

GEWORKBENCH 1.8.0

USER MANUAL

(DECEMBER 2009)



caBIG[™] *cancer Biomedical
Informatics Grid*[™]

an initiative of the National Cancer Institute



Columbia University
Joint Centers for Systems Biology
Center for Computational Biology and Bioinformatics
Herbert Irving Cancer Center

National Centers for Biomedical Computing - MAGNet
National Cancer Institute Center for Biomedical Informatics
and Information Technology (NCI-CBIIT)

caBIG[®]

AMDeC - Academic Medicine Development Company

Manual Revision History:

Version 1.0 October 14th, 2004
Version 1.1 June 21st, 2005
Version 1.2 August 1st, 2006
Version 1.3 September 14, 2006
Version 1.4 January 9, 2008
Version 1.5 February 28, 2008
Version 1.5.1 April 28, 2008
Version 1.6 October 23, 2008
Version 1.8 December 28, 2009

Copyright and License

geWorkBench v1.8.0

SOFTWARE LICENSE AGREEMENT

Copyright 2004-2009 Columbia University.

This software was developed by Columbia University in conjunction with First Genetic Trust and the National Cancer Institute, and so to the extent government employees are co-authors, any rights in such works shall be subject to Title 17 of the United States Code, section 105.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

2. The end-user documentation included with the redistribution, if any, must include the following acknowledgment:

"This product includes software developed by the Columbia University, First Genetic Trust and the National Cancer Institute."

If no such end-user documentation is to be included, this acknowledgment shall appear in the software itself, wherever such third-party acknowledgments normally appear.

3. This license does not authorize the incorporation of this software into any proprietary programs.

4. THIS SOFTWARE IS PROVIDED "AS IS," AND ANY EXPRESSED OR IMPLIED WARRANTIES, (INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE) ARE DISCLAIMED. IN NO EVENT SHALL THE COLUMBIA UNIVERSITY OR THEIR AFFILIATES BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

5. Below is the list of all third party software used in geWorkbench and their license information.

This product includes software developed by the Apache Software Foundation. Batik, Xerces, and Xalan are part of Apache XML project. Byte Code Engineering Library, POI, Jakarta Commons are part of Jakarta project, Axis is part of Apache Web Services project. Log4J is part of Apache Logging Services project. ObjectRelationalBridge is part of the Apache DB project. All aforementioned Apache projects are trademarks of The Apache Software Foundation. For further open source licensing issues pertaining to Apache Software Foundation, visit:

<http://www.apache.org/LICENSE>

This product includes software developed by NCI Center for Bioinformatics (NCICB). caBIO is part of the caCORE project. caArray is cancer array informatics project. For more information, visit:

http://ncicb.nci.nih.gov/core/caBIO/technical_resources/core_jar/license

<http://ncicb.nci.nih.gov/download/caarraylicense.jsp>

This product may include the following software:

Cytoscape by the Institute for Systems Biology, University of California at San Diego, Memorial Sloan-Kettering Cancer Center and Institut Pasteur.

NetX by J. Maxwell, ODE For Java by Tim Schmidt.

OpenJGraph by Jesus M. Salvo, Jr.

Java Excel API by Andy Khan.

JMOL by molvisions.com

BioJava by BioJava.org.

JSCi by Mark Hale.

Ensemble for Java by the Sanger Institute and the European Bioinformatics Institute.

JGraph by JGraph Ltd.

These software products are licensed under the Lesser General Public License. For more information,

visit:

<http://www.gnu.org/copyleft/lesser.html>

This product may include the following software:

Bayesian Network tools in Java by Kansas State University.

Java Hidden Markov Models (JAHMM) by Jean-Marc François.

JFreeChart by David Gilbert.

the Ostermiller utils by Stephen Ostermiller.

Weak by the University of Waikato

These software products are licensed under the General Public License. For more information, visit:

<http://www.gnu.org/copyleft/gpl.html>

This product may include the following software:

ArrayExpress by the European Bioinformatics Institute.

Ogsa from Globus Alliance.

JDOM by Jason Hunter and Brett McLaughlin.

Looks by Karsten Lentzsch.

PureTLS by Eric Rescorla.

SkinLF by Frédéric Lavigne.

Jaxen by The Werken Company.

Dom4J by MetaStuff, Ltd.

Piccolo by the University of Maryland.

Ontologizer 2.0 by Peter Robinson and Sebastian Bauer.

These software products are licensed under the BSD or BSD style License. For more information, visit:

<http://www.gnu.org/philosophy/license-list.html#OriginalBSD>

This product may include following public domain software:

AntLR by Terence Parr.

Distributions by the University of Edinburgh
Java Matrix Package by MathWorks and NIST.
SplashBitmap by Kai Blankenhorn

This product may include the following software:

AspectJ by the Eclipse Foundation.

JUnit by Erich Gamma and Kent Beck.

AntLR by Terence Parr.

Distributions by the University of Edinburgh

Java Matrix Package by MathWorks and NIST.

WSDL4j by IBM, Inc.

These software products are licensed under the Common Public License. For more information, visit:

<http://www.eclipse.org/legal/cpl-v10.html>

This product may include the following software:

Eleritec Docking Framework by Marius. This software is under MIT license. For more information, visit: <http://www.eleritec.net/>

This product may include the following software:

NetComponents by Original Reusable Objects, which is under its own license. For more information, visit:

<http://www.savarese.org/oro/downloads/NetComponentsLicense.html>

This product may include the following software:

ARACNE by Andrea Califano's lab at Columbia University (<http://wiki.c2b2.columbia.edu/califanolab>).

This software has its own license, provided below in this document. geWorkbench users should use ARACNE in agreement with the terms of this license.

All other product names mentioned herein and throughout the entire project are trademarks of their respective owners.

<i>geWorkbench 1.8.0 Development Project Members¹</i>		
<i>Development</i>	<i>Documentation</i>	<i>Program Management</i>
Zhou Ji	Kenneth Smith	Aris Floratos
Kiran Keshav	Aris Floratos	Kenneth Smith
Yih-Shien (Mark) Chang	Mary VanGinhoven	Zhou Ji
Thomas Garben		Kiran Keshav
Michael Honig		
Nikhil Reddy Podduturi		
Udo Többen		
Oleg Shteynbuk		
Min You		
Meng Wang		
¹ All contributors are current or former members of the Joint Centers for Systems Biology, Columbia University, New York, NY.		

<i>Contacts and Support</i>	
caBIG [®] Molecular Analysis Tools Knowledge Center (MATKC)	https://cabig-kc.nci.nih.gov/Molecular/KC/index.php/Main_Page
geWorkbench User and Developer Discussion Forums (hosted by the MATKC)	https://cabig-kc.nci.nih.gov/Molecular/forums/
geWorkbench project website	http://www.geworkbench.org

Table of Contents

TABLE OF CONTENTS	I
1 INTRODUCTION TO THE MANUAL	1
1.1 CHANGES IN MANUAL VERSION 1.8.....	1
1.2 CHANGES IN MANUAL VERSION 1.6.....	2
1.3 CHANGES IN MANUAL VERSION 1.5.1.....	2
1.4 CHANGES IN MANUAL VERSION 1.5.....	2
1.5 CHANGES IN MANUAL VERSION 1.4.....	2
1.6 GETTING STARTED WITH GEWORKBENCH	2
1.7 DOCUMENT TEXT CONVENTIONS	3
2 OVERVIEW OF THE SOFTWARE.....	5
2.1 MOTIVATION	5
2.2 INTRODUCTION TO GEWORKBENCH.....	5
2.3 THE GEWORKBENCH APPROACH TO INTEGRATED GENOMICS	6
2.4 COMPONENTS OF GEWORKBENCH	7
2.5 SUPPORTED DATATYPES.....	7
3 VISUAL INTERFACE AND DATA MANAGEMENT.....	9
3.1 LAYOUT OF THE GEWORKBENCH INTERFACE	9
<i>Menu Bar</i>	<i>10</i>
<i>Data management area (1).....</i>	<i>10</i>
<i>Set selection and management (3).....</i>	<i>10</i>
<i>Visualization and Analysis tools (2 and 4).....</i>	<i>10</i>
3.2 ONLINE HELP.....	10
3.3 WORKING WITH DATA FILES	11
3.3.1 Workspaces – A brief overview.....	11
3.4 THE PROJECT FOLDERS COMPONENT	12
3.4.1 Example of opening a local microarray data file.....	15
3.4.2 Other file operations – Merging and renaming	18
4 QUERYING CAARRAY	20
4.1 SEARCHING CAARRAY USING MAGE ANNOTATIONS	20
5 COMPONENT CONFIGURATION MANAGER.....	27
5.1 OVERVIEW.....	27
5.2 INDIVIDUAL CONTROLS.....	28
6 ANALYSIS COMPONENT FRAMEWORK.....	30
6.1 OVERVIEW.....	30
6.2 LAYOUT OF THE ANALYSIS FRAMEWORK	30
6.2.1 Lists.....	30
6.2.2 Analysis Actions.....	30
6.2.3 Analysis Parameters	31
6.3 CREATING SAVED PARAMETER SETS	31
6.4 INTERPLAY OF PARAMETERS AND LIST.....	32
7 MICROARRAY DATA ANALYSIS	33
7.1 SET SELECTION (THE MARKERS/ARRAYS/PHENOTYPES) COMPONENTS.....	33
7.1.1 Marker Sets.....	33
7.1.2 Set Activation and Manipulation.....	36
7.1.3 Array/Phenotype Sets.....	37
7.1.4 The Commands Menu.....	37

7.2 THE VIEW WINDOW.....	38
7.2.1 The Microarray Viewer.....	39
7.2.2 The Expression Profiles Tool.....	42
7.2.3 The Color Mosaic View.....	42
7.2.4 The Tabular View.....	44
7.2.5 The Image Viewer.....	45
7.2.6 Scatter Plot.....	46
7.2.7 Expression Value Distribution.....	47
THE ANALYSIS/ANNOTATION WINDOW.....	49
7.3 FILTERING OPERATIONS.....	49
7.3.1 Normalization Tools.....	51
7.3.2 Dataset History.....	52
7.4 THE ANALYSIS TOOLS.....	52
7.4.1 Hierarchical Clustering:.....	53
7.4.2 Self Organizing Map (SOM).....	55
7.4.3 The Dataset Annotation Tool.....	56
7.4.4 The Experiment Info Tool.....	57
8 DIFFERENTIAL EXPRESSION (T TEST).....	59
8.1 OVERVIEW.....	59
8.2 PREPARATION.....	59
8.3 T-TEST PARAMETERS.....	59
8.3.1 P-value.....	59
8.3.2 Alpha corrections.....	59
8.3.3 Degrees of Freedom.....	59
8.3.4 Classification.....	59
8.3.5 Set Analysis Parameters.....	61
8.3.6 t-Test Results.....	62
9 MARKER ANNOTATIONS.....	65
9.1 OVERVIEW.....	65
9.2 SUBMIT QUERY.....	65
9.3 PATHWAY AND GENE ANNOTATIONS.....	66
9.4 CANCER GENE INDEX.....	69
10 SEQUENCE RETRIEVAL.....	75
10.1 OVERVIEW.....	75
10.2 PREREQUISITES.....	75
10.3 EXAMPLE - RETRIEVING SEQUENCES FOR A LIST OF GENE MARKERS.....	75
10.3.1 Obtaining a set of markers.....	75
10.4 SAVING THE SEQUENCES TO AN EXTERNAL FASTA FILE.....	79
11 SEQUENCE ALIGNMENT.....	81
11.1 OVERVIEW.....	81
11.2 BLAST.....	81
11.3 BLAST JOB SETUP.....	82
11.3.1 Prerequisites.....	82
11.3.2 Query sequences.....	82
11.3.3 Parameters - Main.....	83
11.3.4 Parameters - Advanced Options.....	84
11.3.5 General controls.....	85
11.4 BLAST RESULTS VIEWER.....	85
11.4.1 Controls.....	87
11.5 SUBMITTING A BLAST JOB.....	87
11.6 EXAMPLE: RUNNING A BLAST SEARCH.....	88
11.7 REFERENCES.....	89

12 PATTERN DISCOVERY	90
12.1 OVERVIEW	90
12.2 TUTORIAL	90
12.2.1 <i>Discovery analysis</i>	91
12.2.2 <i>Hierarchical analysis</i>	94
12.2.3 <i>Exhaustive analysis</i>	95
12.3 VISUALIZATION OF PATTERN DISCOVERY RESULTS	96
12.4 COMPONENT VISUAL ELEMENTS	96
12.5 REFERENCES	99
13 PROMOTER ANALYSIS	100
13.1 OVERVIEW	100
13.2 JASPAR CORE DATABASE	100
13.3 WORKING WITH THE PROMOTER GRAPHICAL INTERFACE	100
13.3.1 <i>Prerequisites</i>	100
13.3.2 <i>Layout</i>	100
13.3.3 <i>TF Mapping tab</i>	101
13.3.4 <i>The LOGO tab</i>	103
13.3.5 <i>The Parameters tab</i>	104
13.3.6 <i>The Sequence tab</i>	105
13.3.7 <i>Implementation details</i>	107
13.4 SCAN IMPLEMENTATION	107
13.4.1 <i>Normalization and the Pseudocount</i>	107
13.4.2 <i>Scoring</i>	108
13.5 EXAMPLE: RUNNING AND VIEWING A SCAN	108
13.5.1 <i>Prerequisites</i>	108
13.5.2 <i>Running the scan</i>	110
13.6 REFERENCES	113
14 ANALYSIS OF VARIANCE (ANOVA)	115
14.1 OVERVIEW	115
14.2 SETTING UP AN ANOVA RUN	115
14.2.1 <i>Prerequisites</i>	115
14.2.2 <i>ANOVA Parameters and Settings</i>	116
14.3 SERVICES (GRID)	118
14.4 WORKING WITH AND VIEWING ANOVA RESULTS	118
14.4.1 <i>Significant markers set</i>	118
14.4.2 <i>The ANOVA result node in the Project Folders component</i>	118
14.4.3 <i>Color Mosaic Viewer</i>	119
14.4.4 <i>Tabular Viewer</i>	120
14.5 DATASET HISTORY	121
14.6 EXAMPLE OF RUNNING ANOVA	121
14.6.1 <i>Prerequisites</i>	122
14.6.2 <i>Loading and preparing the example data</i>	122
14.6.3 <i>Choosing array groups</i>	122
14.6.4 <i>Setting up the parameters and starting ANOVA</i>	123
14.6.5 <i>Results</i>	124
14.7 REFERENCES	124
15 ARACNE	127
15.1 OVERVIEW	127
15.2 SETTING UP AN ARACNE RUN	128
15.2.1 <i>Prerequisites</i>	128
15.2.2 <i>Parameters and Settings</i>	128
15.3 SERVICES (LOCAL VS GRID)	133

15.3.1 <i>Special Note on running in PREPROCESSING mode on caGRID</i>	133
15.4 VIEWING ARACNE RESULTS.....	133
15.5 DATASET HISTORY	135
15.6 EXAMPLE OF RUNNING ARACNE.....	135
15.6.1 <i>Prerequisites</i>	135
15.6.2 <i>Loading the example data</i>	135
15.6.3 <i>Setting up the parameters and starting ARACNe</i>	136
15.7 REFERENCES.....	137
16 USING CAGRID ANALYTICAL SERVICES.....	138
16.1 OVERVIEW.....	138
16.2 SERVICES TAB.....	138
16.2.1 <i>Local/Grid</i>	138
16.2.2 <i>Change Index Service</i>	139
16.2.3 <i>Change Dispatcher</i>	139
16.2.4 <i>Grid Services</i>	139
16.2.5 <i>Service Details</i>	140
16.3 RUNNING A GRID JOB	140
16.4 FURTHER ASPECTS OF RUNNING GRID JOBS	141
17 CYTOSCAPE	143
17.1 OVERVIEW.....	143
17.2 LAYOUT OF THE CYTOSCAPE COMPONENT.....	143
17.3 SELECTING NODES IN CYTOSCAPE	145
17.4 SET OPERATIONS ON NETWORKS.....	147
17.5 PROJECTING MARKER SETS ONTO CYTOSCAPE.....	148
17.6 ALTERING THE VIEW IN CYTOSCAPE.....	149
17.7 NETWORK COMMANDS	150
17.7.1 <i>Edit Network Title</i>	151
17.7.2 <i>Create View</i>	151
17.7.3 <i>Destroy View</i>	151
17.7.4 <i>Destroy Network</i>	151
APPENDIX A. ERROR MESSAGES/INDICATORS AND PROBLEM RESOLUTIONS.....	153
APPENDIX B. GLOSSARY.....	155

1 Introduction to the Manual

This manual is intended for the users of geWorkbench. It is directed at the bench scientist and bioinformatician. It explains the basic principles, design goals, and uses of geWorkbench, centered around the single or joint analysis of gene expression microarray and sequence data. A separate Installation Guide is available.

While extensive explanations of how to use the software are given in this manual, further tutorial information can be found on-line at www.geworkbench.org. The tutorials at that site are the primary documentation for geWorkbench. This manual is based in large part on the web-based tutorials.

This manual will cover the basic operations of geWorkbench and its core components. This revision of the manual pertains primarily to release 1.8.0. New modules for geWorkbench continue to be developed, and manual pages for them are made available at www.geworkbench.org.

Documentation for following components is only available in the online Wiki tutorials:

- [Cellular Networks KnowledgeBase](#)
- [Classification](#)
- [Gene Ontology Term Over-representation](#)
- [Jmol](#)
- [Mark-Us](#)
- [Master Regulator Analysis](#)
- [MatrixREDUCE](#)
- [MINDy](#)
- [Pudge](#)

1.1 Changes in manual version 1.8

The manual has been updated to reflect changes included in version 1.8.0 of geWorkbench. Extensive tutorials on new features are available at www.geworkbench.org. The primary source of documentation is now the Tutorial section on the wiki. This manual has been updated to incorporate chapters from the Tutorials that pertain to core functionality of geWorkbench. The following sections were so updated: Local and remote (caArray) file open, Marker Annotations (now includes Cancer Gene Index), Sequence Alignment (BLAST), Promoter, ANOVA, ARACNe and

Grid Services. New sections include the Component Configuration Manager, Differential Expression and Sequence Retrieval. The section on caScript is no longer included.

1.2 Changes in manual version 1.6

Figures updated to reflect changes in many of the graphical components.

1.3 Changes in manual version 1.5.1

The caArray query mechanism has changed slightly. It no longer uses the MAGE-OM query mechanism. Instead, a Java API is used. Some aspects of the geWorkbench graphical interface used for forming a query against caArray were simplified. Screenshots for several other components of the geWorkbench GUI were updated.

1.4 Changes in manual version 1.5

This release of the manual adds a chapter on the ANOVA component for Analysis of Variance calculations.

1.5 Changes in manual version 1.4

This release of the manual incorporates material that was previously released as a separate supplement, entitled “Advanced Services”. Topics include:

1. use of geWorkbench within the context of the caGRID infrastructure. Several analytical routines already supported directly within geWorkbench have been developed as formal caGRID services, with an appropriate service interface present within geWorkbench. They are initially intended to be used in the analysis of microarray data. They are:
 - a. Hierarchical Clustering
 - b. SOM (Self-Organizing Maps)
 - c. ARACNE (a gene network reverse-engineering tool)
2. a query interface for caARRAY which allows searches on available annotation fields
3. use of the caSCRIPT scripting language developed specifically for geWorkbench to automate the running of repetitive or complex tasks.

1.6 Getting Started with geWorkbench

To get started with geWorkbench you may refer to the following sections of this manual:

- Review Chapter 2 for a brief overview of the software
- Review Chapter 3 to learn about the Graphical User Interface, including basic file operations.
- Refer to Chapters 4 through 18 cover the basic a description of how to use the core modules of geWorkbench.
- For information on remote access to services via caGrid, consult Chapter 17.

Detailed instructions and step-by-step tutorials on how to install and run geWorkbench are available online at <http://www.geworkbench.org/>.

1.7 Document Text Conventions

The following table shows various typefaces to differentiate between regular text and menu commands, keyboard keys, and text that you type. This illustrates how conventions are represented in this guide.

Table 1-1 Document Conventions

Convention	Description	Example
Bold & Capitalized Command Capitalized command > Capitalized command	Indicates a Menu command Indicates Sequential Menu commands	Admin > Refresh
TEXT IN SMALL CAPS	Keyboard key that you press	Press ENTER.
TEXT IN SMALL CAPS + TEXT IN SMALL CAPS	Keyboard keys that you press simultaneously	Press SHIFT + CTRL and then release both.
Boldface type	Options that you select in dialog boxes or drop-down menus. Buttons or icons that you click.	In the Open dialog box, select the file and click the Open button.
<i>Italics</i>	Used to reference other documents, sections, figures, and tables.	<i>caCORE Software Development Kit 1.0 Programmer's Guide</i>
<i>Italic boldface type</i>	Text that you type	In the New Subset text box, enter <i>Proprietary Proteins</i> .
Courier typestyle	Used for filenames, directory names, commands, file listings, source code examples and anything that would appear in a Java program, such as methods, variables, and classes.	URL_definition ::= url_string
Note:	Highlights a concept of particular interest	Note: This concept is used throughout the installation manual.

Warning!	Highlights information of which you should be particularly aware.	Warning! Deleting an object will permanently delete it from the database.
{}	Curly brackets are used for replaceable items.	Replace {root directory} with its proper value such as c:\cabio

2 Overview of the Software

2.1 Motivation

Recent advances in high-throughput genomic technologies, spurred on in part through the Human Genome Project, have opened the flood-gates to many different types of biological data. For example, NBCI provides open access to genome sequences of over 1000 organisms; nucleotide and protein sequences (e.g., GenBank, RefSeq, Swiss-Prot, PIR etc.), 3D macromolecular structures; population study data sets, catalogs of human disease genes, genetic markers or tagged-sites database (SNP, EST, STS), molecular modeling and genome mapping information. These developments directly influence biomedical research. However, making use of this cornucopia of information is difficult for investigators because most laboratories lack the tools to integrate the data into their own studies.

Although a large selection of bioinformatics software tools is available, these have been developed as individual software programs and do not readily interface with other software. Differences in application design, programming language used for implementation, and input/output requirements restrict their use to certain operating systems, and/or impose data reformatting requirements. Furthermore, management of any complex biological data (e.g. combining output from two different gene-expression clustering tools) usually requires custom programming, because even though concepts such as a gene expression cluster are well understood and ubiquitous in the literature, their representation has not been standardized.

2.2 Introduction to geWorkbench

geWorkbench is an open-source platform for bioinformatics data analysis . It supports a growing collection of self-contained software modules for management, analysis and visualization of a range of biological research data. It also provides integration of external databases and services into the local desktop client. The overriding goal of geWorkbench is to provide biomedical researchers with a user-friendly application that can link the analysis of disparate data types. It is an extension of a project originally sponsored by the National Cancer Institute Center for Bioinformatics (NCICB) to develop tools for microarray data analysis (caWorkBench).

geWorkbench has been primarily constructed for analysis of data derived from gene expression microarray experiments, and allows pulling in many different resources to this end, including sequence, gene ontology, promoter analysis, and standard analytic techniques such as the t-test, hierarchical clustering, and gene network reverse-engineering.

geWorkbench has a modular, component-based design. New modules can easily be written and added as the need arises. A primary aim is to allow easy integration of different forms of data analysis. Such integration removes the common hindrance of needing to reformat data for each different type of analysis undertaken.

Extensive documentation and training material for geWorkbench can be found on its main website at <http://www.geworkbench.org/>. There are wiki-based tutorials there for almost all components of the application. These tutorials are more applied in nature than the material in the printed manual. The software can be downloaded via links found on the 'Download' section of that site. Those links refer to the actual archival location of the software, which is the GForge site maintained by the NCICB. All official releases of the software can be downloaded from that site.

This manual provides a detailed view of the modules that make up the core functionality of geWorkbench. Tutorials and examples are available on the application website, www.geworkbench.org. The application download area can be reached from that site or directly from <http://gforge.nci.nih.gov/projects/geWorkbench/>. Information about additional modules not covered in this manual can also be found at www.geworkbench.org.

2.3 The geWorkbench Approach to Integrated Genomics

We believe that biomedical researchers will be best served by the establishment of a standardized, fully integrated bioinformatics software infrastructure (such as geWorkbench) that supports not only heterogeneous data and models, but also algorithms, management, and visualization tools that can be seamlessly integrated and distributed within the biomedical scientific community. It is this realization that is the central motivation of the geWorkbench framework. The latter, then, attempts to address the following needs:

1. Sharing not just a growing set of biological data types and data sets, but also a growing set of application software tools to manipulate them.
2. Allowing different modules to interact with each other based on their semantic compatibility (that is, the programmatic interfaces they implement). This is like having two individuals that communicate not just because they have a "vocabulary" translating individual words into their native languages but because they know how they are assembled into meaningful sentences and concepts (semantics).
3. Supporting automatic event-driven computations and data analysis and visualization workflows in a distributed environment that operates transparently to the end-users.

2.4 Components of geWorkbench

geWorkbench: geWorkbench is a Java application which is run on the User's local Windows, Macintosh or Linux workstation. This main application also serves as a front-end client to a number of external computational and data services. Such services already present in geWorkbench include the ability to run BLAST jobs on NCBI servers, and to retrieve gene, pathway and sequence information from sources such as UC Santa Cruz and the NCICB. A built-in, interpreted, Java-like scripting language, caSCRIPT, can be used to automate tasks within geWorkbench.

caGRID: caGrid is a project sponsored by the National Cancer Institute to link and make available data stored in Cancer Centers throughout the United States. . caGrid provides a mechanism for all data and parameters passed on the grid to be of known, registered types, to facilitate interoperation. geWorkbench can be used as a client for such grid services and several such modules have been implemented.

caArray: caArray is a MIAME-supportive database system for microarray data developed by the National Cancer Institute. geWorkbench supports querying against such databases and retrieval of expression information.

Central to data integration in geWorkbench is a mechanism that allows independently built tools and data sources to communicate in a meaningful fashion. This mechanism, termed "component semantics interoperability," facilitates construction of complex biomedical applications from simple components, much like building complex assemblies from Lego pieces. It is implemented through the exchange or broadcast of well-defined messages.

2.5 Supported Datatypes

geWorkbench currently is oriented towards the integration of microarray gene expression and sequence data. Examples of data types handled by include:

- microarray gene expression data (Affymetrix, GenePix)
- Sequence data
 - (e.g., DNA, RNA, protein sequences)
- complex multi-dimensional data-types
 - biochemical pathways
 - gene regulatory pathways

Examples of external data sources and services provided through geWorkbench include

- NCBI BLAST
- Server-side implementation of pattern discovery algorithms.
- UC Santa Cruz (GoldenPath) genome sequence retrieval.
- EBI protein sequence retrieval.
- Access to NCI databases including
 - CGAP gene annotations
 - BioCarta pathway diagrams
 - caArray gene expression data
 - Pathway Interaction Database
 - Cancer Gene Index

Note: Previous versions of this program appeared under the names BioWorks and caWorkbench

3 Visual Interface and Data Management

3.1 Layout of the geWorkbench Interface

Figure 3-1 shows a screenshot of geWorkbench’s graphical interface. The workspace is divided into 4 resizable panels whose functionality is further defined by the folder tabs running across the top of each panel. Each panel can be arbitrarily resized by clicking on an edge of that panel’s frame and dragging the mouse. In addition, the triangular shaped wedges (on the left sides of the horizontal separators and at the top of the vertical separator) can be clicked on to maximize that frame vertically and/or horizontally.

Each of these configurable panels is described in more detail in the sections that follow; the purpose of this section is to provide an overall orientation. Moving from left to right and from top to bottom, these panels include a *Project* window (1), a *View* window (2), a *Selection* window (3), and an *Analysis/Annotation* window (4).

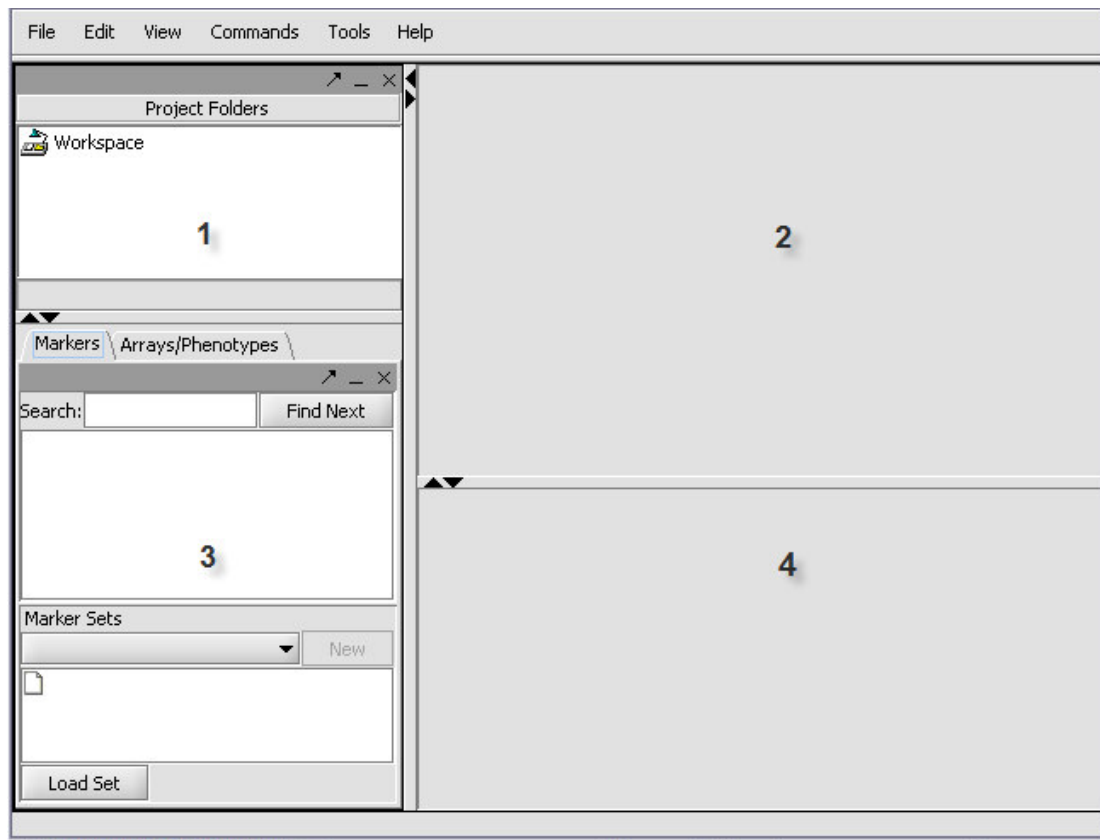


Figure 3-1 Layout of the geWorkbench Graphical Interface

Menu Bar

The GUI provides a menu bar at top with a standard choice of commands. Many commands that are available in the menu bar are also available by right-clicking on data objects.

Data management area (1)

Working with geWorkbench involves creating a project within the top-level workspace. Open data files and the results of data transformation or analysis are stored within a project. A workspace can contain more than one project at a time, allowing data to be organized as desired. A workspace and all the projects and data within it can be saved and later reloaded.

Set selection and management (3)

A key feature of geWorkbench is the ability to work with defined sets of markers or arrays. This allows subsets of data to be analyzed, and allows for passing of selected subsets of data between different components. For example, the t-test can be used to create a list of markers showing a significant difference in expression between two states, and this list can then be used to retrieve relevant sequences or annotations.

Visualization and Analysis tools (2 and 4)

geWorkbench works such that only the visualization and analysis components relevant to the type of dataset currently selected in the Project Folders area (1) are displayed through tabs in their respective areas (2 and 4). Thus choosing a microarray dataset will result in a different set of tabs being displayed as compared with those seen when a nucleotide sequence file is selected. When a new data file is loaded, or an analysis produces a new data set, not only is it added to the Project area (1), but an appropriate viewer in the Visualization area (2) is automatically selected.

3.2 Online Help

Figure 3-2 shows the **Online Help** interface. **Online Help** is found as a menu item under **Help** on the top menu bar. **Online Help** is provided for all geWorkbench modules which have been included in a formal release. They focus on the actual use of particular controls within a given module, e.g. button actions, definition of parameters etc.

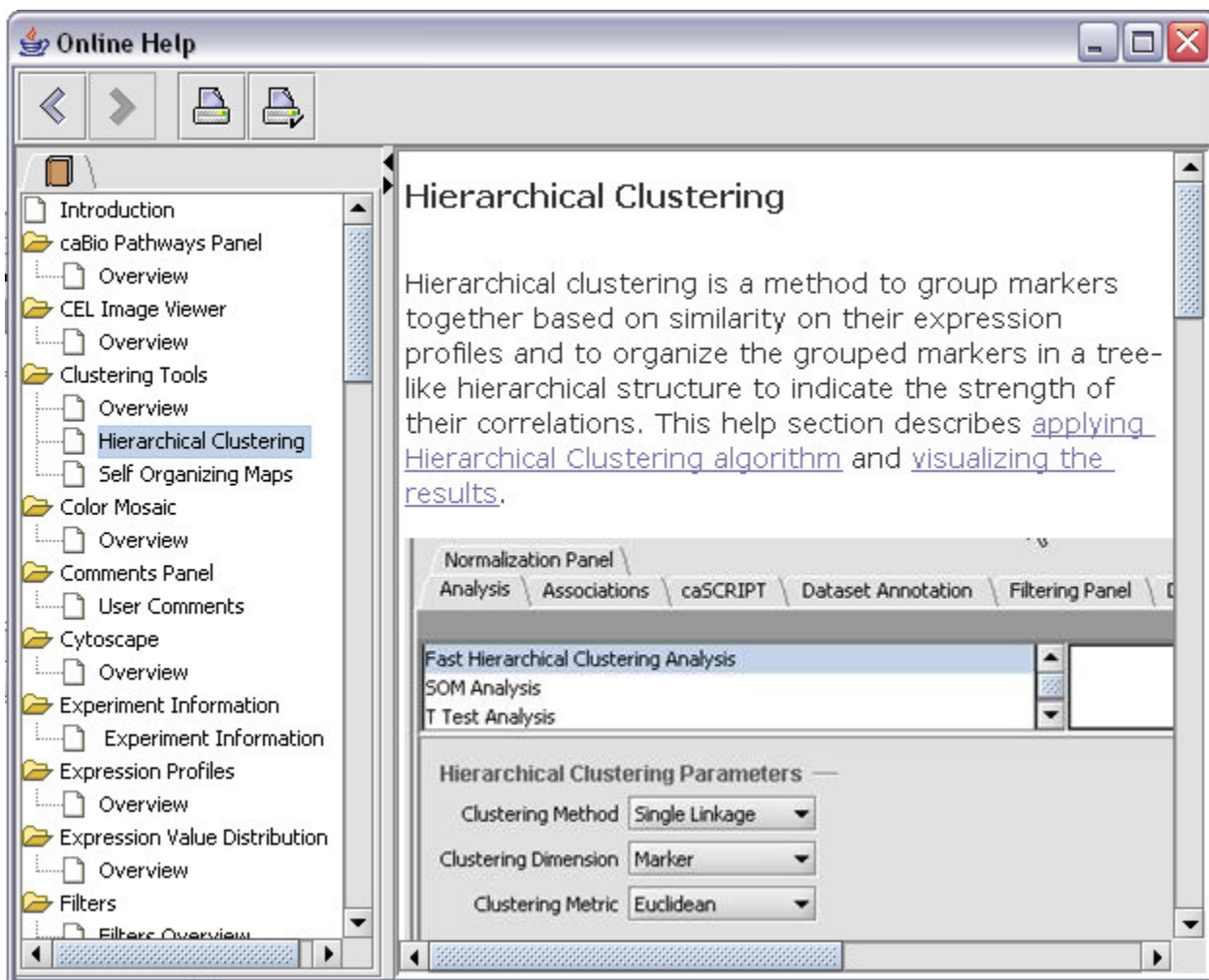


Figure 3-2 Online Help

3.3 Working with Data Files

3.3.1 Workspaces – A brief overview

The top level of organization of data in geWorkbench is the Workspace. A Workspace can contain any number of Projects, which are used to organize data and results.

geWorkbench organizes data files using a *workspace/project* paradigm. A project is analogous to a “virtual” folder, as it allows individual data sets to be grouped together without modifying their physical storage locations. Once a project’s data sets have been defined, it is possible to open, save, or close all data sets in that project with a single action.

A typical use of the project facility is to associate multiple data sets from the same experiment with one another in a single project folder. In addition to loaded data files, other types of data generated during the session—e.g., images, results from various steps

analysis such as clustering, etc.—are also saved associated with their parent dataset in a particular project.

Multiple projects can, in turn, be managed within a single workspace. A user can create a project in a workspace, delete an existing project from a workspace, or rename a project. The application supports handling an arbitrary number of projects in a single workspace.

In summary, a workspace may contain multiple projects, which may themselves contain a variety of raw, filtered, normalized, or otherwise annotated microarray data sets. Workspaces are saved as files with a *.wsp* extension. Projects can only be saved and/or accessed as part of a workspace

In addition to the controls provided within each individual component, a main menu bar appears at the top of the screen. The first menu option, **File**, is used to manage the opening, creating, deleting, and saving of workspaces, projects, and files. Most of the options included in the main menu bar are applicable to the Project and Marker/Phenotype windows, and are described below.

3.4 The Project Folders component

The Project Folders component provides a centralized area for managing projects and files in the current workspace. Operations on this window are controlled by the **File** and **Edit** options in the main menu bar. Most of these menu options can also be accessed by right clicking the mouse when it is located over an appropriate element in the Project or Marker/Phenotype window.

The operations described here all manipulate the workspace, projects, files, and images that are visible in the Project Folders component. Only a single workspace can be open at one time, but multiple projects, files, and images can be managed within that workspace. The Project Folders component uses a hierarchical treelike structure to manage these elements, as shown by the example in Figure 3-3.

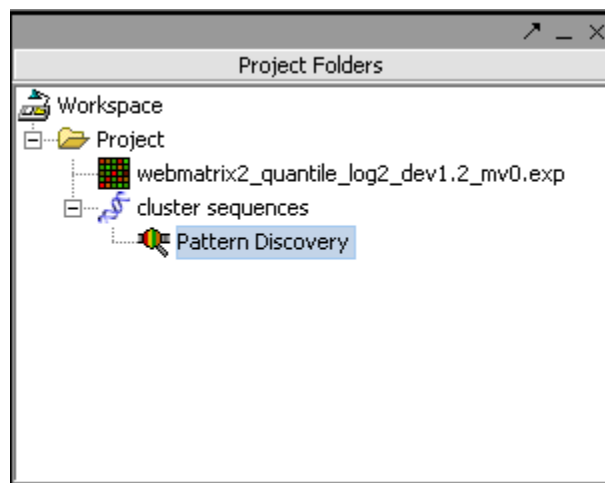


Figure 3-3 An Example Project Tree

Used to capture a “working session,” the workspace is represented by a folder icon at the very top of the file hierarchy into which all of the data generated during a user session can be subsumed. Items contained in a workspace include projects, which may themselves contain a variety of raw, filtered, normalized or otherwise annotated microarray data sets.

Multiple projects can be accommodated under one workspace heading. Multiple data sets and derivatives thereof can be grouped within a single project. It is important to note, however, that some operations require data sets to be part of the same project, and in some cases, in the same file. For instance, two microarrays cannot be viewed side by side unless they have been merged into one file. Moreover, two data sets cannot be merged into one file unless they are included in the same project.

illustrates the expanded options for the first three operations under **File**, i.e., **Open**, **Save**, and **New**. These operations can be selectively applied to open, save, or create workspaces, projects, and files. (a) shows the submenus for opening a data file or a workspace. Note in contrast in (b), only workspaces can be explicitly saved here. Files, once open in a Project, are saved as part of their workspace. (c) illustrates creating a new workspace or project.

Selecting the **File→Open→File** operation without having first defined a project in which to open that file will generate a prompt advising the user to first select a project in the Project window.

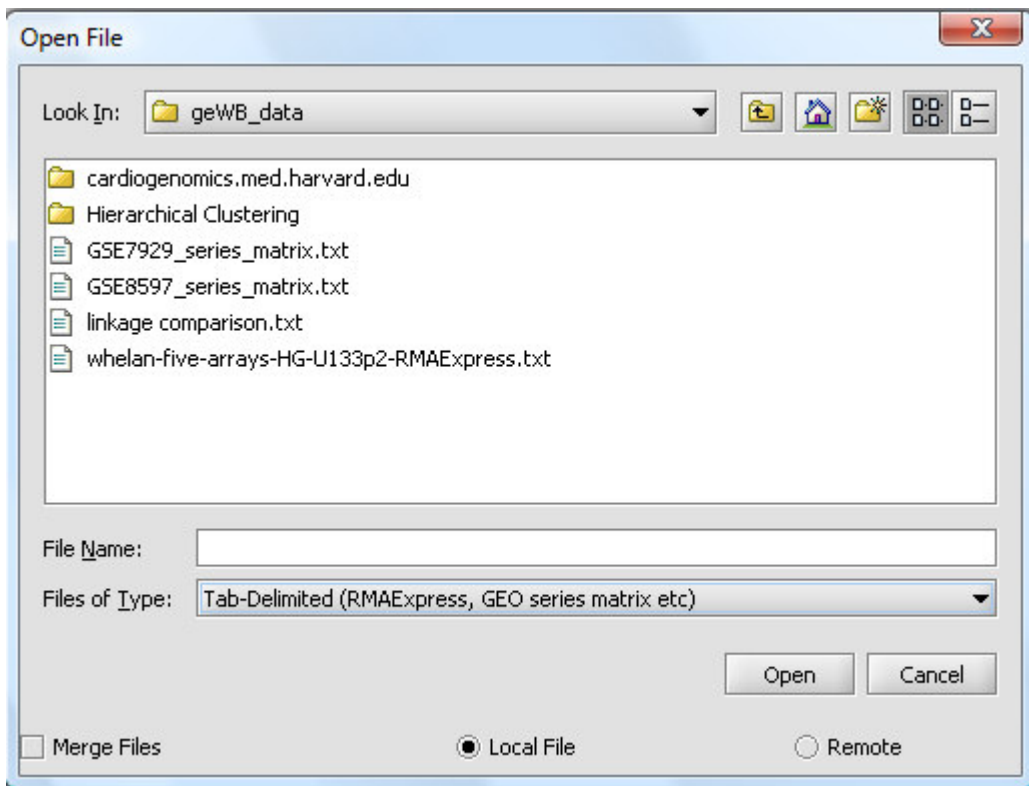
When the application starts, a blank workspace is created. A new project can be created in the workspace by selecting **File→New→Project** from the drop-down menu (

The default option in the pop-up "Open File" dialog box is to open a local file, as indicated by the pre-selected radio button at the bottom of the dialog box in Figure 3-4(a). Selecting the appropriate file type from the pull-down list in the local file dialog box will display all files of that type. Selecting a file from the list of those available and clicking on the **Open** button or double-clicking will then close the file dialog box and add that file to the current project. A small set of example datasets are available with the download package, in the *Sample Data* directory, and additional examples are available as part of the Tutorial dataset at www.geworkbench.org.

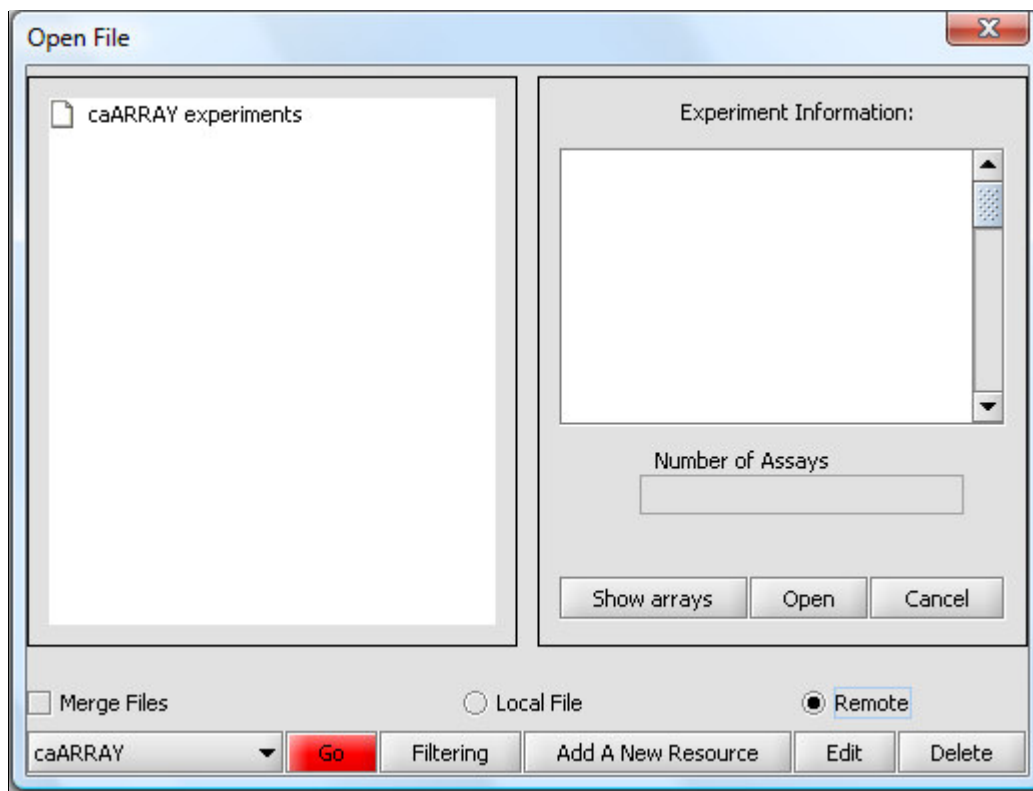
Alternatively, to access remote data sets stored on a remote source such as NCI's caArray server, you must first select the **Remote** radio button at the bottom of the dialog box, as shown in Figure 3-4(b). As seen there, the left-hand panel lists the files available for selection, and a scrollable text window on the right displays information about the currently selected experiment.

To open a file, begin by right clicking on the selected experiment in the left panel. Selecting a file from that list and pressing **Open** in the right panel will import that file to the current project.

In summary, the **File→Open** command can be used to add files to a selected project and to open workspaces. Files can be opened from a local disk, and in the case of microarray gene expression data from a remote instance of caArray, and are always saved as part of a project. Finally, opening or creating a new workspace will drop any unsaved work that has been done in the current workspace, so be sure to save your work before performing either of these actions.



(a)



(b)

Figure 3-4 The Local (a) and Remote (b) Open File Dialog Boxes

3.4.1 Example of opening a local microarray data file

3.4.1.i Prerequisites

Certain aspects of the functionality of geWorkbench currently depends on microarray annotation files. For example, such files are supplied by Affymetrix for their microarray chips. Due to licensing restrictions, these Affymetrix files are not distributed with geWorkbench as part of formal releases. The examples in this portion of the User Manual do not depend on annotation information. If you nonetheless would like to work with the full functionality of geWorkbench, the relevant file for the dataset used in this manual can be downloaded from the Affymetrix.com support web site. The file name has the form “HG_U95Av2.na29.annot.csv”, where “29” is the version number.

3.4.1.ii Example

geWorkbench includes sample data files. In the example below we will open a set of Affymetrix MAS5 format files. These files are part of the geWorkbench tutorial data set (see http://wiki.c2b2.columbia.edu/workbench/index.php/Download#Tutorial_data)

Opening files in a new Project (see Figure 3-5)

1. Right-click on Workspace and select **New Project**.
2. Right-click on **Project** and select **Open File(s)**.
3. By default, the file browser should open in the geWorkbench data directory. Navigate instead to the directory to which you downloaded and extracted the tutorial data files, and find the directory “cardiogenomics.med.harvard.edu”. Select **File of Type** to be **Affymetrix MAS5/GCOS**.
4. In the figure below, a set of MAS5 format files are selected.
5. Check the “Merge Files” checkbox so that all files will be merged into a single dataset in geWorkbench.

