In silico discoveries for biomedical sciences

Haagen, H. van

Summary

Text-mining is a challenging field of research initially meant for reading large text collections with a computer. Text-mining is useful in summarizing text, searching for the informative documents, and most important to do knowledge discovery. Knowledge discovery is the main subject of this thesis. The hypothesis that knowledge discovery is possible started with the work done by Swanson. He made, as a first finding, links between Raynaud's disease and fish oil using intermediate medical terms to relate them to each other. This principle was formalized in the A-B-C concept. A and C are not directly related to each other but via an intermediate concept B that needs to be discovered.
Tex data can be extended by adding other non textual data such as microarray experiments. Then we are in the field of data-mining. The final goal is to do all kinds of discoveries with computer (in silico) using data sources in order to assist biology research to save time and discover more.

In chapter two we introduced the techniques that are mainly used for the rest of the research. We explained what we mean by concept based text-mining. A concept is an unambiguous unit of thought, meaning that we all agree talking about the same thing. We described how we can define associations between two concepts and the strength of association using statistics. We continued how the direct associations were used to form these A-B-C triplets and used in our text-mining approach called concept profiles. A concept profile is a summary of all concepts related to the main concept stored in a vector with weights. With vector algebra we calculate multiple indirect, A-B-C, links between two concepts. These indirect links we call implicit links. We introduced the statistics used to evaluate our methods such as ROC curves, retrospective analyses, and prioritizers. We concluded what would be the further of text-mining and how it fits into the World Wide Web.

In chapter three we did a large scale analysis on the prediction of protein-protein interactions (PPIs) taken from several protein databases. We compared our concept based text-mining system and concept profiles with MEDLINE, which is word based, and the direct relationship method used by STRING. By direct relationship method we mean that only information is used for two concepts if they co-occur in abstracts, e.g. A-C links, and no C. This is the classical way of doing text-mining. The direct relationship method only detected around 30% of the known PPIs in MEDLINE, concluding that 70% should be detected with indirect or implicit links. The concept profiles outperformed any other method that was based on direct

relations only (Area under ROC curve of 0.90 for concept profiles compared to 0.69 for other methods).
Subsequently we did a retrospective analyses to see if PPIs, added in databases between 2005 and 2007, could be predicted before 2005. Concept profiles showed much better prediction results.
The most interesting result from this analysis was to confirm one prediction in the lab. We made a prioritized list for the protein CAPN3 and predicted PARVB as top candidate with no direct link with CAPN3. It was confirmed with three independent lab experiments that these to proteins physically interact.

We continued on the prediction of PPIs in chapter four. We added five non-textual databases to the text-mining part. This should increase the prediction accuracy and lower the number of false positives. Hence we shifted our analysis from text-mining to data-mining. In this analysis again we used STRING as benchmark. We evaluated different ways to combine data sources. The best method appeared to be Fisher's method for combining single sided p-values. We examined how well our data-mining system was able to predict meaningful protein pairs using three case studies. The first case study was on Dysferlin. This protein showed little information in additional databases and had its information mostly within text. The second case study was on the huntingtin protein. A previous published study of 60 up to 120 putative interaction partners with HTT was used as test data. The prediction of these test samples outperformed that of the STRING method. The last case study on PKD1 showed that adding other databases is also useful for solving homonym problems occurring in text-mining.

In chapter five we switched to another semantic type combination that of the gene-disease and we evaluated how well text-mining is able to predict these kind of relationships. In contrast to the PPI study where we had a large positive set of PPIs, we only evaluated small sets of gene-diseases. This was because it was hard to collect good samples. We generated a new set of 18 known gene-disease pairs known and we used two sets used to evaluate the gene prioritize Endeavour. One contains 10 monogenic diseases and the other six polygenic diseases. We only did roll back analysis to simulate if we could predict these gene-disease pairs. We were able to rank the test gene 2-fold higher than Endeavour on the polygenic diseases.
In this study we delved more into the implicit or indirect links between gene and disease and reasoned if the link was logical. One case study predicted the gene CENPJ when mutated causes Seckel syndrome. Our system was able to rank this gene on position 14 out of more than 12,000 genes before the landmark paper about CENPJ-Seckel syndrome was published.
From these examples and the examples from chapter three is became a burning question how much knowledge can be extracted from text using implicit links, *i.e.*

how much information in the whole of MEDLINE is implicit. We did an analysis on all gene disease pairs and calculated match scores and p-values for each match score. We plotted the p-value against the fraction of explicit links and the fraction of implicit links. We were stunned that for significant scores p-value<0.05 the amount of implicit information already succeeded 80%, concluding that the vast majority of information is implicit.

In chapter six we concluded that implicit information extraction really pays of and that there is far more information in text that we could imagine. However text-mining and data-mining still have their limitations. The best way to solve the shortcomings of the methods is by community annotation. The accuracy of a text-mining system can be increased or even pushed to 100% by manual curation on the internet by millions of users. The ironic thing is that every analysis started *in silico* but ends with the refinement using manual annotation, although it is done by millions of users.