W616–W620 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki500

GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data

Juan M. Vaquerizas[1], Lucía Conde[1], Patricio Yankilevich[1], Amaya Cabezón[1], Pablo Minguez[1], Ramón Díaz-Uriarte[1], Fátima Al-Shahrour[1], Javier Herrero[1,2] and Joaquín Dopazo[1,3,*]

[1]Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain, [2]Ensembl Team, EMBL-EBI, Hinxton, Cambridge, UK and [3]Functional Genomics Node, INB, Centro de Investigación Príncipe Felipe, Autopista del Saler 16, 46013 Valencia, Spain

Received February 14, 2005; Revised April 9, 2005; Accepted May 3, 2005

ABSTRACT

The Gene Expression Profile Analysis Suite, GEPAS, has been running for more than three years. With >76 000 experiments analysed during the last year and a daily average of almost 300 analyses, GEPAS can be considered a well-established and widely used platform for gene expression microarray data analysis. GEPAS is oriented to the analysis of whole series of experiments. Its design and development have been driven by the demands of the biomedical community, probably the most active collective in the field of microarray users. Although clustering methods have obviously been implemented in GEPAS, our interest has focused more on methods for finding genes differentially expressed among distinct classes of experiments or correlated to diverse clinical outcomes, as well as on building predictors. There is also a great interest in CGH-arrays which fostered the development of the corresponding tool in GEPAS: InSilicoCGH. Much effort has been invested in GEPAS for developing and implementing efficient methods for functional annotation of experiments in the proper statistical framework. Thus, the popular FatiGO has expanded to a suite of programs for functional annotation of experiments, including information on transcription factor binding sites, chromosomal location and tissues. The web-based pipeline for microarray gene expression data, GEPAS, is available at http://www.gepas.org.

INTRODUCTION

GEPAS, which stands for Gene Expression Profile Analysis Suite, is a web tool designed and oriented to the analysis of DNA microarray gene expression experiments. The emphasis in the development of new tools for GEPAS has been driven by the requirements of data analysis in the most active fields using microarray technologies, which are, without doubt, biomedical applications [e.g. (1–4)]. As a consequence, much stress has been put on the implementation of proper methods for gene selection, predictors, CGH-arrays and functional annotation of experiments. More classical data analysis approaches, such as clustering, have also been incorporated into GEPAS, as well as different options for data preprocessing.

GEPAS has been conceived as an integrated web-based pipeline for the analysis of gene expression patterns where different methods can be used within an integrated interface that provides a user-friendly environment to end users. The way in which the methods are connected has been designed to guide the user by suggesting all the available possibilities to continue with the analysis and to prevent possible inappropriate uses of the tools.

GEPAS, which was originally the backbone of the pipeline of microarray data analysis of the CNIO, was made public three years ago and first published in 2003 (5,6). In the years since, GEPAS has become a *de facto* standard for many researchers and its use has undergone a spectacular growth. In terms of the scope of analysis, GEPAS is the most complete web-based resource that can be found nowadays.

Our aim is to keep GEPAS ‘living’ by the continuous addition of new algorithms. Here we report the new modules, some trends observed in its use and some novelties.

*To whom correspondence should be addressed. Tel: +34 96 328 96 80; Fax: +34 96 328 97 01; Email: jdopazo@ochoa.fib.es
Present address:
Juan M. Vaquerizas, Lucía Conde, Pablo Minguez, Fátima Al-Shahrour and Joaquín Dopazo, Bioinformatics Department, Centro de Investigación Príncipe Felipe, Autopista del Saler 16, 46013, Valencia, Spain

