Introduction to Biological Databases

1. Introduction

As biology has increasingly turned into **a data-rich science**, the need for storing and communicating large datasets has grown tremendously. The obvious examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and macromolecular NMR. A new field of science dealing with issues, challenges and new possibilities created by these databases has emerged: **bioinformatics**.

Bioinformatics is the application of Information technology to store, organize and analyze the vast amount of biological data which is available in the form of sequences and structures of proteins (the building blocks of organisms) and nucleic acids (the information carrier). The biological information of nucleic acids is available as sequences while the data of proteins is available as sequences and structures. Sequences are represented in single dimension where as the structure contains the three dimensional data of sequences.

Sequences and structures are only among the several different types of data required in the practice of the modern molecular biology. Other important data types includes metabolic pathways and molecular interactions, mutations and polymorphism in molecular sequences and structures as well as organelle structures and tissue types, genetic maps, physiochemical data, gene expression profiles, two dimensional DNA chip images of mRNA expression, two dimensional gel electrophoresis images of protein expression, data A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. There are two main functions of biological databases:

- **Make biological data available to scientists.**
  - As much as possible of a particular type of information should be available in one single place (book, site, and database). Published data may be difficult to find or access and collecting it from the literature is very time-consuming. And not all data is actually published explicitly in an article (genome sequences!).
- **To make biological data available in computer-readable form.**
  - Since analysis of biological data almost always involves computers, having the data in computer-readable form (rather than printed on paper) is a necessary first step.

Data Domains

- Types of data generated by molecular biology research:

  – Nucleotide sequences (DNA and mRNA)

  – Protein sequences

  – 3-D protein structures

  – Complete genomes and maps

- Also now have:
  – Gene expression
  – Genetic variation (polymorphisms)

2. Biological Databases

When Sanger first discovered the method to sequence proteins, there was a lot of excitement in the field of Molecular Biology. Initial interest in Bioinformatics was propelled by the necessity to create databases of biological sequences.

Biological databases can be broadly classified into sequence and structure databases. Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure database is applicable to only Proteins. The first database was created within a short period after the Insulin protein sequence was made available in 1956. Incidentally, Insulin is the first protein to be sequenced. The sequence of Insulin consisted of just 51 residues (analogous to alphabets in a sentence) which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases (individual units of nucleic acids) was found out. During this period, three dimensional structures of proteins were studied and the well known Protein Data Bank was developed as the first protein structure database with only 10 entries in 1972. This has now grown in to a large database with over 10,000 entries. While the initial databases of protein sequences were maintained at the individual laboratories, the development of a consolidated formal database known as SWISS-PROT protein sequence database was initiated in 1986 which now has about 70,000 protein sequences from more than 5000 model organisms, a small fraction of all known organisms. These huge varieties of divergent data resources are now available for study and research by both academic institutions and industries. These are made available as public domain information in the larger interest of research community through Internet (www.ncbi.nlm.nih.gov) and CDROMs (on request from www.rcsb.org). These databases are constantly updated with additional entries.

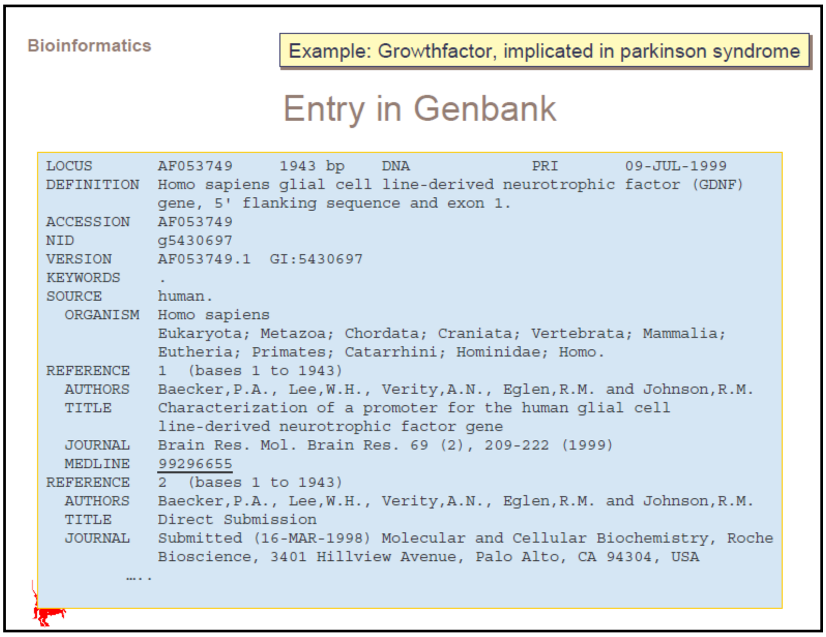
Databases in general can be classified in to **primary**, **secondary** and **composite** databases. A **primary** database contains information of the sequence or structure alone. Examples of these include Swiss-Prot & PIR for protein sequences, GenBank & DDBJ for Genome sequences and the Protein Databank for protein structures.

A **secondary** database contains derived information from the primary database. A secondary sequence database contains information like the conserved sequence, signature sequence and active site residues of the protein families arrived by multiple sequence alignment of a set of related proteins. A secondary structure database contains entries of the PDB in an organized way. These contain entries that are classified according to their structure like all alpha proteins, all beta proteins, etc. These also contain information on conserved secondary structure motifs of a particular protein. Some of the secondary database created and hosted by various researchers at their individual laboratories includes SCOP, developed at Cambridge University; CATH developed at University College of London, PROSITE of Swiss Institute of Bioinformatics, eMOTIF at Stanford.

