An Insight of Biological Databases Used in Bioinformatics

1

Vaibhav D. Bhatt, Monika Patel, and Chaitanya G. Joshi

Abstract

Collections of life sciences information from scientific investigations, high-throughput experiment technology, available literature, and computational analysis are called biological databases. It contains information from research areas comprising genomics, microarray gene expression, proteomics, phylogenetics, metabolomics, gene function, structure, localization and similarities of biological sequences. In a nutshell, databases are libraries for storage and representation of biological data obtained from the scientific community which converts data into knowledge. Utmost biological databases are available from websites that categorize data which operators can browse through the data online. Due to the vast amount of data generated by high-throughput DNA sequencers in the investigation of genome, transcriptome, and exome sequences of various organisms in current times, the biological data has stored with an exponential rate. The availability of enormous amount of biological data (sequences as well as structural) has generated a need for managing, storing, and retrieving this huge data. This chapter reviews current knowledge of the different types of databases available with examples of their file formats.

Keywords

Biological sequences · High-throughput DNA sequencers · Transcriptome and exome sequences

V. D. Bhatt (✉)
Department of Pharmaceutical Sciences, Saurashtra University, Rajkot, Gujarat, India

M. Patel · C. G. Joshi
Department of Animal Biotechnology, College of Veterinary Science and Animal Husbandry, Anand Agricultural University, Anand, Gujarat, India

G. Wadhwa et al. (eds.), *Current trends in Bioinformatics: An Insight*,
https://doi.org/10.1007/978-981-10-7483-7_1

1.1 Introduction

Databases are the convenient system to properly store, search, and recover several types of data. A database helps to easily handle and share large amount of data and supports large-scale analysis by easy access and data update (Liu and Özsu 2009).

Due to the vast amount of data generated in experiments of genome, transcriptome, and exome sequences of various organisms in current times, the biological data has stored with an exponential rate. The availability of enormous amount of biological data (sequences as well as structural data) has generated a need for managing, storing, and retrieving this huge data.

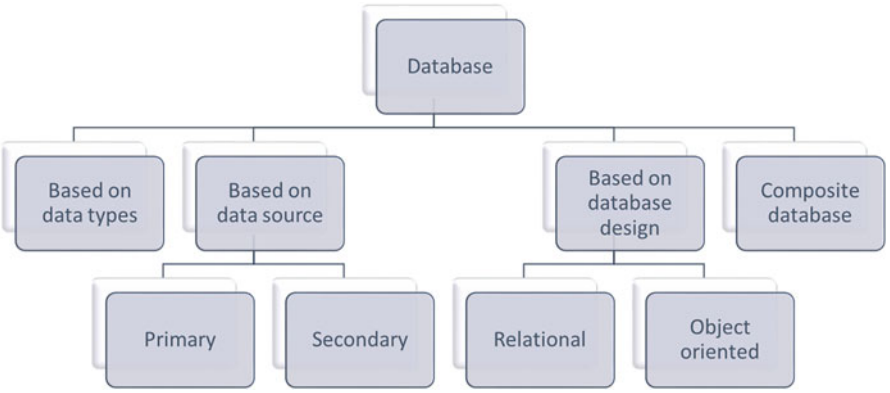
Therefore the biological databases have come into existence as invaluable sources for the biological community. In a nutshell, databases are libraries for storage and representation of biological data obtained from the scientific community which converts data into knowledge.

1.2 History

A book published in 1965, *Atlas of Protein Sequences and Structures*, *was the first biological database* by Margaret Dayhoff and colleagues, and further they have published other editions of the book in the 1970s; however the first edition was limited to 65 sequences only (Dayhoff and Foundation 1973, 1976; Foundation 1972).

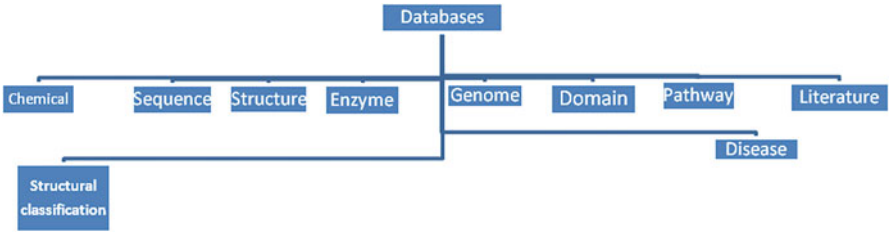
With the discovery of the integrated circuit, the powerful and reliable third generation computers are became the choice of storage of biological databases for scientists. An English scientist Tim Berners-Lee in 1989 invented the "World Wide Web" (WWW) which is the primary tool people use to interact on the Internet and is the way to access all biological databases. Production of high throughput sequencing machines leads production of data rich science, needs an interdisciplinary arena to develop software tools which is used to understand biological data. The field of science with the involvement of computer, statistics and engineering to study biological data is called Bioinformatics.

1.3 Classification of Biological Databases



1.3.1 Databases Based on Data Types

This database was divided into several databases; some of the databases were discussed below in detail.



1.3.1.1 Sequence Databases

Sequence databases contain both nucleic acid and protein sequences. First we will discuss about nucleotide sequence repositories.

(I) Nucleic Acid Sequence Database

There are three main nucleotide sequence repositories:

(A) GenBank
(B) European Molecular Biology Laboratory (EMBL)
(C) DNA Data Bank of Japan (DDBJ)

Raw nucleic acid sequences are stored in these databases and make available through Internet sources. Initially, these databases worked independently, but later the *International Nucleotide Sequence Database Collaboration* (INSDC, http://insdc.org) was developed to maintain collaboration between DDBJ, GenBank, and EMBL (Fig. 1.1). These databases started exchanging their data through constant communication between the team at each collaborating organizations in order to access the sequences present in all three different formats.

(A) *GenBank*

GenBank is a collection of raw and annotated nucleotide as well as protein information. GenBank is maintained and accessed through the National Center for Biotechnology Information (NCBI). Every 2 months a new release is made. It is maintained by NCBI as part of the INSDC (Benton 1990). There are approximately 137384889783 bases, from 149819246 sequence records in the GenBank release 188.0 on February 15, 2012. Type "insulin" in the search tab on the GenBank home page to view list of sequences of insulin gene, partial or complete from different organisms (Fig. 1.2).

