GenBank: Implementation Model and Retrieval Techniques

Navreet Kaur

Computer Engineering Department, Punjabi University
University College of Engineering, Patiala

[1]navreet_9@yahoo.co.in

***Abstract*— The Human Genome Project (HGP) has generated an enormous amount of sequence data. The GenBank database is the main genome database holding the vast biological data and is the most popular database used by the scientists all over the world. With the genome information becoming richer and complex the underlying data model is coming into picture. Understanding the internal structure of the database is vital for extracting the hidden information in the sequences from the database for various research purposes. Also the understanding paves way for building more efficient and advanced data storage model. In this paper the data model and retrieval techniques of GenBank database are explained in length. The paper targets the Computer Scientists using the GenBank. It aims to provide an in depth explanation of the architecture and design of GenBank database.**

***Keywords*— bioinformatics, databases, relational model, genome, flat files**

I. INTRODUCTION

The development of high throughput advanced technologies and knowledge in the field of bioinformatics has led to the generation of massive amount of biological genome data. This data is useful for biologists/scientists all over the world for research purposes and new discoveries in the fields of genomics, healthcare, agriculture, proteomics, molecular phylogenies etc. The generated data can be useful for the mentioned research purposes only if it can be safely stored for later use. The enormity of the available data makes it difficult to store this available data even in the form of books leaving aside the scientists mind capturing all the data. The sequence information generated by the human genome research, initiated in 1988 if compiled in books, the data would run into 200 volumes of 1000 pages each and reading alone (ignoring understanding factor) would require 26 years working around the clock [1]. An intelligent computerized system which provides efficient, safe, long term storage with the guarantee of efficient, easy retrieval and modification techniques is needed to store the huge amount of generated data. The databases which are computerized record keeping systems very well fit into the above requirements of an Intelligent System. They provide efficient storage, data access and modification capabilities round the clock. The databases can be rightly termed to form the base of the pyramid of bioinformatics. Gauging the importance of databases for biological purposes a large number of biological databases have been developed since the inception of the field of bioinformatics. The count of databases has reached to 1512 as of now. These databases are heterogeneous in structure, placed at different locations, based on different data models and store varied types of biological data ranging from nucleotide sequences to protein 3D structures. GenBank is one of the main Genome database which is heavily used by researchers and scientists. It falls into the category of sequence databases which is a subset of genome databases, other being Map databases, model organism databases, bibliographic databases, databases of databases in the set. A good understanding of the GenBank data model and access techniques is useful in exploiting the full potential of stored sequence data for purposes of research like protein structure prediction, drug design etc.

II. GENBANK SEQUENCE DATABASE

GenBank is an open access repository of genetic sequences and is one of the main genome databases in the world. GenBank traces its history back to the year 1979. In the mid 1970's a large amount of DNA sequence data was being generated by biologists and required computerized storage of data. To meet the needs of storing the huge amount of growing data, the Los Alamos Sequence Library in 1979 was created by Walter Goad, which in 1982 became GenBank. In the same year the responsibility of managing GenBank was shifted to the newly developed NCBI (National Center for Biotechnology Information). Originally, GenBank was a repository for nucleotide sequence data, but it was expanded to include EST data, protein sequence data, 3-D protein structure, taxonomy, and links to the biological literature (MEDLINE) [3]. It provides links (hard and neighbouring) to other resources (literature, related sequences, etc.) and databases. GenBank consists of publicly available nucleotide sequences of almost 260,000 formally defined species [7]. GenBank is part of the International Nucleotide Sequence Database Collaboration, along with its two partners, the DNA Data Bank of Japan (DDBJ, Mishima, Japan) and the European Molecular Biology Laboratory (EMBL) nucleotide database from the European Bioinformatics Institute (EBI, Hinxton, UK). The three databases exchange information on daily basis. This is done so that all the three databases have the same updated information contents and available uniformly across the world. A full release of GenBank occurs

on a bimonthly schedule with incremental (and non incremental) daily updates available by anonymous FTP.

.

