BMC Bioinformatics

Software

ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization

Enrico Glaab, Jonathan M Garibaldi and Natalio Krasnogor*

Address: School of Computer Science, Nottingham University, Jubilee Campus, Wollaton Road, Nottingham, UK

Email: Enrico Glaab - enrico.glaab@cs.nott.ac.uk; Jonathan M Garibaldi - jmg@cs.nott.ac.uk; Natalio Krasnogor* - nxk@cs.nott.ac.uk

* Corresponding author

Published: 28 October 2009

BMC Bioinformatics 2009, **10**:358 doi:10.1186/1471-2105-10-358

Received: 8 May 2009

Accepted: 28 October 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/358

Abstract

Background: Statistical analysis of DNA microarray data provides a valuable diagnostic tool for the investigation of genetic components of diseases. To take advantage of the multitude of available data sets and analysis methods, it is desirable to combine both different algorithms and data from different studies. Applying ensemble learning, consensus clustering and cross-study normalization methods for this purpose in an almost fully automated process and linking different analysis modules together under a single interface would simplify many microarray analysis tasks.

Results: We present ArrayMining.net, a web-application for microarray analysis that provides easy access to a wide choice of feature selection, clustering, prediction, gene set analysis and cross-study normalization methods. In contrast to other microarray-related web-tools, multiple algorithms and data sets for an analysis task can be combined using ensemble feature selection, ensemble prediction, consensus clustering and cross-platform data integration. By interlinking different analysis tools in a modular fashion, new exploratory routes become available, e.g. ensemble sample classification using features obtained from a gene set analysis and data from multiple studies. The analysis is further simplified by automatic parameter selection mechanisms and linkage to web tools and databases for functional annotation and literature mining.

Conclusion: ArrayMining.net is a free web-application for microarray analysis combining a broad choice of algorithms based on ensemble and consensus methods, using automatic parameter selection and integration with annotation databases.

Background

DNA microarray experiments provide a powerful means to improve our understanding of diseases with a genetic basis or contribution. Commercial microarray chips for highly accurate diagnosis of several cancers are already available on the market [1,2] and pharmaceutical companies are using DNA-chip technology to identify new drug targets.

The fast accumulation of gene expression data in public online databases and the great variety of available analysis methods, however, also pose new challenges. Integrating data from different sources, choosing appropriate normalization, analysis and cross-validation methods and selecting suitable parameters requires substantial time and effort. Since different algorithms have different strengths and similar data from independent studies is often availa-

ble, it is desirable to combine multiple methods and/or data sets to obtain more robust and accurate results. This creates ample opportunities for ensemble methods and cross-study normalization techniques.

Although statistical programming frameworks like R [3] and Matlab [4] allow users to develop and apply complex scripts for expression data analysis, they are difficult to use for non-experts and there is a high risk of deviating from standard guidelines. To obviate the need for specialized programming skills and manual software installations, several web-based tools for gene expression analysis have been presented in recent years. Currently available integrative online analysis services include *GEPAS* [5], *Expression Profiler* [6], *ASTERIAS* [7], *EzArray* [8], *CARMAweb* [9], *MAGMA* [10], *ArrayPipe* [11], *RACE* [12], *WebArray* [13] and *MIDAW* [14]. These web-based systems provide methods for a multitude of data pre-processing and analysis purposes ranging from image analysis, missing value imputation, single-study normalization, gene filtering and gene name conversion to higher-level analysis methods for clustering, gene selection and gene annotation, prediction, data visualization and gene set enrichment analysis, among others.

Additionally, numerous web-applications have been developed and optimized for single, specific analysis tasks, e.g. biclustering of genes and samples [15], co-clustering of genes with similar functional annotation [16], framework inference for regulatory networks [17] and cross-species clustering [18]. Although various tools provide a choice and comparison between different algorithms for one analysis task, to the best of our knowledge, currently no integrative analysis software enables the user to easily combine multiple methods together using ensemble learning and consensus clustering techniques. Previous studies have shown that microarray analysis can profit from ensemble feature selection, ensemble prediction and consensus clustering methods both in terms of robustness and accuracy [19-22], suggesting that there is significant potential still to be exploited with these approaches.

