Genome Expression Pathway Analysis Tool - Analysis and Visualization of Microarray Gene Expression Data under Genomic, Proteomic and Metabolic Context

Markus Weniger[1], Jörg Schultz[1]
[1]Department of Bioinformatics, Biocenter, University of Würzburg, D-97074 Würzburg, Germany

Abstract
A typical analysis of microarray gene expression data consists of a number of steps, including normalization, filtering and annotation of the data, followed by various group prediction (unsupervised clustering) or classification (supervised clustering) methods. Although there exist a large number of data analysis tools available for these steps, most of them lack the ability to help with the interpretation of the results. For a deeper insight of the biological meaning, information like gene function, chromosomal location, affected pathways, protein interactions and literature references provide a useful start for further research.

For this reason we developed a web-based Tool, GEPAT, offering an integrated analysis of transcriptome data under genomic, proteomic and metabolic context. GEPAT imports various formats of oligonucleotide arrays, cDNA arrays and data tables in csv format. Upload of multiple files at once is possible. Data is stored password-protected in private user space on our server, allowing access to data from any computer around the world. A gene annotation database was build based on the UniGene and Ensembl databases, allowing gene identification for the probes on the microarray chip. Following data import, various missing value imputation and normalization methods can be applied to the data, for both oligonucleotide and cDNA arrays. Additional information, such as CGH data for the samples, can also be specified.

GEPAT offers various analysis methods for gene expression data. Hierarchical, k-means and PCA clustering methods allow group detection in probes and samples. With these detected groups or a predefined group set, a linear model based t-Test can be used to identify differences in gene expression between groups. An M/A-Plot, filtering and sorting on fold-change, p-value and probe variance allow a quick identification of differentially expressed genes, a GO category enrichment analysis shows categories with elevated number of differentially expressed genes.
For an interpretation of the results, GEPAT uses data from the Ensembl database and provides information about gene names, chromosomal location, GO categories and enzymatic activity for each probe on the chip. Gene interaction and association data from the STRING database, overlaid with analysis results, e.g. fold change values, can be used to find functionally related genes. The enzymatic information for the genes is used to overlay analysis results onto KEGG pathway maps, allowing an overview of the regulation of metabolic pathways. To check for chromosomal aberrations a chromosome overview can be generated, showing analysis results and optional CGH data on the chromosome set. For a further investigation of gene functions, a tree view of the Gene Ontology graph has been implemented.

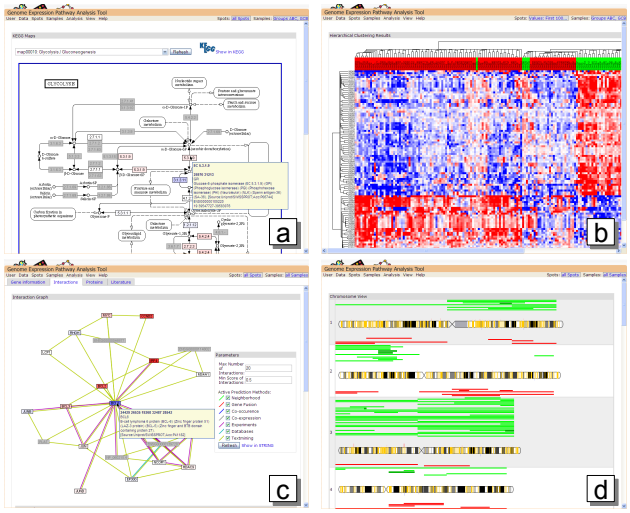

Fig.1: Different views of GEPAT: a) shows a KEGG pathway overlayed with t-Test results, b) shows hierarchical clustering results, c) shows gene associations overlayed with t-Test results, d) shows chromosome overview with CGH data for 82 samples

For easy usage of GEPAT we provide an application-like environment in the web browser. Drop-down-Menus and dialog windows provide the look and feel of a desktop application. Computational intensive analysis tasks are directed to our cluster network, allowing a large number of users at the same time.

GEPAT is freely accessible for academic or non-profit users at
http://bioapps.biozentrum.uni-wuerzburg.de/GEPAT/