biological databases + file formats

Biological databases emerged as a response to the huge data generated by low-cost DNA sequencing technologies. One of the first databases to emerge was GenBank, which is a collection of all available protein and DNA sequences. It is maintained by the National Institutes of Health (NIH) and the National Center for Biotechnology Information (NCBI). GenBank paved the way for the Human Genome Project (HGP). The HGP allowed complete sequencing and reading of the genetic blueprint. The data stored in biological databases is organized for optimal analysis and consists of two types: raw and curated (or annotated). Biological databases are complex, heterogeneous, dynamic, and yet inconsistent. The inconsistency is due to the lack of standards at the ontological level.

Why are these Important?

Earlier, databases and databanks were considered quite different. However, over the time, database became a preferable term. Data is submitted directly to biological databases for indexing, organization, and data optimization. They help researchers find relevant biological data by making it available in a format that is readable on a computer. All biological information is readily accessible through data mining tools that save time and resources. Biological databases can be broadly classified as sequence and structure databases. Structure databases are for protein structures, while sequence databases are for nucleic acid and protein sequences.

Kinds of Biological Databases

Biological databases can be further classified as primary, secondary, and composite databases.

Primary databases contain information for sequence or structure only. Examples of primary biological databases include:

- Swiss-Prot and PIR for protein sequences
- GenBank and DDBJ for genome sequences
- Protein Databank for protein structures

Secondary databases contain information derived from primary databases. Secondary databases store information such as conserved sequences, active site residues, and signature sequences. Protein Databank data is stored in secondary databases. Examples include:

- SCOP at Cambridge University
- CATH at the University College of London
- PROSITE of the Swiss Institute of Bioinformatics
- eMOTIF at Stanford

Composite databases contain a variety of primary databases, which eliminates the need to search each one separately. Each composite database has different search algorithms and data structures. The NCBI hosts these databases, where links to the Online Mendelian Inheritance in Man (OMIM) is found.

The Future

Because of high-performance computational platforms, these databases have become important in providing the infrastructure needed for biological research, from data preparation to data extraction. The simulation of biological systems also requires computational platforms, which further underscores the need for biological databases. The future of biological databases looks bright, in part due to the digital world.

With a large number of biological databases available, the need for integration, advancements, and improvements in bioinformatics is paramount. Bioinformatics will steadily advance when problems about nomenclature and standardization are addressed. The growth of biological databases will pave the way for further studies on proteins and nucleic acids, impacting therapeutics, biomedical, and related fields.

What is SGD in bioinformatics?

The **Saccharomyces Genome Database** (SGD) provides comprehensive integrated biological information for the budding yeast Saccharomyces cerevisiae along with search and analysis tools to explore these data, enabling the discovery of functional relationships between sequence and gene products in fungi and higher organisms.
The Saccharomyces Genome Database (SGD) provides Internet access to the complete Saccharomyces cerevisiae genomic sequence, its genes and their products, the phenotypes of its mutants, and the literature supporting these data. The amount of information and the number of features provided by SGD have increased greatly following the release of the S.cerevisiae genomic sequence, which is currently the only complete sequence of a eukaryotic genome. SGD aids researchers by providing not only basic information, but also tools such as sequence similarity searching that lead to detailed information about features of the genome and relationships between genes. SGD presents information using a variety of user-friendly, dynamically created graphical displays illustrating physical, genetic and sequence feature maps. SGD can be accessed via the World Wide Web at http://genome-www.stanford.edu/Saccharomyces/

