PubMed: bridging the information gap

Johanna McEntyre, David Lipman

Review

Synthèse

Drs. McEntyre and Lipman are with the National Center for Biotechnology Information, the National Library of Medicine, the National Institutes of Health, Bethesda, Md.

CMAJ 2001;164(9):1317-9

On-line literature searches of bibliographic databases such as PubMed (www.ncbi.nlm.nih.gov/entrez/query.fcgi) are now integral to the lives of clinicians. A huge amount of knowledge can be gleaned from even a basic PubMed search, while the use of advanced functions can add speed and focus. The 11 million or so abstracts that constitute PubMed undoubtedly make it one of the most significant barrier-free on-line biomedical resources. However, the scientific abstract is but one flavour of information that we use in our professional lives: full-length research articles, clinical trials databases, molecular biology data and so on all contribute to a rich information landscape.

One challenge of the information age is to make the best use of new technologies to present and integrate these diverse information resources in ways that were not previously possible. Forging links between disparate pieces of information can potentiate new ideas that may cross traditional disciplinary boundaries. At the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/), one of our goals is to weave together published information — be it scientific articles, books or molecular sequence data — in such a way that the information is enhanced and the potential for making discoveries is increased. Although PubMed is a mighty stand-alone service, it is significantly enriched by the addition of links to related resources. These can contribute background, refinement and depth. In this short article, we will discuss how we are in the process of integrating PubMed with other information resources in order to build a layered approach to biomedical data.

Links to molecular biology data

When an author submits a paper reporting a new gene sequence, it is usually a prerequisite of acceptance that the sequence is submitted to one of the public databases such as Genbank (www.ncbi.nlm.nih.gov/Genbank/index.html). The new sequence is then cited in the article in the form of an accession number, ensuring that any interested researchers have access to the sequence. Genbank entries in turn usually cite one or more research papers to add biological context to the data, so the GenBank sequence records incorporate links to the PubMed abstracts of the cited papers. We can make use of these to enable reciprocal links from PubMed back to the sequence information, not only for nucleotide sequences, but also for protein sequences and protein structures in their respective public repositories. In individual PubMed records, the links are labelled Nucleotide, Protein or Structure (Fig. 1). In this way, the biological data provide foundation and depth to PubMed.

Books linked to PubMed

Just as PubMed abstracts can provide biological context for molecular data, textbooks can expand on concepts mentioned in abstracts. Abstracts are rich in information, but they are usually somewhat esoteric and do not attempt to explain the terms or concepts used. A new project at NCBI that involves linking books to PubMed might address this shortfall.

The idea arose out of a consideration of the new approach on a genomic scale to biomedical research: techniques that put thousands of genes on one microchip mean that investigators have to transcend the boundaries of their traditional area of expertise and look at biomedical function outside their normal remit. The addition of background information not only assists these researchers but is also enriching for nonexperts and students.

Books will be hooked up to PubMed in such a way that hyperlinked phrases in the PubMed abstract lead to the most relevant sections of the book(s). Currently, there is just one book available in this way: the general molecular and cellular biology textbook

Molecular Biology of the Cell.[1] All the content accessible from the NCBI Web site is freely available as stand-alone sections, but it is not possible to browse the book in its entirety.

In the near future, we plan to add more books, which will include more medical and specialist texts. As our library of books grows, the coverage will expand to include