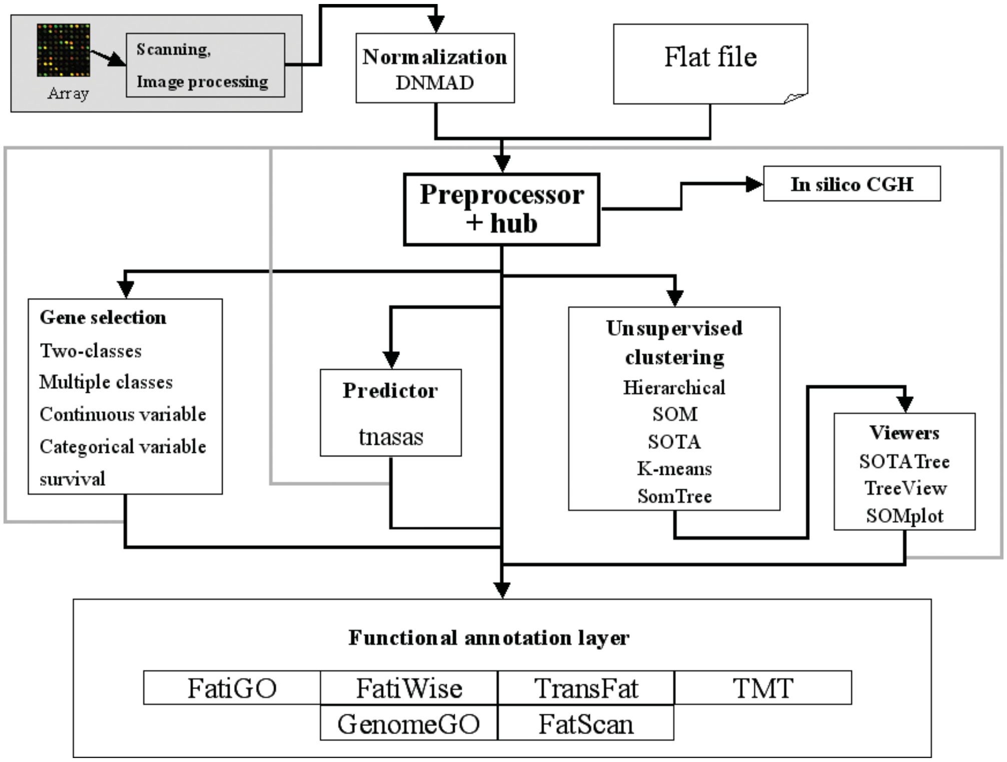

Figure 1. The GEPAS pipeline. The figure summarizes the most important features of the GEPAS pipeline. Black arrows show the flow of information from the raw data to the three main types of analysis: CGH-array, unsupervised clustering and supervised analysis (gene selection or predictors). Functional annotation is possible from the latter two options. Grey arrows represent the possibility to re-analyse parts of the experiments.

SCOPE OF GEPAS

As previously mentioned, GEPAS is experiment oriented. This means that facilities for data manipulation such as rows and columns management are deliberately absent in its design. With the exception of the module DNMAD, which can take as input Genepix (Axon instruments) GPR files from a scanner (see below), GEPAS accepts as input data already preselected (usually coming from a database) in a very simple format: a tab-delimited text file containing genes in rows and experiments in columns (except the first column, which contains the identifiers for the genes).

Several preprocessing facilities are provided. These are normalization along with different kinds of data transformation such as missing value imputation, filtering of 'flat' patterns and extraction of genes based on functional properties.

GEPAS permits two main types of experimental designs: those oriented towards class discovery, for which different clustering methods are available, and those related to supervised questions, which include mainly gene selection and building predictors. GEPAS includes two tools for dealing with both of these problems.

In addition, there is a great interest now in tools that allow CGH-arrays to be handled. GEPAS includes a module for mapping either genomic or mRNA hybridizations over the corresponding chromosomal locations, with different facilities for data visualization.

Finally, GEPAS provides a module for functional annotation of experiments that includes the popular FatiGO (8), as well as a variety of new tools.

GEPAS AT A GLANCE

GEPAS includes a number of interconnected tools implemented as individual modules that can be used either independently or within the pipeline (Figure 1). Since the previous version (6), GEPAS has undergone a number of technical improvements which have not had much impact on its external aspect but have notably changed its performance. Internal links among modules have been improved and redesigned in order to avoid wrong pathways in the pipeline. Some CPU-intensive modules have been moved to dedicated computers (in particular DNMAD, Pomelo and Tnasas). The structure of GEPAS is as follows.

