CHAPTER 1
Proteins

Proteins perform a variety of functions, including enzymatic catalysis, transporting ions and molecules from one organ to another, nutrients, contractile system of muscles, tendons, cartilage, antibodies, and regulating cellular and physiological activities. The functional properties of proteins depend on their three-dimensional structures. The native structure of a protein can be experimentally determined using X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, electron microscopy, etc. Over the past 40 years, the structures of more than 53,000 proteins (as of May 12, 2009) have been determined. On the other hand, the amino acid sequences are determined for more than eight million proteins (as of May 5, 2009). The specific sequence of amino acids in a polypeptide chain folds to generate compact domains with a particular three-dimensional structure. Anfinsen (1973) stated that the polypeptide chain itself contains all the information necessary to specify its three-dimensional structure. Deciphering the three-dimensional structure of a protein from its amino acid sequence is a long-standing goal in molecular and computational biology.

1.1 Building blocks

Protein sequences consist of 20 different kinds of chemical compounds, known as amino acids, and they serve as building blocks of proteins. Amino acids contain a central carbon atom (C_α), which is attached to a hydrogen atom, an amino group (NH_2), and a carboxyl group (COOH) as shown in **Figure 1.1.** The letter R in **Figure 1.1** indicates the presence of a side chain, which distinguishes each amino acid.

1.1.1 Amino acids

Amino acids are naturally of 20 different types as specified by the genetic code emerged from DNA sequences. Furthermore, nonnatural amino acids occur, in rare cases, as the products of enzymatic modifications after translocation. The major difference among the 20 amino acids is the *side chain* attached to the C_α through its fourth valance. The variation of side chains in 20 amino acids is shown in **Figure 1.2**. These residues are represented by conventional three- and one-letter codes. Most of the databases use single-letter codes.

The amino acids are broadly divided into two groups, hydrophobic and hydrophilic, based on the tendency of their interactions in the presence of water

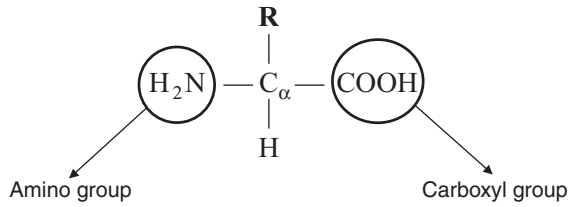


Figure 1.1 Representation of amino acids. R is the side chain that varies for the 20 amino acids.

molecule. The hydrophobic residues have the tendency of adhering to one another in aqueous environment. Generally, amino acids, Ala (A), Cys (C), Phe (F), Gly (G), Ile (I), Leu (L), Met (M), Val (V), Trp (W), and Tyr (Y), are considered as hydrophobic residues. In this category, Ala, Ile, Leu, and Val contain aliphatic side chains; Phe, Trp, and Tyr contain aromatic side chains; and Cys and Met contain sulfur atom. Gly has no side chain, and it has hydrogen (H) at the fourth position. Two Cys residues in different parts of the polypeptide chain but adjacent to each other in the three-dimensional structure of a protein can be oxidized to form a disulfide bridge. The

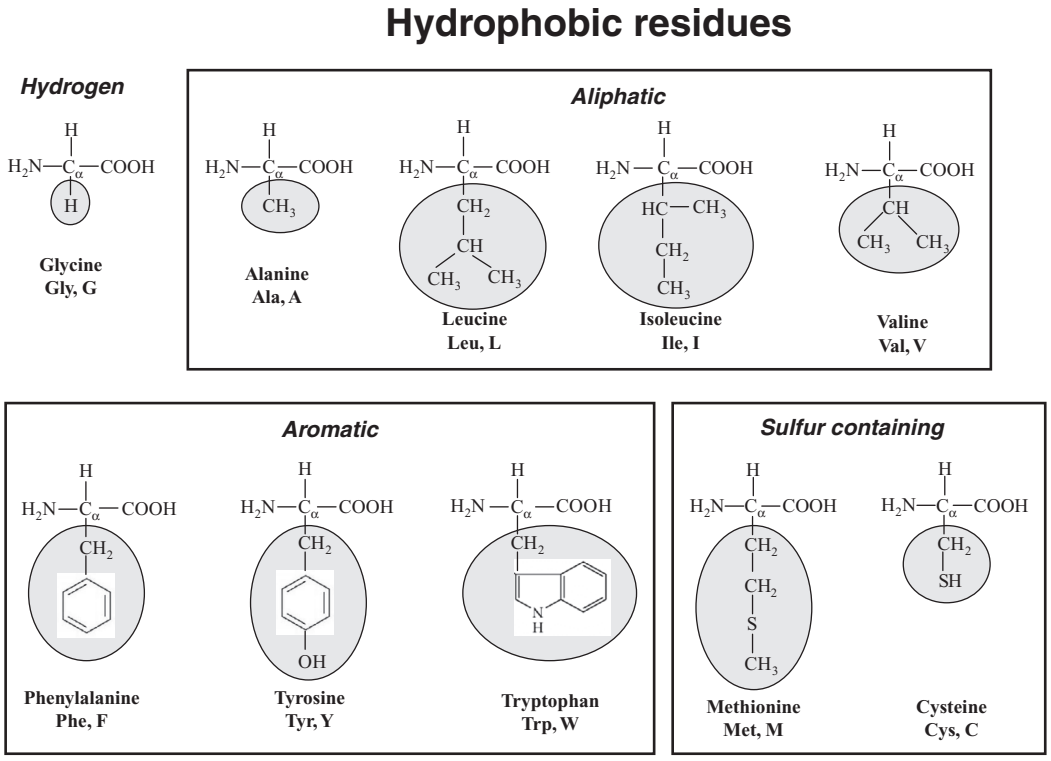


Figure 1.2 The common 20 amino acids in proteins. The three- and one-letter codes for the amino acids are also given. The amino acids are classified into hydrophobic (hydrogen, aliphatic, aromatic, and sulfur containing) and hydrophilic (negatively charged, positively charged, and polar). The side chains are marked with oval boxes.

