BIOINFORMATICS

Mining the bibliome

The secret to successful mining lies in exploiting new resources and efficiently extracting and converting their contents into a valuable and useful commodity. Bioinformaticists (miners of a different sort) have long known of a potentially rich seam of information — the published literature — but have struggled with finding ways to efficiently mine it. Now a new automated text-processing approach to this problem has proved that meaningful biological information can be mined from the literature, while highlighting the language barriers that hinder its extraction. Importantly, it also shows how text-derived information can guide microarray data analysis.

The authors began by selecting human gene identifiers — names, symbols and synonyms — from gene databases, such as the HUGO nomenclature database and LocusLink. The titles and abstracts of ~10 million MedLine records in PubMed were then searched for the occurrence of these identifiers, which represented 13,712 human genes. The selected records were then re-analysed for the occurrence of a second gene, on the assumption that if two genes co-occur in a MedLine record it reflects an underlying biological relationship between them. For their results to be meaningful, Jenssen *et al.* had to prove this assumption to be correct.

This they approached in several ways. By manually assessing 1,000 gene pairs — 500 that co-occur in one record only (low weight) and 500 that co-occur more frequently (high weight) — they found that 40% of the low-weight and 29% of the high-weight gene pairings had no biological basis. These errors predominantly occurred when genes shared their symbols with other genes or with general acronyms. Jenssen *et al.* also found that their gene pairs were under-represented in certain databases. For example, PubGene — their database of gene pairs — contained only 51% of the human, interacting protein pairs present in the Database of Interacting Proteins. This under-representation was again caused by inconsistently used gene names, and by genes having different names to their products and by being absent from a record's title or abstract.

This literature mining produced not only linked gene pairs but also gene networks — could these help explain the co-regulated gene clusters that emerge from microarray gene-expression studies? Jenssen *et al.* found that genes that were co-expressed in two published microarray experiments (by Alizadeh *et al.* and Iyer *et al.*) were also represented in literature-derived subnetworks. Annotating these networks with MeSH (medical subject heading) terms provided further insight into the biological basis of the gene clusters. Furthermore, genes in co-regulated clusters neighboured other loci in the literature-derived networks, providing additional loci for analysis.

Although the authors acknowledge the limitations of their approach, many stem from the inconsistent use of language and gene nomenclature and affect similar text-based bioinformatics strategies. Standardized database indexing and the use of controlled vocabularies are future considerations towards making text-based data-mining a realistic goal.

Jane Alfred

References and links

ORIGINAL RESEARCH PAPER Jenssen, T.-K. *et al.* A literature network of human genes. *Nature Genet.* **28**, 21–28 (2001)

WEB SITE PubGene