A web-server for integrative microarray and gene set analysis

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

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 and providing access to several algorithms for each of these tasks based on a single, easy-to-use interface.

Workflow
The ArrayMining server consists of three PHP-modules linked to the R statistical programming environment and to several online annotation databases (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.

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 automatic selection of clusters in microarray data using a number of clustering algorithms that are executed in parallel. For all algorithms the number of clusters is determined automatically based on the clusters’ validity indices. Various visual aids are available to interpret the results, e.g., a 2D-principle components plot, a Silhouette plot and 3D interactive visualisations created using our software package VRMLGen [4].

Clustering Analysis
To analyze gene expression data with unknown sample labels our web-service features both partition-based clustering methods (e.g. SCMI, 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-principle components plot, a Silhouette plot and 3D interactive visualisations created using our software package VRMLGen [4].

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.

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).

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.).

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