# CAGED v 1.0

## Cluster Analysis of Gene Expression Dynamics

User Manual

This User Manual is integral part of the computer program CAGED and it is subject to the same terms and conditions of use set forth in the License Agreement. In particular, installation and use of this program and its documentation are restricted to academic and research use by current members of academic and other non-profit organization. This documentation cannot be distributed without permission.

Microsoft, Windows, Microsoft Word are registered trademarks of Microsoft Corporation.

# Table of Contents

# Introduction

Welcome to CAGED and thank you for using it. CAGED is an acronym for Cluster Analysis of Gene Expression Dynamics and names a computer program developed to analyze temporal gene expression data.

This help includes a brief Overview of the program and of this help system. It also provides a description of the Basic Concepts necessary to use the program and understand its functionalities.

The Graphic User Interface of CAGED is designed as a Wizard interface, which takes the user through a number of steps (called screens) by gathering the necessary information at each step. Once you are familiar with the concepts behind the program, you can start looking at the Quick Guide, which takes you through a screen-by-screen description of the Wizard interface of CAGED. A Glossary of terms and a Bibliography of suggested readings are provided to help you to get a faster and deeper grasp on the program.

If you still have problems using the program, you can Contact us for help and the address in the next section. We strongly suggest you to read the terms and conditions of the License Agreement you subscribe by using CAGED.

## Overview

The term CAGED stands for Cluster Analysis of Gene Expression Dynamics. CAGED is a computer program performing Bayesian clustering on temporal gene expression data.

What: Temporal data are observations displaced along time. When studying the, say, cell-cycle of budding yeast during sporulation, we measure the expression values of a set of genes every hour for 24 hours. The resulting database is a sequence of observations, called time series, for each gene. The aim of a clustering program, like CAGED, is to group together these time series according to some similarity measure.

Note: Although CAGED is designed for temporal data, it can be used as a Bayesian clustering program on a-temporal expression data, because the statistical model of a-temporal data is a special case of the statistical model of temporal data, where data have no past (i.e. the Markov order is equal to 0).

How: CAGED is based on a method called Bayesian Clustering by Dynamics, which identifies the most probable set of clusters given a set of time series. Intuitively, we imagine that the time series we observe are generated by a set of stochastic processes. Imagine a set of EKG taken from a set of patients, some healthy, some with different cardiac diseases. We expect the EKG tracks of healthy patients to be generated by the same process (i.e. the process of an healthy organ), and we expect the tracks of the diseased patients to be similar according to their disease. We do not know how many processes are responsible for the data we observe, and we do not know which time series have been generated by a process. CAGED identifies the set of processes which are more likely to be responsible for the observed time series and assign each time series to the process which is most likely to be responsible for it. The result of the cluster analysis are a set of clusters grouping together the time series generated by the same process and a statistical model of the prototypical time series of each cluster. A description of the method underlying CAGED is available from the Methods section.

Why: There are several available clustering methods for gene expression data. What makes CAGED different is:

1. The ability to handle temporal gene expression profiles as time series.
2. The ability to cluster time series without a predefined similarity threshold.
3. The ability to identify a statistical model of the prototypical time series of each cluster.
4. The ability to assess whether the information is enough to discriminate among different clustering models (i.e. time series to different clusters).
5. The ability to provide a global measure of goodness of fit (in our case, the posterior probability of a clustering model) to assess how a clustering model explains the data.

A more detailed description of CAGED characteristics is available from the Features section.

## Requirements

There are some basic requirements you, your use of the program, and your hardware must meet in order to use CAGED.

### Legal requirements

As a user of the program, you must meet the following requirements:

1. The program can be used for academic, research, and educational purposes only.
2. Under no circumstances, CAGED can be used for diagnostic, clinical or otherwise medical purposes.
3. You must be a current member of an academic or other nonprofit organization and use the program for the aims and scopes of the institutions he or she belongs.

### Hardware requirements

Your computer must meet at least the following requirements:

1. 256 MB RAM
2. 700Mhz Processor
3. 30 MB of free disk space
4. Microsoft Windows 9x/NT/ME/2000

Note: These are minimum requirements to run the program at all. For very large databases, this configuration may not be sufficient. The performance of the program heavily depends on the available computer resources.

## License Agreement

Permission is granted herein to use this software for academic purposes only to current members of academic and non-profit organizations. Commercial use, unauthorized redistribution or reverse engineering of this program, or any part of it, are strictly prohibited. To use this program for commercial purposes, contact the developers at the address in the About box.

Any use for clinical, diagnostic, or otherwise medical purposes is absolutely forbidden.

This program is provided with no warranty whatsoever, and no responsibility is accepted by authors, sponsors, and copyright holders for any damage or loss deriving as a direct or indirect consequence of its use.

By installing and using this program, you accept the terms and conditions set forth in this License Agreement. This program is protected under US copyright law and international treaties. Unauthorized use, reproduction and distribution of this program, or any portion of it, may result in civil and criminal penalties.