Similarly, it would be desirable not only to combine different algorithms but also different data sets for a common organism and phenotype. Although currently available cross-study normalization methods are based on simplified assumptions and limited in applicability and accuracy, various successful applications [23,24] have shown that the benefits of an increased sample size can outweigh the loss of information due to the normalization process.

For these reasons, we have developed a new web-application that provides access to multiple algorithms for each of the most common tasks in statistical microarray analysis, namely gene selection, sample clustering, sample classification and gene set analysis, based on a single, easy-to-use interface. In contrast to other web-tools, in which the results of individual methods are made available, here, ensemble feature selection, ensemble prediction and consensus clustering approaches are provided. Likewise, instead of using only data from a single study, different cross-study normalization methods are made available to integrate similar data from different studies and compare the results based on density and quantile-quantile plots.

Apart from these combinations of data sets and methods within an analysis module, different modules have been interlinked, enabling for example the integration of gene set analysis with classification or cross-study analysis with gene selection or clustering. Other new features include access to an in-house developed rule-based evolutionary classification algorithm, automatic parameter selection mechanisms on all modules, the availability of specific cancer-related gene sets for enrichment analysis in addition to gene sets from KEGG and GO, and a 3D-VRML-visualization of clustering results using the authors' new R software package "vrmlgen" [25].

Since the above methods and features are not available on other microarray-related web-tools, and similarly, other tool sets include methods distinct from our system, we see our service as a complement rather than an alternative to existing services.

In the following we provide an overview of the workflow and describe all features in detail.

Implementation and workflow

The ArrayMining.net tool set consists of five main modules for microarray analysis: *Cross-Study Normalization*, *Gene selection*, *Class Discovery*, *Class Assignment* and *Gene Set Analysis*. Each of these modules features multiple analysis methods accessible through a unified web-interface. The user can upload his own data in tab-delimited text-file format or as zip-compressed Affymetrix CEL-files which will be automatically extracted, normalized and summarized using the Robust Microarray Analysis (RMA) method [26]. Alternatively, various example data sets have been made available directly on the webpage and access to the GEO database [27], the largest public microarray data base, is provided on the class discovery module. After submitting an analysis task, an output webpage containing the downloadable results as plots, tables, VRML-files etc. is generated. Depending on the chosen module and algorithm the data can be forwarded to further analysis modules and will be interlinked with annotation data from external web-tools and data bases.

ArrayMining.net is based on software written in the programming languages R [3] and C++ and a PHP-interface combining all implementations together on an Apache web server. The system uses in-house algorithms and implementations as well as standard packages from the Bioconductor project [28]. All modules are easily extensible and the authors encourage users to contribute with feature requests or their own analysis scripts. A regularly updated illustration of the workflow and features on our server is available online (see Availability section). Below we describe each of the modules in detail.

Cross-study normalization module

Current microarray studies often only contain a small number of samples, resulting in limited robustness and reliability of statistical analyses. To alleviate this problem five cross-study normalization methods have been made available on ArrayMining.net to combine samples from two different studies: An approach based on linked gene- and sample-clustering (XPN [23]), an empirical Bayes method (EB [29]), a median rank score based method (MRANK [24]), an outlier-removing discretization technique (NorDi [30]) and a quantile discretization procedure (QDISC [24]). While the first three methods provide continuous-valued outputs, the last two are based on discretization to filter out noise, exploiting the fact that for higher-level analysis often only a general categorization of gene expression levels in different conditions is required (e.g. "unaltered", "up"- or "down"-regulated), but potentially resulting in a higher loss of biological information. The input data sets can originate from different microarray platforms, but the associated gene sets need to overlap significantly and the samples should be derived from the same tissue type under comparable biological conditions. As a result, the combined data can be downloaded or forwarded to other modules, and density and quantile-quantile plots are generated to compare different algorithms.