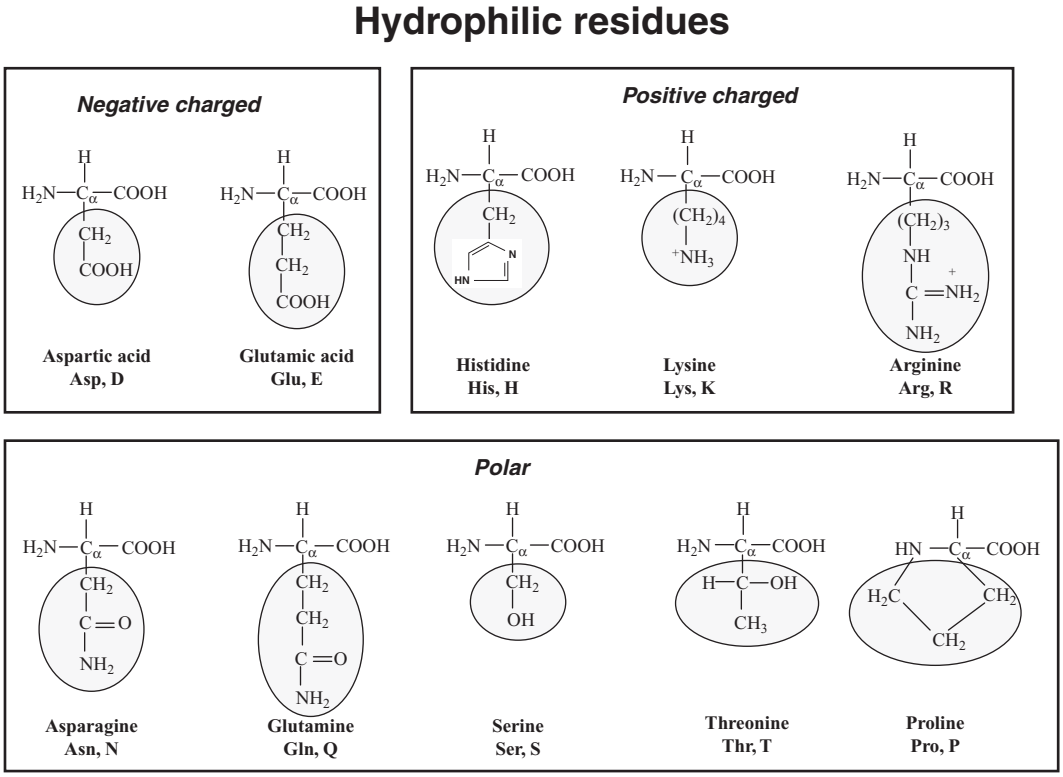


FIGURE 1.2 (*Continued*)

formation of disulfide bridges in protein structures stabilizes the protein, making it less susceptible to degradation.

Amino acids, Asp (D), Glu (E), His (H), Lys (K), Asn (N), Pro (P), Gln (Q), Arg (R), Ser (S), and Thr (T), are classified as hydrophilic residues. In this category, Asp and Glu are negatively charged; His, Lys, and Arg are positively charged; and others are polar and uncharged.

1.1.2 Formation of peptide bonds

The carboxyl group of one amino acid interacts with the amino group of another to form a peptide bond by the elimination of water (**Figure 1.3**). Amino acids are joined end-to-end during protein synthesis by the formation of such peptide bonds. The peptide bond (C—N) has a partial double-bond character due to resonance, and hence there is no rotation about the peptide bond. In **Figure 1.3,** the peptide is represented as a planar unit with the C═O and N—H groups positioning in opposite directions in the plane. This is called *trans*-peptide. There is another form, *cis*-peptide in which the C═O and N—H groups point in the same direction. To avoid steric hindrance, the *trans* form is frequently presented in protein structures for all amino acids except Pro, which has both *trans* and *cis* forms. The *cis* prolines are found in bends of the polypeptide chains.

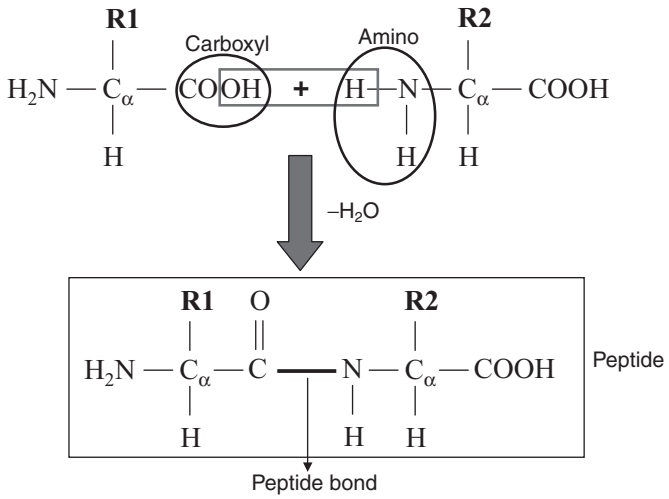


Figure 1.3 Formation of a peptide bond by the elimination of a water molecule.