TIGR (The Institute for Genomic Research) Microbial Database

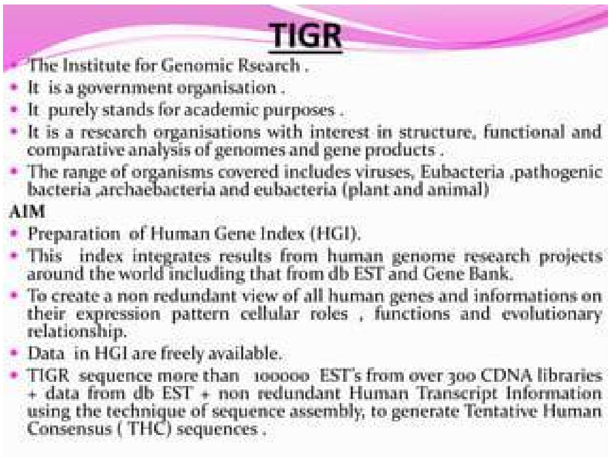
What you can do:

Retrieve genetic information of published microbial genomes and chromosomes and those in progress

Highlights:

- Provides a collection of curated databases containing DNA and protein sequence, gene expression, cellular role, protein family, and taxonomic data for microbes, plants and humans.
  - The CMR (Comprehensive Microbial Resource) contains analysis on completed microbial genome sequencing.

The Institute for Genomic Research (TIGR) is a non-profit research institute located in Rockville, Maryland. The primary interest of TIGR is the sequencing of the genomes and the subsequent analysis of the sequences in prokaryotic and eukaryotic organisms. J. Craig Venter founded TIGR in 1992

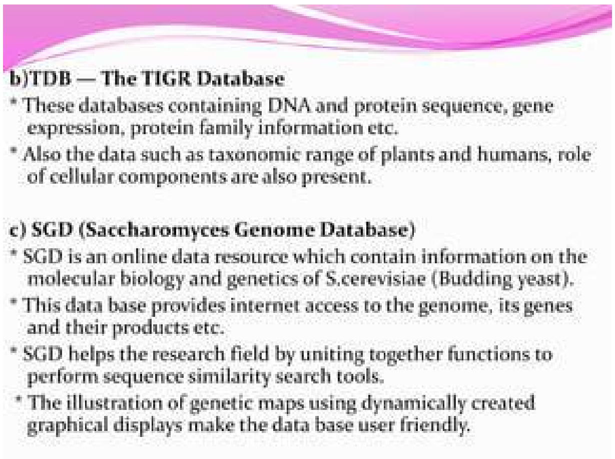


TIGR

- The Institute for Genomic Rsearch .
- It is a government organisation .
- It purely stands for academic purposes .
- It is a research organisations with interest in structure, functional and comparative analysis of genomes and gene products .
- The range of organisms covered includes viruses, Eubacteria ,pathogenic bacteria ,archaebacteria and eubacteria (plant and animal)

AIM

- Preparation of Human Gene Index (HGI).
- This index integrates results from human genome research projects around the world including that from db EST and Gene Bank.
- To create a non redundant view of all human genes and informations on their expression pattern cellular roles , functions and evolutionary relationship.
- Data in HGI are freely available.
- TIGR sequence more than 100000 EST's from over 300 CDNA libraries + data from db EST + non redundant Human Transcript Information using the technique of sequence assembly, to generate Tentative Human Consensus (THC) sequences .



b)TDB — The TIGR Database

* These databases containing DNA and protein sequence, gene expression, protein family information etc.
* Also the data such as taxonomic range of plants and humans, role of cellular components are also present.

c) SGD (Saccharomyces Genome Database)

* SGD is an online data resource which contain information on the molecular biology and genetics of S.cerevisiae (Budding yeast).
* This data base provides internet access to the genome, its genes and their products etc.
* SGD helps the research field by uniting together functions to perform sequence similarity search tools.
* The illustration of genetic maps using dynamically created graphical displays make the data base user friendly.

AceDB

Acedb is a database system developed specifically for handling genome and bioinformatic data, it includes many powerful tools for the manipulation, display and annotation of genomic data.

It was developed by Richard M. Durbin and Jean Thierry-Mieg in 1989. AceDB stands for a **C. elegans database.**_Although AceDB was initially created as a database specifically for the nematode worm it has also come to mean the database software itself, which has been used to store information for other species.