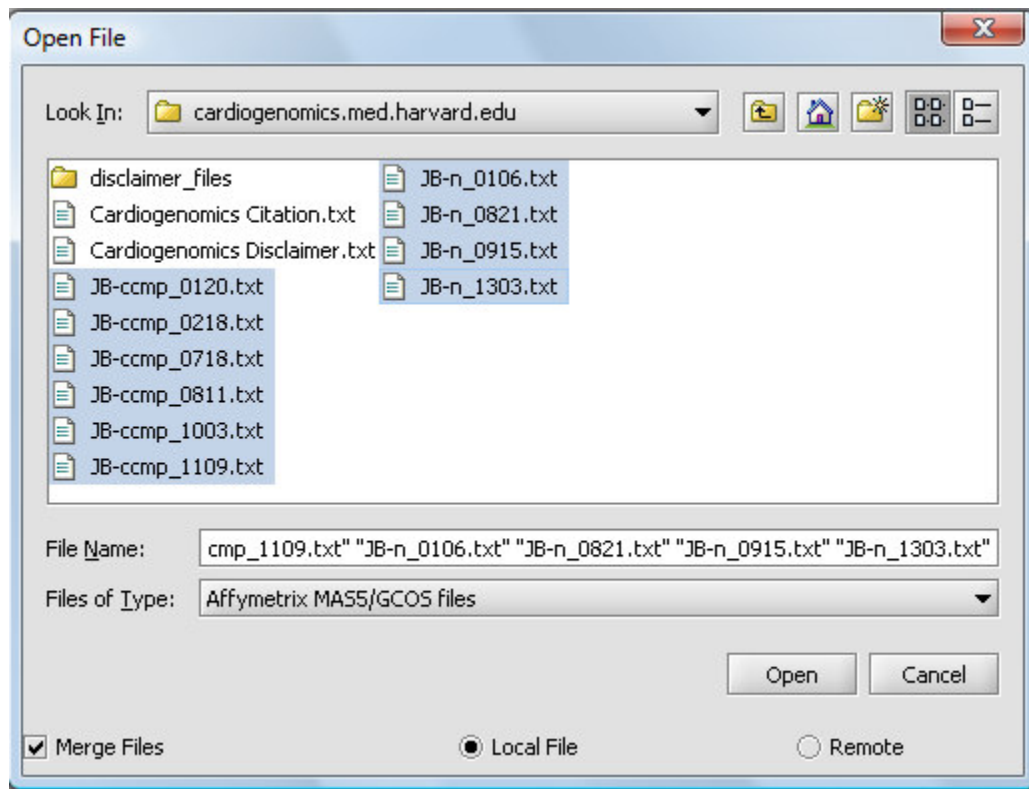


Figure 3-5 Opening a file in a project

6. A box with information about annotation files will appear. Click **Continue**.
7. Annotation files are not distributed with geWorkbench. Instructions for obtaining current Affymetrix annotation files are available in the FAQ section of the geWorkbench.org website. If you have obtained an annotation file, navigate to the directory containing it (Figure 3-6). Select

the desired annotation file and press the **Open** button. If you do not have the file, just press **Cancel** and proceed without the annotation file.

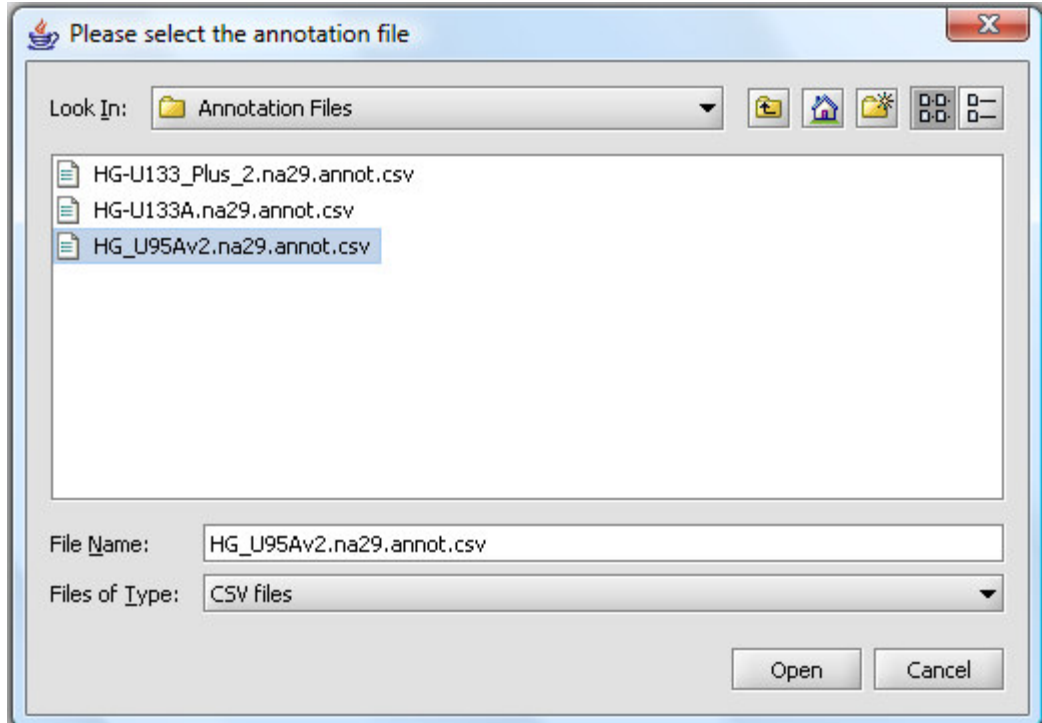


Figure 3-6 Opening the annotation file

The opened data file is now shown within the **Project Folders** area at upper left in the GUI. Components relevant to acting on microarray data will now appear in the interface Figure 3-7.

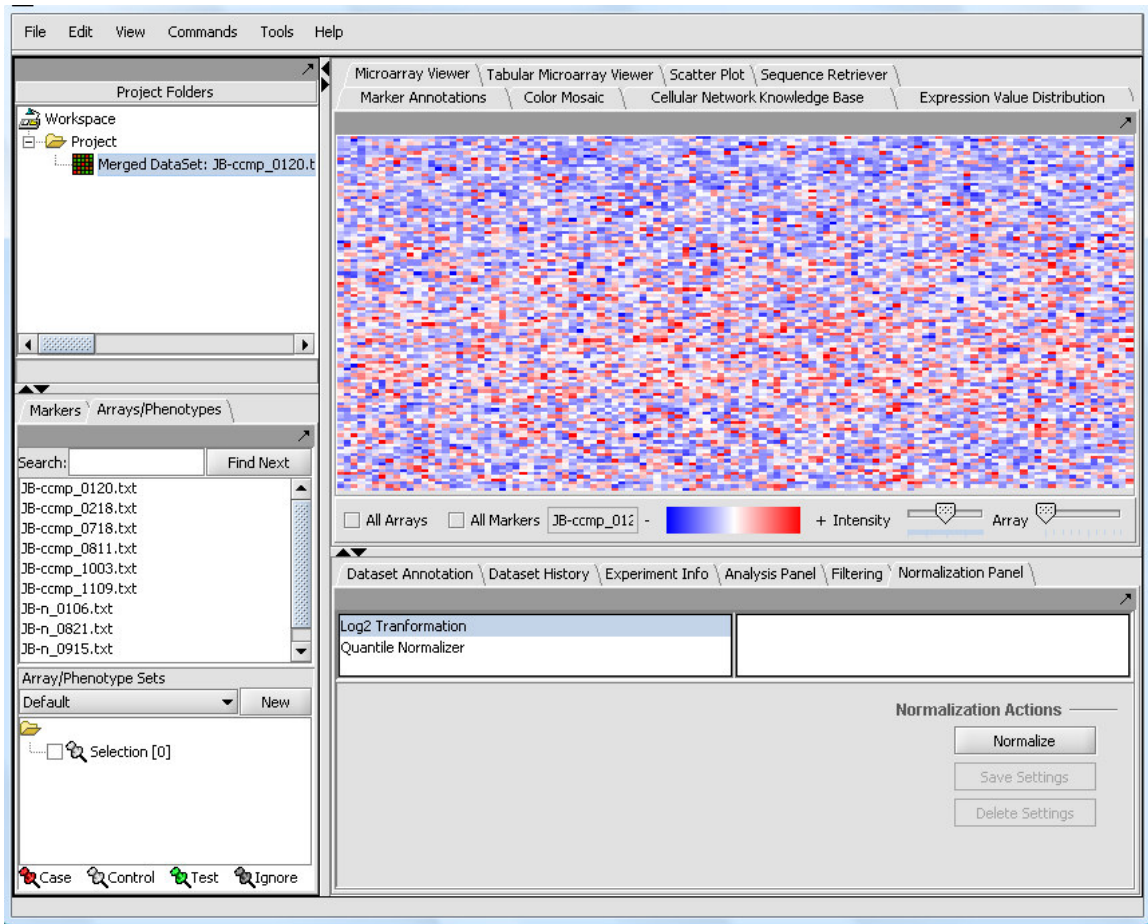


Figure 3-7 geWorkbench GUI showing all microarray-related modules

3.4.2 Other file operations – Merging and renaming

The next two options in the **File** pull-down menu are **Export** and **Merge Datasets**. The export option can be used to save the data in a data file or a selected image (see Section 3.1.1) in a new format. The **Merge Datasets** operation combines two or more microarray data sets generated using the same platform to produce a single set (Chapter 3). *Only those data sets included in the same project and being of the same array type can be merged.*

Figure 3-8 shows the expanded drop-down menus associated with the next two file operation. The **File→Remove** option is used to remove images and files from a project, and projects and marker panels from a workspace. In all cases, the user must first select the object to be removed in the Project window before executing the operation.

It is possible to rename projects and data files from within the Project Folder component by right-clicking on the desired element and using the shortcut pop-up menu. The drop-down **Edit→Rename** menu in the main menu bar also provides this option.

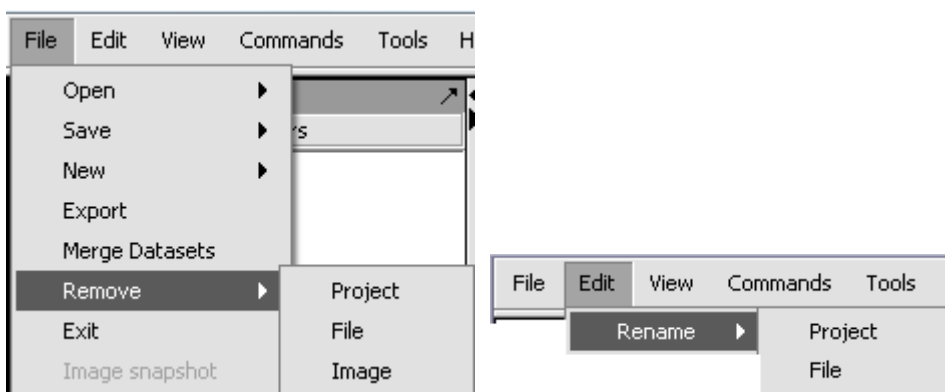


Figure 3-8 Removing and Renaming

The **File→Remove** option closes geWorkbench.

The last option in the **File** drop-down menu is the **Image snapshot** operation, which is used to take snapshots in the View window and is described in Section 7.2.5 . For convenience, all of the file operations are summarized in Table 3-2.

Table 3-1

Command	Arguments	Action
File→Open	files, workspaces,	Opens a new workspace, or a file in the current workspace. Clicking the “Merge Files” checkbox will combine the selected microarray files into a single dataset.
File→Save	workspaces	Saves the workspace
File→New	workspaces, projects,	Creates a new workspace, or a new project in the current workspace.
File→Export	images, files	Saves an image or data set in a new format.
File→Merge Datasets	files	Merges two or more microarray data sets to generate a single combined set.
File→Remove	projects, files, images	Removes files and/or images from a project, or projects from a workspace.
File→Exit		Exit geWorkbench. A save workspace dialog is shown first.
File→Image snapshot	objects in the View window	Takes a snapshot of an object in the View window.

Table 3-2 Summary of File Operations in the Main Menu bar

4 Querying caARRAY

This chapter describes how geWorkbench can query remote instances of a caARRAY database. caARRAY currently uses a Java API which allows searching on a number of common annotation data fields, such as species, array type and tissue type.

4.1 Searching caARRAY using MAGE annotations

caARRAY is a microarray gene expression repository developed by the NCICB which supports storing and querying of annotated datasets. The annotations are consistent with those defined in the MAGE (Microarray Gene Expression) model. It should be noted that actual datasets may be only partially or sparsely annotated.

In the current implementation, geWorkbench can query against four types of annotations supported by caARRAY:

- Tissue type
- Chip Platform (e.g. Affymetrix, Agilent etc.)
- Organism
- Principal Investigator

The following example illustrates constructing a query against caARRAY.

1. Create a new project.
2. Right-click on the **Project** entry and select **Open File(s)**
3. Click on the “**Remote**” radio button (Figure 4-1). This will cause the Open File popup to switch to the remote file interface

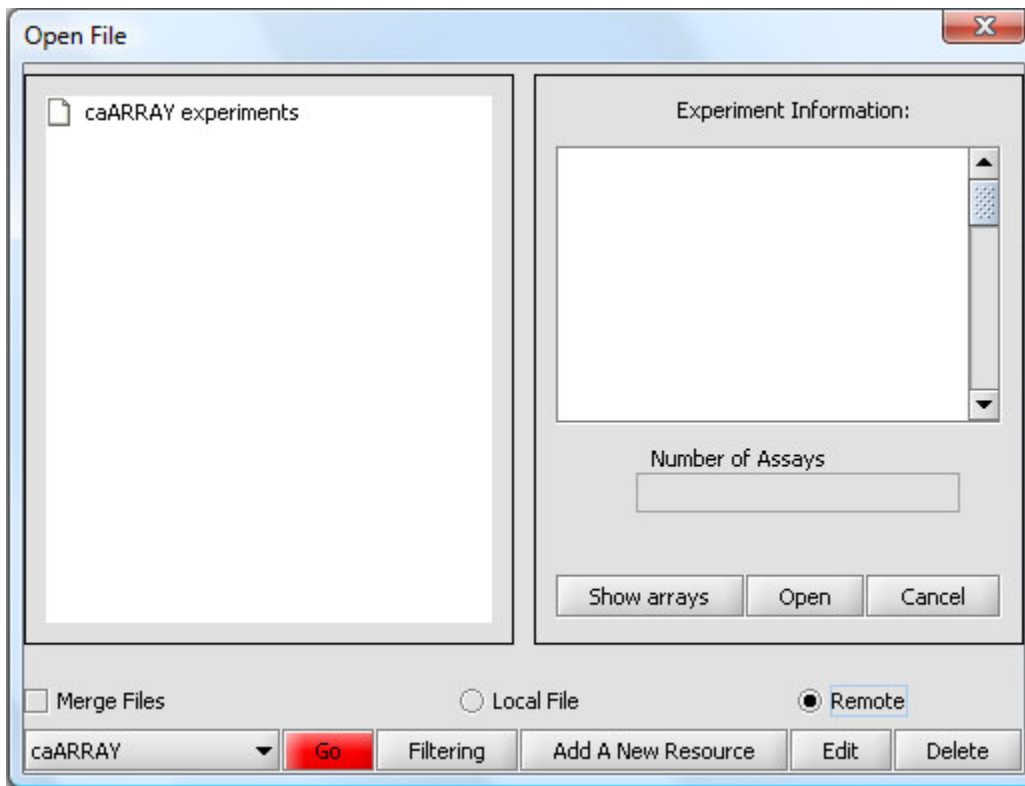


Figure 4-1 The Open File interface

4. You can add a new caArray resource (**Add A New Resource** button), or edit an existing one (**Edit** button), using the respective buttons. Either results in a properties editor window appearing as shown in Figure 4-2. Here we show the entry for the NCI public instance of caArray:



Figure 4-2 Adding/Editing a caARRAY Service entry

Once a source has been chosen, clicking on the red **Go** button will retrieve all available experiments. If instead you wish to query for just specific types of experiments, you can use the **Filter** button instead to construct a query.

5. Click on **Filter** (see above in Figure 4-1) to build a keyword search. The query builder appears (Figure 4-3).

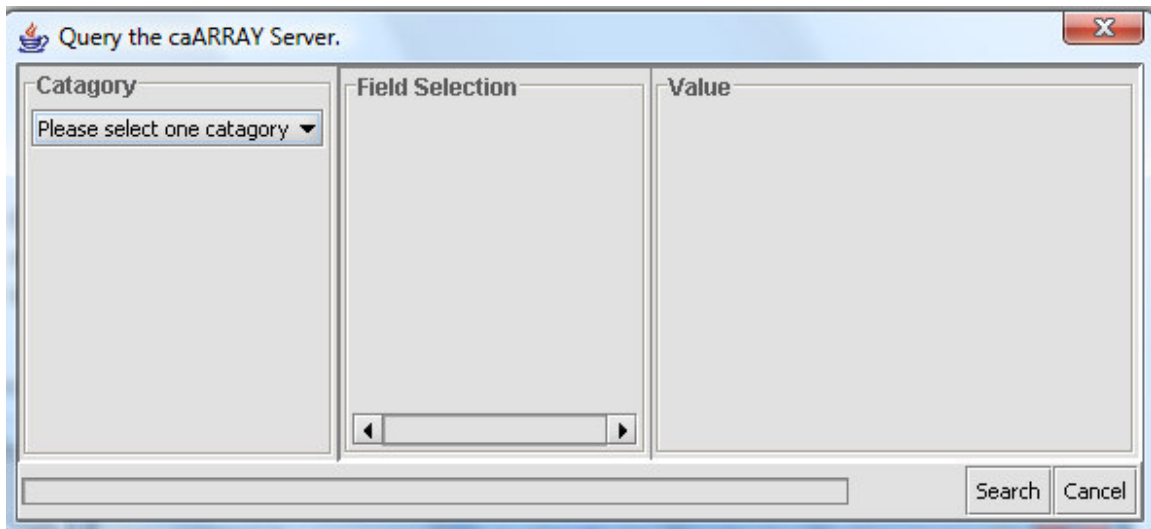


Figure 4-3 The caARRAY query interface

6. Under Category, select **Experiments** (Figure 4-4). The available search field types will be displayed. Here we will search on **Organism**. Highlighting this field shows available organism types for all of the experiments loaded into the database. Here we have selected human.

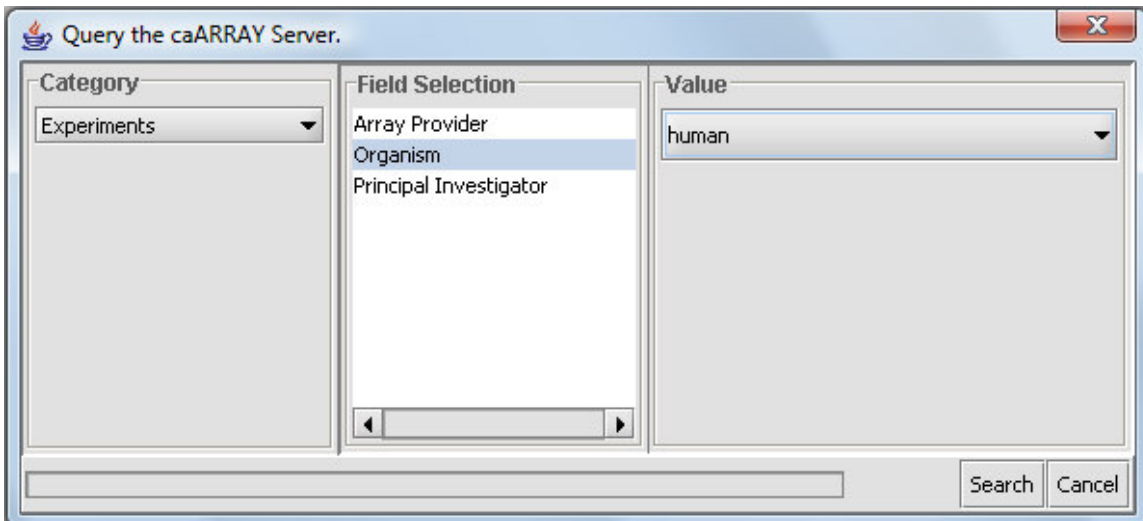


Figure 4-4 Constructing a new caARRAY query

7. Click **Search**. (Figure 4-4). A progress bar may appear. (Figure 4-5).

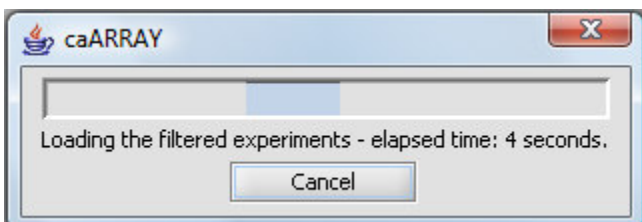


Figure 4-5 Progress Bar

8. Experiments matching the search term are returned (Figure 4-6). Select an experiment and click **Show Detail**. This will display a list of the available bioassays associated with the experiment.

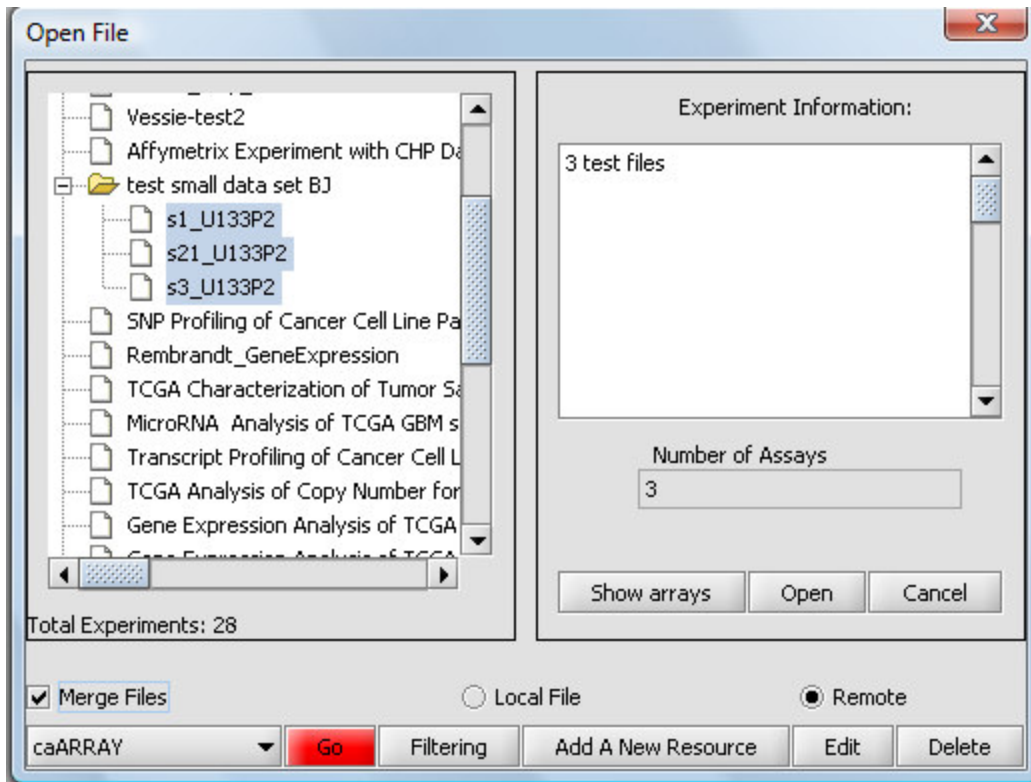


Figure 4-6 Retrieving the list of bioassays for an experiment

9. Select the desired bioassays for retrieval(Figure 4-6).
10. If you wish to merge the files into a single dataset during download, check the **Merge Files** checkbox.
11. To retrieve the selected bioassays click **Open**. A dialog box will appear asking which quantitation type to retrieve. Here we select the primary signal derived from an Affymetrix CHP type datafile (Figure 4-7).

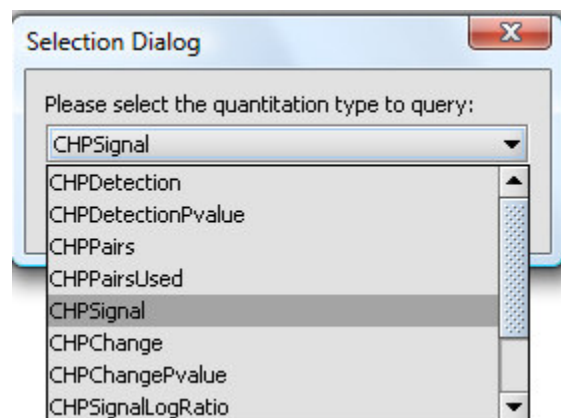


Figure 4-7 Selecting the quantitation type

12. The returned bioassays are shown in the Project Folders component (Figure 4-8).

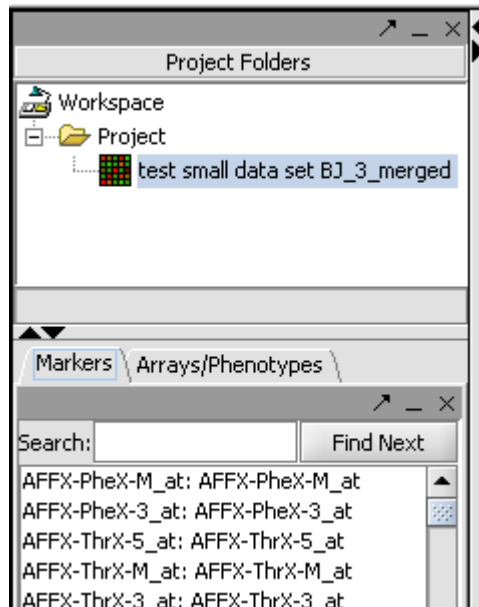


Figure 4-8 The merged dataset retrieved from caARRAY as displayed in the Project Folder

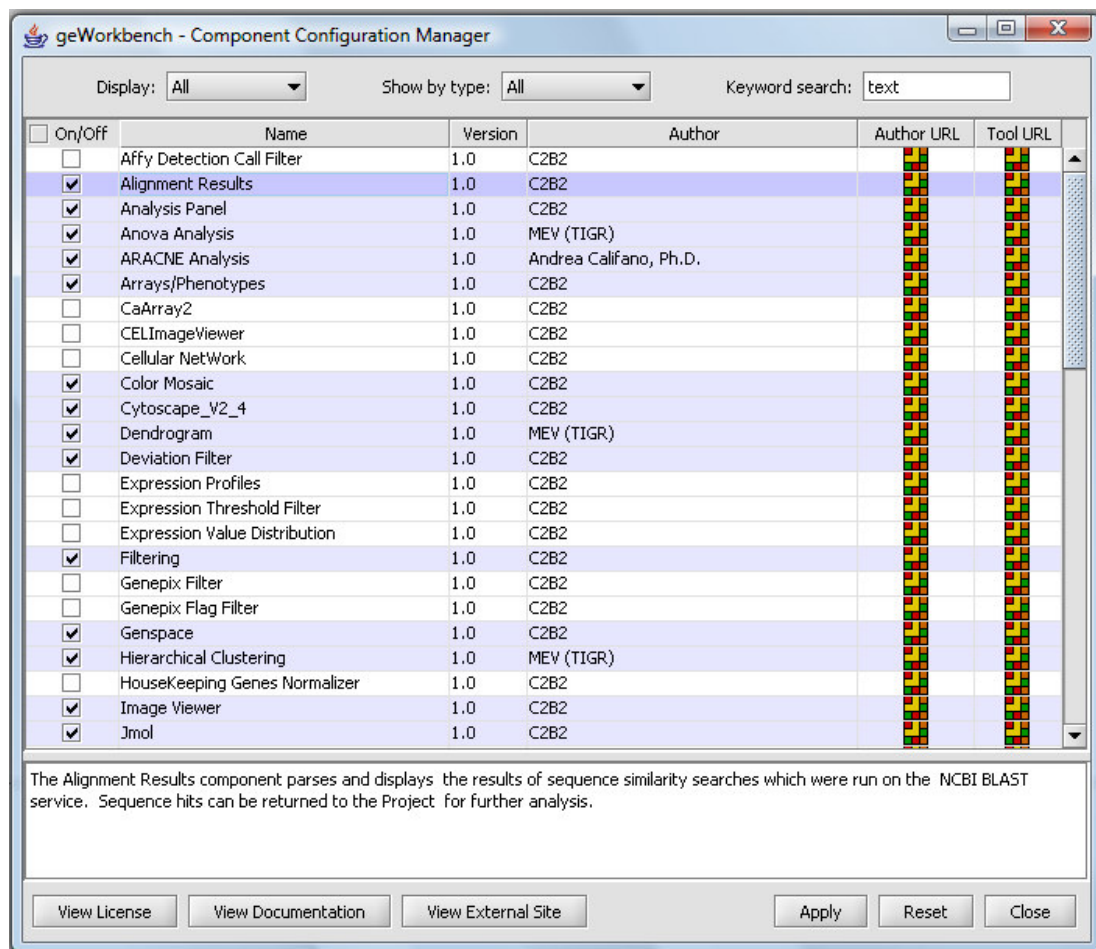
13. The merged dataset can be renamed if desired by right-clicking on it and selecting **Rename**.
14. Did you forget to check the **Merge** checkbox before download? You can merge the files after download by selecting menu item **File -> Merge Datasets**.

5 Component Configuration Manager

5.1 Overview

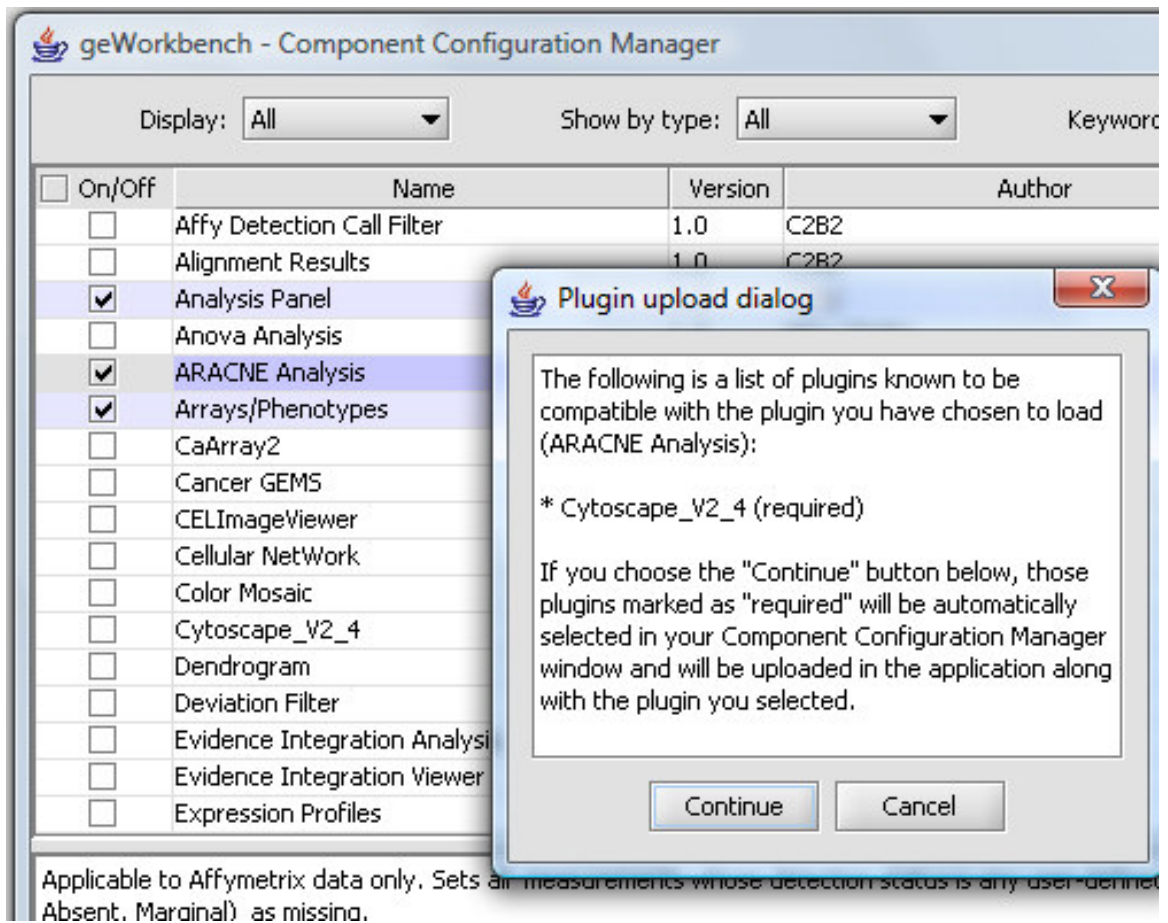
geWorkbench supports the dynamic loading and unloading of individual components for analysis and visualization. The Component Configuration Manager (CCM) lists each available component, provides descriptive and license information, and controls loading and unloading. The ability to customize the loaded components may be useful to those who intend only to use particular types of analysis. It may in some cases also increase the responsiveness of the application, as potentially fewer viewers may need to respond to particular data events (new data set, marker set activation etc.).

As shipped, the CCM may appear as in the figure below, with only basic components configured. Individual components can be added by clicking on their check boxes.



As each component is added, additional components may also be required, e.g. adding an analysis may require a particular viewer. These dependencies are all handled in the CCM.

An example is shown in the following figure, where adding the ARACNE component also requires adding the Cytoscape component.



If a component requires agreeing to a license, the license will also be displayed at this time, with the option to accept or reject it.

5.2 Individual controls

- **Display** : All, Only Loaded, Only Unloaded - selects which components to display.
- **Show by type**: All, Parsers, Analysis plugins, Visualizers - filters displayed components by their category of action.
- **Name**: Component name
- **Version**: Component version
- **Author**: Component author or source of primary calculation code.
- **Author URL**: some components have a URL to the author's website configured.
- **Tool URL**: some components have a URL to their own website configured.

- **View license** - If a license is required for a component, it can be displayed by highlighting the component in the list and pressing this button.
- **View documentation** - Provided for displaying extra documentation about the component if available.
- **View external site** - same as the Tool URL link.

- **Apply** - if any changes have been made to the check-boxes indicating which components are to be loaded, pushing "Apply" causes the changes to actually be made.
- **Reset** - If changes have been made but "Apply" has not been pushed, set all checkboxes back to their previous state.
- **Close** - Close the CCM window.

On/Off – removed in geWorkbench 1.8.0.

6 Analysis Component Framework

6.1 Overview

Most analysis routines are located in the command area in the lower right quadrant of geWorkbench. There they share a common framework and a common method for saving parameter settings.

6.2 Layout of the analysis framework

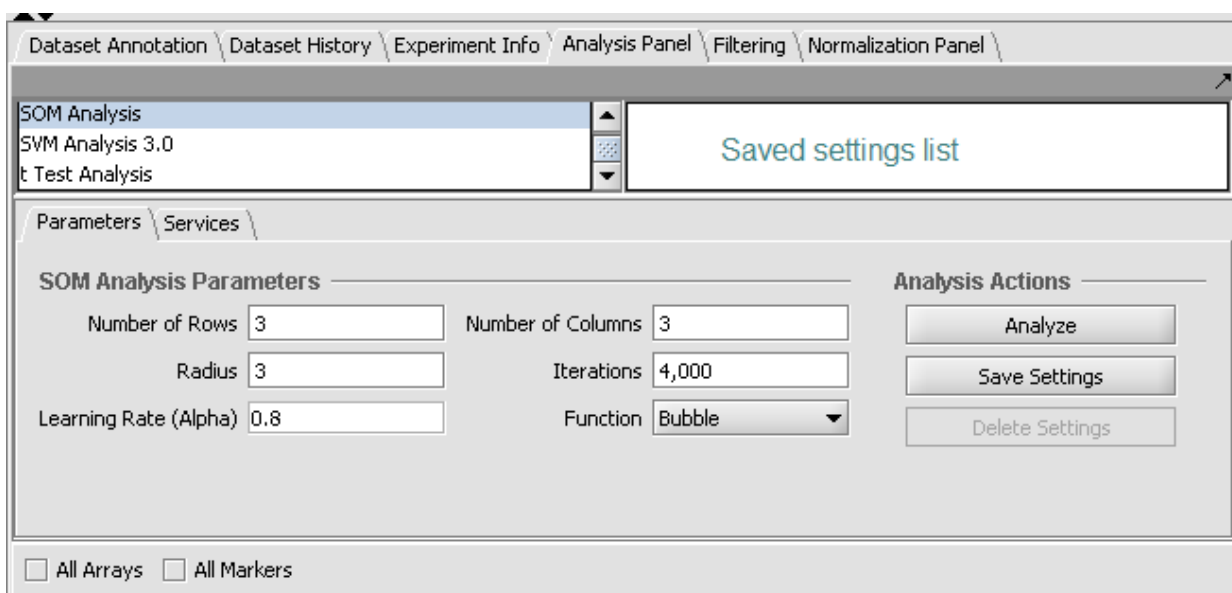


Figure 6-1 - The analysis framework (using as an example SOM)

6.2.1 Lists

At upper left of the framework is a list of available analysis routines (as loaded using the [Component Configuration Manager](#)). At upper right is a list (labeled "Saved settings list") which can store named sets of customized parameter settings for each different analysis.

6.2.2 Analysis Actions

At right are three buttons shared by all analysis components:

6.2.2.i Analyze

Launches the currently selected analysis with the specified parameters.

6.2.2.ii Save Settings

Save the current parameter settings to a named set in the settings list.

6.2.2.iii Delete Settings

Delete the currently highlighted settings entry from the list.

6.2.3 Analysis Parameters

The parameters and settings for the currently selected analysis component are located in the lower-left portion of the framework.

6.3 Creating saved parameter sets

The current parameter settings can be saved by pushing the **Save Settings** button. A dialog will appear asking for a name for the new set of saved parameters.

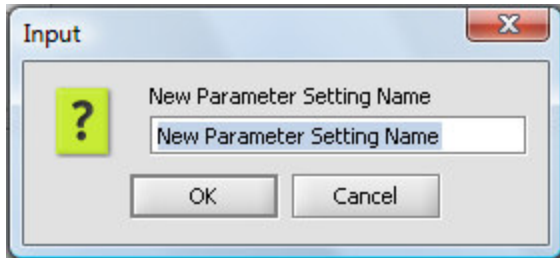


Figure 6-2 - Naming a new parameter set.

In this figure, the default settings have been saved to a set called "Default Settings".

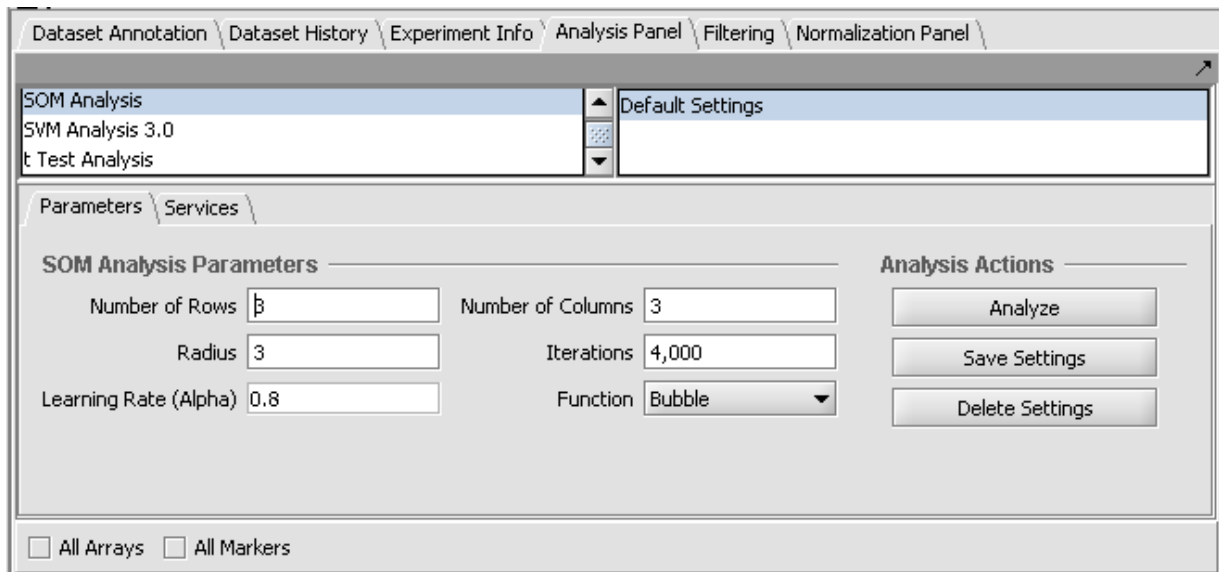


Figure 6-3 - A new parameter set.

The next figure illustrates changing a couple of parameters and saving them to a new set called "New Settings".

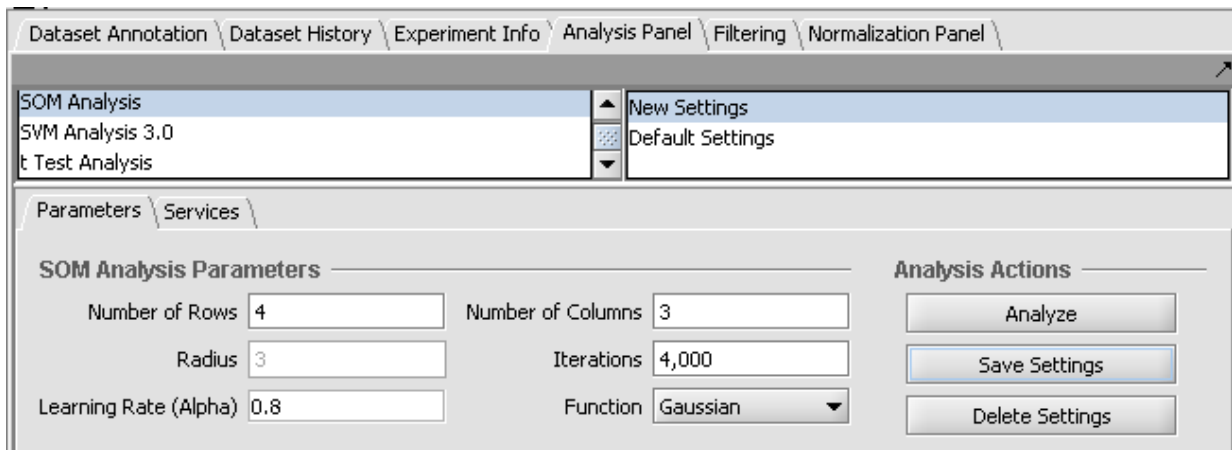


Figure 6-4 - A second parameter set added.

6.4 Interplay of parameters and list

The analysis parameters can be restored to their saved values by selecting the desired item on the list.

For most analysis components, when the parameters shown in the graphical interface are made to exactly match those stored in a list entry, that entry will be selected (highlighted). If any change is made so that the parameters no longer match a stored entry, then no entry will be highlighted.

7 Microarray Data Analysis

geWorkbench provides a comprehensive and extensible suite of open-source desktop software tools that can be applied to the analysis, visualization, and annotation of microarray data. In addition to the analysis and visualization tools routinely found in microarray software tools today, geWorkbench provides an enhanced environment via its integration with the Cancer Bioinformatics Infrastructure Objects (caBIO)¹. This integration provides geWorkbench users with access to publicly available microarray data on a remote NCI server; to the CGAP web site's gene annotation pages, and to the pathway visualization diagrams generated by BioCarta. This last capability allows users to view the observed microarray data in the context of metabolic and signal transduction pathways.

The workbench is intended to support a variety of the input formats in which microarray data are found; its open-ended design supports the extension of the software to accept additional formats as needed. The present version of geWorkbench supports Affymetrix (.txt, MAS 4.0/5.0), Expression Matrix (.exp) and GenePix (.gpr) files. A simple plug-in framework allows users to further define and use any input format they wish. Similarly, this plug-in framework supports the addition of any number of user-defined filters, normalizers, and analysis algorithms.

This chapter provides an overview of a rather complex software suite, and assumes that the user has some experience with microarray data analysis. The discussion which follows outlines procedures for loading data files, for using visualizations, and for annotating data.

7.1 Set Selection (the Markers/Arrays/Phenotypes) components

7.1.1 Marker Sets

The term *marker* is used generically to represent several different things in geWorkbench. When working with microarrays, the term marker refers to a gene probe (in other cases, it can be individual items from other data sets, such as sequences). The definition of what constitutes a gene probe in turn depends on the type of microarray platform. On Affymetrix platforms, gene probes are oligonucleotides synthesized on the microarray chip *in situ*. On other platforms (e.g. GenePix), gene probes are oligonucleotides or cloned DNA fragments deposited and immobilized on the substrate by various techniques.

As soon as a specific microarray is selected in a project folder, the entire complement of markers on that array is displayed in the *Selection* area under the **Markers** tab. For example, in

¹ The NCICB Technical Guide provides a detailed description of the caBIO project and its application programming interface (API).

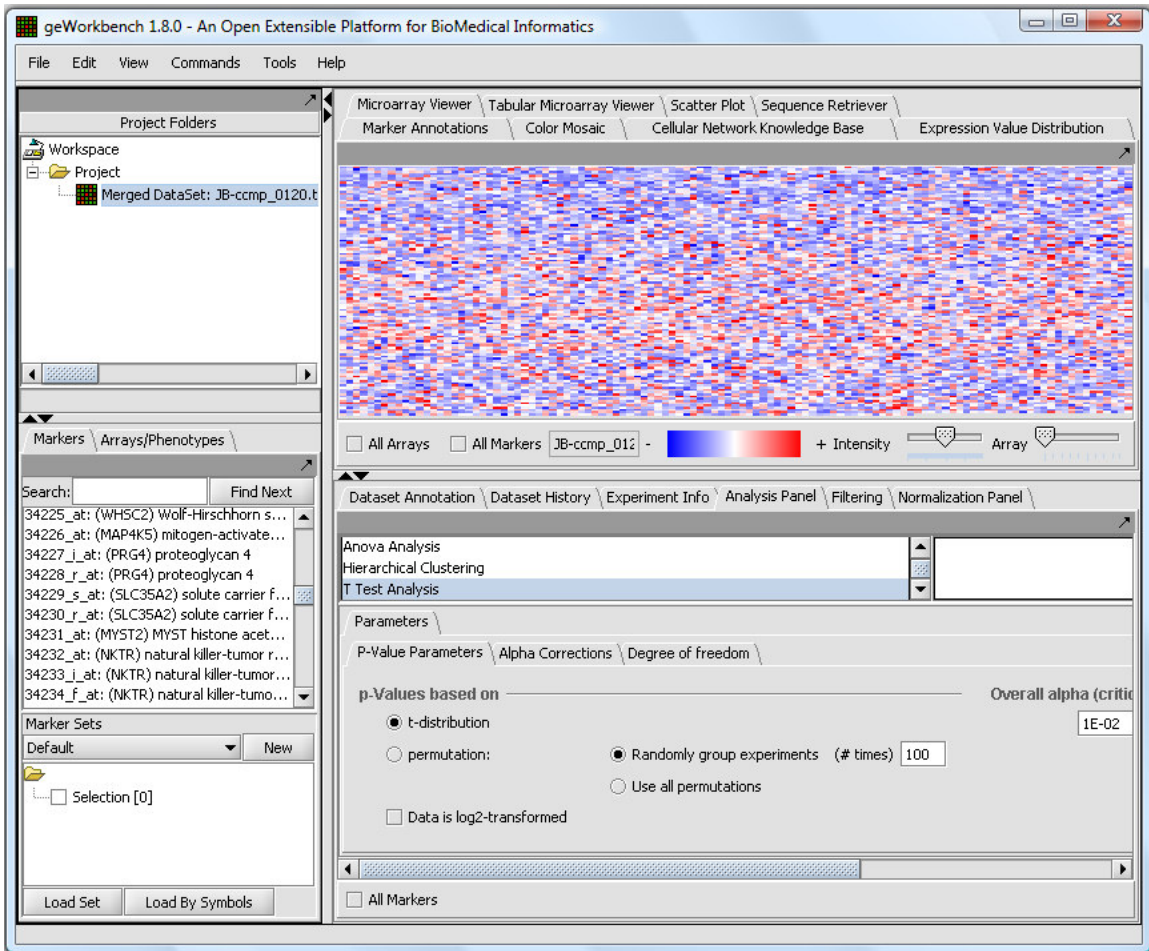


Figure 7-1, the user has just created a new project and loaded a set of files which were merged into a single dataset. .

