InfoCodex eDiscovery – Delivering High Quality Results

The large pharma enterprise Merck & Co. Inc has conducted a comprehensive experiment in 2011 to investigate the power of InfoCodex for the *discovery of novel biomarkers and phenotypes* for specific diseases by analysing large sets of medical and clinical publications (PubMed etc.). The experiment has proven that the InfoCodex technology indeed can identify new potential biomarkers, enabling accelerated and targeted research (submitted for publication in a joint paper).

The assessment of the discovery results is generally done through a standard precision/recall metrics, and hence this metrics was used for evaluating the efficacy of the proposed semantic methods in recognizing biomarkers/phenotypes. However, this assessment method can be used only for the comparison of detected biomarkers/phenotypes with those already known. For a statistically significant certification of specific novel biomarkers/phenotypes, there are no reliable tools available.

The objective of the Merck experiment was a proof-of-concept for the automatic discovery of **potential novel** biomarkers/phenotypes. The assessment of the identified candidates had to be carried out by human subject matter experts (SME).

Because of intellectual property restrictions imposed by Merck, the most promising novel biomarkers discovered by InfoCodex and validated by an SME have not been published. However, to demonstrate the soundness and high quality of the InfoCodex results, two positive findings can be reported.

1. Melatonin receptor 1B (MTNR1B) has been identified by InfoCodex as a potential phenotype for obesity and diabetes. In March 2011, a PubMed search for “MTNR1B” AND “obesity” returned 9 documents, while the search for “MTNR1B” AND “obesity” AND “phenotype” did not return any documents. However, through its automated analysis of these 9 documents in conjunction with all other available documents, InfoCodex identified MTNR1B as an obesity phenotype. To illustrate the associative identification of MTNR1B as an obesity phenotype, InfoCodex automatically selected two of the 9 documents (PMID: 20200315, 19088850). A human inspection of the two abstracts selected by InfoCodex would indeed identify MTNR1B as a phenotype for obesity.

Now (January 2013), the PubMed search including the criterion "phenotype" finds two new documents identifying MTNR1B as an obesity phenotype:

- "A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium" by the Washington University School of Medicine, PMID: 21386085 (April 2011)
- "A low frequency variant within the GWAS locus of MTNR1B affects fasting glucose concentrations: genetic risk is modulated by obesity" by the University of Oklahoma Health Sciences Center, PMID: 21558052 (Nov. 2012).

2. As reported in http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3192645/pdf/dst-05-0784.pdf (Institute for Clinical Research and Development, Mainz, Germany, Mai 2011), not only insulin is a known diabetes phenotype, but also proinsulin is indeed a robust diabetes biomarker. This was correctly identified by the InfoCodex semantic engine.