Features included:

- a graphical user interface with many specific displays and tools for genomic data
- open-source software
- package that implements a simple web browser interface allowing the database to accessed from anywhere
- Interfaces easily with perl, Java and CORBA
- Tools for comparative genome analysis including The Oxford Grid, The Pairwise Chromosome Map, The One-to-Many Chromosome Map, The Species Grid, Translocation Grid
- built-in ability to handle phylogenetic data
- easily links to outside applications
- developed to run under the Unix operating system, using X-Windows for graphics, with a local copy of the database files

Development on AceDB appears to have ceased as the interface is still using old technology and many of the sites that originally used it have upgraded to newer software

PDBsum

PDBsum is a database that provides an overview of the contents of each 3D macromolecular structure deposited in the Protein Data Bank. The original version of the database was developed around 1995 by Roman Laskowski and collaborators at University College London.

Full name: **Pictorial database of 3D structures in the Protein Data Bank**. Description: PDBsum is a database of mainly pictorial summaries of the 3D structures of proteins and nucleic acids in the Protein Data Bank.

The PDBsum database at http://www.biochem.ucl.ac.uk/bsm/pdbsum was created in 1995 . Its aim was to provide an at-a-glance summary of the molecules contained in each PDB entry (i.e. protein and DNA/RNA chains, small-molecule ligands, metal ions and waters), together with annotations and analyses of their key structural features. Thus, for each PDB entry there is a corresponding summary web page in PDBsum, accessible by the four-character PDB identifier.

The GenBank file format

The Genbank format allows for the storage of information in addition to a DNA/protein sequence. It holds much more information than the FASTA format. Formats similar to Genbank have been developed by ENA (EMBL format) and by DDBJ (DDBJ format).

the first section includes the entry's LOCUS, DEFINITION, ACCESSION and VERSIONand denoted by ORIGIN, These five elements are the essential parts of the GenBank format.\

The rest of the sections are added information that, although important, are not essential and could be missing. Note, we could not lack the LOCUS field for a GenBank file, or it could not be recognised as such a file. The non-essential parts of the entry contain what is commonly known as metadata, and can include more detailed information about the organism, cross-references to other databases, and even a list of publications in which this entry is featured in. The FEATURES part of the entry describes important characteristics of the entry's sequence such as presence of coding sequences, proteins, etc. This section is less human-friendly, and it may contain fields that do not make any sense to the untrained eye. But don't worry, this parts are mainly intended to be read by a computer program. Finally, at the end of the file, we find the actual sequence that could be DNA or protein. Note that the last line of the entry has a "//". These two characters are very important and indicate the end of the entry/file. Although it might be clear to you that this is the end of the file because there is nothing else underneath, computer programs have to be told when to stop reading.

So within one file, we have a wealth of information, from the nucleotide sequence of the genome to the publications that are related to this genome entry and cross-references to other databases.

GenBank format (GenBank Flat File Format) consists of an annotation section and a sequence section. The start of the annotation section is marked by a line beginning with the word "LOCUS". The start of sequence section is marked by a line beginning with the word "ORIGIN" and the end of the section is marked by a line with only "//".

What is a flat file?
A flat file is a collection of data stored in a two-dimensional database in which similar yet discrete strings of information are stored as records in a table. The columns of the table represent one dimension of the database, while each row is a separate record.

The information stored in a flat file is generally alphanumeric with little or no additional formatting. The structure of a flat file is based on a uniform format as defined by the type and character lengths described by the columns. A flat file format is **a table with a single record per line**. FASTA and other file formats are an example of a flat file format in bioinformatics.

DDBJ flat file format
The database is a collection of "entry" which is the unit of the data. The entry submitted to DDBJ is processed and publicized according to the DDBJ format for distribution (flat file). The flat file includes the sequence and the information of submitters, references, source organisms, and "feature" information, etc. The "feature" is defined by The DDBJ/ENA/GenBank Feature Table Definition to describe the biological nature such as gene function and other property of the nucleotide sequence.