A protein chain is formed by several amino acids in which the amino group of the first amino acid and the carboxyl group of the last amino acid remain intact, and the chain is said to extend from the amino (N) to the carboxyl (C) terminus. This chain of amino acids is called a polypeptide chain, main chain, or backbone. Amino acids in a polypeptide chain lack a hydrogen atom at the amino terminal and an OH group at the carboxyl terminal (except at the ends), and hence amino acids are also called **amino acid residues** (simply residues). Nature selects the combination of amino acid residues to form polypeptide chains for their function, similar to the combination of alphabets to form meaningful words and sentences. These polypeptide chains that have specific functions are called **proteins**.

1.2 Hierarchical representation of proteins

Depending on their complexity, protein molecules may be described by four levels of structure (Nelson and Cox, 2005): primary, secondary, tertiary, and quaternary (**Figure 1.4**). Because of the advancements in the understanding of protein structures, two additional levels such as supersecondary and domain have been proposed between secondary and tertiary structures. A stable clustering of several elements of secondary structures is referred to as a supersecondary structure. A somewhat higher level of structure is the domain, which refers to a compact region and distinct structural unit within a large polypeptide chain.

1.2.1 Primary structure

Primary structure describes the linear sequence of amino acid residues in a protein. It includes all the covalent bonds between amino acids. The relative spatial arrangement of the linked amino acids is unspecified.

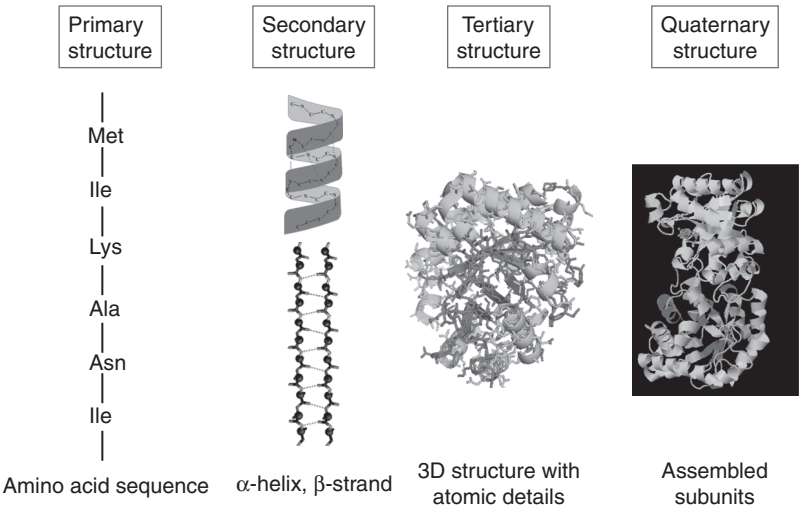


Figure 1.4 Structural organization of proteins.

1.2.2 Secondary structure

Secondary structure refers to regular, recurring arrangements in space of adjacent amino acid residues in a polypeptide chain. It is maintained by hydrogen bonds between amide hydrogens and carbonyl oxygens of the peptide backbone. The major secondary structures are α-helices and β-structures.

The α-helical conformation was first proposed by Linus Pauling and co-workers (1951), and a typical α-helix is shown in **Figure 1.5**. In this structure, the polypeptide backbone is tightly wound around the long axis of the molecule, and R groups of the amino acid residues protrude outward from the helical backbone. The repeating

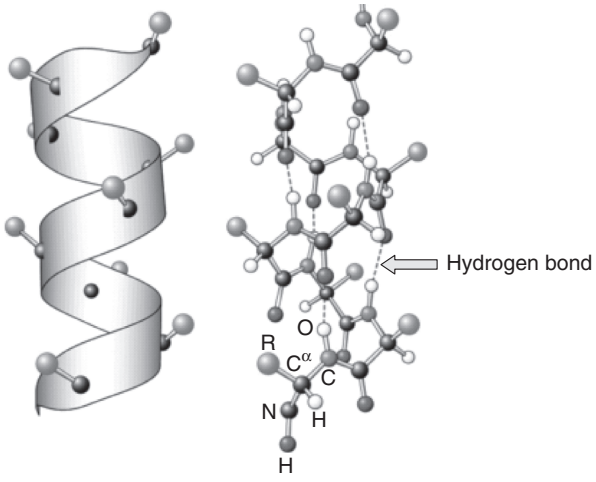


Figure 1.5 Structure of a typical α-helix. The hydrogen bonds between the residues n and $n + 4$ are shown as dotted lines. Figure was taken as a screenshot from the Web, http://www.food-info.net/uk/protein/structure.htm

