nature genetics
volume 28 no. 1 may 2001

Community watch

Microarray analysis yields a wealth of information that presents the geneticist with a vexatious challenge: where does one go with a cluster of genes defined by similar or identical expression levels in different tissues, or during adaptation to a new environment? How does one single out genes that warrant further analysis? Three studies[1–3] (including one presented by Eivind Hovig and colleagues[3] on page 21) describe automatic methods to ferret out hidden patterns in the literature—similar to the way in which microarray analysis reveals patterns in the chaos that is gene expression. As such, they may help to direct sensible enquiry downstream of microarray studies. They also embrace issues central to genetic and genomic studies: language, standards and access.

Hovig and colleagues[3] developed a program called PubGene and used it to search the titles and abstracts of MEDLINE entries from 1966 to mid 2000. They assigned numbers or 'weights' to pairs of gene symbols: the higher the number, the greater the frequency of their co-occurrence (co-occurrence refers to the appearance of both gene symbols in the same MEDLINE entry). In calling up controlled vocabularies linked to gene symbols—in the form of Medical Subject Headings (MeSH) keywords and Gene Ontology (GO) terms[4]—they were able to mine qualitative information that has been organized to permit automated query.

Is this method useful? Its application to two publicly available microarray data sets and comparison with gene pairs described in entries of two other databases (Online Mendelian Inheritance in Man and the Database of Interacting Proteins) indicates that it is. And yet, as both the authors and Daniel Masys[5] (see page 9) point out, the method has limitations.

A substantive limitation is the irregular use of gene symbols in the literature. Gene symbols representing human and mouse genes are established, through consultation with researchers, by nomenclature committees run by the Human Genome Organization and The Jackson Laboratory, respectively. Whereas there is tight correlation between mouse and human gene symbols (for example, the mouse ortholog of *FGF8* is *Fgf8*), gene symbols of orthologous genes in zebrafish, *Caenorhabditis elegans* and other model organisms are less likely to correspond; partly because these organisms have not had the benefit of dedicated nomenclature researchers. The evolution of literature-mining tools, together with the growth of model-organism communities, provides a strong incentive for a greater adherence to standardized nomenclature where it already exists, and renewed efforts to ensure logical connections between the gene symbols of different species where they are lacking.

The mission of the GO consortium is analogous to that of the nomenclature committees: to establish a standard set of words ('controlled vocabularies') that describe information relevant to gene function and remain useful as knowledge of gene and protein function changes. Three sets of vocabularies (called ontologies) have been established. These fall under the headings of molecular function, biological process and cellular component, and provide a means of systematically representing gene function in databases. The vocabularies have been established through dialogue between groups with expertise in the biology of different organisms and who seek feedback from their respective communities. Currently, *Drosophila*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *C. elegans*, *Arabidopsis* and the mouse are represented in the GO database. (It is said that *Escherichia coli* is soon to join.)

Critics of the initiative point out that biology is complicated and that categorizing gene function across species is like trying to herd editors: the only immutable law that governs biology is evolution and its consequences are unpredictable. Certainly, it is a tough job to annotate gene function (and gene symbols) across species or even within the same species. This is not, however, a good argument for not trying, when the benefits of success would seem substantive and the alternatives are not obvious.

And, if adherence is a measure of acceptance, it would seem that GO has achieved some success. Celera, Incyte and AstraZeneca use GO vocabularies. Proteome Inc. (which has now merged with Incyte) was commissioned by the National Center for Biotechnology Information to provide GO annotation to its reference sequences (RefSeq), and both the *Drosophila* and Riken annotation jamborees (the latter convened to annotate mouse cDNAs) made use of the scheme.

It is clear that tools like PubGene will be more powerful with access to full text; there are often valuable nuggets of information provided in the main body of a research article but not referred to in the title and abstract. But, more text means more 'noise'—groups of gene symbols that have more distant biological connections than those typically appearing in an abstract and title could obscure useful information. Analyzing individual paragraphs separately might be one means of overcoming this limitation. In any case, assessing the value of full text to literature-mining is currently unfeasible given the huge amounts of computing power necessary and the need for private subscriptions to gain access to full text of most research articles.

PubGene, the GO initiative and other efforts to establish controlled vocabularies are testimony to the power of *in silico* analysis and cast a new light on the commodity that is published text. The extent to which researchers are able to use published text—as a tool—depends on publishing strategy, and poses interesting questions for the publishers of on-line scientific literature. It is one that will become more interesting as literature-mining tools evolve.

1. Shatkay, H., Edwards, S., Wilbur, W. & Boguski, M.S. in *Intelligent Systems for Molecular Biology* 317–328 (AAAI Press, San Diego, 2000).
2. Masys, D.R. *Bioinformatics* **17**, 1–8 (2001).
3. Jenssen, T.K., Lægreid, A., Komorowski, J. & Hovig, E. *Nature Genet.* **28**, 21–28 (2001).
4. Ashburner, M. *et al. Nature Genet.* **25**, 25–29 (2000).
5. Masys, D.R. *Nature Genet.* **28**, 9–10 (2001).