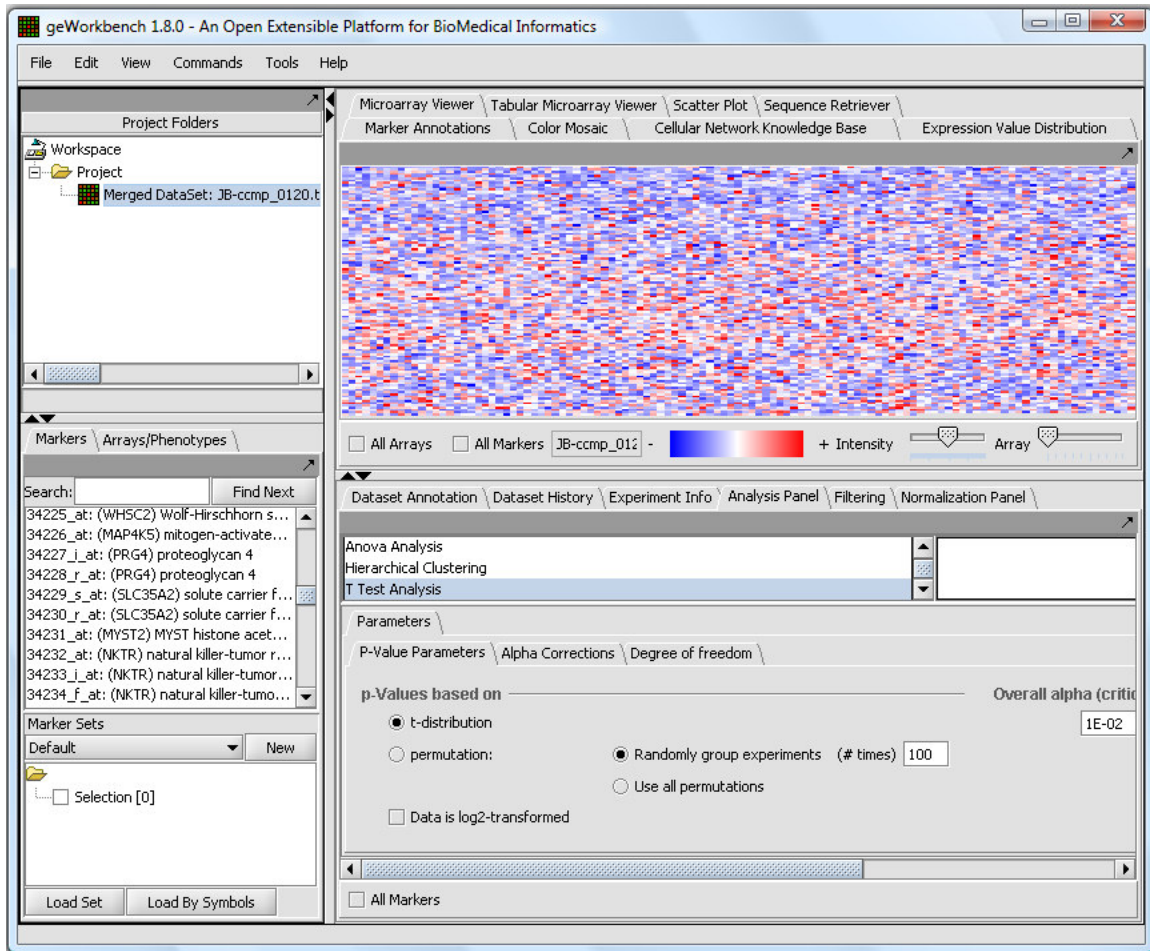


Figure 7-1 geWorkbench Display Immediately After Loading a Data File

A *marker set* is a user-defined grouping of several markers, e.g. gene probes. Typically, these sets probe for genes of specific interest, of importance for certain disease or developmental states, or for characteristic changes in gene expression that may be hallmarks of a tumor in a particular tissue.

A master list of all gene probes in the currently selected data set is displayed under the **Markers** tab, and a color-coded image of the corresponding gene expression measurements is shown in the View window's **Microarray Viewer**. Individual or groups of gene probes can be selected from this master list and added to smaller marker sets for use in a specific study. The sets are managed in the smaller window immediately below the master list. A marker set can be saved by right-clicking on it and selecting **Save**. Clicking on the **Load Set** button loads saved sets.

Any number of sets can be created, and they can be grouped as desired. A new group can be created using the **New** button in the Sets area below the master list. A new set can be created manually by selecting markers in the master list and right-clicking, then selecting **Add to Set** from the pop-up menu. A prompt will appear, for the set name. More

commonly, new sets of markers will be returned from an analysis step, for example from hierarchical clustering or a t-test. Figure 7-2 shows a set called “cluster tree_84_markers” containing a set of 84 gene probes., and a second set containing 12 probes.,

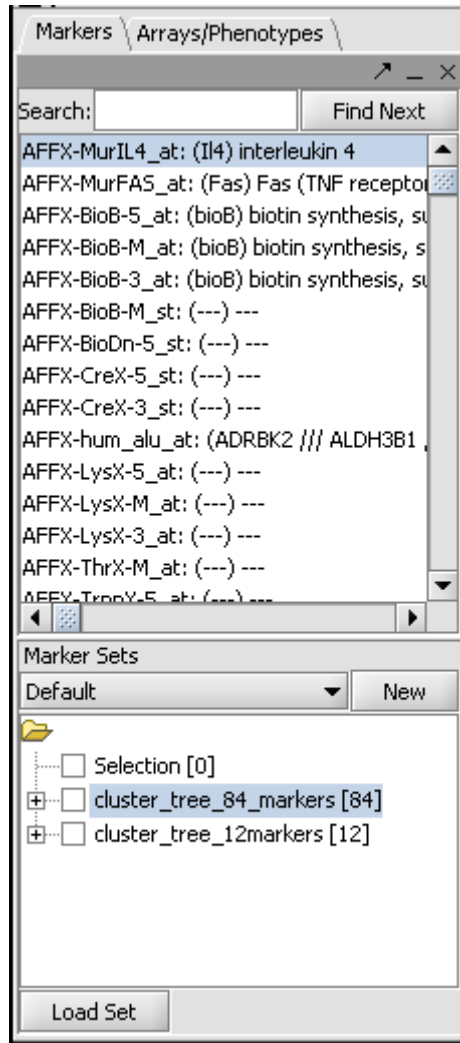


Figure 7-2 - Two different marker sets in the default group.

7.1.2 Set Activation and Manipulation

A gene set, once created, can be *activated*. This notifies other components that this subset of data is available for analysis or visualization as a distinct entity. Any number of sets can be in the activated state at one time. A set is activated simply by checking the box before its name, or from a right-click menu or from the top menu bar Commands

As described in the next section, most visualization tools, including in the **Microarray Viewer** provide **All Arrays and All Markers** checkboxes. By default neither is checked. In this state, if no sets are activated, then all Markers/Arrays are considered active

implicitly. If any set has been activated, then only that set will be used, unless the respective “All” button is checked..

A set’s **Activate** option can be used to activate all of the markers or arrays in that set. Similarly, the **Deactivate** option will deactivate all markers or arrays in the set.

As mentioned in the previous section, it is also possible to save and load panel sets independent of the workspace where they were created. While it does not make sense to load a panel set generated from a data set that is not currently loaded, this facility can be useful when several saved workspaces share common data. The **Load Set** option allows users to load panel sets defined outside the current workspace. The user has the option of assigning meaningful names to the sets, using the **Rename** operation.

Finally, individual marker or array sets can also be explicitly renamed, activated, deactivated and deleted from the group. All of the elements listed in the *Marker/Phenotype* windows have pop-up menus associated with them.

7.1.3 Array/Phenotype Sets

geWorkbench uses the term *phenotype* to refer to any user-defined grouping of microarrays. These microarrays will often share some common property that in most cases is phenotypic, although this is not a requirement. For example, one such “phenotype” might represent a disease state such as tumor tissue samples, with a second “phenotype” defined as a collection of experiments performed on normal tissue samples.

Like the **Markers** component, the **Arrays/Phenotypes** component has two portions: the top portion lists the arrays included in the selected data set, and the bottom portion (titled “Array/Phenotype Sets”) lists any user-defined array groupings and sets.

For data sets involving a single array, only that array is present in the top portion. But in the case of multiarray data sets, the display becomes more interesting: each experiment that was included in the set is displayed as a potentially separate phenotype.

Analogous to the procedures for selecting markers or gene probes into gene panels, arrays are selected and grouped—according to the user’s preferences—into phenotype sets. Each array in the top portion of the phenotype window has an associated pop-up menu with options **Add To Set** and **Clear Selection**.

7.1.4 The Commands Menu

Many of the commands for manipulating marker and phenotype sets are also available from the **Commands** menu option in the main menu bar, as shown in Figure 7-3.

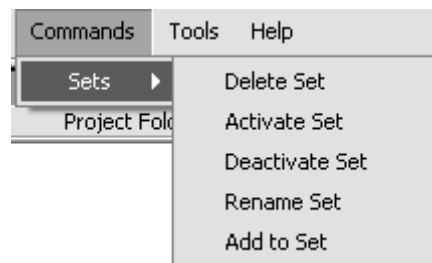


Figure 7-3 The Commands Menu

7.2 The View Window

The View window is in a sense the main work area as it provides all of the visualization tools in geWorkbench. Folder tabs running across the top of the screen provide access to these tools, a basic selection of which are summarized in Table 7-1 and described in more detail below.

Table 7-1 Microarray Visualization Tools in the View Window

<u>Visualization Tool</u>	<u>Description</u>
caBIO Pathways	Displays BioCarta pathway diagrams for selected genes.
Color Mosaic	A color mosaic representation of measurements, with each array in one column and each probe in one row.
Dendrogram	Displays tree-structured diagrams (dendrograms) reflecting the results of hierarchical clustering analysis.
Expression Value Distribution	Display a graph of the distribution of expression values for a set of markers from a particular hybridization.
Expression Profiles	Displays the expression of genes across several arrays/hybridizations.
Image Viewer	Displays snapshot images taken from whole screen views.
Marker Annotations	Allows users to retrieve and display CGAP annotations for genes within a marker panel.
Microarray Viewer	Displays expression measurements as spots over a red-green color spectrum (absolute scale) or a red-blue spectrum (relative scale).
Scatter Plot	Allows the plotting of a single microarray chip or marker against other chips or markers in the project. Useful for presenting a visual picture of markers that have changed under different experimental conditions.
SOM Clusters	Displays the results of self-organizing map cluster analysis.
Tabular Microarray Panel	Presents the numerical values of the expression measurements in a table format; each row represents an individual probe and the columns display the signal and background intensities and intensity ratios

It is important to note that the viewing and analysis tools displayed depend on the type of data currently selected in the Project Folders area.. In addition, the windows vary greatly in information content and small displays can sometimes prohibit a complete view of data

for the more complex windows. In order to maximize the display of information it is often useful to detach the display window, by clicking on the arrow facing up and to the right on the view panel. See the image below that shows a detached promoter window. The arrow in the top right now points to the bottom right, indicating the window is detached.

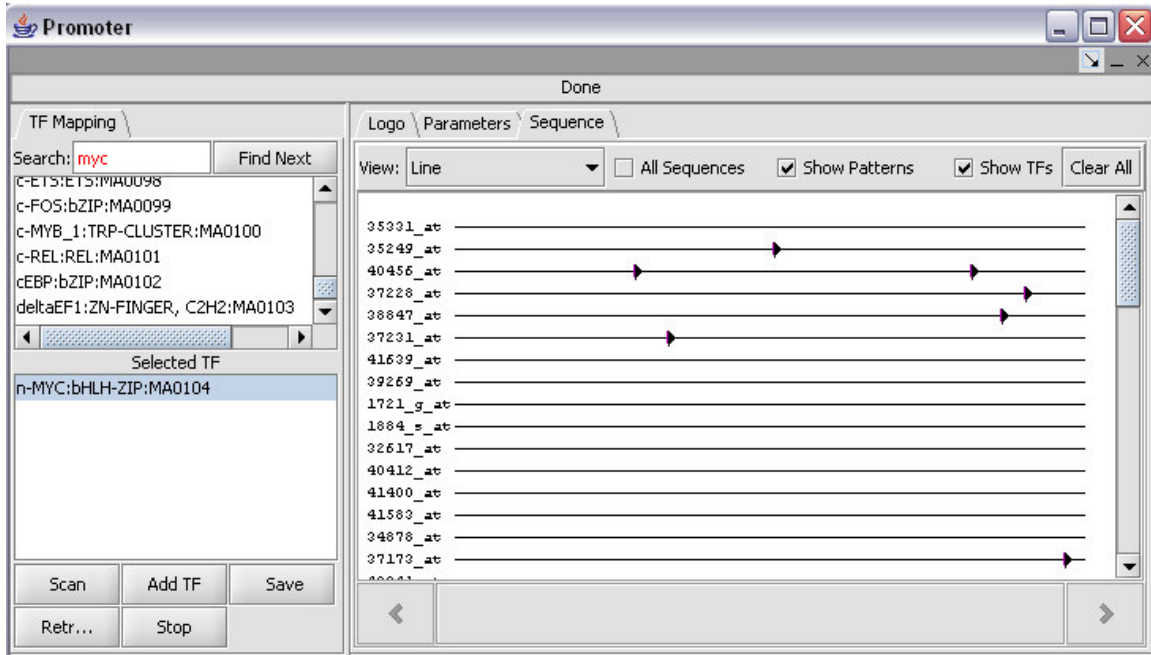


Figure 7-4 A detached window.

Some of the visualization tools are only applicable to data sets involving more than one array; others are enhanced by applying filters and/or normalizations to the data, and two of the tools are only applicable to clustered data—the Dendrogram and SOM Clusters tools. This section provides a quick tour of the general capabilities of those tools that can be applied to unclustered data.

7.2.1 The Microarray Viewer

The **Microarray Viewer** is the default visualization tool in the View window, and is displayed when the application is first started. As each new microarray data file is opened, that data set becomes the currently selected one, and the data is displayed in the **Microarray Viewer**. The image displays color-coded levels of gene expression, using the absolute color scale these vary gradually from red (positive values) through black (zero) to green (negative values). The interpretation of course depends on the specifics of the data loaded. The density of the data points in this screen is determined by the number of probes on the array. The Microarray Viewer has four controls at the bottom of the panel, shown in Figure 7-5.

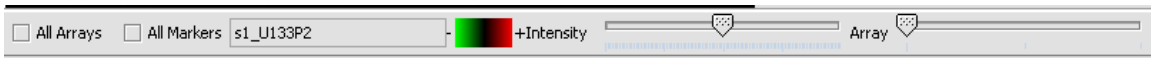


Figure 7-5 Graphical Controls in the Microarray Panel

The two checkboxes to the left, **All Arrays** and **All Markers**, determine which data points are included in the display. If neither is checked, then the entire data set is shown. The **All Arrays** control is useful when working with data sets comprising multiple arrays. In this case, only those arrays that are included in a currently *activated* phenotype set will be displayed (see previous section). Similarly, de-selecting the **All Markers** checkbox can be used to view only those probes that are currently included in an activated marker set.

The scrollbar to the right of the two checkboxes is active only when a multi-array data set is being viewed. In this case the individual microarrays are displayed from left to right, and the scroll bar can be used to jump from one microarray to the next. The entry point to each of the individual chip displays is indicated by a tick mark on the scrollbar.

Right-clicking on the panel will provide the following menu items:

1. **Show Marker**: highlights a particular marker in the display for reference purposes.
2. **Remove Marker**: unselects markers selected by **Show Marker**.
3. **Image Snapshot**: save an image under the dataset node in the **Project Tree**.

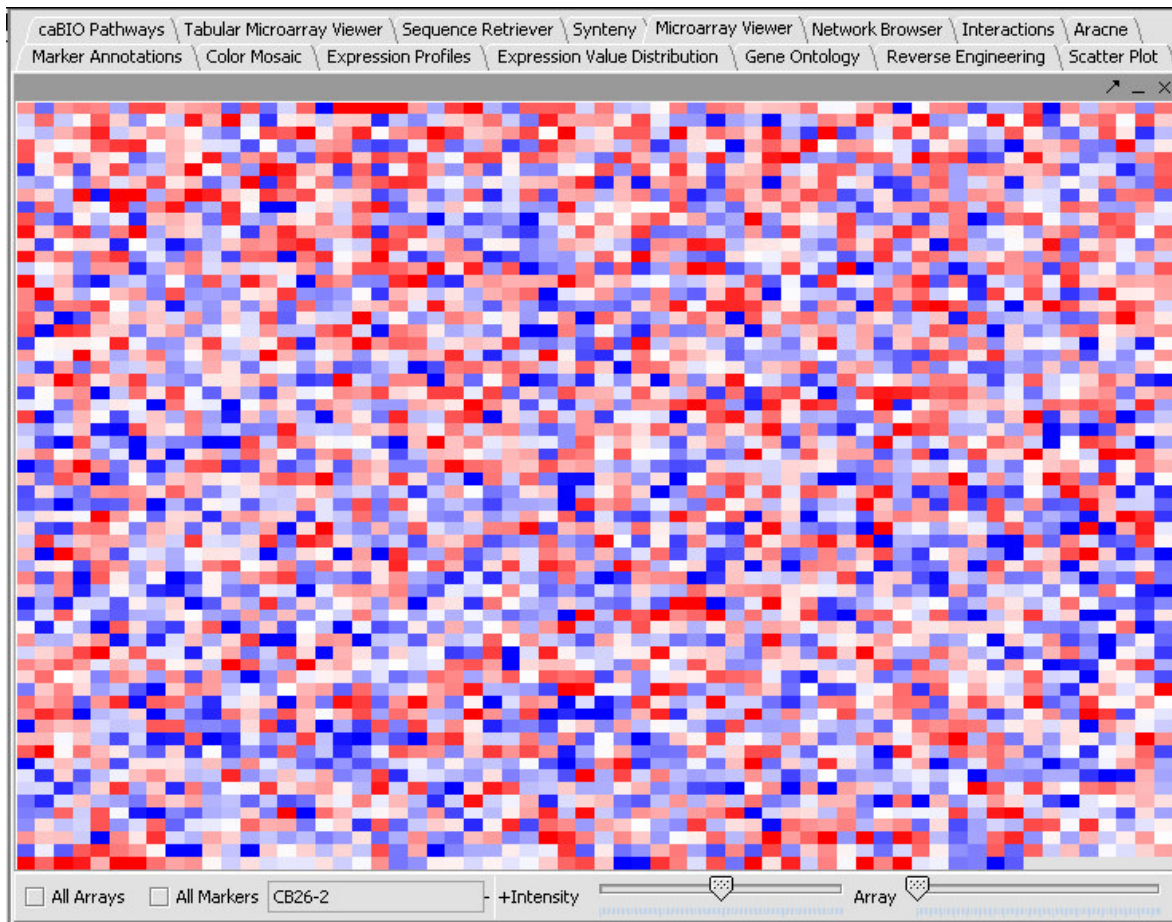


Figure 7-6 Display of a single filtered Affymetrix array with about 3,000 features in the Microarray viewer, using the relative display preference.

The **Microarray Viewer** provides an overview of the chip(s) under investigation and can be used for ascertaining the quality of the data—i.e., the uniformity of the hybridizations, the compatibility of intensities between chips, and so forth. Each feature on the chip can be accessed with a small cursor box and highlighted. Left-clicking the mouse will then highlight the corresponding probe in the master list contained in the Gene Panel window. This association between the **Microarray Viewer** and the **Markers** component facilitates the selection of individual probes for inclusion on explicit marker sets, as the user can then right click the selected probe and simply select **Add to Set**.

Only those gene sets that are currently activated will be displayed when the **All Markers** checkbox is unchecked. Any number of smaller predefined marker sets can be activated and displayed in this zoomed view.

Phenotype sets are used to create experiment groups. For instance, in a multi-array data set containing arrays from both normal and tumor tissue, these samples can be divided into appropriate sets, which can then be used for example in setting up statistical tests. Like their marker set counterparts, phenotype sets can be selectively activated and displayed using the controls described above.

7.2.2 The Expression Profiles Tool

The **Expression Profile** view makes it possible to visualize changes in the gene expression levels across different hybridizations. This is useful especially in the analysis of time course or dose response experiments. Since the tool generates a graphical representation of *relative* expression levels across two or more arrays, the **Expression Profile** view can only be applied to datasets containing multiple arrays.

After loading, , merging and normalizing the desired data sets, the user may also wish to apply marker and phenotype panels in order to zoom in on the expression behavior of a subset of genes and/or hybridizations.

Figure 7-7 shows a sample expression profile,” A set of 84 genes showing similar expression was returned following hierarchical clustering analysis and activated.

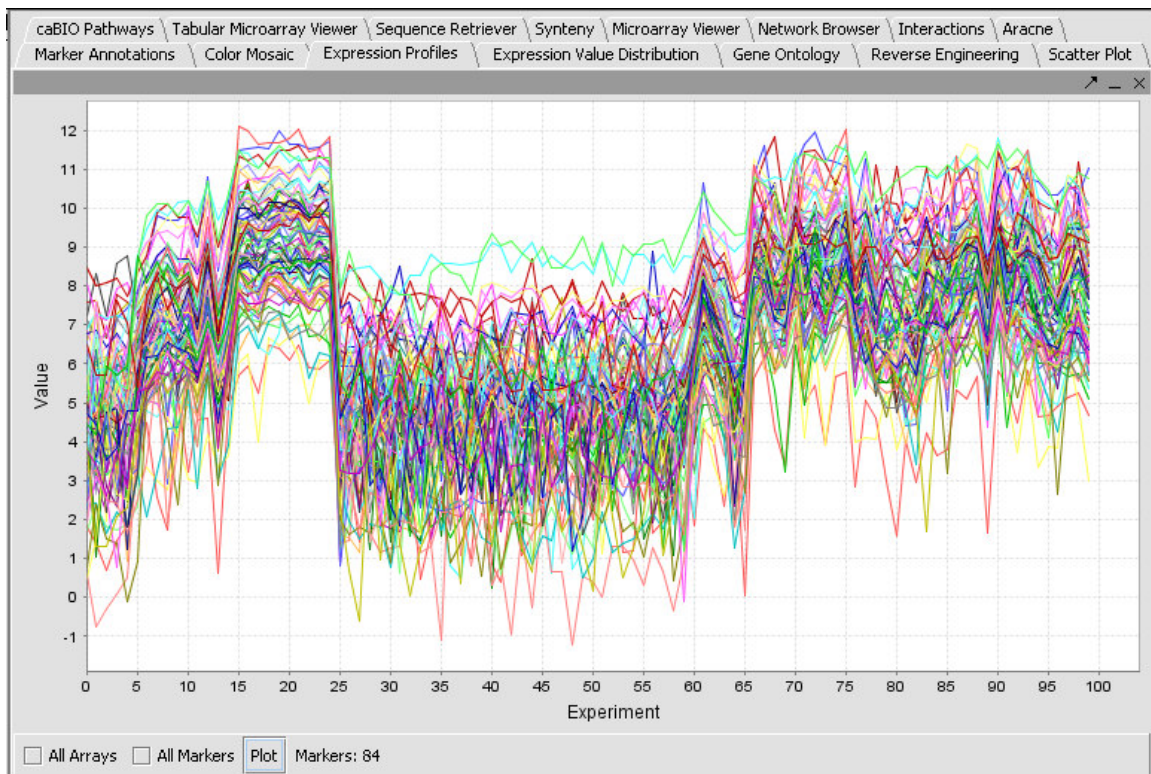


Figure 7-7 An Expression Profile Using Phenotype and Marker Panels

7.2.3 The Color Mosaic View

When a multiple array data set is viewed in the Color Mosaic view, the gene expression levels across all of the microarrays are displayed as a color coded image, where each column in the image corresponds to one of the microarrays, and each row corresponds to a particular gene probe.

Figure 7-8 shows a color mosaic generated for a single microarray data file (the mosaic is displayed only when the “Display” button is activated)..The graphical controls provided in the Color Mosaic window can be seen in

Figure 7-8 and in Figure 7-9, and include **the following checkboxes and buttons:**

Table 7-2

<u>Button/ Checkbox</u>	<u>Description</u>
Display	Display selected data.
All Arrays	Will display all arrays even if array sets are activated.
All Markers	Will display all markers (probes) even if marker sets are activated.
Abs	(not used)
Accession	Displays accession Numbers
Ratio	(not used)
Label	Displays gene names
Pat	(not used)

A mouse over tool tip controls for changing the height and width of the displayed genes, and a slider for modifying the intensity of the color codings.

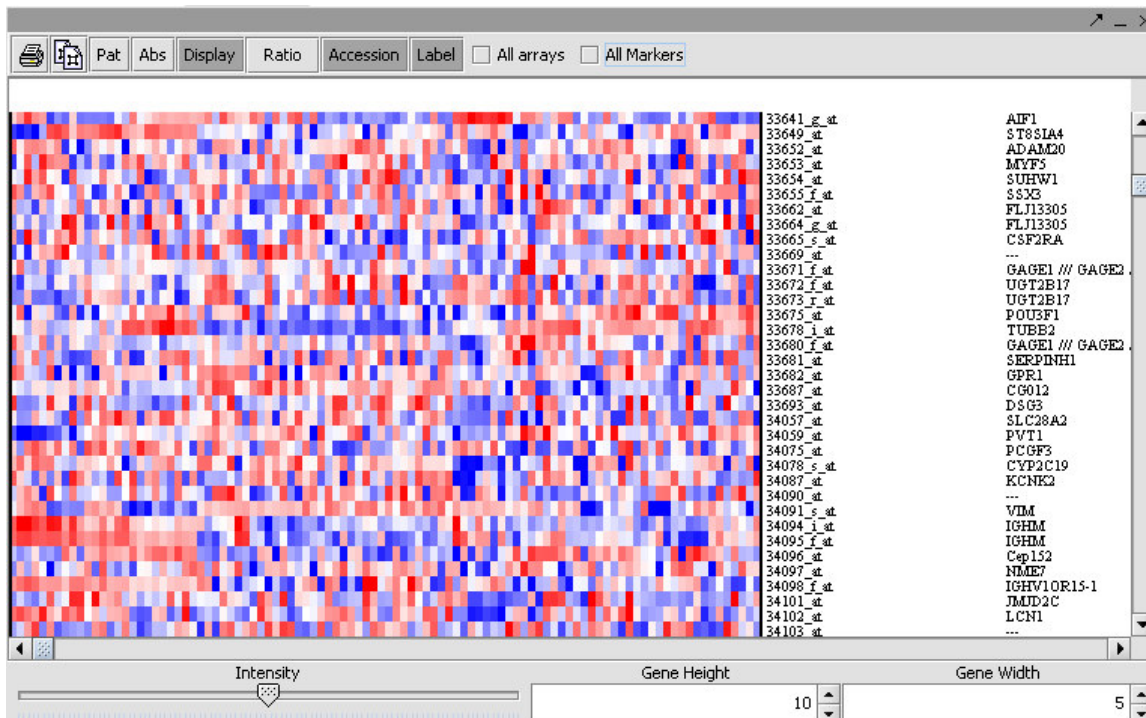


Figure 7-8 - A Color Mosaic View of a Single Data Set

The height and width controls increase or decrease the dimensions of the tiles as well as the associated labels when these are displayed. The slider increases or decreases the thresholds used to define the color codings. Clicking on a tile in the mosaic will highlight the corresponding gene probe in the Marker Panel window, where it can be picked and added to a marker panel if desired.



Figure 7-9 - Graphical controls for the Color Mosaic View

7.2.4 The Tabular View

Like the **Color Mosaic**, the **Tabular View** can be used to obtain a side-by-side comparison of the observed intensities for each gene probe over multiple chips. The display in this case however, shows the numerical values in a simple table format.

As with most of the other panels in the View window, checkboxes are provided for displaying selected phenotype and marker sets.

Figure 7-10 shows a **Tabular View** generated for the *webmatrix.exp* data set included with the distribution in the *example data* directory.

Marker	CB26-2	CB511	CB512	CB1171	CB1193	N 4-13	N 4-14	N 4-7	N1
AFFX-MurIL...	4.15	4.81	1.98	4.42	4.13	2.93	2.18	0.8	2
AFFX-MurF...	2.69	3.99	6.03	4.96	0.77	4.76	5.08	1.91	5
AFFX-BioB-...	8.4	6.64	7.28	7.94	7.32	7.97	6.67	6.29	7
AFFX-BioB-...	9.62	7.05	7.59	8.43	8.07	8.3	6.94	6.66	7
AFFX-BioB-...	8.64	5.39	4.64	7.47	6.85	7.91	6.55	5.67	5
AFFX-BioB-...	4.21	5.86	6.71	5.82	4.38	6.53	6.16	5.58	5
AFFX-BioDn...	7.5	6.6	5.01	7.98	7.1	6.5	5.48	6.35	4
AFFX-CreX-...	7.17	6.92	6.23	6.65	7.01	7.01	4.77	4.8	6
AFFX-CreX-...	6.81	7.42	6.53	7.48	8.07	7.94	6.25	6.27	3
AFFX-hum_...	12.06	10.87	10.92	8.23	8.4	12.31	10.42	11.11	10
AFFX-LysX-...	4.28	4.44	2.3	3.37	1.33	1.57	4.5	4.2	4
AFFX-LysX-...	2.47	2.93	3.4	2.92	5.05	4.62	3.61	4.53	2
AFFX-LysX-...	4.01	3.92	4.13	1.48	3.49	-0.02	0.05	3.5	2
AFFX-ThrX-...	1.62	4.4	2.4	1.77	4.22	3.67	3.67	5.35	4
AFFX-TrpnX...	3.47	5.3	5.52	3.15	3.86	5.21	4.8	4.5	3
AFFX-TrpnX...	3.55	2.05	1.78	0.09	-0.27	1.92	3.1	4.87	2
AFFX-HUMI...	4.59	7.48	7.58	6.61	7.03	9.59	8.57	7.74	8

Figure 7-10 The Tabular View

Note – values filtered out by applying one of the masks described in Section 7.3 will be highlighted in yellow.

7.2.5 The Image Viewer

Several of the visualization tools provide a means of capturing a snapshot of a selected region of the display. For example, right clicking on any point in the **Microarray Viewer** or the **Dendrogram** component will cause a pop-up text control to appear, with one of the options being **Image Snapshot**. Left-clicking on this control will create a snapshot of whatever is currently visible in the display view, and store that image under the associated data file in the Project Folders component.

Figure 7-11 shows a snapshot captured from the **Microarray Viewer**. The snapshot is stored in the **Project Folders** under the dataset from which it originated. An image can be saved in BITMAP, TIFF, JPEG, or PNG format by selecting the image in the Project window, and selecting the **Export** option from the drop-down file menu.- Note that when the image file is selected, the only viewer type available is the Image Viewer – no tabs are present.

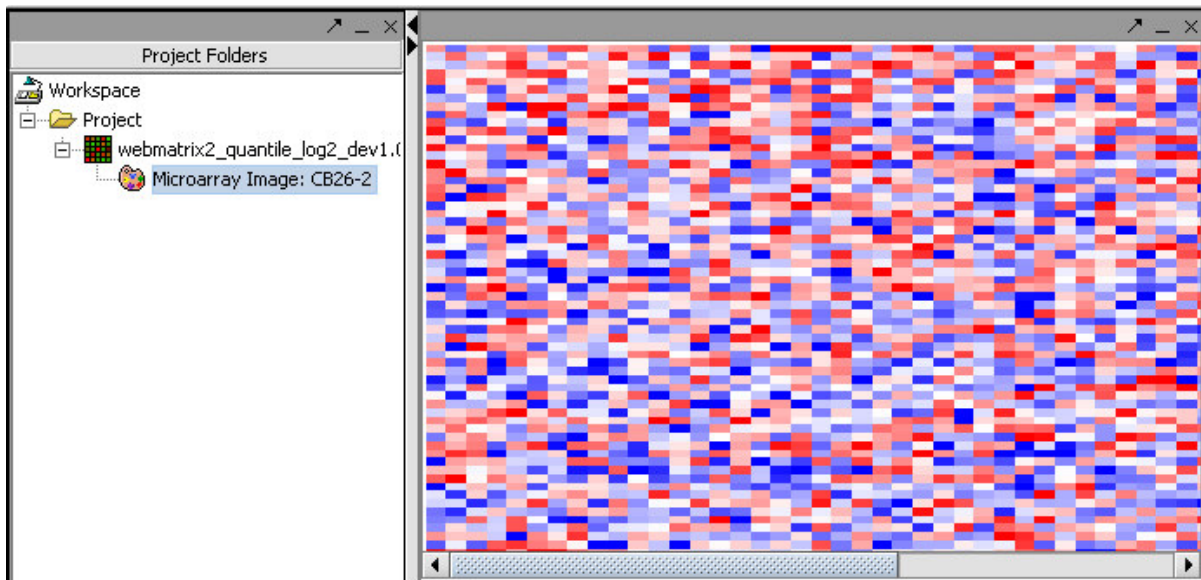


Figure 7-11 The Image Viewer

The steps used to capture an image and to view and save the stored snapshot are:

1. In the **Project** window, select the data set for which you would like to capture an image.
2. In the View window, select a visualization tool.
3. Right click on any point in the tool's display and left click on **Image Snapshot**.
4. In the **Project** window, expand the associated data set (if is not already open) by clicking on the +-like icon to the left of the data set.
5. Click on the stored image to bring it up in the **Image Viewer**.

6. Select the image in the **Project** window, and use the **Export** option from the file menu to save it in BITMAP, TIFF, JPEG, or PNG format.

7.2.6 Scatter Plot

The scatter plot feature of geWorkbench is useful for a visual comparison of up to 6 pairs of markers or arrays. The user selects an initial marker or array which is plotted on the x-axis of the Scatter plot graph and highlighted in black on the arrays/phenotype or marker selection panel. The user can then select a second marker or array which is plotted on the y-axis and highlighted in grey on the arrays/phenotype or marker selection panel. If markers are selected as the axes, then each point represents an array. If arrays are chosen for the axes, then each point represents a marker. Additional markers or arrays that are selected are also highlighted in grey, and are used as the y-axis for additional graphs. A light blue color is shown in the panel when a marker or array is deselected from the list.

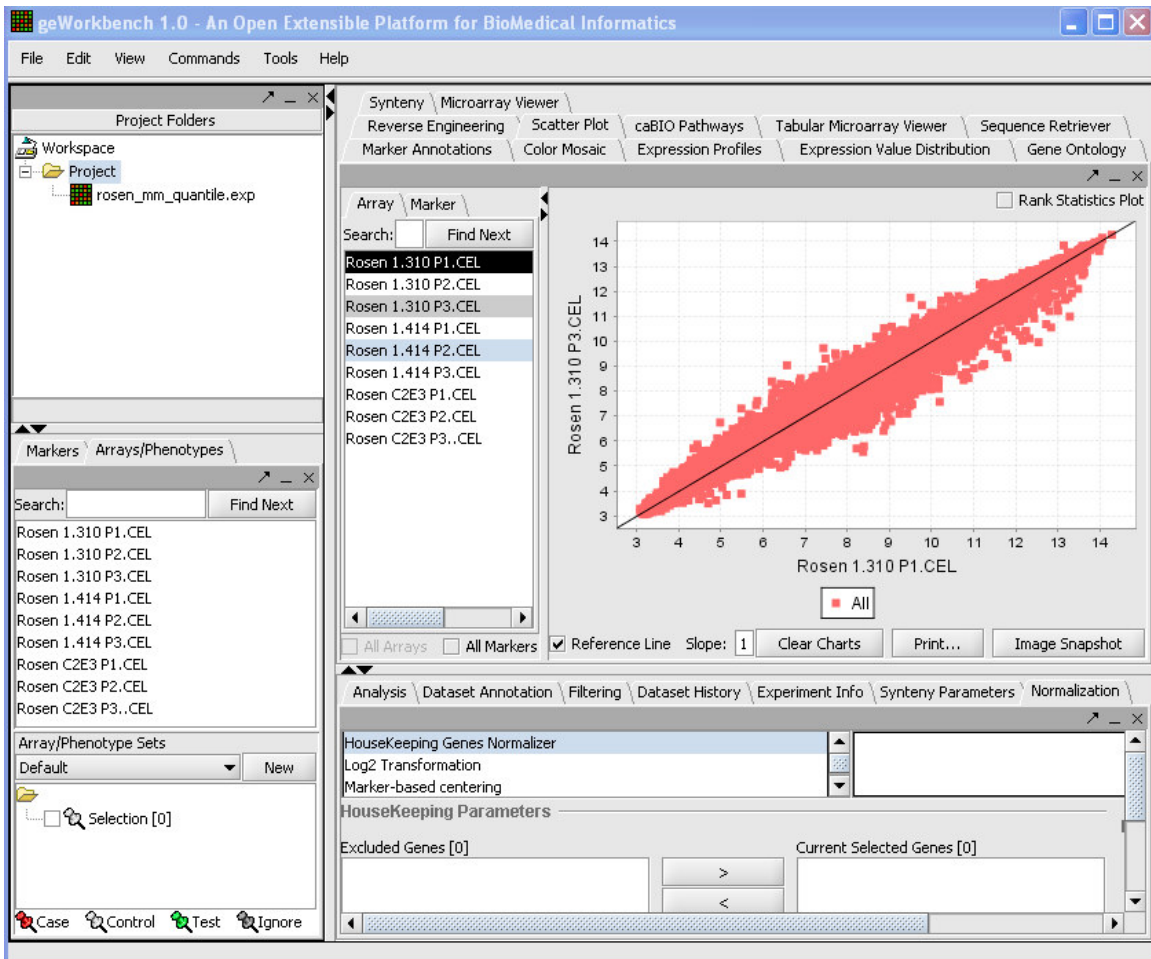


Figure 7-12 Scatter Plot

The picture above shows the ScatterPlot in geWorkbench. In general one would use either the Array or Marker tab in this component along with the search functionality in order to find the desired marker or arrays to be plotted. Then it is simply a matter of

picking the x-axis marker or array, and then selecting the desired marker or array for the y-axis to plot it against. As usual, the the “All Markers” or “All Arrays” checkboxes can be used to override any activated sets of Markers or Arrays..

The checkbox for “Rank Statistics Plot” transform the data from being plotted by expression value to being plotted by their rank. The “Clear Charts” button removes all charts, the “Print” button allows for printing, and the “ImageSnapshot” button takes a snapshot of the Scatterplot. The “Reference Line” checkbox simply adds or removes the reference line from the Scatterplot.

7.2.7 Expression Value Distribution

The expression value distribution component plots the expression values of microarray markers for the selected hybridized array or arrays. Both can be specified in the selection panel in the markers and arrays/phenotypes tabs respectively and should immediately be plotted on the graph. The expression value distribution is useful for determining the difference in expression levels between sets of markers under difference conditions, a T-Test can be used to detect markers with significantly different expression. The controls for the expression value distribution are shown below.

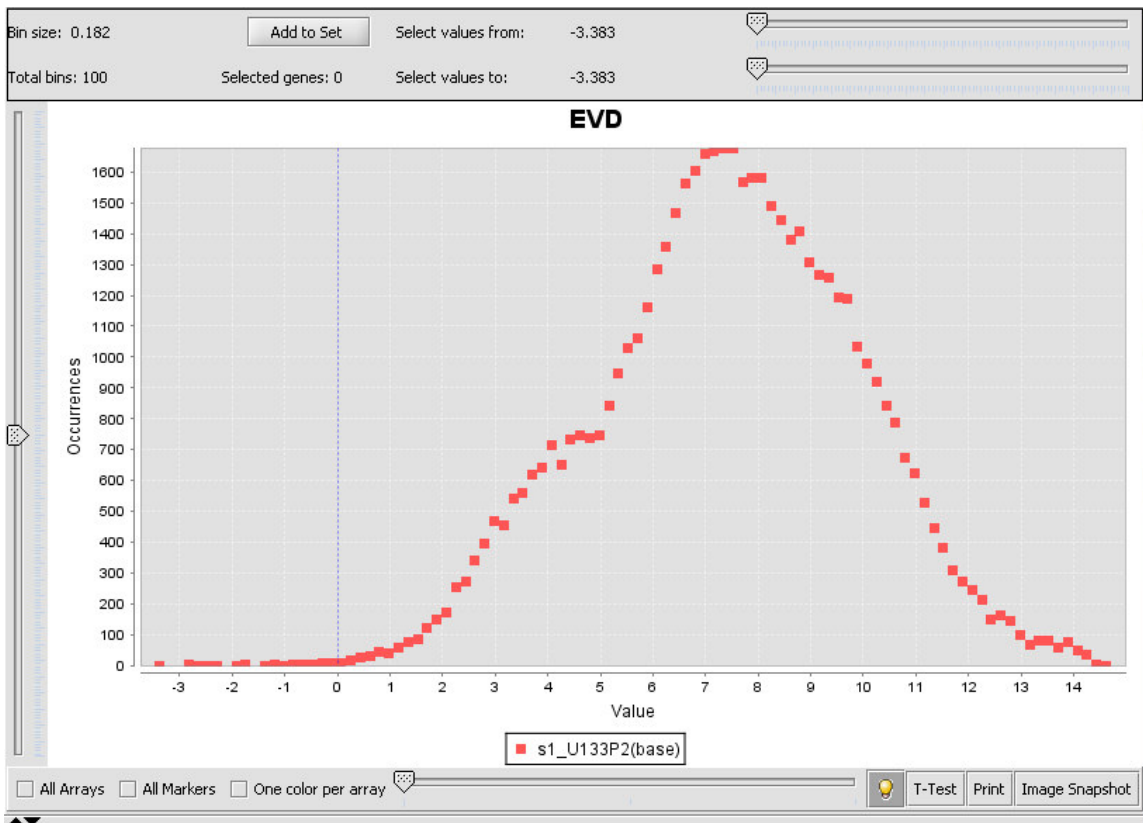


Figure 7-13 Expression Value Distribution

Controls:

Bin size and Total Bins: Each expression range on the EVD corresponds to a “bin”. The size of each bin and total bins are displayed at the upper left corner of the component. The number of total bins equals $(m - M) / \text{number of bin size}$.



Tooltip: When a user scrolls over a plot, the enabled tool tip displays the name of the array. The tooltip display can be disabled by selecting the tooltip icon. The disabled tooltip icon appears grayed out.

Change the base array: The slider can be used to navigate through the microarrays in the selected dataset. As the user moves the slider to the next tick which represents a microarray, the grid is updated with positions representing the marker values in the Microarray being displayed. The base array always labels with the red color.

All Markers and All Arrays Checkbox: These checkboxes override any activated sets of Markers or Arrays, causing all to display.

One Color per Array: The EVD lines and plots appear color-coded, based on the color preferences of the arrays panel. All the arrays in a set are shown with the same color. The user can modify the color display to reflect a unique color per array by selected the checkbox **One Color per Array**.

Legend: A legend appears at the bottom of the plot indicating which color and shape corresponds to which panel. Modification of image display preferences are described in details in Preferences. The base array always lists as the first one.

Image Snapshot: Clicking on this button creates image snapshots that are saved in the project.

Zoom: Zooming in can be done by left or right- clicking and dragging down and to the right over a region of the image. Zooming out is done similarly but dragging up and to the left..

T-Test: This button computes a t-test statistic for each marker and updates the graph title from EVD to T-Test and disables the base array slider. Case and control panels must be created and classified (see Marker Sets section for additional information).

Print: Prints the displayed EVD. The print dialog pop up is displayed to support printer selection.

Selected Values from/to: The starting/ending X axis locations that allow markers to be selected graphically and filtered. The number of current selected genes is displayed next to the Total bins. “Add to Set” button adds the selected genes to the Markers component.

The Analysis/Annotation Window

The tab-indexed tools in this last component area include facilities for filtering, normalizing and analyzing data, along with components for viewing the history of operations that have been performed on a data set and general experiment and annotation information. All of the filtering, normalization, and analysis tools include a **Save Settings** option, which saves the parameters used in the analysis or processing step with the workspace.

Filters are used to remove data points when some data quality or signal criteria are not met. As a result of applying a filter, the status call of a questionable data point may be reset to "missing," or alternatively, the data point may be eliminated altogether from the data set. In the later case, all measurements for that marker (across all chips in the data set being filtered) will be eliminated. In contrast, normalizers do not change the status or remove individual markers, but re-scale the observed intensities, usually in preparation for some type of analysis. Filtering or normalizing is done directly on a dataset in geWorkbench, and does not create a new copy of the data. Copies of data in a particular state can be created manually by saving a dataset to a file.

7.3 Filtering Operations

The **Filtering Panel** contains several filters that allow the software to set certain values to missing. For instance, the *Affy detection call* filter allows the user to filter out undesirable values on the basis of the Affymetrix calls ("present," "absent" or "missing").

Figure 7-14 shows the result of applying this filter to remove from analysis all markers with a call of "absent". In the **Microarray Viewer** display of the filtered data, all of the missing data points now appear as yellow squares in the heat map. Table 7-3 summarizes the filters that are available from a pull-down list in the **Filtering** component. *Please note that all filtering and normalization functions change the original dataset and do not keep it intact.*

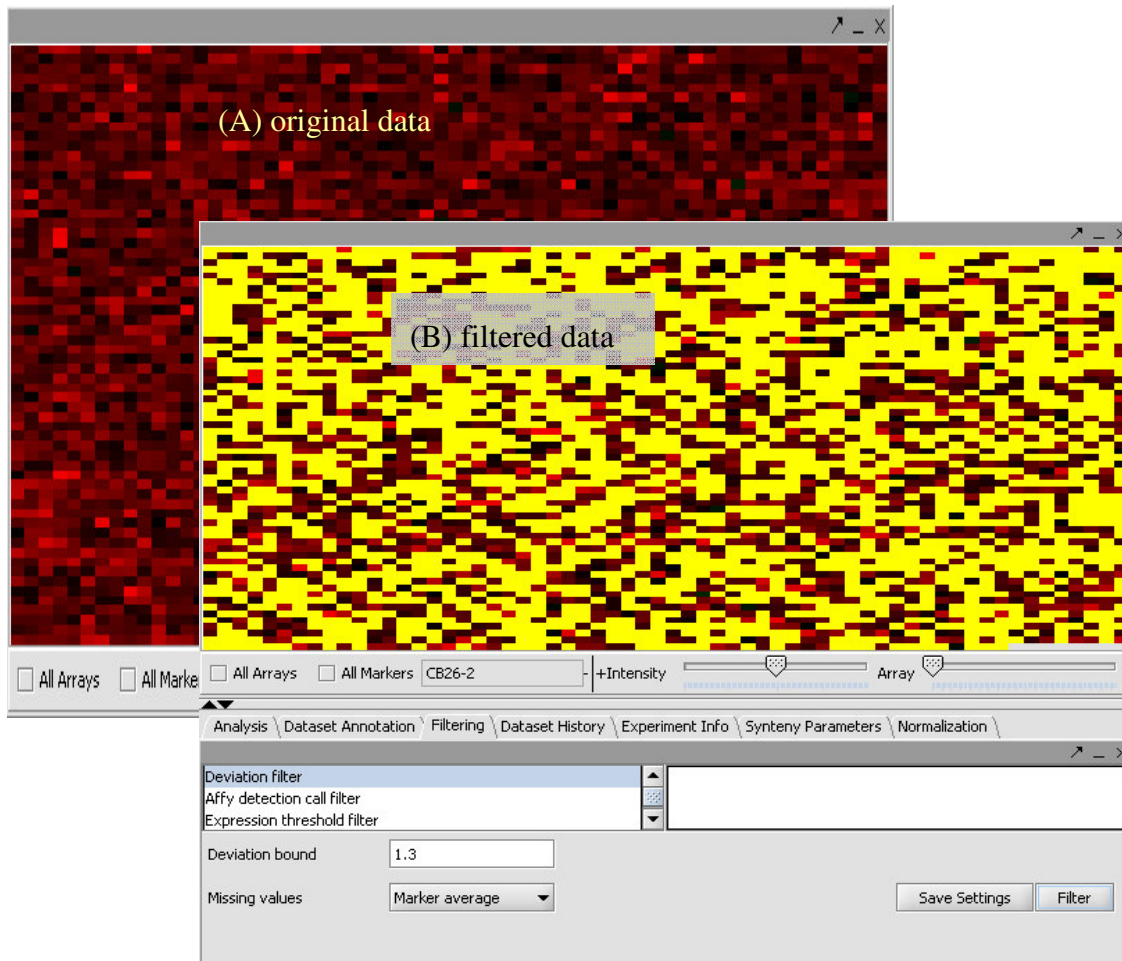


Figure 7-14 Microarray Viewer showing data before (A) and after (B) filtering.