This Agreement is governed by the United States laws.

## Credits

### The CAGED Team

Marco Ramoni, Children's Hospital Informatics Program, Harvard Medical School

Paola Sebastiani, Department of Mathematics and Statistics, University of Massachusetts, Amherst

Isaac Kohane, Children's Hospital Informatics Program, Harvard Medical School

### Special thanks to

Stefano Monti, Whitehead Institute, Massachusetts Institute of Technology

Alberto Riva, Children's Hospital Informatics Program, Harvard Medical School

### Software Credits



## Contacts

If you need to contact the developers to report a bug, ask for support or any other reason related to CAGED, please use this address:

CAGED Project
Children's Hospital Informatics Program
Harvard Medical School
Enders Research Building, Fifth Floor
320 Longwood Avenue
Boston, MA 02115

Phone: (617) 355-7424
Email: caged@chip.org
URL: http://genomethods.org/caged

# Basic Concepts

This section introduces some basic concepts behind CAGED methods and implementation.

This section is divided into four parts:

Theory: a brief summary of the theory behind CAGED methodology.

Methods: an outline of the analytical methods used by CAGED.

Features: a summary of the features of CAGED.

File Formats: a description of the file formats used and generated by CAGED.

This brief introduction is just a limited survey of CAGED methodology. For more detailed descriptions, consult the references in the Bibliography section.

## Theory

The theory underlying CAGED, called Bayesian Clustering by Dynamics, is based on two main concepts: clustering and time series. We explore the two notions in turn.

### Clustering

Clustering is the task of partitioning a set of individuals, in our case gene expression profiles, according to some similarity measure.

CAGED is designed to handle gene expression profiles. For CAGED, an individual is a set of repeated measurements of expression level of gene in different conditions. CAGED takes as input a database of observations on a set of genes. Typically, these observations are collected using microarray technology or other parallel gene expression measurement methods. Each gene is measured once in each, say, microarray and it is characterized by the set of measurements obtained in each microarray.

The standard approach to this problem is to select a similarity measure - such as correlation, Euclidean distance, or some measure of informational distance - decide an arbitrary similarity threshold, and group together those gene profiles which are more similar than prescribed by the similarity threshold.

CAGED takes a somewhat different approach to the problem. It takes a Bayesian approach. A Bayesian approach assumes that the data we observe are generated by a set of unknown processes. Imagine a set of EKG taken from a set of patients, some healthy, some with different cardiac diseases. We expect the EKG tracks of healthy patients to be generated by the same process (i.e. the process of an healthy organ), and we expect the tracks of the diseased patients to be similar according to their disease. We do not know how many processes are responsible for the data we observe, and we do not know which time series have been generated by a process.

We have now a new definition of similarity: two individuals (e.g. gene profiles) are similar if the same process generates them. Each process will therefore correspond to the cluster of all individuals generated by that process. This definition spares us the effort of defining an arbitrary similarity threshold to decide whether two individuals are similar. The similarity measure we need is also somewhat different: rather than a pair-wise comparison between two individuals, we need a global measure to decide which is the set of processes responsible for the data we observe and which process is responsible for each single individual.

In a Bayesian framework, we can define this similarity measure as the posterior probability of a clustering model - i.e. a way of clustering the individuals into groups - given the observed data. In principle, the task of CAGED is simply the one of exploring all ways of combining the observed individuals, computing its posterior probability given the observed data, and selecting the most probable one. In turn, this clustering model will also be a representation of the generating processes. So, by choosing the most probable clustering model we are actually selecting the most probable set of processes responsible for the observed data.

### Clustering Time series

One of the important features of CAGED is the ability to handle these data when the repeated measurements are not independent observations but rather observations of the same gene in different time points. Standard clustering methods of gene expression data typically assume that all the expression measurements of each gene are independent from one other. While this assumption of independence holds when expression measures are taken from independent biological samples (e.g. tissues from different patients, animals or cell-lines), it is known to be no longer valid when the observations are actually realizations of a temporal process, where each observation depends on its past. These kinds of temporally oriented data are termed time series.

In order to capture this dependency between observations in a computationally amenable representation, we model these time series as autoregressive models. A time series is said to follow an autoregressive model if each observation is some function of n previous observations. In other words, we assume that each point in the time series depends only on a limited set of n previous observations. The number n of previous observations (i.e. the length of the relevant past) is called Markov order of the autoregressive model. In particular, CAGED assumes that that the observations are stationary time series of continuous values in which each observation is a linear function of the previous n observations. Stationary, in this case, means that, if we could observe a time series to go to infinity, the values of the time series would periodically fluctuate around a mean value.

The main task of CAGED is to cluster these time series following the principles described in this section. The practicalities of this procedure are summarized in the Methods section.

## Methods

Recall from the Theory section that the Bayesian solution to the clustering task is to choose the clustering model with maximum posterior probability given the observed data. The task of exploring all possible clustering models, however, is unfeasible because the number of ways of combining time series into groups grows exponentially with the number of time series in the database.

