nature webdebates

Blurring the boundaries between scientific 'papers' and biological databases

Mark Gerstein
Molecular Biophysics & Biochemistry Department and Computer Science Department, Yale University, Connecticut, USA
Jochen Junker
Molecular Biophysics & Biochemistry Department

Mark Gerstein

Jochen Junker

Until now, it has largely been overlooked that there is little difference between retrieving an article from an on-line journal and downloading an entry from a large genome or protein database. In the future the distinction between database and journal will progressively blur as people become accustomed to flitting back and forth between deposited data, visualization tools and linked sections of on-line publications. Exactly how will this interaction be structured? It will be relatively uncommon for future scientists to surf the table-of-contents pages of on-line journals in the way they would have browsed paper publications in the library, or as most do now on the Web. The bulk of electronic access to literature will in future be through data-integration services.

We expect that complex scientific data sets will become tightly integrated and entwined with the literature, with the interface to publications moving away from simple keyword search models to one reflecting the structure of biological information itself. People will increasingly browse databases arranged around chromosomal location, biochemical pathways and structural interactions that are linked to relevant articles, or parts of articles such as individual paragraphs, tables or figures. One might 'fly through' a large three-dimensional molecular structure, such as the ribosome, where various surface patches would be linked to publications describing associated chemical binding studies (a prototype of such a system, Riboweb, is under development at Stanford).

The PubMed and Entrez services of US National Center for Biotechnology Information (NCBI), provide a somewhat less dramatic but currently much more usable example of this data integration. They allow 'one-stop shopping' of much of the biological literature, using simple searches based on matching words in the title or abstract. Entries in the system are manually tagged with controlled vocabulary keywords (called MeSH terms) and automatically linked to related articles based on overall word frequencies. Most importantly, the articles are also linked to related data sets, such as gene sequences or protein structures. A search on data can therefore be complemented with links to relevant papers, and *vice versa*.

In the future, this integrated approach will be developed to a much greater extent to allow sophisticated full-text queries on the entire body of biological literature. One objective would be to allow users to retrieve articles based on their overall popularity, measured in terms of download frequency or citations, and what they, in turn, cite. Going further, one can envision navigating through a customized 'meta-journal' with articles retrieved daily on the basis of a specific reader's interests. Some of these features are currently available, in limited form, in resources such as the Science Citation Index, Beilstein's list of syntheses, Chemical Abstracts and the Amazon.com on-line bookstore. Amazon, in particular, artfully combines an individual's selections with those of people overall by constantly prompting with "readers who liked this book also liked..."

Like their human counterparts, computer programs or agents will increasingly cross the divide between data and literature, doing similarity computations and large-scale surveys on articles, as they now do on databases. Because of the ever-increasing scale of the literature, automatic computer analysis of articles will become necessary in certain contexts. For instance, even though it will not be possible to read articles about each of the ~30,000 genes in the human genome, one will still want to get an overview of the publications on the genome.

How might this blurring of journals and databases practically affect future information resources? Large central repositories, such as the Protein Data Bank (PDB) or SwissProt, may develop into integrated information resources, encompassing both standardized tabular data and free-text articles. In particular, as there are considerably fewer protein structures or complete genome sequences than articles about them, the database report on a structure or genome makes a convenient place to 'link' the many articles and 'boutique' databases annotating and referring to it. Because of this and from the general benefits of integrated searching, large central databases may become 'portals' into biology - in the same sense that Yahoo, Google and other search sites have accomplished this function for the Web as a whole.

We may see database sites organized increasingly like journals and database curators increasingly perform functions similar to those of journal editors and reviewers, which will give the curators quite influential roles. Thought should be given now to mechanisms for having some formal oversight over the whole curatorial process, as one now sees in the traditional peer-review and editing process for paper journals.

Conversely, at the other end of the spectrum from this extreme centralization, one might imagine a more decentralized approach to information architecture. Here biological information would be distributed more broadly over the whole Web into a looser structure of federal databases. This better reflects the general spirit of the Internet and has the important technical advantage of being more readily scaleable to larger amounts of data. In particular, it reduces the influence that individual database curators have over the presentation of information for everyone. However, it requires serious thought into developing interoperability between various resources.

The merging of databases and literature indicates an approach to one of the major scientific challenges of the next decade: how to annotate the human genome. Various proposals have been put forward, mirroring the issues discussed for information architecture above - for instance, a small group gathering at a centralized annotation jamboree, or a distributed, Web-based system that would allow anyone to contribute annotations with a 'smart browser' that would merge all the efforts. However, a better solution might be to extend the capabilities of the biological science literature.