Table 7-3 The Filtering Panel Toolset

<u>Filter</u>	<u>Description</u>
Missing values	Discards all markers that have “missing” measurements in at least <i>n</i> microarrays, where <i>n</i> is defined by the user. Another filter must first be applied however, in order to generate the missing values upon which this filter can operate.
Deviation	Sets as missing all markers whose measurements deviate below a given value across all microarrays.
Expression threshold	Sets as missing all markers whose measurements are inside (or outside) a user-defined range.
Affy detection call	Applicable to Affymetrix data only. Sets all measurements whose detection status is any user-defined combination of A, P or M as

	missing (Absent, Present, Marginal).
2-channel threshold	Applicable to 2-channel arrays (Genepix) data only. Defines applicable ranges for each channel, and sets all values for which either channel intensity is inside (or outside) the defined range as missing.
Genepix Flags	Removes values flagged in Genepix file

7.3.1 Normalization Tools

Before comparing multiple microarrays, the user must first ensure that the observed values therein have been made “comparable” through a process of normalization. The normalization panel offers the user several gene-centric or array (tissue)-centric methods that are summarized in Table 7-4.

Table 7-4 The Normalization Panel Toolset

<u>Normalization Tool</u>	<u>Description</u>
Missing value calculation	Replaces every missing value with either the mean value of that marker across all microarrays or with the mean measurement of all markers in the microarray where the missing value is observed.
Log2 Transformation	Applies a log2 transformation to all measurements in a microarray.
Threshold Normalizer	All data points whose value is less than (or greater than) a user-specified minimum (maximum) value are raised (reduced) to that minimum (maximum) value
Marker-based centering	Subtracts the mean (median) measurement of a marker profile from every measurement in the profile
Array-based centering	Subtracts the mean (median) measurement of a microarray from every measurement in that microarray.
Mean-variance normalizer	For every marker profile, the mean measurement of the entire profile is subtracted from each measurement in the profile and the resulting value is divided by the standard deviation.
Housekeeping Normalizer	Genes Uses a set of house-keeping genes to normalize expression of all genes on all arrays such that the averaged expression value of house-keeping genes is constant across all microarrays.

Normalization Tool

Description

Quantile Normalization

Assumes distribution of probe intensities is nearly the same in all samples and calculates the quantile of each value and normalizes it against a reference chip.

7.3.2 Dataset History

geWorkbench provides a convenient system for electronic tracking of all actions. As noted, each time a new file or image is generated, that file appears in the **Project Tree Window** as a new node occurring beneath the data set from which it was derived. In addition, the **Dataset History** window displays a list of all of the operations that were performed on both the currently selected data set as well as on all of its “parent” data sets in the **Project Tree**. Figure 7-15 shows the history window for a data normalization workflow.

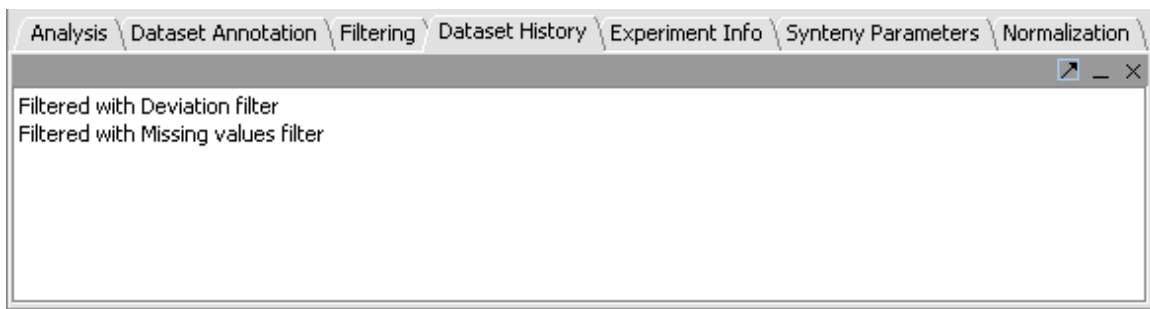


Figure 7-15 The Dataset History Window

7.4 The Analysis Tools

The Analysis area (Figure 7-16) contains access to a number of clustering and statistical analysis tools.. Analysis parameters can be saved to the file system. Analysis results are displayed in the separate viewing region of the application, and datasets are placed in the Project Folders area beneath their parent dataset.. Some of the algorithm implementations are adapted from TIGR's MEV⁵ software.

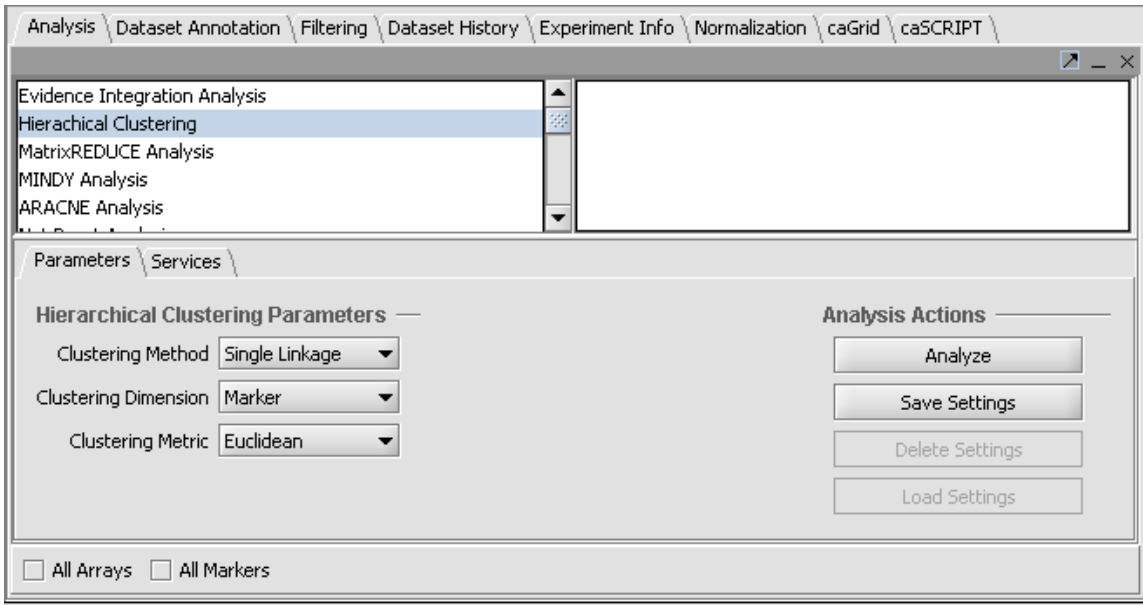


Figure 7-16 The analysis component.

7.4.1 Hierarchical Clustering:

geWorkbench provides two clustering methods in the **Analysis** panel, Hierarchical clustering and SOM (Self-Organizing Maps). Hierarchical clustering groups markers on the basis of similarities in their expression profiles, and outputs a hierarchical tree that can be viewed in the **Dendrogram** component (see Figure 7-17). Clustering can also be done in the array dimension if desired.

The parameters used for this analysis method are Clustering Method, Clustering Dimension and Clustering Metric. The clustering dimension determines whether clustering should be done against the markers, arrays, or both. The remaining two determine how the clustering is performed and are described further in the following:

The clustering methods provided are shown in Table 7-5:

Table 7-5 Hierarchical clustering methods

<u>Hierarchical Tool</u>	<u>Description</u>
Single linkage	The distances are measured between each member of one cluster each member of the other cluster. The minimum of these distances is considered the cluster-to-cluster distance.
Average linkage	The average distance of each member of one cluster to each member of the other cluster is used as a measure of cluster-to-cluster distance.

Hierarchical Tool

Description

Total Linkage

The distances are measured between each member of one cluster each member of the other cluster. The maximum of these distances is considered the cluster-to-cluster distance.

Distances metrics available for cluster computations are:

1. Euclidean
2. Pearson's Correlation
3. Spearman Rank

The Dendrogram view is very flexible. The image can be scrolled. The cluster tree can be saved as an image. The sizes of the grid in the mosaic representing genes can be altered. The color intensities can be altered. A zoom feature allows subtrees to be selected, viewed, and saved as Marker sets. Mouse-over information (expression value) can be toggled with the light bulb button at lower right. Finally, image snapshots of the dendrogram can be taken.

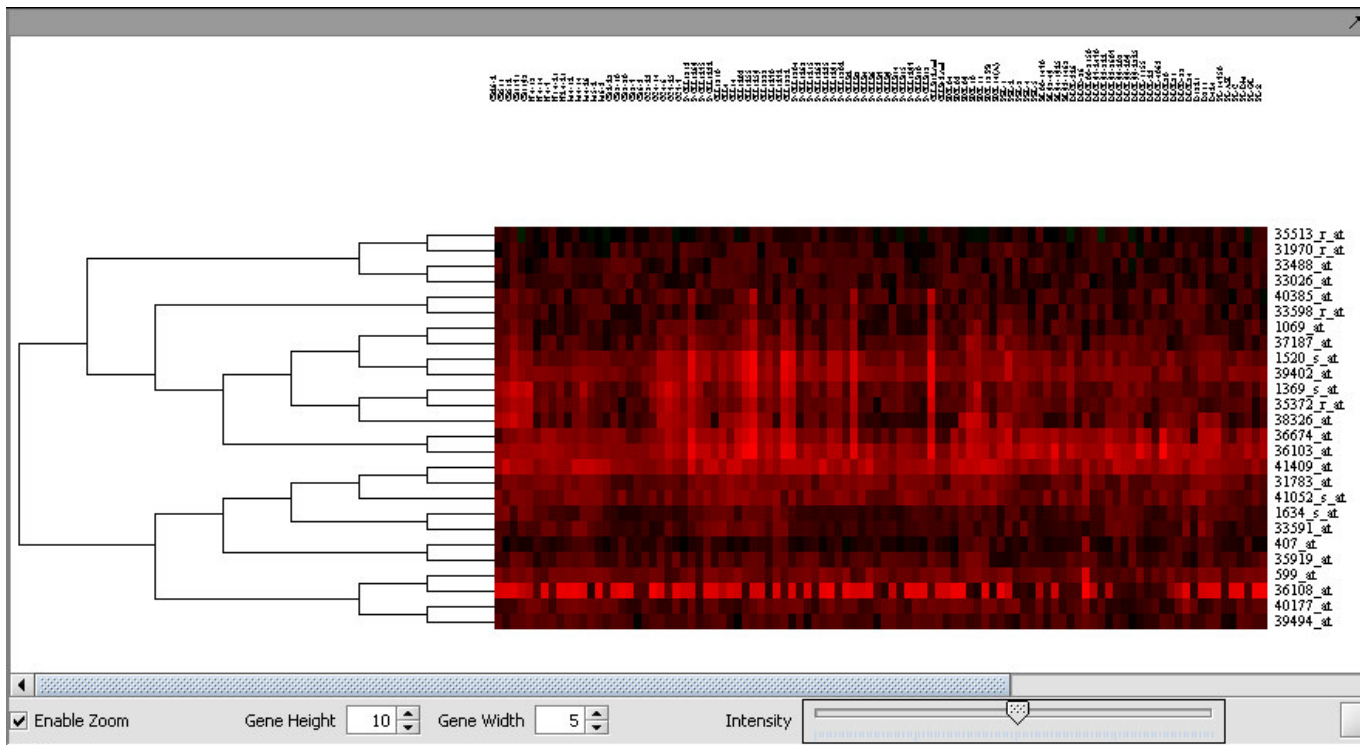


Figure 7-17 An example of viewing a subtree from Hierarchical Clustering in the Dendrogram component.

7.4.2 Self Organizing Map (SOM)

An implementation of SOM³ is provided. The available parameters are:

Table 7-6 SOM parameters

<u>SOM Tool</u>	<u>Description</u>
Rows	The number of rows that the user desires in the resulting SOM.
Columns	The number of columns that the user desires in the resulting SOM .
<u>Radius</u>	When using the bubble neighborhood parameter this float value is used to define the extent of the neighborhood. If an SOM vector is within this distance from the winning node (the cluster to which an element has been assigned) then that node (and SOM vector) is considered to be in the neighborhood and its SOM vector is adapted.
Iteration	The number of times the dataset will be presented to the Map. Each expression element will be presented this number of times to train the nodes.
Alpha	This value is used to scale the change of individual SOM vectors when a new expression vector is associated with a node.
Functions	<p>The neighborhood options indicate the conventions (formulas) used to update (adapt) an SOM vector once an expression vector has been added into a node's neighborhood.</p> <p>Bubble: This option uses the provided radius (see above) to determine which surrounding SOM nodes are in the neighborhood and therefore are candidates for adaptation. When this option is selected the Alpha parameter for scaling the adaptation is used directly as provided from the user.</p> <p>Gaussian: This option forces all SOM vectors in the network to be adapted regardless of proximity to the winning node. In this case the Alpha parameter is scaled based on the distance between the SOM vector to be adapted and the winning node's SOM vector.</p>

SOM analysis uses self-organizing neural nets to identify genes with similar expression patterns, and maps expression profiles into the cells of user defined grids. The **SOM Clusters** View (Figure 7-18) can then be used to explore the resulting maps. It contains a grid of profile graphs depicting each cluster in the SOM results. The number of cells

seen in the results grid equals the product of the number of rows and columns provided in the analysis input. Clusters resulting with no entries are shown as empty profile graphs in the plot. Each profile in a grid cell corresponds to the gene profile of the corresponding gene in the input dataset. Each cluster can be viewed alone by selecting the **Show selected** checkbox and selecting the appropriate grid cell. Finally, image snapshots of the SOM grid as well as zoomed-in images can also be taken.

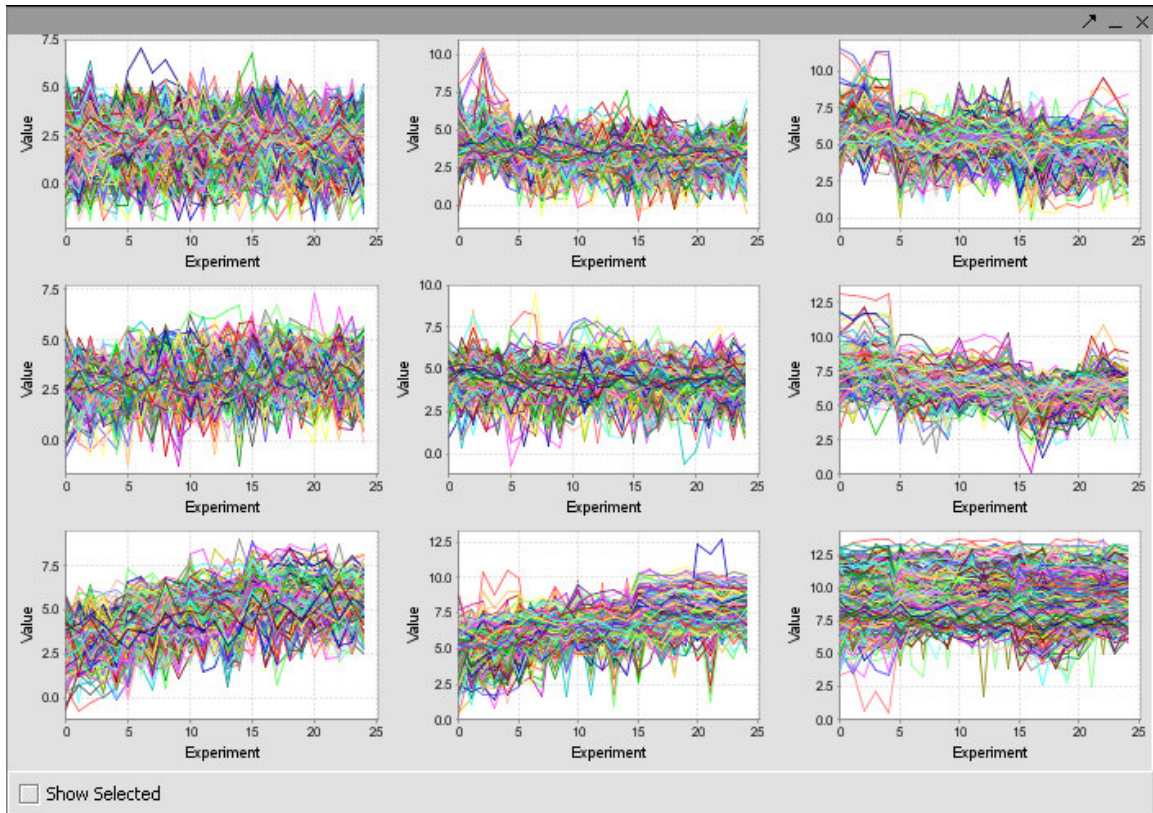


Figure 7-18 The SOM Clusters View window

geWorkbench's implementation of these algorithms is based on their implementation in the [Multi Experiment Viewer \(MEV\)](#) platform, which is freely available from The Institute for Genomic Research (TIGR).

7.4.3 The Dataset Annotation Tool

This panel provides a simple text window for adding any textual information that the user wishes to associate with a particular data set. Examples might include annotations found on the CGAP web site, questions that arise during the analysis which the user may wish to pursue at a later time, actions that were taken that are not otherwise tracked by geWorkbench, etc.

Cut and paste operations are supported to facilitate importing and/or exporting text to and from this window. In particular, as the "parent" data set's annotations are not inherited by

the “child” nodes, the user may wish to copy and paste some of these as new data sets are derived. Any text entered in this window will be saved and retrieved with the experiment when the workspace is reopened.

geWorkbench supports user annotation/comments for images that appear in the Project Folders component. First an image must be created by one of the components, for instance through the Expression Value Distribution component as described in section 3.2.8. Then the user can attach an annotation by right mouse clicking on the generated image to create a new annotation label for the image. The user can also use Edit -> Rename -> File to accomplish the same task.

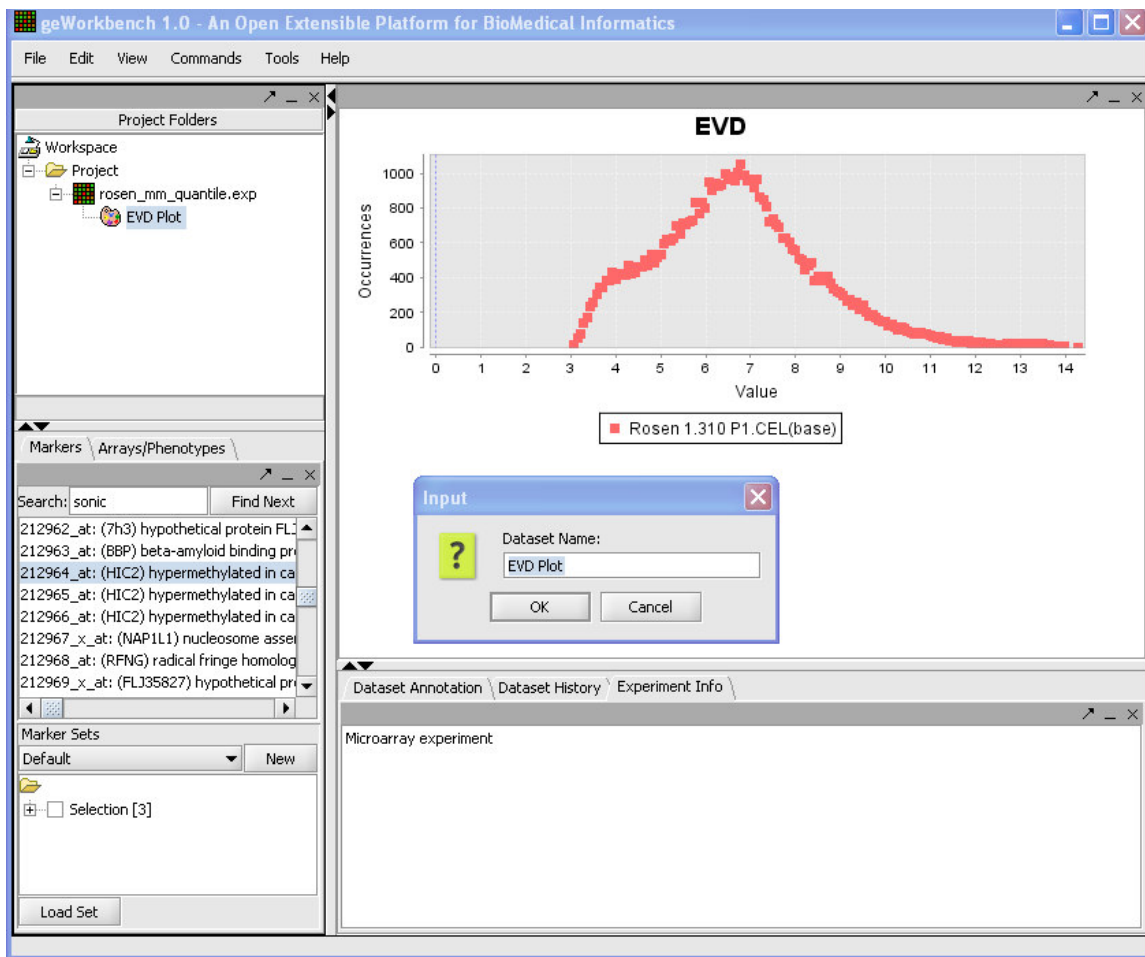


Figure 7-19 Adding a dataset annotation

7.4.4 The Experiment Info Tool

This read-only text window displays the textual preamble that precedes the data in most experiments. While it is not possible to modify the text in this display, the user can copy that text if desired into a **Dataset Annotation** panel. For example, when two

independent data sets are merged to form a new data set, the latter has no experiment information associated with it. Using copy and paste operations, the user can copy the experiment information from each of the original data sets into the **Dataset Annotation** window for the merged data.

Software Tools and References:

1. **GeneChip Arrays.** 2003. Affymetrix. First and most comprehensive whole human genome expression array. <http://www.affymetrix.com>
2. **Eisen, M.B., Spellman P.T., Brown P.O., and Botstein D.** 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863-14868.
3. **Kohonen, T.** 1997. *Self-organizing Maps.* Springer-Verlag, Berlin.
4. **Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., and Golub T.R.** 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907-2912.
5. **Saeed A.I., Sharov V., White J., Li J., Liang W., Bhagabati N., Braisted J., Klapa M., Currier T., Thiagarajan M., Sturn A., Snuffin M., Rezantsev A., Popov D., Ryltsov A., Kostukovich E., Borisovsky I., Liu Z., Vinsavich A., Trush V. and Quackenbush J.** 2003. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34(2):374-8.
6. **Covitz P.A., Sahni H., Gustafson S., and Buetow K. National Cancer Institute Center for Bioinformatics.** 2002. Cancer Bioinformatics Infrastructure Objects (caBIO): An open-source, object oriented API for biomedical informatics. Objects in Bio- & Chem-Informatics: <http://lsr.omg.org/oibc2002/>
7. **GenePix Scanner.** Axon Instruments. <http://www.axon.com>
8. **Microarray Gene Expression Data Society (MGED).** Microarray and Gene Expression (MAGE). <http://www.mged.org/Workgroups/MAGE/mage.html>
9. **Cancer Bioinformatics Objects (caBIO).** National Cancer Institute Center for Bioinformatics. <http://ncicb.nci.nih.gov/core/caBIO>
10. **Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J.E. Hughes, E. Snesrud, N. Lee, and J. Quackenbush.** 2000. A concise guide to cDNA microarray analysis. *BioTechniques* 29:548-556.

8 Differential Expression (t test)

8.1 Overview

A t-Test analysis can be used to identify markers with statistically significant differential expression between two sets of microarrays. The t-test determines, for each marker, if there is a significant difference between the two groups (e.g. case and control). To perform this analysis, you must classify the sets, set the analysis parameters and view the results in the visualization components. The t-test implementation in geWorkbench offers several options for multiple testing correction and evaluation of the test statistic. A detailed description of the t-Test parameters is also available in online help.

8.2 Preparation

Use the file "webmatrix_quantile_log2_dev1.2_mv0.exp", which is contained in the downloadable zip archive tutorial_data.zip. See the [Download](#) area.

The result screenshots below were generated using an earlier dataset, which was obtained by filtering with a deviation bound of 1.0. The dataset currently supplied was created using a deviation bound of 1.2, so results will differ slightly from those shown.

For tips on loading data files, see the section [Tutorial - Projects and Data Files](#).

8.3 t-Test Parameters

8.3.1 P-value

The p-value can be estimated from 1. the t-statistic (the default) or 2. by permutation.

8.3.2 Alpha corrections

For multiple testing (alpha) correction, the following options are offered: 1. no correction 2. Standard Bonferonni Correction 3. Adjusted (step down) Bonferonni Correction. 4. Additional methods are available if the p-value is being estimated by permutation.

8.3.3 Degrees of Freedom

Group variances can be declared as: 1. unequal (Welch approximation) (the default) 2. Equal.

8.3.4 Classification

The desired sets of arrays should be activated in the Arrays/Phenotypes component. This is done by checking the boxes by the desired Sets.

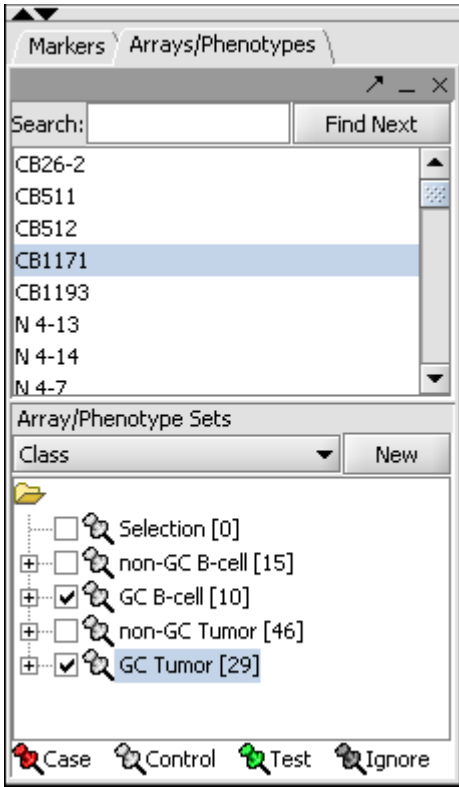


Figure 8-1 - Activation of array sets

The t-test requires two groups of microarrays to compare. geWorkbench distinguishes the two groups by one being labeled as "Case". By default, all others are considered as control. Note that in the Arrays/Phenotypes component, more than one set of arrays can be marked "Case". All remaining (activated) arrays will then be in the "Control" group.

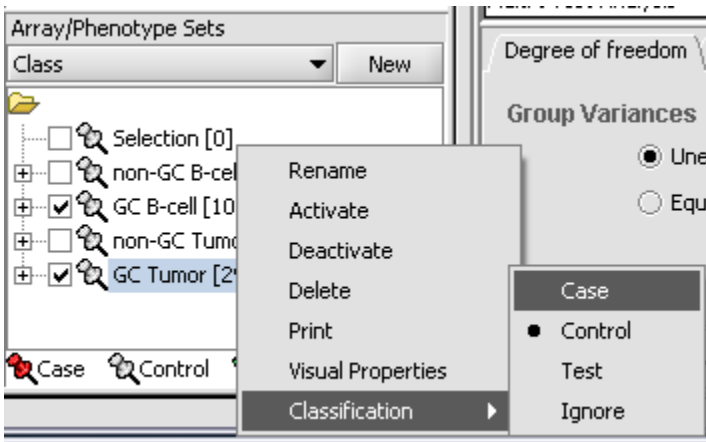


Figure 8-2 - Setting case/control status of array sets

The thumbtack image next to activated Array Sets is colored red.

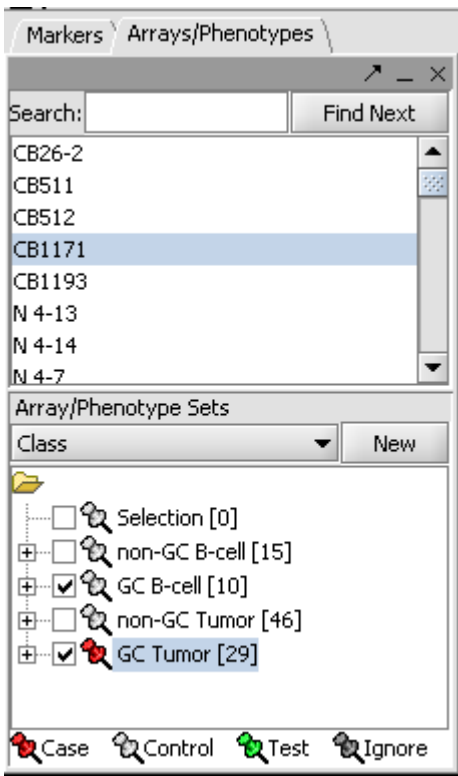


Figure 8-3 – An array set marked as "Case"

8.3.5 Set Analysis Parameters

- From the Analysis Panel, select **T-Test Analysis**.
- Various parameters can be adjusted as desired. Here we will use the Standard Bonferonni method, which is the strictest.
- Alpha-corrections tab: Standard Bonferonni.

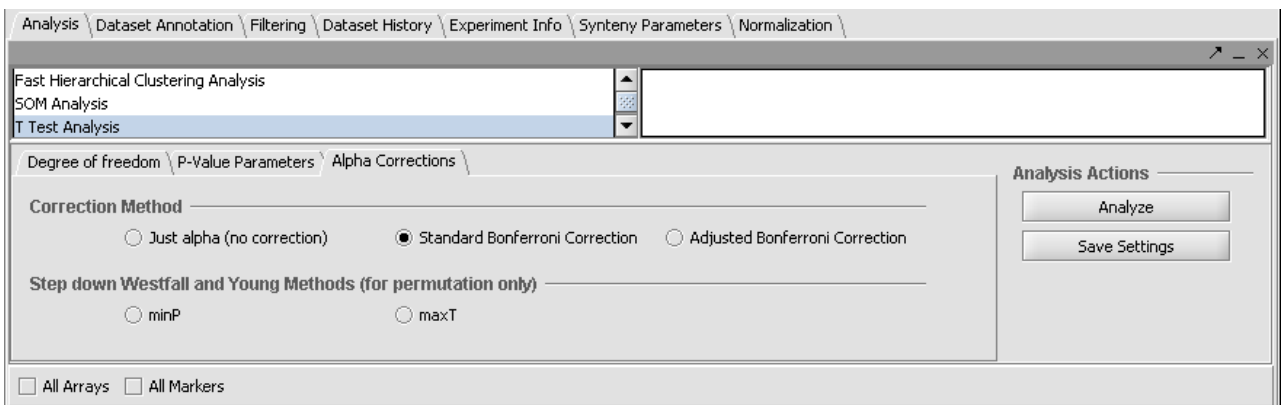


Figure 8-4 - Multiple testing corrections

- P-Value Parameters tab: p-values based on t-distribution. Note that the default alpha (critical p-value) is set to 0.01.

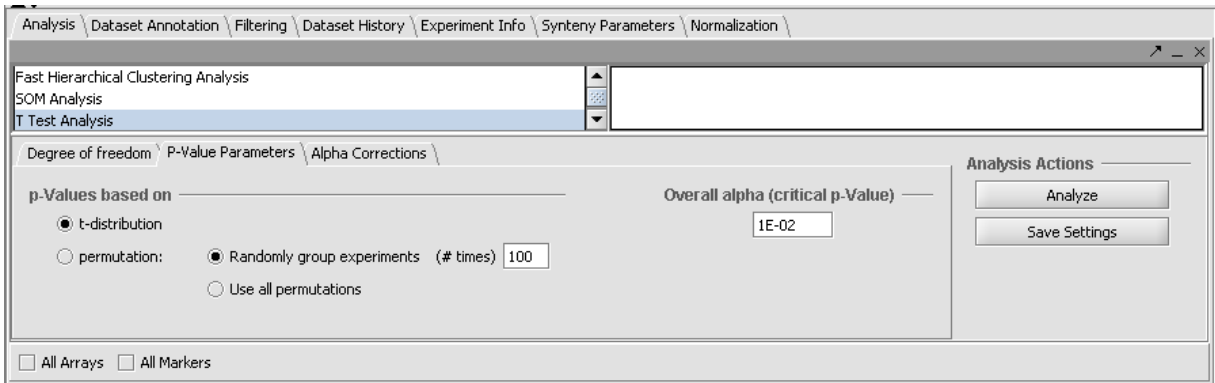


Figure 8-5 - P-value parameters

- Degree of Freedom tab: Welch approximation - unequal group variances.

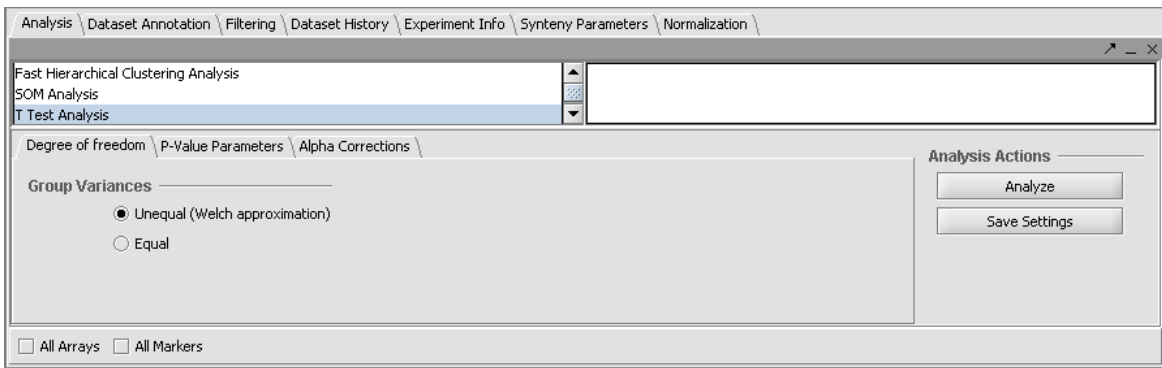


Figure 8-6 - Degrees of freedom

After all the parameters have been set, click Analyze. The results will be returned in three locations: The Project Folder, the Markers component, and the Visualization area.

8.3.6 t-Test Results

The result is placed into the Projects Folder as a child of the microarray dataset that was analyzed.

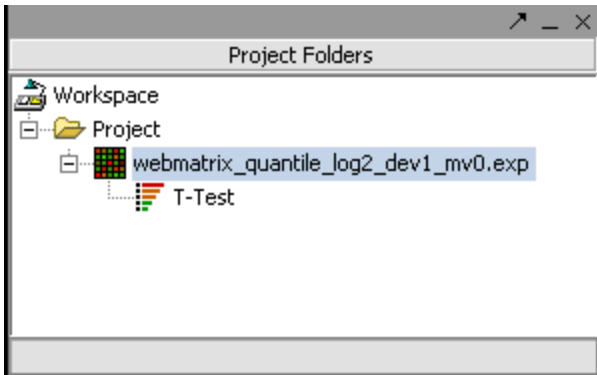


Figure 8-7 – T-test result node in Project Folder

The results are displayed by default using the Volcano Plot visualizer.

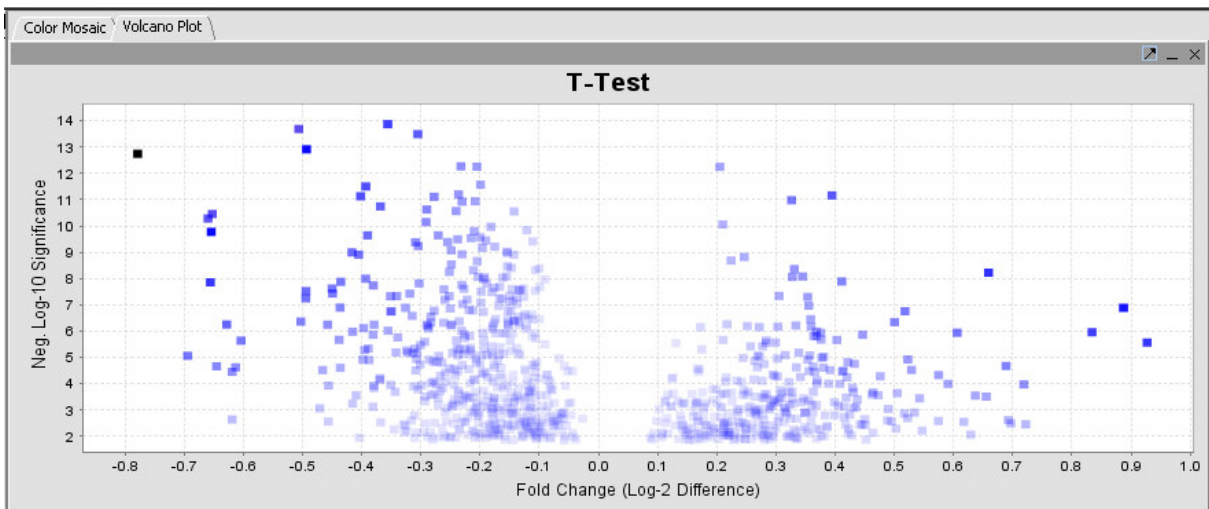


Figure 8-8 - T-test Volcano Plot

The adjacent tab provides a Color Mosaic showing all of the arrays and the p-value calculated for each marker. It also can display annotation for each marker.

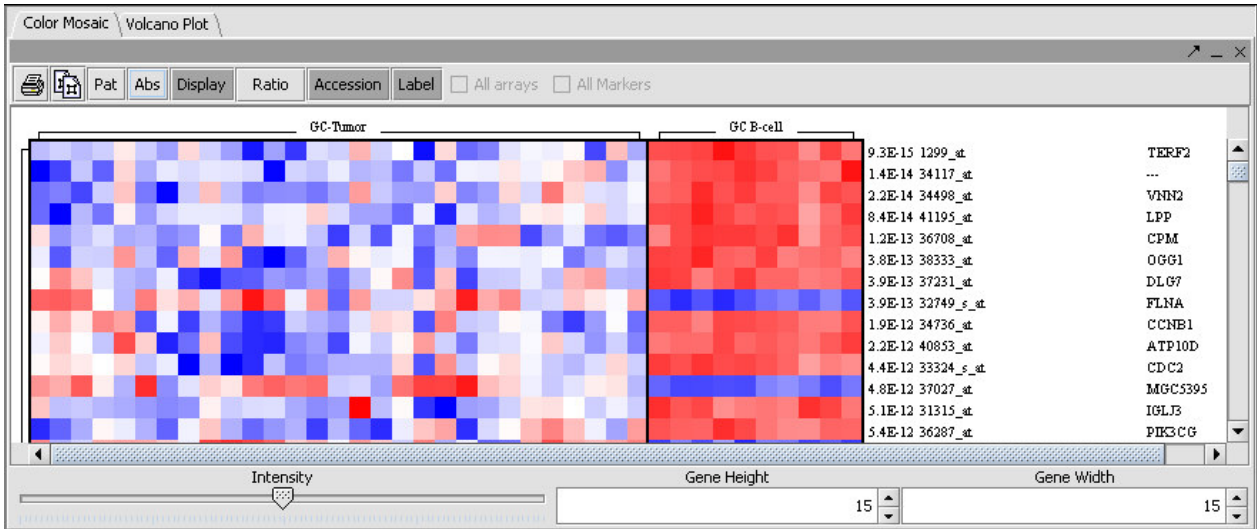


Figure 8-9 - T-test result displayed as a color mosaic

The set of markers which met the minimum significance criterion are placed into a new Marker Set labeled "Significant Genes" in the Markers component. The number of markers is shown also.

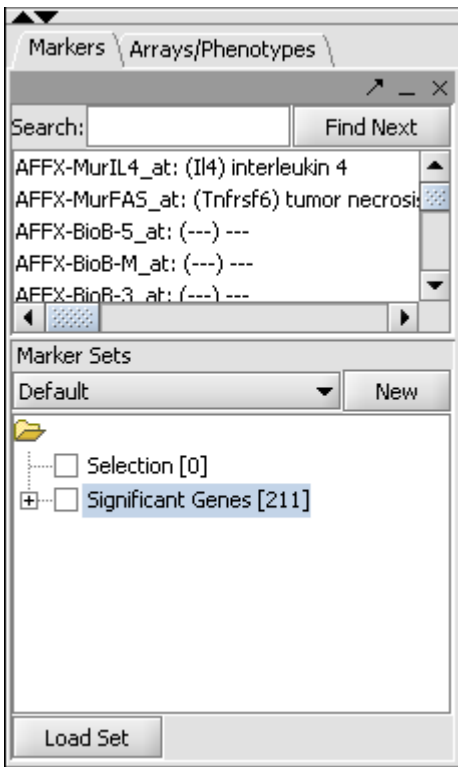


Figure 8-10 - T-test significant genes returned to a marker set.

9 Marker Annotations

9.1 Overview

The Marker Annotations component enables the retrieval of biological annotation information for a collection of genes. For every gene, the following data can be retrieved:

- A set of pathways containing the gene.
- A set of gene-disease and gene-compound associations derived from the literature articles.

All annotations are retrieved from remote servers maintained by the National Cancer Institute (NCI). The data in those server come from the following sources:

- **Pathways:** NCI's Pathway Interaction Database (PID). PID pathways come from 3 sources: BioCarta, Reactome and "NCI-Nature Curated". Information about the PID and each of the contributing sources is available at: http://pid.nci.nih.gov/userguide/database_content.shtml. These pathways are stored in servers used by the Cancer Gene Anatomy Project (CGAP, <http://cgap.nci.nih.gov/>).
- **Gene-disease/compound associations:** the Cancer Gene Index (CGI) data base. The reported associations are extracted from article abstracts using a combination of automatic text mining, semi-automatic verification, and manual curation. Project details are available at: <http://ncicb.nci.nih.gov/NCICB/projects/cgdcp>.

9.2 Submit Query

The Marker Annotations module will retrieve information for all markers that belong to activated marker sets, or, more precisely, for the genes corresponding to those markers:

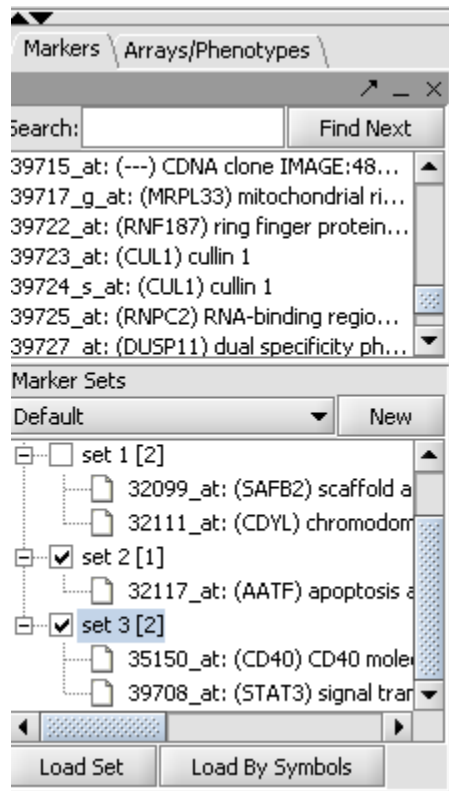


Figure 9-1 - Marker set activation

E.g., in the example shown above, information will be retrieved about the genes AATF, CD40, STAT3. Checkboxes at the bottom of the component's user interface can be used to specify which data source(s) to query: CGAP, CGI or both. For CGAP, the associated drop-down can be used to designate the target organism for which annotations are retrieved: human (the default) or mouse. Clicking the "Retrieve Annotations" button initiates the communication with the NCI servers:

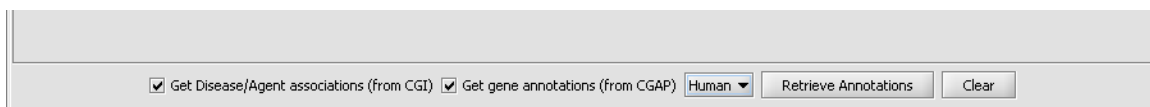


Figure 9-2 - Choosing annotation types for retrieval

9.3 Pathway and Gene Annotations

The "Annotations" tab presents a summary listing of the annotations retrieved from CGAP:

Marker	Gene	Pathway
32117_at	AATF	h_rb_1pathway
35150_at	CD40	h_dcPathway
35150_at	CD40	h_asbcellPathway
35150_at	CD40	h_cd40
35150_at	CD40	h_th1th
35150_at	CD40	h_bbccl
35150_at	CD40	h_blym
39708_at	STAT3	h_ptp1bpathway
39708_at	STAT3	h_pdqfPathway
39708_at	STAT3	h_edfPathway
39708_at	STAT3	h_TPOPathway
39708_at	STAT3	h_il6Pathway
39708_at	STAT3	h_met_pathway
39708_at	STAT3	h_il2_1pathway
39708_at	STAT3	h_il6_7pathway
39708_at	STAT3	h_stat3Pathway
39708_at	STAT3	h_il12_2pathway
39708_at	STAT3	h_il23pathway
39708_at	STAT3	h_hdac_classi_pathway

Get Disease/Agent associations (from CGI)
 Get gene annotations (from CGAP)
 Human

Figure 9-3 - List of retrieved annotations

The listing contains at least one row for each gene which annotation information is available for. If a gene is associated with more than one pathways, then one row for every pathway is listed (e.g., as is the case above for CD40 and STAT3). Every row displays the marker (i.e., probeset) id, the corresponding gene name and the name of the associated pathway. Clicking on a pathway brings up a popup menu offering a number of options:

- **View Diagram:** available only for BioCarta pathways. Such pathways are accompanied by images offering a graphical/artistic rendition of the pathway. Selecting the "View Diagram" option will display this image within the "Pathway" tab.
- **Add pathway genes to set:** extracts the pathway genes for which there are associated probes in the microarray set currently selected by the user and places all such probes in a new marker set within the "Markers" component (by default, the marker set is named after the pathway).
- **Export genes to CSV:** creates a new text file containing a listing of all pathway genes. The file format (csv = comma separated values) is compatible with Microsoft Excel.

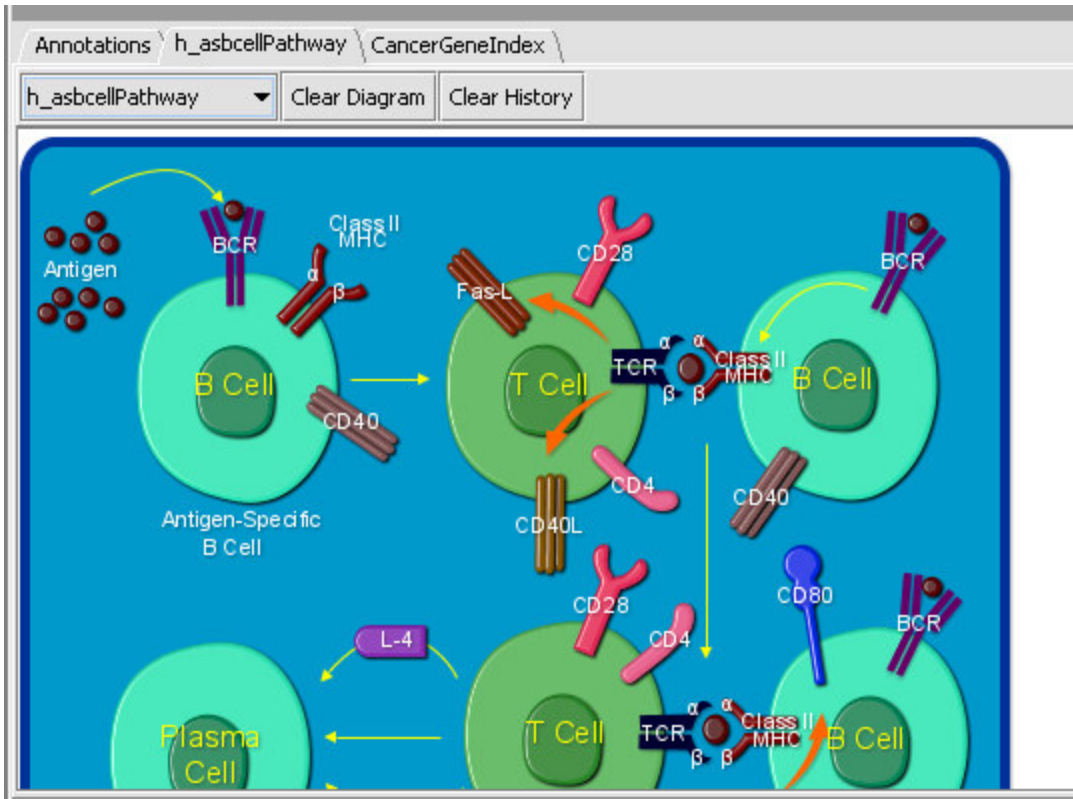


Figure 9-4 - BioCarta pathway from caBIO

A BioCarta pathway image is displayed above after selecting the "View Diagram" option from the "Annotations" tab. The drop-down box, on the top left corner above the diagram, shows the name of the currently displayed diagram. The component keeps a history of all BioCarta diagrams selected by the user; using the drop-down it is possible to switch among the corresponding pathway images. The "Clear Diagram" button clears the currently displayed diagram. The "Clear History" button both clears the currently displayed diagram and removes all pathway history information from the pathway name drop-down box.