Gene selection module

Identifying differentially expressed genes is a common starting point for the biological interpretation of microarray data. Our gene selection module enables the comparison and combination of a diverse choice of methods for this purpose: The Empirical Bayes t-statistic (*eBayes*) [31,32], the Significance Analysis in Microarrays method (*SAM*) [33], a correlation-based combinatorial feature selection approach (*CFS*) [34], a ranking method based on Random Forest classification (*RF-MDA*) [35] and a Partial-Least-Squares based filter (*PLS-CV*) [36] using the weight vectors defining the first latent components in cross-validated PLS-models. To exploit the synergies of different algorithms, we have implemented a method to compute aggregated gene ranks from the sum of ranks of individual methods (*ENSEMBLE*). The resulting outcome reports provide a ranked list of genes, in which known gene identifiers become clickable navigation items, referring the user to related entries in functional annotation databases and literature search engines. Additionally, box plots and heat maps (see Fig. 1 and 2) visualize the expression values of top-ranked genes across different sample-groups. If the supplied data uses common gene identifiers (Entrez Gene ID, NCBI GI accession, Unigene ID, RefSeq Genomic ID, etc.), the list of selected genes can be forwarded to external analysis tools, e.g. the functional annotation clustering service of the DAVID web database [37].

Class discovery module

Clustering methods allow experimenters to identify natural groupings among microarray samples based on their expression patterns across the genes. To account for the great variety of existing scoring and search space exploration methods, our class discovery module includes both partition-based and hierarchical clustering algorithms, an evaluation based on multiple validity indices and a consensus clustering method. Currently, the partition-based clustering methods available are *k-Means*, *PAM* [38], *SOM* [39] and *SOTA* [40], and the hierarchical clustering methods are *Average Linkage Agglomerative Clustering*, *Divisive Analysis Clustering* and a combination between the

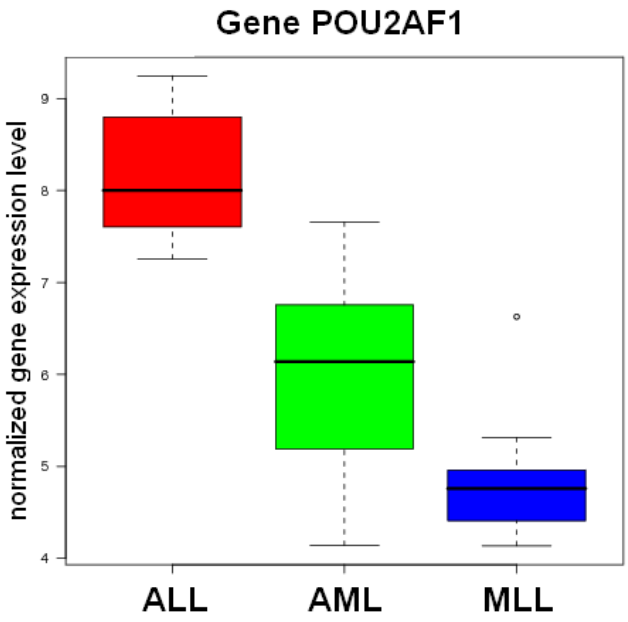

Figure 1
Boxplots. Example of a boxplot illustrating the spread of a gene's expression values across three classes of leukemia samples: Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML) and Mixed Lineage Leukemia (MLL) (data set by Armstrong et al. [55], [see Additional file 1 for further details on this and other differentially expressed genes]).