Preprocessing

DNMAD (9) is for normalization using print-tip loess (10,11) (http://www.bioconductor.org), with different possibilities. Some additional options have been included in this new version: the possibility of using a spot's flags, optional use of background subtraction and the possibility of using global loess (instead of print-tip). We have also included a better management of flagged dots, new diagnostic plots (the density plots for either raw or background-corrected red and green channels) and automatic dye-swap. DNMAD can take as input Genepix (Axon instruments) GPR files.

Preprocessor (12) performs some preprocessing of the data (log-transformations, standardizations, imputation of missing values, etc.). Data can also be filtered on the basis of their functional labels [GO terms (13)] using the Knowledge Filtering module (6).

IDconverter, a new module, maps lists of accession numbers and identifiers among different clone, gene or protein standards. IDconverter includes distinct levels of information such as gene level (gene HUGO name, Ensembl gene, Unigene cluster, LocusLink, RefSeq, gene location, gene description), clone level (Affymetrix, GenBank accession number, IMAGE Clone ID) and protein level (SwissProt, TrEMBL, now UNIPROT). Chromosomal locations are obtained from Ensembl.

Analysis

Unsupervised clustering includes different methods such as aggregative clustering (14), SOTA (15,16), SOM (17), *K*-means (18) (which is a new addition in this version of GEPAS) and SOM-Tree (19).

Supervised analysis includes

(i) *Gene selection*. Analysis of genes differentially expressed between two or more classes, related to a continuous experimental factor (e.g. the concentration of a metabolite) or to survival is performed by the module Pomelo (6). Different methods for multiple testing adjustment are included (20–22).

(ii) *Predictors*. The module Tnasas (for 'This is not a substitute for a statistician') implements a simple, although effective, way of building class predictors from microarray data. The error rate is computed taking into account the effect of gene selection and is not biased downwards by the 'selection bias' problem so common in many microarray studies [e.g. (23,24)].

For the *analysis of CGH-arrays*, given the growing interest in microarray-based CGH (array CGH) (25), we have expanded the capabilities of the InSilicoCGH tool, which allows the mapping of the results of microarray hybridizations onto chromosome coordinates. The InSilicoCGH module has been designed for the simultaneous analysis of genomic and mRNA hybridizations on the same expression array. It can also deal with BAC-arrays. We have added a new option: the zoom. This magnifies the view of the desired chromosomal location in order to facilitate detection of the precise position of chromosomal gains and losses; in general, it allows hybridization values at gene level to be viewed in more detail. Figure 2 is a screenshot of the zoom tool.

Functional analysis of experiments

Functional annotation of microarray experiments is an important aspect of analysis that very few packages incorporate. Several modules for functional annotation of microarray