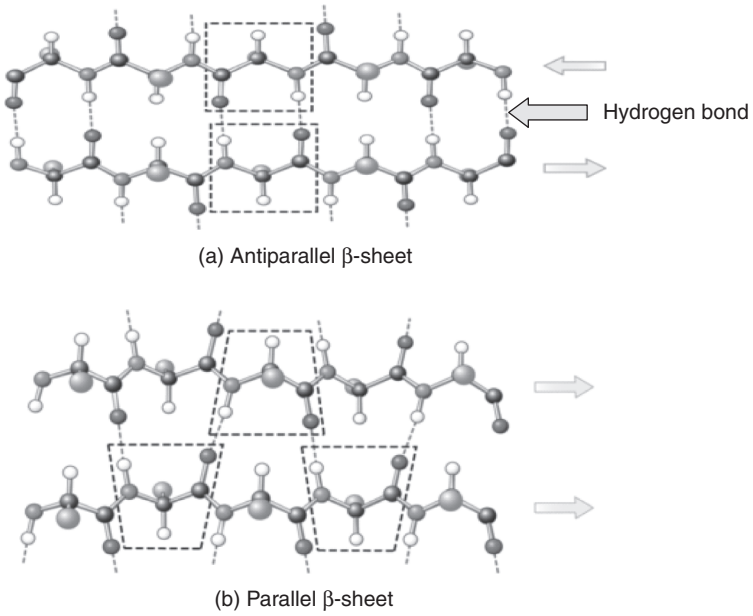


Figure 1.6 Structures of (a) antiparallel and (b) parallel β-sheets. The dotted lines show the hydrogen bonds between amino acid residues. The arrows indicate the directions of the polypeptide chain, from N- to C-terminal. Figure was taken as a screenshot from the Web, http://www.food-info.net/uk/protein/structure.htm.

unit is a single turn of a helix, which extends about 0.54 nm along the axis, and the number of amino acid residues required for one complete turn is 3.6. In an α-helix, each carbonyl oxygen (residue, n) of the polypeptide backbone is hydrogen bonded to the backbone amide hydrogen of the fourth residue further toward the C-terminus (residue, $n + 4$). The hydrogen bonds, which stabilize the helix, are nearly parallel to the long axis of the helix.

The other common secondary structure is β-structure that includes β-strands and β-sheets. β-strands are portions of the polypeptide chain that are almost fully extended, and several β-strands constitute β-sheets. β-sheets are stabilized by hydrogen bonds between carbonyl oxygens and amide hydrogens on adjacent β-strands (**Figure 1.6**). In β-sheets, the hydrogen bonds are nearly perpendicular to the extended polypeptide chains. The β-strands may be either parallel (running in the same N- to C-terminal) or antiparrallel (running in opposite N- to C-terminal directions).

In a polypeptide chain, the α-carbon atoms of adjacent amino acids are separated by three covalent bonds arranged as C_α—C—N—C_α. In these bonds, rotation is permitted about the N—C_α and C_α—C bonds, and the torsional angles are conventionally denoted as ϕ and ψ, respectively. Every secondary structure is described completely by these two torsional angles that are repeated at each residue. The allowed values for ϕ and ψ can be shown graphically by simply plotting these values known as Ramachandran plot (Ramachandran et al. 1963). **Figure 1.7** shows the conformations that are permitted for most amino acid residues in Ramachandran plot.

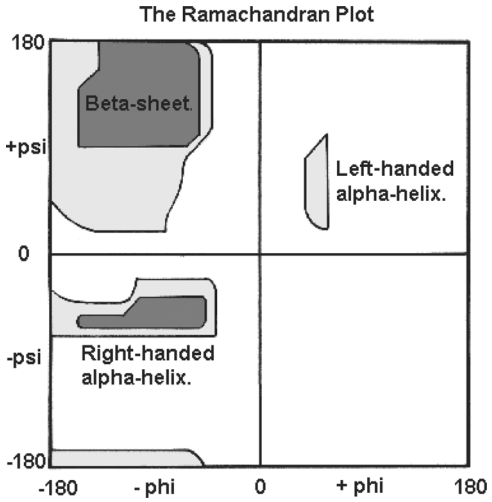


Figure 1.7 Ramachandran plot showing the allowed regions of α-helical and β-strand conformations. Figure was taken as a screenshot from the Web, http://swissmodel.expasy.org/course/text/chapter1.htm.

1.2.3 Tertiary structure

Tertiary structure refers to the spatial relationship among all amino acids in a polypeptide; it is the complete three-dimensional structure of the polypeptide with atomic details. Tertiary structures are stabilized by interactions of side chains of nonneighboring amino acid residues and primarily by noncovalent interactions. The formation of tertiary structure brings the amino acid residues that are far apart in the primary structure close together.

1.2.4 Quaternary structure

Quaternary structure refers to the spatial relationship of the polypeptides or subunits within the protein. It is the association of two or more polypeptide chains into a multisubunit or oligomeric protein. The polypeptide chains of an oligomeric protein may be identical or different. The quaternary structure also includes the cofactor and other metals, which form the catalytic unit and functional proteins.

1.3 Structural classification of proteins

Proteins are broadly classified into two major groups: fibrous proteins, having polypeptide chains arranged in long strands, and globular proteins, with polypeptide chains folded into a spherical or globular shape.

1.3.1 Fibrous proteins

Fibrous proteins are usually static molecules and play important structural roles in the anatomy and physiology of vertebrates, providing external protection, support, shape, and form. They are water insoluble and are typically built upon a single, repetitive structure assembled into cables or threads. Examples of fibrous proteins