File formats

- CRAM format
- FASTA format
- FASTQ format
- NeXML format
- Nexus format
- Pileup format
- SAM format
- Stockholm format
- VCF format

LOCUS
locus name, sequence length, molecule type, molecular form, division, the date of last release
Locus Name
Locus name is a unique ID of the entry in the database. In DDBJ, since July 1996, the locus name has been assigned the same as accession number.
Length of Sequence
Notice: No information is available on the Master record of MGA data.
Molecule Type
According to the value of /mol_type qqualifier for source feature, it is described as DNA, RNA, mRNA, rRNA, tRNA, or cRNA.
Molecular Form
This column indicates whether molecular form of nucleotide sequence is "linear" or "circular". If the entry is the full length of circular form, "circular" is appeared.
Division
DDBJ classifies entries into 21 divisions
The date of last release
The current publicized date is described. If the entry is updated and reopened to public site, this date will be changed.
DEFINITION
The definition briefly describes the information of gene(s). "DEFINITION" is constructed by each of the three data banks in accordance with standard rules in principle.However, in the case of EST or GSS submission using Mass Submission System, DDBJ will sometimes ask submitters to construct "DEFINITION".

ACCESSION
This line shows accession number of the entry data. A unique accession number is issued to the data submitter by each of the three data banks. The accession number is composed of 1 alphabet character and 5 digits (ex. A12345) or 2 alphabet characters and 6 digits (ex. AB123456). The former style was used in 1980s, but later the latter style was introduced because

of data explosion.
The alphabet part is called “prefix”.

VERSION
This line consists of an accession number and a version number, like “AB123456.1”, in which the digit(s) after the period is a version number.

DBLINK
The DBLINK line is used to link other databases for BioProject, BioSample accession numbers, Sequence Read Archive Run accession numbers and so on. DDBJ has replaced the PROJECT line by DBLINK line format since 2009 to expand for other data resources than projects.

KEYWORDS
The KEYWORDS lines were used for indexing (gene) and (product) names in the past.For now, KEYWORDS lines are used to indicate the detail category of the data (EST, TSA, HTC, etc) information about experimental method, “finishing level” of genome sequencing and else, if necessary.

SOURCE
This line shows the scientific name (and common name, if defined) on organism from which the sequence is obtained and an organelle type if the sequence is derived from an organelle other than the nucleus.

ORGANISM
The organism name and its phylogenic lineage from which the sequence is obtained are described.

REFERENCE 1
The information of submitter(s) is described as REFERENCE 1 (except old entries and some CON entries).
In the case of Nucleotide Sequence Submission System, REFERENCE 1 is processed with the information entered

AUTHORS
Submitter(s) of the entry is/are indicated in principle. Submitter is responsible for the data and can update it.

TITLE
“Direct Submission” is indicated to follow the standard form.

JOURNAL
At first, “Accept Date” of the entry is indicated. “Accept Date” is defined as the date when DDBJ have received the acceptable data to assign accession number in principle. Even if the entry is updated, “Accept Date” is NOT changed. Then, the information about the address and the affiliation of “Contact Person” is indicated.

REFERENCE 2
The information of references related to the submitted sequence is indicated on REFERENCE line (other than (REFERENCE 1). Since REFERENCE 2 indicates the publication status of the sequence, the reference which does not describe about the submitting sequence is indicated as REFERENCE 3 or after, not as REFERENCE 2.

COMMENT
The information about an entry that can not be described using FEATURES or the other fields. For instance, if submitter has the other affiliation to REFERENCE 1, it can be described on COMMENT line.

FEATURES
Biological features of a submitted sequence data are described with “Feature” key (the biological nature of the annotated feature), “Location” (the region of the sequence which corresponds to Feature), and “Qualifier” (supplementary information about Feature). In principle, EST or GSS entries are not described with any features except the “source” key.

source
Identifies the biological source of the specified span of the sequence.