In the "Annotations" tab, it is also possible to click on a gene name and explore functional annotation information from a number of sources (Entrez, CGAP, GeneCards):

35150_at	CD40	h_dcPathway
35150_at	CD40	h_th1th2Pathway
35150_at	CD40	h_bcellPathway
35150_at	CD40	h_blymphocytePathway
35150_at	CD40	h_cd40Pathway

Go to Entrez for CD40
Go to CGAP for CD40
Go to GeneCards for CD40

Figure 9-5 - Retrieving gene entries from Entrez

9.4 Cancer Gene Index

For many genes, there are hundreds of records in the CGI database. Retrieving all those records at once can be a very time consuming operation, especially if the query involves many genes. To avoid very long waits, the retrieval of the data occurs in 2 stages. In the first stage, at most 10 records for each association type (gene-disease/gene-compound) are being fetched (for each query gene). Data retrieved as displayed in the CancerGeneIndex tab:

Marker	Gene	Disease	Role	Sentence	Pubmed
39708_at	STAT3	melanoma(6)
39708_at	STAT3	leukemias(1)
39708_at	STAT3	breast cancers(1)
39708_at	STAT3	prostate carcinomas(1)
39708_at	STAT3	hepatocellular carcinomas(1)

Marker	Gene	Agent	Role	Sentence	Pubmed
39708_at	STAT3	epo(4)
39708_at	STAT3	imatinib(2)
39708_at	STAT3	rapamycin(1)
39708_at	STAT3	dextran sulfate sodium(1)
39708_at	STAT3	hsp70(1)
39708_at	STAT3	curcumin(1)

Figure 9-6 - Retrieving additional records from the Cancer Gene Index

The user interface is divided in two regions, sharing the same overall structure and functionality: the table on the left displays gene-disease associations, while the table on the right is used for reporting gene-compound associations. To avoid redundancy, the paragraphs that follow describe only the gene-disease table. The exact same description applies in the case of the gene-compound table.

Each table row represent an association between a query gene and a disease. The first column contains the name of the probeset associated with the query gene. For any given gene, if its corresponding probeset name appears in bold-face, this means that there are more gene-disease records on the CGI server that have yet to be fetched (beyond the 10 records acquired in the initial retrieval stage). The disease name within a row is followed by a number in parentheses. This number indicates how many records (among those fetched) support the reported association. E.g, in the image above, the first row indicates that there are 6 distinct records linking STAT3 to melanoma. A detailed listing of these 6 records is available by "expanding" the row. This can be achieved by right-clicking on the row and selecting "expand" from the ensuing popup menu (Figure 9-6). The resulting display is shown in

Marker	Gene	Disease	Role	Sentence	Pubmed
39708_at	STAT3	melanoma	Gene_Associated_With_Disease	Gene therapy with dominant-ne...	10537273
39708_at	STAT3	melanoma	Gene_Product_Expressed_In_Tissue	Prolonged DEX treatment of mel...	12589944
39708_at	STAT3	melanoma	Gene_Product_Expressed_in_Disease	Prolonged DEX treatment of mel...	12589944
39708_at	STAT3	melanoma	Gene_Product_Has_Biochemical_Function	Interleukin-6-resistant melanom...	11442760
39708_at	STAT3	melanoma	Gene_Product_Malfunction_Associated_...	The expression of dominant ne...	11463377
39708_at	STAT3	melanoma	Gene_Product_js_Pathway_Element	Significantly, melanoma cells un...	12370822
39708_at	STAT3	breast cancers(1)
39708_at	STAT3	hepatocellular carcinomas(1)
39708_at	STAT3	leukemias(1)
39708_at	STAT3	prostate carcinomas(1)

Gene therapy with dominant-negative Stat3 suppresses growth of the murine melanoma B16 tumor in vivo.

Get Disease/Agent associations (from CGI) Get gene annotations (from CGAP) Human

Figure 9-7.

Marker	Gene	Disease	Role	Sentence	Pubmed
39708_at	STAT3	melanoma	Gene_Associated_With_Disease	Gene therapy with dominant-ne...	10537273
39708_at	STAT3	melanoma	Gene_Product_Expressed_In_Tissue	Prolonged DEX treatment of mel...	12589944
39708_at	STAT3	melanoma	Gene_Product_Expressed_in_Disease	Prolonged DEX treatment of mel...	12589944
39708_at	STAT3	melanoma	Gene_Product_Has_Biochemical_Function	Interleukin-6-resistant melanom...	11442760
39708_at	STAT3	melanoma	Gene_Product_Malfunction_Associated_...	The expression of dominant ne...	11463377
39708_at	STAT3	melanoma	Gene_Product_js_Pathway_Element	Significantly, melanoma cells un...	12370822
39708_at	STAT3	breast cancers(1)
39708_at	STAT3	hepatocellular carcinomas(1)
39708_at	STAT3	leukemias(1)
39708_at	STAT3	prostate carcinomas(1)

Gene therapy with dominant-negative Stat3 suppresses growth of the murine melanoma B16 tumor in vivo.

Get Disease/Agent associations (from CGI) Get gene annotations (from CGAP) Human

Figure 9-7 – Expanded listings for a gene from the Cancer Gene Index.

Each detailed record contains 3 additional pieces of information:

- **Role:** a curator-assigned description of the kind of association being reported. The values in this column come from a controlled vocabulary (developed to support the CGI database creation effort).
- **Sentence:** the actual article abstract sentence used to derive the reported gene-disease association. The full sentence is displayed in the text area at the bottom portion of the interface (it is also available as a tool tip text, by mousing over the "Sentence" column).
- **Pubmed:** the Pubmed ID of the source article. Clicking on the Pubmed link brings up in the web browser the corresponding Pubmed abstract page (in this example, the sentence used for deriving the gene-disease association is actually the paper title):

The screenshot shows the PubMed website interface. At the top, there is the NCBI logo and the PubMed logo with the URL www.pubmed.gov. Below this is a navigation bar with links for All Databases, PubMed, Nucleotide, Protein, Genome, and Structure. A search bar contains the text 'PubMed' and a 'Go' button. Below the search bar are buttons for Limits, Preview/Index, History, Clipboard, and Details. A display settings section shows 'AbstractPlus' selected, 'Show 20', 'Sort By', and 'Send to'. A status bar indicates 'All: 1' and 'Review: 0'. The main content area shows a search result for '1: Cancer Res. 1999 Oct 15;59(20):5059-63.' Below the result is a linkout box with the text 'Gene therapy with dominant-negative Stat3 suppresses growth of the murine melanoma B16 tumor in vivo.' and a list of authors: Niu G, Heller R, Catlett-Falcone R, Coppola D, Jaroszeski M, Dalton W, Jove R, Yu H. The text below the authors describes the research at the H. Lee Moffitt Cancer Center and Research Institute, University of South Florida. A 'Related articles' sidebar on the right lists related topics like 'Overexpression of transducing factor', 'Down-regulation of transcription factor', and 'Novel signaling pathway'.

Figure 9-8 – PubMed linkout from a Cancer Gene Index entry.

It is also possible to link out to the NCI thesaurus in order to see the definition of the disease being associated with a gene; this is achieved by right-clicking on the table row for the gene-disease association and selecting the "Link to NCI_Thesaurus" option from the popup:

The screenshot shows the NCI Thesaurus website. The browser address bar shows the URL http://ncit.ncl.nih.gov/NCIBrowser/ConceptReport.jsp?dictionary=NCI_Th. The page header includes the National Cancer Institute logo and the text 'National Cancer Institute' and 'U.S. National Institutes of Health | www.cancer.gov'. The main content area is titled 'Concept Details' and 'Bookmark this page'. It shows a table with the following data:

Identifiers:	
name	Melanoma
code	C3224

Below the table, there is a section titled 'Relationships to other concepts:' with a link to 'Disease_Has_Abnormal_Cell' and a link to 'Melanoma Cell'. The left sidebar contains a search box with 'Quick Search' and 'Advanced Search' tabs, a 'Max Results: 25' dropdown, and a 'Go!' button. Below the search box is a 'Concepts visited (during this session):' section with a dropdown menu showing 'Melanoma'. At the bottom of the sidebar are 'QUICK LINKS' for 'EVS HOME' and 'NCICB HOME'.

Figure 9-9 – NCI Thesaurus link-out from a Cancer Gene Index entry.

It should also be noted that the user interface allows ordering and filtering of the data (the latter can be very useful if there are many records being displayed):

- **Ordering:** the table rows can be sorted alphabetically by the contents of any column, by clicking on a column heading.
- **Filtering:** the drop down boxes that appear above the columns "Marker", "Gene", "Disease", and "Role" contain one value for each distinct entry within those columns. They can be used to select only records that contain the designated values. Of note is the drop-down associated with the "Disease" column:

Marker	Gene	Disease	Role	Sentence	Pubmed
39708_at	STAT3	breast cancers (1)	Gene_Associated_With_Disease	Gene therapy with dominant-ne...	10537273
39708_at	STAT3	hepatocellular carcinomas (1)	Gene_Product_Expressed_In_Tissue	Prolonged DEX treatment of mel...	12589944
39708_at	STAT3	leukemias (1)	Gene_Product_Expressed_in_Disease	Prolonged DEX treatment of mel...	12589944
39708_at	STAT3	melanoma (1)	Gene_Product_Has_Biochemical_Function	Interleukin-6-resistant melanom...	11442760
39708_at	STAT3	prostate carcinomas (1)	Gene_Product_Malfunction_Associated_...	The expression of dominant ne...	11463377
39708_at	STAT3	melanoma	Gene_Product_is_Pathway_Element	Significantly, melanoma cells un...	12370822
39708_at	STAT3	breast cancers(1)
39708_at	STAT3	hepatocellular carcinomas(1)
39708_at	STAT3	leukemias(1)
39708_at	STAT3	prostate carcinomas(1)

Figure 9-10 – Filtering on values in a column in the Cancer Gene Index.

The parentheses next to a disease name indicate how many distinct genes (among those included in the user query) are associated with this particular disease.

It should be noted that these numbers are calculated using only fetched records. As mentioned above, the first stage of the information retrieval will fetch at most 10 records per gene. The remaining records associated with a gene can be retrieved from the CGI server by right-clicking on a table row corresponding to the gene and by selecting the popup menu option "retrieve" all. After all gene-disease association records for a given gene have been fetched, the bold-face type of its associated probeset name is removed:

Annotations \ Pathway \ CancerGeneIndex											
Marker	Gene	Disease	Role	Sentence	Pubmed	Marker	Gene	Agent	Role	Sentence	Pubmed
39708_at	STAT3	tumor(353)	39708_at	STAT3	epo(39)
39708_at	STAT3	cancer(126)	39708_at	STAT3	ethanol(34)
39708_at	STAT3	cancers(119)	39708_at	STAT3	pma(28)
39708_at	STAT3	tumors(119)	39708_at	STAT3	genistein(18)
39708_at	STAT3	prostate ca...	39708_at	STAT3	curcumin(14)
39708_at	STAT3	melanoma(63)	39708_at	STAT3	sulindac(13)
39708_at	STAT3	myeloma(51)	39708_at	STAT3	wortmannin...
39708_at	STAT3	multiple mye...	39708_at	STAT3	piceatanno(...
39708_at	STAT3	breast canc...	39708_at	STAT3	atra(10)
39708_at	STAT3	metastasis(43)	39708_at	STAT3	tpa(9)
39708_at	STAT3	human panc...	39708_at	STAT3	h2o2(9)
39708_at	STAT3	leukemia(31)	39708_at	STAT3	tyrosine kin...
39708_at	STAT3	human pros...	39708_at	STAT3	rituximab(9)
39708_at	STAT3	prostate car...	39708_at	STAT3	rapamycin(8)
39708_at	STAT3	pancreatic c...	39708_at	STAT3	glucocorticol...
39708_at	STAT3	breast cardi...	39708_at	STAT3	ionomycin(8)
39708_at	STAT3	human hepa...	39708_at	STAT3	egcg(8)
39708_at	STAT3	gastric canc...	39708_at	STAT3	antisense oli...

Figure 9-11 – Result of retrieving all records for a gene in the Cancer Gene Index.

Finally, by clicking on the "Export" button at the bottom of the user interface, the contents of the gene-disease and the gene-compound tables displayed within the CancerGeneIndex tab can be exported as comma separated values text files for further analysis or/and visualization by spreadsheet software.

10 Sequence Retrieval

10.1 Overview

geWorkbench contains a number of modules that allow DNA or protein sequences to be visualized and analyzed. Sequences can be loaded from a local disk as a FASTA format file, or can be retrieved from a remote resource. Here we discuss retrieval of sequences from the network.

Once a set of sequences has been obtained, it can be used for several types of analysis in geWorkbench, including searching using known promoter motifs ([Promoter Analysis](#)), running [BLAST](#) searches, or looking for common motifs using [Pattern Discovery](#).

Nucleotide sequences are obtained directly from the UC Santa Cruz Golden Path database. Amino-acid sequences are retrieved from the European Bioinformatics Institute (EBI).

10.2 Prerequisites

- A microarray dataset must be loaded.
- An annotation file must be associated with the microarray dataset at the time it is loaded. At present, only Affymetrix-format annotation files can be read in. These files can be obtained for Affymetrix chip types from [affymetrix.com](#). For exact instructions, please see the geWorkbench FAQ page: [FAQ](#)

10.3 Example - retrieving sequences for a list of gene markers

10.3.1 Obtaining a set of markers

Sequences can be retrieved for any set of markers of interest. For this example we have loaded the tutorial data file BCell-100.exp and selected the last 10 markers into a new Marker Set:

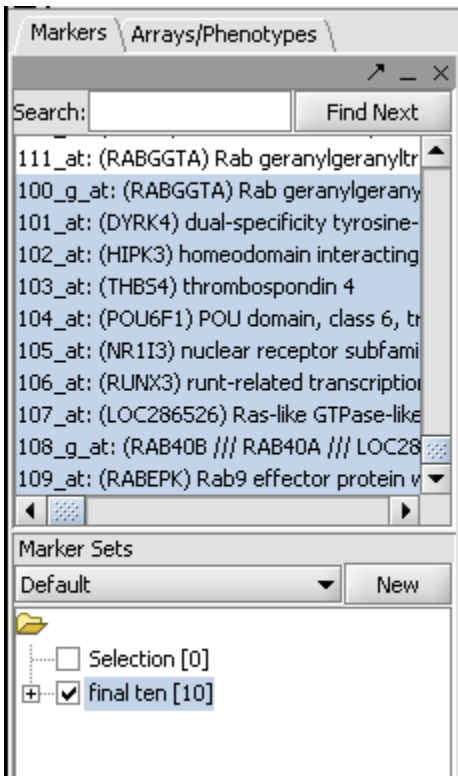


Figure 10-1 Activating a marker set for sequence retrieval.

When the set is activated (through use of the check box) the selected marker set will appear in the Sequence Retriever component:

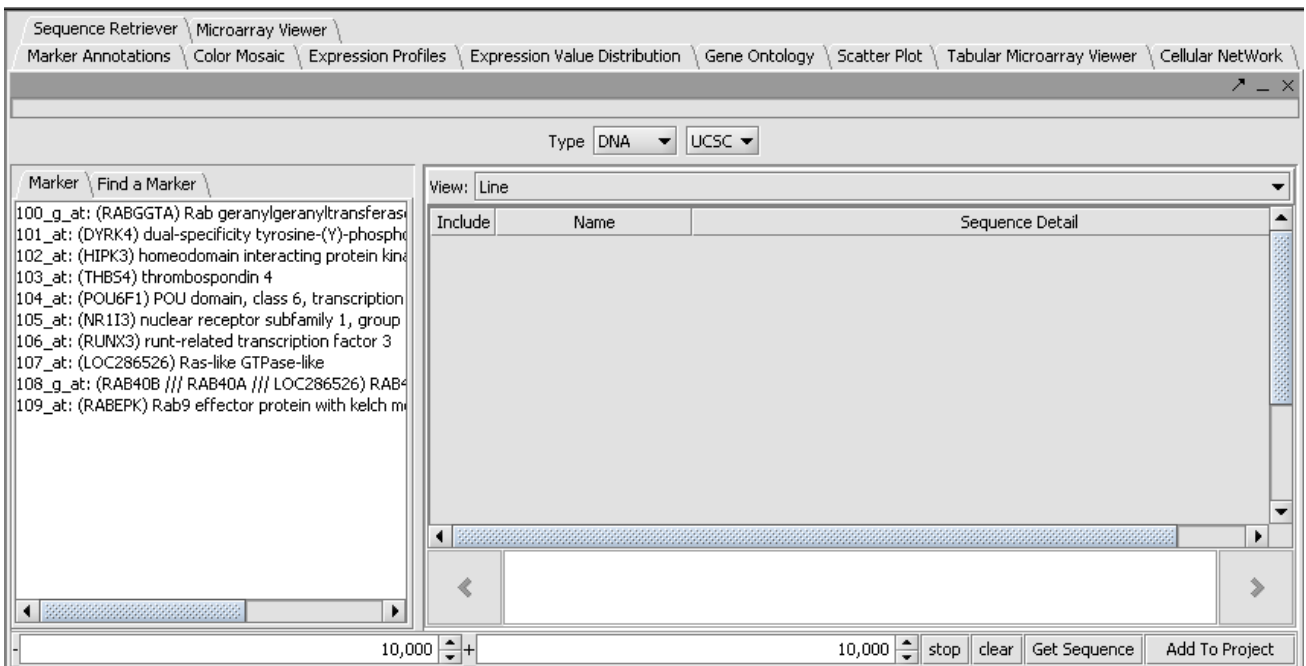


Figure 10-2 - Activated markers appear in the Sequence Retriever component

We will retrieve DNA sequences from Santa Cruz and leave the default settings of $\pm 10,000$ relative to the start of transcription. After pressing Get Sequence the sequences are downloaded:

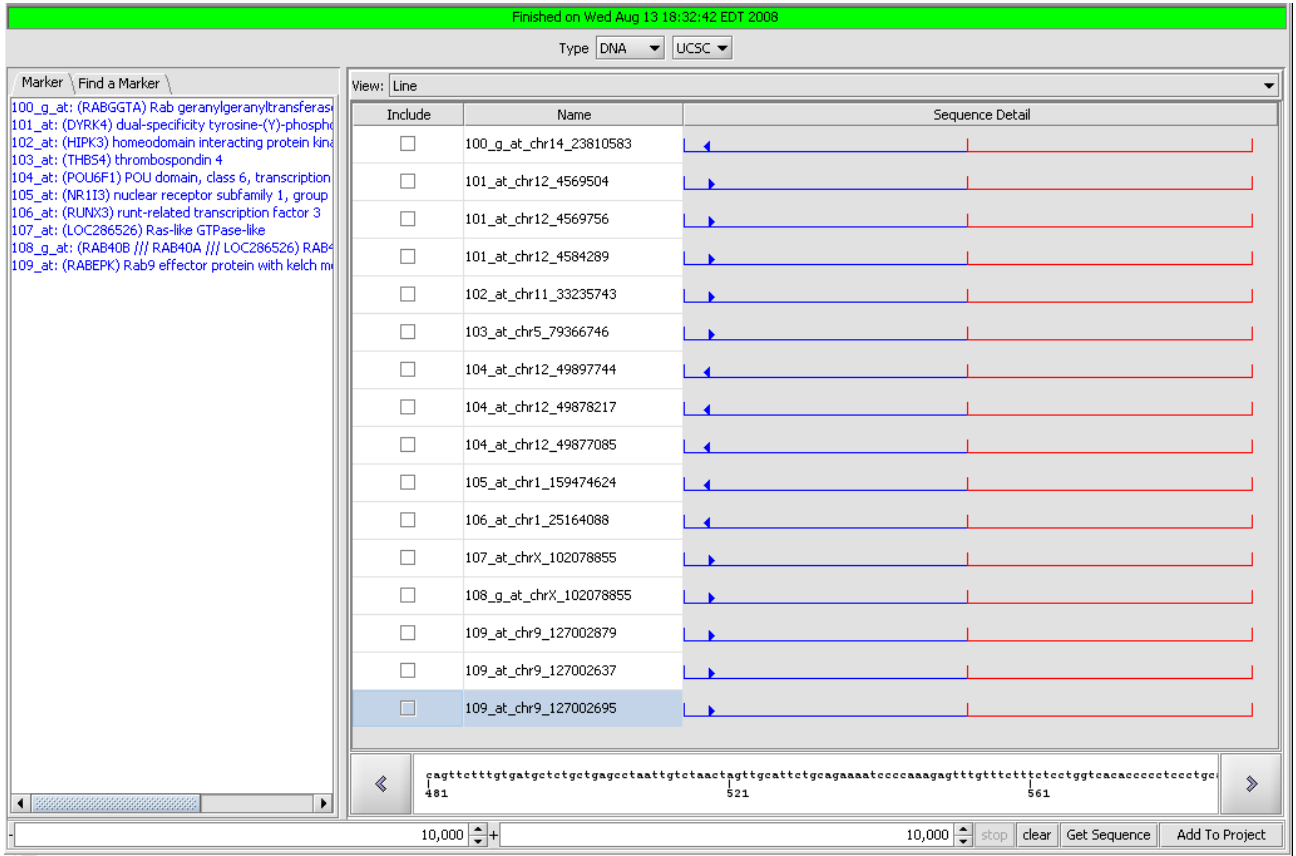


Figure 10-3 - Sequences retrieved for activated markers.

Note that for several of the markers more than one sequence has been retrieved. All sequences associated with a given gene symbol are retrieved.

Double-clicking on one of the lines shows the sequence detail:

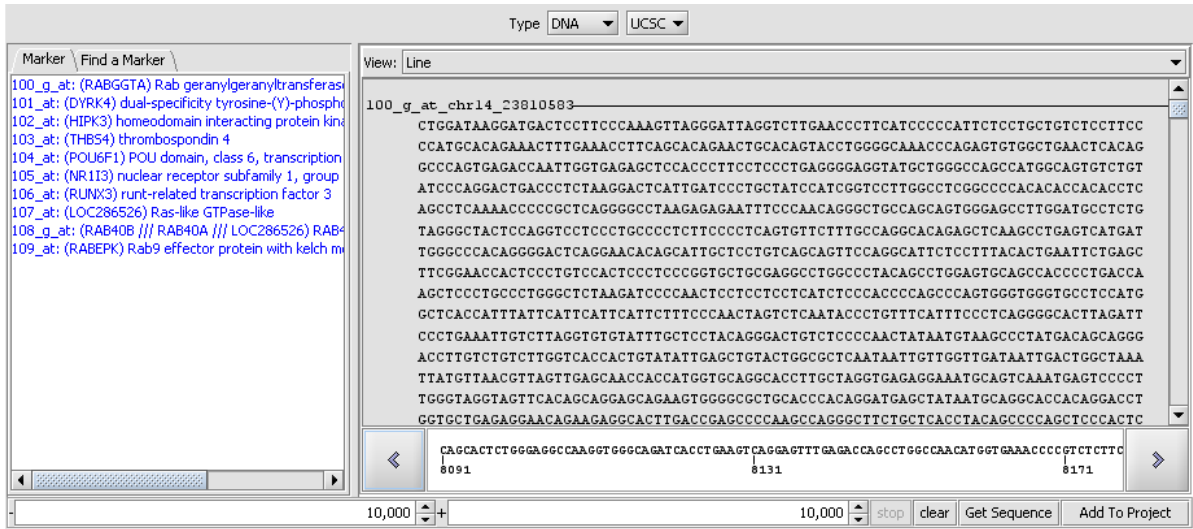


Figure 10-4 - Detailed sequence view.

The component provides check boxes which allow sequences of interest to be selected and added to the Project Folders component as a data node:

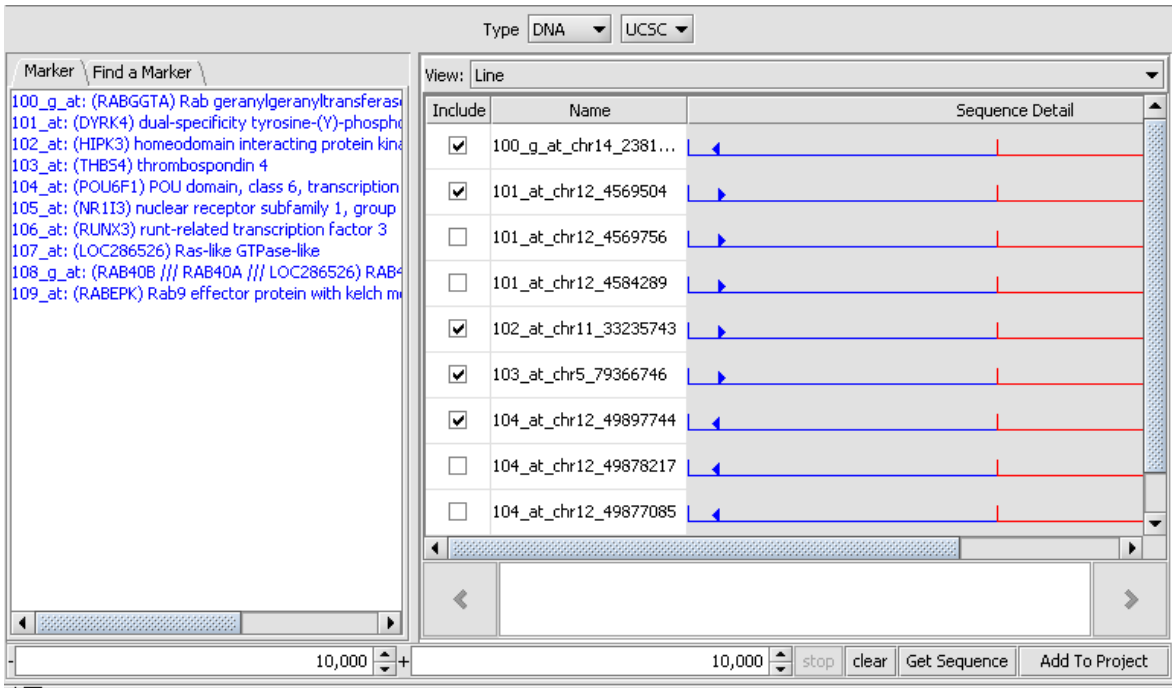


Figure 10-5 - Selecting sequences to add to the Project.

When Add to Project is pushed, the user is asked for a name for the new data node:

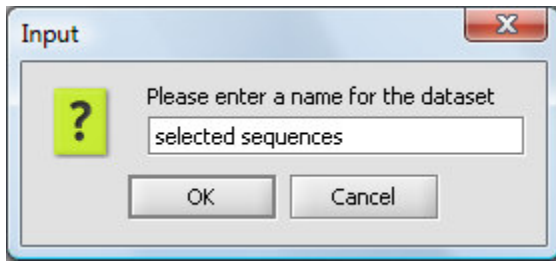


Figure 10-6 - Naming the new set of sequences.

The resulting node is placed into the Project Folder as a child of the original dataset:

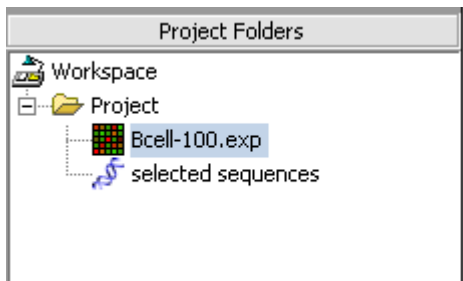


Figure 10-7 - The new sequence set added to the Project.

Note that when this node is added, the Viewing area of the geWorkbench GUI will now show components that support working with sequences. However, the Sequence Retrieval component will no longer be visible! You must select the Project or the sequence's parent object to see the Sequence Retrieval component again.

10.4 Saving the sequences to an external FASTA file

1. Right-click on the "Selected Sequences" entry you made in the Project Folders component.
2. Select **Save**.
3. Enter a suitable name and save the file.

11 Sequence Alignment

11.1 Overview

The comparison of sequences is central to the study of genes and genomes, and can illuminate function, regulation and evolution. For example, highly conserved protein coding sequences imply structural similarity and tend to be causally related to a common function. Therefore, in studying a gene or other sequence, it is important first to detect all significant similarities between the encoded amino acid or nucleic acid query sequence and any accumulated protein or nucleic acid sequence data, for example that contained in the national sequence databases.

11.2 BLAST

The BLAST algorithm is found in the Sequence Alignment tab, located in the command area of geWorkbench (lower right quadrant). The Sequence Alignment tab appears when a protein or DNA sequence is loaded and selected in the Project Folders component. BLAST is currently the only alignment option supported.

The BLAST algorithm is used to find similarities between nucleotide or amino acid query sequences and sequences held in a database. It is often used to give clues to the function of a sequence based on its similarity to already characterized sequences.

geWorkbench runs BLAST by submitting jobs to the NCBI server. NCBI-supported sequence databases and search algorithms can be selected in the user interface (arrows). There is no provision at this time for running a local BLAST job on the client desktop machine.

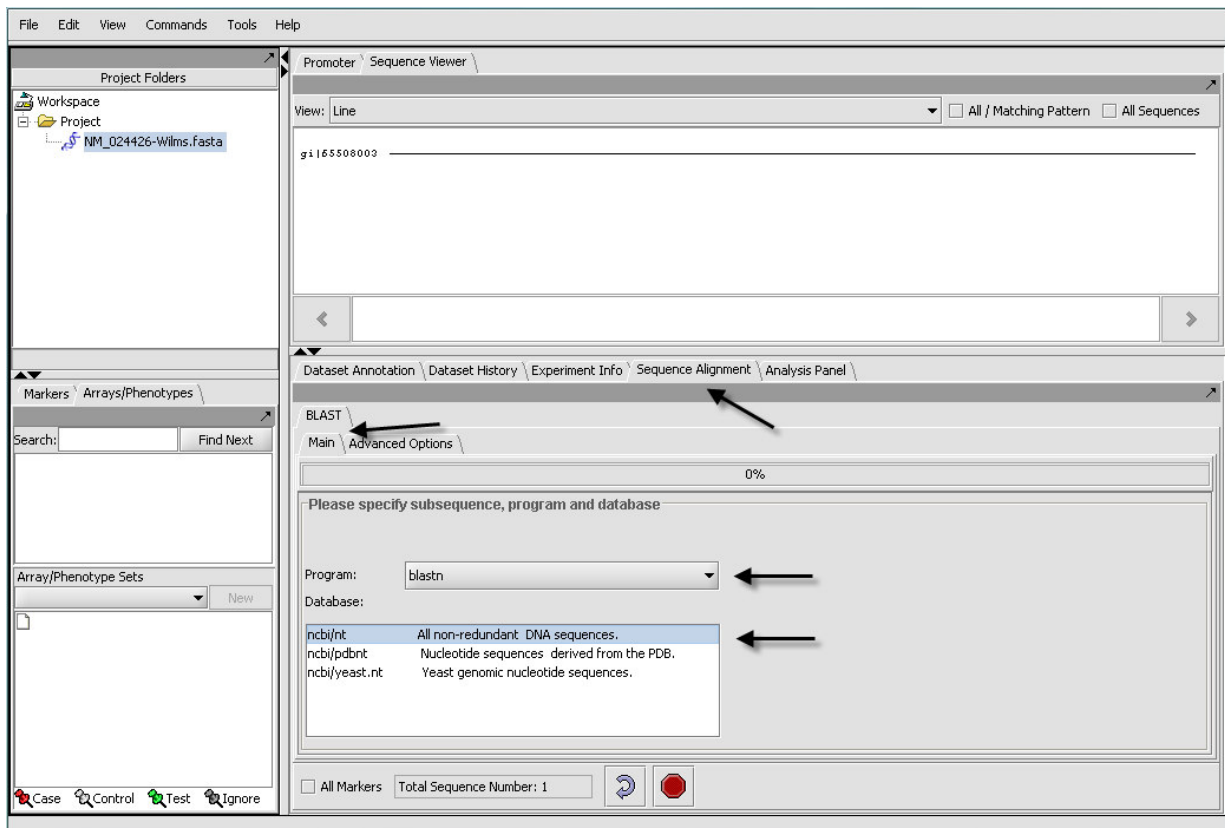


Figure 11-1 - The Sequence Alignment component showing the BLAST main tab.

11.3 BLAST job setup

11.3.1 Prerequisites

- The Sequence Alignment component must be loaded in the geWorkbench [Component Configuration Manager](#).
- A protein or nucleotide sequence must be loaded in the Project Folders component.

11.3.2 Query sequences

BLAST accepts nucleotide or amino-acid query sequences in the FASTA format. A query file can contain one or multiple sequences. The file can be loaded from disk using the File Open command, or may have been placed into the Project Folders component by another component such as the Sequence Retriever, or as a result of a previous BLAST run.

11.3.3 Parameters - Main

11.3.3.i Algorithms

The user must make sure that the algorithm chosen matches the type of query sequence (protein or nucleotide) that has been loaded. Some of the algorithms translate a nucleotide query, a nucleotide database, or both into amino acid sequence before executing the query. Searching in the amino-acid space is more sensitive for certain types of query, as it ignores synonymous, non-functional changes in nucleotide sequence.

11.3.3.i.1 For protein query sequences:

- **blastp** - Compares an amino acid query sequence against a protein sequence database.
- **tblastn** - Compares a amino acid query sequence against a nucleotide database translated in all reading frames.

11.3.3.i.2 For nucleotide query sequences:

- **blastn** - Compares a nucleotide query sequence against a nucleotide sequence database.
- **blastx** - Compares a nucleotide query sequence translated in all reading frames against a protein sequence database.
- **tblastx** - Compares the 6 frame translations of a nucleotide query sequence against the six frame translations of a nucleotide sequence database.

11.3.3.ii Databases

Standard protein and nucleic acid databases maintained at NCBI are supported. The appropriate databases for the search algorithm chosen will be displayed.

11.3.3.ii.1 For nucleic acids:

- **ncbi/nt** - all non-redundant DNA sequences.
- **ncbi/pdbnt** - nucleotide sequences derived from the PDB database (protein 3D structure database).
- **ncbi/yeast.nt** - yeast genonic sequences.

11.3.3.ii.2 For proteins:

- **ncbi/nr** - all non-redundant protein sequences
- **ncbi/pdbaa** - protein sequences from the PDB database (protein 3D structure database).
- **ncbi/swissprot** - sequences from Swiss-Prot, a primary reference database.
- **ncbi/yeast.aa** - translations of yeast genomic coding regions.

11.3.4 Parameters - Advanced Options

- **Expect** - Expected number of chance matches in a random model.
- **Word size** - The length of the seed that initiates an alignment.
- **Matrix** - Various scoring matrices are available for protein queries. A default matrix is used for DNA searches. The chosen matrix assigns a score for aligning any possible pair of residues. (A future release will include match/mismatch scores for DNA searches).
- **Gap Costs** - The pull-down menu shows the available choices of gap costs for the chosen matrix. Increasing the gap costs will tend to decrease the number of gaps in returned alignments. This is currently only implemented in geWorkbench for protein searches.
- **Low Complexity** - filter out low compositional complexity sequence.
- **Mask lower case** - filter out sequence which is in lower case in the FASTA query sequence.
- **Mask for lookup table only** - masks low-complexity sequence only while constructing the lookup table used by the initial hit-find phase of BLAST. The second phase, hit extension, is not affected and hits can be extended through low-complexity sequence. NCBI notes that this option is experimental and subject to change.
- **Human repeats filter** - masks human repeats (LINE's and SINE's). This option can speed searches involving long query sequences or databases containing sequences with many repeats.
- **Display result in your web browser** - geWorkbench will display the HTML page returned by NCBI BLAST in your web browser as well as within its own display.

Please see the [NCBI BLAST Help page 1](#) and [NCBI BLAST help page 2](#) for further details on these options.

Nucleotide Search parameters (default):



The screenshot displays the 'BLAST' application window, specifically the 'Advanced Options' tab for nucleotide searches. The 'Expect' parameter is set to 10, 'Word size' is 11, and the 'Matrix' is 'dna.mat'. The 'Low Complexity' and 'Display result in your web browser' options are checked, while 'Mask lower case', 'Mask for lookup table only', and 'Human Repeats Filter' are unchecked.

Figure 11-2 - BLAST Advanced Options (Nucleotide).

Protein Search parameters (default):



Figure 11-3 - BLAST Advanced Options (amino acid)

11.3.5 General controls

- **All Markers** - if selected, use all sequences loaded, overriding any activated sets in the Marker Sets component.
- **Total Sequence Number** - indicates how many sequences have been selected for query.
- **Curling arrow** - start BLAST search
- **Stop sign** - stop BLAST search (if pushed, geWorkbench will not wait for or retrieve the BLAST results).

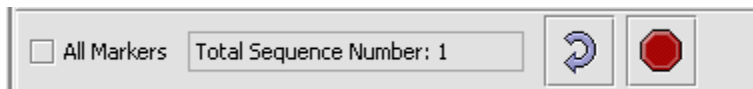


Figure 11-4 - BLAST job controls

11.4 BLAST Results Viewer

When the BLAST search results are returned they are placed in a new node in the Project Folders component as a child of the query sequence used. You can mouse over the result set to see how many sequences are in it.

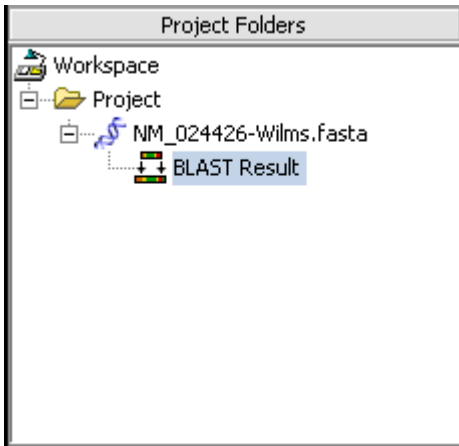


Figure 11-5 - BLAST result node in Project Folder

Each different hit is listed on a line in the results table, shown below. Note that a query sequence can hit a database target sequence in more than one place, resulting in multiple alignments displayed per target hit. The results viewer also shows statistics for each hit, including the E-value, start position and length of the hit, and the percent identity.

If the "Display result in your web browser" option was chosen, then the browser will open with the HTML formatted results.

In the pane at left in the picture below, the name of the input query sequence is shown, e.g. the gi number of a Genbank sequence. If there had been more than one query sequence, then this pane would show a list of query sequence names, allowing you to select the results to be viewed.

The screenshot shows the BLAST job result viewer. On the left, the search term 'gi|65508003' is entered. The main window displays a table of search results with columns: db, Name, Description, e-value, start point, align length, %ide..., and Include. Below the table, a detailed view of a hit is shown for 'ref|NM_024426.3| UEG Homo sapiens Wilms tumor 1 (WT1), transcript variant D, mRNA'. The detailed view includes the score (4828 bits), identities (3029/3029), and a sequence alignment between the query and subject.

db	Name	Description	e-value	start point	align length	%ide...	Include
ref	NM_024...	Homo sapiens Wilms tumor 1 (WT1), transcript...	0.0	1	3029	100	<input type="checkbox"/>
ref	NM_024...	Homo sapiens Wilms tumor 1 (WT1), transcript...	0.0	1	3020	99	<input type="checkbox"/>
ref	NM_024...	Homo sapiens Wilms tumor 1 (WT1), transcript...	0.0	1	2978	98	<input type="checkbox"/>
ref	NM_000...	Homo sapiens Wilms tumor 1 (WT1), transcript...	0.0	1	2969	98	<input type="checkbox"/>
emb	X51630.1	Human Wilms tumor WT1 mRNA for zinc finger prot...	0.0	109	2895	100	<input type="checkbox"/>
ref	XM_001...	PREDICTED: Macaca mulatta similar to Wilm...	0.0	458	53	97	<input type="checkbox"/>
ref	XM_001...	PREDICTED: Pan troglodytes Wilms tumor 1 ...	0.0	379	2387	99	<input type="checkbox"/>
gb	M30393.1	Human Wilms' tumor (WT33) protein mRNA, 3...	0.0	1	2308	100	<input type="checkbox"/>
dbj	AK09316...	Homo sapiens cDNA FLJ35849 fis, clone TESTI20...	0.0	282	2139	97	<input type="checkbox"/>
dbj	AK29173...	Homo sapiens cDNA FLJ77569 complete cds, high...	0.0	1	2114	99	<input type="checkbox"/>

>ref|NM_024426.3| UEG Homo sapiens Wilms tumor 1 (WT1), transcript variant D, mRNA
 GENE ID: 7490 WT1 | Wilms tumor 1 [Homo sapiens] (Over 100 PubMed links)

Score = 4828 bits (5354), Expect = 0.0
 Identities = 3029/3029 (100%), Gaps = 0/3029 (0%)
 Strand=Plus/Plus

```

Query 1      CCAGGCAGCTGGGGTAAGGAGTTCAAGGCAGCGCCACACCCGGGGGCTCTCCGCAACCC 60
            |||
Sbjct 1      CCAGGCAGCTGGGGTAAGGAGTTCAAGGCAGCGCCACACCCGGGGGCTCTCCGCAACCC 60
  
```

Buttons: Load, Reset, Select All, Add Selected Sequences to Project, Only Add Aligned Parts

Total hits for all sequences are 50.

Figure 11-6 - BLAST job result viewer.

11.4.1 Controls

11.4.1.i Within the list of returned hits

- **Include** check boxes - when checked, selects these sequences for import into the Project Folders component.

11.4.1.ii At the bottom of the pane

- **Load** - load a HTML format BLAST file into the viewer.
- **Select All** - mark as checked all the "Include" boxes.
- **Reset** - uncheck all "Include" boxes.
- **Add Selected Sequences to Project** - for each hit whose "Include" box is checked, add its sequence to a sequence node in the Project Folders component.
- **Only Add Aligned Part** - for each hit whose "Include" box is checked, add to the Project Folders component only the portion of its sequence which aligned with the query sequence .

11.4.1.iii In the query list

- **Search** - input a text search to find entries in the list of queries.
- **Find Next** - search for the next occurrence of the entered text.

11.5 Submitting a BLAST job

- Press the **curved arrow** submit button. The adjacent **Stop** button will terminate the search (geWorkbench will not wait for or retrieve the BLAST results).



Figure 11-7 - The submit and stop job buttons

- Once the search has been submitted, a progress bar in the "Main" tab will indicate first that the sequence is being uploaded and then that the job is running.

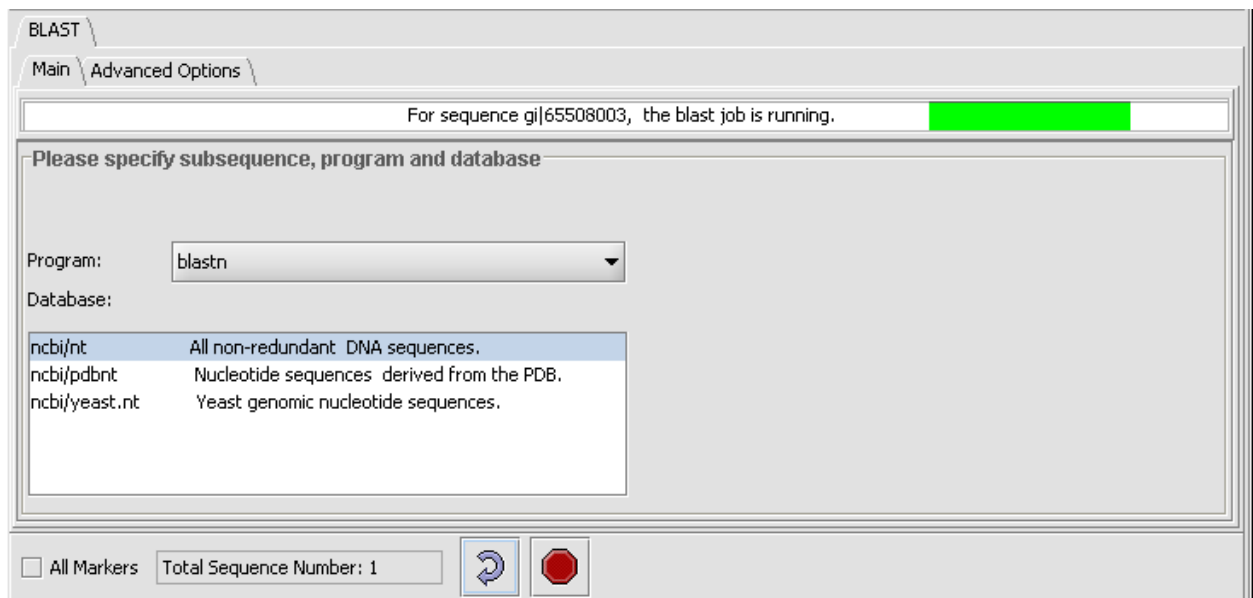


Figure 11-8 - A running BLAST job.

11.6 Example: Running a BLAST search

Two Genbank sequence files in FASTA format are included in the geWorkbench data/public_data folder: a nucleotide sequence, "NM_024426-Wilms.Fasta", and its protein sequence, "NP_077744-Wilms.fasta".

For a simple search using the nucleotide query file, one can select the blastn program and search against the ncbi/nt non-redundant database of nucleotide sequences.

- From the geWorkbench "File" menu, select "**Open->File**".
- Select a file type of FASTA.
- Navigate to data/public_data within the geWorkbench distribution and select the file "NM_024426-Wilms.Fasta".
- Press "**Open**".
- (The above steps can also be accomplished by right-clicking on a Project node and selecting "**Open File(s)**" and following the same steps.
- In the Project Folders component, make sure the sequence file just read in is selected. This will activate those components that can work with sequence data.
- In the Commands Area click on the **Sequence Alignment** tab.
- Select the **BLAST** tab, and under it the **Main** tab.
- For program select **blastn**.
- For database, select "**ncbi/nt** - the complete non-redundant nucleotide database. For a faster search, one could select the ncbi/pdbnt database instead, which is much smaller.

Note: The text field at the bottom of the Sequence Alignment component shows the number of sequences that have been selected. If you have a Fasta file that has multiple sequences, you can select the ones you want in the Markers component and activate this selection, letting you search on a subset. You can search on all sequences in a file by clicking the All Markers checkbox.

- Click on the **Advanced Options** Tab
- Change the **Expect Value** to 0.01. This sets the cutoff for which BLAST hits will be displayed.
- Make sure "dna mat" is selected for the **Matrix**.
- Leave the **Display result in your web browser** checked.

- Hit the "**curving arrow**" run button. The job will be submitted and the results returned as shown in the sections above.

11.7 References

BLAST

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402.

12 Pattern Discovery

12.1 Overview

Sequence Pattern Discovery is the process of identifying nucleotide or amino acid arrangements, also called motifs that are enriched in a set of sequences. Such motifs may identify regions that have been preserved by evolution and which therefore may play a key functional or structural role. geWorkbench currently provides three modes of Sequence Pattern Discovery: Regular Discovery, Hierarchical Discovery, and Exhaustive Discovery.

Regular Discovery is based on the algorithm SPLASH (Califano, A., 2000); it generates a list of all regular expression patterns (*motifs*) that satisfy a user-defined minimum support and a minimum density criteria. The former determines the minimum number of times a pattern must occur in the sequence set to be reported. This can also be expressed as the minimum percent of sequences that must contain the pattern. The latter determines how sparse the pattern can be, in other words the minimum number of matching characters k (any character except for the dot character “.”) over a window of predefined length w .

SPLASH-based motif discovery is extremely efficient and can process most large protein super-families in a few minutes on a conventional workstation. Discovery is uniquely effective in identifying sparse patterns using extremely low-density constraints, and the results obtained with Discovery can provide the core for a large number of more specific local alignments.

Exhaustive Discovery starts from a relatively high minimum support (e.g. patterns occurring in 75% of the sequences) and it progressively reduces the support, until a statistically significant pattern is discovered. Discovered patterns are reported and then masked in the sequence set so that they are no longer discovered. Then the process continues iteratively until the minimum support reaches a lower user-defined limit. Exhaustive Discovery, thus, produces a list of non-overlapping motifs in order of support.

Hierarchical Discovery (note – this feature is not available in geWorkbench 1.8) is based on the top-down clustering algorithm CASTOR (Liu and Califano, 2003); it proceeds similarly to Exhaustive discovery, except that each time a pattern is reported, the set is split into sequences containing and sequences not containing the pattern. Discovery continues hierarchically in the individual split subsets. This produces a binary tree of sequence sets and associated patterns. Discovery stops when the sets become smaller than a user-defined limit or when statistically significant patterns can no longer be discovered. In addition, HMM models are also generated.