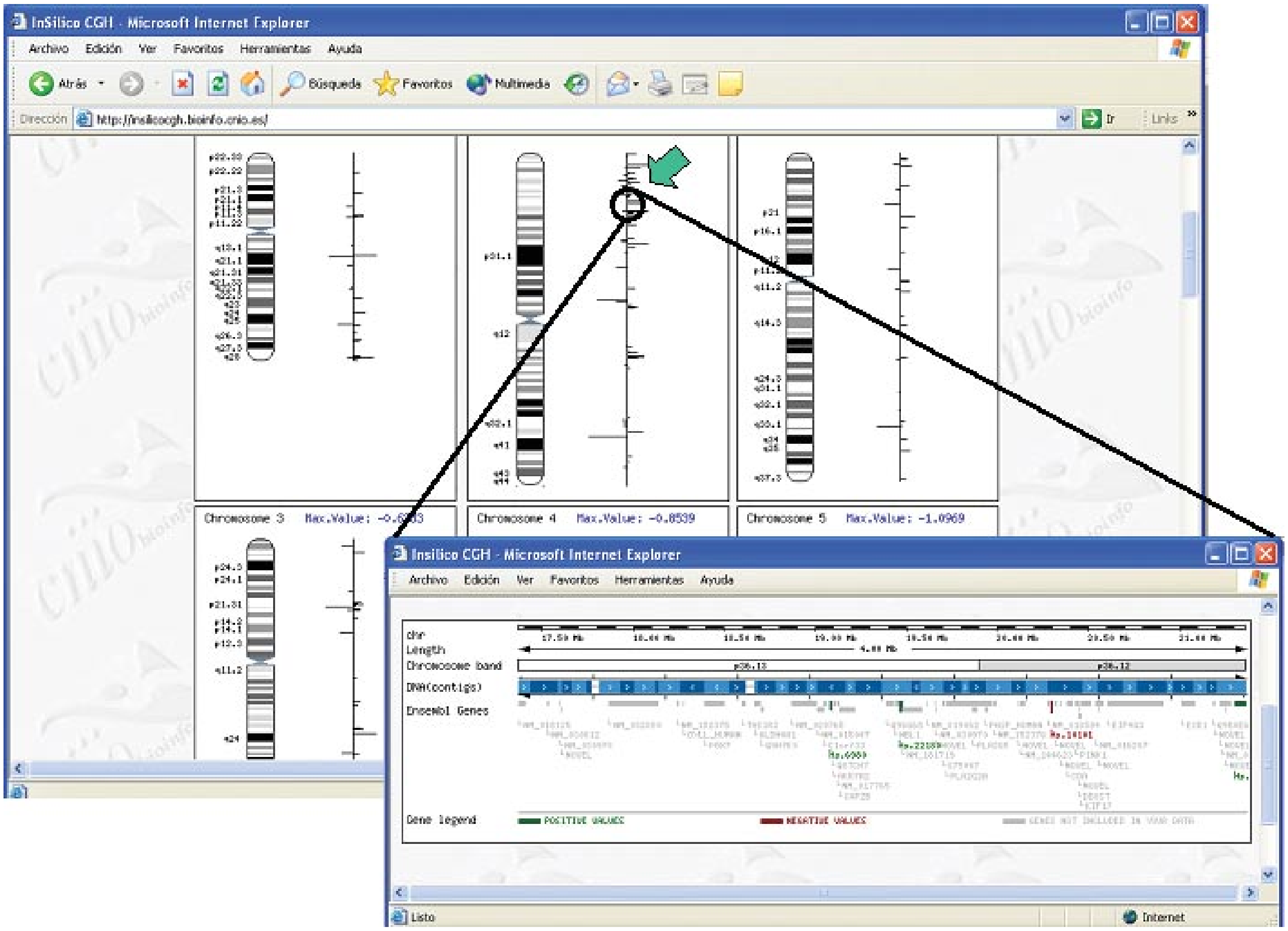
InSilico CGH - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás
Búsqueda
Favoritos
Multimedia
Dirección http://insilicocgh.bioinfo.cnio.es/
Ir
Links
Chromosome 3 Max.Value: -0.6
Chromosome 4 Max.Value: -0.8539
Chromosome 5 Max.Value: -1.0969
Insilico CGH - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Chr
Length
Chromosome band
DNA(contigs)
Ensembl Genes
Gene legend
POSITIVE VALUES
NEGATIVE VALUES
GENES NOT INCLUDED IN YOUR DATA
Listo
Internet

Figure 2. The zoom tool of InSilicoCGH in action. Clicking on the desired chromosomal region produces a pop-up window with a zoom facility. The user can freely move around the point chosen and can easily visualize in detail the hybridization values. Borders of deleted or amplified regions can be precisely defined in this way.

experiments are available. These programs, along with similar ones, are discussed in an accompanying paper by our group (7)

(i) *FatiGO* (8) allows significant asymmetrical distributions of GO terms between groups of genes to be found.
(ii) *FatiWise* (6) does the same with InterPro motifs (26), KEGG (27) pathways and SwissProt keywords, when available.
(iii) *TransFAT* performs the same operations for putative transcription factor binding sites in the promoter regions of genes as predicted by the program Match (28), from the Transfac® database (29).
(iv) *TMT*, the Tissues Mining Tool, is a web application to extract significant information related to the differential expression of two sets of genes in tissues.
(v) *FatiScan* allows the detection of modest but coordinate changes in gene expression values by applying the FatiGO algorithm to lists of genes ordered according to their differences in expression.

All these tools, in addition of being connected to GEPAS (because of its obvious usefulness for the analysis of microarray data), are grouped as an independent resource called Babelomics (7). Babelomics has, at its general purpose, the facilitation of functional annotation in any type of high-throughput experiments (proteomics, interactomics, massive sequencing, etc.).