Composite database amalgamates a variety of different primary database sources, which obviates the need to search multiple resources. Different composite database use different primary database and different criteria in their search algorithm. Various options for search have also been incorporated in the composite database. The National Center for Biotechnology Information (NCBI) which hosts these nucleotide and protein databases in their large high available redundant array of computer servers, provides free access to the various persons involved in research. This also has link to OMIM (Online Mendelian Inheritance in Man) which contains information about the proteins involved in genetic diseases.

2.1 Primary Nucleotide Sequence Repository – GenBank, EMBL, DDBJ

These are three chief databases that store and make available raw nucleic acid sequences. GenBank is physically located in the USA and is accessible through NCBI portal over internet. EMBL (European Molecular Biology Laboratory) is in UK and DDJB (DNA databank of Japan) is in Japan. They have uniform data formats (but not identical) and exchange data on daily basis. Here we will describe one of the database formats, GenBank, in detail. The access to GenBank, as to all databases at NCBI is through the Entrez search program. This front end search interface allows a great variety of search options.

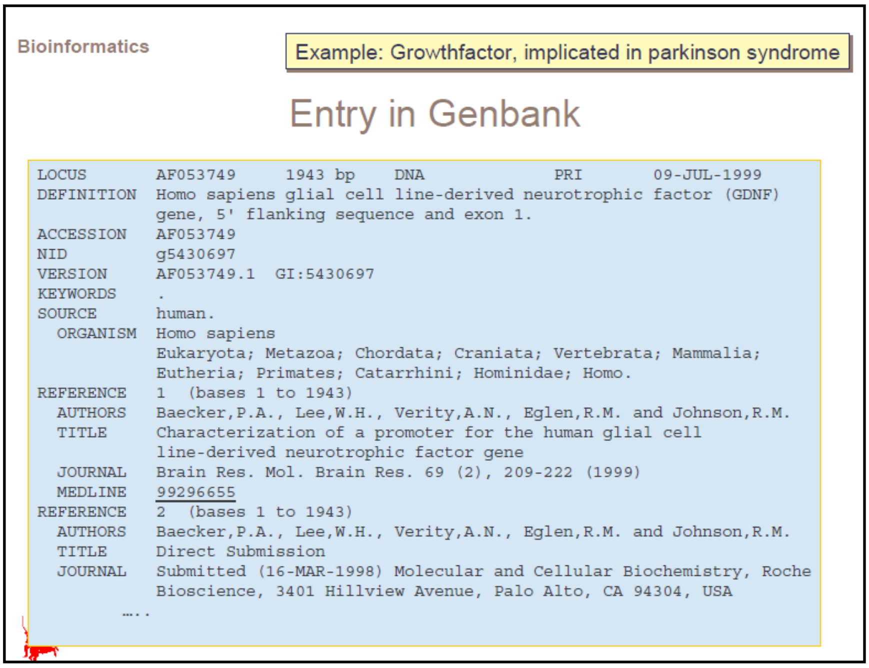


Bioinformatics

Example: Growthfactor, implicated in parkinson syndrome

Entry in Genbank

```
LOCUS       AF053749     1943 bp    DNA             PRI       09-JUL-1999
DEFINITION  Homo sapiens glial cell line-derived neurotrophic factor (GDNF)
            gene, 5' flanking sequence and exon 1.
ACCESSION   AF053749
NID         g5430697
VERSION     AF053749.1  GI:5430697
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
            Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1943)
  AUTHORS   Baecker,P.A., Lee,W.H., Verity,A.N., Eglen,R.M. and Johnson,R.M.
  TITLE     Characterization of a promoter for the human glial cell
            line-derived neurotrophic factor gene
  JOURNAL   Brain Res. Mol. Brain Res. 69 (2), 209-222 (1999)
  MEDLINE   99296655
REFERENCE   2  (bases 1 to 1943)
  AUTHORS   Baecker,P.A., Lee,W.H., Verity,A.N., Eglen,R.M. and Johnson,R.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-MAR-1998) Molecular and Cellular Biochemistry, Roche
            Bioscience, 3401 Hillview Avenue, Palo Alto, CA 94304, USA
         …..
```