12.2 Tutorial

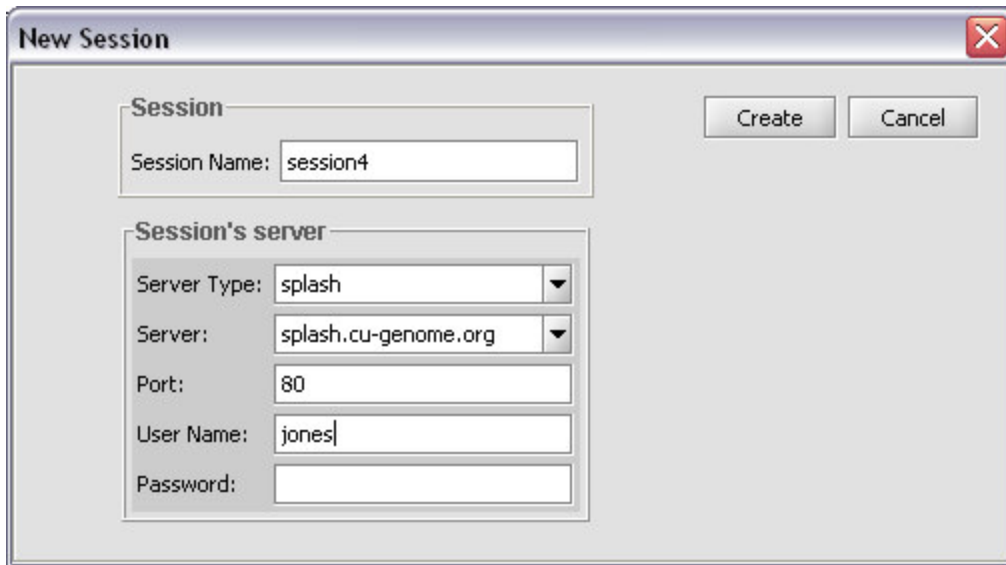
In this tutorial we will present examples of Normal, Hierarchical, and Exhaustive analyses. To demonstrate Normal, we will load a dataset of 254 amino acid sequences containing H1 histone sequences and attempt to discover a common motif in at least 95% of the

sequences. (This file, H1H5_HistoneDB_NHGRI.fasta is available for download from the geWorkbench website as part of the tutorial data). To demonstrate Hierarchical, and Exhaustive analyses, we will search the same dataset using different constraints. We will then describe three panels that can display the resulting patterns. These are Sequence, Position Histogram, and Patterns(Promoter).

12.2.1 Discovery analysis

In this example, we begin the Discovery process by loading the H1-histone sequence dataset from local files into the Project Folders component. After loading the Project Folders, you can see the list of sequence IDs for the dataset by clicking on the Markers tab in the Selection Area.. You can also see a graphic display of the dataset by opening the Sequence component in the Visual Area. In this view each sequence is represented by a line proportional to its length, preceded by the sequence ID. Now open the Pattern Discovery component in the Analysis area. In the default view, you will see the Normal radio button has been selected and the Basic sub-panel is open with default parameter values. Change the settings in the text boxes to Support: 80%, Min. Tokens: 7, Density Window: 12, and Density Tokens: 4. Next, open the Advanced sub-panel tab and uncheck Exact Only to activate the BLOSUM50 similarity matrix.

To begin the Discovery process, press Execute icon . This brings up a New Session dialog box.



The image shows a 'New Session' dialog box with the following fields and controls:

- Session:** Session Name: session4
- Session's server:**
 - Server Type: splash
 - Server: splash.cu-genome.org
 - Port: 80
 - User Name: jones
 - Password: (empty)
- Buttons:** Create, Cancel

Figure 12-1 Pattern Discovery dialog box

Enter the following: **Server: splash.cu-genome.org, port: 80.** For the session and User Name, you can enter any convenient values (if in the future a login is required to access remote servers, then a valid User Name and Password will need to be entered). Note, subsequent searches in the same session will not elicit the dialog box. Then press **Create**.

Looking back at the **Pattern Discovery** panel, you will see the following series of progress bar text messages: **Uploading, Processing seeds, Discovering, Collating, and Done**. The **Discovery Table** will then fill with information on the discovered motifs. The table in Figure 12-2 shows that seven similar patterns were found. If you select a motif, it will be highlighted in blue. Here, the motif `[NDE][RK].G.S...[ILM].[RK].[ILMV]` was selected. The table shows that it is found once in each of 209 of the 254 input sequences, spans 14 tokens and contains 7 full character tokens. Right-clicking on the table elicits a pop-menu, described in Section .12.4

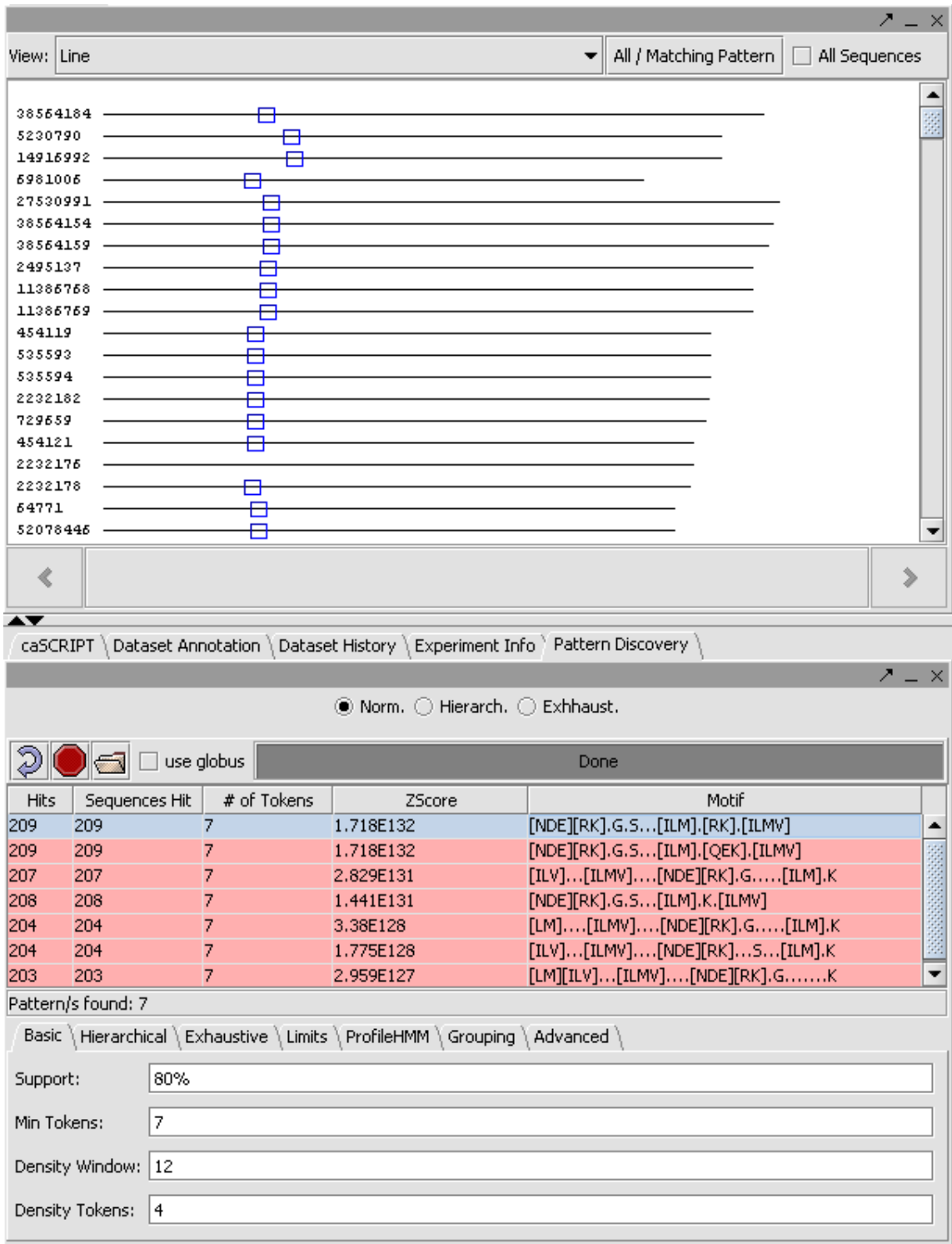



Figure 12-2 Normal pattern discovery

12.2.2 Hierarchical analysis

For an explanation of hierarchical analysis see [“http://www.research.ibm.com/splash/Hyerarchical/HierarchicalDiscovery.htm”](http://www.research.ibm.com/splash/Hyerarchical/HierarchicalDiscovery.htm).

To perform Hierarchical analysis on the histone dataset:

1. Select the Hierarc radio button.
2. Then set the Basic constraints as in the Normal example:
3. Leave the Advanced constraints as they were.
4. In the Hierarchical sub-panel, further options can be adjusted, but we will use the default values (Min. Cluster Size: 10, and Min. Pattern Number: 10).
5. Press the Execute icon  to initiate the search (Note – this search may take a long time). A gray progress bar is displayed while the search is in progress.
6. Results: Motifs and their frequency are listed in the Display Area as nodes in expandable folders (Figure 12-3).

The primary node shows the total number of sequences (254) in the dataset. Expanded folders show motifs and number. Folders expand until hierarchical constraints are reached.

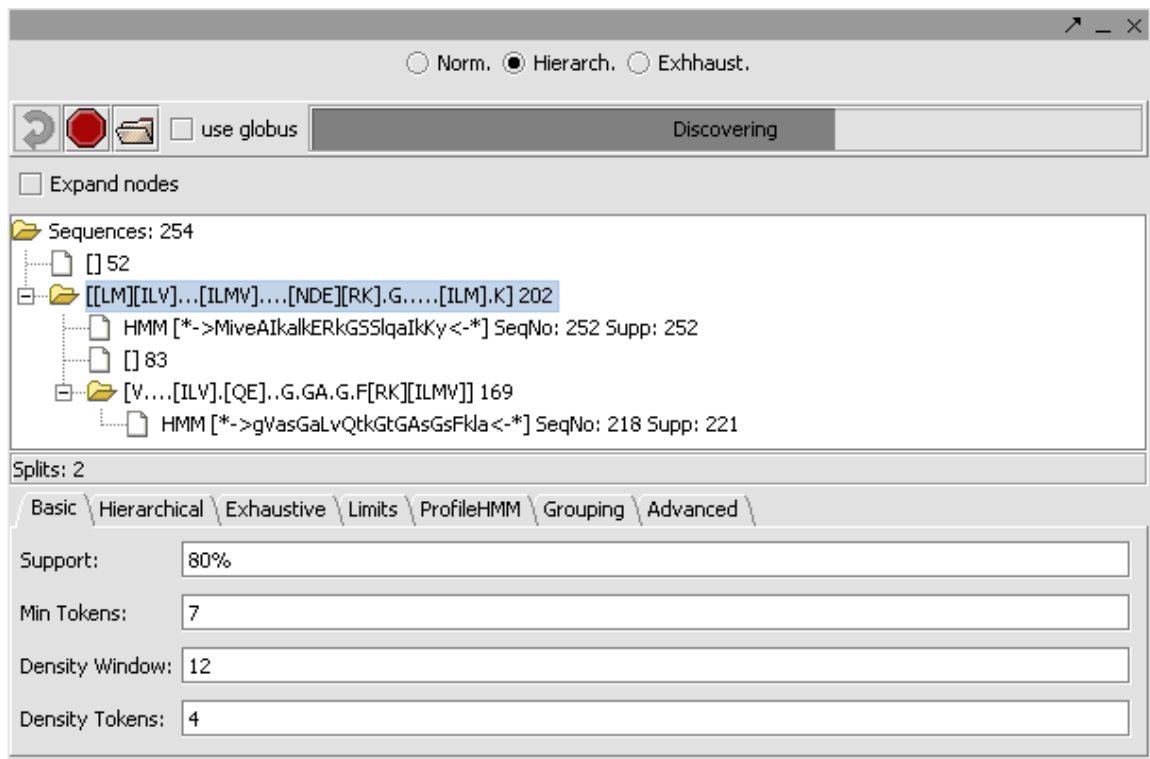


Figure 12-3 Hierarchical pattern discovery

12.2.3 Exhaustive analysis

In this third example, Exhaustive analysis was applied for Pattern Discovery on the same dataset, using the same Basic and Advanced constraints as in the hierarchical example. However, it is possible to set additional constraints, specified in the Exhaustive sub-panel. Here, we left the default parameters for Dec. Support(%) at 5 and Min Support at 10% (Figure 12-4).

The screenshot shows the caSCRIPT software interface. The top panel displays a list of sequences with highlighted motifs. The bottom panel shows the 'Exhaustive' sub-panel with a table of discovered patterns and their parameters.

Hits	Sequences Hit	# of Tokens	ZScore	Motif
202	202	8	1E300	[LM][ILV]...[ILMV].....
173	173	9	1E300	[ILV].Q..G.GA.G.[FY]...

Pattern/s found: 2

Basic \ Hierarchical \ Exhaustive \ Limits \ ProfileHMM \ Grouping \ Advanced \

Dec. support (%): 5 Min. Support: 10%

Dec. density support: 0 Min. Pattern Number:

By occurrence: occurrence

By sequence: sequence

Figure 12-4 Exhaustive pattern discovery in progress, two non-overlapping results highlighted and displayed in the Sequence component.

12.3 Visualization of Pattern Discovery Results

Various aspects of selected motifs can be effectively displayed in the **Sequence**, **Position Histogram**, and **Promoter** components in the **Visual Area**. This is applicable to all three of the **Discovery** modes. To display the motifs in the **Position Histogram** as in Figure 12-5 select the motifs as in Figure 12-2, press the **Position Histogram** tab, and press **Plot Position**. The same motifs are displayed as in **Sequence** component shown above (Figure 12-4). Notice, each motif is represented by a specific color and these colors are the same in both displays..

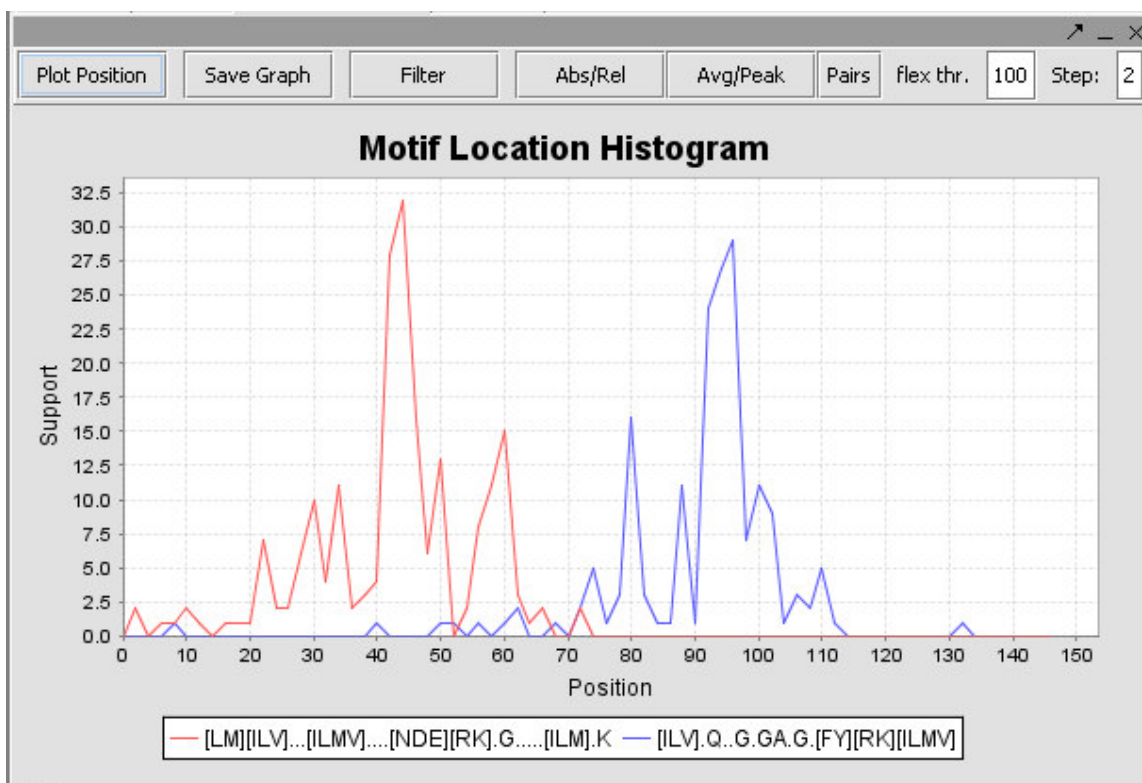


Figure 12-5 Position Histogram

12.4 Component Visual Elements

This section describes the various visual elements of the **Pattern Discovery** component (see the [SPLASH](#) page at IBM for further details).

Display Area: this is where input datasets and search results are shown either in graphical or textual representation.

Normal: this radio button implements Normal Pattern Discovery.

Hierarc: this radio button implements Hierarchical analysis.

Exhaustive: this button implements Exhaustive analysis.



Execute: this command button brings up **New Session** dialog box to start **Discovery**.



Stop: this command button stops a search in progress.



(Load): this command button loads **Discovery** results from a local file.

Progress Bar: in the narrow band above the **Display Area**, an animated slider pulsates back and forth as long as the search is ongoing.

The following are columns in the **Sequence Discovery Table**:

Hits: this is the total number of times a motif appears in a sequence dataset.

Sequences Hit: this is the total number of different sequences in which the motif is found.

of Tokens: this is the number of full-character tokens in the motif.

Zscore: this is a measure of how often the motif would be found in a random set of sequences of the same size and composition as the project dataset.

Motif: this is a sequence of tokens, which may be full character or wildcard. Periods (...) correspond to wild cards. Parentheses identify “either/or” tokens that satisfy the BLOSUM matrix.

The following are elements in the pop-up menu, elicited by right-clicking on the **Discovery Table**:

Mask Pattern: this menu item filters sequences containing selected motifs, so that they are not re-discovered or displayed in the **Sequence Panel** when **Discovery** is executed.

Unmask all Patterns: this menu item removes all the masks applied by **Mask Pattern**.

Save Patterns (Regex Only): this menu item saves the sequences of the motifs to a local file.

Save Selected Patterns: this menu item saves selected motifs from the **Discovery Table** to a local file. The saved table can be re-loaded into **geWorkbench** by pressing the **Load** button

Save All Patterns: this menu item saves the complete **Discovery Table** to a local file. The saved table can be re-loaded into **geWorkbench** by pressing the **Load** button

Add Patterns to Project: this menu item saves the results in the **Discovery Table** as a node in the **Project Folder**.

The following are visual elements in the **Basic** subpanel:

Basic: this tab opens a subpanel for defining motif parameters.

Support: this sets the minimum number or % of sequences in the set containing the shared motif. Type in a % sign to indicate %, e.g. 80% or 80 sequences (no percent sign).

Min. Tokens: sets the minimum number of density tokens in the density window.

Density window: is a window within the motif that counts tokens and wild cards.

Density tokens: are the full character tokens (not wildcards) in the density window.

The following are visual elements in the **Hierarchical** subpanel:

Hierarchical: this tab opens a subpanel for setting **Hierarchical** specific parameters,

Min. Cluster Size: this sets a lower limit to number of sequences that will be searched for a shared motif.

Min. Pattern Number: this sets a lower limit to the number of sequences that must contain a shared motif to be included in the hierarchy.

The following are parameters in the **Exhaustive** subpanel:

Dec support (%):this sets the size of intervals by which support level is decremented in successive searches (default is 5).

Min Support: this sets the lower limit on the percentage of sequences that must contain a specific motif (default is 10%).

Dec. density support: INACTIVE

Minimum Pattern Number: this sets a lower limit on the number of motifs in a cluster.

By occurrence: INACTIVE

By sequence: INACTIVE

The following are visual elements in the **Limits** subpanel:

Limits: this tab opens a subpanel for setting maximum pattern number and run time.

Max Pattern Number: this limits the number of patterns to discover.

Max Run Time (sec): this limits search time.

The following are visual elements in the **ProfileHMM** subpanel:

ProfileHMM: this tab opens a subpanel for setting parameters for profile-hidden Markov models (pHMMs).

Entropy Threshold: N/A

Conserved Region Extension: bases on either side of conserved region to be considered.

Sliding Window Size: bases considered in sequence being searched.

The following are visual elements in the **Grouping** subpanel:

Grouping: this tab opens a subpanel for setting Grouping parameters.

Type: feature is disabled.

Size: N/A

The following are visual elements in the **Advanced** subpanel:

Advanced: this tab opens a subpanel for setting **Advanced** parameters.

Exact Only: this check box, if unchecked, activates use of BLOSUM matrices.

Count sequences: this check box allows you to sort patterns by number of occurrences, number of distinct sequences in which they occur, length, or Zscore.

ZScore: calculate and use the ZScore (which is a measure of the statistical significance) to filter patterns to display,

BLOSUM50: this is the default substitution, or similarity, matrix used for polypeptide motif discovery; others available in the scroll panel are **BLOSUM100** and **BLOSUM150**

Similarity Threshold: a measure of the stringency of the search.

Minimum ZScore: **minimum value of ZScore for pattern to be significant.**

12.5 References

Califano A. SPLASH: structural pattern localization analysis by sequential histograms. *Bioinformatics* **16**:341-57 (2000).

13 Promoter Analysis

13.1 Overview

The Promoter component scans one or more sequence profiles against nucleotide sequences that the user has loaded into geWorkbench. Motifs from the JASPAR database of transcription factor binding sites are included with the component. Additional motifs can be added by the user.

The Promoter component will also display the results of hits found in the Pattern Discovery component.

13.2 JASPAR CORE database

The promoter component of geWorkbench includes version 3.0 of the JASPAR CORE database (<http://jaspar.genereg.net/>). It contains 138 curated, non-redundant profiles. These "profiles are derived from published collections of experimentally defined transcription factor binding sites for multi-cellular eukaryotes. The database represents a curated collection of target sequences" ([JASPAR Documentation](#)).

The datafile used "MATRIX_DATA.txt" can be found at http://jaspar.genereg.net/html/DOWNLOAD/mysql/JASPAR_CORE_2008/.

The profiles represent counts of how many times each of the four nucleotide bases occurs at a particular position in the aligned promoter sequences.

13.3 Working with the Promoter graphical interface

13.3.1 Prerequisites

- To use the Promoter component, first check that it has been loaded in the Component Configuration Manager.
- The Promoter component appears when a data node of type "sequence" has been loaded in the Project Folders component.
- The Promoter component appears in the upper-right quadrant of geWorkbench, in the "Visual Area".

13.3.2 Layout

The figure shows the display of profile "TBP:TATA-box:MA0108" from the list of those included in JASPAR. The main features of the component include

- The **TF Mapping** tab at left. This area allows profiles to be searched for and selected for use in a sequence scan.
- The **Logo**, **Parameters** and **Sequence** tabs at right. These provide respectively visual display of the profile, the parameters, and the scan results.
- Control buttons at bottom left to manage scans and results.

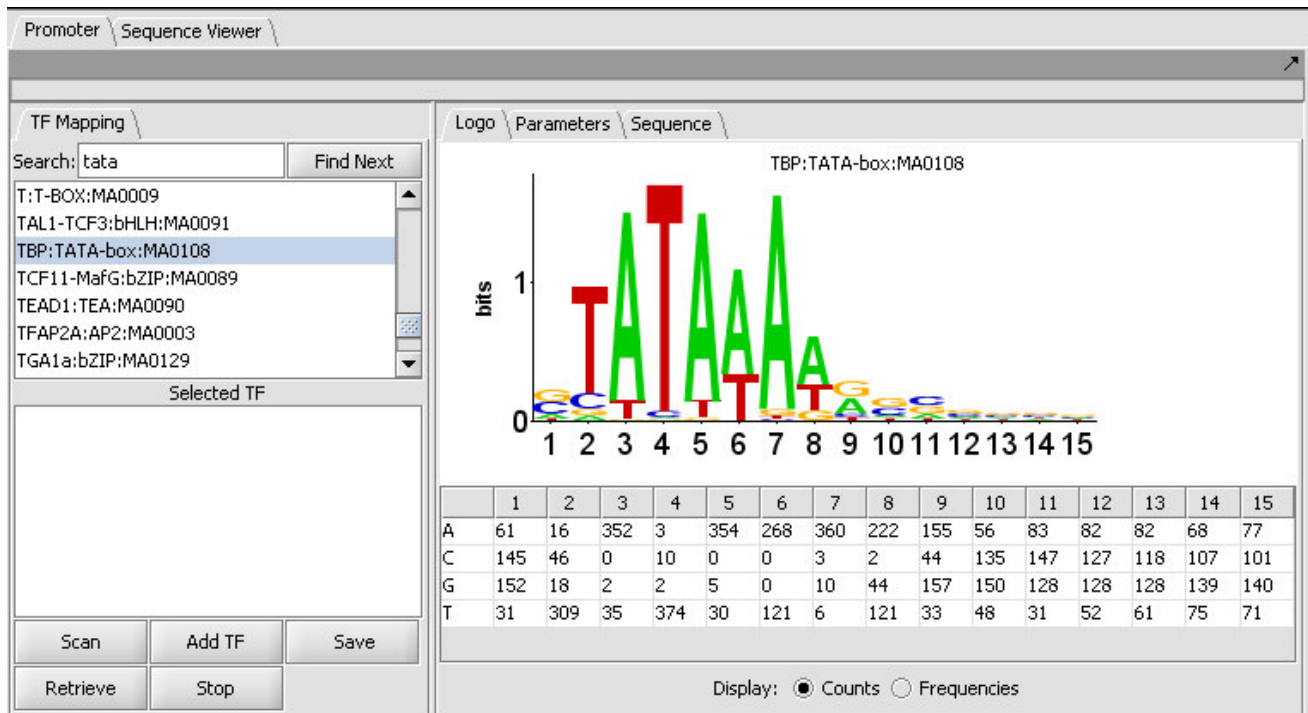


Figure 13-1 - The Promoter component showing the Logo display.

Further details of each part of the component are provided below.

13.3.3 TF Mapping tab

13.3.3.i List of available transcription factors

TF list - The TF list contains the names of available transcription factor binding site sequence profiles from JASPAR and those that have been loaded from files using Add TF. Double-click on TF signatures in the TF list box (upper box) to add TF's to the Selected TF (lower box). This lower list displays transcription factors which can be searched against the available genomic sequences by clicking on Scan. Double-clicking on a TF name clears it from the Selected TF list and returns it to the TF List.

Multiple profiles can be moved to the selected list for scanning.

- **Search** - Enter a portion of a TF name in the Search text box. The list scrolls to and highlights the TF containing the text string. Click **Find Next** to highlight the TF containing the text string. If nothing matches the text string, the entered text is changed to red.
- **Find Next** - continue scanning the list for the next occurrence of the search string.

13.3.3.ii Selected TF

This list contains the profiles that will be used in the next scan. Entries on the Selected TF list can be moved back up to the "available TFs" list by again double-clicking on their entry.

Here the TATA-box entry has been moved to the selected list.

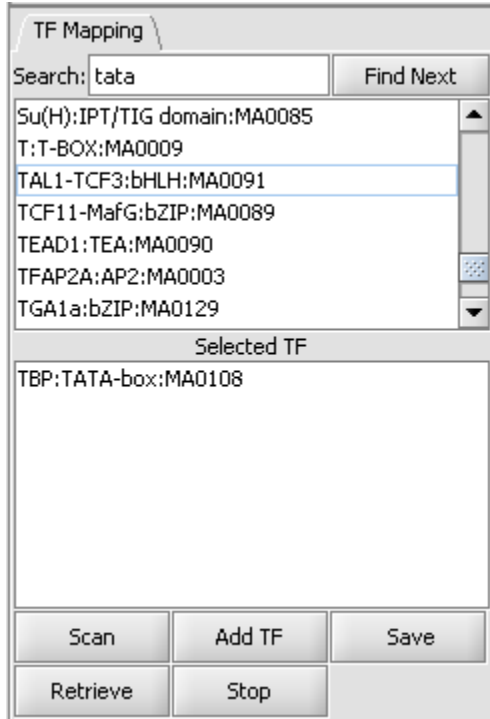


Figure 13-2 - Searching for transcription factor motif by name.

13.3.3.iii Controls

- **Scan** - Scans the sequences in the Selected TF list against the available genomic sequences. If the Selected TF list is empty, the system displays an error message.
- **Add TF** - load a new profile from a file into the list of available profiles. This is not a permanent addition; it remains loaded only for the current invocation of geWorkbench. See the "Profile File Format" entry below for details.
- **Save** - Saves to file a list of hits by a profile to a nucleotide sequence, including the sequence identifier, the transcription factor name and the start and stop points of the match along the sequence, as shown here:

gil65508003

ATHB5:HOME0-ZIP:MA0110 2104 2113

ATHB5:HOME0-ZIP:MA0110 2115 2106

ATHB5:HOME0-ZIP:MA0110 2882 2891

ATHB5:HOME0-ZIP:MA0110 2893 2884

- **Retrieve** - Not implemented.
- **Stop** - Stop the current scan.

13.3.3.iv Profile file format

A profile in the form of a count matrix can be loaded from an external file. The profile should consist of a tab delimited series of counts, one for each position in the profile. It should consist of four lines, in the order A, C, G, T. There are no header lines or row labels, just the numeric matrix. For example, here is a profile showing the first six columns:

```
0 12 0 0 1 0
```

```
49 0 20 23 3 45
```

```
0 37 29 2 45 4
```

```
0 0 0 25 0 0
```

Because the normalization step uses the count total (sequences aligned to generate the profile), loading a frequency matrix is not currently supported.

13.3.4 The LOGO tab

13.3.4.i LOGO display

The LOGO display implements the method of Schneider and Stephens (1990) to display the information at each position in a motif. Briefly, the total height of the column of letters at a position shows the information available, on a scale of 0 to 2 bits (the information needed to represent the 4 possible nucleotide bases at each position). The relative heights of each letter in a column show their individual contribution to the information at that position.

The LOGO display in geWorkbench implements the "small sample correction" described by Schneider, the magnitude of which depends on the number of sequences aligned to generate the profile. The correction is subtracted from the calculated information content at each position, with a minimum value (floor) of zero being displayed.

13.3.4.ii Table display

A table is used to show the numeric data from which the LOGO diagram is generated. The table depicts each position in the profile as a column, and has a row for each of the four nucleotide bases A, C, G and T. The user can choose to display the data either as the original counts or as frequencies.

- Display: Counts or Frequencies

13.3.5 The Parameters tab

Logo Parameters Sequence

Parameters:

PValue / 1K:

Use Thr. 13K Set

Iterations:

Pseudocount: sqrt(n)

Results:

	Total hits	Sequences with hits
Expected:	<input type="text" value="0"/>	<input type="text" value="0"/>
Actual:	<input type="text" value="0"/>	<input type="text" value="0"/>
Enrich. p-value:	<input type="text" value="1.0"/>	<input type="text" value="1.0"/>
% with hits:		<input type="text" value="0%"/>

	5' hits	3' hits
Expected:	<input type="text" value="0"/>	<input type="text" value="0"/>
Actual:	<input type="text" value="0"/>	<input type="text" value="0"/>

Figure 13-3 - The Promoter motif search parameters tab.

13.3.5.i Background sequence and scoring threshold determination

A background sequence is used to estimate an appropriate scoring threshold. This background can be generated in two ways.

1. determine base composition of input sequence and from this generate random sequence.
 2. (13K) - use a set of 13,000 promoter sequences as background.
- The length of background sequence scanned is given by $1000 * \text{Iterations} / \text{PValue}$.
 - The threshold value is calculated by scanning the background sequence with the profile and finding the top 100 scores. The 100th score is used as the threshold.
 - Calculated p-values are Bonferroni corrected and also corrected for duplicates in the list of 100 top scores.

- Positive and negative strands are scanned and values above threshold are reported.

13.3.5.ii Parameters

- PValue / 1K -
- Use Thr. - Use threshold - if checked, use a user-input threshold rather than a calculated threshold for scoring a match.
- 13K Set - If not checked (default), use the random background sequence described above. If checked, use the 13K sequences as background.
- Iterations -
- Pseudocount - a small-sample correction factor (default 1.0). See above description.

13.3.5.iii Results

13.3.5.iii.1 Total hits and Sequences with hits

Total hits counts all hits regardless of how many times one sequence is hit.

- Expected - number of hits expected by chance.
- Actual - observed number of hits.
- Enrich. p-value - p-value for chance of getting this outcome by chance.
- % with hits

13.3.5.iii.2 5' hits and 3' hits

- Expected - Expected number of hits
- Actual - Actual number of hits.

13.3.6 The Sequence tab

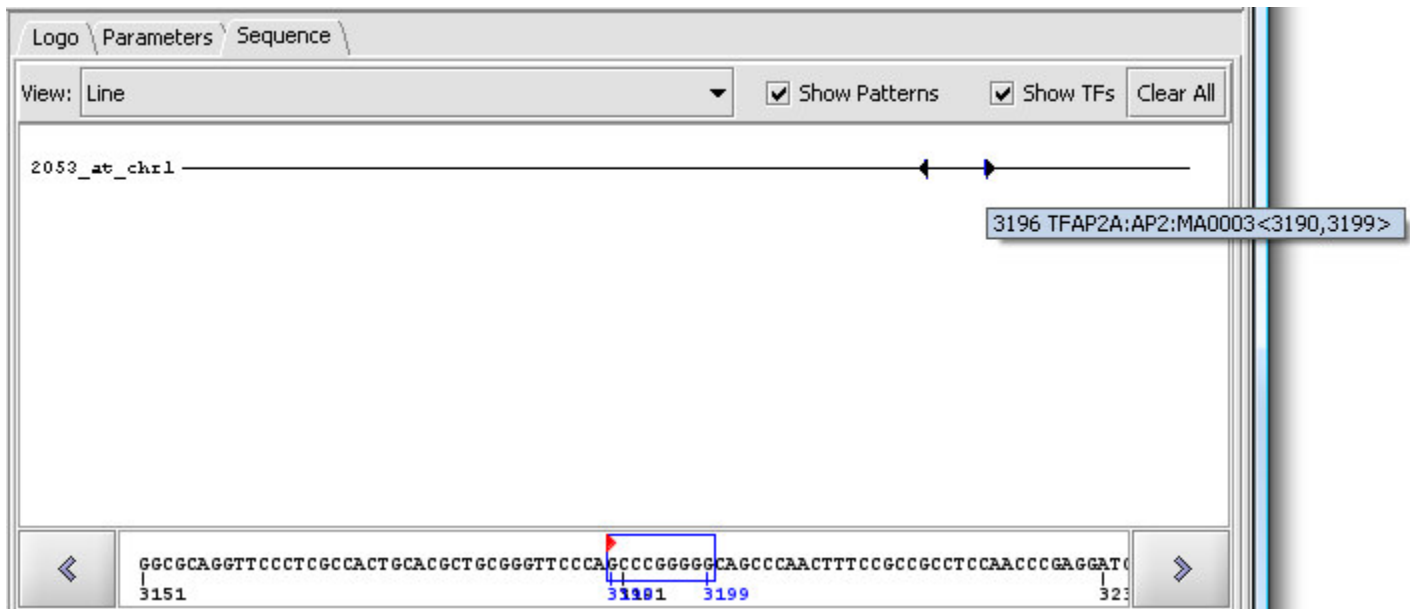


Figure 13-4 - Viewing motif hits on a sequence.

The Sequence tab can display either a line or a full character representation of the sequence which was searched against. Clicking on a position along the line or character representation will cause that portion of the sequence to be displayed in the detail box at the bottom of the component. This box also displays numbers representing position along the sequence, relative to the start of that particular sequence (not its genomic location). Both the character and detail views will show the location and extent of any profile match to the sequence.

If matches are found, the sequence will include blocks in various colors with solid arrows indicating the match orientation (forward or reverse complement). Individual hits can be identified by positioning the mouse pointer over them, which will display a tool tip. Clicking on an area with a match will show it in the Sequence Detail at the bottom with the hits shown as boxes around the characters.

The tooltip format is as follows: numeric position, Transcription Factor name <numeric position of the first character of the pattern, numeric position of last matching pattern character>.

- **View - Line or Full Sequence.** Line represents the sequence as a simple line, with any hits positioned along it. Full shows the entire sequence as characters.
- **Show Patterns** - display hits from Pattern Discovery (this is a separate component, not part of the Promoter component). Implementation note - these hits are represented in the "Active Patterns" data-structure.
- **Show TFs** - show hits from a search in this component. Implementation note - these hits are represented in the "Active TFs" data-structure.
- **Clear All** - clear all hits from the sequence window (and from the associated data structures). Note this will also clear the two adjacent check boxes. The relevant box must be re-checked to see further results.

The full character display shows any hits in white with a red background, and a small red arrow marks the start of the match and its direction.

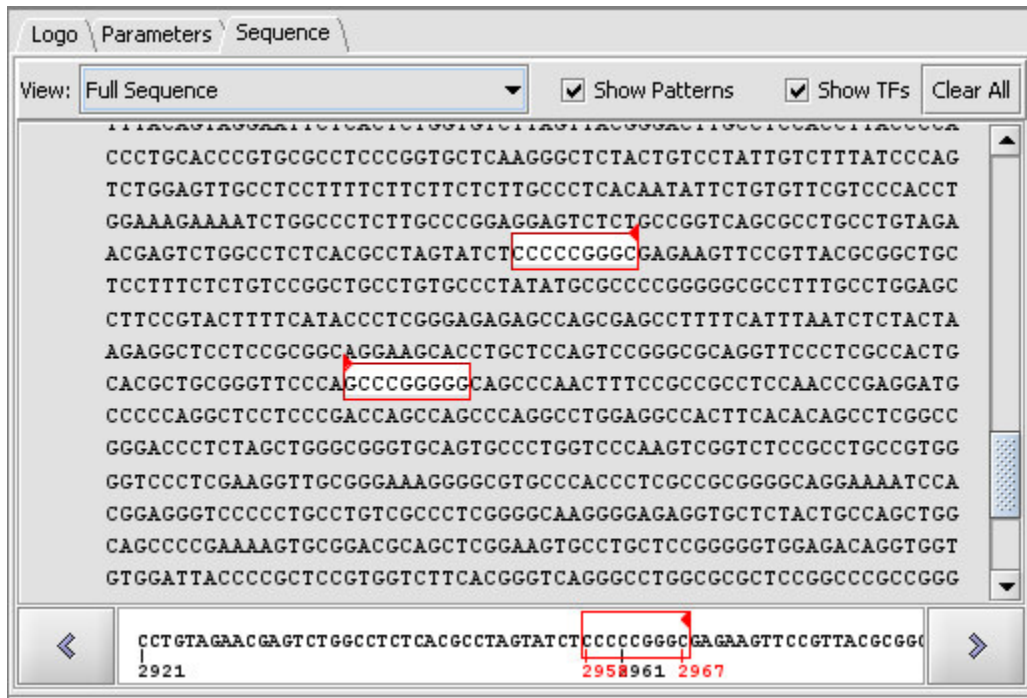


Figure 13-5 - Viewing motif hits on the full sequence display

13.3.7 Implementation details

1. Each time that a transcription factor (TF) matching operation is run, the "Active TFs" data structure is **AUGMENTED** with the results of the discovery operation (i.e., contents due to previous runs are maintained). The "Active TFs" data structure is not affected by pattern discovery.
2. Each time that a Pattern Discovery analysis is run, the contents of the "active patterns" structure are **REPLACED** with the results of the discovery operation (i.e., contents due to previous runs are cleared).

13.4 Scan Implementation

13.4.1 Normalization and the Pseudocount

The count matrices are normalized to frequencies using an algorithm which includes a "pseudocount" (see [Nishida 2009](#)). The pseudocount is a way to compensate for the effects of small sample sizes in the original observations used to generate the profiles. Nishida et al. studied how to determine an appropriate value for the pseudocount. They found that the optimal values were independent of the sample size and were correlated with the entropy of the original matrices. They say that this implies that the less-conserved the binding site, the larger a value should be used for the pseudocount. They find that 0.8 is a good value "for practical uses". They do not recommend use of the square root of the total count.

geWorkbench allows a pseudocount factor to be directly entered, or it can be selected to be the square root of the total count of sequences used to generate the profile. Prior to geWorkbench

1.8.0, setting the pseudocount to the square root of the total counts was directly coded and not changeable. The current default is to set the pseudocount equal to 1.0.

The normalization formula used in calculating frequencies is then, where b is the pseudocount, and $\text{counts}(i, j)$ is the observed count in a particular entry in the matrix,

$$\text{freq}(i, j) = (\text{counts}(i, j) + b/4) / (\text{totalCounts} + b).$$

The resulting frequency matrix is used in the subsequent scan.

Because the pseudocount is a settable parameter, the frequency matrix is recalculated for each scan from the original counts.

13.4.2 Scoring

- Calculated p-values are Bonferroni corrected and also corrected for duplicates in the list of 100 top scores found during the background scan.
- Positive and negative strands are scanned and values above threshold are reported.

13.5 Example: Running and viewing a scan

13.5.1 Prerequisites

The Promoter component is only available when a sequence has been loaded, either from disk or for example using the Sequence Retriever component to obtain genomic sequence.

For this example, we will obtain upstream genomic sequence for CDH2, the N-Cadherin gene. However, using the Sequence Retriever component requires that a microarray dataset and its annotations be loaded. Here, we will use the JB-ccmp_0120.txt file, which is an Affymetrix HG-U95Av2 MAS5 format text file and is part of the geWorkbench [tutorial dataset](#).

The Affymetrix HG-U95Av2 annotation file can be obtained from the Affymetrix website. Please see instructions on the [geWorkbench FAQ](#).

1. In the Project Folders component, load the file JB-ccmp_0120.txt as type MAS5/GCOS.
2. When prompted, associate the HG-U95Av2 annotation file.
3. In the Markers component, search for the gene name "CDH2" using the Find Next button. On this chip type, marker 2053_at represents the CDH2 gene.
4. You can double-click on the marker to add it to the default "Selection" set. Or you can right-click on it and add it to a named set, such as "Cadherins". This is depicted below.
5. "Activate" (check the box next to) the set to which you added the marker.

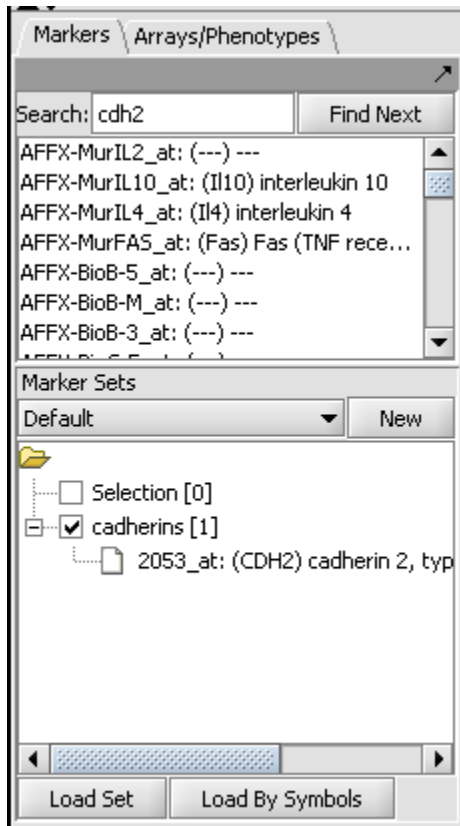


Figure 13-6 - Example: Choosing a marker for obtaining genomic sequence

6. Any activated markers will appear in the Sequence Retriever component, as shown below.
7. Set the retrieval limits to 2000 base pairs up- and downstream from the transcription start site.
8. Make sure the retrieval type is set to **DNA**, **UCSC** (the Santa Cruz genome sequence database).
9. Click "**Get Sequence**".

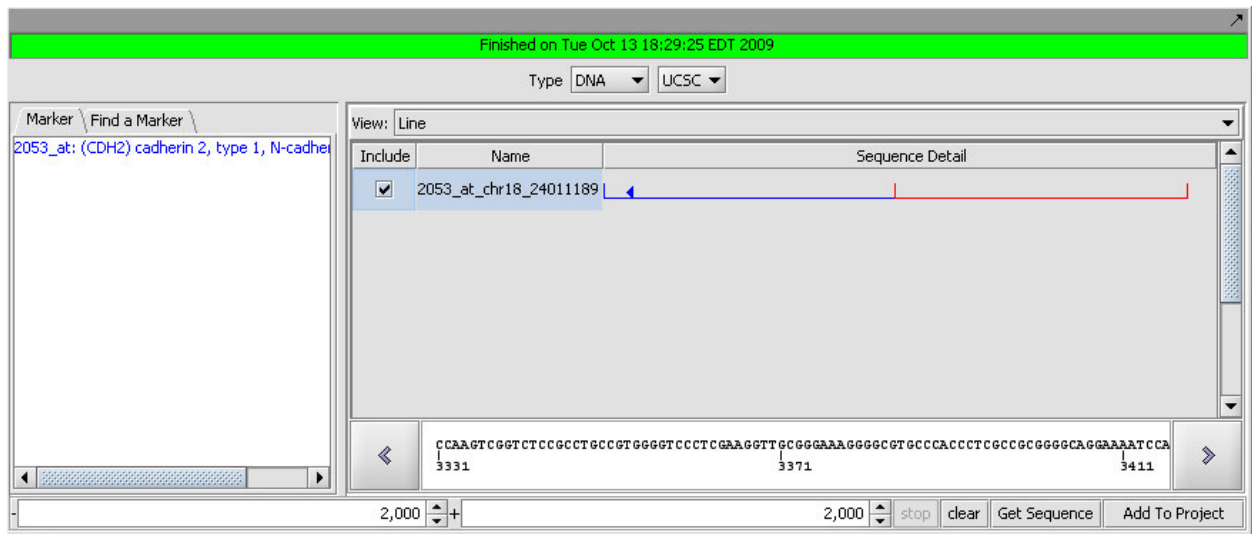


Figure 13-7 - Example: Genomic sequence retrieval.

10. Once the sequence has been retrieved, check the box next to the sequence and then hit the button "**Add to Project**". We now have the genomic sequence available for other components to use.

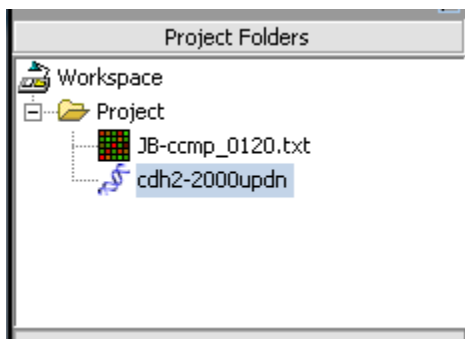


Figure 13-8 - Example: retrieved sequence in the Project Folder.

13.5.2 Running the scan

1. In the Promoter component, search for "ap2", which corresponds to transcription factor activator protein 2 alpha.
2. Double-click on the TFAP2A entry to move it down to the search list.

