) and one without (Sporulation.txt).

The files have to be tab delimited with the extension .txt