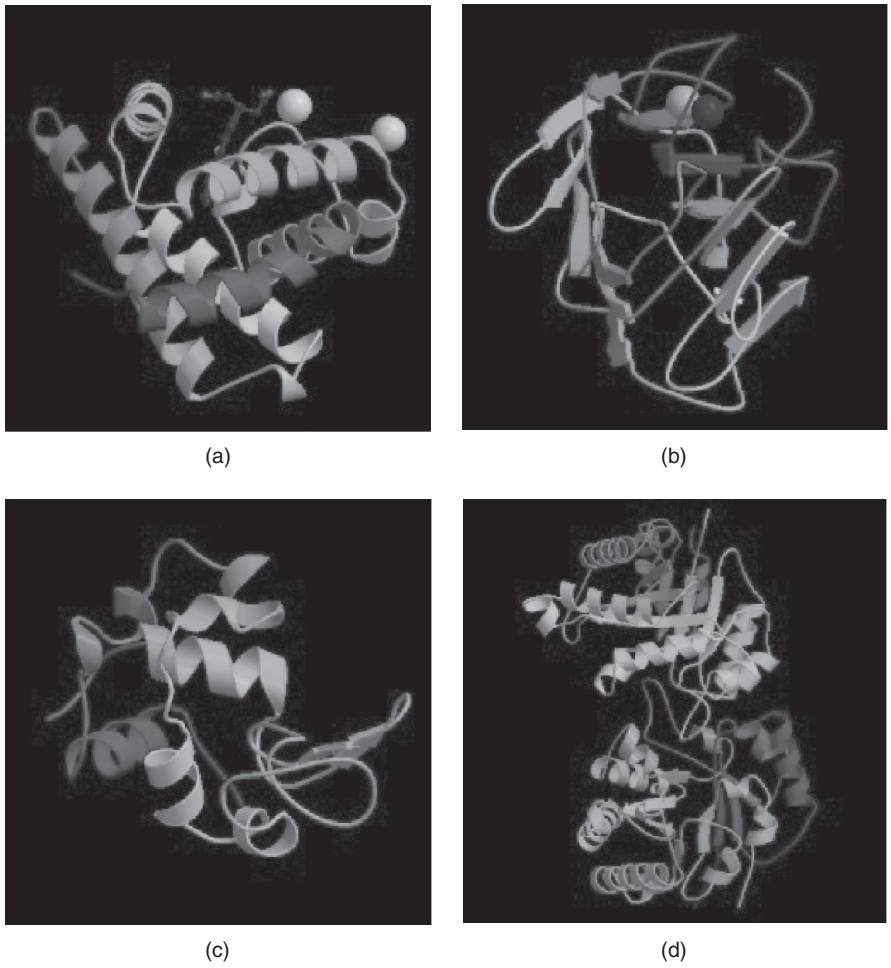


Figure 1.8 Ribbon diagram for four typical protein structures in different structural classes (a) all-α (4MBN), (b) all-β (3CNA), (c) α + β (4LYZ), and (d) α/β (1TIM). Figure was adapted from Gromiha and Selvaraj (2004).

are α-keratin, the major component of hair and nails, and collagen, the major protein component of tendons, skin, bones, and teeth.

1.3.2 Classification of globular proteins

Globular proteins are categorized into four structural classes: all-α, all-β, α + β, and α/β (Levitt and Chothia, 1976). The ribbon diagrams illustrating the structures in each class are shown in **Figure 1.8**.

The all-α and all-β classes are dominated by α-helices (α > 40% and β < 5%) and by β-strands (β > 40% and α < 5%), respectively (**Figures 1.8a** and **b**). The α + β class contains both α-helices (>15%) and antiparallel β-strands (>10%) that do not mix but tend to segregate along the polypeptide chain (**Figure 1.8c**). The α/β class proteins (**Figure 1.8d**) have mixed or approximately alternating segments of α-helical (>15%) and parallel β-strands (>10%).

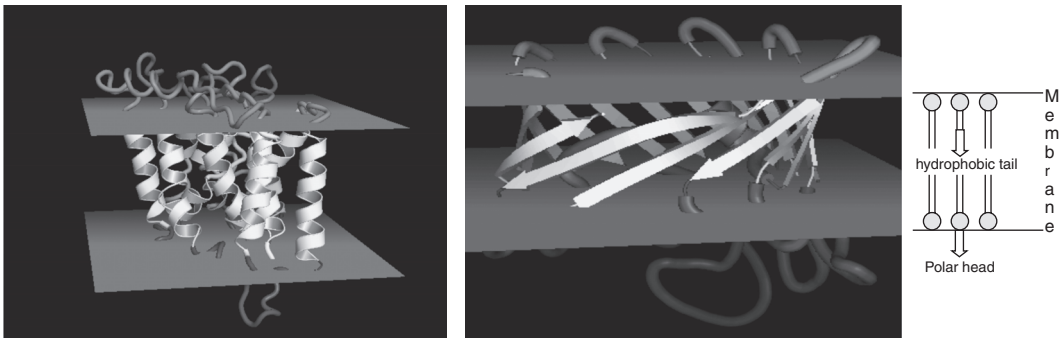


Figure 1.9 Representation of (a) α-helical and (b) β-barrel membrane proteins. The membrane spanning regions are shown within the disc. Protein structures were taken from Protein Data Bank of Transmembrane Proteins (http://pdbtm.enzim.hu/).

1.3.3 Membrane proteins

Membrane proteins, which require embedding into the lipid bilayers, have evolved to have amino acid sequences that will fold with a hydrophobic surface in contact with the alkane chains of the lipids and polar surface in contact with the aqueous phases on both sides of the membrane and the polar head groups of the lipids (**Figure 1.9**). In genomes, 30% of the proteins are suggested to be membrane proteins, and most of the transmembrane helical and strand proteins are identified as targets for drug design. Membrane proteins perform a variety of functions, including cell–cell signaling and mediating the transport of ions and solutes across the membrane. They are of two kinds: (i) transmembrane helical proteins in which they span the cytoplasmic membrane with α-helices (White and Wimley, 1999) and (ii) transmembrane β-barrel proteins that traverse the outer membranes of gram-negative bacteria with β-strands (Schulz, 2003). **Figure 1.9** shows the structures of membrane proteins with these two different motifs, α-helices and β-strands.