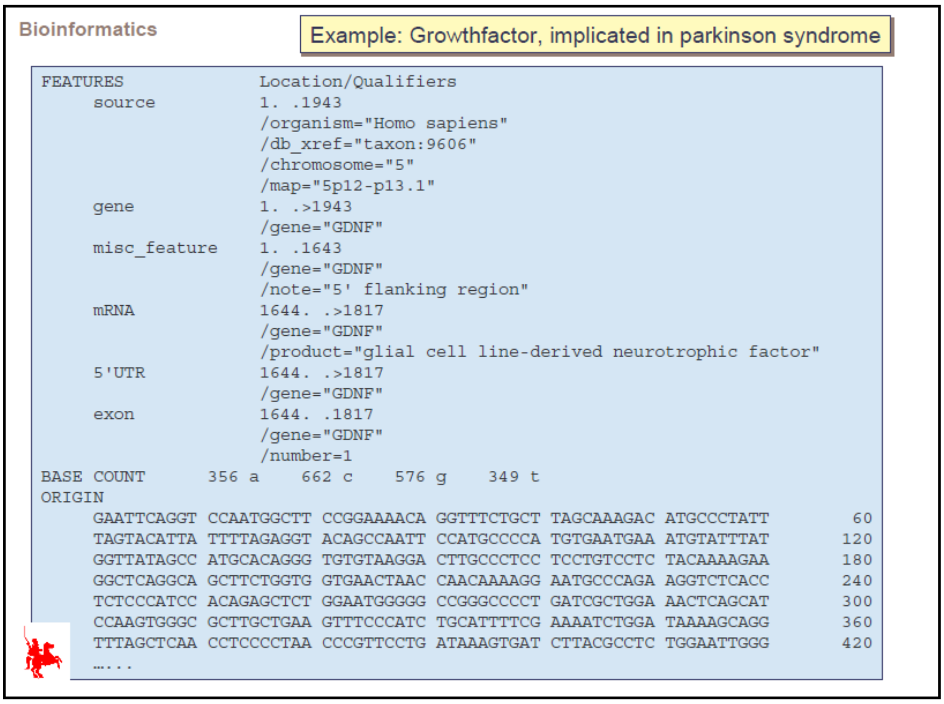


Bioinformatics

Example: Growthfactor, implicated in parkinson syndrome

Entry in Genbank

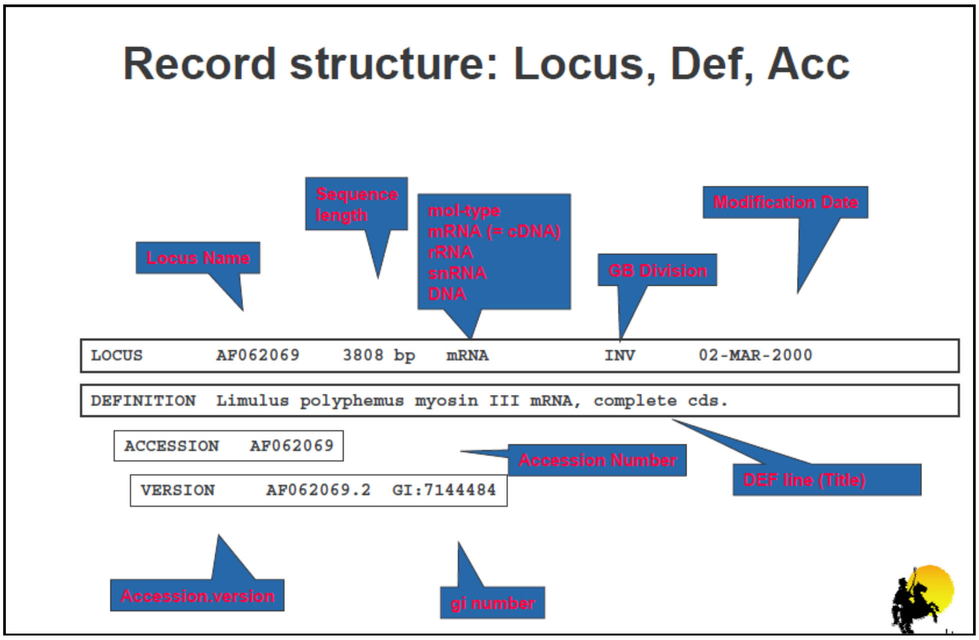
```
LOCUS       AF053749     1943 bp    DNA             PRI       09-JUL-1999
DEFINITION  Homo sapiens glial cell line-derived neurotrophic factor (GDNF)
            gene, 5' flanking sequence and exon 1.
ACCESSION   AF053749
NID         g5430697
VERSION     AF053749.1  GI:5430697
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
            Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1943)
  AUTHORS   Baecker,P.A., Lee,W.H., Verity,A.N., Eglen,R.M. and Johnson,R.M.
  TITLE     Characterization of a promoter for the human glial cell
            line-derived neurotrophic factor gene
  JOURNAL   Brain Res. Mol. Brain Res. 69 (2), 209-222 (1999)
  MEDLINE   99296655
REFERENCE   2  (bases 1 to 1943)
  AUTHORS   Baecker,P.A., Lee,W.H., Verity,A.N., Eglen,R.M. and Johnson,R.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-MAR-1998) Molecular and Cellular Biochemistry, Roche
            Bioscience, 3401 Hillview Avenue, Palo Alto, CA 94304, USA
        …..
```



Bioinformatics

Example: Growthfactor, implicated in parkinson syndrome

```
FEATURES             Location/Qualifiers
     source          1. .1943
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="5"
                     /map="5p12-p13.1"
     gene            1. .>1943
                     /gene="GDNF"
     misc_feature    1. .1643
                     /gene="GDNF"
                     /note="5' flanking region"
     mRNA            1644. .>1817
                     /gene="GDNF"
                     /product="glial cell line-derived neurotrophic factor"
     5'UTR           1644. .>1817
                     /gene="GDNF"
     exon            1644. .1817
                     /gene="GDNF"
                     /number=1
BASE COUNT      356 a    662 c    576 g    349 t
ORIGIN
     GAATTCAGGT CCAATGGCTT CCGGAAAACA GGTTTCTGCT TAGCAAAGAC ATGCCCTATT       60
     TAGTACATTA TTTTAGAGGT ACAGCCAATT CCATGCCCCA TGTGAATGAA ATGTATTTAT      120
     GGTTATAGCC ATGCACAGGG TGTGTAAGGA CTTGCCCTCC TCCTGTCCTC TACAAAAGAA      180
     GGCTCAGGCA GCTTCTGGTG GTGAACTAAC CAACAAAAGG AATGCCCAGA AGGTCTCACC      240
     TCTCCCATCC ACAGAGCTCT GGAATGGGGG CCGGGCCCCT GATCGCTGGA AACTCAGCAT      300
     CCAAGTGGGC GCTTGCTGAA GTTTCCCATC TGCATTTTCG AAAATCTGGA TAAAAGCAGG      360
     TTTAGCTCAA CCTCCCCTAA CCCGTTCCTG ATAAAGTGAT CTTACGCCTC TGGAATTGGG      420
     …...
```



