



MetaMine User Manual

version 1.1

September 9, 2002

Copyright 2002 Silicon Genetics. All rights reserved. GeneSpring, GeneSpider, GenEx, GeNet, MetaMine, ScriptEditor and MicroSift are trademarks of Silicon Genetics. All other products, including but not limited to Affymetrix GeneChip®, Affymetrix Global Scaling™, GenBank, Microsoft Excel®, Microsoft Notepad®, Pico™, SimpleText© and Adobe FrameMaker®, are the trademarks of their respective holders.

Related Documents

GeneSpring User Manual

GeNet User Manual

Scripts & the Script Editor

Table of Contents

1	Getting Started With MetaMine	
	Installing MetaMine	1-2
	System Requirements	1-2
	Installation.	1-2
	Configuration	1-3
	MetaMine Analyses	1-4
	List of Analyses	1-4
2	Using MetaMine	
	The MetaMine Interface	2-2
	MetaMine Console	2-2
	MetaMine Preferences in GeNet.	2-2
	Emails	2-3
	Search Reports Page	2-3
	MetaMine Results Display	2-4
	Full Reports.	2-4
	Availability of Results	2-6

Getting Started With MetaMine

MetaMine™ is designed to discover valuable information among the incredibly large sets of gene expression data. It searches through databases of gene expression data and reports interesting findings using email messages.

MetaMine allows groups of researchers to:

- identify new facts that relate to questions they pose and questions they have yet to conceive
- minimize the possibility of overlooking potentially valuable results
- extract more information out of their existing data repositories
- ensure that multiple research teams are aware of each other's interesting data
- perform standardized analyses on each and every sample loaded into GeNet, with minimal human effort
- retrieve interesting results based upon customized user profiles

MetaMine maximizes the effectiveness of research teams by automating the lead generation process. Rigorous analyses that would normally require countless employee hours can now be completed automatically by MetaMine. Designed to interface with GeNet™, MetaMine is capable of analyzing data from multiple GeneSpring™ users.

Reports are searchable on MetaMine at any time. All MetaMine reports are archived and can be searched from within GeNet. Researchers who did not initially receive the email reports can easily look up reports that they find interesting.

Installing MetaMine

System Requirements

Before you install MetaMine, GeNet must already be installed, configured, and running. You must also have the MetaMine .jar and preferences files and a valid license key from Silicon Genetics.

Windows:

- Windows 2000
- Pentium IV or better
- 2-3 GB RAM
- 80GB HDD

Macintosh:

- MacOS 8.1 or higher
- Power PC or better
- MRJ 2.2.5
- 1-2 GB RAM
- 80GB HDD

Unix:

- Most common Unix OSes (Linux and Solaris recommended)
- A JVM that supports JDK1.1 or later
- 2-3 GB RAM
- 80GB HDD

Installation

1. Create a directory for MetaMine, i.e., C:\MetaMine, and a directory within it called data.
2. Place the following files in the new directory:
 - MetaMine.jar
 - MetaMine Preferences
3. Within this directory, create a new subdirectory named data.
4. Create a script to execute the following command:

```
java -Xmx1024m -cp MetaMine.jar MetaMine
```

On some systems it may be necessary to provide a full path to the Java executable.

5. Open MetaMine Preferences in a text editor and specify the directory in which MetaMine data will be kept, i.e., C:\MetaMine\data. The directory you specify must already exist. MetaMine will not create this directory automatically.

- Open the GeNet preferences file in a text editor and specify the address and port number on which you will run MetaMine.

Do not modify any of the other settings yet.

- Start MetaMine by executing the script you created.
- When the MetaMine console appears, click the "Preferences" button and set the other options there. For information on MetaMine Preferences, see "Configuration" on page -3

Configuration

To configure MetaMine, click the **Preferences** button on the MetaMine console window.

Data Directory	c:\Program Files\SiliconGenetics\MetaMine\data	Browse
MetaMine Port Number	400	
GeNet Address	192.168.1.100	
Desired Memory Use (MB)	96	
Disk Cache Size (MB)	100	
Refresh References After (days)	60	
SMTP Server Address	smtp.sigenetics.com	
Email 'From' Address	haystack@sigenetics.com	
Number of Processors	1	

Figure 1-1 MetaMine Preferences

On the Preferences screen, you can specify the following information:

- **Data Directory**—The directory in which MetaMine saves data. You can use the Browse button to select a new directory, if desired.
- **MetaMine Port Number**—The port number on which the MetaMine server should run.
- **GeNet Address**—The IP address of your GeNet server.
- **Desired Memory Use**—The amount of memory MetaMine should try to stay within.
- **Disk Cache Size**—The amount of disk space, in MB, to use for caching data.
- **Refresh References After**—How often to check for new references on genes.
- **SMTP Server Address**—The address of the SMTP server through which MetaMine will route outgoing email.
- **Email 'From' Address**—The email address that should appear in the 'From' line of emails sent by MetaMine.
- **Number of Processors**—The number of processors in the computer.

