BMC Bioinformatics

Software

GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments

Andreas Keller*†1, Christina Backes†1, Maher Al-Awadhi1, Andreas Gerasch2, Jan Küntzer1, Oliver Kohlbacher2, Michael Kaufmann2 and Hans-Peter Lenhof1

Address: 1Center for Bioinformatics, Saarland University, Saarbrücken, Germany and 2Wilhelm Schickard Institute for Computer Science, Eberhard Karls University, Tübingen, Germany

Email: Andreas Keller* - ack@bioinf.uni-sb.de; Christina Backes - cbackes@bioinf.uni-sb.de; Maher Al-Awadhi - maal5003@stud.uni-saarland.de; Andreas Gerasch - gerasch@informatik.uni-tuebingen.de; Jan Küntzer - kuentzer@bioinf.uni-sb.de; Oliver Kohlbacher - oliver.kohlbacher@uni-tuebingen.de; Michael Kaufmann - mk@informatik.uni-tuebingen.de; Hans-Peter Lenhof - lenhof@bioinf.uni-sb.de

* Corresponding author †Equal contributors

Published: 22 December 2008

BMC Bioinformatics 2008, **9**:552 doi:10.1186/1471-2105-9-552

Received: 17 June 2008
Accepted: 22 December 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/552

Abstract

Background: High-throughput methods that allow for measuring the expression of thousands of genes or proteins simultaneously have opened new avenues for studying biochemical processes. While the noisiness of the data necessitates an extensive pre-processing of the raw data, the high dimensionality requires effective statistical analysis methods that facilitate the identification of crucial biological features and relations. For these reasons, the evaluation and interpretation of expression data is a complex, labor-intensive multi-step process. While a variety of tools for normalizing, analysing, or visualizing expression profiles has been developed in the last years, most of these tools offer only functionality for accomplishing certain steps of the evaluation pipeline.

Results: Here, we present a web-based toolbox that provides rich functionality for all steps of the evaluation pipeline. Our tool GeneTrailExpress offers besides standard normalization procedures powerful statistical analysis methods for studying a large variety of biological categories and pathways. Furthermore, an integrated graph visualization tool, BiNA, enables the user to draw the relevant biological pathways applying cutting-edge graph-layout algorithms.

Conclusion: Our gene expression toolbox with its interactive visualization of the pathways and the expression values projected onto the nodes will simplify the analysis and interpretation of biochemical pathways considerably.

Background

Recent biotechnological advances provide the basis for high-throughput techniques that allow for measuring the expression of thousands of genes or proteins simultaneously. Both, the sheer size of the resulting data sets and its noisiness necessitate powerful automatic procedures for normalizing and evaluating these expression profiles. cDNA microarrays that allow for quantifying the expression levels of a wide variety of transcripts have become one of the most important experimental data source in the

life sciences. Usually, transcript levels are measured under different conditions, resulting in two or more sets of expression profiles that have to be compared and analyzed in order to detect differentially expressed genes. Thereby, biochemical categories and pathways that exhibit different expression activities and thus different biochemical behavior can be detected.

For the statistical evaluation of gene sets, many standalone as well as web-based tools have been implemented over the past years [1]. The long list of published programs includes FatiGO [2], BiNGO [3], and GOstat [4] that analyze only enriched Gene Ontologies [5]. For microarry data, ErmineJ [6], CRSD [7], or GSEA-P [8] have been proposed. Other tools allow for the analysis of arbitrary experimental data (e.g. WebGestalt [9], Babelomics [10], or GeneTrail). Another class of approaches focuses on the pre-processing of microarray data and provides only basic statistical analysis, but does no offer methods for gene set enrichment analysis: PMmA [11] was one of the first tools for the detection of differentially expressed genes. The program NMPP [12] is tailored for the preprocessing of self-designed NimbleGen microarray data. Other tools, as AMDA [13] offer clustering methods and functional annotation of the differentially regulated genes. More examples of tools focusing on preprocessing and basic statistical evaluation are ArrayPipe [14], one of the first web-based application, or GEPAS [15], which provides clustering methods and can correlate its results to diverse clinical outcomes. Most recently, Morris et al. [16] described a comprehensive collection of perl modules for microarray management and analysis. However, none of these tools provide a dynamic graphical representation of the detected pathways. This has to be done manually using one of the existing network visualization tools. One of the most popular visualizers with a large user and developer base is Cytoscape [17], which also offers a plugin architecture allowing to extend the functionality, e.g., for integrating data analysis methods. Other visualization tools for biological interaction data are VisANT [18], which has been designed specifically for the integrative visual data-mining of biological pathways, and OSPREY [19], which has been developed to explore large networks.