Record structure: Locus, Def, Acc

Sequence length

mol-type
mRNA (= cDNA)
rRNA
snRNA
DNA

Modification Date

Locus Name

GB Division

```
LOCUS       AF062069     3808 bp    mRNA              INV       02-MAR-2000
DEFINITION  Limulus polyphemus myosin III mRNA, complete cds.
ACCESSION   AF062069
VERSION     AF062069.2  GI:7144484
```

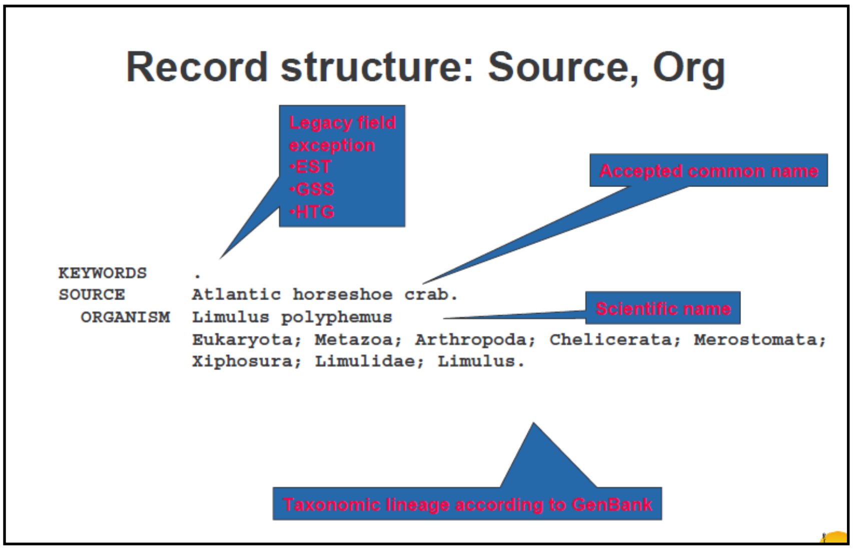
Accession Number

DEF line (Title)

Accession.version

gi number

The word accession number defines a field containing unique identification numbers. The sequence and the other information may be retrieved from the database simple by searching for a given accession number. Taking the field names in order, we have first all the word ‘LOCUS’. This is a GenBank title that names the sequence entry. Apart for accession number, it also specifies the number of bases in the entry, a nucleic acid type, a codeword PRI that indicates the sequence is from primate, and the date on which the entry was made. PRI is one of the 17 keyword search that are used to classify the data. The next line of the file contains the definition of the entry, giving the name of the sequence. The unique accession number came next, followed by a version number in case the entries have gone through more than one version.



Record structure: Source, Org

Legacy field exception
•EST
•GSS
•HTG

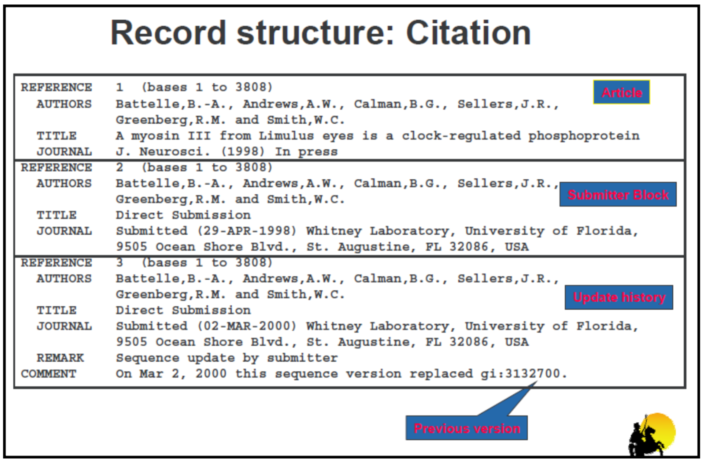
Accepted common name

```
KEYWORDS    .
SOURCE      Atlantic horseshoe crab.
  ORGANISM  Limulus polyphemus
            Eukaryota; Metazoa; Arthropoda; Chelicerata; Merostomata;
            Xiphosura; Limulidae; Limulus.
```

Scientific name

Taxonomic lineage according to GenBank

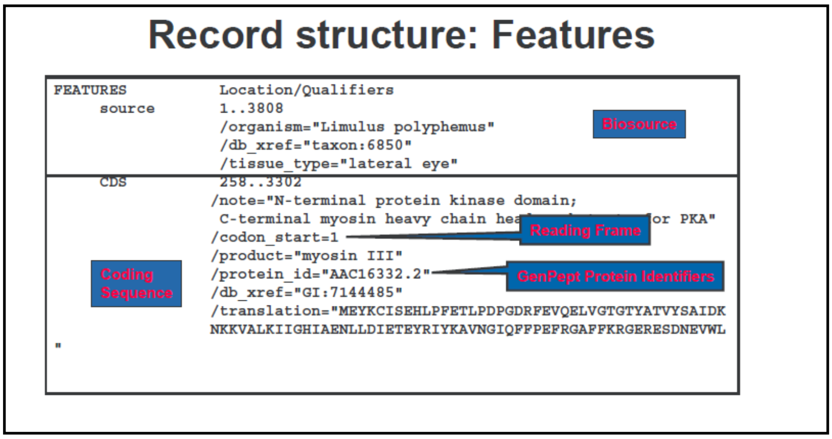
The next item is a list of specially defined keywords that used to index the entries. Next come a set of SOURCE records which describe the organism from which sequence was extracted. The complete scientific classification is given. This is followed by publication details.