Once you are done specifying these preferences, click **OK** to save your changes and return to the MetaMine console.

MetaMine Analyses

MetaMine automatically analyzes experiments that are uploaded to GeNet. Each experiment that is uploaded is analyzed in exactly the same way, regardless of which user uploaded it, what genome it is part of, which users might be interested in it, or any other factors.

The list of analyses that MetaMine performs is fixed. MetaMine determines what to analyze, when, and how. When an analysis is completed, any potentially interesting results are permanently archived to disk. MetaMine flags a result as “potentially interesting” if the result is statistically significant and might conceivably be interesting to some person at some time. This does not mean that the result is actually interesting to any user at the present time. Whether a result is archived is determined without regard to the interest profiles that users have created.

All archived results are available to users (depending on their access level) through the GeNet web interface. In addition, MetaMine automatically sends email to users notifying them of results that closely match their interest profiles. These emails contain links back to the GeNet web interface for viewing reports about the results.

The following section describes the analyses performed by the current release of MetaMine. Additional analyses will be available in future releases. The description of each analysis includes whether a similar analysis is available in GeneSpring. If any additional information beyond the experiment is required for the analysis, it is also stated in the description.

Note that all current analyses relate to one and only one experiment. There are no analyses that combine information from two different experiments and analyze them jointly. However, some analyses do combine an experiment with other information on the GeNet server, such as gene lists and pathways.

List of Analyses

- **Determine whether an experiment contains any samples for which no data at all is available for any gene.**

This usually indicates that a mistake was made in loading the data (such as specifying the wrong column in the data file) or normalizing it (such as specifying a cutoff value so high that all data falls below it).

There is no analysis in GeneSpring that specifically does this check. On the other hand, the lack of any data is easily apparent when examining a graph of the All Samples interpretation for the experiment.

- **Using the default interpretation for the experiment, look for sets of samples which are marked as being replicates, but whose expression values are significantly different from each other.**

This may indicate an error in data collection (for example, contamination of a biological sample) such that supposed “replicate” samples are not actually replicates of each other. Alternatively, it may simply indicate that the default interpretation does not accurately reflect the replicate structure of the experiment.

This analysis does not directly correspond to any analysis in GeneSpring. On the other hand, similar information can be obtained using experiment trees and the Predict Parameter Values->Crossvalidate Training Set command.

- **Perform a hierarchical clustering of genes based on the experiment's default interpretation. Then compare every cluster to every standard gene list in GeNet to see if there is a significant overlap.**

This corresponds to GeneSpring's auto-annotation feature for gene trees. This feature most useful when a good collection of standard gene lists (such as the Gene Ontology lists) is present on GeNet for that genome.

- **Perform a hierarchical clustering of genes based on the experiment's default interpretation, then use each cluster as the basis of a search for potential regulatory sequences.**

This analysis can only be done when a sequence is available for the genome. In principle, a user could do the same thing in GeneSpring by creating a gene tree, creating gene lists from its branches, and executing the Find Potential Regulatory Sequences command on each one. In practice, that would take far too much work to be practical in most cases. It could, however, be automated with a script.

- **For each parameter defined for the experiment, attempt to find genes whose expression values can be used to predict the value of that parameter.**

This is equivalent to the Predict Parameter Values tool in GeneSpring.

- **For each pathway stored in GeNet for this genome, attempt to find new genes that could potentially fit on the pathway based on their expression profiles in the experiment's default interpretation.**

A gene is considered to "fit on the pathway" if there is some position on the pathway where it could be placed such that it is similar to nearby genes, and significantly less similar to more distant genes. This analysis is only useful when a good collection of pathways is available on GeNet.

This analysis roughly corresponds to the Find Genes Which Could Fit Here command in GeneSpring. This analysis is more highly automated, however. Whereas the GeneSpring command requires the user to specify the location on the pathway where candidate genes should be placed, the MetaMine analysis automatically tries to determine an optimal candidate location for every gene.