Here, we present the first framework that integrates data retrieval, pre-processing, gene set enrichment analysis, and network visualization. Our tool, called GeneTrailExpress (GTXP), represents a pipeline tailored for mining information from microarray experiments that offers rich functionality for all crucial steps of microarray evaluation. Notably, the gene set analysis of GTXP relies on our tool GeneTrail [20].

Results and Discussion

Our web-based application GeneTrailExpress integrates all steps of a microarray analysis pipeline, as the workflow shown in Figure 1 outlines. GTXP guides the user through data retrieval, normalization, gene scoring, and the selection of biological categories for gene set analysis. After the gene set analysis has been carried out, the results are presented as a list of significant categories and pathways. Finally, the computed pathways can be visualized using a novel graph visualization tool called BiNA (Biological Networks Analyzer).

Data integration

To perform gene set analyses, a variety of biochemical data extracted from heterogeneous databases is required, including regulatory and metabolic pathways from KEGG [21] and TRANSPATH [22], Gene Ontologies (GO) [5], and many more. Since GTXP imports most of these data sets from the biochemical network library BN++ [23,24] and the underlying database BNDB [25], the user only needs to load up the expression profiles to be analyzed. To this end, our tool offers a database connection to the NCBI Gene Expression Omnibus (GEO) [26]. Of course, the user can also upload his own flatfiles containing expression profiles.

Pre-processing

For the different types of analyses, including normalization and gene scoring, various statistical methods are offered. To this end, we implemented a comprehensive C++ module that handles the statistical pre-processing of the expression profiles. Several normalization techniques are provided, as mean value normalization, median value normalization, or a normalization of mean and variance. The distributions of expression values before and after normalization are presented via bar charts.

Furthermore, several scoring functions for the computation of the differential expression are available: mean fold-change, median fold-change, unpaired t-test, paired t-test, Wilcoxon Mann-Whitney test, ANOVA, and Wilcoxon Rank-Sum test. The distribution of resulting scores is shown as a histogram.

Additionally, a list of all transcripts sorted by their score is generated. A brief summary on the scoring methods and application prerequisites can be found on the GTXP web interface. To test the stability and correctness of the implemented statistical tests, we cross-checked the results of GTXP with those of *R*, a widely used programming language for statistical computations.

Gene Set Analysis

For the statistical evaluation of gene sets we apply our gene set analysis tool GeneTaril [20] that offers both com-