In terms of its internal architecture, GEPAS is a collection of programs mainly written in C++, although some were written in other programming languages such as R [DNMAD (9)] or PERL [Preprocessor (12)]. These modules are interconnected by PERL wrappers.

A PIPELINE OF MICROARRAY DATA ANALYSIS TOOLS

The efficiency of a modular package such as GEPAS lies largely in its degree of integration of the different data analysis tools. Users can move through a complete pipeline of data analysis in a transparent way, without needing to perform any reformatting operation. In addition, a properly designed workflow can help to prevent possible wrong operations in microarray data analysis owing to misconceptions. Figure 1 illustrates the structure of the GEPAS pipeline. Raw data can be loaded and normalized. Several data transformation options are available through the Preprocessor tool. Depending on the particular problem addressed, data can be directed to any of the three main types of analysis: CGH-array, unsupervised clustering and supervised analysis (gene selection or predictors). A functional annotation is possible from the last two options. GEPAS has been designed in a way that prevents possible misuses of the methods implemented in the package.

TRAINING PROGRAMME AND GEPAS

In addition to the tools, a collection of on-line tutorials that can be used to learn the use of the tools or as a part of a course is available on the GEPAS web page. The structure of the tutorials includes some theory, a guided example and several examples based on publicly available datasets. There are tutorials for (i) normalization using DNMAD, (ii) data preprocessing using the Preprocessor tool, (iii) data clustering using the different algorithms available (UPGMA, SOM, SOTA), (iv) selection of differentially expressed genes using the Pomelo tool and (v) functional annotation using FatiGO.

The tutorials are currently used on different courses, such as a masters in bioinformatics (Spain) and the international FCUL-IGC Post-Graduate Programme in Bioinformatics (http://bioinformatics.fc.ul.pt/).

CONSOLIDATION OF GEPAS AS A WIDELY USED PACKAGE

Our records indicate that, since March 2004, GEPAS has been used to analyse >76 000 experiments, with a daily average of almost 300 uses (statistics can be checked at http://bioinfo.cnio.es/docus/webalizer/ on the different pages for GEPAS, and the particular pages for Pomelo, Tnasas, DNMAD and FatiGO, which are independently monitored). Compared with last year's records (35 000 experiments per year with a daily average of 130) (6), there has been a clear increase in the use of the tool. The distribution of users has also changed. Whereas one year ago it was used more by Spanish researchers (25%), followed by US (.edu and .net domains) (15%), French (10%), UK (5%) and other users (Japanese, German, Dutch, etc.) (6) the profile of users during this last year has changed to 23% US (.edu and .net), 9% French, 6% Spanish, 5% UK and others. These figures suggest that GEPAS seems to be becoming more popular among US-based researchers. Obviously the usage in all countries has increased, since the remainder of the percentages appear to maintain the same level while the absolute number of uses has increased 2-fold.

CONCLUSIONS

Despite the availability of many programs and packages for microarray data analysis, there are still many aspects of the analysis with poor or incomplete coverage. There are a number of options for analysing DNA microarray data (see e.g. http://www.dnamicroarrays.info/software.html). Most of the software available for microarray data analysis focuses on unsupervised cluster methods, which, in many cases, are used for inadequate purposes (23). There are also different initiatives such as BASE (30), Bioconductor (31) and BRB tools (http://linus.nci.nih.gov/BRB-ArrayTools.html), but these are in some cases dependent on a particular computer operating system and usually require from the user previous training in statistics. GEPAS can be considered the most complete web-based resource that can be found nowadays.

Since the first release (5,6), GEPAS has avoided the temptation to become a list of as many methods as possible and evolved really to cope with new challenges that have emerged in the field of microarray data analysis. Much work has been invested in the implementation of a useful workflow. GEPAS provides the user with an integrated environment in which modules can be found for different types of analysis that respond to real analysis demands. Modules are connected in such a way as to avoid improper use of the tools.

From a technical point of view, GEPAS has been designed with the intention of taking full advantage of the properties of the web: connectivity, cross-platform compatibility and remote usage. The modular architecture allows the addition of new tools and facilitates the connectivity of GEPAS from and to other web-based tools.