- **For each standard gene list stored on GeNet, attempt to find subclusters of genes within that list based on expression profiles in the experiment's default interpretation.**

A subcluster is a set of genes whose profiles are similar to each other, but different from other genes in the list. This may be an interesting result, since standard gene lists usually represent groups of related genes whose behavior is expected to be similar to each other. This analysis is only useful when a good collection of standard gene lists is present on the GeNet server. There is no comparable analysis in GeneSpring.

Using MetaMine

There are five different “user interfaces” by which MetaMine interacts with users in various ways: the MetaMine console, the MetaMine Preferences page in GeNet, emails which are sent to the user, the Search Results page in GeNet, and the “MetaMine Results” display in GeNet.

The MetaMine Interface

MetaMine Console

Only the MetaMine system administrator sees this interface. It appears only on the server on which MetaMine is installed. This interface is used to start and stop MetaMine, configure various preferences (memory use, disk cache, etc.), and see an overview of what MetaMine is currently doing and how many analyses it has run.

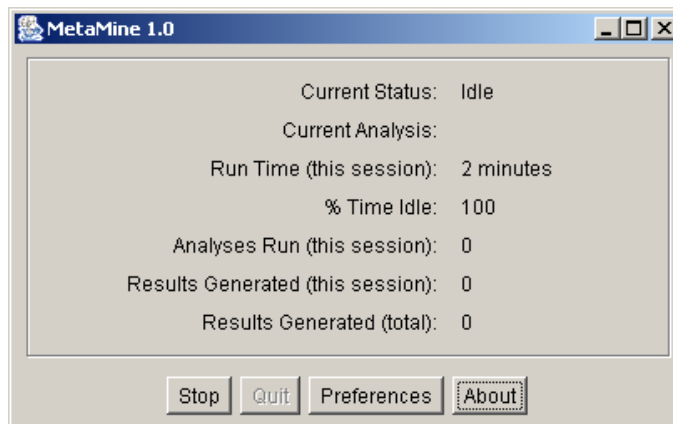


Figure 2-2 The MetaMine Console

MetaMine Preferences in GeNet

This is part of the GeNet web interface. It allows users to describe what things they are interested in: genes, gene lists, pathways, research groups, etc. They also can specify a maximum number of emails they wish to receive from MetaMine in a single day.

GeNet EXPRESSION DATA MANAGEMENT SILICON GENETICS

METAMINE > PREFERENCES Help

SAVE CHANGES CANCEL

Email Address for administrator:

None set.

Use the Personal Profile page to change your email address.

What is the maximum number of emails you want to receive from MetaMine per day?

1

Do you want to receive email in HTML format?

Yes No

How interested are you in each of the following genomes?

Affy Test Genome	Not Interested	<input checked="" type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	Very Interested
Bigfoot	Not Interested	<input checked="" type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	Very Interested
HG_U95A_version2	Not Interested	<input checked="" type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	Very Interested
HU35K ABCD	Not Interested	<input checked="" type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	Very Interested
HUFL	Not Interested	<input checked="" type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	Very Interested

Done Internet

Figure 2-3 MetaMine Preferences in GeNet

Emails

Every night, MetaMine looks through every result it has ever generated and compares them to the interest profiles for every user. It then attempts to determine which results a user would be most interested in (and has not already received email about), and sends brief letters describing them. These email contain URLs that can be used to view detailed reports on the results.

Search Reports Page

This is part of the GeNet web interface. It allows the user to search for MetaMine results by a variety of criteria: date generated, result number, genome, specific data that the result relates to, etc. The user can then view detailed reports for the matching results.

