LEGIBILITY NOTICE

A major purpose of the Technical Information Center is to provide the broadest dissemination possible of information contained in DOE's Research and Development Reports to business, industry, the academic community, and federal, state and local governments.

Although a small portion of this report is not reproducible, it is being made available to expedite the availability of information on the research discussed herein.

JAN 1 1988

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W 7405-ENG-36

LA-UR--87-4022

DE88 004315

TITLE THE GENBANK DATABASE AND THE FLOW OF SEQUENCE DATA
FOR THE HUMAN GENOME

AUTHOR(S) Christian Burks

MASTER

SUBMITTED TO Proceedings of the Brookhaven National Laboratory
Symposium on the Human Genome (Long Island, NY, Sept.
14-16, 1987).

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes.

Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

**THE GENBANK DATABASE AND THE FLOW
OF SEQUENCE DATA FOR THE HUMAN GENOME***

by

Christian Burks

Theoretical Biology and Biophysics (T-10)

Theoretical Division

Los Alamos National Laboratory

Los Alamos, NM 87545 U.S.A.

TEX Reformatter 9.5

* Work supported by the U.S. Department of Energy

THE GENBANK DATABASE AND THE FLOW OF SEQUENCE DATA FOR THE HUMAN GENOME

by Christian Burks

In the 1960's and 1970's DNA and RNA first began to be sequenced and, for a time, the amount of available data was quite small. Within one's domain of interest it was possible to retain the totality of relevant data in one's head or, if not, to keep track in a single notebook or file folder of the sequences and the patterns within those sequences. Also, because it took so much time to determine the sequence data, the information usually concerned a very limited region of the DNA that, by the time the sequence itself had become available, it had been very roughly? characterized experimentally. So the sequence data often was simply a confirmation, or perhaps a stitching together, of a great wealth of information already known about a particular region.

With the re?dutionized sequencing technology that appeared in the mid to late 1970's, very quickly it became impossible to keep track of interesting data either in one's head or in one's laboratory notebook. In addition, we are now approaching the point where it is no longer true that experimental approaches and techniques are able to keep up with the appearances of new sequence data. When people began to develop nucleotide sequence databases, they were thinking primarily of the problems associated with data storage and memory. Increasingly we must structure the data and develop access tools such that we can find and characterize the regions that our research interests dictate.

In the late 1970's the National Science Foundation (NSF) sponsored a workshop at Rockefeller University to discuss the development of DNA sequence databases; a number of agencies, including the National Institutes of Health (NIH) and the European Molecular Biology Laboratory (EMBL), sponsored subsequent meetings to discuss this. The outcome was the establishment of EMBL Data Library in Heidelberg and the GenBank database in the United States.¹ GenBank is funded through a contract with the National Institute of General Medical Sciences (NIGMS) with a number of co-sponsors, who provide various levels of funding for the project. The

¹ "GenBank" is a registered trademark of the U. S. Department of Health and Human Services.

project's charter is to provide a computer database of all published DNA and RNA sequences longer than 50 nucleotides as well as related bibliographic and biological information. The data are available in magnetic tape, floppy disk, on-line, and book form. Data distribution is handled by Intelligenetics, Inc. (IG)² and Los Alamos National Laboratory (LANL)³ who collect and organize the data. We collaborate in our data collection efforts with the EMBL Data Library. They cover the journals published in Europe while we cover the journals published in North America; we have also recently begun collaborating with the DNA Data Bank of Japan at the National Genetics Institute in Mishima.