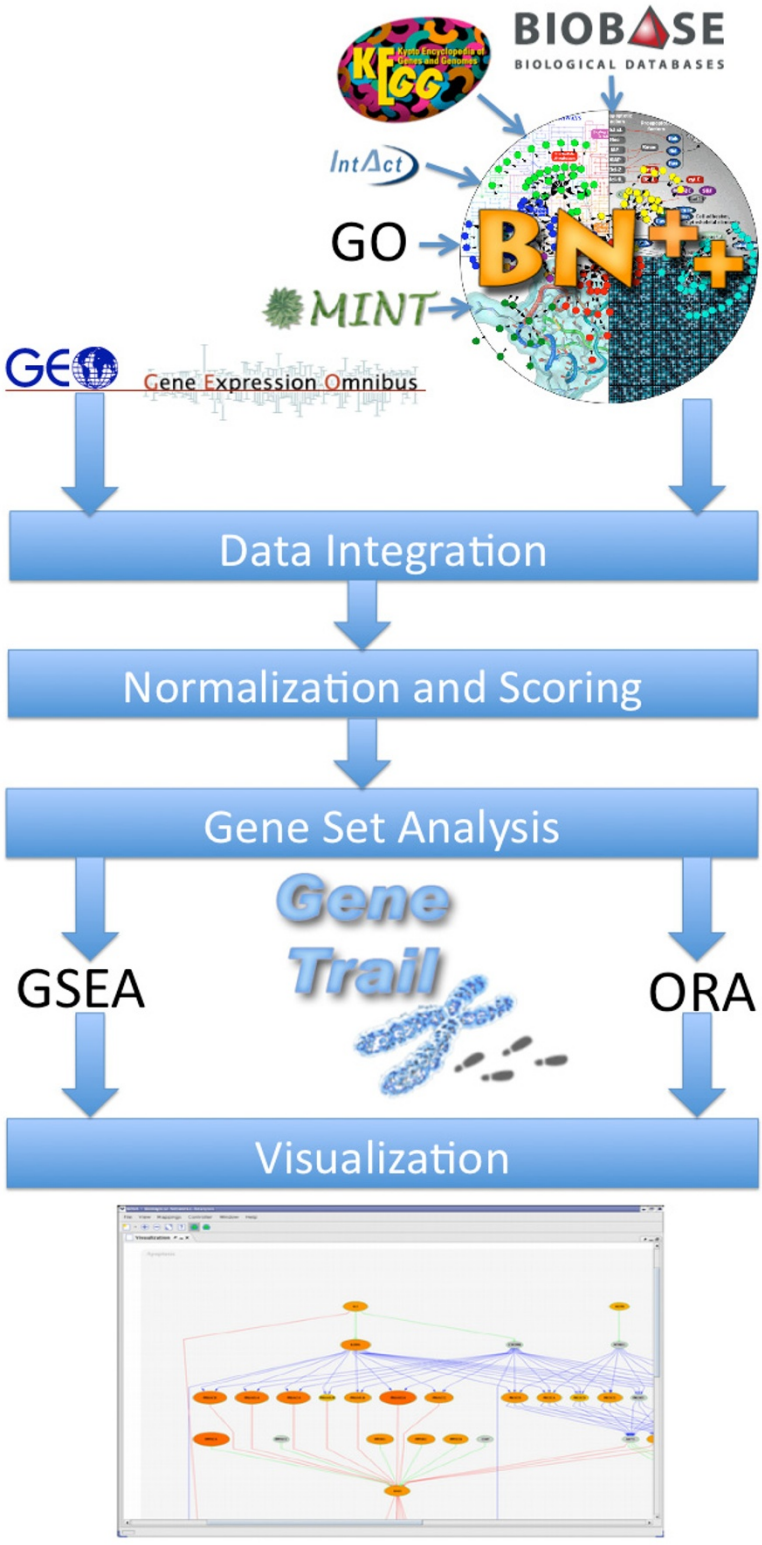

Figure 1
GTXPs Workflow.

mon statistical approaches. The first method, the so-called "Over-Representation Analysis" (ORA), compares the set of interest to a reference set. When considering a certain biochemical category as a GO term, ORA tries to detect if this category is over-represented or under-represented in the respective gene set and computes its significance either by Hypergeometric test or by Fisher's test. The second method, which is cutoff-free, is called "Gene Set Enrichment Analysis" (GSEA). Here, the input set is sorted by some specific criteria (e.g., gene expression values). When considering an arbitrary functional category, GSEA tests if the genes in the set that belong to the category are randomly distributed or accumulated on top or on bottom of the sorted input list. While other tools estimate the GSEA p-values by non-parametric permutation tests, GeneTrail computes exact p-values by an efficient dynamic programming algorithm [27]. For a more precise description of both methods, GSEA and ORA, we refer to [20]. Other strengths of GeneTrail include the support of many organisms (among others *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana* and *Staphylococcus aureus*) and many biological categories (among others KEGG and TRANSPATH pathways, Gene Ontologies, transcription factors from TRANSFAC and sequence analyses). To integrate the diversity of data is realized by using the biochemical network library BN++ [23,24]. As comprehensive data source, BN++ can grasp a plenty of information of the underlying database BNDB [25].

