Text Mining Methods as Computational Biology Tools

Alfonso Valencia
Structural Biology and BioComputing Programme
Spanish National Cancer Research Centre (CNIO)
valencia@cnio.es

During the last few years many new Information Extraction and Text Mining methods have been developed and many of them are accessible on the web. Still, we do not have many examples of their integration with those commonly applied to biological problems in Genomics and Systems Biology, despite the current general recognition of the need to use extensively and systematically the information directly extracted from textual sources.

My group has been working in integrating Text Mining approaches in large-scale projects, together with other complementary experimental and bioinformatics methods. In particular in the ENFIN project we have developed new approaches to collect information on proteins interacting with proteins known to form part of the human spindle body complex and to systematically score them by the likelihood of their implication in the formation of the spindle. The predictions of this Text Mining method, combined with those of a collection of other methods based on sequence and structure analysis, have been followed up by detailed experimental verification including in situ localization assays and iRNA screenings. Furthermore, we have developed a Text Mining system to assist human experts in the annotation of spindle related proteins that have allowed us to generate a large collection of validated proteins and text pieces. This new system is now being use to train and test the sequence / structure based prediction methods.

For these, and other, applications of Text Mining it is crucial to have an accurate estimation of the capacity of the current systems. The BioCreative II challenge organized by CNIO, MITRE and NCBI in collaboration with the MINT and INTACT databases (http://biocreative.sourceforge.net, Genome Biology, August 2008 Special Issue) provides such an overview. BioCreative II was organized in two tasks:

1. gene name identification and normalization, where many systems were able to achieve a consistent 80% balanced precision / recall.
2. protein interaction detection, which was divided into four sub-tasks:
   (a) ranking of publications by their relevance on experimental determination of protein interactions
   (b) detection of protein interaction partners in text
   (c) detection of key sentences describing protein interactions and
   (d) detection of the experimental technique used to determine the interactions.

The results were good in the categories of publication raking, detection of experimental methods, and highlighting of relevant sentences, while they pointed to persistent problems in the correct normalization of gene/protein names. It is interesting to notice that the typical performance of the best Text Mining methods is not very different from that of many standard bioinformatics methods, for example structure prediction and protein docking methods. Furthermore, BioCreative has channeled the collaboration of several teams for the creation of the first Text Mining meta-server (The BioCreative Meta-server, Leitner et al., Genome Biology 2008 BioCreative special issue). We are now working in the preparation of BioCreative III, with particular focus in fostering the creation of Text Mining systems that can be integrated in Genome analysis pipelines.