Example of GenBank Format

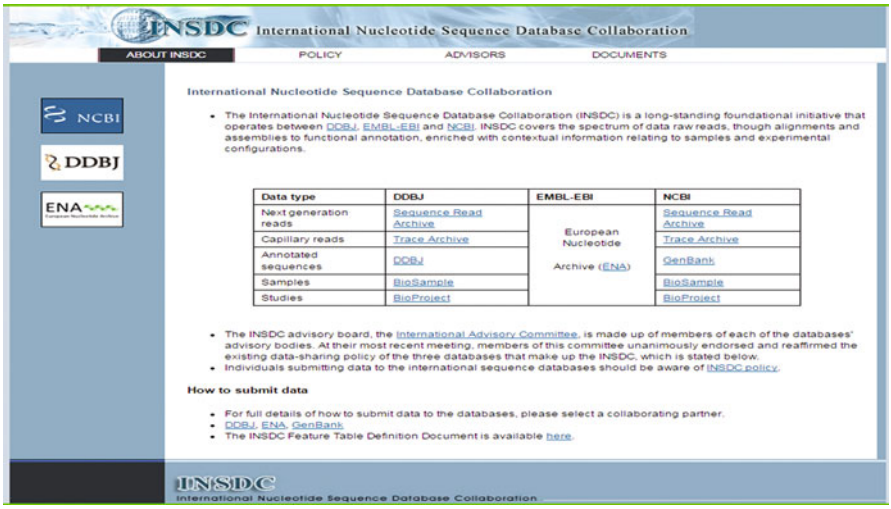


Fig. 1.1 The home page of International Nucleotide Sequence Database Collaboration (INSDC) (http://insdc.org)

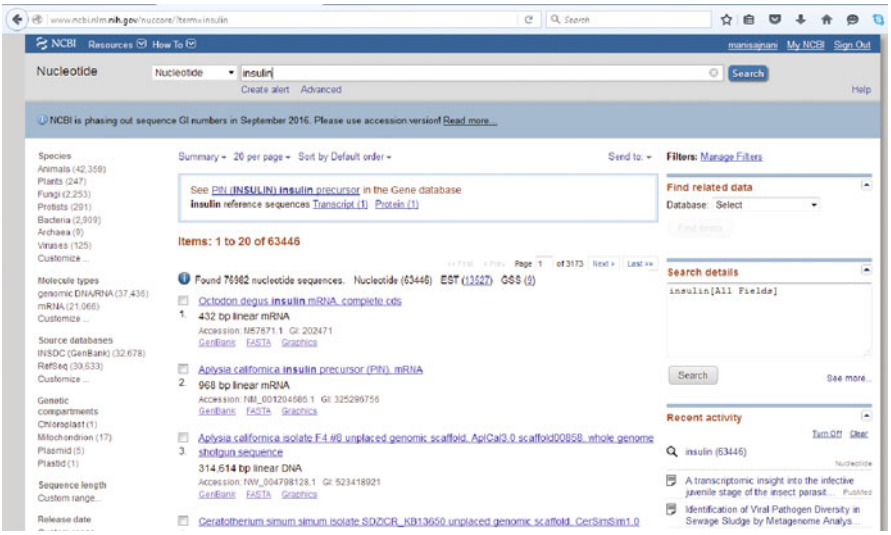


Fig. 1.2 Using GenBank to query insulin sequences (http://www.ncbi.nlm.nih.gov/nuccore/?term=insulin)

Octodon degus insulin mRNA, complete cds

GenBank: M57671.1

FASTA Graphics

Go to:

```
LOCUS       OCOINS                  432 bp    mRNA    linear   ROD 27-APR-1993
DEFINITION  Octodon degus insulin mRNA, complete cds.
ACCESSION   M57671
VERSION     M57671.1
KEYWORDS    insulin; insulin alpha-chain; insulin beta-chain; insulin
            connecting peptide.
SOURCE      Octodon degus (degu)
  ORGANISM  Octodon degus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Hystricognathi; Octodontidae; Octodon.
REFERENCE   1  (bases 1 to 432)
  AUTHORS   Nishi,M. and Steiner,D.F.
  TITLE     Cloning of complementary DNAs encoding islet amyloid polypeptide,
            insulin, and glucagon precursors from a New World rodent, the degu,
            Octodon degus
  JOURNAL   Mol. Endocrinol. 4 (8), 1192-1198 (1990)
   PUBMED   2293024
COMMENT     Original source text: Octodon degus pancreas, cDNA to mRNA.
FEATURES             Location/Qualifiers
     source          1..432
                     /organism="Octodon degus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10160"
                     /tissue_type="pancreas"
     gene            1..432
                     /gene="insulin"
     CDS             42..371
                     /gene="insulin"
                     /codon_start=1
                     /product="insulin"
                     /protein_id="AAA40590.1"
                     /translation="MAPWMHLLTVLALLALWGPNSVQAYSSQHLCGSNLVEALYMTCG
                     RSGFYRPHDRRELEDLQVEQAELGLEAGGLQPSALEMILQKRGIVDQCCNNICTFNQL
                     QNYCNVP"
     sig_peptide     42..113
                     /gene="insulin"
     mat_peptide     114..200
                     /gene="insulin"
                     /product="insulin B-chain"
     mat_peptide     207..293
                     /gene="insulin"
                     /product="insulin C-peptide"
     mat_peptide     300..368
                     /gene="insulin"
                     /product="insulin A-chain"
     regulatory      414..419
                     /regulatory_class="polyA_signal_sequence"
                     /gene="insulin"
     polyA_site      432
                     /gene="insulin"
ORIGIN
        1 gcattctgag gcattctcta acaggttctc gaccctccgc catggccccg tggatgcatc
       61 tcctcaccgt gctggccctg ctggccctct ggggacccaa ctctgttcag gcctattcca
      121 gccagcacct gtgcggctcc aacctagtgg aggcactgta catgacatgt ggacggagtg
      181 gcttctatag accccacgac cgccgagagc tggaggacct ccaggtggag caggcagaac
      241 tgggtctgga ggcaggcggc ctgcagcctt cggccctgga gatgattctg cagaagcgcg
      301 gcattgtgga tcagtgctgt aataacattt gcacatttaa ccagctgcag aactactgca
      361 atgtccctta gacacctgcc ttgggcctgg cctgctgctc tgccctggca accaataaac
      421 cccttgaatg ag
//
```

Format Explanation

GenBank format includes *locus name* which is similar to the accession number and unique to the entry, and it is followed by sequence length. In our example sequence length is 587 bp. Definition includes description of source organism, gene/protein name, and other details about sequence.