III. GENBANK DATA MODEL

As genome information becomes richer and more complex, more of the real, underlying data model is appearing in common representations such as GenBank files. So it is important to understand the underlying data model [4]. GenBank database is not based on a single model but a combination of Flat Files, ASN.1 based structure and Relational model. GenBank initially stored the information in the form of Flat Files which were the elementary unit of information and was the legacy file format. But now Flat Files are used for the purpose of presenting output to the users. The modern Abstract Syntax notation (ASN.1) is now the generic data specification file format and is used as the standard format which specifies the structures of how information is actually stored in GenBank. It promotes data interoperability. The ASN.1 structured data is further internally stored in the form of relational tables (GenBank ID database). The GenBank sequence entries are stored and maintained in the Sybase relational database management system (RDBMS) [2]. Figure 1. shows the flow of information in genbank. The data can be inputted from various sources but before submission to genbank needs to be converted in the standard ASN.1 format. The ASN.1 structured data is than entered into the main ID database (relational model of genbank). The stored data can be retrieved by users through Entrez interface or via FTP from NCBI. The Output can be viewed in various formats such as FASTA, Flat Files, XML etc.

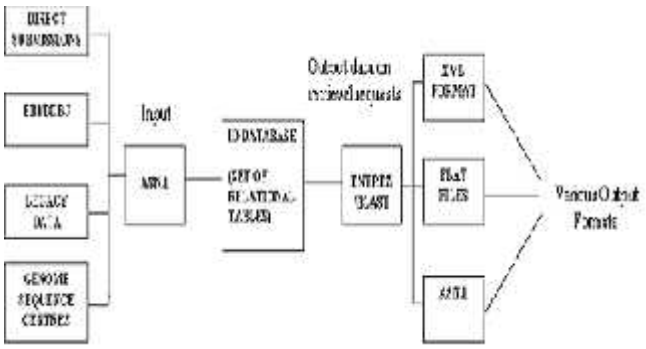


Fig. 1 Flow of information through genbank database

A. GenBank Flat Files

A Flat file based database is a database which consists of only one table containing all the information. The table consists of many columns (fields) and Rows (records). The fields can have different names and data types which impose restrictions on the type of data that can be entered. Amongst the fields one field can be chosen to serve as the Index field (Key Field) out of the various fields. The index field always has a unique value. It facilitates searching and easy retrieval. The fields (columns) in a single row combine to form a record. As in a table several rows may be present, so we have several records which have different values. The fields in the records can be separated by various delimiters like spaces, tabs, commas etc [6]. As mentioned above, the flat file format is now used for presenting information to users and for data exchange. Flat files are the de facto data interchange standard for Molecular Biology data. Many tools biologists are accustomed to (e.g. BLAST and FASTA) work only with flat files. As a consequence, most Molecular Biology databases provide their entire contents in one or more flat files [5]. The NCBI provides output from querying the GenBank in the form of Flat Files to the users. Gen Bank Flat Files (GBFF) are easy to read, and also there is no information loss after data import. The GBFF can be separated into three parts: the header, which contains the information (descriptors) that apply to the whole record; the features, which are the annotations on the record and the nucleotide sequence itself. All major nucleotide database flat files end with // on the last line of the record. [4]. Below a description of the fields of the Gen Bank File format is explained for the beta haemoglobin of Homo Sapiens record.

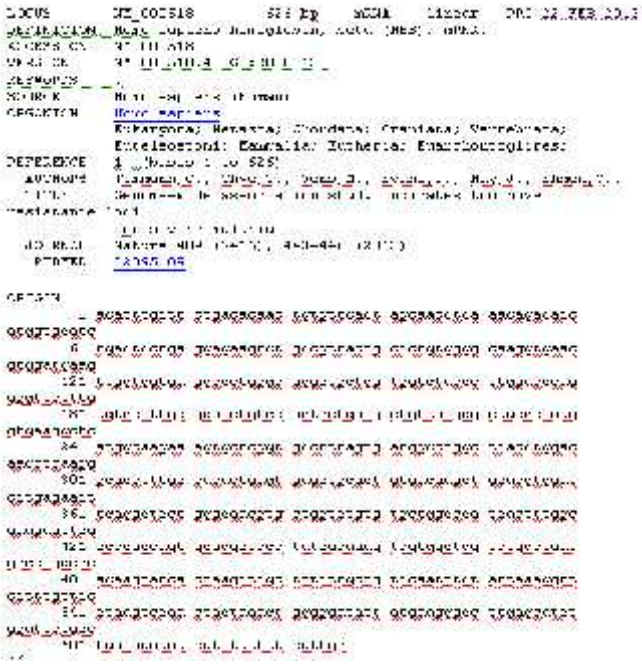


Fig. 2 An excerpt from GenBank for beta haemoglobin of Homo Sapiens.