GTXP enables the user to carry out GSEA and ORA analyses by including GeneTrail. For GSEA, the entire sorted gene list is used as input. For ORA, the gene list has to be separated in a test and a reference set. To this end, our tool provides different options: the user can decide to take the first x genes in the list, the first x percent of genes, or all genes with a score above or below a threshold as test set. In each case, the reference set contains all genes that are not included in the test set. For both gene set analysis approaches, GSEA and ORA, the biological categories to be analyzed can be chosen via a menu. After the gene set analysis has been carried out, the significant catgories are listed, sorted by the respective p-values.

Network Visualization using BiNA

As discussed in the Background section, several tools for network visualization have been published in the last decade. We have developed BiNA, the Biological Network Analyzer, a visual analytics tool for biochemical networks. While a detailed description of BiNA is beyond the scope of this work, we will sketch its architecture and highlight its special features that are the reason for using BiNA in this project. BiNA consists of two parts, the platform and a plugin system. While the platform as central element of BiNA contains the graphical user interface and many common utilities, it does not have any possibilities for displaying or analyzing networks. For this task, we developed a powerful plugin structure, which plays an important role, both for the visualization of networks and also for the integration of BiNA into the BN++ framework. Besides the standard Java version, we also implemented a Java Webstart of BiNA allowing the seamless integration into websites.

BiNA builds upon our integrative system BN++ and the underlying comprehensive data warehouse BNDB. This warehouse system ensures a full semantic integration of many databases, including KEGG and TRANSPATH. Since GeneTrail relies on the same data warehouse system, the usage of BiNA ensures that the user gets visual representations of exactly the data that are analyzed by our gene set analysis tool. Since GTXP uses the Webstart version of BiNA, GeneTrail adds for each significant network a link on the results page. By following this link, the user directly generates a visualization of the respective network. To integrate the pathway data, we equipped BiNA with an SQLite interface to the BN++ database BNDB. If a pathway visualization is started for the first time, BiNA and all available topological network information are downloaded (about 40 MB) and stored on the local hard drive. Whenever BiNA is used again, a version control is carried out ensuring that the newest version of BiNA and the pathway topology information are available on the local disk. Thereby, an efficient visualization is guaranteed, even if the respective networks are large.

A key feature of BiNA is the comprehensive set of available graph layout algorithms. It includes most standard graph layouts (e.g., organic, circular, and hierarchical), but, in addition, also provides biologically inspired graph layouts, implementing the drawing conventions common in textbooks and allowing for a dynamical visualization of the networks using the static KEGG layout information. Moreover, BiNA provides convenient interactive analysis and navigation capabilities. Among others, BiNA allows to map arbitrary scalar data, like expression data, onto the biological networks. If a significant network is visualized by BiNA, the genes on the path are directly colored by their scores, facilitating the interpretation of the statistical evaluation. Figure 2 shows BiNA's graphical user interface visualizing a real biological example. A GSEA of lung cancer expression data reveals overexpression of lung cancer genes in the Cell Cycle, indicated by the red-colored genes.

Conclusion

In this study, we present GeneTrailExpress, a toolbox that helps researchers to analyze and interpret expression data. The user is intuitively guided through all analysis steps of the pipeline. A main strength of our application is the integrated graph visualization tool that enables the user to

