Utilizing literature for biological discovery

Lars J. Jensen[1,2], Jasmin Saric[3] and Peer Bork[1,2]

[1] European Molecular Biology Laboratory, D-69117 Heidelberg, Germany
[2] Max-Delbrück-Centre for Molecular Medicine, D-13092 Berlin, Germany
[3] European Media Laboratory GmbH, D-69118 Heidelberg, Germany

One of the major challenges in bioinformatics is to make biological discoveries by integrating the large quantities of heterogeneous information that we have available today. Although these data sources include genome sequences, microarray expression measurements, protein–protein interaction screens and other high-throughput data types, the single largest resource of scientific information is the vast number of scientific publications.

Mapping gene names across species

One of the many obstacles to integration of all these data is the use of different gene identifiers in different data sources. Establishing which of the identifiers that refer to the same gene is a very labor intensive task; this is in particular true for gene symbols and gene names, however, deciphering these is essential for text mining as most publications exclusively use these exclusively. To integrate large-scale data sets from different sources, one also has to resolve the many different database identifiers, *e.g.* SWISS-PROT identifiers, systematic ORF names, and UniGene cluster identifiers. To facilitate data integration, we have developed a resource of synonymous gene names and database identifiers for the most common eukaryotic model organisms (`http://www.bork.embl.de/synonyms/`) (1), which also provides information on orthologous between the species.

Figure 1 shows the global gene name overlap between different species. When only including the points for three or more species a perfect fit is obtained with an exponential function. Many of the gene names that are found in three or more species refer to entirely unrelated genes in the different genomes, *e.g.* the name `GLP1` refers to no less than five different proteins from four species. It is noteworthy that only half of the gene names that are shared by all six species refer to proteins from only one orthologous group.

Text mining of scientific publications

Although it is in some cases posible via MeSH terms to utilize Medline for data mining without parsing natural language (2), this is not the general case. To parse abstracts, we identify word and sentence boundaries using Tokeniser by Helmut Schmid, perform part-of-speech tagging of each sentence with TreeTagger (`http://www.ims.uni-stuttgart.de/~schmid/`), and use CASS (`http://www.vinartus.net/spa/`) for chunking. To improve the accuracy on Medline abstracts, in particular sentences that mention interactions, TreeTagger was

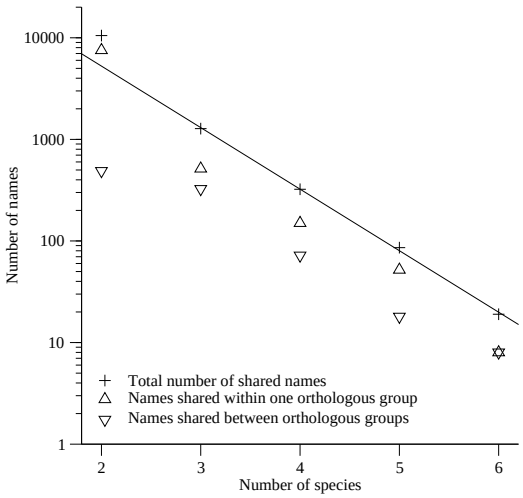

Figure 1: **Number of gene name occurrences follows a Poisson distribution.** The number of species in which each gene name occurs was counted and the distribution was plotted along with number of cases where occurrences can be mapped to one/multiple orthologous groups.

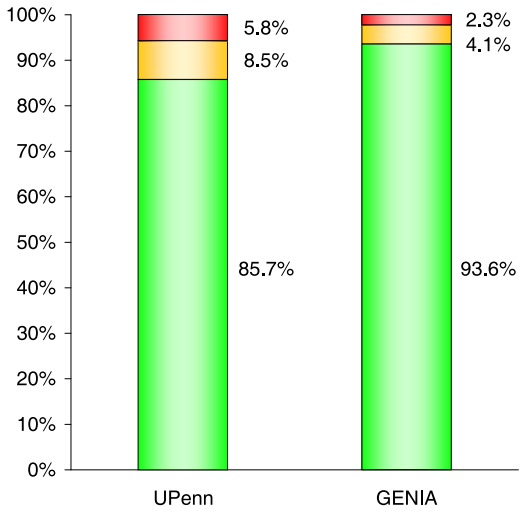

Figure 2: **Retraining of TreeTagger improves performance.** The percentages of correctly tagged tokes (green), questionable tags (yellow), and clear tagging errors (red) are shown using both the standard English parameter file (UPenn) and our retrained tagger (GENIA).

retrained on the manually tagged GENIA 3.0 corpus, also expanding the dictionary with the synonyms list described above. Retraining reduced the error rate two-fold (Figure 2).

Based on the tagged and chunked text corpus, we have developed hand-coded lexico-syntactic patterns for identification of named entities such as genes, promoters, binding sites, and transcription factors. Finally, we build a grammar for recognizing interaction types (*e.g.* "activation") from the relevant named entities. Through integration of the extracted interactions with Chromatin-IP and DNA microarray expression data, it is possible to automatically annotate regulatory networks derived high-throughput experiments with literature references for already known interactions.

As full text electronic access to journals is becoming more and more common, using of more than just abstracts should be considered when doing text mining. For this purpose, the introduction and discussion are the most promising the parts of scientific publications (3). We intend to apply our method to a corpus of full text articles in the near future.

References

[1] L. J. Jensen and P. Bork. *unpublished*, 2003.

[2] C. Perez-Iratxeta, P. Bork, and M. A. Andrade. *Nature Genetics*, 31:316–319, 2002.

[3] P. K. Shah, C. Perez-Iratxeta, P. Bork, and M. A. Andrade. *BMC Bioinformatics*, 4:20, 2003.