Record structure: Citation

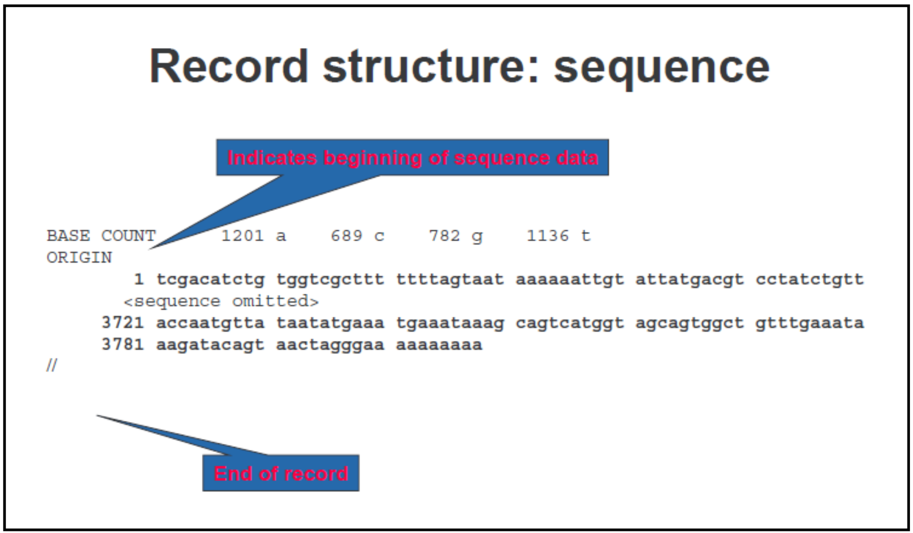
```
REFERENCE   1  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     A myosin III from Limulus eyes is a clock-regulated phosphoprotein
  JOURNAL   J. Neurosci. (1998) In press
REFERENCE   2  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-APR-1998) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
REFERENCE   3  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-MAR-2000) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
  REMARK    Sequence update by submitter
COMMENT     On Mar 2, 2000 this sequence version replaced gi:3132700.
```

In the beginning, sequences were extracted from the published literature and painstaking entered in the database. Each entry was therefore associated with a publication. The features table includes coding region, exons, introns, promoters, alternate splice patterns, mutation, variations and a translation into protein sequence, if it code for one. Each feature may be accompanied by a cross-reference to another database. After the feature table, a single line gives the base count statistics for the sequence. Finally comes the sequence itself. The sequence is typed in lower cases, and for ease of reading, each line is divided into six columns of ten bases each. A single number on the left numbers the bases.



Record structure: Features

```
FEATURES             Location/Qualifiers
     source          1..3808
                     /organism="Limulus polyphemus"
                     /db_xref="taxon:6850"
                     /tissue_type="lateral eye"
     CDS             258..3302
                    /note="N-terminal protein kinase domain;
                     C-terminal myosin heavy chain hea              or PKA"
                    /codon_start=1
                    /product="myosin III"
                    /protein_id="AAC16332.2"
                    /db_xref="GI:7144485"
                    /translation="MEYKCISEHLPFETLPDPGDRFEVQELVGTGTYATVYSAIDK
                    NKKVALKIIGHIAENLLDIETEYRIYKAVNGIQFFPEFRGAFFKRGERESDNEVWL
"
```

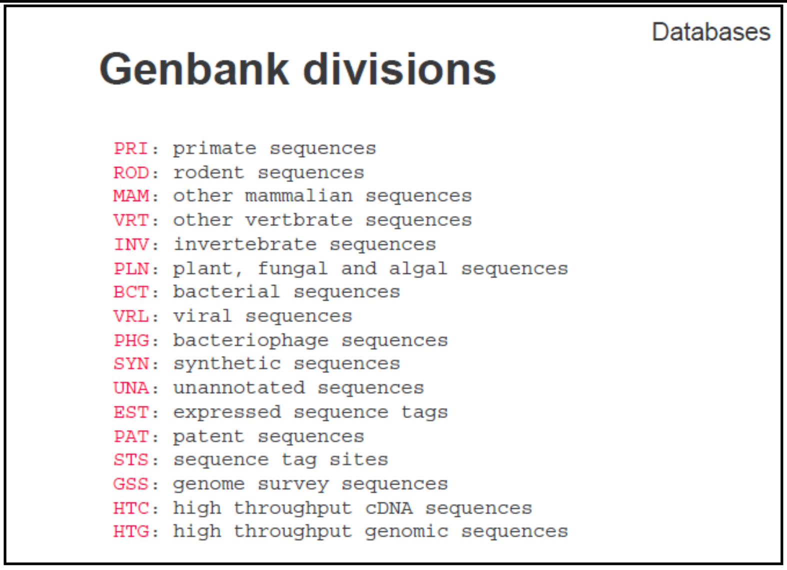


Record structure: sequence

Indicates beginning of sequence data

```
BASE COUNT      1201 a     689 c     782 g     1136 t
ORIGIN
        1 tcgacatctg tggtcgcttt ttttagtaat aaaaaattgt attatgacgt cctatctgtt
       <sequence omitted>
     3721 accaatgtta taatatgaaa tgaaataaag cagtcatggt agcagtggct gtttgaaata
     3781 aagatacagt aactagggaa aaaaaaaa
//
```

End of record



Databases

Genbank divisions

```
PRI: primate sequences
ROD: rodent sequences
MAM: other mammalian sequences
VRT: other vertbrate sequences
INV: invertebrate sequences
PLN: plant, fungal and algal sequences
BCT: bacterial sequences
VRL: viral sequences
PHG: bacteriophage sequences
SYN: synthetic sequences
UNA: unannotated sequences
EST: expressed sequence tags
PAT: patent sequences
STS: sequence tag sites
GSS: genome survey sequences
HTC: high throughput cDNA sequences
HTG: high throughput genomic sequences
```

The above description does not cover all the fields used in GenBank, but only the more important ones.