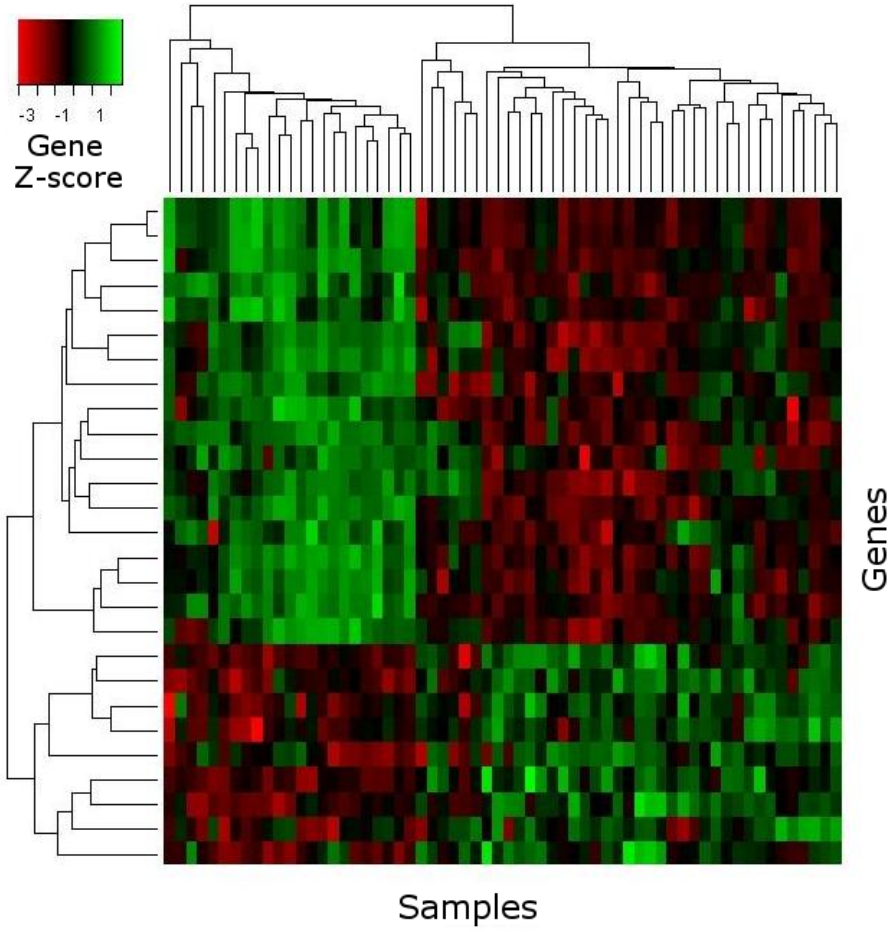

Figure 2
Heat map. Example of a heat map visualizing the expression values of selected genes (rows) across samples (columns).

agglomerative and divisive approach, *Hybrid Hierarchical Clustering* [41]. To combine the information content from multiple clusterings into a single representative solution, we have implemented our own consensus clustering approach, which maximizes a score for the agreement between sample-pair assignments of the consensus clustering and all input clusterings using a fast simulated annealing approach [42]. This method was developed based on experiences from earlier work on protein structure similarity clustering [43], which showed that consensus methods can increase the robustness and reliability of statistical analyses on biological data sets. For each algorithm the number of clusters is estimated automatically by means of multiple validity indices and a refined estimate can be obtained by combining all pairs of algorithms and validity indices. Optionally, different types of data standardization and two gene filtering methods can be applied prior to the analysis. This includes a classical variance-based filter as well as a recently published parameter-free method, which can distinguish between uncorrelated, uninformative genes and regulators with high correlation to other genes [44]. An alternative filtering approach is to first use the gene set analysis module (see below) to extract "meta-genes" representing biological pathways and forward this data to the class discovery module. As a result for each analysis, the user will obtain a tabular summary of the calculated validity indices and clustering results and various graphical outputs including a silhouette-plot [45], a 2D principal components plot and a 3D VRML-visualization (see Fig. 3), including density estimation contour surfaces based on an Independent Component Analysis of the data and our software-package "vrmlgen" for the R programming language [3] (freely available at Ref. [25] and the official R package archive, CRAN).

Class Assignment module

An important goal behind microarray analysis is to improve the diagnosis of diseases with genetic components by predicting the disease type based on labeled

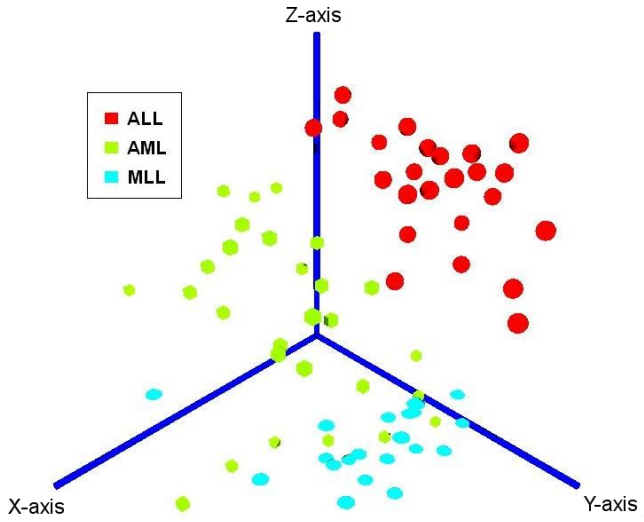

Figure 3
Independent Component Analysis. Example of a VRML-visualization for an Independent Component Analysis (data set by Armstrong et al. [55]).