What data are in the database? There are the nucleotide sequences, "a", "c", "g", and "t", representing the four monomeric constituents of DNA and RNA. If our only task were to incorporate strings of these monomers into the database, there would not be the great concern for whether or not we will be able to keep up with the inflow of data. The majority of our efforts go into developing the annotation that goes along with the sequences entered into the data library. This includes information about the bibliographic context of the sequences (i.e., where was it published, what was the citation for the article, the title of the article, and who were the authors. We include information about the physical context of the sequence, what organism it was taken from and which tissue; we enter any information that is available to allow one to pin-point how this sequence relates to other known sequences. We include all the mapping information, where on the chromosome the sequence occurs, the name of the gene, and what other genes it is close to. Finally, there is information about the functional context of the sequences. Many of the sequences in GenBank correspond to protein-coding regions and regions that code for other products, and information is given naming these regions. For instance, if someone is interested in hemoglobin sequences, then these can be pulled out of the database, together with data indicating the coding regions spanned, as one can pull out the DNA itself and translate it to get protein-coding regions.

A sample entry from the database is shown in Fig. 2. At the bottom is the sequence itself, which was published in the two articles listed. The entry has the citation or bibliographic information for the sequence, and the functional context,

² Contact: GenBank; Intelligenetics, Inc.; 700 El Camino Real East; Mountain View, CA 94040 (telephone 415-962-7364, E-mail genbank@blonet.20.arpa)

³ Contact: GenBank; T-10, MS L-710; Los Alamos National Laboratory; Los Alamos, NM 87545 (telephone 505-666-2177, E-mail genbank@lanl.gov)

showing that it codes for actin protein. Under the heading source and organism is information about the taxonomy, the name of the species, and the tissue used. This entire entity is called an entry. GenBank now has citations to about 9000 literature articles in the literature. There are about 14,000 entries in the database, covering about 15 million nucleotides. This is impressive, but if it is measured in the context of the numbers that will become available, it is obvious that this is quite small.

Once the data is in the database, how does one get it out? There's a two-part answer to that. One is that the data is stored and organized in such a way that it becomes relatively easy to extract subsets of the data, either simply by matching an input sequence and pulling all those examples out of the data library or by retrieving sequences and other information based on the annotation tags that we attach to them. Thus, one might want all sequences that code for hemoglobin. The information is labelled and structured in ways that are amenable to automatic extraction, using software. To some extent visual examination can be used first, and then the search can be made automatically using computer programs that go through and find the appropriate "tags" in response to the particular query.

Figure 3 illustrates the various ways that one can access the database now. GenBank supplies an on-line system. There are a number of regional computer resources serving the molecular biology community that have on-line services; BIONET is one example. Then there are a number of distributed software packages, such as Intelligenetics and DNA Star, and a number of pharmaceutical companies who are introducing packages with a set of software that often includes the database. Finally, you can write to GenBank and get your own copy of the database on magnetic tape and write your own software to do tailor-made searches of the database.

What are the uses of the database? The first two, probably a great majority of the use in the last five years, are described by DeLisi and Gilbert (this volume). The third category is going to be increasingly important. People have been using the database just as you would use a library. They see an article in a journal and want to get the sequence data from that article; therefore they use the citation to pull the sequence out of GenBank. One step up from that is where people might have established a sequence and want to know if there are other sequences like that in the data library. So that information is used to query GenBank and pull out related sequences. It is akin to treating the database as something of an unknown domain where, with the biological information, the physical and functional

context information that is provided, one can look for relationships between various sequences in the database that have not been experimentally determined. One can begin to delve for and mine out of the data such relationships, looking at GenBank as an entity to itself. It might be something as simple as a consensus sequence that occurs consistently in front of a particular family of genes. It might be something much more complex that ties in the name of the organism and the function or the name of the protein along with the particular sequences that are tagged by those items.

Finally, there is the human genome, and there are on the order of 130 sequences that one might call complete chromosomes or complete genomes. There is a big qualification here. The complete chromosomes and genomes that have been sequenced and are present in GenBank are all extremely parasitic entities that rely for their replication cycle and their function on the host genome in which they reside. Therefore, if one goal of sequencing a complete genome is to get a picture of the information set necessary to run an organism, then these are not particularly good examples because they rely so much on the host organisms. On the other hand, in their limited way they do represent a complete data set. For those organisms we really care about, we know very little; for example, *E. coli*. If one does a rough calculation, in GenBank there is about 20% of the complete genome, and this is the sum of data from different strains that have been determined; it corresponds to somewhere between 400 and 500 genes. For human sequences there is something less than one tenth of one percent of the complete genome in the data line frame.

