nature

webdebates

GenBank - a model community resource?

Jo McEntyre and David J. Lipman

**National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Building 38A, 8600 Rockville Pike
Bethesda, MD 20894, USA**

GenBank has become a household name among biologists. They all benefit from having free access to the 16 billion base pairs of primary DNA sequence and the related molecular information that has been submitted to this shared resource by the international scientific community. The information either goes directly to GenBank or is submitted via its counterparts in Europe -- the European Bioinformatics Institute in Cambridge (EBI) -- and Japan -- the DNA Data Bank of Japan (DDJB). GenBank demonstrates that, even in the fiercely competitive world of science, researchers recognize that contributing to large, shared data sets ultimately benefits everyone. The shared resource that is created is an indispensable tool that is greater than the sum of its parts.

Scientists have shown a willingness to place data in a community archive for the common good, knowing that it can be freely used by anyone. Moreover, all leading journals have adopted a policy that requires sequences to be deposited in the public databases, and the corresponding access numbers to be cited in published articles. All publicly funded laboratories now consider it *de rigueur* to contribute sequence data to GenBank within 24 hours of its generation, even if there is no accompanying research paper.

As a result, GenBank now houses sequence from over 900 complete genomes, including the draft human genome, and some 95,000 species. This is a real treasure-trove, and considering it as a whole is key to the effective handling of information. New sequence data can be deposited directly to EBI, and DDJB or GenBank. The three databases are synchronized daily so data submitted to any one of them is available in all three within 24 hours. Each database has its own format for submitting data, but agreed conversion protocols allow information to be swapped seamlessly.

Entries within GenBank have a common language, so we can easily identify and delineate the many fields within an individual database such as comments, journal citations and feature annotations (i.e. coding regions). Imposing basic rules about data structure ensures that we can tell whether "AC" or "CA" are an author's initials, or part of a nucleotide sequence.

GenBank uses a data format known as [Abstract Syntax Notation One](#) (ASN.1). This is a structured language similar in many ways to Extensible Markup Language (XML), which has become the language of choice for structuring Web data and is used in many publishing ventures. GenBank records can now be downloaded in either XML or ASN.1.

For the scientist, the GenBank approach to data management offers several advantages -- the delay in accessing the latest information is minimal, and the data is free; anyone may use it as they please. Our strict formatting rules enable search software to be written and facilitate re-use of the data.

BLAST (Basic Local Alignment Search Tool) software, used to find similarities between sequences, is a good example of the success of this approach. We believe that defined data structures are needed to allow software to work effectively. The beauty of BLAST is that it enables discovery by computing on the core information, sequence data. Because it searches the sequence directly, BLAST does not need to know such basic surrogates as the name of the gene, or its synonyms, to identify matches elsewhere. Clues as to the function of an unknown stretch of sequence can be grasped within seconds by matching it to other, better-characterized sequences in the database.

BLAST used in conjunction with more structured information adds further sophistication. For example, as there is a field in a GenBank record for recording DNA features like gene-coding regions, this means that BLAST can be "instructed" to search only the <5% of the human genome that codes for genes, rather than sift through all 3 billion bases. The latest versions of BLAST also allow the user to combine a sequence search with a regular text Boolean query of GenBank, allowing such comparisons as "my sequence" vs. all those from marsupials or even "my sequence" vs. all those by "Smith".

On a larger scale, commercial companies can freely download and use GenBank locally to develop new products in a secure environment. NCBI, Celera Genomics and the University of California at Santa Cruz each created an assembly of the human genome, based either partly or entirely on publicly available data.

Perhaps the most oft-cited criticism of GenBank as a primary sequence database is that "there's a lot of rubbish in there". Of course, any open system that does not practice some rigorous form of peer review is bound to have more errors and less desirable elements present. However, several quality-control mechanisms are integrated into the system. The daily synchronization of data among the three collaborating sites requires content and syntax to be consistent, and re-use of the data by others leads to the discovery of technical glitches. GenBank also encourages users and contributors to send feedback and update records, to remove vector contamination, for example. It is true, though, that the proportion of records corrected is small.

The availability of this archive of primary sequence information gives others the opportunity to provide different "added value" views of this basic information. More refined layers are created through curation, further organization and analysis. Several projects, both free and subscription-based, provide such a service. These include many of the organism-specific databases, SwissProt, Genecards, Pfam and the Reference Sequence project at NCBI.

Perhaps there is no need for all this data to be in one place. It is possible to simply post the information you wish to share on your own website, and have a search engine find it. This obviously does not create a stable archive, but it would work for reading records one-by-one, so long as they existed on the individual's website. However, if one does wish to build a stable archive and create sophisticated software search tools, then it is essential to have a repository with a consistent data structure as a basis. If all sequence data were distributed on the websites of the individuals who generated it, then BLAST's usefulness would be compromised. GenBank is a centralized repository in that it is available as a single unit in a uniform format. However, in terms of distribution it is quite the opposite. Not only can all the data also be found in the EMBL and DDBJ databases, but also on countless other non-profit and commercial sites, in all sorts of different guises. A uniform format means that anyone who wishes to convert the information to their preferred tagging system so they can use and display it as they wish needs to write just one piece of conversion software to do so.

It is also tempting to consider where database publishing and traditional journal publishing intersect. Most significantly, we happen to be at an historical juncture where information-delivery technologies are merging. The tagging technologies used in online publishing and databases are very similar, and this will undoubtedly allow us to forge better links between molecular information and its detailed analysis in research papers. Moreover, some laboratories are now generating huge data sets, such as microarray data, that cannot be accommodated in traditional-style research articles and this will necessitate a further blurring of the boundaries between the written word and molecular information.

No one would deny that GenBank and its collaborating databases have proved to be fantastically useful resources, perhaps in ways that few anticipated at their inception over a decade ago. The lessons learnt from the GenBank model of data management -- that a collective archive can contribute to everybody's science -- may be useful ones to consider in these days of Internet publishing. Only time will tell.