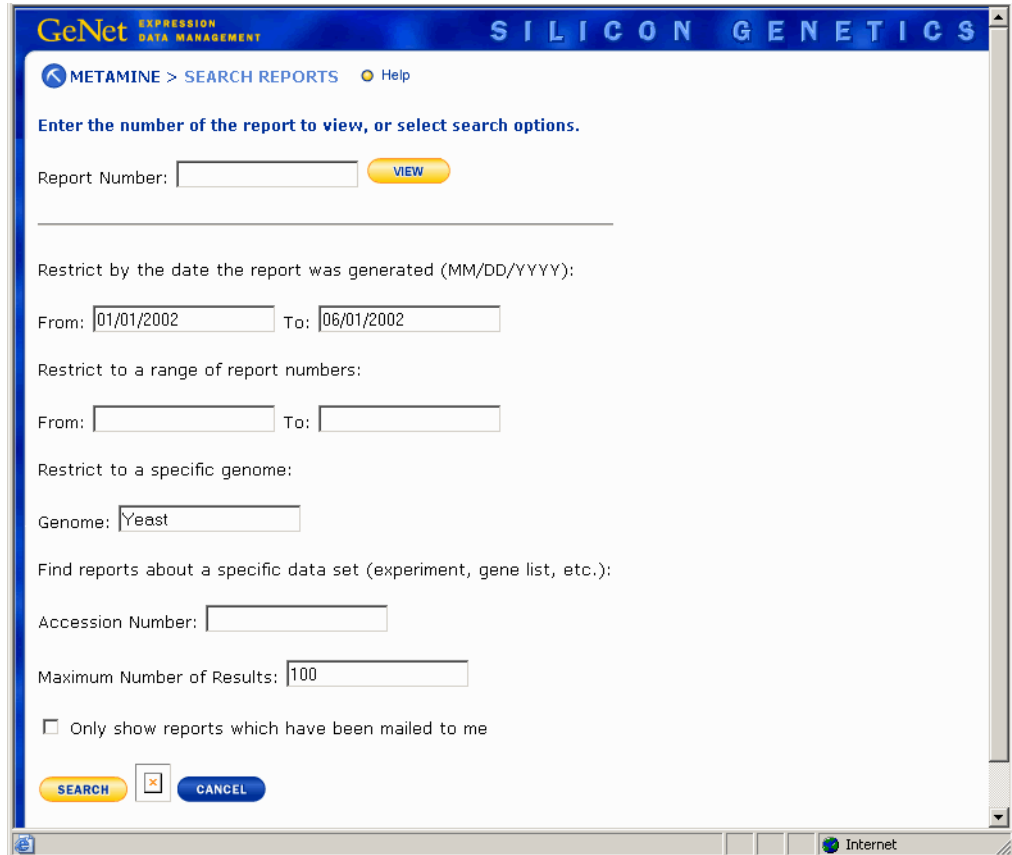


Figure 2-4 The MetaMine Search Reports screen in GeNet

MetaMine Results Display

This is part of the GeNet web interface. In the “Display” menu in the GeNet navigator frame, one of the options that can be selected is “MetaMine Results”. This gives the user a list of every MetaMine result that refers to the user’s currently selected data (the current experiment, gene list, pathway, etc.). Once again, the user can then jump to the detailed report on any of the results.

Full Reports

These are summaries of the results of interesting analyses. Such reports include:

- A description of the type of analysis that produced the interesting result
- Graphs depicting the behavior of the genes in the report.
- A list of references (from PubMed) for each of the genes identified by the report.

Some MetaMine results involve “new” data objects, which were created as part of the analysis and are not actually stored in GeNet. For example, an analysis result based on doing a hierarchical clustering of genes and then analyzing the genes in a cluster involves the list of genes in that cluster. Although this gene list is not actually stored in the GeNet repository, the user can still use the GeNet web interface to examine it, display graphs of it, etc. This is done by selecting a link in the detailed report for a result. The user can also

use the Download command in GeNet to transfer it to their local computer, and then import it into GeneSpring.

Availability of Results

Once an analysis result has been generated and saved to disk, it is kept forever. There are factors that may prevent it from actually being available, however.

First, the data it is based on must still exist in GeNet. Every result is based on an experiment, and in some cases may be based on other data as well such as gene lists or pathways. If any of the data on which a result is based is deleted from the GeNet server, that result will no longer be accessible.

Second, GeNet access permissions are used in determining whether a result should be accessible to a given user. If a user does not have read permission for all of the data on which the result is based, that result is inaccessible to them. It will never be mailed to them, it will not show up in lists of search results, etc.

Index

A

analyses 1-4
analysis summaries 2-4

C

configuration 1-3
console 2-2

E

emails 2-3

F

full reports 2-4

G

GeNet preferences 2-2

I

installing 1-2
 system requirements 1-2
interface 2-2
 console 2-2
 emails 2-3
 reports 2-4
 results 2-4

M

MetaMine overview 1-1

P

preferences in GeNet 2-2

R

reports 2-3
results
 availability of 2-6
results display 2-4

S

search reports 2-3

system requirements 1-2