2.2 Primary Protein Sequence Repositories

PIR-PSD or protein information resource – protein sequence database, at the NBRF (National Biomedical Research Foundation, USA), and SWISS-PROT at the SBI (Swiss Biotechnology Institute), Switzerland are protein sequence databases.

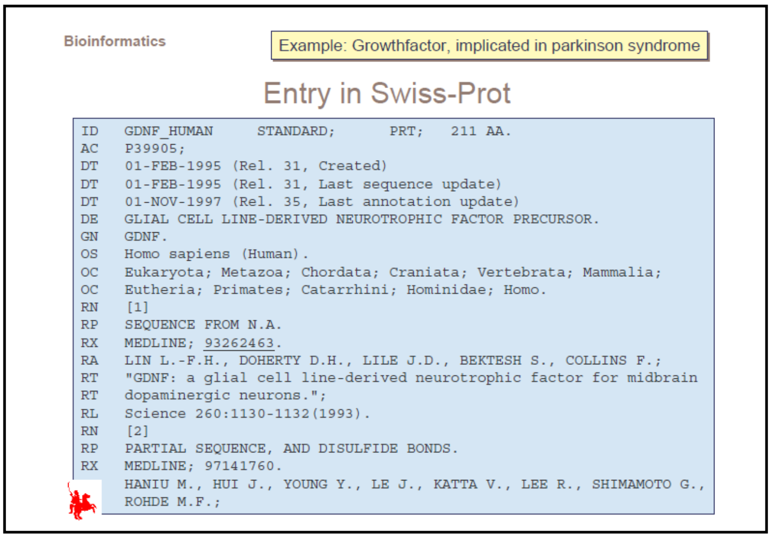
The PIR-PSD is a collaborative endeavour between the PIR, the MIPS (Munich Information Centre for Protein Sequences, Germany) and the JIPID (Japan International Protein Information Database, Japan). The PIR-PSD is now a comprehensive, non-redundant, expertly annotated, object relational DBMS. It is available at http://pir.georgetown.edu/pirww. A unique characteristic of the PIR-PSD is its classification of protein sequences based on the super family concept. Sequence in PIR-PSD is also classified based on homology domain and sequence motifs. Homology domains may correspond to evolutionary building blocks, while sequence motifs represent functional sites or conserved regions. The classification approach allows a more complete understanding of sequence function structure relationship.

The other well known and extensively used protein database is SWISS-PROT(http://www.expasy.ch/sprot). Like the PIR-PSD, this curated proteins sequence database also provides a high level of annotation. The data in each entry can considered separately as core data and annotation. The core data consists of the sequences entered in common single letter amino acid code, and the related references and bibliography. The taxonomy of the organism from which the sequence was obtained also forms part of this core information. The annotation contains information on the function or functions of the protein, post-translational modification such as phosphorylation, acetylation, etc., functional and structural domains and sites, such as calcium binding regions, ATP-binding sites, zinc fingers, etc., known secondary structural features as for examples alpha helix, beta sheet, etc., the quaternary structure of the protein, similarities to other protein if any, and diseases that may rise due to different authors publishing different sequences for the same protein, or due to mutations in different strains of an described as part of the annotation.

Lines of code in SWISS-PROT database:

Code	**Expansion**	**Remarks**
ID	Identification	Occurs at the beginning of the entry. Contains a unique name for the entry, plus information on the status of the entry. If it has been checked and conforms to SWISS-PROT standards, it is called STANDARD.
AC	Accession numbers	This is a stable way of identifying the entry. The name may change but not the AC. If the line has more than one number, it means that the entry was constituted by merging other entries.
DT	Date	There are three dates corresponding to the creation date of the entry and modification dates of the sequence and the annotation respectively
DE	Description	Lines that start with the identifier contain general description about the sequence.
GN	Gene name	The name of the gene (or genes) that codes for the protein
OS, OG,OC	Organism name, Organelle, Organism classification	The name and taxonomy of the organism, and information regarding the organelle containing the gene e.g. mitochondria or chloroplast, etc.
RN, RP,RX,RA RT,RL	Reference number, Position, comments, cross-reference, authors, title and location.	Bibliographic reference to the sequence. This includes information (following the code RP) on the extent of work carried out b the authors.
CC	Comments	These are free text comments that provide any relevant information pertaining to the entry.
DR	Database cross-reference	This line gives cross-references to other databases where information regarding this entry is also found. As for example to structural information for the protein in the PDB.

KW	Keywords	This line gives a list of keywords that can be used in indexes. Search programs very often simply go through such indices to identify required information
FT	Features Table	These lines describe regions or sites of interest in the sequence, e.g. post-transitional modifications, binding sites, enzyme active sites and local secondary structures
SQ	Sequence Header	This line indicates the beginning of the sequence data and gives a brief summary of its contents.