- *Accession number* is the unique identifier of the sequence (NM_013564).
- *Version* is similar to accession number, but whenever a change occurs in sequence data, the version increases by 1. In our example, version is NM_013564.7; this indicates that sequence has been changed seven times.
- *GI (GenInfo Identifier)* number also runs parallel to the accession number and version system. A new GI is allotted, if the sequence has been changed and the version has increased by unity. In our example, GI is 365192585.
- *Keywords* are words or expressions about sequence. The keyword field contains a dot if nothing is provided.
- *Source* contains name of the organism from which the sequence has been derived.
- *Organism* is a related sub-keyword of source and contains the scientific name of the organism along with the lineage as described in NCBI taxonomy database.
- *Reference* contains the publication by the authors of the sequence.
- *Authors* contain list of authors in the same order as appears in publication.
- *Title* shows the title of published/unpublished work.
- *Journal* contains MEDLINE abbreviations of the journal name where the work is published.
- *PubMed* field provides the PubMed identifier (PMID) of that article.
- *Comment* points out the change occurred in the submitted sequence.
- *Features* provide information about genes and their products, segment of biological significance in the submitted sequence, as well as other characteristics.
- *Gene* provides gene length and gene name and its function and synonyms. CDS represents coding sequence which codes for protein sequence.
- *Origin* contains the sequence data. Finally, GenBank record ends with // sign.

Sequence Submission to GenBank

Sequence submission is done by using different tools available at NCBI. Few of them are:

BankIt: direct submissions are made to GenBank using it (www.ncbi.nlm.nih.gov/WebSub/?tool=genbank).
Sequin: it is a stand-alone submission platform (www.ncbi.nlm.nih.gov/Sequin/).
tbl2asn: it is a command-line program, used for submission of large batches of sequences and complete genomes (www.ncbi.nlm.nih.gov/genbank/tbl2asn2).

Table 1.1 Various databases and software tools of NCBI for sequence analysis

NCBI			
Tools			Databases
Sequence Submission	Sequence	Data mining	Literature
	Analysis		Nucleotide
Sequin	BLAST	Entrez	Protein
BankIt	Blink	My NCBI	Structure
tbl2asn	Stand-alone BLAST	LinkOut	Genome
			OMIM
			SNP
Barcode Submission Tool	e-PCR	Citation	Books
		Matcher	Domain
	ORF Finder		Chemical
			Expression
	Map viewer		Other databases
	Tax plot		
	Trace archive		

Barcode Submission Tool: it is a WWW-based tool for the submission of sequences and trace read data (http://www.ncbi.nlm.nih.gov/WebSub/?tool=barcode).

National Center for Biotechnology Information (NCBI)

NCBI was started in 1988, as a part of the US National Library of Medicine (NLM) located at Bethesda, Maryland. It is a division of the National Institutes of Health and is directed by David Lipman. The responsibility of NCBI is to make available the GenBank nucleotide sequence database since 1992. NCBI is playing a very remarkable role for biological scientists by making available various public databases and software tools for sequence analysis (Table 1.1). GenBank manages with individual laboratories and other sequence databases like those of the EMBL and the DDBJ. Meanwhile in 1992, NCBI has developed to run other databases in addition to GenBank ((US) 2013). The home page of NCBI is shown in Fig. 1.3.

Databases and Tools of NCBI

Database Retrieval Tool

Entrez (www.ncbi.nlm.nih.gov/Entrez/) in Fig. 1.4 is a primary text search engine which comprises of 40 molecular and literature databases. It extracts huge information from the PubMed database, such as DNA and protein sequences and structure, gene, genome, genetic variation, and gene expression.

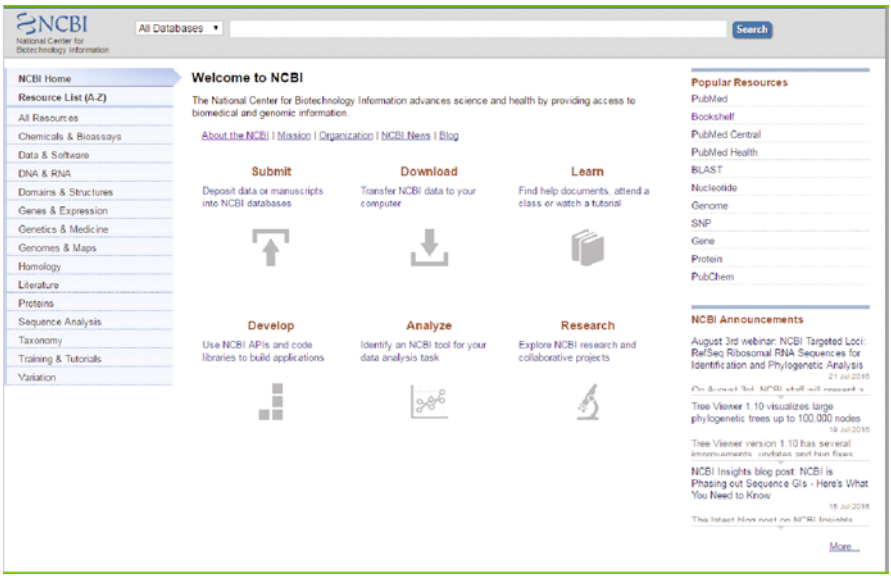


Fig. 1.3 The home page of National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/)

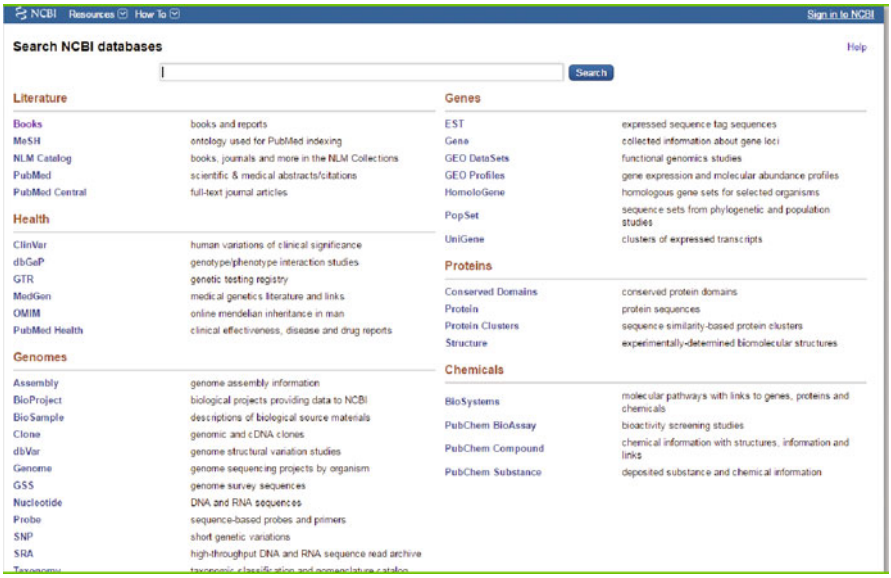


Fig. 1.4 The home page of Entrez (www.ncbi.nlm.nih.gov/Entrez/)