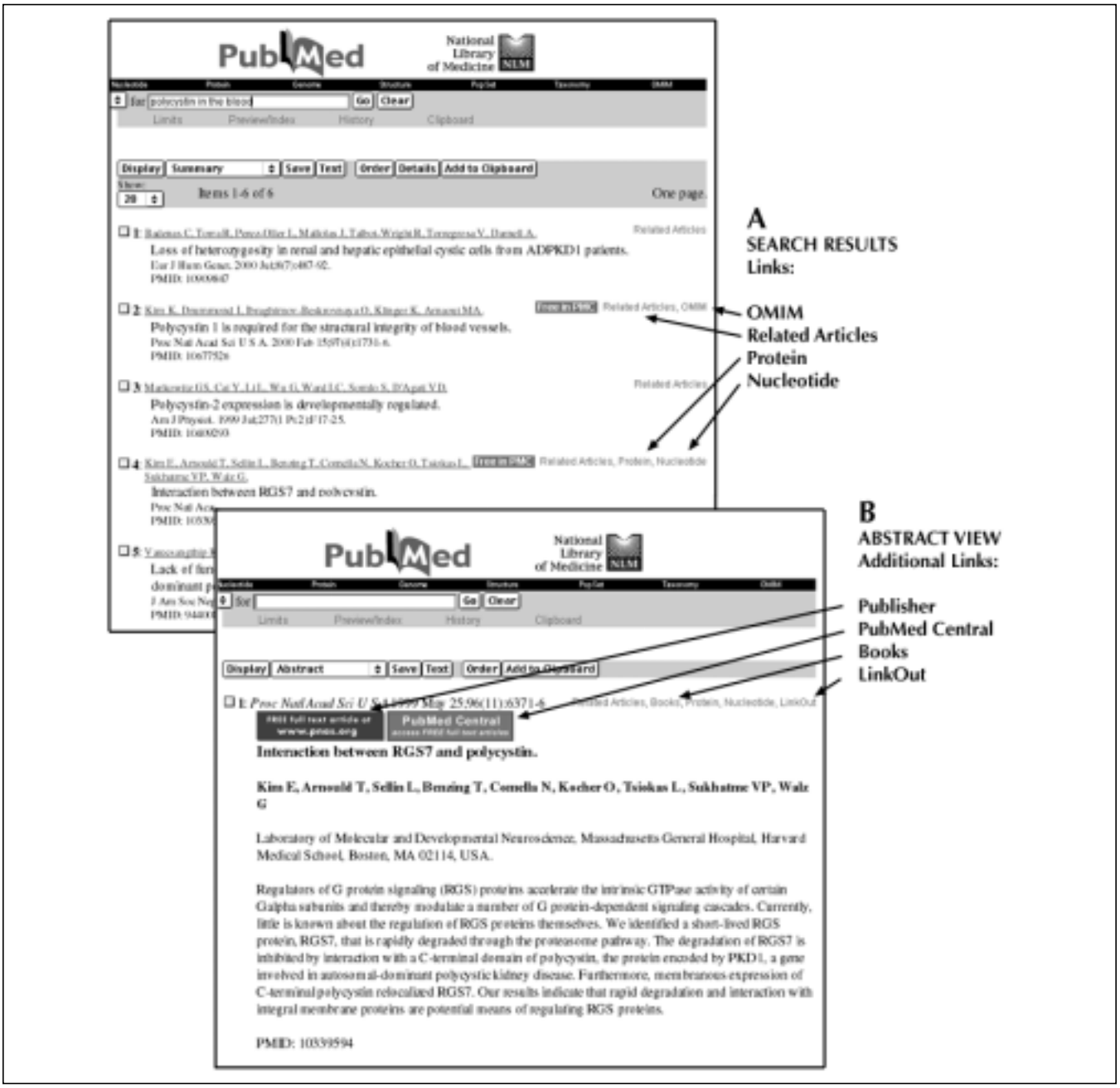

Fig. 1: Mutations in the polycystin 1 (PDK1) gene can cause autosomal dominant polycystic kidney disease (ADPKD). ADPKD is associated with life-threatening vascular abnormalities. A. A PubMed search for "polycystin in the blood" finds several articles, of which 2 have free full text available in PubMed Central. There are direct links to NCBI resources such as OMIM and protein and nucleotide sequences. The Related Articles link represents a precomputed list of articles that are similar to those found through the original search and provides an easy way to expand a PubMed search. B. The abstract view of the PubMed record reveals more links. In this case, the full text of the article is freely available both from the publisher's Web site and from PubMed Central. Clicking on the Books link will display the same abstract, except that some of the terms will be hyperlinked to sections of the book(s) that are significant for that phrase. The LinkOut summary page displays a list of resources that are related to the PubMed article currently being viewed. These will usually include hyperlinks to the publisher of the article, plus other resources, such as MEDLINEplus or NCBI's Genes and disease Web site.

more concepts and to provide a choice of level and approach to research topics. Furthermore, the book sections could be linked back to other resources at NCBI, such as sequences or structures, where appropriate.