Bioinformatics

Example: Growthfactor, implicated in parkinson syndrome

Entry in Swiss-Prot

```
ID   GDNF_HUMAN     STANDARD;      PRT;   211 AA.
AC   P39905;
DT   01-FEB-1995 (Rel. 31, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   01-NOV-1997 (Rel. 35, Last annotation update)
DE   GLIAL CELL LINE-DERIVED NEUROTROPHIC FACTOR PRECURSOR.
GN   GDNF.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 93262463.
RA   LIN L.-F.H., DOHERTY D.H., LILE J.D., BEKTESH S., COLLINS F.;
RT   "GDNF: a glial cell line-derived neurotrophic factor for midbrain
RT   dopaminergic neurons.";
RL   Science 260:1130-1132(1993).
RN   [2]
RP   PARTIAL SEQUENCE, AND DISULFIDE BONDS.
RX   MEDLINE; 97141760.
     HANIU M., HUI J., YOUNG Y., LE J., KATTA V., LEE R., SHIMAMOTO G.,
     ROHDE M.F.;
```

Both PIR-PSD and SWISS-PROT have software that enables the user to easily search through the database to obtain only the required information. SWISS-PROT has the SRS or the sequence retrieval system that searches also through the other relevant databases on the site, such as TrEMBL.

TrEMBL (for Translated EMBL) is a computer-annotated protein sequence database that is released as a supplement to SWISS-PROT. It contains the translation of all coding sequences present in the EMBL Nucleotide database, which have not been fully annotated. Thus it may contain the sequence of proteins that are never expressed and never actually identified in the organisms.

2.3 Derived or Secondary databases of nucleotide sequences

Many of the secondary databases are simply sub-collection of sequences culled from one or the other of the primary databases such as GenBank or EMBL. There is also usually a great deal of value addition in terms of annotation, software, presentation of the information and the cross-references. There are other secondary databases that do not present sequences at all, but only information gathered from sequences databases.

An example of the former type of database is the FlyBase or The Bereley Drosophila Genome Project (http://www.fruitfly.org). A consortium sequenced the entire genome of the fruit fly *D. Melanogaster* to a high degree of completeness and quality.

Another database that focuses on a single organism is ACeDB. More than a database, this is a database management system that was originally developed for the *C. Elegans* (a nematode worm) genome project. It is a repository of not only the sequence, but also the genetic map as well as phenotypic information about the *C. Elegans* nematode worm.

The comprehensive Microbial Resource maintained by TIGR (The Institute for Genomic Research) at http://www.tigr.org allows access to a database called Omniome. This contains all the focus on one organism. Omniome has not only the sequence and annotation of each of the completed genomes, but also has associated information about the organisms (such as taxon and gram stain pattern), the structure and composition of their DNA molecules, and many other attributes of the protein sequences predicted from the DNA sequences. The presence of all microbial genomes in a single database facilitated meaningful multi-genome searches and analysis, for instance, alignment of entire genomes, and comparison of the physical proper of proteins and genes from different genomes etc.

A database of the genomes of mitochondria and other such organelles is available at the Organelle Genome Database at the University of Montreal, Canada, and is called GOBASE (http://megasun.bch.umontreal.ca/gobase).

2.4 Derived or Secondary databases of amino acid sequences - Subcollections

Another family of a database focussed on a particular family protein is GPCRGB (http://rose.man.pozen.pl/aars/). These are transmembrane protein used by cells to communicate with the outside world. They are involved in vision, smell, hearing, taste and feeling.GPCRGB is in fact more than a collection of sequences of the protein family. It includes additional data on multiple sequences alignments. Ligands and ligands binding data, 3D models, mutation data, literature reference, disease patterns, cell lines, protocols, vectors etc. It is fully integrated information system with data, and browsing and query tools.

MHCPep (http://wehih.wehi.edu.au/mhcpep/) is a database comprising over 13000 peptide sequences known to bind the Major Histocompatibilty Complex of the immune system. Each entry in the database contains not only the peptide sequence, which may be 8 to 10 amino acid long, but in addition has information on the specific MHC molecules to which it binds, the experimental method used to assay the peptide, the degree of activity and the binding affinity observed , the source protein that, when broken down gave rise to this peptide along with other, the positions along the peptide where it anchors on the MHC molecules and references and cross links to other information.

The CluSTr (Cluster of SWISS-PROT and TrEMBL proteins at http://ebi.ac.uk.clustr) database offers an automatic classification of the entries in the SWISS-PROT and TrEMBL databases into groups of related proteins. The clustering is based on the analysis of all pair wise comparisons between protein sequences.

Similar to CluSTRr is the COGS or Cluster of Orthologous Groups of database that is accessible at http://ncbi.nlm.nih.gov/COG. An orthologous group of proteins is one in which the members are related to each other by evolutionary descent. Such orthology may not be just from one protein to another, and then to another and so on down the line. It may involve one-to-many ad many-to-many evolutionary relationships, and hence the term ‘groups’. COGS is thus a database of phylogenetic relationships. The approximately 2500 groups have been divided into 17 broad categories. The utility of COGS, as of CluSTr, is that it helps assign function to new protein sequences without going through tedious biochemical discovery processes.

2.5 Derived or Secondary databases of amino acid sequences – Patterns and Signature

A set of databases collects together patterns found in protein sequences rather than the complete sequences. The patterns are identified with particular functional and/or structural domains in the protein, such as for example, ATP binding site or the recognition site of a particular substrate. The patterns are usually obtained by first aligning a multitude of sequences through multiple alignment techniques. This is followed by further processing by different methods, depending on the particular database.

PROSITE is one such pattern database, which is accessible at http://www.expasy.ch/prosite. The protein motif and pattern are encoded as “regular expressions”. The information corresponding to each entry in PROSITE is of the two forms – the patterns and the related descriptive text. The regular expression is placed in a format reminiscent of the SWISS-PROT entries, with a two letter identifier at beginning of the each line specifying the type of information the line contains. The expression itself is placed on line identified by “PA”. The entry also contains references and links to all the proteins sequences that contains that pattern. The related descriptive text is placed in a documentation file with the accession number making the connection to the expression data.