Therefore, CAGED implements a form of hierarchical clustering which iteratively merges time series into clusters, so that each cluster groups the time series generated by the same process.

Although this agglomerative hierarchical clustering procedure makes the search feasible, the computational cost may still be too high to make the program scalable to large databases. In order to further reduce the computational effort, CAGED uses a similarity-driven heuristic search procedure. This heuristic procedure starts by computing a similarity measure between each pair of time series and evaluates the clustering model in which the two nearest time series are merged in the same cluster. A profile for this cluster is computed by averaging the time series in the cluster and the posterior probability of the resulting clustering model is computed. If the posterior probability of this clustering model is higher than the clustering model in which the two series are separated, the merging is accepted. The Dendrogram window shows the Bayes factor of each merging, that is, how many times more probable is the clustering model in which the two time series are merged than the clustering model in which they are kept separated.

When the merging is accepted, the algorithm goes on trying to merge other time series. If the merging does not increase the posterior probability of the clustering model, it is rejected and the merging between more distant time series is considered. If one of these merges increases the posterior probability of the

clustering model, the new merging is accepted and the program goes on trying to merge other time series. If none of the possible merges increases the posterior probability, then the algorithm stops and returns the set of clusters found so far.

Note that the clustering procedure is actually performed on the posterior probability of the clustering model and the similarity measure is only used to increase the speed of the search process and to prevent the algorithm from falling into local maxima. CAGED implements four type of similarity measures and their description is available in the Cluster Model section of this help.

As a matter of facts, CAGED does not use directly the posterior probability of the clustering model but rather a quantity proportional to it called marginal likelihood, taken on a logarithmic scale. Since the marginal likelihood is directly proportional to the posterior probability, the Bayes factor of two clustering model can be computed exactly.

## Features

These are main features of the method underlying CAGED. Click on the blue arrows to expand each description.

### The ability to handle temporal data

Temporal data are observations displaced along time. When studying the, say, cell cycle of budding yeast during sporulation, we measure the expression values of a set of genes every hour for 24 hours. The resulting database is a sequence of observations, called time series, for each gene. The aim of a clustering program, like CAGED, is to group together these time series according to some measure of similarity. Standard clustering methods for gene expression data assume that each expression measure in a time series is independent the previous observations. CAGED doesn't. Under a certain set of assumptions, CAGED takes into account the past of each observation in order to reconstruct the dynamics of a gene expression time series.

### No arbitrary threshold

Since CAGED uses the posterior probability of a clustering model and selects the most probable clustering model given the data, it does not require to set an arbitrary threshold on a similarity measure to decide whether two series must be grouped together in the same cluster.

### The ability to identify statistical models

In the process of generating a clustering model, CAGED actually constructs a statistical model for each cluster, which can be used to assess the goodness of fit of the statistical model and to predict the evolution of the genes on the basis of their generating process.

### The ability to assess information content

Since CAGED does not use arbitrary thresholds, it will discriminate between two time series if and only if there is enough information in the database to support the separation.

### The ability to take the context into account

The scoring metric used by CAGED to merge time series into clusters - i.e. the posterior probability of a clustering model given the data - is a global measure of fit, which takes into account the whole clustering model rather than a single merge of a pair of time series.

## File Formats

There are four file types used or generated by CAGED.

### Database Format

CAGED takes as input a database in ASCII tab delimited format. The columns in the database are separated by TAB characters and the rows are separated by RETURN character.

The structure of the database looks like this:

| | Gene name | Accession number | 0HR | 15MI |
|---|---|---|---|---|
| 1 | EST W95909 | W95909 | 1.0 | 0.72 |
| 2 | SID487537 H.sapiens mRNA for s | AA045003 | 1.0 | 1.58 |
| 3 | SID486735 Human peptidyl-proly | AA044605 | 1.0 | 1.1 |
| 4 | Homo sapiens protein 4.1-G mRNA | W88572 | 1.0 | 0.97 |
| 5 | SID469959 EST AA029909 | AA029909 | 1.0 | 1.21 |
| 6 | SID381721 EST AA059077 | AA059077 | 1.0 | 1.45 |
| 7 | SID471855 Lumican | AA035657 | 1.0 | 1.15 |
| 8 | EST AA180272 | AA044619 | 1.0 | 1.32 |

The database must be structured as follows:

Columns:

First column: A description of each gene in the database.

Second column: Gene accession number to an available on-line database, such as GenBank, UniGene or OMIM.

Other columns: The measured expression value of each gene. Each column represents an experimental condition or a time point.

Rows

First row: The first cell must be labeled "Gene name", the second cell must be labeled "Accession number". The remaining cells can be named with a label denotative of the relative time point or the experimental condition.

Other rows: Each cell contains the expression measurements for each gene in the database.

### CAGED Format

Once an analysis is finished, CAGED can save the results in a compact and fast loading binary format. You can reload the saved sessions by choosing the second option in the Getting Started screen of the interface.