training data. The third module on our web-server is therefore dedicated to supervised learning methods, including various common methods for microarray sample classification (*SVM* [46], *RF* [35], *PAM* [38] and *kNN*). We also provide access to an in-house developed rule-based machine learning approach, BioHEL [47], which learns structured classification rule sets, known as "decision lists", by applying a genetic algorithm within an iterative rule learning (IRL) framework. BioHEL has previously been shown to achieve high prediction accuracies on complex biological data sets [48], while being based on easily interpretable "if-then-else"-rules. The prediction methods can be evaluated and compared based on the widely accepted external two-level cross-validation methodology [49], using automatic parameter optimization within a nested cross-validation. As with the other modules, an ensemble of algorithms is available both for selection and prediction to obtain more robust results. Moreover, since prediction models derived from training data of a single study can typically not be applied to samples from other platforms and laboratories, the combination of cross-study normalization (see above) with prediction provides a means to obtain more general models based on a larger sample size.

The results for an analysis contain various performance measures for evaluation and Z-scores for the genes that were most frequently selected across different cross-validation cycles. To obtain more insights on these genes, similar analysis plots and annotation tools are available as for the gene selection module.

Gene Set Analysis module

Two common problems in microarray analysis are high noise levels for single genes and a high number of redundant or uninformative genes. Using gene set analysis (GSA) to aggregate functionally related genes into gene sets and summarizing their expression values to a robust "meta"-gene expression vector is a promising approach to overcome some of these limitations [50]. Moreover, differentially expressed gene sets can provide insights on the differences between the biological conditions of the samples on the level of molecular modules and biochemical pathways. Our gene set analysis module provides access to three functional annotation sources to identify functionally related genes in a data set and extract corresponding gene sets: The Gene Ontology Database [51], the KEGG data base [52], and a collection of 37 cancer-related gene sets from the van Andel Institute in Michigan [53]. Alternatively, users can specify their own gene sets using the gene identifiers for the data set of interest. Since common non-parametric GSA methods are often computationally expensive or provide only rough estimates of a gene's significance score, we compute p-values based on the parametric PAGE-method [53], requiring a minimum gene set size of approx. 10 genes. To adjust for multiple testing, the Benjamini-Hochberg method [54] is used.

Summarized meta-gene expression vectors for a gene set are obtained by transforming the expression levels using Principal Component Analysis (PC-GSA) or Multidimensional Scaling (MDS-GSA).

The outcome is presented as a ranked list of gene sets and additionally contains box plots and heat maps similar to those on the gene selection module. Meta-gene expression values derived from the gene sets can be downloaded or forwarded to other analysis modules, e.g. to be used as predictors in sample classification.

Results

Providing example results for all modules and algorithms on ArrayMining.net would exceed the scope of this paper. However, we have included some results obtained with the well-known three-class leukemia data set by Armstrong et al. [55] [see Additional file 1]. This includes the list of the 30 top-ranked genes using the ENSEMBLE gene selection method, as well as a heat map, box plots, a ranked list of cancer gene sets from the gene set analysis module, sample classification results, a VRML-file visualizing the results from an Independent Component Analysis computed on the class discovery module (shown in Fig. 3) and a discussion of all results. In summary, nearly all of the selected genes with available annotation data are known or likely to be differentially expressed in different leukemia types. An example box plot for a top-ranked gene - the transcriptional regulator POU2AF1, which has

been implicated in lymphoma and leukemia development [56] - is shown in Fig. 1. We also show results for a combination of two modules, obtaining an average sample classification accuracy of 87% (external 10-fold cross-validation) on the Class Assignment module when using meta-genes derived from the gene set analysis module as robust input features. On the class discovery module, the clustering and validity methods were able to perfectly distinguish two leukemia subtypes in the data, Mixed Lineage Leukemia (MLL) and Acute Lymphoblastic Leukemia (ALL), while the samples for a third subtype, Acute Myeloid Leukemia (AML), could only partly be separated from the other two groups. However, when visualizing the pre- filtered data using an Independent Component Analysis, the three leukemia groups were well separated in 3D-space with only a small overlap between the MLL and the AML group, although all results were generated in a fully automatic process [see Additional file 2 for a VRML-visualization of the data].