1.4 Databases for protein sequences

Recombinant DNA techniques have provided tools for the rapid determination of DNA sequences and, by inference, the amino acid sequences of proteins from structural genes. The number of such sequences is increasing exponentially, and these sequences have been deposited in the form of database, generally, known as protein sequence databases. Specifically, Georgetown University, Washington, D.C., USA, developed the database, Protein Information Resource (PIR). The Swiss Institute of Bioinformatics and European Bioinformatics Institute developed SWISS-PROT and TrEMBL databases. Recently, progress has been made to set up a single worldwide database of protein sequence and function, UniProt, by unifying PIR, SWISS-PROT, and TrEMBL database activities.

1.4.1 Protein Information Resource

PIR has evolved from the Atlas of Protein Sequence and Structure established in the early 1960s by Margaret O. Dayhoff (Dayhoff et al. 1965). It produces the largest, most comprehensive, annotated protein sequence database in the public

domain, the PIR International Protein Sequence Database, in collaboration with the Munich Information Center for Protein Sequences and the Japan International Protein Sequence Database (Barker et al. 2000). It is freely available at http://pir.georgetown.edu/. PIR offers a wide variety of resources mainly oriented to assist the propagation and standardization of protein annotation on three major aspects: (i) PIRSF, protein family classification system; (ii) iProClass, integrated protein knowledgebase; and (iii) iProLink, literature, information, and knowledge. The iProClass database provides value-added information reports on protein sequences, structures, families, functions, interactions, expressions, and modifications. The sequence information of a specific protein can be searched with simple "Text search" in iProClass **(Figure 1.10a)**. The search yielded 10 proteins, and the correct one has been selected with a click on the left-side box. It is also possible to save the results as a table or in FASTA format. The result obtained for the search with "Human lysozyme" is shown in **Figure 1.10b**. It includes general information

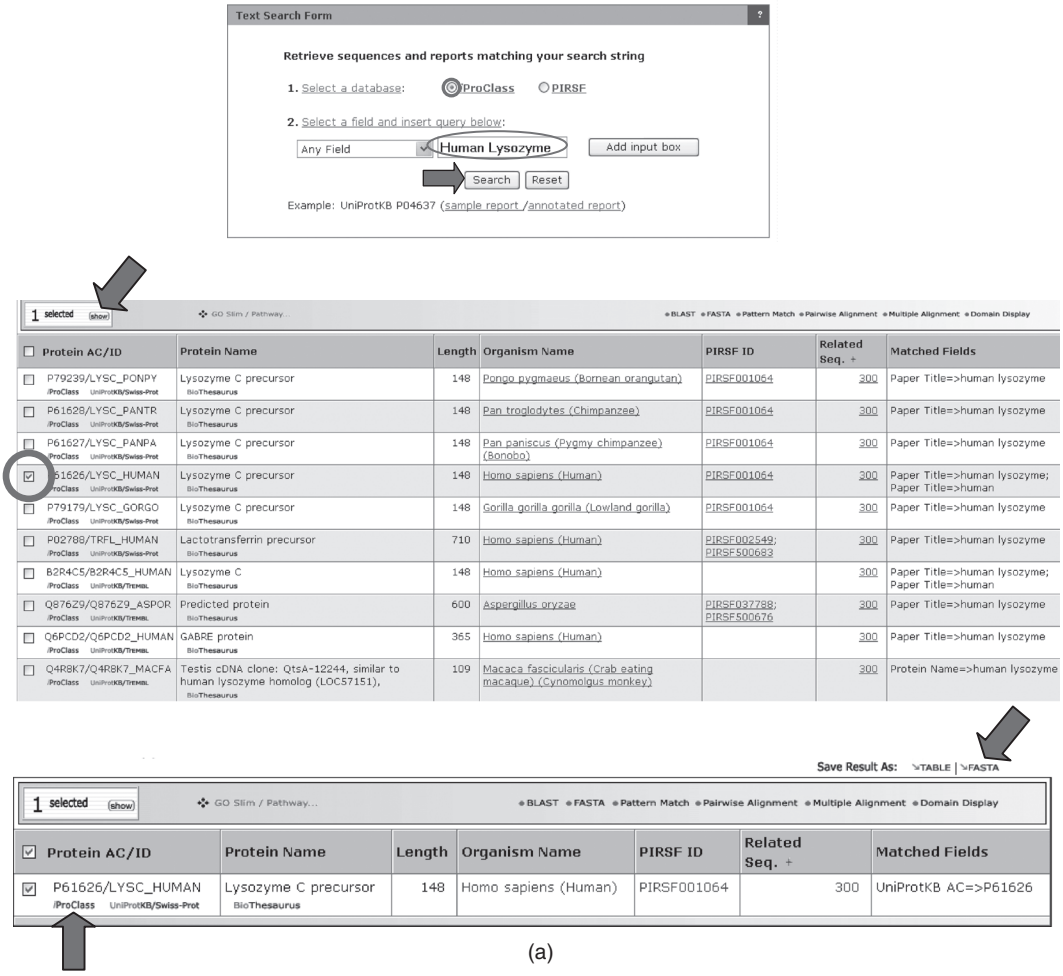
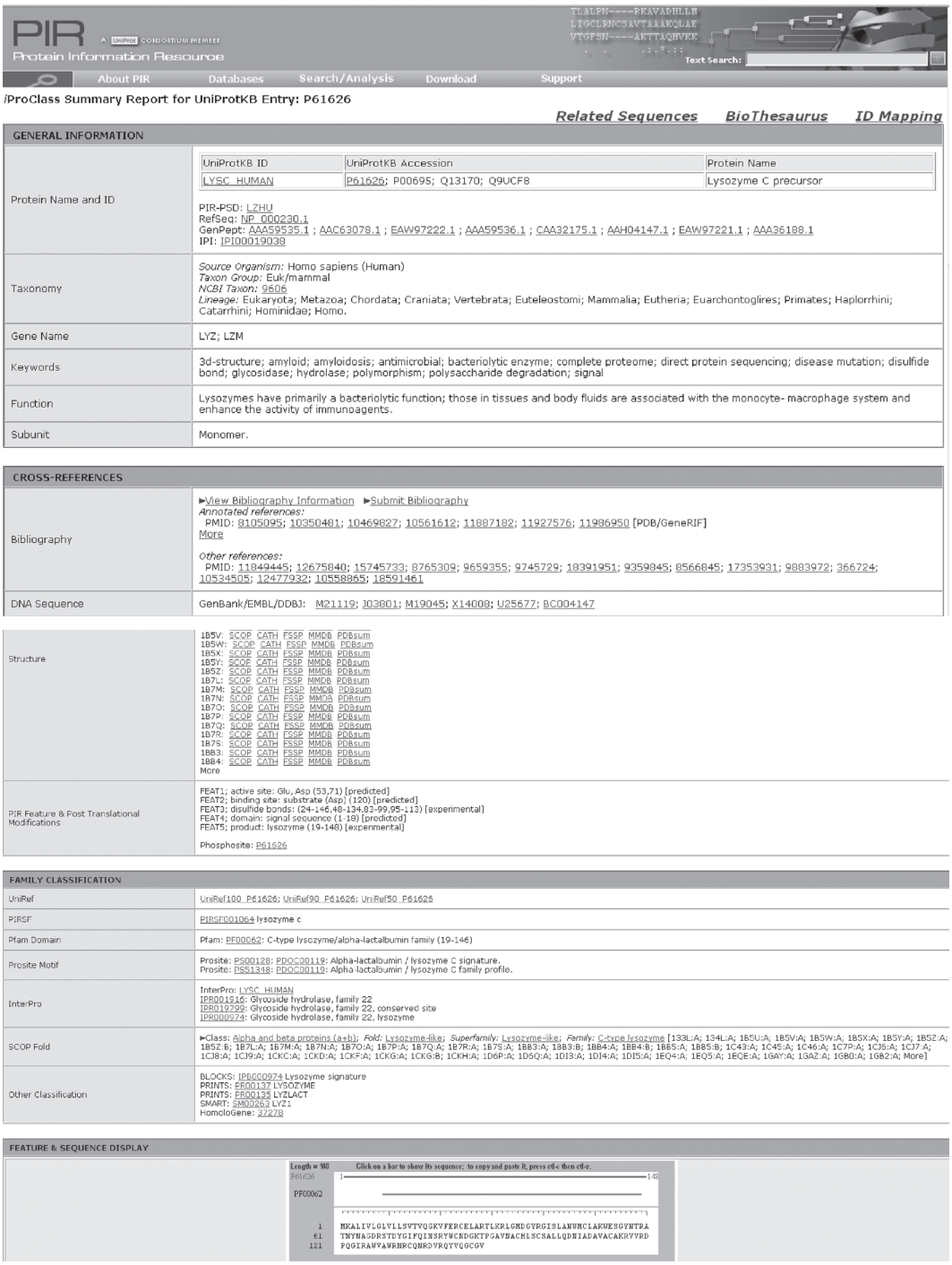