Note that although the current journal system is decentralized, most research articles adhere to common standards that make them ideal for annotation. In particular: each article associates a bit of information with a distinct time and place and with specific, responsible parties; attentive scholarly referencing and footnoting provide a precise way to connect bits of annotation and allow for continuous 'updates'; peer review and editing provide a proven quality-control mechanism; publication is an established indicator of scientific productivity, providing an incentive for literature annotation, whereas database submissions are often regarded as a thankless chore.

The main drawback of current journal article formats is that they are not very 'computer-parsable' or

suitable for bulk annotation of thousands of genes. However, a 'literature annotation standard' that could readily be interpreted by computers could be established by adding sections of highly structured text to each article and linking subparts of an article to relevant database identifiers.

More specifically, the data tables associated with an article could be considerably expanded. Certain tables could be marked as 'bulk' and then downloaded separately in a standardized format. A number of tables in a paper could be linked together through common identifiers into a static relational database that would encapsulate molecular coordinates, annotations on many genes (for example, lists of all the membrane proteins in the worm genome) and the results of functional genomics experiments. Moreover, standardized identifiers could allow tables in one paper to refer precisely to those in another, allowing economical large-scale annotation without replicating information.

This joining of tables obviously highlights the importance of developing universal systems of identifiers for 'biological entities' such as genes and molecules. This is as much a social issue as a scientific one, reflecting the many different reasons things get named as they do. However, many naming problems in biology, such as for genes or species, have a restricted enough scope that complete, self-consistent solutions are possible. This sort of problem is routinely solved in the commercial world where identifiers for tens or hundreds of thousands of dynamically changing objects are needed - for example, the ticker symbol given to each of the ~5,000 stocks on the New York Stock Exchange and the standard code given each of the ~50,000 products in a large grocery store.

Developing an expressive, controlled vocabulary that describes the highly variable properties of the underlying objects, for example, the function of genes, is more problematical because there are substantial scientific issues of how to describe precisely open-ended concepts such as gene function. Progress has been made in building controlled vocabularies (for example, GO, MIPS, Enzyme, GenProtEC for gene function, scop and cath for protein structure, and the National Library of Medicine's UMLS for disease states and clinical indications) but still much work remains to be done. (For further information on functional and structural classification including pointers to the many on-line resources, see 'Parts list'.) One issue that remains is whether objects should be classified in a strict hierarchy, where they can only belong to a single major category (for example, scop) or a more flexible directed acyclic graph (DAG), where they can belong to multiple major categories (for example, GO).

From the perspective of the data tables in a paper, the associated article text would represent high-quality documentation of such things as the meaning of each column and the allowed values in each field - all important meta-data, in the jargon of the database world. All of these tables and associated text would be fixed in content at the time of publication, avoiding the inconsistencies that arise from referencing dynamically changing entities. The data would be updated in the normal way through publishing further papers or perhaps simple addenda to existing ones.

In addition to functioning as metadata, the readable text of on-line articles may differ in other fundamental ways in the future. Because of the much lower distribution costs of the Internet, the length of on-line articles will be less restricted, even in the most selective journals. Hopefully, this will encourage a more thorough and explanatory writing style. Counterbalancing this tendency to verbosity is hypertext, which will allow authors to link their articles to supplementary material within bulk data tables, on their own websites or in external databases. This will enable them to condense the main text, making it less technical and moving details to linked sections. Perhaps different 'views' of an article could be requested from a journal: short, full, extended and so forth.

The complete transition envisioned here from paper to integrated electronic journals will take a while. Some of the reasons are technological - for example, waiting for faster networks and 'books-on-demand' printers. Others are economic: in the long term the shift to electronic journals offers great efficiencies, potentially saving money for scientists and the journal users. However, in the short term, it may redistribute funds in the delicate world of academic publishing, causing discomfort. It is thus only being marginally embraced by the commercial journals, which are still trying to figure out how to make money on-line. (The financial aspects of on-line publishing are discussed in detail in many other commentaries in this series.)

With whatever financing scenario adopted, the transition to integrated on-line literature will take place. As a concrete demonstration of this reality it is worth noting that most physics disciplines, with the prominent exception of biophysics, have already moved from traditional print journals to a hybrid system, where electronic versions of papers are made freely available from a single on-line preprint server at Los Alamos as well as from archival journals. Integration between literature and data is also proceeding apace in astronomy, with the creation of global 'virtual observatories'.

Paper will continue to have a place for a while, as printouts will be read on the sofa or over coffee. But journals as we know them will no longer be delivered to our doorsteps, except for one or two that we buy for a good read. And eventually 'paper' itself may be 'electronic' with the advent of working 'e-ink' displays.