1) Header- A Dissection

- Locus - It is the first field of the GBFF and consists of several Locus fields. It defines the subject of the record using various parameters like Locus Name, Date Modified etc.
  - Locus Name – The Locus Name in the example is NM_000518. The locus name was originally designed to help group entries with similar sequences: the first three characters usually designated the organism; the fourth and fifth characters were used to show other group designations, such as gene product; for segmented entries, the last character was one of a series of sequential integers.[8].

- Sequence Length – In the example the length is 626 base pairs It is number of nucleotide base pairs. The minimum length required for submission is 50 bp, although there might be some shorter records from past years [8]. There is no restriction on the maximum length, though a restriction of 350kb is placed on an individual genbank record.
- Molecule Type – In the example the molecule type is mRNA. It is used to tell the type of biological molecule. It can be a DNA or RNA molecule. The acceptable mol types are DNA, RNA, tRNA, rRNA, mRNA,and uRNA and are intended to represent the original biological molecule [4].
- GenBank Division – In the example the division is PRI. The GenBank database is divided into 18 divisions [8]. These divisions can be organism based PRI , ROD or technology based like EST, HTG.
- Modification Date – In the above example the modification date is 22 – FEB – 2013. It represents the date when the record has been last modified.
- Definition - It consists of a summary of the biology of the record. It appears in the FASTA and BLAST output. An agreement between the collaborative databases has been reached specifying a general syntax for specifying the description.
- Accession Number – It is an identifier used for uniquely identifying the sequence record. At this time, accession numbers exist in one of two formats: the ''1 + 5'' and ''2 + 6'' varieties, where 1 + 5 indicates one uppercase letter followed by five digits and 2 + 6 is two letters plus six digits. Most of the new records now entering the databases are of the latter variety.
- Version – It is used to represent the number of times a record is modified. It's given by ACCESSION.VERSION. Each time the record changes it is incremented by 1.
- GenInfo Identifier - It runs parallel to Version and is given a new number each time the record is modified.
- Keywords - Word or phrase describing the sequence. If no keywords are included in the entry, the field contains only a period [8].
- Source – It contains the common name used for the organism. It may have the scientific name used for the organism.
- Organism – It consists of the scientific or the formal name of the organism and the lineage.
- Reference - Publications by the authors of the sequence that discuss the data reported in the record [8]. It consists of the Authors, Title, Journal and the PubMed Identifier. The Reference has the beginning and end Base number. A sequence may have more than 1 reference.
- Comment – It is the last part of the header section and contains detailed notes and comments about the sequence record.

2) *Features*: Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features [8].

3) Sequence: It contains the original sequence whose length is as specified in the sequence length field. It is represented as string of base pairs (A, C, T, G). Origin specifies the beginning of the sequence.

B. GenBank ASN.1 Standard

The NCBI data model is often confused as NCBI ASN.1 data model. Abstract Syntax Notation (ASN.1) is the data description language in which all the sequence data at NCBI is structured [9]. It defines the structured format in which the data for the biological entities like DNA, genes will be submitted and entered in the GenBank database. The data entering genbank may come through various sources like direct submissions through Sequin, BankIt, Legacy Data systems, DDBJ/EBI and may have different formats. ASN.1 helps to reliably encode data coming from various sources in a same standard format which permits for computers and software systems to easily work on that data and exchange data between databases independent of the original source of data. ASN.1 allows a detailed description of both the sequences and the information associated with them, e.g., author names, source organism, and sequence features [9]. ASN.1 is designed for a computer to read and is amenable to describing complicated data relationships in a very specific way. NCBI describes and processes data using the ASN.1 format [4]. Based on the ASN.1 data format, a number of human-readable formats and tools like Entrez, GenBank, and the BLAST databases are produced. Without the existence of a common format such as this, the neighboring and hard-link relationships that Entrez depends on would not be possible. NCBI's asn2ff (ASN.1 to flatfile) will convert an ASN.1 file into a variety of flatfiles [4]. Thus ASN.1 is the format of data for computers and is a complex format for humans to read. To provide a simple format for humans to read the biological information flat files are used for presenting the output to users

