EMBO *reports*

viewpoint

Mining the bibliome: searching for a needle in a haystack?

New computing tools are needed to effectively scan the growing amount of scientific literature for useful information • *by Les Grivell*

Writing in 1985 in a committee report for the US National Academy of Sciences, Harold J. Morowitz (George Mason University, VA) argued that biological research had reached a point where 'new generalizations and higher order biological laws are being approached, but may be obscured by the simple mass of data' (Morowitz, 1985). Now, 16 years later, his warning has proven to be not exaggerated. In 1985, the *total* number of sequence entries in the EBI nucleotide database was around 5000. In 2001, the number of entries added to the database *per day* was around five times this number. And the increasingly wider application of data-intensive technologies, such as DNA and protein chips, high-throughput protein three-dimensional structure determination and real-time molecular and cellular imaging, have confirmed fears, rational or otherwise, that biologists are likely to be swamped by a digital tsunami of data.

But amongst the many prophesies of doom, relatively little attention has been paid to the consequences of the growing amount of scientific literature. One reason for this neglect may be the fact that this increase has been less dramatic than that of sequence and other databases. It is, nevertheless, still impressive, as evidenced by the latest release notes for the US National Library of Medicine's Medline bibliographic database (www.nlm.nih.gov/databases/databases_medline.html), which stores metadata for more than 11 million articles from some 4300 refereed journals. Another reason may be that electronic access, both to metadata and to full text, has made it considerably easier to search for and use scientific literature. Entrez-PubMed, for instance, NCBI's simple web-based search system, allows scientists to search Medline for bibliographic information, find related publications and, depending on the journal and date of original publication, retrieve the full text of the article, all without leaving their desk. To those of us who started our research careers using hand-written index cards with holes punched into them to allow the selection of article categories by insertion of a knitting needle, PubMed and other services have dramatically short-circuited the route to literature.

Amongst the many prophesies of doom, relatively little attention has been paid to the consequences of the growing amount of scientific literature

Nevertheless, there is no reason for complacency. Given the complexity and sheer magnitude of the task of searching the vast amount of literature and other databases for a certain piece of information, it is necessary to develop improved computer-based tools to aid the human expert. Also, this information is often scattered throughout the published literature and it first must be translated into computer-readable form and associated with the data records to which they are referring. Furthermore, as PubMed is limited to abstracts and keywords, it may miss important information elsewhere in the text of the article.

So far, the availability of full-text articles in digital form, mostly alongside their paper equivalents, has contributed only fractionally to this goal. Such 'online' editions are usually electronic versions of the original papers and available only in formats that are suited primarily to human perusal, such as pdf, html or tif, which limits the possibilities for computer searching and retrieval of full text. Their paper-based origin also makes it difficult for computer-based search algorithms to retrieve and analyse data from other articles or databases that are only referred to.

Additionally, the freedom of natural language provides a considerable challenge for algorithms to extract meaningful information from natural text. And some of the major problems that still limit computer-aided full-text searches have not even been tackled yet. The detection of gene symbols and names, for instance, remains difficult, as researchers have seldom followed logical rules. In some organisms—the fruit fly *Drosophila* is an example—scientists have enjoyed applying gene names with primary meaning outside the biological domain. Names such as *vamp*, *eve*, *disco*, *boss*, *gypsy*, *zip* or *ogre* are therefore not easily recognised as referring to genes (Proux *et al.*, 1998).

Also, both synonymy (many different ways to refer to the same object) and polsemy (multiple meanings for a given word) cause problems for search algorithms. Synonymy reduces the number of recalls of a given object, whereas polsemy causes reduced precision. Another problem is ambiguities of a word's sense. The word insulin, for instance, can refer to a gene, a protein, a hormone or a therapeutic agent, depending on the context. In addition, pronouns and definite articles and the use of long, complex or negative sentences or those in which information is implicit or omitted pose considerable hurdles for full-text processing algorithms.

Despite these problems, initial attempts to extract textual information from scientific abstracts have had moderate success, mainly because they have been limited to specialised biological domains. One example is the field of protein–protein interactions, in which the use of language is simple and direct and in which a number of specialised biological terms abound (Jenssen *et al*., 2001; Ono *et al*., 2001). But these studies are encouraging in two respects. Database annotators now have sets of controlled, discriminatory terms that they can use to improve the quality of annotations. And these initial attempts provide a tantalising glimpse of the power of applying data mining and extraction techniques to combined literature and factual data.