Fig. 1.5 The home page of European molecular biology laboratory (http://www.embl.org/)

(B) *European Molecular Biology Laboratory (EMBL)*

The *European Molecular Biology Laboratory* (*EMBL*) (http://www.embl.org/) in Fig. 1.5 is a molecular biology organization which is maintained by 20 European countries, with Australia as associate member state. It is an intergovernmental organization created in 1974. It develops and maintains a large number of databases, and scientists can access the data free of cost. This research laboratory functions from five different locations, the main laboratory, the European Bioinformatics Institute (EBI), Heidelberg, Germany, is a hub for bioinformatics research and services, directed by Dr. Rolf Apweiler and Dr. Ewan Birney. It is a part of INSDC, which includes DDBJ and GenBank. Typing insulin gene at EMBL search engine produced a result in Fig. 1.6.

EMBL File Format

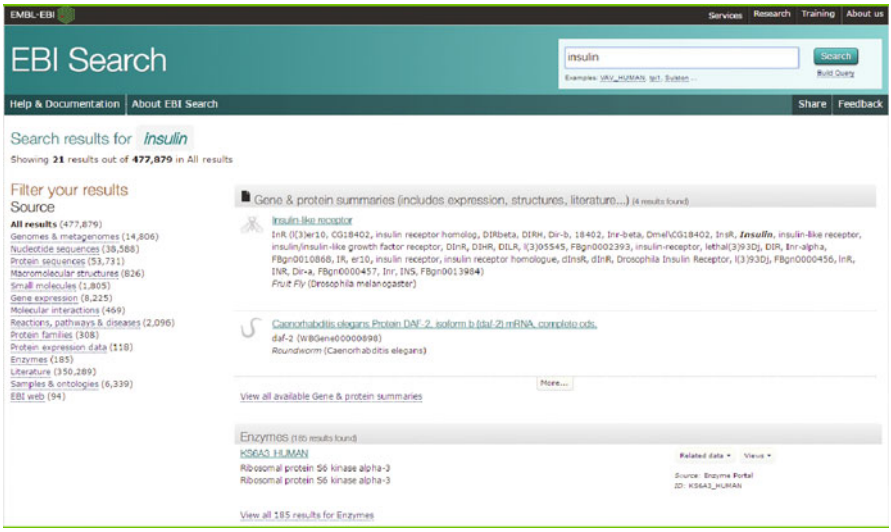


Fig. 1.6 Insulin gene search at European molecular biology laboratory website (https://www.ebi.ac.uk/ebisearch/search.ebi?query=insulin&db=allebi&requestFrom=searchBox)

```
ID   AH002190; SV 2; linear; genomic DNA; STD; ROD; 782 BP.
XX
AC   AH002190; M25583; M25585;
XX
DT   13-JUN-2016 (Rel. 129, Created)
DT   13-JUN-2016 (Rel. 129, Last updated, Version 1)
XX
DE   Rattus norvegicus insulin 2 (INS2) gene, complete cds.
XX
KW   insulin.
XX
OS   Rattus norvegicus (Norway rat)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC   Muridae; Murinae; Rattus.
XX
RN   [1]
RP   1-782
RX   DOI; 10.1111/j.1749-6632.1980.tb47271.x.
RX   PUBMED; 6249167.
RA   Lomedico P.T., Rosenthal N., Kolodner R., Efstratiadis A., Gilbert W.;
RT   "The structure of rat preproinsulin genes";
RL   Ann. N. Y. Acad. Sci. 343:425-432(1980).
XX
DR   MD5; 2b03b65970e00e00d50a5054fad8125c.
XX
CC   On or before Jun 10, 2016 this sequence version replaced gi:204949,
CC   gi:204950, gi:204951.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..782
FT                   /organism="Rattus norvegicus"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:10116"
FT   gene            1..739
FT                   /gene="INS2"
FT   exon            <1..46
FT                   /gene="INS2"
FT                   /number=1
FT   intron          47..165
FT                   /gene="INS2"
FT                   /number=1
FT   CDS             join(180..366,541..686)
FT                   /codon_start=1
FT                   /gene="INS2"
FT                   /product="insulin 2"
FT                   /note="precursor"
FT                   /protein_id="AAA41440.1"
FT                   /translation="MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGE
FT                   RGFFYTPMSRREVEDPQVAQLELGGGPGAGDLQTLALEVARQKRGIVDQCCTSICSLYQ
FT                   LENYCN"
FT   sig_peptide     180..251
FT                   /gene="INS2"
FT   exon            180..366
FT                   /gene="INS2"
FT                   /number=2
FT                   /note="first expressed exon"
FT   mat_peptide     252..341
FT                   /gene="INS2"
FT                   /product="beta chain"
FT   mat_peptide     join(348..366,541..614)
FT                   /gene="INS2"
FT                   /product="insulin 2 connecting peptide"
FT   intron          367..>410
FT                   /gene="INS2"
FT                   /number=2
FT   gap             411..510
FT                   /estimated_length=unknown
```

```
FT   intron          <511..540
FT                   /gene="INS2"
FT                   /number=2
FT   exon            541..739
FT                   /gene="INS2"
FT                   /number=3
FT   exon            541..>686
FT                   /gene="INS2"
FT                   /number=3
FT                   /note="preproinsulin 2"
FT   mat_peptide     621..683
FT                   /gene="INS2"
FT                   /product="insulin 2"
FT                   /note="alpha chain"
XX
SQ   Sequence 782 BP; 136 A; 212 C; 173 G; 161 T; 100 other;
     cccagcccta agtgaccagc tacagtcgga aaccatcagc aagcaggtat gtactctcca        60
     aggtgggcct agcttcccca gtcaagactc caaggatttg agggacgctg tgggctcttc       120
     tcttacatgt acctttttgct agcctcaacc ctgactatct tccaggtcat tgttccaaca       180
     tggccctgtg gatccgcttc ctgcccctgc tggccctgct catcctctgg gagccccgcc       240
     ctgcccaggc ttttgtcaaa cagcaccttt gtggttctca cttggtggaa gctctctacc       300
     tggtgtgtgg ggagcgtgga ttcttctaca cacccatgtc ccgccgcgaa gtggaggacc       360
     cacaaggtaa gctctgctcc tgaattctat cccaagtgct aactaccctg nnnnnnnnnn       420
     nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn       480
     nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn tgccctgtgc tgacatgacc tccctggcag       540
     tggcacaact ggagctgggt ggaggcccgg gggccggtga ccttcagacc ttggcactgg       600
     aggtggcccg gcagaagcgc ggcatcgtgg atcagtgctg caccagcatc tgctctctct       660
     accaactgga gaactactgc aactaggccc accactaccc tgtccacccc tctgcaatga       720
     ataaaacctt tgaaagagca ctacaagttg tgtgtacatg cgtgcatgtg catatgtggt       780
     gc                                                                     782
//
```

