

DNA Microarray Data Analysis

IIRIS HOVATTA, KATJA KIMPPA, ANTTI LEHMUSSOLA, TOMI PASANEN,
JANNA SAARELA, ILANA SAARIKKO, JUHA SAHARINEN, PEKKA TIIKKAINEN
TEEMU TOIVANEN, MARTTI TOLVANEN, MAUNO VIHINEN AND GARRY WONG
EDITORS JARNO TUIMALA AND M. MINNA LAINE

CSC

CSC – Scientific Computing Ltd. is a non-profit organization for high-performance computing and networking in Finland. CSC is owned by the Ministry of Education. CSC runs a national large-scale facility for computational science and engineering and supports the university and research community. CSC is also responsible for the operations of the Finnish University and Research Network (FUNET).

All rights reserved. The PDF version of this book or parts of it can be used in Finnish universities as course material, provided that this copyright notice is included. However, this publication may not be sold or included as part of other publications without permission of the publisher.

© The authors and
CSC – Scientific Computing Ltd.
2005

Second edition

ISBN 952-5520-11-0 (print)

ISBN 952-5520-12-9 (PDF)

<http://www.csc.fi/oppaat/siru/>

<http://www.csc.fi/molbio/arraybook/>

Printed at
Picaset Oy
Helsinki 2005

List of Contributors

Iiris Hovatta
National Public Health Institute
Haartmaninkatu 8
FI-00290 Helsinki
Finland

Katja Kimppa
PerkinElmer Life Sciences and Analytical Sciences
- Wallac Oy
P.O.Box 10
FI-20101 Turku
Finland

M. Minna Laine
CSC, the Finnish IT center for science
Keilaranta 14
FI-02101 Espoo
Finland

Antti Lehmussola
Tampere University of Technology
P.O.Box 553
FI-33101 Tampere
Finland

Tomi Pasanen
University of Helsinki
P.O.Box 68
FI-00014 University of Helsinki
Finland

Janna Saarela
Biomedicum Biochip Center
Haartmaninkatu 8
FI-00290 Helsinki
Finland

Ilana Saarikko
University of Helsinki
P.O.Box 68
FI-00014 University of Helsinki
Finland

Juha Saharinen
National Public Health Institute
Haartmaninkatu 8
FI-00290 Helsinki
Finland

Pekka Tiikkainen
VTT
P.O.Box 106
FI-20521 Turku
Finland

Teemu Toivanen
Centre for Biotechnology
Tykistökatu 6
FI-20521 Turku
Finland

Martti Tolvanen
Institute of Medical Technology
Biokatu 8
FI-33520 Tampere
Finland

Jarno Tuimala
CSC, the Finnish IT center for science
Keilaranta 14
FI-02101 Espoo
Finland

Mauno Vihinen
Institute of Medical Technology
Biokatu 8
FI-33520 Tampere
Finland

Garry Wong
A. I. Virtanen -institute
University of Kuopio
FI-70211 Kuopio
Finland

3 Web extra: Cluster analysis of microarray information

Mauno Vihinen and Jarno Tuimala

3.1 GeneSpring and clustering

GeneSpring offers five different clustering and classification algorithms, one of which is a supervised method. These are hierarchical clustering (gene and experiment trees), K-means clustering, self-organizing maps, principal component analysis, and parameter value prediction. Clustering tools are invoked from the menu *Tools->Clustering*, the supervised classification method can be found from *Tools->Predict parameter values*, and the principal component analysis is located in *Tools->Principal components analysis*.

3.1.1 Clustering tool

The clustering tool has four fields, which need to be filled in before running the analysis (Figure 3.1). From top to down, the first box indicates the gene list to be clustered. Initially the list is the same that was highlighted when the clustering tool was invoked, but it can be changed from the navigator bar on the left. Next box contains the information about the experiment to be clustered. This can also be easily changed from the navigator bar on the left. The pull-down menu offers the choice of three clustering algorithms (hierarchical, K-means and SOM). The setting for the current analysis is in the box below the pull-down menu. For K-means, the number of clusters needs to be specified, as well as the number of iterations and the desired measure of similarity.

GeneSpring has several different similarity measures, which fall into the following categories: correlation, confidence, and distance. The selection of the similarity measure should be given some thought, because it significantly affects the generated results. Pearson's correlation emphasizes both over- and underexpressed genes, and the Standard correlation finds especially overexpressed genes. Spearman's correlation is highly similar to Pearson's correlation except it uses ranks for

the calculation of the correlation coefficient (and is thus a nonparametric measure of correlation). Distance measures the euclidian distance between two gene expression profiles. It is calculated as the square root of averaged squared deviation of the profiles. Spearman's confidence measures the probability of getting a correlation of S or higher by chance alone, if the true correlation is zero.