To those of us who started our research careers using hand-written index cards, PubMed and other services have dramatically short-circuited the route to literature

However, much more work is required before other, more complex, areas can be tackled and the methods can be applied to the full text of scientific articles. Indeed, the benefits of full-text searches are unlikely to be limited just to database annotators—efficient searching and discovery of document relationships are cases in point. There may be many situations when the user does not know in advance which combinations of search terms will be sufficiently specific or discriminatory to yield the most relevant hits on a topic of interest. Results may be lost because their titles or abstracts only contain synonyms of the search terms. Or, in order to reduce the number of results returned to a manageable level, a user may be forced to err on the side of greater specificity at the risk of lost sensitivity, for example by the use of multiple terms combined in Boolean logic, such as ‘apoptosis AND mitochondria’. In such cases, the inability to recognise synonyms, or to differentially weight the importance of individual terms, might lead to unjustified inclusion of some hits or exclusion of others.

Although it is only the user who ultimately defines which articles are relevant, various literature retrieval systems now offer some assistance in the form of a document-neighbouring function that is usually based on a measurement of similarity between clusters of documents. For example, PubMed provides each relevant document found with pointers to other articles with a statistically similar usage of discriminatory words in the abstract.

Now consider how much more powerful this would be if it were based on full text, rather than on a 100–200 word summary. This is no idle dream: such neighbouring algorithms were developed more than 30 years ago and have reached high levels of sophistication. One of the pioneers in this area was Gerald Salton from Cornell University (NY), who devised the Vector Space Model (Salton and Buckley, 1991) for document comparisons and developed it further for automatic text analysis, theme generation and summarisation of computer-readable texts (Salton *et al.*, 1994). Salton’s algorithm breaks documents down into simple terms, usually individual words that are indexed and their frequencies calculated. Each term is then assigned a weight that reflects its potential to distinguish any one document from the remainder in the collection. In practice, those terms that are neither too frequent nor too rare turn out to be the best discriminators. The algorithm provides a vector of weighted terms, $D_i = (d_{i1}, d_{i2}, \ldots, d_{it})$, where d_{ik} represents an importance weight for term T_k attached to document D_i. Each document is now characterised by a unique vector that allows easy and quick comparisons with other document vectors. Other approaches include Latent Sematic Indexing (Deerwester *et al.*, 1990) and Concept Indexing (Karypis and Eui-Hong, 2000), which are often used in combination with the Vector Space Model.

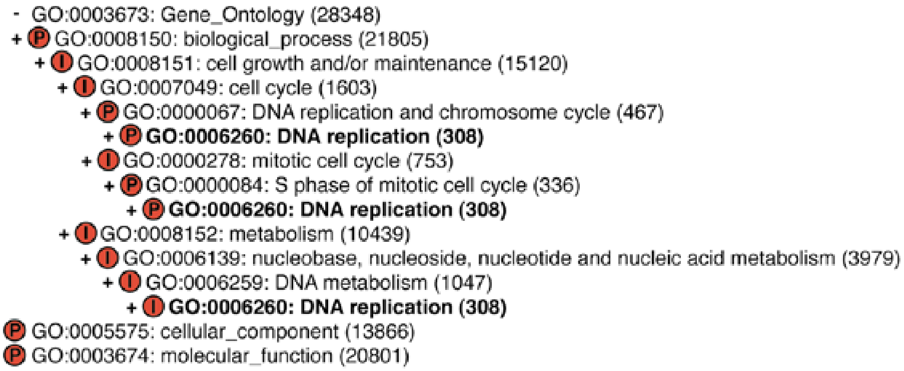
The freedom of natural language provides a considerable challenge for algorithms to extract meaningful information from natural text

Full-text analysis of scientific literature is technically feasible. But, apart from the considerable computational resources required to index terms and pre-compute statistical relationships for several million articles, there are a number of other important issues that have to be addressed when scaling article comparisons from abstract to full-text level. One obstacle is that scientific journals are owned by a large number of different publishers. The discussion raised by the recent PubMed Central and Public Library of Science initiatives (Eisen and Brown, 2001) shows that, in the absence of radical changes to current e-publishing models, content will continue to be scattered over many differently owned sites. Computational tools thus must be able to cope with the analysis of text distributed across multiple locations.

Furthermore, to be of maximal utility, such tools must be capable of both using and providing contextual information. Like a number of major search engines on the world-wide web, they should allow the user to define search terms containing word combinations or phrases and present the results together with the context in which such phrases are used. They should also make use of controlled vocabularies, or, better still, ontologies, which, by defining concepts and logical rules, allow computational methods to describe, organise, interpret and visualise knowledge (Figure 1). In the context of literature analysis, this would permit retrieval of related articles from different fields of research. Such tools should also be able to compare documents in different ways, thereby allowing flexible ranking options. Last but not least, they should allow searches in different human languages.