No discussion of books linked to PubMed would be complete without mentioning Online Mendelian Inheritance in Man (OMIM) (www.ncbi.nlm.nih.gov/Omim/). Based on the book *Mendelian Inheritance in Man*[2] edited by Dr. Victor McKusick and his colleagues at Johns Hopkins University in Baltimore, and elsewhere, OMIM is a searchable database of information on the molecular genetics of disorders that have a hereditable factor. The information is presented as cross-linked summaries of basic and clinical information, which are constantly updated and extensively referenced. Links to OMIM are created in the same way as they are for nucleotides and proteins (i.e., by the reciprocal linking of PubMed articles cited in OMIM records), rather than by the phrase-matching method used for *Molecular Biology of the Cell*. However, the end result is the same: background information is provided.

LinkOut

If you are looking to expand a search to resources outside those provided at the NCBI Web site, you might try using LinkOut (www.ncbi.nlm.nih.gov/entrez/journals/loftext_noprov.html) (Fig. 1). LinkOut allows other information providers to associate themselves with a particular abstract (or collection of abstracts) and construct links to related content. For example, the researchers who wrote an article could make a link to their home page, or a support organization for people with a particular disorder could create a link back to their site from all the abstracts about that disorder. Currently, this service is being used by a small number of pioneers, including a few libraries, organism-specific databases and MEDLINEplus (www.medlineplus.gov/), which is a database of up-to-date medical information produced by the National Library of Medicine. The technology used for LinkOut also enables publishers of articles in PubMed to have a link from the abstract back to their Web site.

Individual PubMed users can specify the links they would like to see displayed by registering their preferences in the NCBI Cubby (available on all PubMed pages). As the use of LinkOut increases, links will be divided into broad categories (e.g., publisher, database), so users can select general classes as well as specific links.

Full text of journal articles and PubMed Central

An extremely valuable addition to PubMed is the ability to link the abstract of a paper directly to its full text. Currently, for this to happen, the journal publisher has to supply PubMed with the the appropriate information about the full-text article, including the URL, and, in most cases, the researcher or an institution has to subscribe to the journal.

Over the next few years, the extent of this service should increase for 2 main reasons. First, libraries that use LinkOut will enable their members to use links that seamlessly connect abstracts to the full text of electronic journals that they hold within their collection. Second, the growing trend of offering free access to the full text of life science research articles means that an increasing number of abstracts will have such links. Many journals published by scientific societies now offer free back issues; for example, there are now about 250 000 free full-text articles from back issues available on-line through HighWire Press. However, knowing which journals have free content, and to which issues the free content rule applies, is not always obvious when using either the HighWire search engine or PubMed.

One of the catalysts of this trend has been PubMed Central — the National Institutes of Health initiative to put the full text of life science journals on-line in a barrier-free manner. In these early days of the project, PubMed Central has signed up a few dozen journals that will make their content freely available through the PubMed Central site. Any article that is available in PubMed Central is clearly marked on the results of a PubMed search, and clicking on the links will take you via the PubMed abstract to the full-text article (Fig. 1).

Making the full text of life science research articles freely available through PubMed is a very natural extension of the NLM's services. In this sense, PubMed acts as a portal and gives cohesion to a growing mass of information.

Conclusion

The trend over the past 10 years in the life sciences has been the accumulation of research information on a vast scale. One coming challenge will be to make sense of the data through documented experiment and annotation. A significant part of making the best use of this wealth is to organize and link the information together in such a way that discovery is facilitated. Ultimately, we should be attempting to make the path from basic research findings to clinical applications as smooth as possible.

Competing interests: None declared.

Contributors: Dr. McEntyre researched, wrote and edited the article. Dr. Lipman provided direction and edited the article.

Acknowledgement: We thank Ed Sequeira for his helpful comments on the manuscript.

References

1. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. *Molecular biology of the cell.* New York: Garland Publishing; 1989.
2. McKusick VA. *Mendelian inheritance in man. Catalogs of human genes and genetic disorders.* 12th ed. Baltimore (MD): Johns Hopkins University Press; 1998.

Reprint requests to: *Dr. Johanna McEntyre, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 45, Office 5C, 45 Center Dr., Bethesda MD 20892, USA; fax 301 480-0109; mcentyre@ncbi.nlm.nih.gov*