CDS
Coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon).

//
”//” is the terminal symbol of the entry.

FASTA format

In bioinformatics and biochemistry, the FASTA format is a text-based format for representing either nucleotide sequences or amino acid sequences, in which nucleotides or amino acids are represented using single-letter codes. The format allows for sequence names and comments to precede the sequences.

What is FASTA full form?
FASTA is pronounced "fast A", and stands for "**FAST-All**", because it works with any alphabet, an extension of the original "FAST-P" (protein) and "FAST-N" (nucleotide) alignment tools.
A sequence in FASTA format begins with a single-line description, followed by lines of sequence data.

What are the main parts of a FASTA file?
The FASTA format is composed of two main parts: (i) the heading line of each sequence, starting with the character “>”, followed by the “specimen ID” and the “species name field” (dived by a vertical bar “|”) and (ii) the nucleotide sequences (a string of A, C, G, T characters).

What are the advantages of FASTA format?
FASTA is a file format used for storing nucleotide and amino acid polymeric sequences and is compatible with a high variety of bioinformatics software. It is used as database for ribosomal RNA sequences but also for eukaryotic reference genomes and protein databases, that can be several gigabytes in size.

What is FASTA vs Genbank format?

The Genbank format allows for the storage of information in addition to a DNA/protein sequence. It holds much more information than the FASTA format. Formats similar to Genbank have been developed by ENA (EMBL format) and by DDBJ (DDBJ format).

Protein Data Bank (file format)

The **Protein Data Bank (PDB) file format** is a textual file format describing the three-dimensional structures of molecules held in the Protein Data Bank, now succeeded by the mmCIF format. The PDB format accordingly provides for description and annotation of protein and nucleic acid structures including atomic coordinates, secondary structure assignments, as well as atomic connectivity. In addition experimental metadata are stored. The PDB format is the legacy file format for the Protein Data Bank which now keeps data on biological macromolecules in the newer mmCIF file format. The PDB file format was invented in 1976 as a human-readable file that would allow researchers to exchange protein coordinates through a database system. Its fixed-column width format is limited to 80 columns, which was based on the width of the computer punch cards that were previously used to exchange the coordinates. Through the years the file format has undergone many changes and revisions.
A typical PDB file describing a protein consists of hundreds to thousands of lines
HEADER, TITLE and AUTHOR records
provide information about the researchers who defined the structure; numerous other types of records are available to provide other types of information.
REMARK records
can contain free-form annotation, but they also accommodate standardized information; for example, the REMARK 350 BIOMT records describe how to compute the coordinates of the experimentally observed multimer from those of the explicitly specified ones of a single repeating unit.
SEQRES records
give the sequences of the three peptide chains (named A, B and C), which are very short in this example but usually span multiple lines.
ATOM records
describe the coordinates of the atoms that are part of the protein. For example, the first ATOM line above describes the alpha-N atom of the first residue of peptide chain A, which is a proline residue; the first three floating point numbers are its x, y and z coordinates and are in units of Ångströms. The next three columns are the occupancy, temperature factor, and the element name, respectively.
HETATM records
describe coordinates of hetero-atoms, that is those atoms which are not part of the protein molecule.

PDB

Filename extension	.pdb, .ent, .brk
Internet media type	chemical/x-pdb
Developed by	Protein Data Bank
Type of format	chemical file format
Container for	Molecule 3D structure, Protein tertiary structure

What does a SwissProt file look like

- The ID record
- The ID line

  The ID (IDentification) line is always the first line of an entry.
- The AC record

  The AC (ACcession number) line lists the accession number(s) associated with an entry.
- The DE record

The DE (DEscription) lines contain general descriptive information about the stored sequence. This information is generally sufficient to identify the protein precisely.

- References

The DR record
The DR (Database cross-Reference) lines are used as pointers to information related to SwissProt entries and found in data collections other than SwissProt.