Since these examples cover only some of the available features, various well-known microarray cancer data sets are available on the different analysis modules to enable the user to more fully explore the capabilities of ArrayMining.net without needing to upload new data.

Conclusion

We have developed a new web-application that provides a simple and fast way to analyze arbitrary DNA-chip data and other high-dimensional data sets. Ensemble, consensus and cross-study normalization methods help to increase the robustness and accuracy of the outcomes, and automatic parameter selection mechanisms and a direct linkage to functional annotation data bases (ENSEMBL, DAVID, etc.) relieve the user of time-consuming routine tasks. For each of the major statistical analysis tasks - feature selection, clustering, prediction and gene set analysis - several analysis methods are available and can be compared, combined or interlinked in many ways. In contrast to other software products for microarray analysis, the user is neither tied to a particular methodology nor needs to understand in detail the inner working of the algorithms. New researchers in the field can use the web-tool without the risk of deviating from standard validation guidelines.

For the next version of the server, we are planning to add a new module for co-expression network analysis and the possibility to integrate additional clinical or biological data.

Availability and requirements

The web-application, video tutorials and an illustration of the features and workflow are freely accessible at http://www.arraymining.net.

Authors' contributions

EG participated in the conceptual design of the web-application, implemented the algorithms and PHP-interface and drafted the manuscript. NK took part in the conceptual design of the web-application and helped to draft the manuscript. JMG helped to draft the manuscript. JMG and NK wrote the grant application upon which this project was built. All authors read, made corrections and approved the final manuscript.

Additional material

Additional file 1

Results for an example analysis with ArrayMining.net. *The document provided contains the results and a discussion for an example analysis of microarray data with ArrayMining.net.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-358-S1.pdf]

Additional file 2

Example VRML-visualization of an Independent Component Analysis. *This file contains an example VRML-visualization of an Independent Component Analysis for the microarray data set by Armstrong et al. [55] (a VRML browser plugin or viewing software is required to open the file).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-358-S2.wrl]

Acknowledgements

We acknowledge support by the Marie-Curie Early-Stage-Training programme (grant MEST-CT-2004-007597), by the UK Engineering and Physical Sciences Research Council (EP/E017215/1) and the Biotechnology and Biological Sciences Research Council (BB/F01855X/1).

References

1. Wittner B, Sgroi D, Ryan P, Bruinsma T, Glas A, Male A, Dahiya S, Habin K, Bernards R, Haber D, *et al.*: **Analysis of the MammaPrint breast cancer assay in a predominantly postmenopausal cohort.** *Clin Cancer Res* 2008, **14(10):**2988.
2. Horlings H, Warmoes M, Kerst J, Helgason H, De Jong D, Van't Veer L: **Successful classification of metastatic carcinoma of known primary using the CUPPRINT.** *J Clin Oncol* 2006, **24:**20028.
3. Ihaka R, Gentleman R: **R: A Language for Data Analysis and Graphics.** *J Comput Graph Stat* 1996, **5(3):**299-314.
4. The MathWorks Inc: *Matlab. Natick, MA* 1998.
5. Tarraga J, Medina I, Carbonell J, Huerta-Cepas J, Minguez P, Alloza E, Al-Shahrour F, Vegas-Azcarate S, Goetz S, Escobar P, *et al.*: **GEPAS, a web-based tool for microarray data analysis and interpretation.** *Nucleic Acids Res* 2008, **31(13):**3461-3467.
6. Kapushesky M, Kemmeren P, Culhane A, Durinck S, Ihmels J, Korner C, Kull M, Torrente A, Sarkans U, Vilo J, *et al.*: **Expression Profiler: next generation-an online platform for analysis of microarray data.** *Nucleic Acids Res* 2004:W465.
7. Diaz-Uriarte R, Alibes A, Morrissey E, *et al.*: **Asterias: integrated analysis of expression and aCGH data using an open-source, web-based, parallelized software suite.** *Nucleic Acids Res* 2007:W75.
8. Zhu Y, Zhu Y, Xu W: **EzArray: A web-based highly automated Affymetrix expression array data management and analysis system.** *BMC Bioinformatics* 2008, **9:**46.
9. Rainer J, Sanchez-Cabo F, Stocker G, Sturn A, Trajanoski Z: **CARMAweb: comprehensive R-and bioconductor-based web**