Sequence Retrieval System (SRS)

SRS (http://srs.ebi.ac.uk/) (Fig. 1.7) is a powerful searching tool to retrieve sequences (and other types of data) and also to perform various operations on retrieved information for EMBL. It is similar to Entrez of NCBI, a search engine for extracting all sort of information available at EMBL.

Sequence Submission at EMBL

There are mainly three tools available for submitting data at EMBL.

1. Webin: for nucleotide sequence submission
2. Sequin: a stand-alone tool for submitting nucleotide sequences to GenBank, EMBL, and DDBJ developed by NCBI
3. Webin-Align: a tool for sequence alignment submission

(C) *DNA Data Bank of Japan (DDBJ)*

DDBJ, (http://ddbj.sakura.ne.jp/) (Fig. 1.8) part of *INSDC*, was established at the National Institute of Genetics (NIG), Japan, in 1986 with the support of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

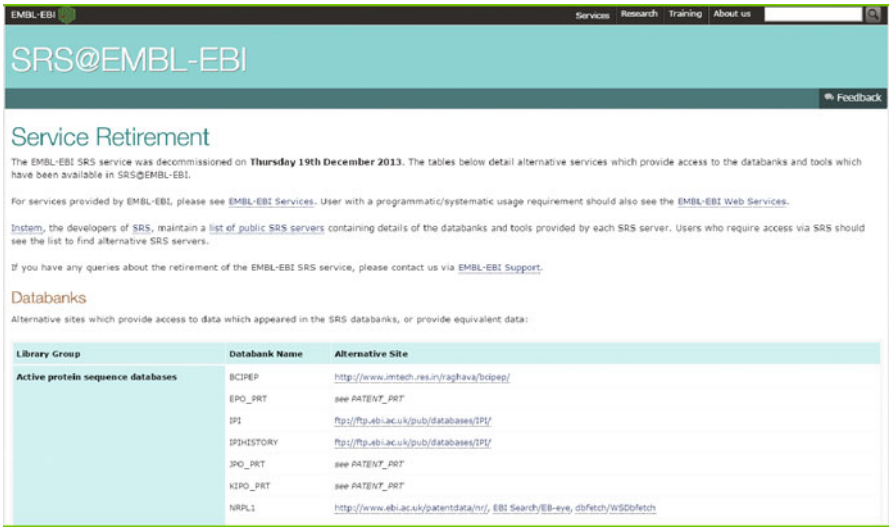


Fig. 1.7 The home page of Sequence Retrieval System (http://srs.ebi.ac.uk/)

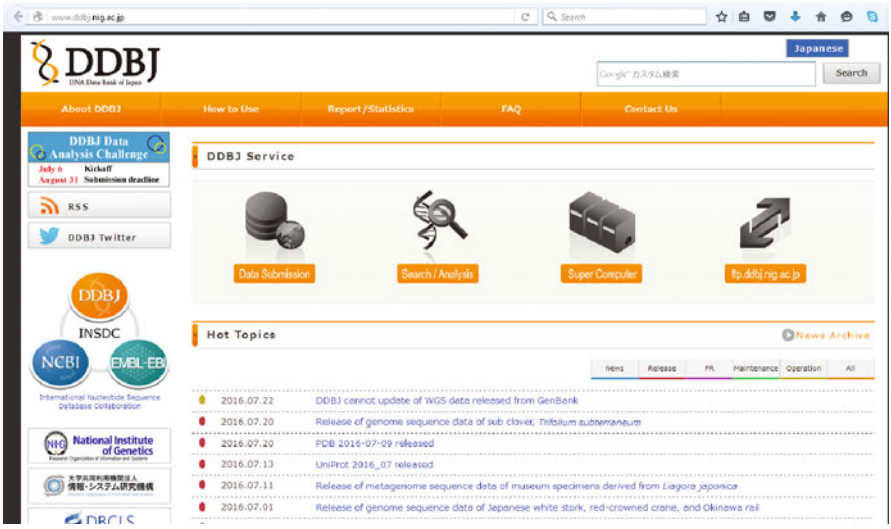


Fig. 1.8 The home page of DNA Data Bank of Japan (http://ddbj.sakura.ne.jp/)

SAKURA

SAKURA (http://sakura.ddbj.nig.ac.jp/top-e.html) is a source for data (nucleotide sequence) submission system through the WWW-based server where one can enter and submit nucleotide sequences and translated amino acid sequences. Since 1995 it is open to the public and scientists community.