### HTML Format

You can save the results of an analysis on a Report file. This report includes a description of the data, an outline of the method and a summary of the results, together with the relevant images. This file is saved in HTML format so that it can be posted on the WWW as supplementary material of a publication or directly edited in a HTML capable editor, such as Microsoft Word.

### Images Format

All the images generated by CAGED during the Analysis process can be saved in PNG or JPG format. PNG (Portable Network Graphics) is the recommended format because of its small footprint, its quality, and the generous maximum limits (2G pixels x 2G pixels).

# Quick Guide

CAGED user interface is structured as a Wizard interface. A Wizard interface is structured as a set of screens guiding the user through the task of the program.

At the bottom of each screen there are six buttons. From left to right they are:

About: pops up the About box, containing License, Contact, and Credits information.

Help: invokes this help file.

Cancel: exits the program.

Back: goes back to the previous screen.

Next: steps forward to the next screen.

Finish: exits the program when the task is completed.

CAGED Wizard interface comprises five screens:

Welcome Screen: The opening screen, containing logo and License Agreement.

Getting Started: The screen in which the user decides the task to undertake: run the program on a database, load a saved session, or open a website (disabled).

Analysis: The screen in which the users sets the parameters for the cluster analysis. The program skips this screen if the user chooses to load a saved session.

Cluster Model: The screen in which the program returns the results of the cluster analysis. From this screen, several windows can be evoked to explore the generated model. These windows are described in the Other Windows section of this help.

Pack and Go!: The final screen of the program, where the user can save the results of the session and generate a report of the analysis in HTML format.

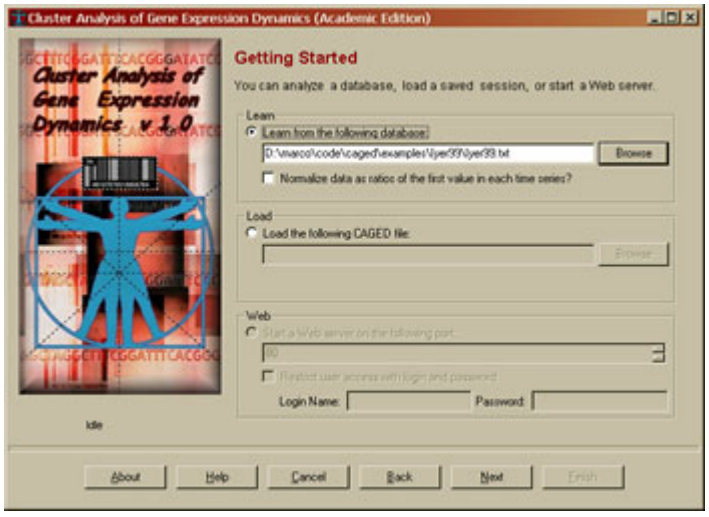
## Welcome Screen

When the program starts, it opens a Welcome screen containing a brief welcome message together with licensing and copyright information.


Hit the Next button to reach the Getting Started screen.

## Getting Started

In the Getting Started screen, the user decides what to do in the session.



The user is given three options:

Learn: Start the cluster analysis of a database. If you choose this option, you must specify the pathname of a valid database. Optionally, you can decide to convert the current database into ratios of the first data point, by selecting the optional thick below the database load slot. If you choose this filter, you can also decide to exclude all the series with no value above or below a fold change.

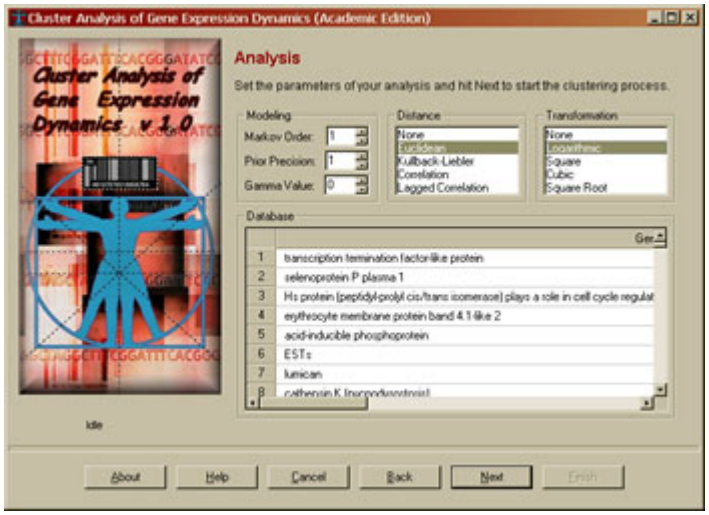
Load: Load a previously saved session in CAGED format. If you do so, the program will skip the Analysis screen and go directly to the Cluster Model screen.

Web: Open a Web site offering the functionalities of CAGED over the World Wide Web (currently disabled).

Hit the Next button to reach the Analysis screen.

## Analysis

The Analysis screen collects the information required to analyze your database.




The screen is divided in four labeled regions.