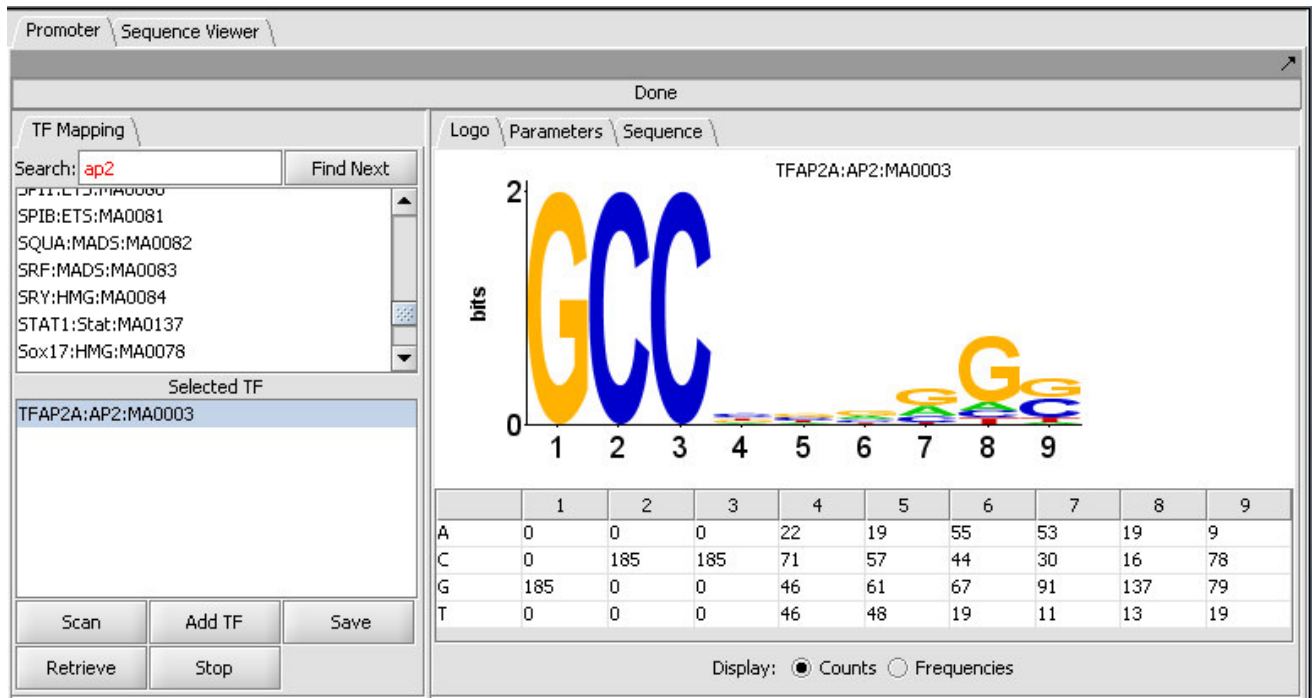


Figure 13-9 - Example: Selecting a motif for scanning.

3. Hit the "Scan" button. The result is displayed in the Sequence tab of the Promoter component as shown here.

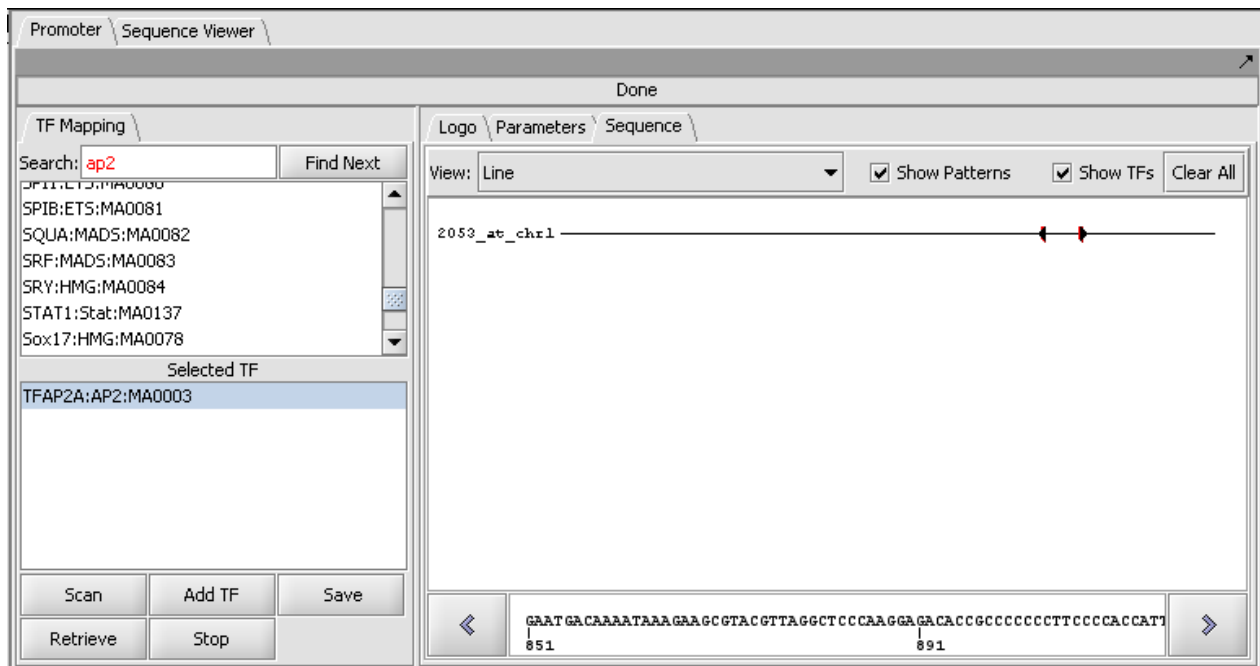


Figure 13-10 - Example: Display of scan results.

4. Setting the View to **Full Sequence** shows the hits in white on the sequence. Red arrows indicate whether the hit is to the forward (right arrow) or reverse (complementary) (left-arrow) strand.

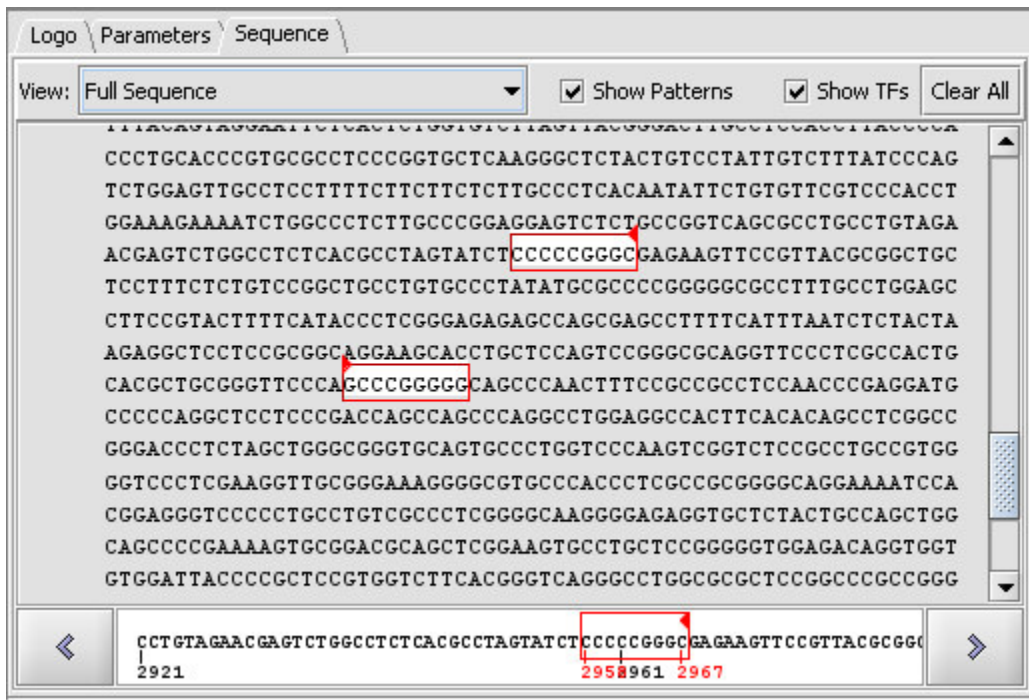


Figure 13-11 - Example: Scan results in Full Sequence view.

5. The parameters tab displays the actual threshold values calculated during the run, and displays the enrichment results.

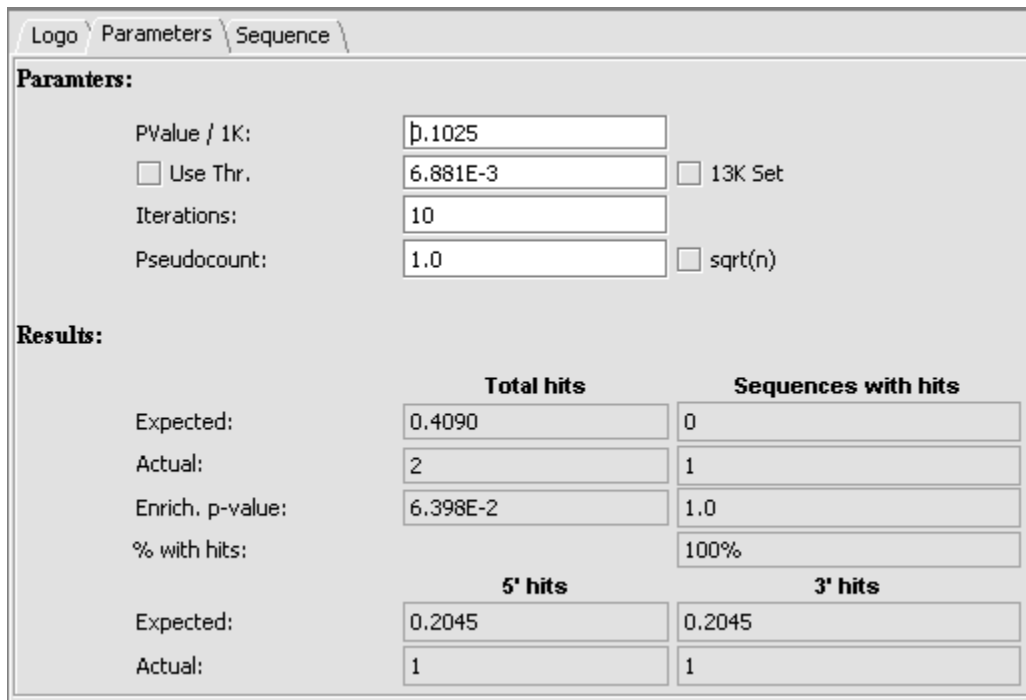


Figure 13-12 – Example: Statistics on the motif scan.

13.6 References

- Lawrence and Reilly (1990) Searching putative regulatory sequences against a collection of known transcription factor DNA-binding signatures represented as a position weight matrices (PWMs) (citation unknown). See perhaps: Lawrence, C. and Reilly, A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7 (1), 41-51. [Link to Abstract](#)
- Nishida K, Frith MC, Nakai K. (2009) Pseudocounts for transcription factor binding sites. *Nucleic Acids Res.* Feb;37(3):939-44. [link to paper](#)
- Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* Jan 1;32(Database issue):D91-4 ([link to paper](#)).
- Schneider TD, Stephens RM. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* Oct 25;18(20):6097-100. ([link to paper](#))
- Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* Jan 1;34(Database issue):D95-7.

(additional)

- Lenhard. B. and Wasserman, W. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*. **18**:1135-6 (2002).
- Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*. **16**:939-45 (1998).

14 Analysis of Variance (ANOVA)

14.1 Overview

The ANOVA (ANalysis of VAriance) algorithm (Zar, 1999) is used to determine whether any significant difference in the means exist in a dataset composed of three or more groups of experimental tests.

The geWorkbench ANOVA component implements a one-way analysis of variance calculation derived from [TIGR's MeV \(MultiExperiment Viewer\)](#) (Saeed, 2003). At least three groups of arrays must be specified by defining and activating them in the Arrays/Phenotypes component. For each chosen marker the routine determines if, at the specified level of significance, any difference in the mean exists in expression values between any of the groups (the null hypothesis is that there is no difference between the groups). Several basic methods of multiple testing correction are offered. The analysis does not indicate between which groups the difference is found, only that one exists.

Those markers for which a significant difference is found are placed into a new set in the Markers component called "Significant Genes". The results are also display as a heat map in the Color Mosaic component.

14.2 Setting up an ANOVA run

14.2.1 Prerequisites

- To use the ANOVA routine, first check that it has been loaded in the [Component Configuration Manager](#).
- ANOVA is found in the list of loaded analysis routines in the lower-right **Commands** quadrant of geWorkbench.

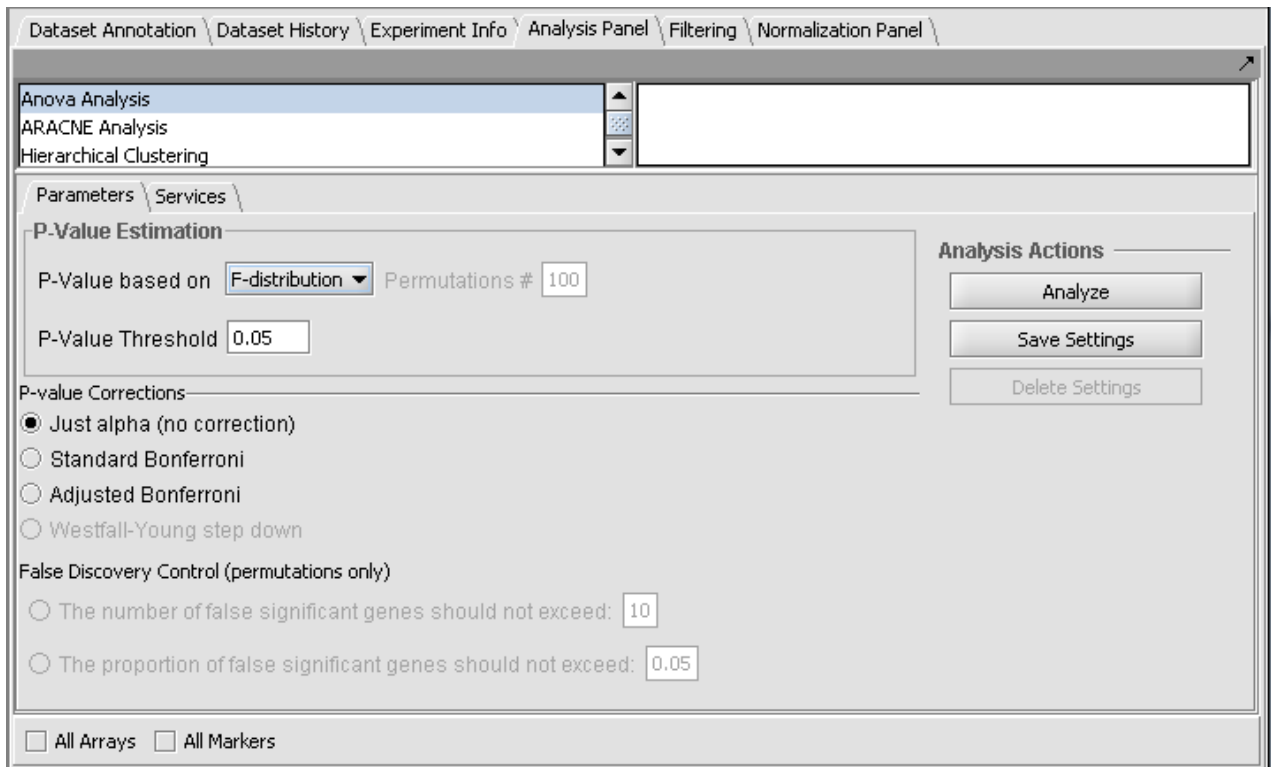


Figure 14-1 - The ANOVA parameters tab.

14.2.2 ANOVA Parameters and Settings

14.2.2.i P-Value Estimation

The P-value represents, for any one test (one marker), the probability of falsely rejecting the null hypothesis - that is, calling a difference real when it is not. It is the probability that an F-statistic at least as large as obtained would occur under the null hypothesis of no difference in means.

- **P-value based on** - Select one of two methods for calculating p-values:
 - **F-Distribution** - The p-value will be calculated using the F-distribution. The F-distribution arises from the ratio of the variances of two normally distributed statistics (chi-squared distributions).
 - **Permutation** - Permutations of the data will be used to generate a distribution against which the significance of the observed difference is judged. The number of desired permutations can be entered. The default number of permutations is 100.
- **P-value threshold** - sets the value of alpha, the critical p-value, for judging whether the null hypothesis can be rejected - that is, whether a difference is regarded as significant.

14.2.2.ii P-value corrections

Several methods for correcting for the effects of performing multiple tests are offered, including Bonferroni and False Discovery Rate control. They differ in how they compare the calculated p-value to the cutoff value of alpha - the critical p-value for determining the significance of an observed difference.

- **Just Alpha** - No correction is performed.
- **Standard Bonferroni** - The cutoff value (alpha) is divided by the number of tests (genes) before being compared with the calculated p-values.
- **Adjusted Bonferroni** - Similar to the Bonferroni correction, but for each successive P-value in a list of p-values sorted in increasing order, the divisor for alpha is decremented by one and then the result compared with the P-value. The effect is to slightly reduce the stringency (increase the power) of the Bonferroni correction. This is a step-down procedure.
- **Westfall-Young Step-Down** - (Dudoit, 2003) Another step-down procedure which adjusts the critical value alpha using a more complex expression. (This correction is only available when the permutation method is chosen for calculating p-values).

14.2.2.iii False Discovery Control

(This correction is only available when the permutation method is chosen for calculating p-values).

Rather than controlling the family-wise error rate (FWER) as do the Bonferroni corrections, that is, the probability of even one false positive occurring in the multiple trials, the false discovery rate calculation controls the rate of false positives. This can result in increased power to detect true differences. See Korn, 2001 and Korn, 2004, if one can accept more false positives. The number of false positives that is acceptable may be an economic decision, based on how many follow-up tests can be performed.

The user must select a limit to the rate of false discoveries as follows and enter the cutoff value in the adjacent text field:

- **The number of false significant genes should not exceed** - An upper limit on the number of false positives (markers falsely called as showing a significant difference), or
- **The proportion of false significant genes should not exceed** - An upper limit on the proportion of false positives.

14.2.2.iv Analysis Actions

- **Analyze** - start the ANOVA analysis
- **Save Settings, Delete Settings** - The geWorkbench analysis framework provides a standard method for saving one or more different sets of parameter settings per each type of analysis component. Please see the [Analysis Framework Tutorial](#) for further details.

- Note - The False Discovery Control parameter fields will only have their values saved if they are actually selected. As they are controlled by radio buttons, only one text field can be active at one time, and hence only at most one of those fields will be saved in any one parameter set.

14.3 Services (Grid)

ANOVA can be run either locally within geWorkbench, or remotely as a grid job on caGrid. See the [Grid Services](#) section for further details on setting up a grid job.

14.4 Working with and Viewing ANOVA Results

14.4.1 Significant markers set

All markers which met the threshold p-value (alpha) cutoff are placed into the "Significant Genes" set in the Markers component. Such sets of markers can be used as the starting point for further characterization and analysis.

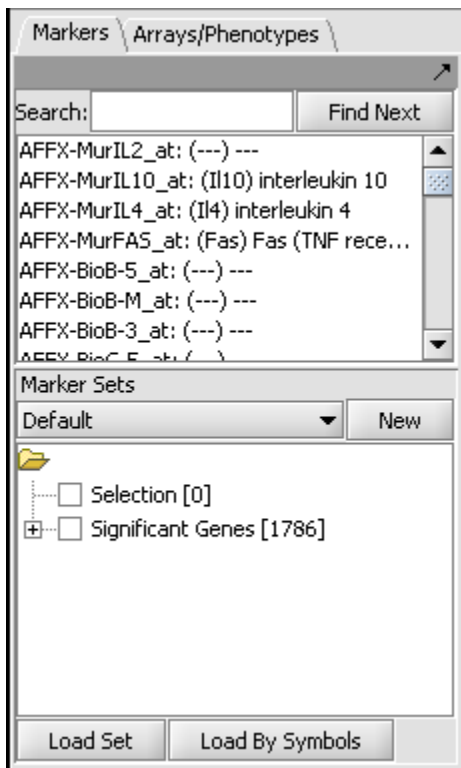


Figure 14-2 - Significant genes returned to the Marker Sets component.

14.4.2 The ANOVA result node in the Project Folders component

When the ANOVA calculation completes, the result node is placed in the Project Folders component. When the result node is selected (highlighted) in the project panel, the results will be displayed in both a tabular form and in the form of a heatmap in the Color Mosaic component.

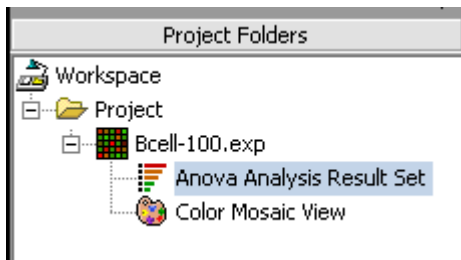


Figure 14-3 - An ANOVA result node in the Project Folder.

Also shown is a "snapshot" node, that is a static picture of the heat map, labeled "Color Mosaic View". It was produced by right-clicking on the Color Mosaic display (see below) and selecting "Take Snapshot".

14.4.3 Color Mosaic Viewer

The [Color Mosaic](#) view displays the results as a heat map, which uses a color spectrum to indicate the relative magnitudes of the expression measurements. The heat map is colored using the currently selected color scheme (Menu->Tools->Preferences->Visualization). A color bar at the bottom shows the range of the color display and its correlation with expression values.

Columns represent individual arrays, and each row represents a marker. The arrays are grouped by the array set to which they belong, with each set labeled at the top of the picture. The markers are initially sorted in order of the calculated p-value, from smallest to largest. The p-values are shown at right in the diagram. Further details are available in the [Color Mosaic tutorial](#).

The heat map depicted below was drawn using the "absolute" setting in the [main Tools->Preferences->Visualization menu](#).

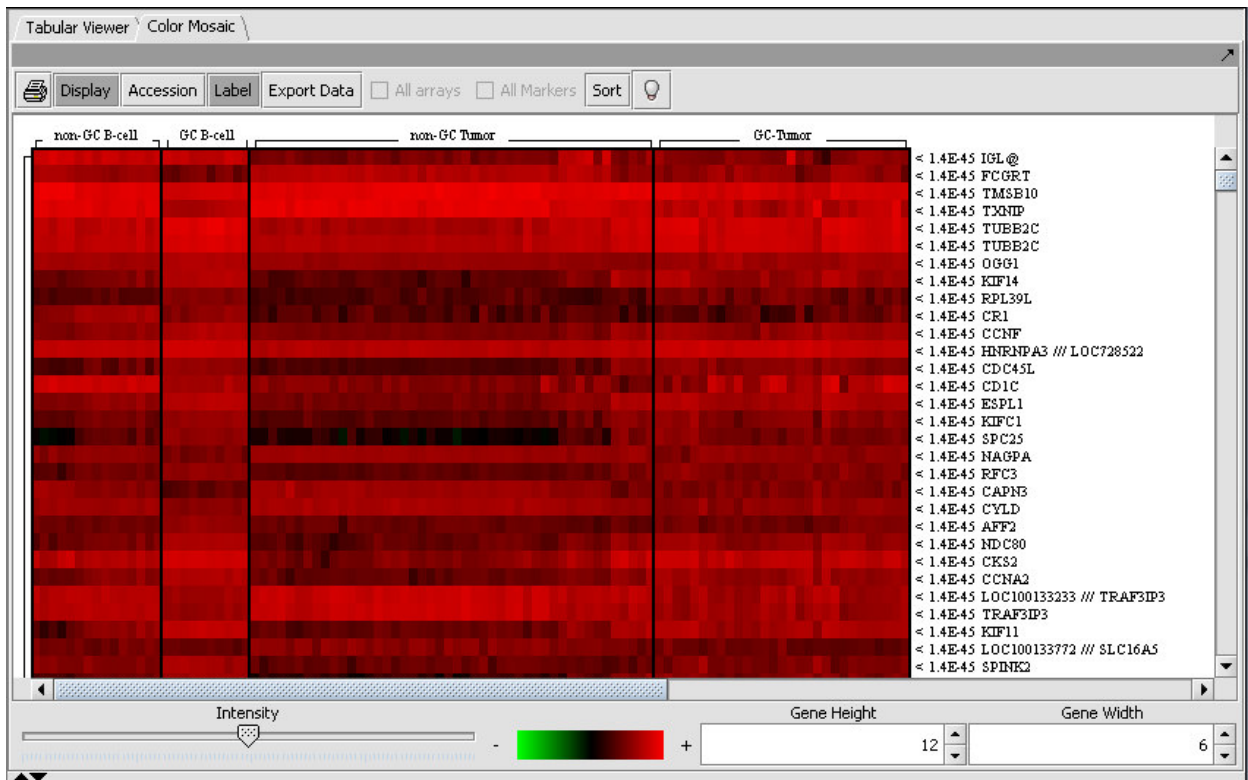


Figure 14-4 - An ANOVA result displayed as a color mosaic.

14.4.4 Tabular Viewer

This Visual Area component displays a read-only spreadsheet view of the significant genes sorted by p-value in ascending order (from most significant to least significant).

14.4.4.i Spreadsheet columns

- **Marker Name** - Shows the gene name if an annotation file has been loaded, otherwise shows the probeset name.
- **F-statistic** - the raw ANOVA score for each marker.
- **P-value** - the probability of observing an F-statistic this large by chance alone, assuming the null hypothesis of no actual differences between sets of arrays.
- **Mean** - the mean expression value for each group of arrays.
- **Std** - the standard deviation for each group of arrays.

14.4.4.ii Controls

- **Display Preference** - this button brings up a panel which controls which of the columns to display. The choices, described in the previous section, are F-statistic, p-value, mean, and standard deviation.
- **Export** - Click on Export in the lower left of the visualization to export this table in .csv format. The export file will contain only the columns displayed.

14.4.4.iii Further customizing the spreadsheet

- **Resize** columns by using the mouse to drag column boundaries.
- **Reorder** columns in the details pane by using the mouse to drag a column heading to the left or right of its original position.
- **Sort** the spreadsheet on a specific column by double clicking on its header. Successive clicks will toggle between ascending order and descending order.

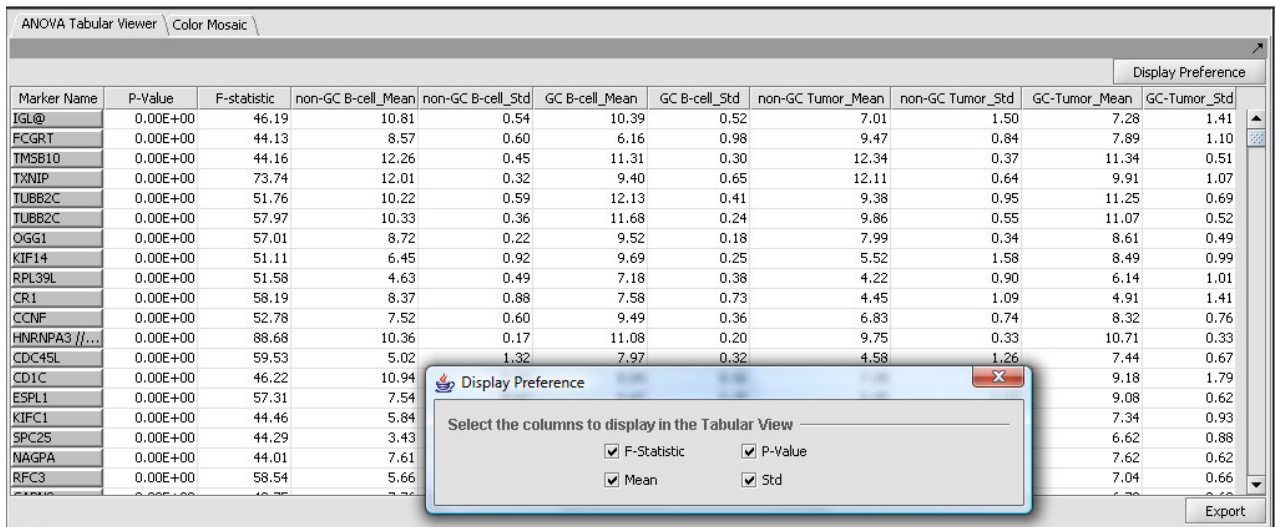


Figure 14-5 - ANOVA Tabular View: Choosing columns to display.

14.5 Dataset History

Details about each run are maintained in the Dataset History component. With the ANOVA result node highlighted in the Project Folders component, the Dataset History display includes the following information:

- P Value estimation method
- P Value threshold
- Multiple testing correction method
- Complete list of arrays in each group analyzed
- Complete list of all markers analyzed.

14.6 Example of running ANOVA

This example uses the Bcell-100.exp dataset available in the data/public_data directory of geWorkbench, and further described on the [Download](#) page. Briefly, this dataset is composed of 100 Affymetrix HG-U95Av2 arrays on which various B-cell lines, both

normal and cancerous, were analyzed. Thus it explores a potentially wide variety of expression phenotypes.

14.6.1 Prerequisites

1. (Optional) Obtain the annotation file for the HG-U95Av2 array type from the Affymetrix NetAffx website (<http://www.affymetrix.com/analysis/index.affx>). The name will be similar to "HG_U95Av2.na29.annot.csv", where na29 is the version number. Loading the annotation file associates gene names and other information with the Affymetrix probeset IDs (see the geWorkbench FAQ for details on obtaining these files).

14.6.2 Loading and preparing the example data

1. Load the Bcell-100.exp dataset into geWorkbench as type "Affymetrix File Matrix". (See [Local Data Files](#)).
2. When prompted, and if desired, load the annotation file.
3. For this example, the data was subjected to quantile normalization followed by log₂ normalization (See [Filtering and Normalizing](#)).

14.6.3 Choosing array groups

The Bcell-100 dataset comes with predefined sets of arrays.

1. In the Arrays/Phenotypes component (at lower left in the geWorkbench GUI), choose the group in the pulldown menu called "Class".
2. Check the box beside each of the four sets of arrays to activate them as shown in the figure below.

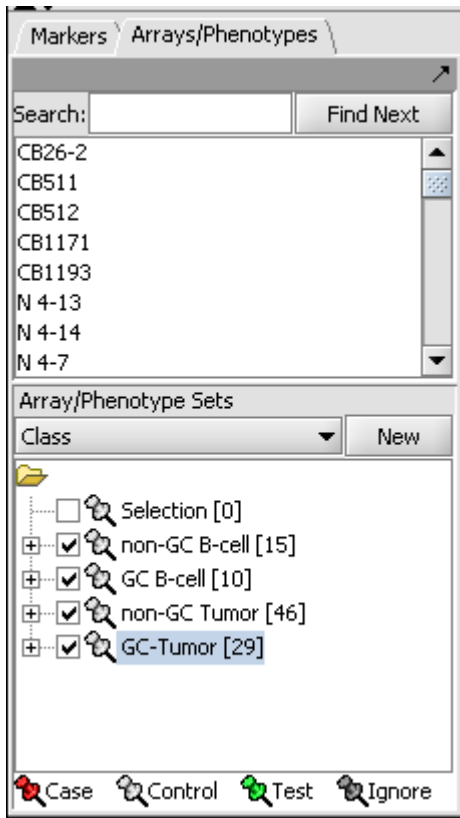


Figure 14-6 – Example: Activating array sets for ANOVA.

14.6.4 Setting up the parameters and starting ANOVA

For this example we will apply a relatively stringent multiple testing correction.

1. Leave the P-value method set to **F-distribution**.
2. Set the **P-Value Threshold** (alpha) to 0.01.
3. For the P-value correction choose **Standard Bonferroni**.
4. Push the **Analyze** button.

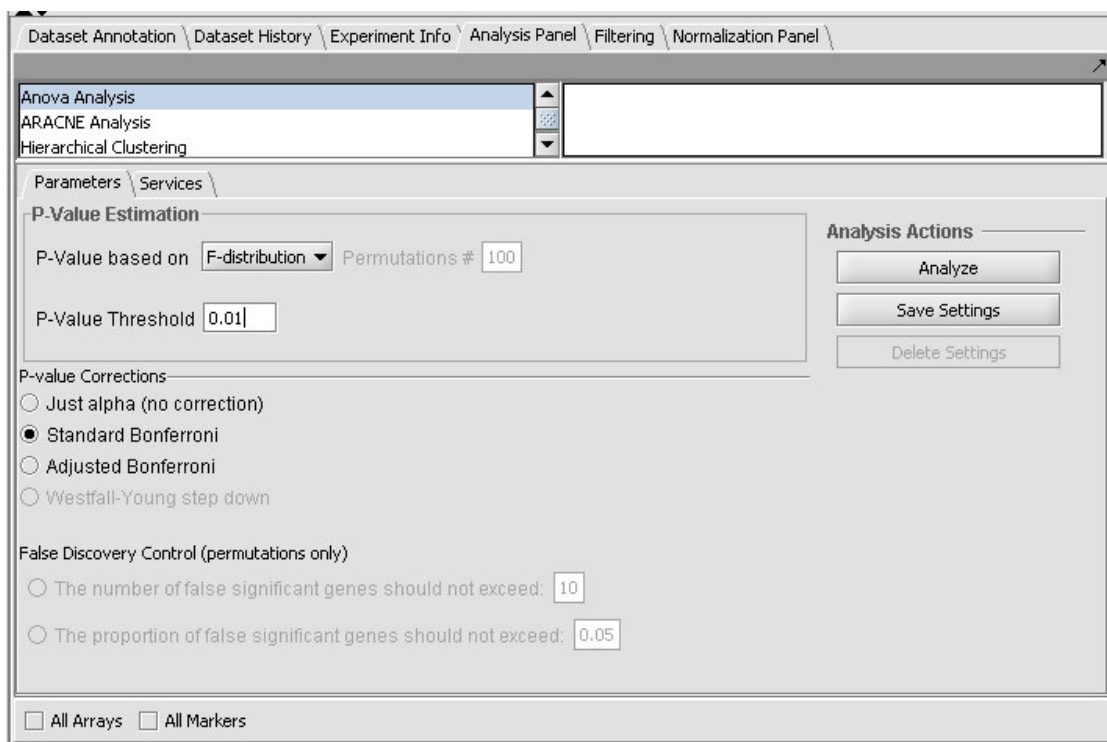


Figure 14-7 – Example: Setting parameters for an ANOVA run.

14.6.5 Results

The result of running ANOVA is a list of markers which meet the specified significance criteria. These markers are placed into a new set in the Markers component called "Significant Genes". The results are also displayed in visual components as detailed above for the Tabular Viewer and the Color Mosaic Viewer ([Viewing ANOVA Results](#)).

14.7 References

TIGR MeV lists the following relevant citations

- Dudoit S., J.P. Shaffer and J.C. Boldrick 2003. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science* 18: 71-103
- Korn, E.L., J.F. Troendle, L.M. McShane, R. Simon (2001). Controlling the number of false discoveries: application to high-dimensional genomic data. Technical report 003, Biometric Research Branch, National Cancer Institute. <http://linus.nci.nih.gov/~brb/TechReport.htm>
- Korn, E.L., J.F. Troendle, L.M. McShane, R. Simon (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* 124: 379-398.
- Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J.

TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*. 2003 Feb;34(2):374-8.

http://www.tigr.org/software/tm4/menu/TM4_Biotechniques_2003.pdf

- Zar, J.H. 1999. *Biostatistical Analysis*. 4th ed. Prentice Hall, NJ., pp 178-182.

Other references:

- Yongchao Ge, Sandrine Dudoit, and Terence P. Speed. Resampling-based multiple testing for microarray data analysis. Technical Report 633. Department of Statistics, University of California, Berkeley. <http://www.stat.berkeley.edu/tech-reports/633.pdf>

15 ARACNe

15.1 Overview

ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) (Basso et al. 2005; Margolin, Nemenman et al. 2006) is an information-theoretic algorithm used to identify transcriptional interactions between gene products using microarray expression profile data. The resulting network is displayed using the Cytoscape component. ARACNe can be used to predict potential functional associations among genes, or to predict novel functions for uncharacterized genes, by identifying statistical dependencies between genes. The results take the form of a matrix of candidate interactions, also called an adjacency matrix, which can be used for further network visualization and analysis. ARACNe has been used to reconstruct networks in mammalian cells through appropriate choice of dataset.

ARACNe performs best with a dataset containing data from 100 to 300 microarrays (see Margolin, Wang et al. 2006) and representing a number of different states of the same cellular system - for example, cells lines of varying phenotype, or cells subjected to a variety of experimental perturbations. Initial work with ARACNe was performed using a large collection (about 340) of B-cell lines of various phenotypes (Basso et al. 2005). A subset of this dataset, derived from 100 arrays, is included with geWorkbench (Bcell-100.exp).

ARACNe can perform two separate calculations:

1. **Mutual Information:** The mutual information (MI) of one or more marker's expression profile(s) is calculated against all other active markers.
2. **Data Processing Inequality (DPI):** The DPI calculation (triangle inequality) is used to remove the weakest interaction (edge) between any three markers. That is, if a MI value is available between each of three possible pairings of three markers, the weakest interaction of the three will be removed from the output. This has the intent of removing indirect interactions. For example, if $A \rightarrow B \rightarrow C$, the interaction $A \rightarrow C$ will likely be weaker than $A \rightarrow B$ or $B \rightarrow C$ and would be removed. A tolerance can be set which relaxes this screening to account for uncertainty in the MI calculation.

Parameters described below allow one to incorporate a list of putative transcription factors and optimize the run to discover targets that they may regulate.

Starting with geWorkbench release 1.7.0, a new version of ARACNe, ARACNe2, is used that supports two important new features.

1. **Algorithm** - A new algorithm, termed "Adaptive Partitioning", which is both much faster and more sensitive, is supported.
2. **Mode** - The user can choose to custom-calculate optimal run parameters for a given dataset.

Further information on ARACNe is available in the References section below.

15.2 Setting up an ARACNe run

15.2.1 Prerequisites

- To use the ARACNe routine, first check that it has been loaded in the [Component Configuration Manager](#).
- ARACNe is found in the list of available analysis routines in the lower-right **Commands** quadrant of geWorkbench.
- A microarray dataset of sufficient size and phenotypic diversity is needed (See the Overview, above).
- Load the microarray dataset into the Project Folders component. If available, associate a gene annotation file with the dataset. This will allow the results to be displayed in consolidated fashion in Cytoscape by gene rather than by marker (individual probeset) name.

15.2.2 Parameters and Settings

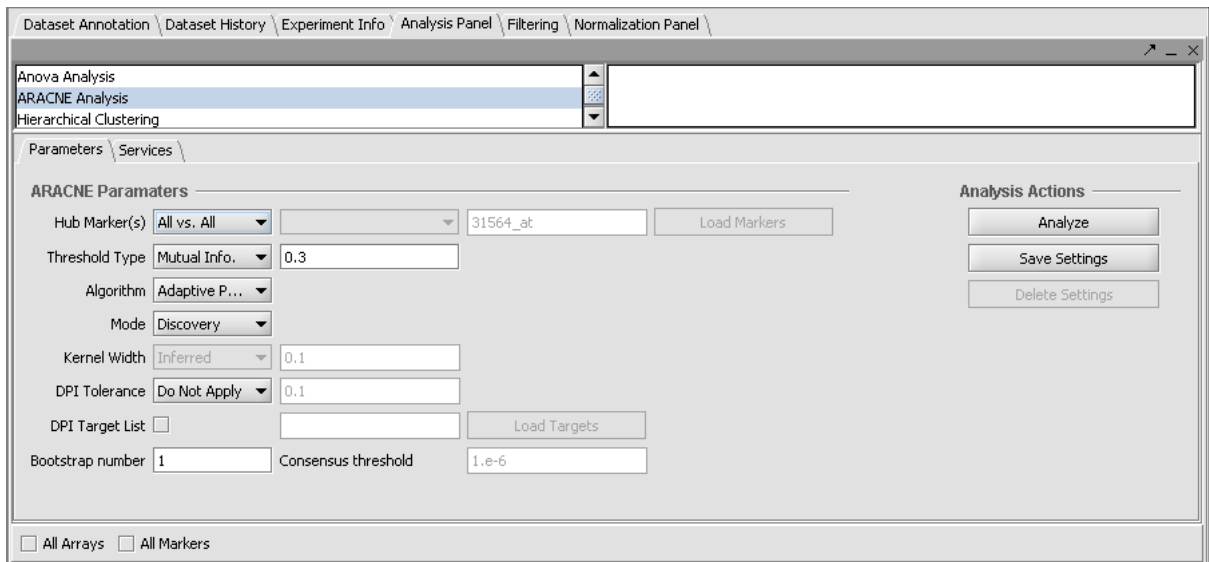


Figure 15-1 - ARACNe parameters

15.2.2.i Algorithms

Two algorithms are offered with which to calculate the pairwise mutual information between markers:

- **Adaptive Partitioning** (default) - should generally be used for all new calculations.
- **Fixed Bandwidth** - previous, slower algorithm using a fixed Gaussian kernel.

15.2.2.i.1 Adaptive Partitioning

Adaptive Partitioning was added with the incorporation of the ARACNe2 code into geWorkbench (as of version 1.7.0). Adaptive Partitioning is much faster than the original Fixed Bandwidth method, and is also considered to produce superior results. Adaptive Partitioning is now the recommended algorithm for all purposes. Unlike the Fixed Bandwidth method, it does not used a fixed kernel-width when calculating the MI.

15.2.2.i.2 Fixed Bandwidth

Fixed Bandwidth was the original algorithm used in ARACNe and is included for compatibility with previous releases. This method uses a kernel-width parameter for a Gaussian function used to calculate the MI.

15.2.2.ii Mode

Used to control the calculation and use of runtime parameters from the input dataset.

- **PREPROCESSING** - calculates the required parameters and writes them to parameter files.
- **DISCOVERY** - The ARACNe mutual information calculation is run. Uses pre-calculated parameter files as needed if they are present.
- **COMPLETE** - Preprocessing and Discovery runs are combined into a single step.

15.2.2.ii.1 PREPROCESSING

In this mode, runtime parameters are calculated, but no MI calculation is performed. Preprocessing for a given combination of dataset and algorithm need be run only once. The results are written to one or two files in the geWorkbench root directory. The names used for these files incorporate both the name of the dataset and the name of the algorithm, and thus are specific to the particular combination. Each time ARACNe is run in Discovery mode, it will look for the dataset-specific parameter files in its root directory. If the files are not found (Preprocessing has not been run), default parameter values will be used.

- Fixed Bandwidth (FBW) algorithm - two files are written to the geWorkbench root directory, one containing parameters for calculating the kernel width, and the other containing parameters for calculating a MI threshold from a specified P-value.
- Adaptive Partitioning (AP) algorithm - only the parameter file for calculating a MI threshold from a specified P-value is written. No kernel-width parameter is used.

15.2.2.ii.2 Preprocessing files included with geWorkbench

Preprocessing as described above was run on the Bcell-100.exp dataset included with geWorkbench. The resulting ARACNe parameter files are also included in the geWorkbench root directory. They will be used by default when the Bcell-100.exp dataset is used in tutorials. Note that if you rerun the preprocessing step, the relevant file(s) will be overwritten.

The parameter files included in geWorkbench are:

- Bcell-100.exp_ARACNe_AP_threshold.txt - Adaptive Partitioning, Pvalue-to-MI threshold conversion parameters.
- Bcell-100.exp_ARACNe_FBW_kernel.txt - Fixed Bandwidth, kernel width calculation parameters.
- Bcell-100.exp_ARACNe_FBW_threshold.txt - Fixed Bandwidth, Pvalue-to-MI threshold conversion parameters.

15.2.2.ii.3 DISCOVERY

The ARACNe mutual information and the DPI (if selected) calculations are run. If dataset-specific parameter files are present, they will be used as needed (based on settings selected for Kernel Width and Threshold).

15.2.2.ii.4 COMPLETE

A preprocessing run will be performed followed immediately by a Discovery run. The dataset-specific parameter files created during the Preprocessing step will be used if needed (based on settings selected for Kernel Width and Threshold).

15.2.2.ii.5 When is preprocessing not needed?

The preprocessing step can be time consuming. If you are for example using Adaptive Partitioning, and decide you do not need to specify a p-value threshold for accepting edges, then you can just set a MI value as the threshold and proceed directly to Discovery mode. This will however make interpreting results more difficult.

If ARACNe does not find the dataset-specific parameter files it needs as described above, it will use by default parameters calculated from the B-cell dataset (see Margolin et al., 2006).

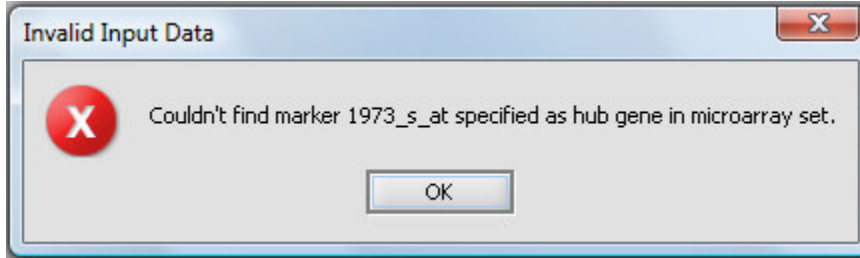
15.2.2.iii Hub Marker(s)

Specifies which gene markers will be treated as "hubs" in the ARACNE mutual information (MI) calculation. The mutual information is calculated for each specified hub marker against all other markers in the submitted dataset. For many uses, it is suggested to use a defined list of known transcription factors as hub markers, rather than using the "All-vs-All" setting.

- **All vs All** - The MI of every pair of markers in the dataset is computed, that is, each is used as a hub.
- **From Sets** - allows a set of markers defined in the Markers component to be chosen from a pulldown menu. Alternatively, the user can type in the names of desired markers directly as a comma separated list.
- **From File** - allows a comma-separated list of markers to be read in from a file by clicking **Load Markers**..

15.2.2.iii.1 Hub marker(s) must appear in active marker set

If a set of markers is activated in the Markers component, rather than using all markers, then the chosen hub marker(s) must also be included in an active set. If the hub marker is missing from the active sets, then an error dialog will be displayed. In the below picture, the marker 1973_s_at was entered into the hub field without being part of a subset of markers that had been activated:



15.2.2.iv Threshold Type

This drop-down specifies the type of threshold to be used and can take the values "Mutual Info" or "P-value". The actual value entered into the adjacent text area is always a number between 0 and 1.

15.2.2.v Kernel width

The Kernel width is a scaling parameter used for fitting a Gaussian function to the data when running the FIXED_BANDWIDTH algorithm only, otherwise this field is disabled. If used, the value can be either inferred or specified directly.

- **Inferred:** If PREPROCESSING has been run on the dataset (mode is set to PREPROCESSING or COMPLETE), the kernel width is calculated directly from those results. If PREPROCESSING has not been run, the kernel width is inferred based on parameters fitted to a large B-cell dataset (Margolin et al, 2006), extrapolated for the number of samples in the dataset being tested.
- **Specify:** The user can enter a value for the kernel width directly, e.g. based on a prior calculation with this dataset.

15.2.2.vi DPI Tolerance

The Data Processing Inequality (triangle inequality) can be used to remove the effects of indirect interactions, e.g. if TF1->TF2->Target, DPI can be used to remove the indirect action of TF1 on the target. Stated another way, the DPI can be used to remove the weakest interaction of those between any three markers. The DPI tolerance specifies the degree of sampling error to be accepted, as with a finite sample size an exact value MI can not be calculated. The higher the tolerance specified, the fewer the edges that will be removed.

- **Do Not Apply** - Do not use the DPI.
- **Apply** - DPI is applied using the threshold value (between 0.0 and 1.0) specified in the adjacent text field. The higher the threshold, the weaker the screening and the more edges will be included in the final output.

***15.2.2.vii* DPI Target List**

The DPI target list can be used to limit the ARACNE calculation to transcriptional networks. It is used to screen out spurious regulatory interaction signals of genes that are tightly co-expressed but are not in a regulatory relationship to each other, for example genes for two proteins that are in a physical complex and hence always produced in the same amounts. A comma-separated list can be typed in, or it can be loaded from an external file. If used, the DPI Target List should contain all markers that are annotated as transcription factors. Signaling proteins could also be included.

- **Details:** If the box is checked, the user selects and loads a file which specifies markers (which should be a list of one or more presumptive transcription factors) which will be given preferential treatment during the DPI edge-removal step. Edges originating from markers on this list will not be removed by edges originating from markers not on this list. However, for DPI calculations where all three markers are members of the list, the weakest connecting edge may still be removed.

***15.2.2.viii* Bootstrapping**

Bootstrap analysis can be used to generate a more reliable estimate of statistical significance for the interactions. Please see Margolin et al. 2006, Nature Protocols, Vol 1, No. 2, pg. 663-672 for further details (full reference below). Briefly, repeated runs of ARACNE are made, with arrays drawn at random from the full dataset with replacement. The same number of arrays is drawn each time as is present in the original dataset. A permutation test is then used to obtain a null distribution, against which the statistical significance of support for each network edge connection (interaction) can be measured.

- **Bootstrap number:** Specifies the number of bootstrapping runs to perform.
- **Consensus threshold** (for bootstrapping only): After the bootstrapping runs are made, a permutation test is used to estimate the significance of interactions. The consensus threshold sets the cutoff point for calling the interactions significant and returning them in the final adjacency matrix

***15.2.2.ix* Array and Marker Set Overrides**

- **All Markers:** checking this box overrides any activated marker set in the Markers component.
- **All Arrays:** checking this box overrides any activated array set in the Arrays/Phenotypes component.