The FT record
The SQ record (the actual sequence)
One would almost forget, but the SwissProt file does also contain a sequence. The format of this sequence part of the file occasionally depends on the search machine used to get that sequence.

SwissProt files are so-called **keyword-organised flat-files**. That means, the file is human readable (which tends to be called a flat-file or an ASCII file) and every line starts with a keyword (in SwissProt that is a two letter code).

The Different Bioinformatics File Types

The FASTA bioinformatics tool was invented in 1988 and used for performing sensitive sequence alignments of DNA or protein sequences.[1] It's associated file type – FASTA format – has become a standard file type in bioinformatics.[2] The rise of sequencing technologies and the development of robust bioinformatics analysis tools have given rise to several others. And if you get involved in using other sequence alignment tools in bioinformatics or other types of sequence analysis, you are sure to encounter and use them extensively.

Sequence formats

FASTA

The FASTA file format is the simplest way of representing nucleic acid of protein sequences using single-letter codes for nucleotides or amino acids.[1,2] For each sequence, there are two lines:

- The first is a sequence identifier, which contains information about the sequence, preceded with a ">" symbol. If you retrieve a sequence from GenBank, SWISS-PROT, BLAS, or another database, the identifier will follow a standardized format.
- The second line in a FASTA file is the nucleotide or amino acid sequence, using single letter IUPAC codes.[3]

These file types, denoted by the .fas extension, are used by most large curated databases. Specific extensions exist for nucleic acids (.fna), nucleotide coding regions (.ffn), amino acids (.faa), and non-coding RNAs (.frn).[4]

A FASTA file can contain one or many sequences. Tools like ClustalW can take FASTA files with multiple sequences to generate an alignment. Converting between FASTA formats and any of the others discussed below can be done with programs like Seqret and MView.[5] Other simple sequence file formats that you may encounter include GCG and IG.

FASTQ

The FASTQ format was developed for and used with next-generation sequencing instruments and builds off of the simplicity of the FASTA format. Information about the quality ("Q" in "FASTQ" stands for quality) of the sequencing reads and base calls are a defining component of the FASTQ file format.

Alignment formats;BAM,sam,cram

Stockholm formats;VCF (Variant Calling Format

Generic feature formats;A GFF (general feature format,The GTF (gene transfer format

Unlabeled formats;PDB file formats contain atomic coordinates and are used for storing 3D protein structures by the Protein Data Bank., MAP (.map file extension) ,CSV (.csv file format) ,PED (.ped file extension)

Why Are There So Many Different Types?

In modern-day, the many different ways of generating and using sequencing data have given rise to the sequence file formats described above. These file formats have their own specific use cases depending on:[4]

- Compatibility with specific software
- Data processing, parsing, and human readability needs
- Efficiency for storage

There are several similarities and differences between several of the file types:

- Generally, all of the different file formats are similarly structured: They contain a header with metadata and a body with lines or fields of data.
- FASTA and FASTQ are used to store raw sequencing data, yet the FASTQ file also holds quality data. FASTA also can store DNA, RNA, and protein sequences, while FASTQ usually only contains DNA sequences.

What are the common file formats in bioinformatics?
The FASTA file format is one of the most widely used bioinformatics file types. FASTQ is also used broadly due to the widespread adoption of next-generation sequencing. Other common file types include SAM, **BAM**, CRAM, BED, **VCF**, GFF, and GTF.

What is flat format in bioinformatics?
A flat file format is a table with a single record per line. FASTA and other file formats are an example of a flat file format in bioinformatics.

What are data types in bioinformatics?
Data types in bioinformatics can be DNA sequences, RNA sequences, amino acid sequences, methylation sequences, three-dimensional protein structures, and more.

Why do we have different sequence file formats?
Scientists are using bioinformatic data for many different purposes, and other file types include different kinds of information concerning a DNA, RNA, or protein sequence. These various file types may be used for compatibility with additional bioinformatics software or storage efficiency.