### Modeling

There are three numerical parameters to be set by the user:

Markov Order: The Markov order is the number of past time points relevant to the present. The value ranges between 0 and 10. See the Methods section for an explanation. When the value is 0, and no value in the past is relevant to

the present, CAGED behaves like a standard clustering method assuming that all observations about a gene expression level are independent.

Prior Precision: The prior precision is the sample size of the imaginary observations upon which the prior distribution is built. It ranges between 1 and 100. We suggest starting with the default number 1 and repeating the analysis for different values to assess the robustness of the generated model with respect to this parameter.

Gamma Value: The gamma value is the rate to zero of the prior precision. Default value is 0. You should not change this value but you may want to run different analysis to check the robustness of your model with respect to this parameter.

Bayes Factor: Imposes a minimum limit on the Bayes factor to accept the merging of two series or two clusters: two series (or clusters) will be merged if the Bayes factor of their merging is at least the value prescribed in this slot.

### Distance

These are the similarity measures used in the heuristic search process. There are five options:

None: No distance used. This option is very dangerous as it allows the program to get into local minima.

Euclidean: The point-wise geometric distance between two time series. This is the default value.

Kullback-Leibler: An entropy-based distance computing the information content of two time series.

Correlation: Standard correlation between two time series.

Lagged Correlation: Correlation measure that takes into account the autoregressive structure of the statistical model. The delay of the correlation measure is the same as the selected Markov order.

These distances are used by the heuristic search process and they are not, per se, responsible for the clustering model. Nonetheless, they may result in different clustering models. You can, however, decide which heuristic is better for your problem by running the program with different distances and choosing the clustering model with the highest marginal likelihood. You can see the marginal likelihood of a clustering model in the window Clustering Properties window in the Cluster Model screen. Transformation Data transformations are functions applied to all data in the database. There are four transformations implemented in CAGED and five options:

None: No transformation. This is the default option.

Logarithmic: Each value is natural logarithm transformed.

Square: Each value elevated to the power of 2.

Cubic: Each value is elevated to the power of 3.

Square Root: Each value is taken as its square root.

Cubic Root: Each value is taken as its cubic root.

The logarithmic transformation is useful when data are actually ratios and you want to treat symmetrically values below and above 1. Do not to use this transformation simply as a smoothing function but only if there is some design reason to do so, such as the case of ratio values.

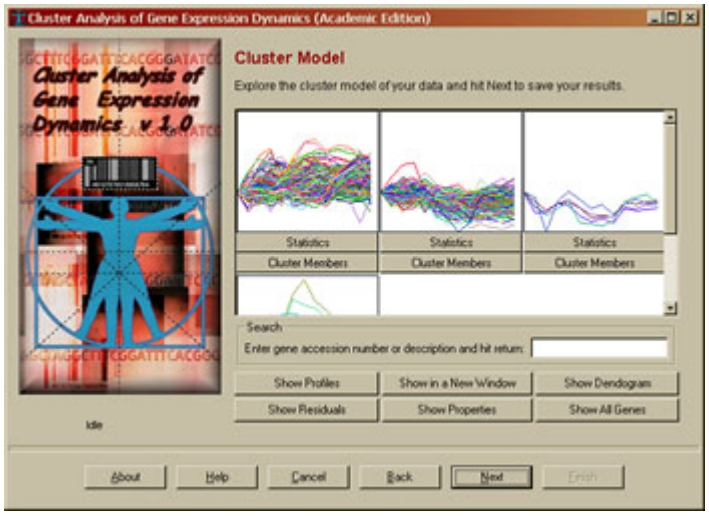
### Database

The database window shows the user the loaded database. This is useful to check to check that the database has been properly loaded.

Hit the Next button to start the analysis process and reach the Cluster Model screen.

## Cluster Model

The Cluster Model screen contains the actual results of your analysis.

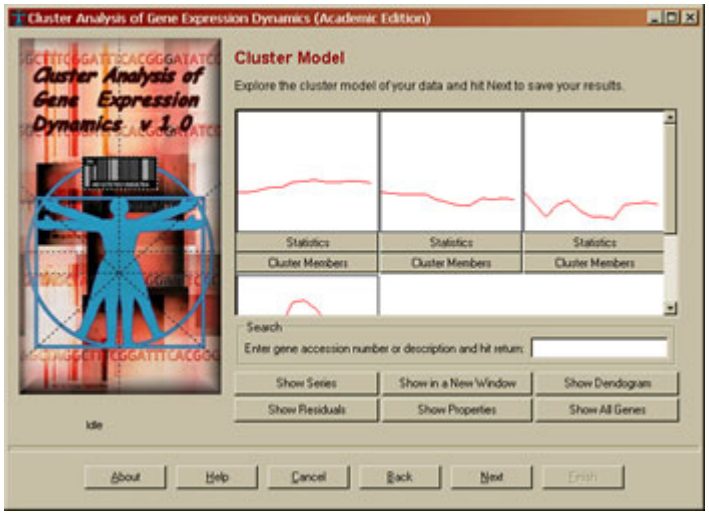



The screen is divided in three main parts:

### Clusters Display

The top part of the screen contains a set of panels, one for each of the discovered clusters. Each panel is composed by three elements:

Display: A drawing region. By default, this draws the time series members of each cluster. A mouse click on the drawing region evokes the Cluster window of the cluster. If you click on the button Show Profiles, the display panels will show the profiles of each cluster:




Statistics: The button Statistics displays a basic description of the statistical model of the cluster in the Statistics window.