In the PRINTS database (http://www.bioinfo.man.ac.uk/dbbrowser/PRINTS), the protein sequence patterns are stored as ‘fingerprints’. A finger print is a set of motifs or patterns rather than a single one. The information contained in the PRINT entry may be divided into three sections. In addition to entry name, accession number and number of motifs, the first section contains cross links to other databases that have more information about the characterized family. The second section provides a table showing how many of the motifs that make up the finger print occurs in the how many of the sequences in that family. The last section of the entry contains the actual finger prints that are stored as multiply aligned set of sequences, the alignment being made without gaps. There is therefore one set of aligned sequences for each motif.

The ProDom protein domain database (http://www.toulouse.inrs.fr/prodom.html) is a compilation of homologous domains that have been automatically identified sequence comparison and clustering methods using the program PSI-BLAST. No identification of patterns is made.. The focus is here to look for complete and self-contained structural domains and the search methods includes signals for such features. A graphical user interface allows easy interactive analysis of structural and therefore functional homology relationships among protein sequences.

A database called Pfam contains the profiles used using Hidden markov models (http://www.sanger.ac.uk/Software/Pfam). HMMs build the model of the pattern as a series of match, substitute, insert or delete states, with scores assigned for alignment to go from one state to another. Each family or pattern defined in the Pfam consists of the four elements. The first is the annotation, which has the information on the source to make the entry, the method used and some numbers that serve as figures of merit. The second is the seed alignment that is used to bootstrap the rest of the sequences into the multiple alignments and then the family. The third is the HMM profile. The fourth element is complete alignment of all the sequences identified in that family.

2.6 Structure Databases

Structure databases like sequence databases comes in two varieties, primary and secondary. Strictly speaking there is only one database that stores primary structural data of biological molecules, namely the PDB. In the context of this database, term macromolecule stretches to cover three orders of magnitude of molecular weight from 1000 Daltons to 1000 kilo Daltons Small biological and organic molecules have their structures stored in another primary structure database the CSD, which is also widely used in biological studies. This contains the three dimensional structure of drugs, inhibitors and fragments or monomers of the macromolecule.

2.6.1 The primary structure database - PDB and CSD

PDB stands for Protein Databank. In spite of the name, PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as antibiotic gramicidin and complexes of protein and nucleic acids. The database holds data derived from mainly three sources. Structure determined by X-ray crystallography form the large majority of the entries. This is followed by structures arrived at by NMR experiments. There are also structures obtained by molecular modelling. The data in the PDB is organized as flat files, one to a structure, which usually means that each file contain one molecule, or one molecular complex.

The Cambridge Structural Database (CSD) was originally a project of the University of Cambridge, which is set up to collect together the published three-dimensional structure of small organic molecules. This excludes proteins and medium sized nucleic acid fragments, but small peptides such as neuropeptides, and monomer and dimmers of nucleic acid finds a place in the CSD. Currently CSD holds crystal structures information for about 2.5 lakhs organic and metal organic compounds. All these crystal structures have been obtained using X-ray or neuron diffraction technique. For each entry in the CSD there are three distinct types of information stored. These are categorized as bibliographic information, chemical connectivity information and the three- dimensional coordinates. The annotation data field incorporates all of the bibliographic material for the particular entry and summarized the structural and experimental information for the crystal structure.

2.6.1.1 Derived or Secondary databases of bimolecular structures

NDB stands for Nucleic acid data bases. It is a relational database of three-dimensional structures containing nucleic acid. This encompasses DNA and RNA fragments, including those with unusual chemistry such as NDB, and collections of patterns and motifs such as SCOP, PALI etc. The structures are the same as those found in the PDB and therefore the NDB qualifies to be called a specialized sub collection. However a substantial amount, and, unlike the PDB, the NDB is much more than just a collection of files. The structure of DNA has been classified into A, B and Z polymorphic forms, based on the information specified by authors. Other classes include RNA structures, unusual structures and protein-nucleic acid complexes. These classes of structures are arranged in the form of an ATLAS of Nucleic Acid Containing Structures, which can be browse and searched to obtain the structure or structures required. Each entry in the atlas has information on the sequence, crystallisation condition, references and details of the parameters and the figures of the merit used in structure solution. The entry has links not only to the coordinated but also to automatically generated graphical views of the molecule. NDB also has also have archives of structural geometries calculated for all the structures or for a subset of them. And finally, the database stores average geometrical parameters for nucleic acids, obtained by statistical analysis of the structures. These parameters are widely used in computer simulations of nucleic acids and their interactions. The NDB may be accessed at http://ndbserve.rutgers.edu/NDB/.

The SCOP database (Structural Classification of Proteins: http://scop.mrc-lmb.cam.ac.uk/scop/) is a manual classification of protein structures in a hierarchical scheme with many levels. The principal classes are the family, the super family and the fold. SCOP is a searchable and browsable database. In other words, one may either enter SCOP at the top of the hierarchy or examine different folds and families as one pleases, or one may supply a keyword or a phrase to be search the database and retrieve corresponding entries. Once a structure, or a set of structures, has been selected, they may be obtained or viewed wither as graphical images. Each entry also has other annotation regarding function, etc., and links to other databases, including other structural classification such as CATH.

CATH stands for Class, Architecture, Topology and Homologous super family. The name reflects the classification hierarchy used in the database. The structures chosen for classification are a subset of PDB, consisting of those that have been determined to a high degree of accuracy.

Conclusion

The present challenge is to handle a huge volume of data, such as the ones generated by the human genome project, to improve database design, develop software for database access and manipulation, and device data-entry procedures to compensate for the varied computer procedures and systems used in different laboratories. There is no doubt that Bioinformatics tools for efficient research will have significant impact in biological sciences and betterment of human lives.