With >76 000 experiments analysed during the last year and a daily average of almost 300 uses, GEPAS can be considered a consolidated tool in the field of microarray data analysis.

ACKNOWLEDGEMENTS

J.M.V. is supported by the Formacion del Personal Investigador fellowship program from the Ministerio de Educacion y Ciencia. L.C. is supported by a fellowship from the Fondo de Investigacion Sanitaria (grant PI020919). P.M. is supported by a grant from Genoma España and Canada Genome. This work is partly supported by grants from Fundación Ramón Areces, Fundació La Caixa, Fundación BBVA and RTICCC from the FIS. The Functional Genomics node of the INB is funded by Fundación Genoma España. IBM awarded us a SUR grant. Funding to pay the Open Access publication charges for this article was provided by Fundacion Genoma España.

Conflict of interest statement. None declared.

REFERENCES

1. van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
2. Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al*. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med*., **8**, 816–824.
3. Albertson,D.G. and Pinkel,D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet*., **12**, 145R–152R.
4. The Tumor Analysis Best Practices Group (2004) Expression profiling—best practices for data generation and interpretation in clinical trials. *Nature Rev. Genet*., **5**, 229–237.
5. Herrero,J., Al-Shahrour,F., Díaz-Uriarte,R., Mateos,A., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS, a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res*., **31**, 3461–3467.
6. Herrero,J., Vaquerizas,J.M., Al-Shahrour,F., Conde,L., Mateos,Á., Santoyo,J., Díaz-Uriarte,R. and Dopazo,J. (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res*., **32**, W485–W491.
7. Al-Shahrour,F., Minguez,P., Vaquerizas,J.M., Conde,L. and Dopazo,J. (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucl. Acids Res*., **33**, W460–W464.
8. Al-Shahrour,F., Díaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms to groups of genes. *Bioinformatics*, **20**, 578–580.
9. Vaquerizas,J.M., Dopazo,J. and Díaz-Uriarte,R. (2004) DNMAD: web-based diagnosis and normalization for MicroArray data. *Bioinformatics*, **20**, 3656–3658.
10. Smyth,G.K., Yang,Y.H. and Speed,T.P. (2003) Statistical issues in microarray data analysis. In Brownstein,M.J., Khodursky,A.B. and Totowa,N.J. (eds), *Functional Genomics: Methods and Protocols, Methods in Molecular Biology*. Humana Press, Vol. 224, pp. 111–136.
11. Dudoit,S. and Yang,H.Y. (2003) Documentation of the Bioconductor's marrayPlots package.
12. Herrero,J., Díaz-Uriarte,R. and Dopazo,J. (2003) Gene expression data preprocessing. *Bioinformatics*, **19**, 655–656.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nature Genet*., **25**, 25–29.
14. Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy*. Freeman, San Francisco.
15. Dopazo,J. and Carazo,J.M. (1997) Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol*., **44**, 226–233.
16. Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.
17. Kohonen,T. (1997) *Self-Organizing Maps*. Springer-Verlag, Berlin.
18. Hartigan,J.A. and Wong,M.A. (1979) A *k*-means clustering algorithm. *Appl. Stat*., **28**, 100–108.
19. Herrero,J. and Dopazo,J. (2002) Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. *J. Proteome Res*., **1**, 467–470.
20. Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*. John Wiley & Sons, New York.
21. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
22. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*., **29**, 1165–1188.
23. Simon,R., Radmacher,M.D., Dobbin,K. and McShane,L.M. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst*., **95**, 14–18.
24. Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.
25. Snijders,A.M., Nowak,N., Segraves,R., Blackwood,S., Brown,N., Conroy,J., Hamilton,G., Hindle,A.K., Huey,B., Kimura,K. *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet*., **29**, 263–264.
26. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
27. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
28. Kel,A.E., Gößling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids. Res*., **31**, 3576–3579.
29. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüß,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*., **28**, 316–319.
30. Saal,L.H., Troein,C., Vallon-Christersson,J., Gruvberger,S., Borg,Å. and Peterson,C. (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol*., **3**, software0003.1–software0003.6.
31. Gentleman,R.C., Carey,V.J., Douglas,M.B., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.