

Combining Text Mining and Microarray Analysis

Christoph M. Friedrich^{1,*} and Michaela Gündel¹

1. Fraunhofer Institute for Algorithms and Scientific Computing (SCAI); Department of Bioinformatics; Schloss Birlinghoven; D-53754 Sankt Augustin; Germany
* Contact author: friedrich@scai.fraunhofer.de

Keywords: Text mining, Microarray workflow, Intracranial Aneurysms

The screenshot shows a web-based application interface for biological data analysis. At the top, there's a navigation bar with links like 'Entry Tree View', 'Select Entry Class to view and search', 'Documents', 'Entries', and 'Analysis'. Below the navigation is a search bar with the query 'Intracranial AHD aneurysms'. A message indicates 'The following entities relating to "Intracranial AHD aneurysms" AND MESH genetics were found in 154 documents.' The main area displays a table with 154 rows of data. The columns include 'Entity', 'Relative Entropy', 'Drug Target?', 'Cytoband', 'Doc Count', and 'Date Reported'. Each row contains a small thumbnail image of a brain scan and several icons for further actions. On the left side of the interface, there's a sidebar with a tree view of categories such as 'Human Genes / Proteins', 'Chromosomal Location', 'Protein', 'non-normalized SNP', 'Normalized SNP', 'Normalized CTF SNP', 'Normalized Normal', 'XPC-Ace', 'Cell', 'Hazardous Chemicals', 'Protein-Disease', 'Protein-Pathway', 'Protein-Target', 'Protein-Phenotype', 'Protein-Subcellular', 'Protein-System', 'Risk Factor', 'Risk Factor for Intracranial A.', 'Risk Factor for Aneurysms', 'Aneurysm in Context', 'Sign', 'Aneurysm Diagnos.', 'General Aneurysm Associates', and 'Location of Intracranial Aneurysm'. There are also buttons for 'Select Confidence' and navigation arrows.

The microarray analysis workflow will be presented in detail as well as “lessons learnt” during the development and use. Additionally it will be shown how text mining is combined with microarray analysis in this Knowledge Environment [5].

Acknowledgements

This work has been partially funded in the framework of the European integrated project @neurIST, which is co-financed by the European Commission through the contract no. IST-027703 (see <http://www.aneurist.org>)

References

1. Gentleman, R.; Carey, V.; Bates, D.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. and Zhang, J. Bioconductor: open software development for computational biology and bioinformatics *Genome Biology*, **2004**, 5, R80 <http://www.bioconductor.org>; last accessed 2009-02-26
2. Gündel, M. *ArrayProcess: Work Flow for Microarrays*; Masters thesis, Life Science Informatics at Bonn-Aachen International Center for Information Technology (B-IT); Germany, **2007**
3. Smyth, G. K. *Limma: linear models for microarray data*. In Gentleman, R. et al. (ed.) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, Springer, **2005**, 397-420
4. Friedrich, C. M.; Dach, H.; Gattermayer, T.; Engelbrecht, G.; Benkner, S., and Hofmann-Apitius, M. *@neuLink: A Service-oriented Application for Biomedical Knowledge Discovery* Proceedings of the HealthGrid 2008, IOS Press, **2008**, 165-172
5. Hofmann-Apitius, M.; Fluck, J.; Furlong, L. I.; Fornes, O.; Kolarik, C.; Hanser, S.; Boeker, M.; Schulz, S.; Sanz, F.; Klinger, R.; Mevissen, H.-T.; Gattermayer, T.; Oliva, B. and Friedrich, C. M. *Knowledge Environments Representing Molecular Entities for the Virtual Physiological Human* Philosophical Transactions of the Royal Society A, **2008**, 366(1878), 3091-3110

Microarrays are well established experimental tools to measure gene expression in biological samples. The Bioconductor project [1] provides a wealth of R packages to analyse the resulting gene expression data. In [2] a semi-automatic microarray analysis workflow with limma [3] as the main analysis engine has been developed. This workflow is part of @neuLink [4], a biomedical Knowledge Discovery application suite. Another module of this application suite allows for Knowledge Discovery in biomedical text sources, namely Medline. Combining prior published knowledge with experimental data allows for example the identification of co-mentioned diseases or drugs in similar published gene expression data.