Understanding the Biological Functions of Gene Sets

Catharina Lippmann[1,2], Jörn Loetsch[1,3] and Alfred Ultsch[2]

[1] lippmanc@mathematik.uni-marburg.de
Fraunhofer Project Group Translational Medicine and Pharmacology (IME-TMP), Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany

[2] ultsch@mathematik.uni-marburg.de
DataBionics Research Group, University of Marburg, Hans-Meerwein-Straße, 35032 Marburg, Germany

[3] j.loetsch@em.uni-frankfurt.de
Institute of Clinical Pharmacology, Goethe - University, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany

Next generation sequencing[1], microarray analysis[2] and genetical database searches, for example, for genes involved in pain[3], produce sets of genes which can contain several hundreds of genes. The central research question for these gene sets is, which biological roles/functions these genes in the organism perform[3]. To answer this question, an Over Representation Analysis (ORA)[4,5] can be applied using the Gene Ontology (GO) knowledge bases[6]. The answer given by an ORA is a directed acyclic graph (DAG)[7]. This DAG is a hierarchical representation of knowledge in form of a graph with nodes containing terms and connected by edges pointing to more detailed descriptions. Such a DAG represents the complete knowledge for the answer and may contain hundreds of GO terms. By its sheer size this obscures the understanding of the main biological functions of the genes[8]. Functional Abstraction derives for this rather unintelligible DAG a small number of topics that highlight different aspects of the functions of the gene set[8]. This allows the identification of new and so far overseen aspects[3]. The abstraction is achieved by first computing a numeric value for all nodes describing the remarkableness of the terms in the DAG. Remarkableness is a function of certainty and information value of a term. For each path from a leave to the root of the DAG the maximum remarkableness identifies an important term, called headline. Subsequently the number of headlines is reduced by subsumption or expanded by detailization using the structural features of the DAG[8]. In this work the division of the complete DAG into separate function specific DAGs which describe the Functional Areas is presented and compared to other approaches on several examples from current research on cancer and pain.

[1] Metzker, M.L. Sequencing Technologies — the next Generation, Nature Reviews Genetics, 11, pp 31-46, 2010.

[2] Quackenbush, J. Computational Genetics: Computational Analysis of Microarray Data, Nature Reviews Genetics, 2, pp 418-427, 2001.

[3] Lötsch, J., Doehring, A., Mogil, J.S., Arndt, T., Geisslinger, G., Ultsch, A. Functional Genomics of Pain in analgesic Drug Development and Therapy, Pharmacol. Ther., 139, pp 60–70, 2013.

[4] Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Müller, R., Meese, E., Lenhof, H.-P. GeneTrail – advanced Gene Set Enrichment Analysis, Nucleic Acids Res., 35: pp 186–192, 2007.

[5] Khatri, P., Draghici, S. Ontological Analysis of Gene Expression Data: current Tools, Limitations, and open Problems, Bioinformatics, 21, pp 3587–3595, 2005.

[6] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: Tool for the Unification of Biology, Nature Genet., 25, pp25-29, 2000.

[7] Zhang, S., Cao, J., Kong, Y.M., Scheuermann, R.H. GO-Bayes: Gene Ontology-based Overrepresentation Analysis using a Bayesian Approach, Bioinformatics, 26, pp 905–911, 2010.

[8] Ultsch, A., Lötsch, J. Functional Abstraction as a Method to Discover Knowledge in Gene Ontologies, PLoS One, 9(2), pp. 90-191, 2014.