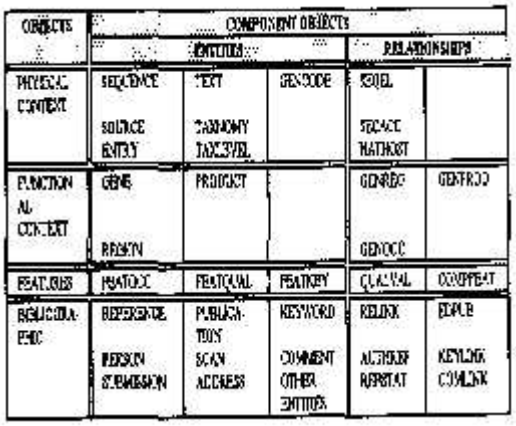
C. GenBank Relational Model

The GenBank is a relational model. The ID database (set of relational tables) stores the ASN.1 structured information. The GenBank database consists of four main objects: physical context data, functional context data, features data and bibliographical data. The above 4 objects are stored in the genbank in a relational model [2]. The relational model of the genbank is used internally by the genbank staff. The Physical context data refers to the biological sources of the sequences used in the database [2]. The functional context data refers to the functions of the nucleic acid molecules [2]. The features data refers to the regions of biological significance reported in the sequence [2]. These can include regions of the sequence that code for proteins and RNA molecules etc. The bibliographical data refers to the references, papers, journals which mention about the sequence. The genbank entries are

stored and maintained in the Sybase relational database management system [6]. Like in relational models, the entries are stored in rows and columns forming a table. The four primary objects have further component objects under them which give a complete description about the primary objects. A simplified ER schema of the Primary objects with its component objects is given in Table1.

TABLE 1.

THE ENTITIES AND RELATIONSHIPS UNDER EACH GENBANK OBJECT [2]



The corresponding schema diagram of ER model of the GenBank is shown in Figure 2.

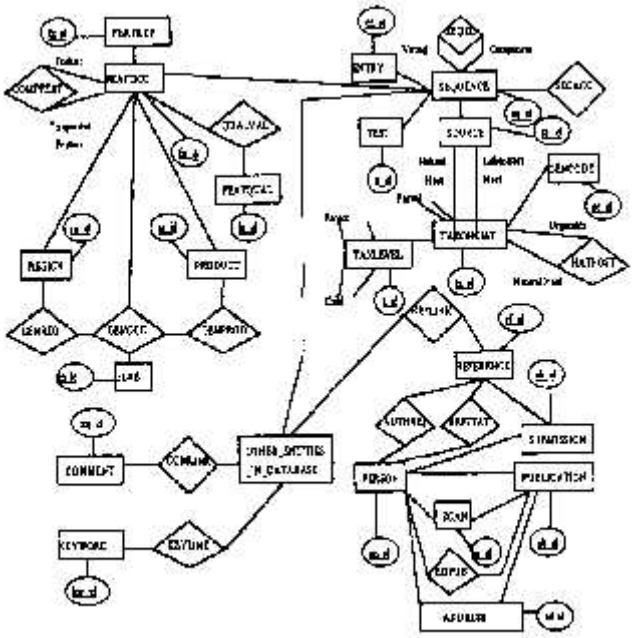


Fig. 2 Simplified Genbank ER diagram [2]

The ER schema diagram can be converted into relational tables. The relational tables are used for storing the sequence related information. In the given ER diagram the boxes are representatives of the main Entities. These entities constitute tables in the Relational Model. The columns (attributes) of the tables are given by circled values of the ER diagram. The diamond shapes represent relationships between the entities and are given by foreign key in relational model. The genbank has the ID database which is a set of relational tables which store all the information of genbank and other data which are included in the entrez (ASN.1 objects and sequence identifier related information).

IV. GENBANK RETRIEVAL TECHNIQUES

GenBank, or any biological database is of little use, unless it provides easy searching and retrieval techniques. Otherwise the sequencing efforts have little purpose, as the information hidden within the million of bases and amino acids can not be deciphered. Entrez the interface developed by NCBI is an effort to provide easy access to the average users. It allows searching and retrieving entries in a useable meaningful format. The Entrez provides relationships and access to many integrated elements like OMIM database, Publications, full text electronic journals, 3D structures, Protein sequences etc. The Entrez can be queried using individual search terms along with Boolean operators like AND, OR, NOT. For efficient searching, keywords can be used corresponding to the fields in the flat file representation. The Entrez searches across all the integrated resources and databases and shows the count of hits found for the search term in all the resources and databases (OMIM, Genes, Protein Structures). Fig. 3 shows the Home Page of Entrez.

