A web-server for integrative microarray and gene set analysis

The University of Nottingham
School of Computer Science, Jubilee Campus, Nottingham, NG8 1BB

Enrico Glaab, Jonathan M. Garibaldi, Natalio Krasnogor
{egg, jmg, nxk}@cs.nott.ac.uk

www.arraymining.net

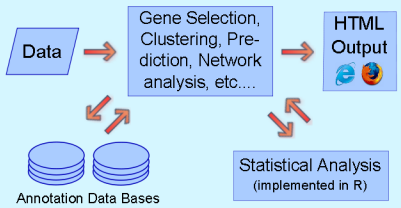
Nature Precedings : doi:10.1038/npre.2011.5552.1 : Posted 18 Jan 2011

1 Introduction

DNA microarray experiments provide a means to understand cancer and genetic diseases on a molecular level, improve diagnosis and identify new drug targets. However, choosing appropriate data processing methods and parameters is a difficult and time-consuming task, particularly for researchers without prior experience in this field.
We present **ArrayMining.net** [1], a free web-service for automatic microarray analysis to address these issues. ArrayMining.net covers several major areas in statistical microarray analysis - **Feature Selection, Clustering, Prediction, Gene Set and Network Analysis** - providing access to several algorithms for each of these tasks based on a single, easy-to-use interface.

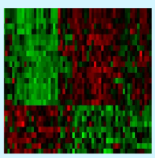
2 Workflow

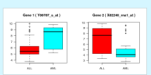
The ArrayMining server consists of **three PHP-modules** linked to the R statistical programming environment and to several online **annotation data bases** (e.g. ENSEMBL [2] and DAVID [3]). Users can upload their own data or use pre-normalized public data sets as input. **Automatic parameter selection** is carried out and all results are combined into a single HTML report.



3 Gene Selection Analysis

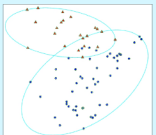
Finding genes which are functionally related to changes in biological conditions can help to improve the understanding of many diseases. Thus, our server provides a diverse choice of gene selection algorithms including **filters, wrappers and ensemble feature selection**.
The resulting web-reports list all selected genes, provide **heat maps** and **box plots** for their expression values and links to **functional annotation** data bases. Selected genes can also be passed over to an external web-tool for functional annotation clustering.

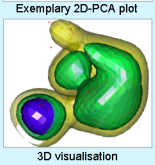

Exemplary heatmap


Box plots for selected genes

4 Clustering Analysis

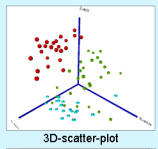
To analyze gene expression data with unknown sample labels our web-service features both **partition-based clustering** methods (e.g. SOM, PAM, k-Means) and various **hierarchical approaches** (e.g. DIANA, AGNES), as well as **consensus clustering**. For all algorithms the number of clusters is determined automatically based on multiple cluster validity indices. Various **visual aids** are available to interpret the results, e.g. a 2D-principe components plot, a Silhouette plot and 3D interactive visualisations created using our software package **VRMLGen** [4].

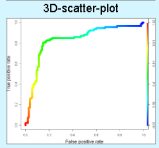

Exemplary 2D-PCA plot


3D visualisation

5 Prediction Analysis

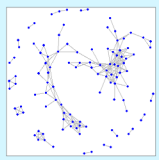
When experimenters wish to categorise new samples based on training data, our prediction module provides access to statistical learners like **SVMs**, **RF, kNN and PAM** in combination with various feature selection methods. All methods can be combined to an **ensemble** and cross-validated accuracies are obtained using the widely accepted **two-level external cross-validation** methodology [5]. Further evaluation statistics and analysis plots assist the user in comparing the performance for different combinations of selection and classification methods.


3D-scatter-plot


ROC-curve evaluation

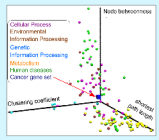
6 Pathway Analysis

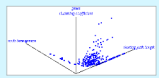
Different genes can have similar functions and their corresponding proteins can occur in the same molecular complex or cellular pathway. ArrayMining.net can map microarray genetic probes onto pathways and identify enriched and differentially regulated pathways (**Gene Set Analysis**), or discover co-expressed genes (**Gene Co-Expression Network Analysis**).


Co-expression sub-networks

7 Topological Analysis

Further biological insights can be obtained by combining microarray and protein interaction data using the analysis module **TopoGSA** [6]. Differentially expressed genes found with ArrayMining.net can be mapped onto **protein interaction networks** for different species (H. sapiens, A. thaliana, D. melanogaster, C. elegans) and topological properties, like the network centrality or the tendency of genes to cluster together in the network, can be compared to other gene sets representing cellular processes and pathways (Gene Ontology, KEGG, BioCarta, etc.).


Visual comparison of gene set topological properties


Topological properties within a single gene set

8 Conclusion

ArrayMining.net is a free web-service for microarray analysis providing:

- **integrative analysis methods (ensemble & consensus techniques)**
- **modular combinations of different analysis types**
- **new approaches for network topological analysis of genes**
- **automatic parameter selection**
- **integration with annotation data bases**

References

[1] E. Glaab, J.M. Garibaldi, N. Krasnogor, *ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization*, BMC Bioinformatics 2009, 10, 358
[2] T.J.P. Hubbard et al., *Ensembl 2009*, Nucleic Acids Research 2009, 37, D690-D697
[3] G. Dennis Jr. et al., DAVID: *Database for Annotation, Visualization, and Integrated Discovery*, Genome Biology 2003, 4(5), 2003-2004
[4] E. Glaab, J. M. Garibaldi, N. Krasnogor, *VRMLGen: An R-package for 3D Data Visualization on the Web*, Journal of Statistical Software 2010, 36(8), 1-18
[5] I.A. Wood, P. M. Visscher, K. L. Mengersen, *Classification based upon gene expression data: bias and precision of error rates*, Bioinformatics 2007, 23(11), 1363-1370
[6] E. Glaab, A. Baudot, N.Krasnogor, A. Valencia, *TopoGSA: network topological gene set analysis*, Bioinformatics 2010, 26(9), 1271-1272

We acknowledge support by Marie-Curie grant MEST-CT-2004-007597, by EPSRC grant EP/E017215/1 and BBSRC grant BB/F01855X/1

EPSRC Engineering and Physical Sciences Research Council
BBSRC bioscience for the future
MARIE CURIE ACTIONS