As yet, none of the current full-text search and analysis tools satisfactorily fulfils all these requirements. At one end of the spectrum, major web engines, such as Google or AltaVista, do a good job of indexing words in texts that are accessible on the net. This also includes documents stored as postscript or pdf files, the latter being a format that has become a *de facto* standard for online versions of articles. Unfortunately, searches on such documents can only be carried out after their conversion back to plain text, a procedure that usually results in some corruption. Luckily, this seems to have little noticeable effect on the results.

But these web engines are necessarily generic. They are unable to deal with synonyms or homonyms specific to a given knowledge domain. And since they do not limit their searches either to particular



```
- GO:0003673: Gene_Ontology (28348)
+ Ⓟ GO:0008150: biological_process (21805)
  + Ⓘ GO:0008151: cell growth and/or maintenance (15120)
    + Ⓘ GO:0007049: cell cycle (1603)
      + Ⓟ GO:0000067: DNA replication and chromosome cycle (467)
        + Ⓟ GO:0006260: DNA replication (308)
      + Ⓘ GO:0000278: mitotic cell cycle (753)
        + Ⓟ GO:0000084: S phase of mitotic cell cycle (336)
          + Ⓟ GO:0006260: DNA replication (308)
    + Ⓘ GO:0008152: metabolism (10439)
      + Ⓘ GO:0006139: nucleobase, nucleoside, nucleotide and nucleic acid metabolism (3979)
        + Ⓘ GO:0006259: DNA metabolism (1047)
          + Ⓘ GO:0006260: DNA replication (308)
+ Ⓟ GO:0005575: cellular_component (13866)
+ Ⓟ GO:0003674: molecular_function (20801)
```

Fig. 1. A section of the Gene Ontology (GO) resource (Gene Ontology Consortium, 2000, www.geneontology.org/), showing the pathways that represent the process of DNA replication as part of a hierarchical tree. Each term is assigned a unique GO identifier. The plus and minus signs indicate that the corresponding terms can be, or have been, expanded to display parent–child relationships. The circled P and I symbols indicate ‘part of’ or ‘is a’ relationships between a particular term and its parent or children. Numbers enclosed in parentheses show the gene associations annotated to this term or to a more specific term below this in the GO tree. The ontology establishes precise, defined relationships between the terms that can be used to implement queries that are much more complex than those possible with simple keywords. As an illustration, this section of the ontology reveals that the process of DNA replication is found as part of three different pathways—cell cycle, mitotic S phase and DNA metabolism—the terms of all of which can be used to recover information about gene products associated with this process. Adapted from Ashburner *et al*. (2001).

domains, or to specific document formats, and as their hit-ranking algorithms are closely guarded secrets, most users tend to directly employ such an engine as a last resort when other approaches have failed. Of more interest for scientists is the ResearchIndex (Citeseer), developed by Steve Lawrence and his colleagues at the NEC Research Institute in Princeton (NJ), which is capable of automatically retrieving relevant publications (www.neci.nec.com/~lawrence/researchindex.html). Provided with a set of keywords, the agent uses generic web search engines to locate and download pdf- and postscript-formatted documents of potential interest. Citeseer then converts these documents to text, parses them to identify semantic features and stores the results in a database that can be subsequently searched by keyword or citation. Articles with similar word usage or common citation patterns can thus be identified and used to build networks of related documents.

Full-text analysis of scientific literature is technically feasible

At the other end of the spectrum is the Collexis indexing engine that was originally developed as part of the AWARE system for clinical and healthcare information networks (Van Mulligen *et al.*, 2000). Based on the Vector Space Model, this engine combines a statistical indexing algorithm with a thesaurus. Currently, it is based on the metathesaurus of the Unified Medical Language System (UMLS) developed at the National Library of Medicine (www.nlm.nih.gov/research/umls/), but that can, in principle, be replaced with an appropriate thesaurus for any other knowledge domain. The UMLS thesaurus consists of three parts: concepts, terms and the words that can be derived from these terms. The indexing algorithm searches for such words in the body of text, references them to the cognate concepts and then computes a fingerprint for each document indexed. Searching and retrieval of documents is simply based on comparisons of these fingerprints, an operation that can be efficiently carried out across a distributed network of document archives. Since the UMLS metathesaurus contains a number of vocabularies that have been translated into various European languages, it is also possible to match documents in different language formats.