We shall be dealing with a great deal more sequence data in the immediate and distant future; Fig. 4 displays several different scenarios for dealing with this. The top graph is a picture of the total accumulated amount of DNA; the dots represent the past, and the solid lines represent the future. The model 4 is the goal in the year 2005 of having the complete human genome. This is expressed in various "guestimates" of how quickly people will be sequencing DNA. Model 1 assumes that the rate will be exactly what it was in the last year; clearly it is a far lower-than-realistic estimate. The middle graph represents the rate of DNA being sequenced each year as a function of the year. Then the final graph shows the change in that rate or the acceleration for these various models. I think that a reasonable estimate involves Model 3, but leaving aside predictions, it is clear that in the future we will be dealing with a lot more data and a much greater accumulation rate.

Today the computer capacity available for storage and manipulation of data is sufficient to accommodate a 10-fold increase. This will not be a limiting factor. The problem is how quickly are the data being accumulated and how quickly we can get the data into GenBank. An answer that is untenable is to add staff in proportion to the amount of data that are appearing. Clearly, as shown in the last figure, that would rapidly become a dominating cost in any aspect of molecular biological research. Instead, we will try to shift to a model for how we conduct a data bank enterprise, which moves away from the notion of the staff at GenBank doing a lot of the entry, characterization and description of data toward a model where that responsibility and effort will lie with the individual scientists who are developing these sequences. Our task will be to provide the tools and assistance that will enable them to stream the data into the data library.

Data has to get from the original experiment into the data library and then back to the user community. Traditionally, the scientist's involvement has ended at the experiment, and the data bank has been responsible for incorporating it into the database. We have had some success in having the authors themselves take a great deal of the burden of this transition. Most of the journals that publish sequence data now request authors to submit their data to the data library. It is our hope that journals will require authors to do that. Very soon NAR will not consider a paper for publication unless the data already have been deposited, in this case in the EMBL Data Library. *Acta Crystallographica* has a similar policy for crystal coordinates.

Figure 5 shows the idealized goals that bring about this shift, enabling us to update the database to keep up with the increased flow of new information. In addition to increasing the involvement of the researchers, we shall move toward eliminating the hard copy step, whether this is publications or reprints/preprints. We want to have data directly from individual laboratories that maintain their data in computer format; this can be streamed electronically into the data library. At present, a substantial amount of data arrives either through electronic mail or magnetic tapes and floppy disks. Also, we shall move to continuous automatic updating of the data library, thus eliminating some of the time lags between data being deposited and being available to the public. Over the last few years several scientists have worked with us, bringing their expertise to bear on the data entries; we then effected any changes in the data library. We would like to build up a cadre of people who are expert in particular areas of genetics and also expert in our complete

database system so that they can work on some of these entries themselves. We also hope to build up the degree to which there is automatic cross-referencing between the databases and hope to tie together the many different data sets to build up a framework so that one can readily jump from data set to data set. An example in the context of GenBank would be that a fairly complete effort was made to include the official human gene names for all the human sequences in the database. That allows one to jump from our database into the human gene mapping library database in New Haven using the human gene names to link those sequences—thus connecting genes that have been sequenced with those genes that have been mapped and are recorded in the human gene mapping library.

We will try to tie together and promulgate standards for the many different software developers that provide access to the data library for their clients. There has been a blossoming of different systems and different standards on what one gets out of GenBank. This has allowed for a lot of software and interesting systems, but has been bad in the sense that not everyone developing software or distributing the data has a good picture of what information is in GenBank nor how it is structured so that it can easily be pulled out.

To close, our effort includes the human genome information resource, which at this point is a research effort into the questions of how one is going to best represent mapping data in order to answer the questions that scientists will want to ask and how to tie the mapping data into sequence data and other biological data sets that will be crucial in getting the most reward out of querying the mapping data and the related sequence data.