service for microarray data analysis. *Nucleic Acids Res* 2006:W498.
10. Rehrauer H, Zoller S, Schlapbach R: **MAGMA: analysis of two-channel microarrays made easy.** *Nucleic Acids Research* 2007:W86.
11. Hokamp K, Roche F, Acab M, Rousseau M, Kuo B, Goode D, Aeschliman D, Bryan J, Babiuk L, Hancock R, *et al.*: **ArrayPipe: a flexible processing pipeline for microarray data.** *Nucleic Acids Res* 2004:W457.
12. Psarros M, Heber S, Sick M, Thoppae G, Harshman K, Sick B: **RACE: remote analysis computation for gene expression data.** *Nucleic Acids Res* 2005:W638.
13. Xia X, McClelland M, Wang Y: **WebArray: an online platform for microarray data analysis.** *BMC Bioinformatics* 2005, **6:**306.
14. Romualdi C, Vitulo N, Favero M, Lanfranchi G: **MIDAW: a web tool for statistical analysis of microarray data.** *Nucleic Acids Res* 2005:W644.
15. Wu C, Fu Y, Murali T, Kasif S: **Gene expression module discovery using Gibbs sampling.** *Genome Inform* 2004, **15:**239-248.
16. Lee J, Sinkovits R, Mock D, Rab E, Cai J, Yang P, Saunders B, Hsueh R, Choi S, Subramaniam S, *et al.*: **Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation.** *BMC Bioinformatics* 2006, **7:**237.
17. Aburatani S, Goto K, Saito S, Toh H, Horimoto K: **ASIAN: a web server for inferring a regulatory network framework from gene expression profiles.** *Nucleic Acids Res* 2005:W659.
18. Lu Y, He X, Zhong S: **Cross-species microarray analysis with the OSCAR system suggests an INSR -> Pax6 -> NQO1 neuro-protective pathway in aging and Alzheimer's disease.** *Nucleic Acids Res* 2007:W105.
19. Saeys Y, Abeel T, Peer Y: **Robust Feature Selection Using Ensemble Feature Selection Techniques.** In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases-Part II* Springer-Verlag Berlin, Heidelberg; 2008:313-325.
20. Tan A, Gilbert D: **Ensemble machine learning on gene expression data for cancer classification.** *Appl Bioinformatics* 2003, **2(3 Suppl):**S75-S83.
21. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data.** *Machine Learning* 2003, **52:**91-118.
22. Swift S, Tucker A, Vinciotti V, Martin N, Orengo C, Liu X, Kellam P: **Consensus clustering and functional interpretation of gene-expression data.** *Genome Biol* 2004, **5(11):**R94.
23. Shabalin A, Tjelmeland H, Fan C, Perou C, Nobel A: **Merging two gene-expression studies via cross-platform normalization.** *Bioinformatics* 2008, **24(9):**1154.
24. Warnat P, Eils R, Brors B: **Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.** *BMC Bioinformatics* 2005, **6:**265.
25. **VRMLGen R software package** [http://bree.cs.nott.ac.uk/vrmlgen]
26. Irizarry R, Bolstad B, Collin F, Cope L, Hobbs B, Speed T: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31(4):**e15.
27. Edgar R, Domrachev M, Lash A: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30:**207-210.
28. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10):**R80.
29. Walker W, Liao I, Gilbert D, Wong B, Pollard K, McCulloch C, Lit L, Sharp F: **Empirical Bayes accomodation of batch-effects in microarray data using identical replicate reference samples: application to RNA expression profiling of blood from Duchenne muscular dystrophy patients.** *BMC Genomics* 2008, **9:**494.
30. Martinez R, Pasquier C, Pasquier N: **GenMiner: Mining Informative Association Rules from Genomic Data.** *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine* 2007:15-22.
31. Lonnstedt I, Speed T: **Replicated microarray data.** *Stat Sin* 2002, **12:**31-46.