DDBJ Format

```
LOCUS       OCOINS                   432 bp    mRNA    linear   ROD 27-APR-1993
DEFINITION  Octodon degus insulin mRNA, complete cds.
ACCESSION   M57671
VERSION     M57671.1
KEYWORDS    insulin; insulin alpha-chain; insulin beta-chain; insulin
            connecting peptide.
SOURCE      Octodon degus (degu)
  ORGANISM  Octodon degus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Hystricognathi; Octodontidae; Octodon.
REFERENCE   1  (bases 1 to 432)
  AUTHORS   Nishi,M. and Steiner,D.F.
  TITLE     Cloning of complementary DNAs encoding islet amyloid polypeptide,
            insulin, and glucagon precursors from a New World rodent, the degu,
            Octodon degus
  JOURNAL   Mol. Endocrinol. 4 (8), 1192-1198 (1990)
   PUBMED   2293024
COMMENT     Original source text: Octodon degus pancreas, cDNA to mRNA.
FEATURES             Location/Qualifiers
     source          1..432
                     /organism="Octodon degus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10160"
                     /tissue_type="pancreas"
     gene            1..432
                     /gene="insulin"
     CDS             42..371
                     /gene="insulin"
                     /codon_start=1
                     /product="insulin"
                     /protein_id="AAA40590.1"
                     /db_xref="GI:202472"
                     /translation="MAPWMHLLTVLALLALWGPNSVQAYSSQHLCGSNLVEALYMTCG
                     RSGFYRPHDRRELEDLQVEQAELGLEAGGLQPSALEMILQKRGIVDQCCNNICTFNQL
                     QNYCNVP"
     sig_peptide     42..113
                     /gene="insulin"
     mat_peptide     114..200
                     /gene="insulin"
                     /product="insulin B-chain"
     mat_peptide     207..293
                     /gene="insulin"
                     /product="insulin C-peptide"
     mat_peptide     300..368
                     /gene="insulin"
                     /product="insulin A-chain"
     regulatory      414..419
                     /regulatory_class="polyA_signal_sequence"
                     /gene="insulin"
     polyA_site      432
                     /gene="insulin"
BASE COUNT           86 a          134 c          119 g           93 t
ORIGIN
        1 gcattctgag gcattctcta acaggttctc gaccctccgc catggccccg tggatgcatc
       61 tcctcaccgt gctggccctg ctggccctct ggggacccaa ctctgttcag gcctattcca
      121 gccagcacct gtgcggctcc aacctagtgg aggcactgta catgacatgt ggacggagtg
      181 gcttctatag accccacgac cgccgagagc tggaggacct ccaggtggag caggcagaac
      241 tgggtctgga ggcaggcggc ctgcagcctt cggccctgga gatgattctg cagaagcgcg
      301 gcattgtgga tcagtgctgt aataacattt gcacatttaa ccagctgcag aactactgca
      361 atgtccctta gacacctgcc ttgggcctgg cctgctgctc tgccctggca accaataaac
      421 cccttgaatg ag
//
```

(II) Protein Sequence Databases

The different protein sequence databases available are the following:

(A) Protein Information Resource
(B) UniProt

(A) *Protein Information Resource (PIR)*

Margaret Dayhoff was the inventor of Protein Information Resource (PIR) in the 1960s at the National Biomedical Research Foundation (NBRF) for investigation of evolutionary relationships among proteins. Analysis tools for protein database are provided by PIR which are freely available to the scientists (George et al. 1997).

In 2002 Protein Information Resource and its worldwide partners, EBI and Swiss Institute of Bioinformatics (SIB), were granted an award from the National Institutes of Health (NIH) to make UniProt, by merging the databases of PIR-PSD, SWISS-PROT, and TrEMBL (Fig. 1.9).

(B) UniProt

It comprises of two sections:

(a) SWISS-PROT
(b) Translated EMBL (TrEMBL)

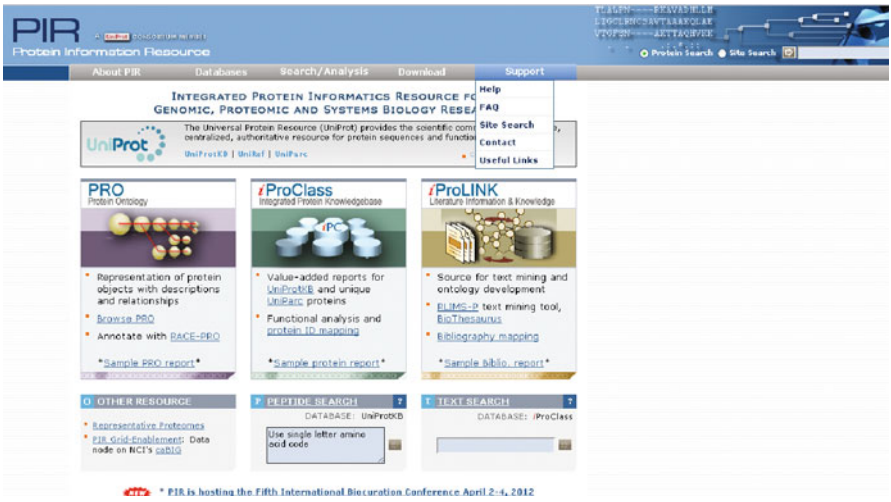


Fig. 1.9 The home page of Protein Information Resource (http://pir.georgetown.edu/)

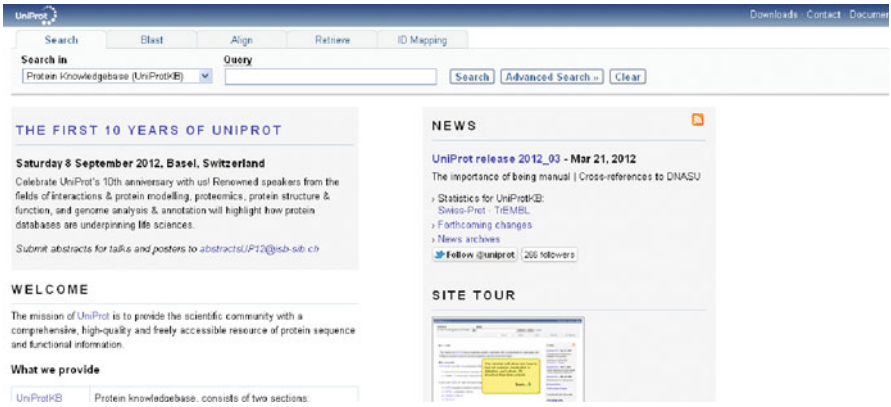


Fig. 1.10 The home page of UniProt (http://www.uniprot.org/)

(a) *SWISS-PROT*

SWISS-PROT (http://www.uniprot.org/) (Fig. 1.10), established in 1896, is the most widely used protein sequence database created by the University of Geneva and the EMBL, collaboratively. After 1994, the collaboration moved to EMBL's UK outstation, the EBI.

SWISS-PROT Format

Each line starts with a two-character line code, which specifies the kind of data contained in the line.