***15.2.2.x* Analysis Actions**

- **Analyze** - start the ARACNE analysis
- **Save Settings, Delete Settings** - The geWorkbench analysis framework provides a standard method for saving one or more different sets of parameter settings per

each type of analysis component. Please see the [Analysis Framework Tutorial](#) for further details.

15.3 Services (Local vs Grid)

ARACNe can be run either locally within geWorkbench, or remotely as a grid job on caGrid. See the [Grid Services](#) section for further details on setting up a grid job.

15.3.1 Special Note on running in PREPROCESSING mode on caGRID

When ARACNew is run in PREPROCESSING mode on a grid node, it writes the parameter files to its execution directory on the grid node and exits. No file is returned to geWorkbench. As currently implemented, the ARACNe server detects the lack of a file to return (normally it returns an adjacency matrix) and reports an error. This error can simply be ignored. If ARACNe2 is run in COMPLETE or DISCOVERY mode this error will not occur because both return adjacency matrices.

15.4 Viewing ARACNe results

- The result of an ARACNe run is an "adjacency matrix". it contains the mutual information value for each pair of markers which exceeded the specified MI threshold. The adjacency matrix is placed into the Project Folders component as a child of the dataset that was analyzed.
- The adjacency matrix can be visualized automatically in the [Cytoscape](#) component, as shown below.

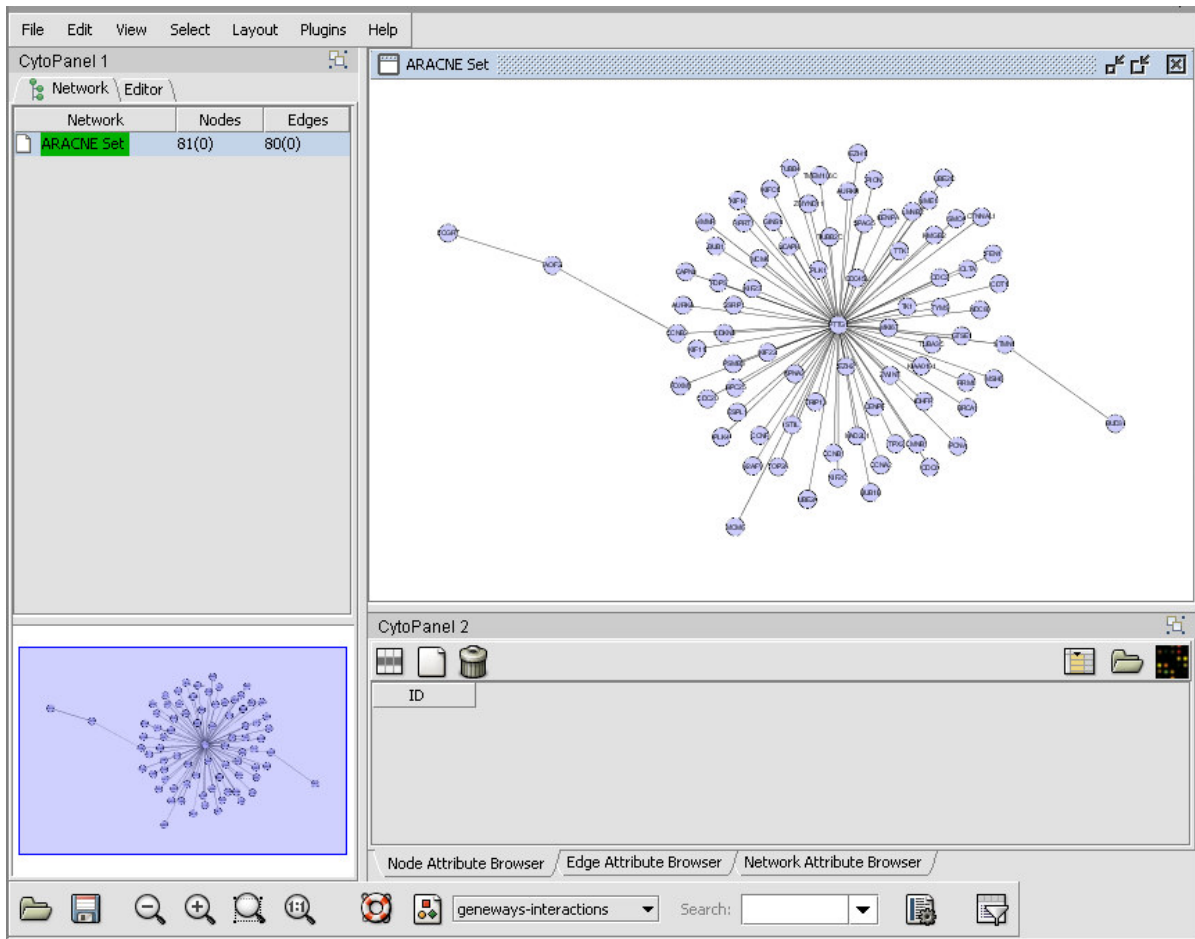


Figure 15-2 - An ARACNe-generated adjacency matrix displayed in the Cytoscape component.

The integration between Cytoscape and geWorkbench allows for two-way interactions between them:

1. Nodes selected in Cytoscape appear in the Marker Sets component in the set "Cytoscape selection".
2. Any set of markers in the Marker Sets component can be projected onto the Cytoscape display, which will cause any matching nodes there to be highlighted.

This interaction is demonstrated further on the [Cytoscape tutorial page](#).

The Cytoscape Viewer maintains a list of networks which it has currently loaded. It allows individual loaded networks to be deleted. However, the network can be reloaded by clicking on its entry in the Project Folders component. Cytoscape controls are more fully described in the [Cytoscape](#) component tutorial.

15.5 Dataset History

Details about each run are maintained in the Dataset History component. With the ARACNe result node highlighted in the Project Folders component, the Dataset History display includes the following information:

- Input file name
- Output file name
- Algorithm
- Mode
- No. bins
- MI threshold
- MI threshold calculated from P-Value - If supplied, the p-value used to set the MI threshold.
- DPI tolerance
- Hub markers
- A listing of the microarrays used.
- A listing of the markers used.

15.6 Example of running ARACNe

This example uses the Bcell-100.exp dataset available in the data/public_data directory of geWorkbench, and further described on the [Download](#) page. Briefly, this dataset is composed of 100 Affymetrix HG-U95Av2 arrays on which various B-cell lines, both normal and cancerous, were analyzed. Thus it explores a potentially wide variety of expression phenotypes.

15.6.1 Prerequisites

1. (Optional) Obtain the annotation file for the HG-U95Av2 array type from the Affymetrix NetAffx website (<http://www.affymetrix.com/analysis/index.affx>). The name will be similar to "HG_U95Av2.na29.annot.csv", where na29 is the version number. Loading the annotation file associates gene names and other information with the Affymetrix probeset IDs (see the [geWorkbench FAQ](#) for details on obtaining these files).

15.6.2 Loading the example data

1. Load the Bcell-100.exp dataset into geWorkbench as type "Affymetrix File Matrix". (See [Local Data Files](#)).
2. When prompted, and if desired, load the annotation file.
3. In the Marker Sets component, load the file "70_TFs_from_HG-U95Av2.na28.csv" (Load Set button). This file is included in the geWorkbench tutorial data.

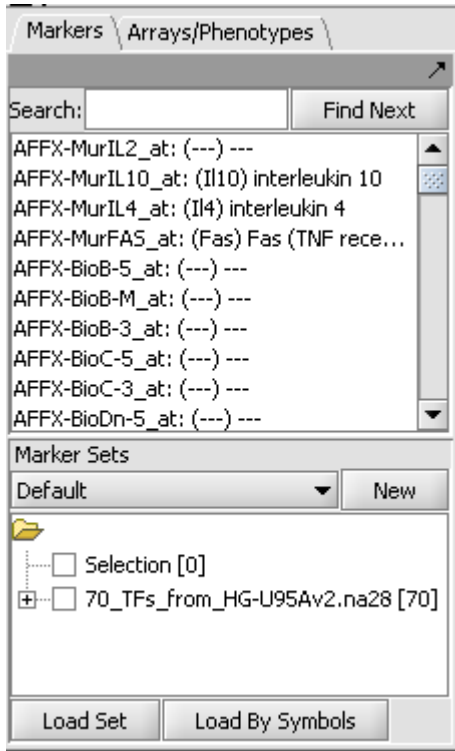


Figure 15-3 – Example: Loading a set of hub markers in the Markers component.

15.6.3 Setting up the parameters and starting ARACNe

1. In the geWorkbench commands area, select the "ARACNe" analysis.
2. Set up the ARACNe parameters as shown
 - **Hub Markers:** From Sets - Choose the set of 70 loaded above "70_TFs_from_HG-U95Av2.na28".
 - **Threshold:** Mutual Info - Set value of 0.5.
 - **Algorithm:** Adaptive Partitioning.
 - **Mode:** Discovery
 - **DPI Tolerance:** Do Not Apply

Parameters \ Services \

ARACNE Parameters

Hub Marker(s) 70_TFs_from_HG-U9...

Threshold Type 0.5

Algorithm

Mode

Kernel Width 0.1

DPI Tolerance 0.1

DPI Target List

Bootstrap number Consensus threshold

All Arrays All Markers

Figure 15-4 – Example: Setting the ARACNe parameters for a Discovery mode run.

3. Press the "**Analyze**" button to launch the job. On a current generation desktop machine expect this example to run for several minutes.
4. The resulting network is the one depicted above in the "Viewing ARACNe results" section.

15.7 References

- Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: Reverse engineering of regulatory networks in human B cells. Nat Genet 2005, 37(4):382-390 ([link to paper](#)).
- Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R and Califano A, (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context, BMC Bioinformatics;7(Suppl.1):S7 ([link to paper](#))
- Margolin, A., Wang, K., Lim, W.K., Kustagi, M., Nemenman, I., and Califano, A. Reverse Engineering Cellular Networks. Nature Protocols (2006), Vol 1, No. 2, pgs. 663-672 ([link to paper](#))

16 Using caGRID Analytical Services

16.1 Overview

In cooperation with caBIG(R), the National Cancer Institute's Cancer Biomedical Informatics Grid program, a number of the geWorkbench analysis components have also been adapted to run as services on caGrid, the primary infrastructure component of caBIG. In accordance with caBIG principles, each has a well-defined object design and a public application programming interface (API) via which data can be exchanged. Annotations describing each service, object and parameter are stored in the caDSR (NCI's Cancer Data Standards Repository), using standard vocabulary terms available from the Enterprise Vocabulary Services (EVS).

Some services are implemented only remotely, such as Mark-Uts, where the grid component serves as an interface to a web service.

Each geWorkbench analysis component that has an associated grid service will show a Services tab in the [Analysis](#) framework, adjacent to the Parameters tab.

16.2 Services tab

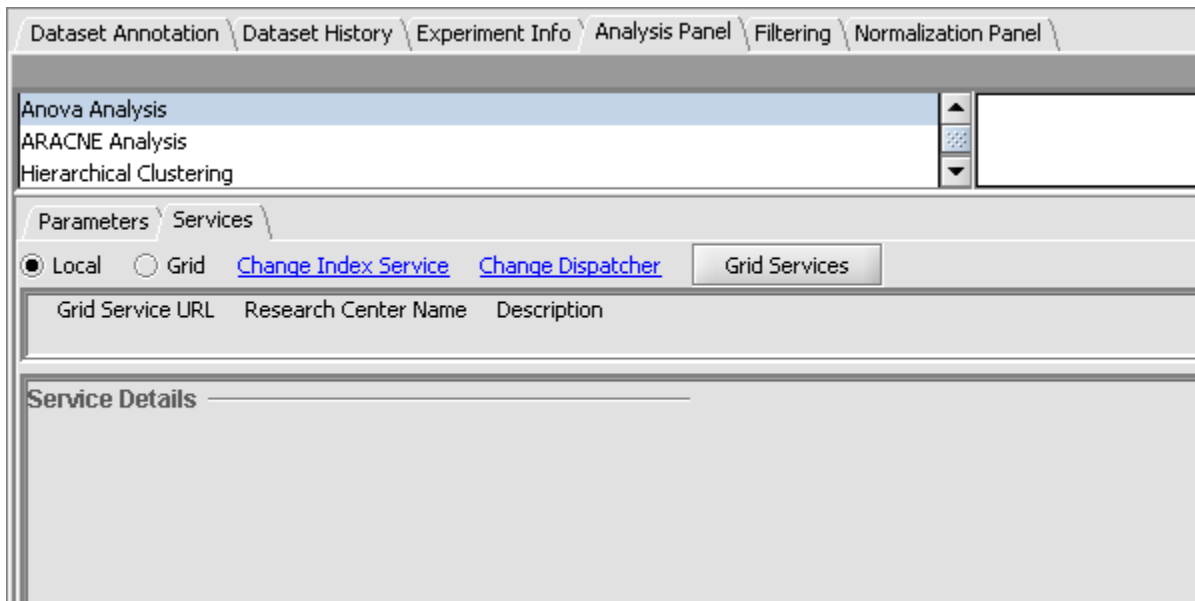


Figure 16-1 - The Grid Services tab in the Analysis Panel.

16.2.1 Local/Grid

Selects whether to run the analysis locally (within geWorkbench on the desktop) or on the grid.

16.2.2 Change Index Service

Index Services maintain lists of available grid services. geWorkbench is delivered with the URL of a Columbia Index Service preconfigured, which provides access to demonstration grid service implementations.

16.2.3 Change Dispatcher

The Dispatcher is a geWorkbench server-side component which provides connectivity between geWorkbench and caGrid. geWorkbench is delivered with the URL of a Columbia Dispatcher Service preconfigured.

16.2.4 Grid Services

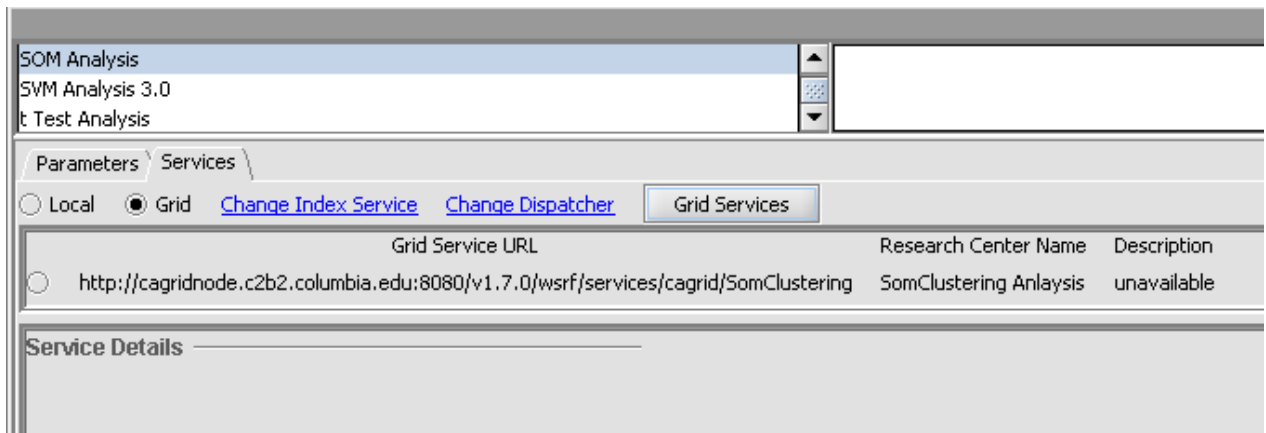


Figure 16-2 - Example of finding a grid service for SOM.

When the Grid Services button is pushed, the list of available services of the desired type will be retrieved from the specified index service. The list will appear in the area below, with each available service preceded by a radio button by means of which it can be selected.

16.2.5 Service Details

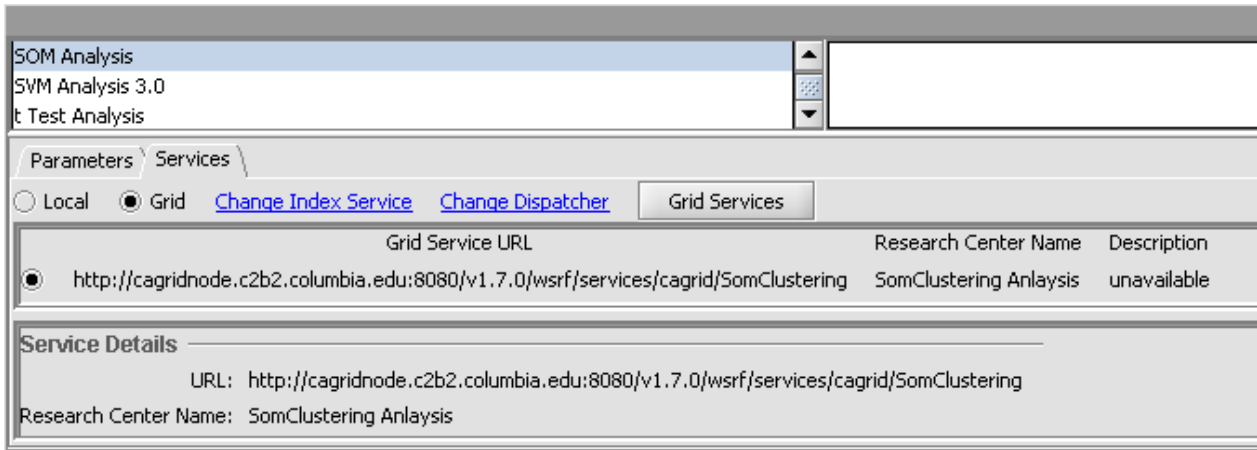


Figure 16-3 - Selecting a grid service and display of service details.

Once a particular grid service has been selected (via its radio button), the details of the service will be displayed in the lower window.

16.3 Running a grid job

1. On the Services tab, select grid. 2. If needed, choose an appropriate Index service and/or Dispatcher service. 3. Push the Grid Services button 4. Once the grid service has been chosen, return to the Parameters tab, and when ready, push the Analyze button.

A dialog will appear asking for a Username and Password. If you possess the appropriate credentials for the service you have selected, enter them here and push OK.

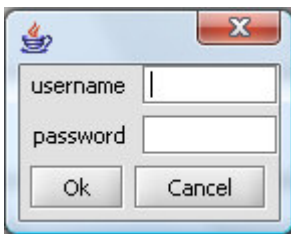


Figure 16-4 - Grid credentials (username/password) entry.

A message will indicate that the job is being submitted.

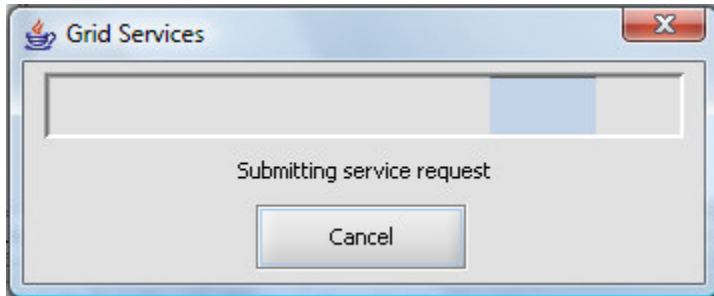


Figure 16-5 - Grid service job progress bar.

While the job is running, a node marked "Pending" will be placed in the Project Folders component, preceded by an hourglass icon. Note that the progress bar that appears when analyses are run locally within geWorkbench will not appear for grid jobs.

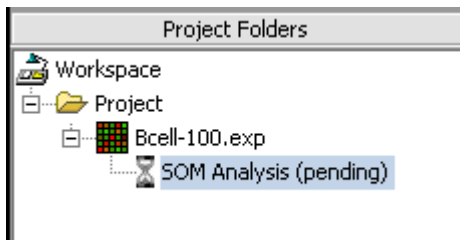


Figure 16-6 - Pending grid job node in Project Folder.

16.4 Further aspects of running grid jobs

1. The grid job, once started, is independent of geWorkbench. The dispatcher component cooperates with geWorkbench to track job status. A geWorkbench workspace containing running grid jobs can be saved and later restored. At the time that the saved workspace is reloaded, geWorkbench will resume monitoring the job for completion, and retrieve the finished results if available.
2. Once a grid job has been started, its execution cannot be canceled from within geWorkbench. However, the "pending" node can be removed from the Project Folders component. In this case, geWorkbench will not receive any results when the calculation actually completes.

References

- (1) Reverse engineering cellular networks. Adam A Margolin, Kai Wang, Wei Keat Lim, Manjunath Kustagi, Ilya Nemenman & Andrea Califano. (2006) *Nature Protocols* **1**, pp 662-671
- (2) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. (2006) *BMC Bioinformatics* **7**,S7.

17 Cytoscape

17.1 Overview

This tutorial includes new functionality specific to geWorkbench version 1.7.0.

Cytoscape (www.cytoscape.org) is a sophisticated network and pathway visualization tool that has been incorporated into geWorkbench as a component. Within geWorkbench, Cytoscape is used to depict putative interaction networks, for example as created from running ARACNe or a Cellular Network Knowledgebase query. Both of these tools return "adjacency matrices", that is, interaction networks, to the Project folders component. Currently, Cytoscape version 2.4.7 used in geWorkbench.

Cytoscape has been integrated into geWorkbench in such a way that it can communicate in both directions with the Markers component.

1. Nodes in a Cytoscape network can be selected individually or by drawing a selection box around them. This will result in the selected nodes being placed into the "Cytoscape selection" set in the Markers component.
2. A set of markers in the Markers component can be labeled with the "tag for visualization" property, which will project that set onto the network depicted in Cytoscape. Those markers in the intersection of the tagged set and the network display will be highlighted in yellow.

The use of Cytoscape and its interactions with geWorkbench are described in the following sections. First we will describe the layout of the Cytoscape graphical interface. The network diagram depicted was calculated using ARACNe in the [ARACNe tutorial](#).

17.2 Layout of the Cytoscape component

The Cytoscape component has 4 main areas:

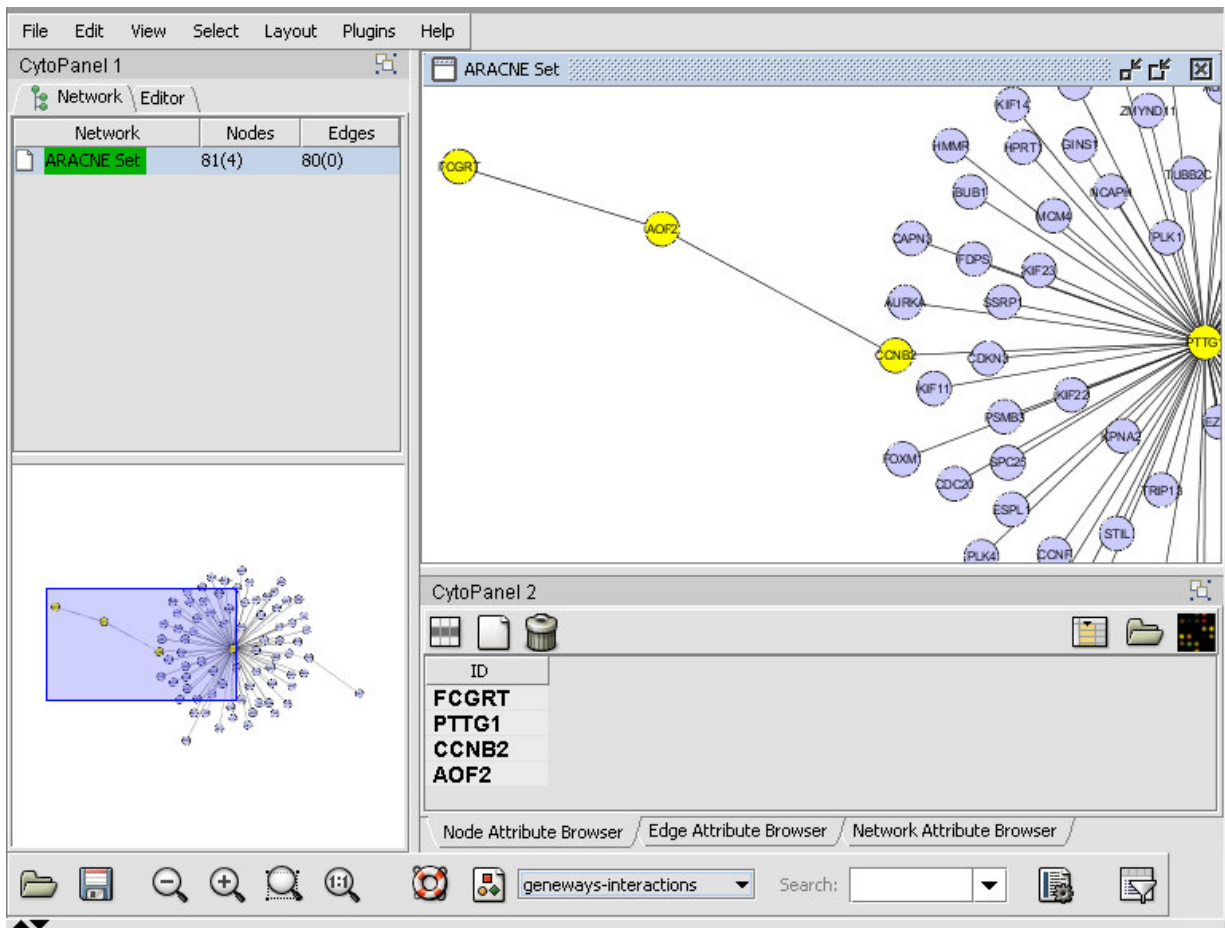


Figure 17-1 - The Cytoscape component in geWorkbench.

1. At upper left is a list of networks that have been loaded into Cytoscape, showing the network name, and the number of nodes and edges. In addition, the numbers in parentheses show the numbers of nodes and edges which have been highlighted (selected) in the network depiction.
2. At upper right is the main network depiction pane. Gene or protein symbols/names will be depicted if available. However, if for example an Affymetrix microarray dataset was read in but no annotation file was associated with it, then only probeset names would appear. Nodes which have been selected are depicted in yellow and are returned to the "Cytoscape selection" set in the Markers component. Edges which have been selected are depicted in red.
3. At lower left is a navigation tool, which shows the entire network and a representation (purple rectangle) of where the viewing pane described above is situated. The purple viewing pane can be moved by using the mouse to visualize different parts of the network as desired. This is done by left-clicking with the mouse in the purple rectangle and moving it.
4. At lower right is a list of selected nodes.

17.3 Selecting nodes in Cytoscape

- Individual nodes and/or edges can be selected in Cytoscape by clicking on them with the mouse.
- To select multiple nodes or edges, hold down the **Shift** key while making the selection.
- Alternatively, a selection box can be drawn around both nodes and edges by left-clicking in the network diagram and selecting the desired targets.

The figure below shows three nodes selected.

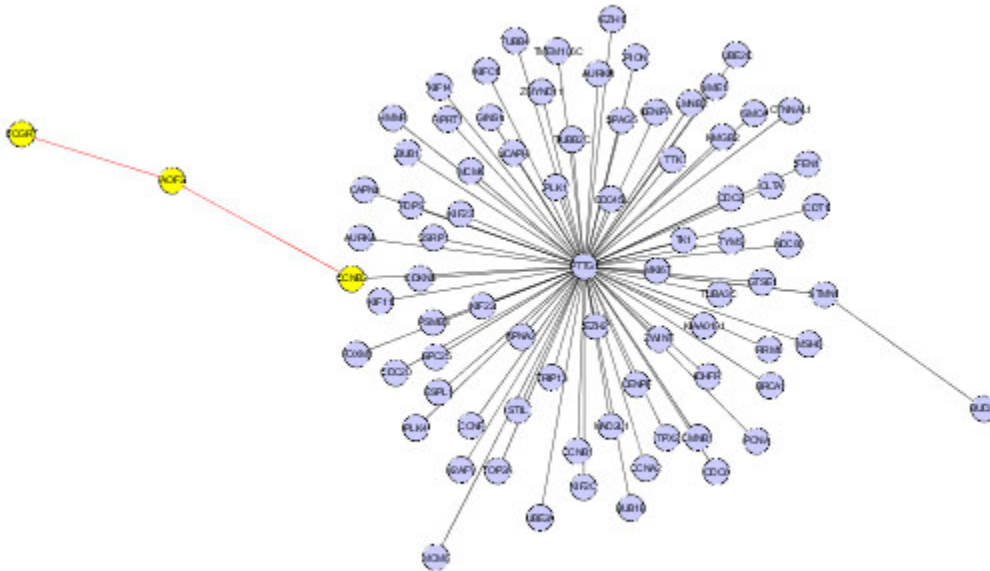


Figure 17-2 - Selecting individual nodes.

The result from selecting the three nodes above is transferred to the Cytoscape selection set in the Markers component. Note that four probesets are hit by the three genes selected, as a gene may be represented by multiple probesets.

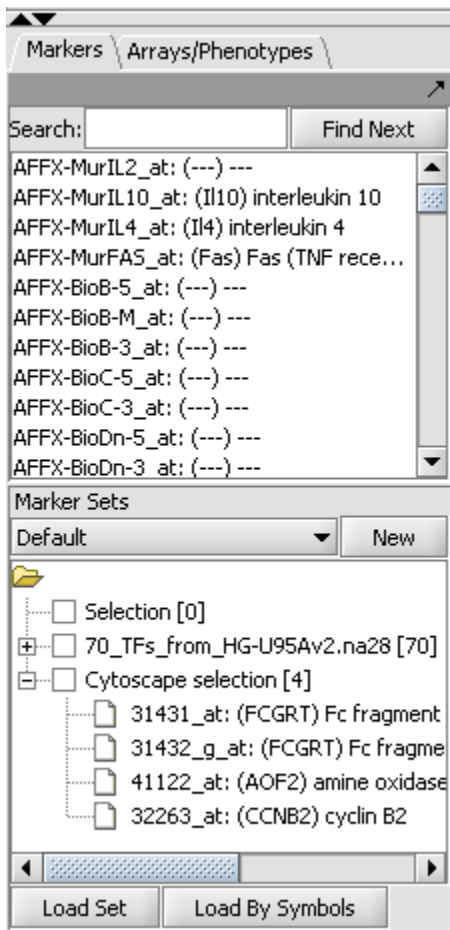


Figure 17-3 - Selected nodes returned to a Marker Set.

If a selection box is drawn on the network with the mouse, both nodes and edges will be selected as shown below:

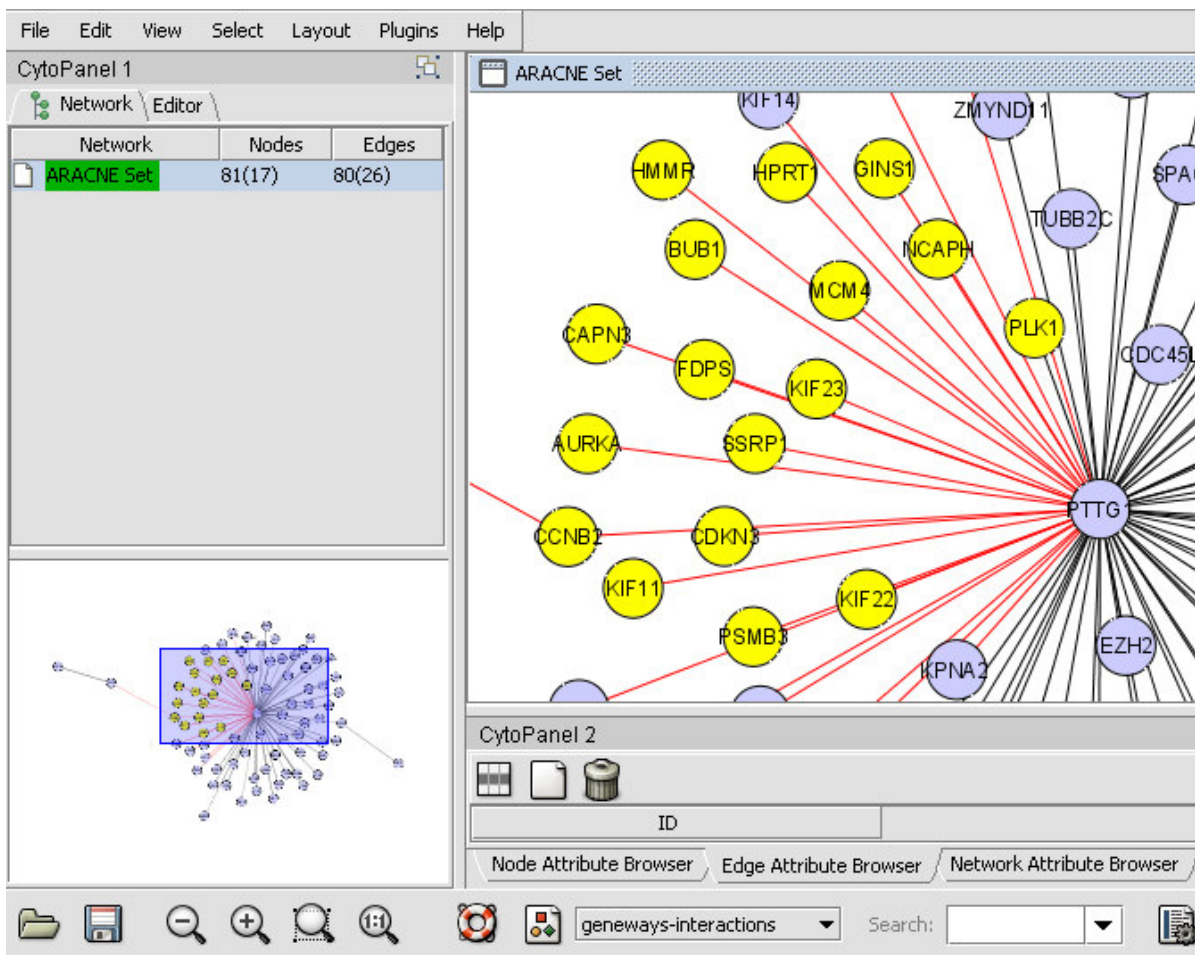


Figure 17-4 – Selection of nodes and edges using a selection box.

17.4 Set operations on networks

As described in the previous section, multiple gene nodes can be selected by holding down the Shift key while left-clicking on each desired node in turn. While the multiple nodes are selected, one can right-click, which will produce a pop-up menu. This menu allows one to choose the set of genes which are either the UNION or INTERSECTION of the those connected directly to the selected nodes.

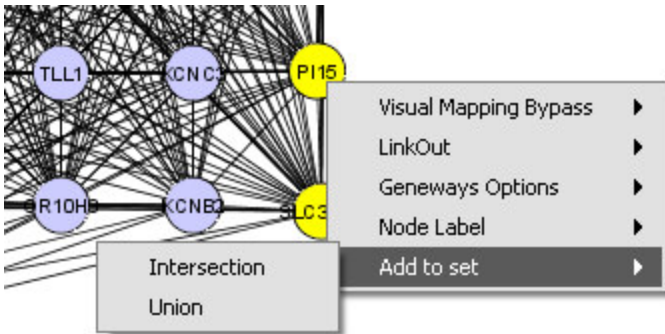


Figure 17-5 - Set operations on nodes selected in Cytoscape.

The genes selected in this way are returned to the Markers component in a set called "Cytoscape Selection". For each gene included in the UNION or INTERSECTION, there may be more than one marker associated with it. If so, all markers belonging to a particular gene will be returned to to the Markers component "Cytoscape Selection" set.

17.5 Projecting marker sets onto Cytoscape

The diagram below illustrates projecting a set defined in the Markers component back onto the Cytoscape network diagram. In this case, the set of transcription factors originally used in the ARACNe run is labeled with "tag for visualization" by right-clicking on it and selecting this menu option.

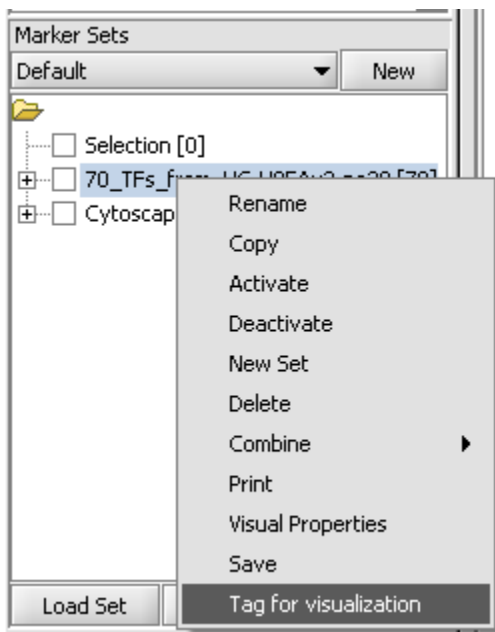


Figure 17-6 - Projecting a marker set into Cytoscape

Three nodes are highlighted in the diagram.

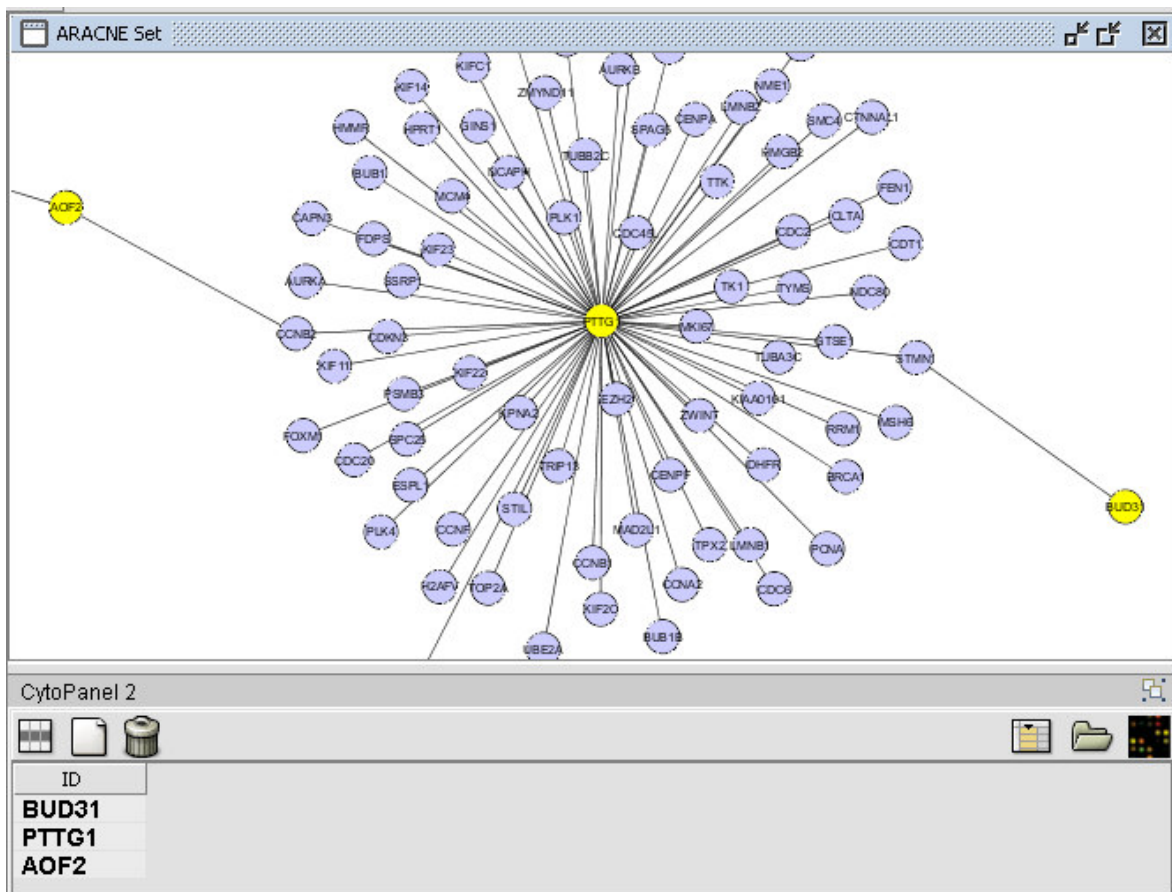


Figure 17-7 - Three genes from the projected set are present in the displayed network.

Note that a marker can potentially be associated with more than one gene. If so, when tagged for visualization, each gene associated with the marker, if present in Cytoscape, would be highlighted.

17.6 Altering the view in Cytoscape

The use of the sliding viewpane at lower left to navigate about the main drawn network has already been mentioned - it can be grabbed and moved by left-clicking on it with the mouse.

There are several more controls arrayed about the lower edge of the Cytoscape component. These include four magnifying glass icons:

- "minus" - zoom out.
- "plus" - zoom in.
- "open" - zoom to display selected region.
- "1:1" - zoom out to display all of current network.

In the image below, the view has been zoomed in:

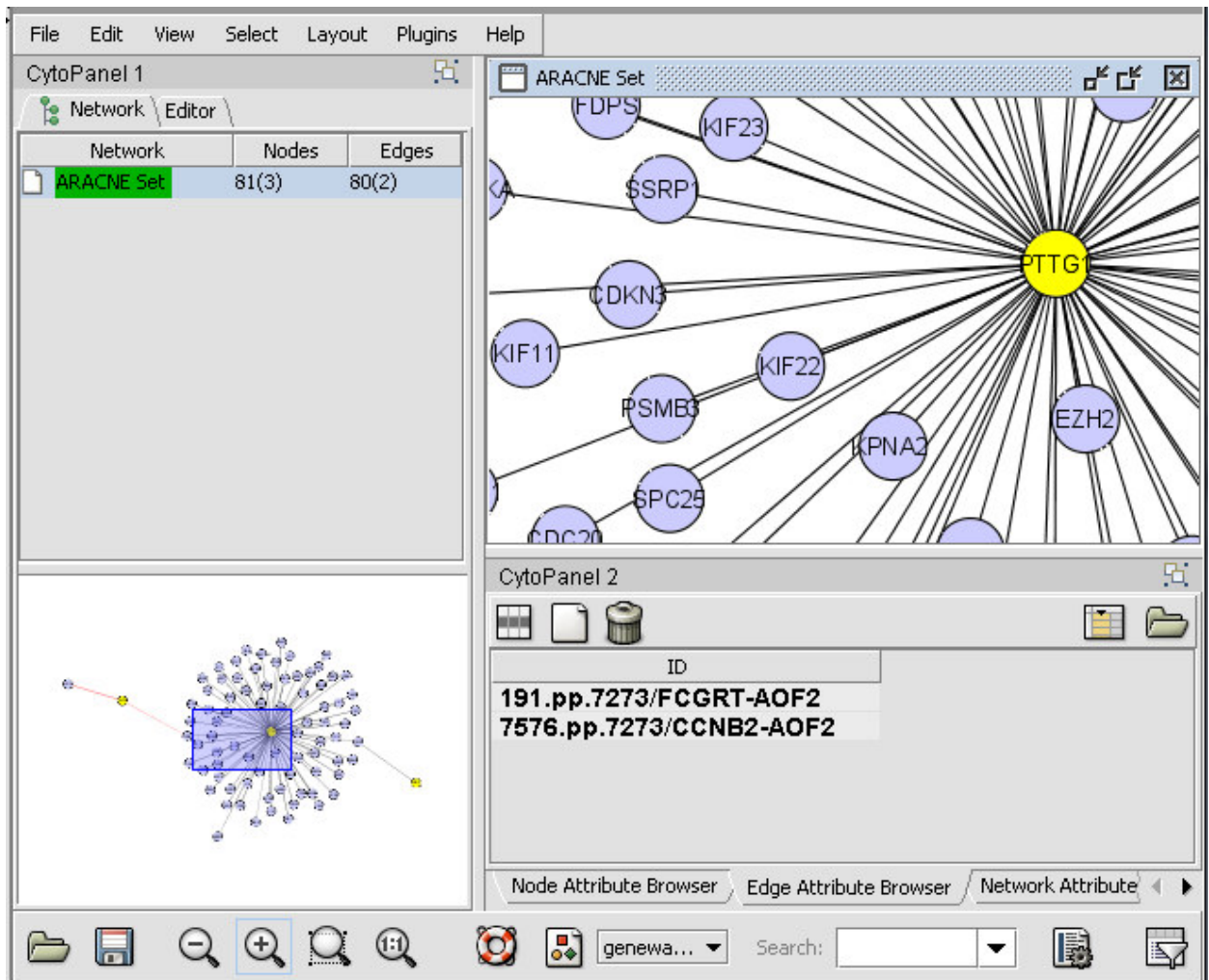


Figure 17-8 - Zooming in on a Cytoscape network view.

17.7 Network commands

Right-clicking on a listed network in Cytoscape will bring up a menu with the following choices:

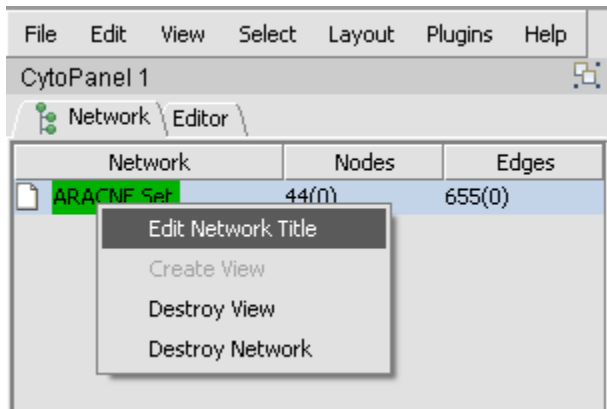


Figure 17-9 - Cytoscape network commands.

17.7.1 Edit Network Title

Edit the title of the network.

17.7.2 Create View

Recreate the network graphics.

17.7.3 Destroy View

Remove the network graphics.]

17.7.4 Destroy Network

Completely remove the network from Cytoscape. Note that this does not remove the network adjacency matrix from the geWorkbench Project Folders component. The network can be recreated in Cytoscape by clicking on the appropriate adjacency matrix in the Project Folders component.

Appendix A. Error Messages/Indicators and Problem Resolutions

Why is the desired caGRID service not available?

Until services are installed in an official caBIG index service, their availability may vary.

After running a large local calculation, my Windows computer seems slow.

We recommend running geWorkbench in a computer with at least 1 GB of memory, while 2 GB or more will greatly increase the size of calculations possible. See also the following entry on increasing the amount of memory allocated to geWorkbench.

How do I increase the memory allocated to geWorkbench?

1. If you are running a packaged distribution of geWorkbench (created using InstallAnywhere), there is a file in the geWorkbench root directory called UILauncher.lax. There is a line there which specifies the Java heap size:

```
lax.nl.java.option.java.heap.size.max=640678989
```

Here it is shown set to about 640 MB. You can experiment with increasing this, subject to the amount of memory in your machine and demands on it from other applications.

2. If you are running geWorkbench from the source distribution using Ant, you can edit the build.xml file found in the geWorkbench root directory to alter the memory requested using the variable **jvmarg**:

```
<target name="run" depends="init" description="Runs geWorkbench.">  
  <java fork="true"  
    classname="org.geworkbench.engine.config.UILauncher">  
    <jvmarg value="-Xmx640M"/>  
    <jvmarg value="-Djava.library.path=lib"/>  
    <arg value="all_release.xml"/>  
    <classpath refid="run.classpath"/>  
  </java>  
</target>
```

Here it is shown requesting 640 MB.

**Where else can I
look for help?**

Please see the main geWorkbench website at <http://www.geworkbench.org/>. Of particular interest will be the following sections:

1. FAQs
2. Known Issues
3. Tutorials

Please also see the **Contacts and Support** section at the beginning of this document.

Appendix B. Glossary

Following is a list of terms and their definitions.

Term	Definition
API	Application Programming Interface
caArray	cancer Array Informatics
caBIG	cancer Biomedical Informatics Grid
caBIO	Cancer Bioinformatics Infrastructure Objects
caCORE	cancer Common Ontologic Representation Environment
caDSR	Cancer Data Standards Repository
caMOD	Cancer Models Database
CDE	Common Data Element
CGAP	Cancer Genome Anatomy Project
CMAP	Cancer Molecular Analysis Project
CVS	Concurrent Versions System
EVS	Enterprise Vocabulary Services
GUI	Graphical User Interface
HTTP	Hypertext Transfer Protocol
JAR	Java Archive
Javadoc	Tool for generating API documentation in HTML format from doc comments in source code (http://java.sun.com/j2se/javadoc/)
MAGE	MicroArray Gene Expression
MAGE-OM	MicroArray Gene Expression - Object Model
MGED	Microarray Gene Expression Data
MO	MGED Ontology
NCI	National Cancer Institute
NCICB	National Cancer Institute Center for Bioinformatics
SDK	Software Development Kit
SQL	Structured Query Language
UI	User Interface
URL	Uniform Resource Locators

December 29, 2009