Editorial

THE NEW BIOLOGICAL LITERATURE

The progress of molecular biology in elucidating the molecular mechanisms of life has been astounding. In the last twenty years, an increasing amount of this progress has depended on the mapping of information from known molecules to those that have been newly sequenced. In many cases, this mapping of information obviates years or decades of laboratory experiments by providing a set of detailed, and relatively easily tested, hypotheses about the function and structure of the newly sequenced molecules.

As this information mapping process has grown in importance, and as the basic sequence and structural databases have grown in size and complexity, the data and annotation in the databases has become an increasingly important source of fundamental information about the molecules they describe. Gradually the sequence and structure databases have become important, or even primary, sources of information about macromolecules. It is common, for example, for a student to obtain his or her first ideas about the structure and function of a macromolecule based on a keyword search of one of the sequence databases, and to use the literature references in the databases as his or her entry point to the conventional literature.

Sequence and structure databases have thus assumed something of the role of a review article. They provide a brief synopsis of molecular structure and function, and they provide pointers to the relevant fundamental literature. But they are review articles with a difference. The powerful indexing and keyword searching facilities provided by database management systems make the publicly available databases a much more scalable way of dealing with the scientific literature than conventional journals coupled with commercial and public indexing systems. As we approach the completion of genomic sequences of several higher metazoans, the online sequence databases provide our only hope of organizing and using this information, guaranteeing the continued importance of these resources as fundamental literature. In the future, the online resources have an even greater promise — direct integration of the data with online analytical tools and simulations.

The power and comprehensiveness of the online data resource for structural and sequence information hides an insidious dark side. In spite of concerted efforts, few, if any, of the major databases have become true community curated resources. Rather than being produced by known experts in the field, and subjected to the peer-review process, the annotations in databases are typically provided by the authors alone, or by annotators with little expertise in the field. Even databases that have established editorial or curatorial boards seldom provide any kind of peer review with the result that annotations may represent viewpoints not shared by the entire field. The reasons for this lack of community involvement are, I think, primarily sociological. When a research scientist invests effort in developing complete annotation for specific entries in a database, the payoff is almost non-existent. No journal articles, no citations that can be included in grant proposals, no recognition acceptable to tenure review committees are generated. It is hardly surprising that few laboratory experts are interested in investing the necessary time.

A new scientific literature, in the form of sequence and structural databases, already exists. It is clear that this literature differs widely from the conventional literature in terms of content, peer review, and ultimately quality. It is also clear that high quality annotation of this data is desirable, and perhaps essential, to the continuation of rapid progress in our understanding of the molecular basis of life.

One course of action is to devote additional resources to the public databases, developing systems whereby community annotations will be perceived and cited in the same way as articles in the conventional literature. Progress along this path has been slow, perhaps because the inclusion of such information in publicly funded databases is perceived as competing with conventional commercial journals. An alternative, currently being followed de facto, is to rely on references to the electronic versions of conventional journals to provide expert level annotation. Since electronic journals may not be freely available to non-subscribers, this kind of annotation is not equally available to all users. In particular, it is unavailable to researchers at poorer institutions and, especially, to those in poorer countries.

Will mechanisms of citation and curation be worked out for the existing databases that allow free access to high quality annotation? Or will there be a tight coupling between databases and the increasingly electronic conventional literature — leading to a privatization of the annotations on sequence and structural data? Or is there some middle ground that balances the demands of commercial publishers and public access to information?

The scientific community needs to make itself heard, for if we do not, commercial and political forces will determine the future of the fundamental data resources.

Michael Gribskov
San Diego Supercomputer Center
University of California
San Diego, USA