Cluster Members: The button Cluster Members displays the list of all genes member of the cluster in the Cluster Members window.

## Search Panel

Type part of a gene description or accession number to display a window with all the genes with the string in their description or accession number.

## Buttons Panel

The third component of the Cluster Model screen is a set of six buttons:

Show Profiles/Series: Choose between the display of the cluster profile or the cluster member series in the Display section of the Cluster Model screen.

Show in a New Window: Pop up an independent Clusters window.

Show Dendrogram: Pop up the Dendrogram window.

Show Residuals: Pop up the Residuals window.

Show Properties: Pop up the Properties window.

Show All Genes: Pop up window containing all genes labeled by their cluster.

Hit the Next button to move to the Pack and Go! screen.

## Pack and Go!

The Pack and Go! screen concludes a CAGED session by offering two ways of saving the results of your session.


The screen is divided in two parts.

### Save in CAGED Format

Save the results in a binary format for fast loading. You can save the results of your session, so you can load and see them again in the future.

### Save in HTML Format

Save a complete report of your results in HTML format. Images are saved in PNG format. You can use this report to publish your results on the WWW or to include sections in your paper.

As the generation of the dendrogram is the most resource demanding operation in the program, low resources machines may be unable to generate a dendrogram for large datasets. A thinkable box makes the inclusion of the dendrogram in the report optional.

Each entry in the dendrogram can be linked to a resource on the Internet, by merging the accession number in the second column of your database to one of the following resources:

UnChip (Affymetrix): Use this item if the second column of your database reports an Affymetrix code. Linked to http://unchip.org.

UniGene: Use this item if the second column of your database reports a UniGene cluster code. Linked to http://www.ncbi.nlm.nih.gov/UniGene.

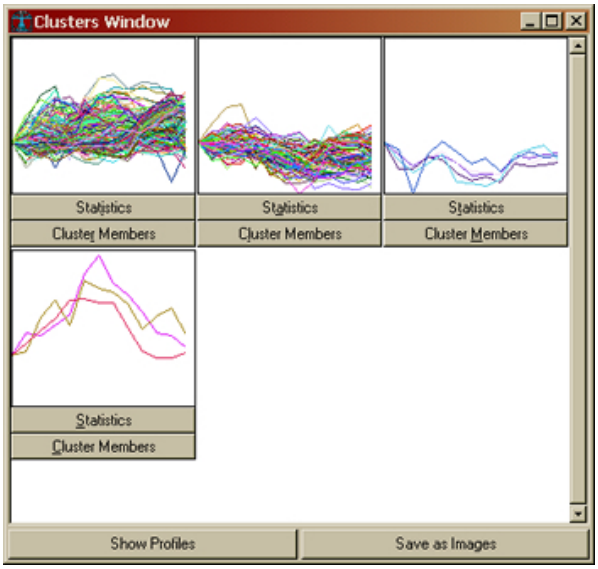
Entrez Sequence Database (GenBank): Use this item if the second column is a GenBank code. Linked to http://www.ncbi.nlm.nih.gov/Entrez/

## Other Windows

There are several other windows in CAGED, besides the main Wizard interface, to allow you a better exploration of the clustering model generated by CAGED.

### Clusters Window

The Cluster Window is the expansion of the Display section of the Cluster Model.




The two buttons at the bottom of the window are:

Show Series/Profiles: Switch between cluster profiles and time series.

Save as Images: Save a cluster image (profile or series) in PNG/JPG format.

There are two Mouse actions:

Left Click: A left mouse click on a cluster evokes an independent window.

Right Click: A right mouse click on the window evokes a Printing menu.

### Cluster Window

The Cluster window shows either the cluster profile or its members. If the box Constrained to global proportions is checked, both profiles and member time series will be drawn proportionally to all the other.

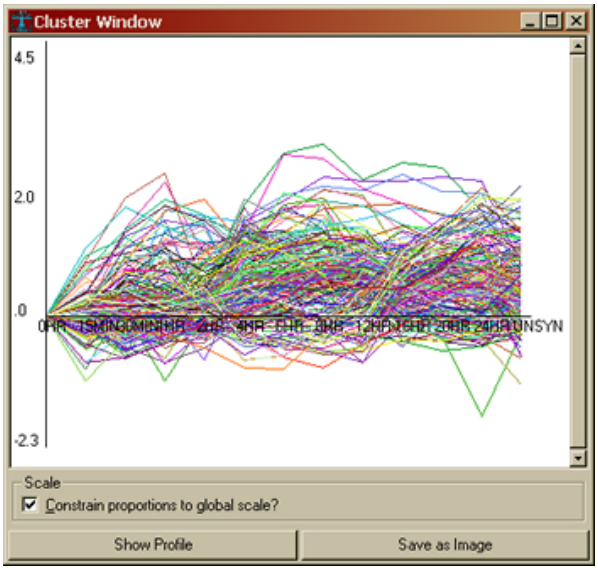
The two buttons at the bottom of the window are:

Show Series/Profiles: Switch between cluster profile and members.