(b) *Translated EMBL*

TrEMBL benefits from the SWISS-PROT format and comprises translations of all coding sequences (CDS) in EMBL. It has two core divisions, designated SWISS-PROT-TrEMBL and REM-TrEMBL.

1.3.1.2 Structure Databases

- PDB (Protein Data Bank)
- MMDB (Molecular Modeling Database)
- VAST (Vector Alignment Search Tool)
- CDD (Conserved Domain Database)
- NDB (Nucleic acid Structure Database)

From the above databases, some of the database is shown below in detail.

(I) Protein Data Bank (PDB)

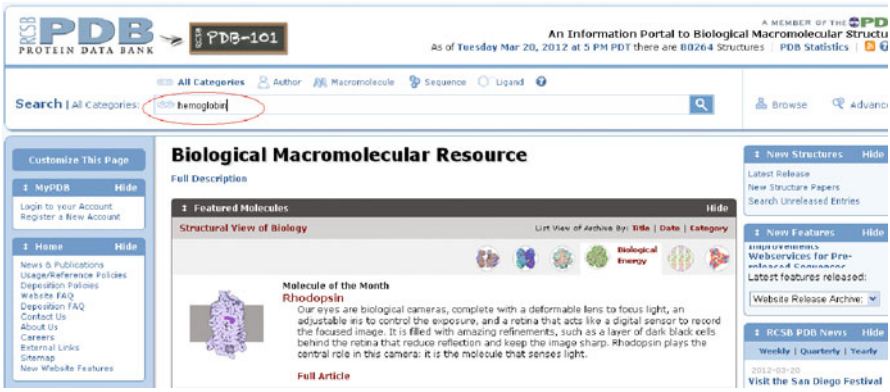


Fig. 1.11 The home page of PDB with the query Hemoglobin (http://www.rcsb.org/pdb/home/home.do)

The PDB (http://www.rcsb.org/pdb/home/home.do) in Fig. 1.11, a source for the three-dimensional structural data of huge biological molecules, includes proteins and nucleic acids. It was established in 1971 by the Research Collaborators for Structural Bioinformatics (RCSB). The data submitted by scientists from different parts of the world are easily without cost available through the Internet. The PDB is supervised by the Worldwide Protein Data Bank (wwPDB) (Berman 2008).

As on March 20, 2012 at 5 PM PDT, there were 80,264 structures. Each structure has been assigned a PDB ID, which contains four characters both alphabets and numerical. The first character is a numeral, while the last three characters can be either numerals or letters. Search results and structure for hemoglobin were showed in Figs. 1.11 and 1.12.

PDB File Format

This format was primarily practiced by the Protein data bank and previously was known as the PDB file format. The PDB also retains data on biological macromolecules, "macromolecular crystallographic information file format" (mmCIF), initiated to be phased in 1996. In the year 2005, an Extensible Markup Language (XML) version of PDBML was described (Westbrook et al. 2005).

Data Deposition Tool of PDB

Auto Dep Input Tool (ADIT) (http://deposit.rcsb.org/adit/) (Fig. 1.13) is developed by RCSB, and it is responsible for depositing structures to PDB in an efficient manner.

(II) *Nucleic Acid Structure Database (NDB)*

Fig. 1.12 Search result of Protein Data Bank (http://www.rcsb.org/pdb/results/results.do?qrid=57082E24&tabtoshow=Current)

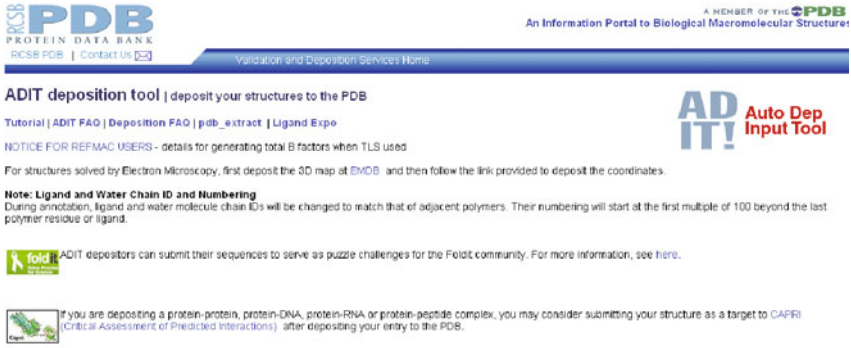


Fig. 1.13 The home page of Auto Dep Input Tool (http://deposit.rcsb.org/adit/)

This database (http://ndbserver.rutgers.edu/) (Fig. 1.14) provides us 3D structures of nucleic acids.

1.3.1.3 Literature Database

Literature databases provide us library of life science work done all over the world. Various literature databases available are the following:

- MEDLINE
- CiteXplore
- OMIM
- Patent abstracts
- FlyBase archives

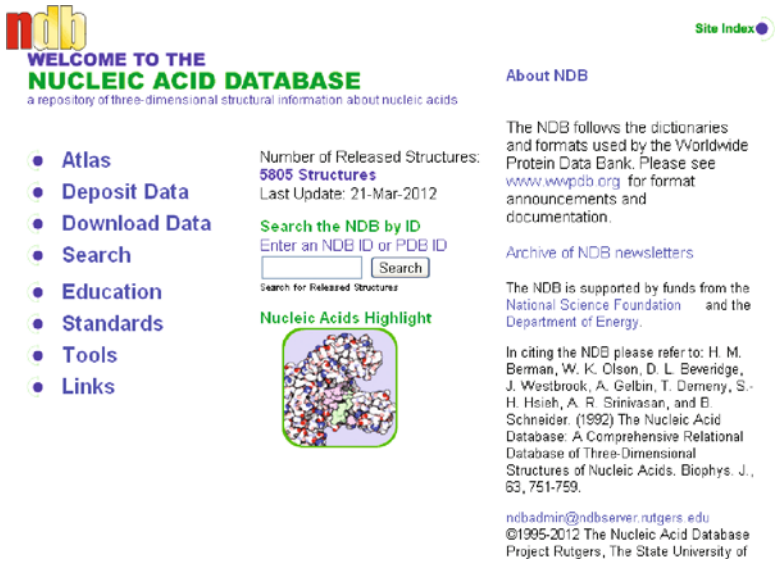


Fig. 1.14 The home page of nucleic acid structure database (http://ndbserver.rutgers.edu/)

