Explaining genome-wide association study results using concept profile analysis and the Kyoto Encyclopedia of Genes and Genomes pathway database

Kristina M. Hettne[1], Harish Dharuri[1], Reinout van Schouwen[1], Peter A.C. 't Hoen[1,2], Barend Mons[1,2], Marco Roos[1,2]

[1] Department of Human Genetics, Leiden University Medical Centre, Leiden, The Netherlands
[2] Netherlands BioInformatics Centre, The Netherlands

Genome-wide association studies (GWAS) with metabolomic phenotypes yield several statistically significant single nucleotid polymorphism (SNP)-metabolite associations (e.g. [1]). The information needed to arrive at an understanding of the mechanistic basis of the association requires integration of disparate structured (such as pathway databases) and unstructured (such as scientific literature) data sources. For example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) RESTful Web services [2] can be used for pathway annotation, and the by us developed concept profile analysis Web services [3] can be used as a source of text-mining based annotation. The concept profile analysis technology uses the vector space model to relate two concepts (such as SNPs and biological process from the Gene Ontology) to each other and measure the strength of the relationship [4].

We evaluated the utility of KEGG pathways and concept profiles in facilitating the biological interpretation of statistically significant SNP-metabolite pairs using the Illig et al. GWAS dataset [1]. Workflows utilizing the KEGG Web services and concept profile analysis Web services were created in the Taverna workbench 2.4 [5] and made available on the workflow collaborative platform myExperiment [6,7]. Our workflow based on the KEGG pathway database was able to map 10 out of the 15 top hits in the Illig et al. study, to genes that participated in pathways relevant to the associated metabolite. The text-mining workflow was also able to reproduce 10 of the 15 manually curated SNP functions, and gave suggested annotations for the remaining five that can serve as material for further investigation and wet lab validation. This gives us a sensitivity measure of 67 % (10/15). This high sensitivity is a validation of the method that seeks to utilize background knowledge present in pathway databases and literature, to make sense of SNP-metabolite pairs from genome-wide association studies of intermediate phenotypes.

1. Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmüller G, Kato BS, Mewes HW, Meitinger T, de Angelis MH, Kronenberg F, Soranzo N, Wichmann HE, Spector TD, Adamski J, Suhre K: A genome-wide perspective of genetic variation in human metabolism. Nature Genetics 2010, 42(2):137-41.
2. http://www.kegg.jp/kegg/rest/keggapi.html
3. http://www.biocatalogue.org/services/3330#overview
4. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA: Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome biology 2008, 9(6):R96.
   http://www.taverna.org.uk
5. http://www.myexperiment.org/workflows/3124
6. http://www.myexperiment.org/workflows/3522