In the future, wider use of structured documents will obviate the need for ‘dumb’ search engines and allow portions of text to be tagged and handled as if they were fields in a database. The tool that will make this possible is XML (eXtended Markup Language). In combination with a document type definition (DTD), XML tags can be custom defined for specific purposes to enhance computer searchability and readability, which is instrumental in the further streamlining of the information flow. In combination with XSL (eXtensible Stylesheet Language), XML-tagged documents can be manipulated and displayed in different forms. The notion of computer-understandable documents is currently being extended in the Semantic Web Activity run by the World Wide Web consortium. A new set of languages is being developed in order to make more web content accessible to machines and to facilitate implementation of automated methods for information searching and retrieval (Berners-Lee, 2001). XML tagging will make intelligent searching of full text feasible, fast and informative, and will allow readers to ‘home in’ on, or retrieve and manipulate, specific parts of a publication.

In his witty and thought-provoking glimpse into an ‘*in silico*’ crystal ball, John Allen (Allen, 2000) warned biologists about the dangers of succumbing to the seductive temptation of abandoning traditional hypothesis-based, deductive approaches to their experimental observations in favour of computer-driven induction applied to the masses of genomic and other data currently at their disposal. I agree with most of his arguments, including the greater part of his introductory paragraph, which at least as far as literature searches are concerned, may be closer to the truth than he thinks:

‘If you are reading this article online, you may have retrieved the file because a search engine found a match to your query, indicating that there is something here you may wish to know. A contextual, semantic search will further confirm this and distil the essence of this article. Searching a genome database is exactly the same. Just as computers are transforming the way we communicate and store information, they are changing the way we discover things worth communicating. In the future, automated discovery will generate new knowledge, take over the process of doing science itself, and tell us what it is that we need to know and understand.

The search engines may, by now, be satisfied with this decoy. So, for those who read beyond titles and first paragraphs: do not believe a word of what you have read so far. The title of this article is irony and its introduction parody...'

I think that it is essential to stress that computer analysis of both text and other forms of information represents an aid to creative scientific analysis and not a replacement of it. Automated discovery will help generate new knowledge, but it will not take over the process of doing science itself. I am therefore confident that a scenario in which automated analysis of the written word will be a commonplace part of every biologist's toolbox, and that the search engines employed for this task will not be decoyed by the word usage in his tongue-in-cheek parody!

References

Allen, J. (2000) *In silico veritas*: data-mining and automated discovery: the truth is in there. *EMBO rep.*, **2**, 542–544.

Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.

Ashburner, M. *et al.* (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **1**, 1425–1433.

Berners-Lee, T. (2001) Scientific publishing on the semantic web. www.nature.com/nature/debates/e-access/articles/bernerslee.htm

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Hashman, R. (1990) Indexing by latent semantic indexing. *J. Am. Soc. Inform. Sci.*, **41**, 391–407.

Eisen, M. and Brown, P.O. (2001) Should the scientific literature be privately owned and controlled? www.nature.com/nature/debates/e-access/articles/eisen.htm

Jenssen, T.K., Laegreid, A., Komorowski, J. and Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.*, **28**, 21–28.

Karypis, G. and Eui-Hong (Sam) Han (2000) Concept indexing: a fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical Report TR-00-0016, University of Minnesota, Minneapolis, MN.

Morowitz, H.J. (1985) Models for biomedical research: a new perspective. Report for the National Academy of Sciences. Committee on Models for Biomedical Research, National Academy Press, Washington, DC.

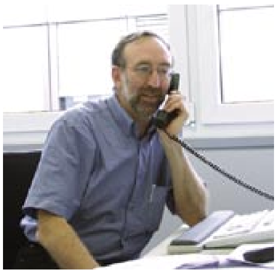
Ono, T., Hishigaki, H., Tanigami, A. and Takagi T. (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.

Proux, D., Rechenmann, F., Julliard, L., Pillet, V. and Jacq, B. (1998) Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. *Genome Inform.*, **9**, 72–80.

Salton, G. and Buckley, C. (1991) Global text matching for information retrieval. *Science*, **253**, 1012–1015.

Salton, G., Allan, J., Buckley, C. and Singhal, A. (1994) Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, **264**, 1421–1426.

Van Mulligen, E.M., Diwersy, M., Schmidt, M., Buurman, H. and Mons, B. (2000) Facilitating networks of information. *Proc. AMIA Symp.*, **2000**, 868–872.



Les Grivell is Manager of E-BioSci, the European electronic publishing initiative.
E-mail: les.grivell@embo.org

DOI: 10.1093/embo-reports/kvf059