1.3.1.4 Pathway Database

To comprehend molecular interactions and chemical reaction networks, the pathway database is used by pathway maps. Various pathway databases available are the following:

- BioCyc database collection comprising EcoCyc and MetaCyc
- KEGG PATHWAY Database (www.genome.jp/kegg/)
- MANET database
- Reactome (Laboratory of Cold Spring Harbor, EBI, Gene Ontology Consortium)

1.3.1.5 Chemical Database

A collection of the chemical information precisely planned is called chemical database. These are the few freely available chemical databases:

- Chemical Entities of Biological Interest (ChEBI)
- PubChem
- Zinc
- eMolecules
- DrugBank

1.3.1.6 Enzyme Database

Enzyme databases cover an extensive range of properties and functions, such as structure, occurrence, kinetics of enzyme-catalyzed reactions, and metabolic function. Various enzyme databases available are the following:

- ExPASy
- BRENDA
- REBASE
- EC enzyme database

1.3.1.7 Disease Database

The disease database provides all disease-related information; it is a cross-referenced index of diseases, symptoms, medications, signs, abnormal investigation findings, etc.

- OMIM
- OMIA

1.3.1.8 Domain Database

Domain database is a database for ancient domains and full-length proteins.

- CDD (Conserved Domain Database)

1.3.1.9 Structural Classification of Protein Database

It provides hierarchical classification of protein structure which defines the evolutionary association between proteins.

- The Structural Classification of Proteins (SCOP) (http://scop.mrclmb.cam.ac.uk/scop/).
- Class, architecture, topology, and homologous superfamily (CATH) is freely available to scientists (www.cathdb.info/).

1.3.1.10 Genome Database

Genome databases are a collection of genome sequences of many species; it interprets and examines them and provides free public access.

- Genome Databases at the National Center for Biotechnology Information (Index)
- Genome Databases at the National Center for Biotechnology Information (Entrez)
- Genome Databases at the National Center for Biotechnology Information (PMGif) Genome List in NIH

- Mitochondrial DNA Database (MitBASE)
- Mouse Genome Informatics
- Plant Genome Project maintained by the National Science Foundation
- Organelle Genome Sequences (PMGif)

1.3.2 Biological Databases Based on Database Source

This database is subdivided into two databases, primary and secondary.

1. *Primary*: databases comprising of data generated experimentally like nucleotide sequences and 3D structures are identified as primary databases.

   Examples are GenBank, DDBJ, EMBL, PIR, PDB, NDB, UniProt, TrEMBL, SWISS-PROT, etc.

2. *Secondary*: it contains databases directly derived from the primary databases.

   Examples are PROSITE, Pfam, Blocks, Prints, SCOP, CATH, OMIM, KEGG, etc.

1.3.3 Composite Databases

It combines various different primary database sources. This makes searching the query more efficient. So, composite database amalgamates various primary databases for easy access.

Examples are OWL, NRDB, MIPSX, SP, and TrEMBL.

1.3.4 Biological Databases Based on Database Design

This database is subdivided into two databases, object-oriented and relational databases.

1.3.4.1 Object Oriented

A database controlling system in which information is characterized in the form of objects. These databases are unlike table-oriented relational databases.

Objects mostly comprise of Attributes and Methods.

How Data Is Stored

There are two methods used for the storage of objects:

- Each object has an exclusive ID and is known as a subclass of a base class, by inheritance to explain attributes.
- For management and object storage, virtual memory mapping has been used.

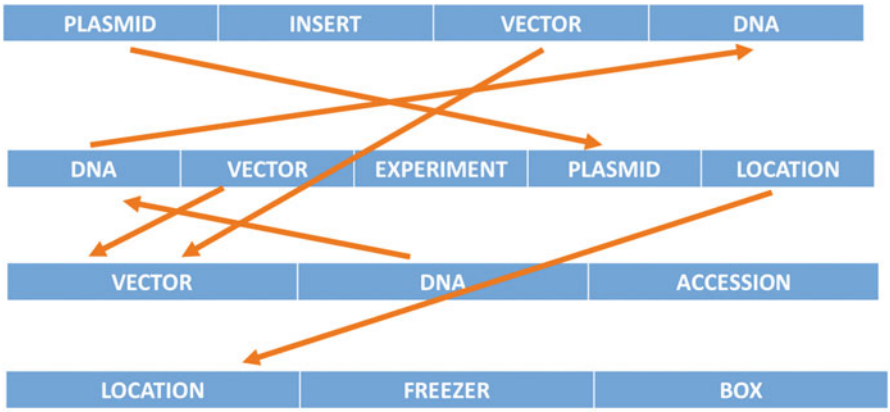


Fig. 1.15 Four tables are shown: plasmid, vector, DNA, and location. Arenas that reference other tables are mentioned to as links. Numerous factors have to be considered when designing a relational database (http://home.cc.umanitoba.ca/)

1.3.4.2 Relational Database

Relational databases can be assumed as comprehensive tables of data. Each record from a flat file could be applied as a row in a table. Although a relational database can be applied in a single large table or "relation," it is often helpful to split the database up into multiple tables (Fig. 1.15).

A benefit of relational databases is that by breaking up the database to various tables, in many circumstances, only one table needs to be rewritten when creating changes in fields. In other cases, addition of a record may need rewriting many or most tables.

References

Benton D (1990) Recent changes in the GenBank on-line service. Nucleic Acids Res 18 (6):1517–1520

Berman HM (2008) The protein data bank: a historical perspective. Acta Crystallogr A64:88–95

Dayhoff MO, N. B. R. Foundation (1973) Atlas of protein sequence and structure: supplement. National Biomedical Research Foundation

Dayhoff MO, N. B. R. Foundation (1976) Atlas of protein sequence and structure. National Biomedical Research Foundation

Foundation N. B. R. (1972) Atlas of protein sequence and structure. National Biomedical Research Foundation

George DG et al (1997) The protein information resource (PIR) and the PIR-International protein sequence database. Nucleic Acids Res 25(1):24–28

Liu L, Özsu MT (2009) Encyclopedia of database systems. Springer US

N. C. f. B. I (2013) The NCBI handbook. In: Mizrachi I (ed) NCBI handbook [Internet], 2nd edn. National Center for Biotechnology Information (US), Bethesda

Westbrook J et al (2005) PDBML: the representation of archival macromolecular structure data in XML. Bioinformatics 21(7):988–992