If there is only one measurement per gene (*i.e.*, one chip), only the distance measure can be used for the clustering of the genes. If there are two replicates of every gene, the Standard correlation can also be used. If there are three replicates, Pearson's correlation becomes available, and with five replicates the confidence measures can be used.

In other words, the decision about the applied similarity measure depends on the biological question you are interested in, and the amount of replicates in your dataset.

After specifying the aforementioned settings, the run can be started by clicking the Start-button on the bottom of the window. When the run has ended, you can name and save the clustering result. The result appears on the main screen of GeneSpring. Thereafter, all the results can be found from the navigator bar under the folder classification. Note that the hierarchical clustering results are stored under two different folders, Gene Trees and Experiment Trees.

After viewing the clustering results, you can get back to the original view by selecting *View->Unsplit window*. Hierarchical clustering results can be dismissed, for example, by selecting *View->blocks*.

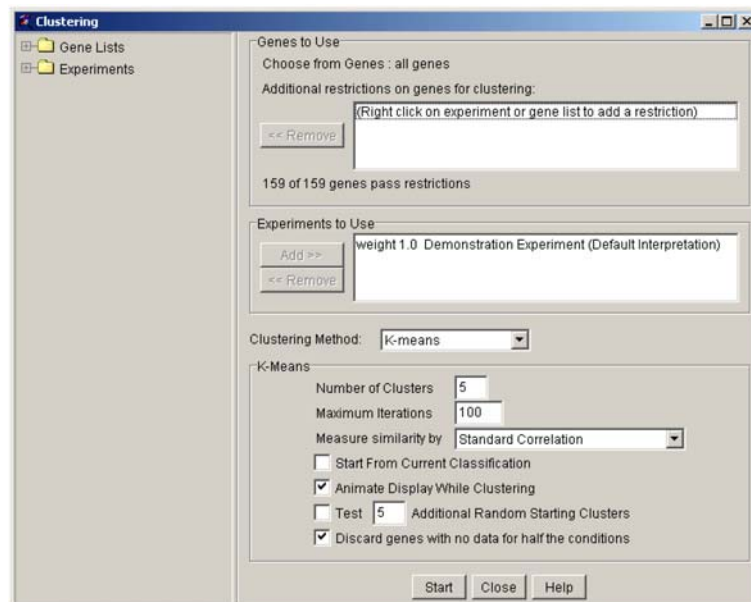


Figure 3.1: *The clustering tool in GeneSpring.*

3.1.2 Principal components analysis tool

Principal component analysis can be used for creating a set of the most significant expression patterns from the data. It can also be used for checking the results of the clustering methods. After selecting *Tools->Principal components method* the analysis is run, and the results are displayed. The opening window contains the most significant profiles. Double-clicking one profile transfers the view to the Gene Inspector, where genes with a similar expression profile can be searched.

The results are displayed in a scatter plot format in the main window. If you want to compare PCA with some clustering results, you can right-click on one clustering result in the navigator bar, and select Set as coloring scheme from the appearing menu. A good clustering result often creates clusters that do not overlap with each other in the PCA scatter plot.

3.1.3 Predict parameter value tool

The predict parameter value tool (Figure 3.2) is used in situations, where we have a certain set of known samples, and based on these, we want to predict in which group the new, unknown samples fall. For example, if we have information about the leukemia type of 60 samples, we can find the genes, which differentiate these leukemia types from each other. After finding the suitable set of genes, the identity of the unknown samples can be predicted. GeneSpring uses the K-means algorithm described by Golub *et al.*

Training and test sets (experiments) need to be specified. GeneSpring also needs to know which parameter contains the information about the groups to be compared (parameter to predict). After cross-validating the test set, the test set can be predicted. The result of the analysis is a prediction of the test set sample identities and a set of genes differentiating the selected groups.

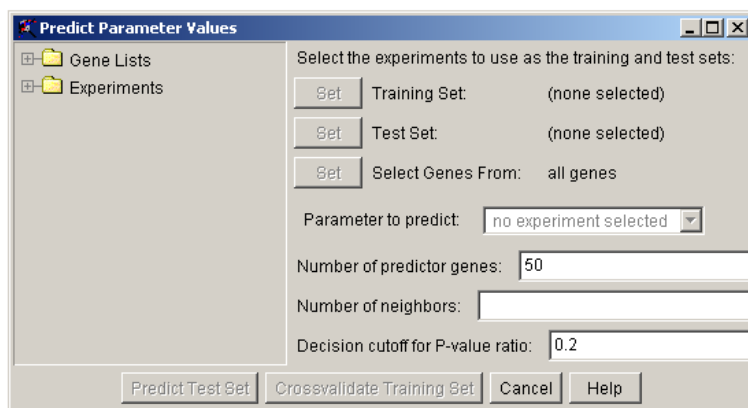


Figure 3.2: The predict parameter value -classification tool in GeneSpring.