Save as Image: Save the cluster image (profile or series) in PNG/JPG format.

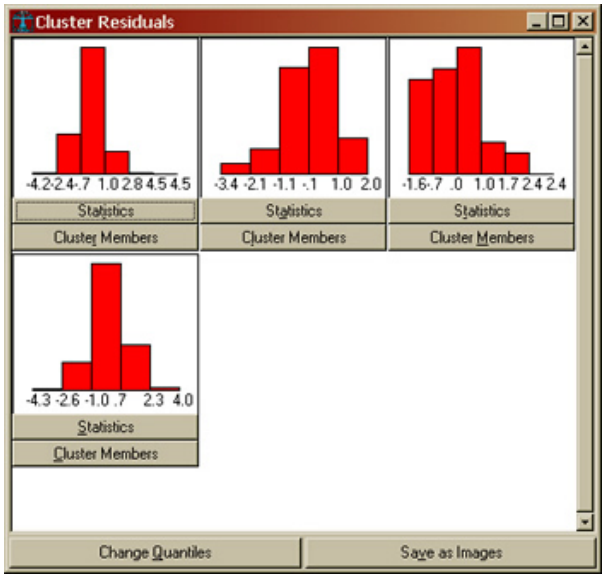
There is only one Mouse action:

Right Click: A right mouse click on the window evokes a Printing menu.




## Residuals Window

The Residuals window shows the residuals of the generated clusters.. They are used for diagnostic purposes and should have an approximately normal distribution. Click on a display to open a window.



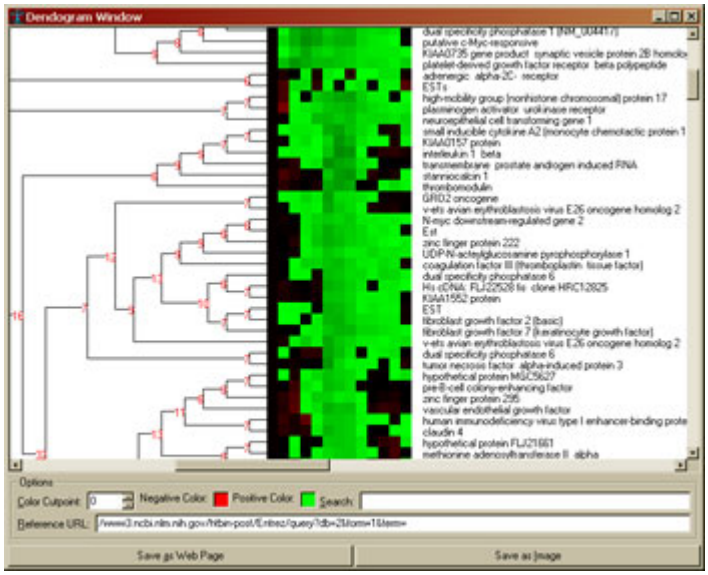
The two buttons at the bottom of the window are:

Change Quantiles: Changes the quatiles of the residual pictures cycling between 2 and 10.

Save as Images: Save each the image of each cluster residuals in PNG or JPG format.

## Dendrogram Window

Dendrograms are a popular representation for hierarchical clustering of gene expression data.




The Dendrogram window is divided in three parts.

Dendrogram viewer: The dendrogram itself is divided in three parts:

Binary tree: A binary tree representing the way in which the time series have been merged by the clustering algorithm. The numbers at the branches of the tree report the Bayes factor of the merging and tells you how many times merging the two time series is making the model more probable than keeping them separated. Bayes factor is reported on a logarithmic scale.

Series intensities: A set of colored squares. Each series is represented by an horizontal sequence of squares. The squares are colored in one of two colors. Each color tells you that the value above or below a certain cutpoint. The intensity of the color tells you how far are the values from the cutpoint. In the picture above, the cutpoint is 1 and the colors are red and green. The higher the value of an observation

above 1, the highest is the intensity of the red. The lowest the value of an observation below 1, the highest is the intensity of the green.

Descriptions: A description of each gene, first column of the database.

Options Panel: A control panel allows you to change the cutpoint, the two colors and allows you to locate a gene by typing any part of its description or its accession number. It also includes a URL reference to query a particular resource using the accession number in the second column of the database. The options are described in the Pack and Go! section.

Button Panel: The buttons save the dendrogram in two formats:

Save as Image: Dendrogram is saved in PNG or JPG format.

