BMC Bioinformatics
IMPACT FACTOR 2.75

Research article

Discovery of novel biomarkers and phenotypes by semantic technologies

Carlo A. Trugenberger, **Christoph Wälti**, **David Peregrim**, **Mark E. Sharp** and **Svetlana Bureeva**

BMC Bioinformatics 2013, **14**:51 doi:10.1186/1471-2105-14-51
Published: 13 February 2013

Abstract (provisional)

Background

Biomarkers and target-specific phenotypes are important to targeted drug design and individualized medicine, thus constituting an important aspect of modern pharmaceutical research and development. More and more, the discovery of relevant biomarkers is aided by in silico techniques based on applying data mining and computational chemistry on large molecular databases. However, there is an even larger source of valuable information available that can potentially be tapped for such discoveries: repositories constituted by research documents.

Results

This paper reports on a pilot experiment to discover potential novel biomarkers and phenotypes for diabetes and obesity by self-organized text mining of about 120,000 PubMed abstracts, public clinical trial summaries, and internal Merck research documents. These documents were directly analyzed by the InfoCodex semantic engine, without prior human manipulations such as parsing. Recall and precision against established, but different benchmarks lie in ranges up to 30% and 50% respectively. Retrieval of known entities missed by other traditional approaches could be demonstrated. Finally, the InfoCodex semantic engine was shown to discover new diabetes and obesity biomarkers and phenotypes. Amongst these were many interesting candidates with a high potential, although noticeable noise (uninteresting or obvious terms) was generated.

Conclusions

The reported approach of employing autonomous self-organising semantic engines to aid biomarker discovery, supplemented by appropriate manual curation processes, shows promise and has potential to impact, conservatively, a faster alternative to vocabulary processes dependent on humans having to read and analyze all the texts. More optimistically, it could impact pharmaceutical research, for example to shorten time-to-market of novel drugs, or speed up early recognition of dead ends and adverse reactions.