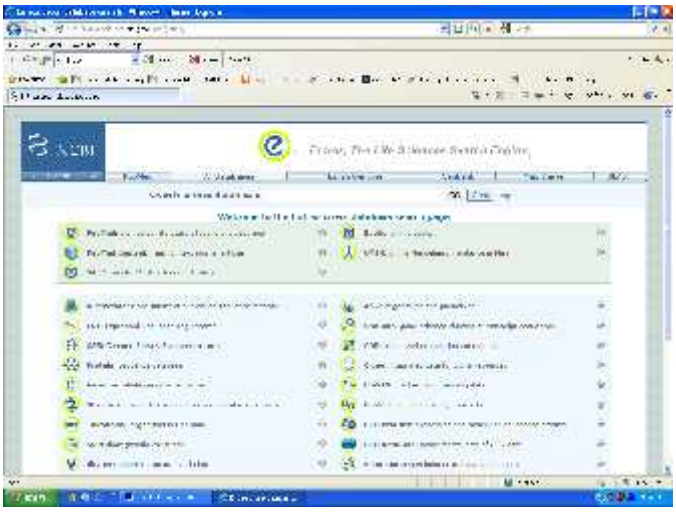
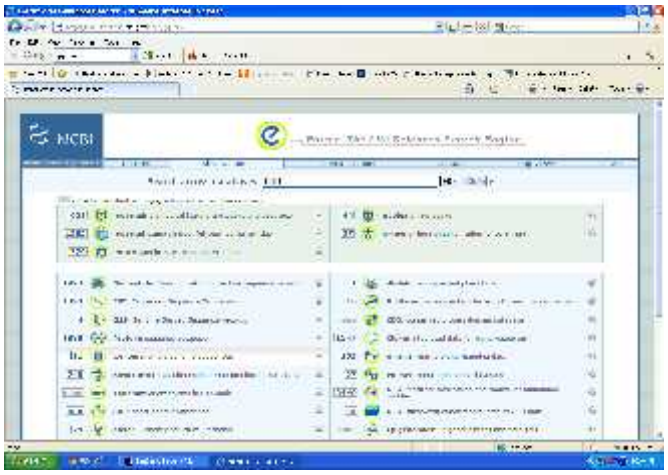


Fig. 3 Home Page of Entrez

The steps for querying a term in Entrez are listed below. We take the example of searching beta haemoglobin of Homosapiens. A mutant of this gene is responsible for the Sickle cell anaemia disease in humans.

- In the search box, enter the term to be searched. We enter HBB. The result counts found are shown in the boxes as seen in the Fig 4. It can be seen that 1993 entries are found in the Nucleotide sequence records.



- The desired database in which the results are to be studied can be selected by clicking on it. The number of results can be narrowed using the limits tab.

Entrez provides an easy and efficient retrieval interface for the users of GenBank. The strength of Entrez lies in the fact that all of this information can be accessed by issuing one and only one query [4]. Searching the genbank for similar nucleotide or protein sequences is handled by the Blast tool developed by NCBI. The Basic Local Alignment Search Tool (BLAST) finds local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

V. CONCLUSIONS

In the paper a detailed study of GenBank database is provided. The data models and retrieval techniques used by GenBank are explained. The paper shows that genbank is not based on a single implementation model but is a combination of Flat Files, ASN.1 and Relational model. All the three are explained in length in the paper. An overview of the retrieval system Entrez and the sequence similarity search tool Blast is also provided.

REFERENCES

[1] M.K.Jhala, C.G.Joshi, T.J.Purohit, N.P.Patel and J.G.Sarvaiya . Role of Bioinformatics in Biotechnology. Technical Report.

[2] I. Kuan Cheang, Y. Bae Choi, and A. Tang, "Overview of the Structures of Heterogeneous Genome Databases", IEEE 27th Annual Hawaii International Conference on System Sciences, 1994.

[3] N. Shamkant, and A. Kogelnik, "The Challenges of Modeling Biological Information for Genome Databases", Conceptual Modeling, LNCS 1565, pp. 168-182, Springer-Verlag 1999.

[4] Andreas D. Baxevanis, B. F. Francis Ouellette, "Bioinformatics – A Practical Guide to the analysis of Genes and Proteins", 2nd ed., John Willey and Sons, New York, 2001.

[5] Francois Bry and Peer Kroeger. A Molecular Biology Database Digest. Technical report, 2001.

[6] Hitomi Hasegawa. Genome Databases: Current Implementation Practices. Technical report, 2008.

[7] Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers, "GenBank," *Nucleic Acids Research.*, vol. 41, pp. D36 – D42, Oct. 2012.

[8] (2013) The NCBI website. [Online]. Available: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html.

[9] Sirotkin K, Tatusova T, Yaschenko E, McEntyre J, Ostell J et al. The Processing of Biological Sequence Data at NCBI. 2002 Oct 9. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21082/