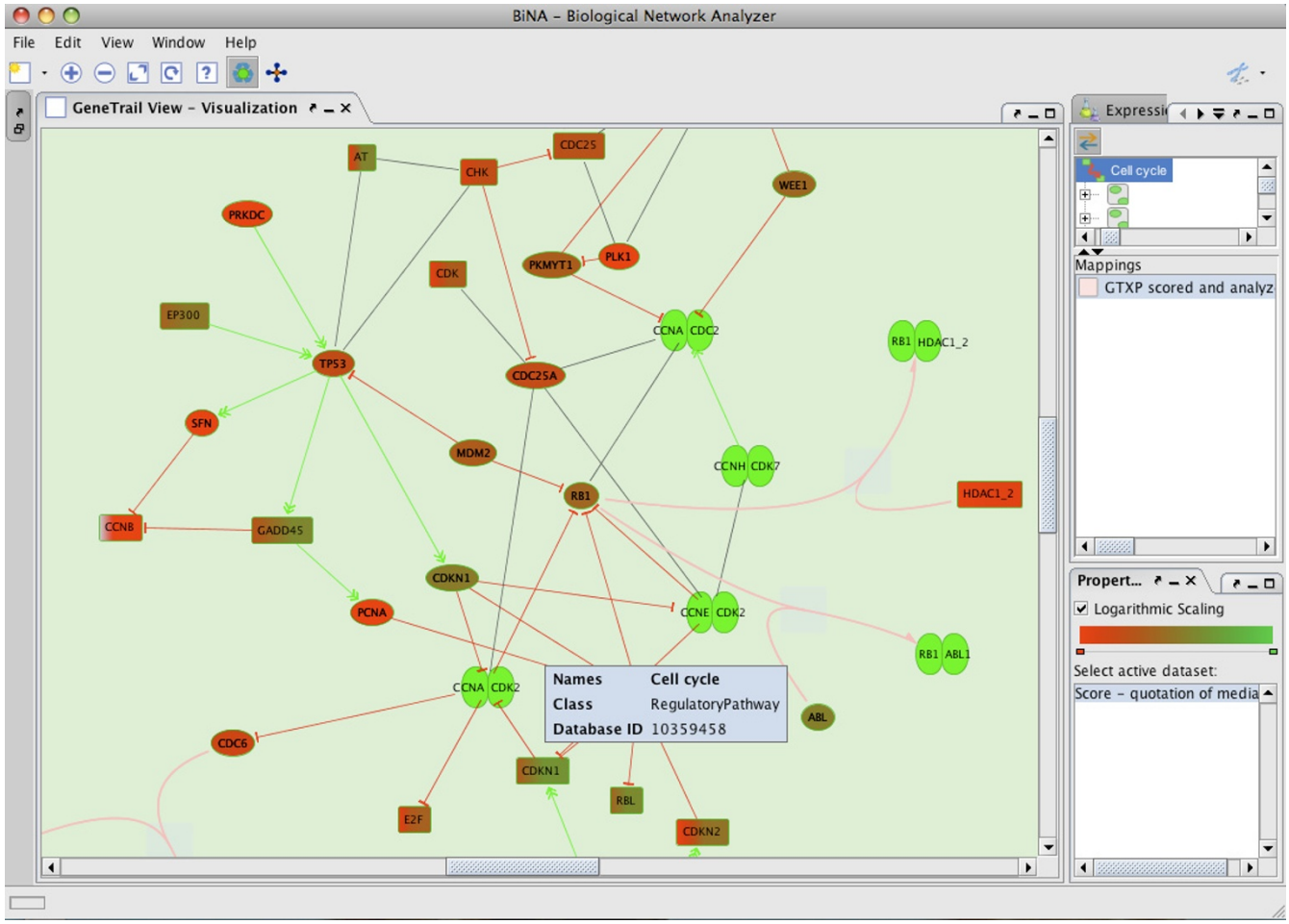

Figure 2
BiNA GUI. For the gene expression omnibus data set GDS1312, containing human lung cancer samples and normal controls, the result of the cell cycle pathway is sown. The performed gene set enrichment analysis computed a p-value of 0.0074, providing evidence for a clear up-regulation of the cell cycle in lung cancer. All genes are colored with respect to their over-expression, the tale green complexes correspond to protein complexes.

draw the relevant biological pathways applying cutting-edge graph-layout algorithms. This interactive visualization of the pathways with the expression values projected onto the nodes facilitates the interpretation of significant findings considerably.

Authors' contributions

AK and CB implemented the gene set enrichment method, AK contributed in writing the manuscript. MA implemented the web interface and the pre-processing. AG developed the BiNA tool. JK contributed to the data integration. MK, OK, and HPL are the senior authors and wrote the manuscript.

Availability and requirements

Project name: GeneTrailExpress

Project homepage: http://genetrail.bioinf.uni-sb.de

Operating system: Platform independent

Programming language: Java, C++, php

Other requirements: JavaWS version 1.6 or higher

Acknowledgements

This work has been funded by DFG Priority Program SPP 1335: LE 952/3-1, KO 2313/3-1, KA 812/13-1

References

1. Nam D, Kim S: **Gene-set approach for expression pattern analysis.** *Brief Bioinform* 2008.
2. Al-Shahrour F, Díaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20:**578-580.

3. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.** *Bioinformatics* 2005, **21:**3448-3449.
4. Beissbarth T, Speed T: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20:**1464-1465.
5. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
6. Lee H, Braynen W, Keshav K, Pavlidis P: **ErmineJ: tool for functional analysis of gene expression data sets.** *BMC Bioinformatics* 2005, **6:**269.
7. Liu C, Lin C, Chen W, Chen H, Chang P, Chen J, Yang P: **CRSD: a comprehensive web server for composite regulatory signature discovery.** *Nucleic Acids Res* 2006, **34:**W571-577.
8. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov J: **GSEA-P: a desktop application for Gene Set Enrichment Analysis.** *Bioinformatics* 2007, **23:**3251-3253.
9. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic Acids Res* 2005, **33:**W741-748.
10. Al-Shahrour F, Minguez P, Vaquerizas J, Conde L, Dopazo J: **BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments.** *Nucleic Acids Res* 2005, **33:**W460-464.
11. Vicentini R, Menossi M: **Pipeline for macro- and microarray analyses.** *Braz J Med Biol Res* 2007, **40:**615-619.
12. Wang X, He H, Li L, Chen R, Deng X, Li S: **NMPP: a user-customized NimbleGen microarray data processing pipeline.** *Bioinformatics* 2006, **22:**2955-2957.
13. Pelizzola M, Pavelka N, Foti M, Ricciardi-Castagnoli P: **AMDA: an R package for the automated microarray data analysis.** *BMC Bioinformatics* 2006, **7:**335.
14. Hokamp K, Roche F, Acab M, Rousseau M, Kuo B, Goode D, Aeschliman D, Bryan J, Babiuk L, Hancock R, Brinkman F: **ArrayPipe: a flexible processing pipeline for microarray data.** *Nucleic Acids Res* 2004, **32:**W457-459.
15. Herrero J, Al-Shahrour F, Díaz-Uriarte R, Mateos A, Vaquerizas J, Santoyo J, Dopazo J: **GEPAS: A web-based resource for microarray gene expression data analysis.** *Nucleic Acids Res* 2003, **31:**3461-3467.
16. Morris J, Gayther S, Jacobs I, Jones C: **A suite of Perl modules for handling microarray data.** *Bioinformatics* 2008, **24:**1102-1103.
17. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Research* 2003, **13:**2498-2504.
18. Hu Z, Mellor J, Wu J, DeLisi C: **VisANT: an online visualization and analysis tool for biological interaction data.** *BMC Bioinformatics* 2004, **5(17):**.
19. Breitkreutz BJ, Stark C, Tyers M: **Osprey: a network visualization system.** *Genome Biol* 2003, **4:**R22.
20. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady Y, Mueller R, Meese E, Lenhof H: **GeneTrail-advanced gene set enrichment analysis.** *Nucleic Acids Res* 2007, **35:**W186-192.
21. Kanehisa M: **The KEGG database.** *Novartis Found Symp* 2002, **247:**91-101.
22. Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D, Michael H, Schwarzer K, Potapov A, Choi C, Kel-Margoulis O, Wingender E: **TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations.** *Nucleic Acids Res* 2006, **34:**D546-551.
23. Sirava M, Schaefer T, Eiglsperger M, Kaufmann M, Kohlbacher O, Bornberg-Bauer E, Lenhof H: **BioMiner-modeling, analyzing, and visualizing biochemical pathways and networks.** *Bioinformatics* 2002, **18(Suppl 2):**S219-230.
24. Kuentzer J, Blum T, Gerasch A, Backes C, Hildebrandt A, Kaufmann M, Kohlbacher O, Lenhof HP: **BN++ – A Biological Information System.** *Journal of Integrative Bioinformatics* 2006, **3:**34.
25. Kuentzer J, Backes C, Blum T, Gerasch A, Kaufmann M, Kohlbacher O, Lenhof H: **BNDB – the Biochemical Network Database.** *BMC Bioinformatics* 2007, **8:**367.
26. Edgar R, Domrachev M, Lash A: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30:**207-210.
27. Keller A, Backes C, Lenhof H: **Computation of significance scores of unweighted Gene Set Enrichment Analyses.** *BMC Bioinformatics* 2007, **8:**290.