Save as Web Page: The content of the Dendrogram is saved in HTML format. Before saving, the user is prompted to enter a title and the template URL for the accession numbers in the second column of the database. The saved HTML page will be active: a click on a gene description takes the browser to the appropriate page in the public database. Details are in the Pack and Go! section.

There are two Mouse actions available on this window:

Left Click: A left mouse click on a series evokes a Web Browser to query the associated database using the Accession number provided for that series.

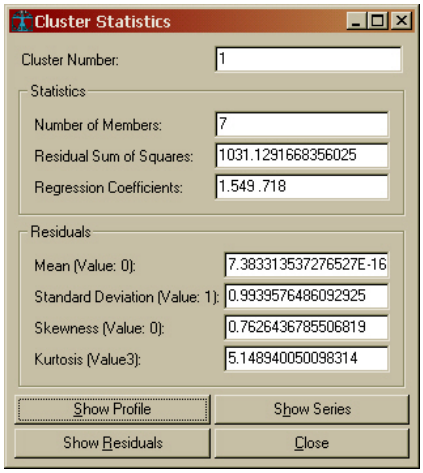
Right Click: A right mouse click on the window evokes a Printing menu.

## Properties Window

Display the general properties of the current clustering model.

## Statistics Window

Display the statistical model of a cluster.



The Statistics display displays the parameters of the statistical model.

The second display, marked Residuals, displays the basic statistics of the standardized residuals to help diagnose the goodness of fit of the model. The values for each item are the values expected for a standard normal distribution.

## Cluster Members Window

List all genes members of a cluster.


The button at the bottom allows the user to save the list on file in ASCII format.

## All Genes Window

List all genes in the database labeled with their cluster membership.


The button at the bottom allows the user to save the list on file in ASCII format.

# Glossary

This is a list of terms used in this manual. For more details, you can refer to literature listed in the Bibliography section.

**Autoregressive model**: a statistical model representing time series. This representation assumes that each point in the time series depends only on a limited set of previous observations. The number of previous observations (i.e. the length of the relevant past) is called Markov order of the autoregressive model.

**Bayes factor**: The ratio between the posterior probabilities of two statistical models. This value tells us how many times more (or less) probable is the first model with respect to the other. In CAGED, this value is usually given on a logarithmic scale.

**Bayesian Clustering by Dynamics**: a clustering method designed to group together a set of time series by discovering the most probable set of clusters responsible for the observed time series.

**Bayesian model selection**: The model selection process of choosing the statistical model with the highest posterior probability.

**Cluster**: a group of individual elements in a database. In the case of CAGED, it is a group of genes, or better, a group of time series representing the temporal profile of a gene, represented by a row in the input database.

**Cluster profile**: Prototype representation of a cluster obtained by averaging the time series in the cluster.

**Clustering**: the process of generating a clustering model by grouping together a set of individuals (e.g. time series).

**Clustering model**: a way of grouping elements of a database. In the case of CAGED, it is a group of genes, or better, a group of time series representing the temporal profile of a gene, represented by a row in the input database.

**Gamma value**: The gamma value is the rate to zero of the prior precision. Default is 0, which represents the case of perfect ignorance.

**Marginal likelihood**: A quantity proportional to the posterior probability of a clustering model. This is the quantity actually used by CAGED to select the most probable set of clusters.

**Markov order**: The number of past time points relevant to the present.

**Model selection**: The process of selecting a particular statistical model.

**Prior distribution**: The probability distribution of an observer before considering any data.

**Prior precision**: The size of the sample upon which the prior distribution is built.

**Posterior probability**: The probability of an event (or a model) given the observation of some evidence.

**Similarity measure**: A numerical measure of how two individuals (e.g. time series) are similar to each other. Similarity measures are used in CAGED as heuristic tools to speed up the search process. There are four similarity measures implemented in CAGED, as described in the Methods section.

**Similarity threshold**: A fixed numeric threshold over a similarity measure to decide whether two individuals (e.g. time series) are actually similar.

**Statistical Model**: A statistical representation of a set of observations.

**Time Series**: a sequence of dependent observations.

# Bibliography

M. Ramoni and P. Sebastiani. Bayesian Methods for Intelligent Data Analysis. In M. Berthold and D.J. Hand, editors, *Intelligent Data Analysis: An Introduction*, Springer, New York, NY, 1999.

*A general introduction to Bayesian data analytic methods. It explains in detail the Bayesian vision and some important concept used in CAGED.*

M. Ramoni, P. Sebastiani and P. Cohen. Bayesian Clustering by Dynamics. *Machine Learning*. To appear.

*The description of Bayesian Clustering by Dynamics. Not really the method used by CAGED but a general framework for it.*

P. Sebastiani and M. Ramoni. Bayesian Clustering of Continuous Time Series. Under review.

*The formal and detailed description of the clustering method used by CAGED.*

M. Ramoni, P. Sebastiani and I. Kohane. Temporal Profiling of Gene Expression Data. Under review.

*The paper describing the application of CAGED to the analysis of temporal profiles of gene expression data.*