32. Smyth G: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**3.
33. Tusher V, Tibshirani R, Chu G, *et al.*: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98(9):**5116-5121.
34. Hall MA: **Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning.** *Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA* 2000:359-366.
35. Breiman L: **Random Forests.** *Machine Learning* 2001, **45:**5-32.
36. Boulesteix A, Strimmer K: **Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.** *Brief Bioinform* 2007, **8:**32-44.
37. Dennis G, Sherman B, Hosack D, Yang J, Gao W, Lane H, Lempicki R: **DAVID: database for annotation, visualization, and integrated discovery.** *Genome Biol* 2003, **4(9):**R60.
38. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *Proc Natl Acad Sci USA* 2002, **99(10):**6567-6572.
39. Kohonen T: *Self-Organizing Maps* Berlin: Springer Verlag; 2001.
40. Herrero J, Valencia A, Dopazo J: **A hierarchical unsupervised growing neural network for clustering gene expression patterns.** *Bioinformatics* 2001, **17(2):**126-136.
41. Chipman H, Tibshirani R: **Hybrid hierarchical clustering with applications to microarray data.** *Biostatistics* 2006, **7(2):**286-301.
42. Szu H: **Fast simulated annealing.** *AIP Conference Proceedings* 1986, **151:**420.
43. Barthel D, Hirst J, Blazewicz J, Burke E, Krasnogor N: **ProCKSI: A decision support system for protein (structure) comparison, knowledge, similarity and information.** *BMC Bioinformatics* 2007, **8:**416.
44. Tritchler D, Parkhomenko E, Beyene J: **Filtering genes for cluster and network analysis.** *BMC Bioinformatics* 2009, **10:**193.
45. Rousseeuw P: **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.** *J Comput Appl Mat* 1987, **20:**53-65.
46. Chang CC, Lin CJ: *LIBSVM: a library for support vector machines* 2001 [http://www.csie.ntu.edu.tw/~cjlin/libsvm].
47. Bacardit J, Burke E, Krasnogor N: **Improving the scalability of rule-based evolutionary learning.** *Memetic Computing* 2009, **1:**55-67.
48. Bacardit J, Stout M, Hirst J, Krasnogor N: **Data Mining in Proteomics with Learning Classifier Systems.** In *Learning Classifier Systems in Data Mining* Edited by: Bull L, Bernado Mansilla E, Holmes J. Springer; 2008:17-46.
49. Wood I, Visscher P, Mengersen K: **Classification based upon gene expression data: bias and precision of error rates.** *Bioinformatics* 2007, **23(11):**1363.
50. Guo Z, *et al.*: **Towards precise classification of cancers based on robust gene functional expression profiles.** *BMC Bioinformatics* 2005, **6:**58.
51. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, *et al.*: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25:**25-29.
52. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27.
53. Kim S, Volsky D: **PAGE: parametric analysis of gene set enrichment.** *BMC Bioinformatics* 2005, **6:**144.
54. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Statist Soc Ser B (Methodological)* 1995, **57:**289-300.
55. Armstrong S, Staunton J, Silverman L, Pieters R, den Boer M, Minden M, Sallan S, Lander E, Golub T, Korsmeyer S: **MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.** *Nat Genet* 2001, **30:**41-47.
56. Galiéque Z, Quief S, Hildebrand M, Denis C, Lecocq G, Collynd'Hooghe M, Bastard C, Yuille M, Dyer M, Kerckaert J: **The B cell transcriptional coactivator BOB1/OBF1 gene fuses to the LAZ3/BCL6 gene by t(3;11)(q27;q23.1) chromosomal translocation in a B cell leukemia line (Karpas 231).** *Leukemia* 1996, **10(4):**579.