Figure 1.10 Text search in iProClass of PIR database: (a) the search with "Human lysozyme," along with intermediate steps, and (b) the information provided at the result page are shown.



(b)

FIGURE 1.10 (*Continued*)

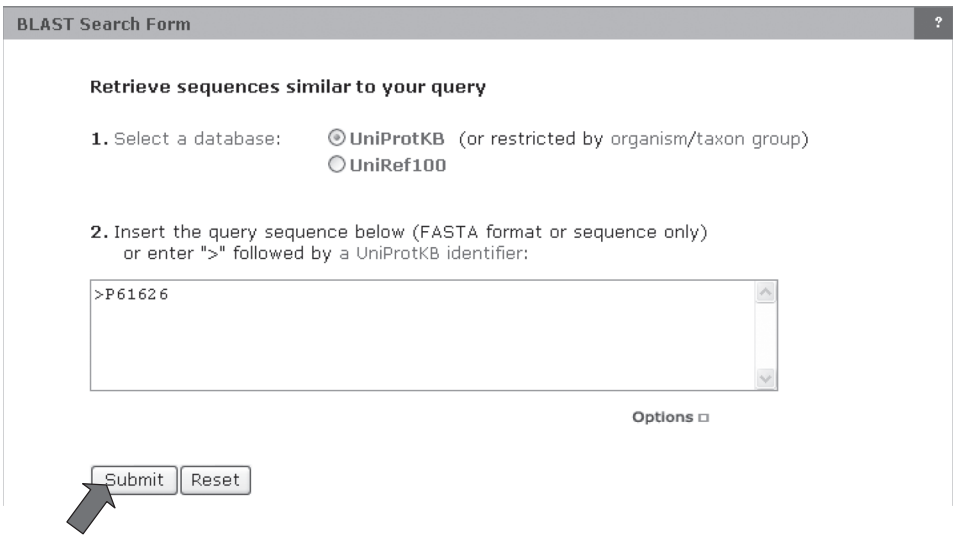


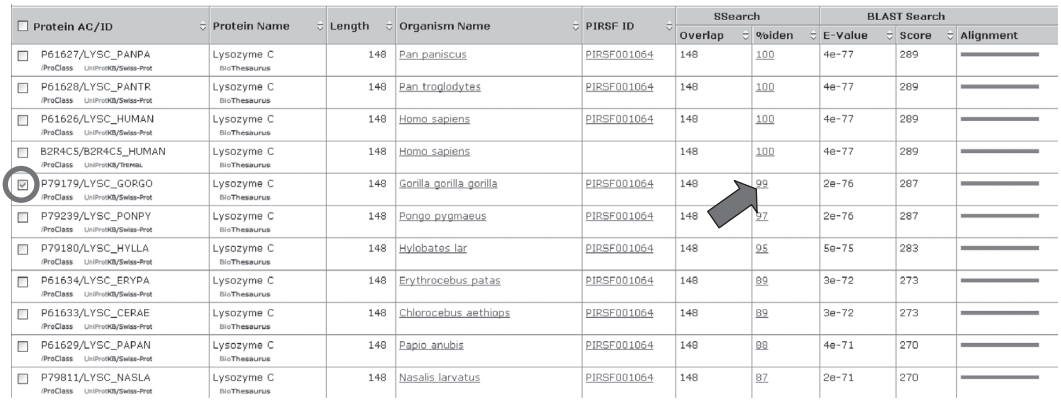
Figure 1.11 Utility of similar search option available in PIR. UniprotKB identifier for human lysozyme, ">P61626" is given as input.

(protein name, taxonomy, gene name, keywords, function, and subunit), cross-references (bibliography, DNA sequence, genome, ontology, function, interaction, structure, and posttranslational modifications), family classification, and feature and sequence display.

It has several features such as similarity search using BLAST and FASTA, peptide match, pattern search, pairwise sequence alignment, and multiple sequence alignment. The similarity search of human lysozyme against UniProtKB (UniProt knowledgebase) using the alignment program BLAST is shown in **Figure 1.11**. It can also be searched using the program FASTA. The partial results obtained with the search option are depicted in **Figure 1.12**. It indicates the sequences and their codes that match the query sequence along with other details, protein name, organism, length, % identity, overlap, e-value, etc (**Figure 1.12a**). Furthermore, it shows the alignment details with other proteins (**Figure 1.12b**). This will be helpful to identify the homologous sequences of any query protein. PIR can also be searched for any specific patterns, for example, alternating hydrophilic and hydrophobic residues as a pattern for β-strands (see **Chapter 2**), and continuous stretches of hydrophobic residues (e.g., AVILLIVWFFGA) in transmembrane helical proteins, etc.

1.4.2 SWISS-PROT and TrEMBL

SWISS-PROT (Bairoch and Apweiler, 1996) is an annotated protein sequence database established in 1986 and maintained collaboratively, since 1987, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library. It is a curated protein sequence database, which strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, posttranslational modifications and variants), a minimal level of redundancy, and a high level of integration with other databases. TrEMBL is a



Protein AC/ID	Protein Name	Length	Organism Name	PIRSF ID	SSearch		BLAST Search		
					Overlap	%iden	E-Value	Score	Alignment
P61627/LYSC_PANPA	Lysozyme C	148	Pan paniscus	PIRSF001064	148	100	4e-77	289	
P61628/LYSC_PANTR	Lysozyme C	148	Pan troglodytes	PIRSF001064	148	100	4e-77	289	
P61626/LYSC_HUMAN	Lysozyme C	148	Homo sapiens	PIRSF001064	148	100	4e-77	289	
B2R4C5/B2R4C5_HUMAN	Lysozyme C	148	Homo sapiens		148	100	4e-77	289	
P79179/LYSC_GORGO	Lysozyme C	148	Gorilla gorilla gorilla	PIRSF001064	148	99	2e-76	287	
P79239/LYSC_PONPY	Lysozyme C	148	Pongo pygmaeus	PIRSF001064	148	97	2e-76	287	
P79180/LYSC_HYLLA	Lysozyme C	148	Hylobates lar	PIRSF001064	148	95	5e-75	283	
P61634/LYSC_ERYPA	Lysozyme C	148	Erythrocebus patas	PIRSF001064	148	89	3e-72	273	
P61633/LYSC_CERAE	Lysozyme C	148	Chlorocebus aethiops	PIRSF001064	148	89	3e-72	273	
P61629/LYSC_PAPAN	Lysozyme C	148	Papio anubis	PIRSF001064	148	88	4e-71	270	
P79811/LYSC_NASLA	Lysozyme C	148	Nasalis larvatus	PIRSF001064	148	87	2e-71	270	

(a)

Sequence Alignment Generated by Similarity Search

```
>>P79179 LYSC_GORGO  Lysozyme C (148 aa)
 s-w opt: 1027  Z-score: 1271.8  bits: 241.3 E(): 1.2e-66
Smith-Waterman score: 1027;  99.324% identity (99.324% ungapped) in 148 aa overlap (1-148:1-148)

               10        20        30        40        50        60
P61626 MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRA
       ::::::::::::::: :::::::::::::::::::::::::::::::::::::::::::
P79179 MKALIVLGLVLLSVMVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRA
               10        20        30        40        50        60

               70        80        90       100       110       120
P61626 TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRD
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
P79179 TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRD
               70        80        90       100       110       120

              130       140
P61626 PQGIRAWVAWRNRCQNRDVRQYVQGCGV
       ::::::::::::::::::::::::::::
P79179 PQGIRAWVAWRNRCQNRDVRQYVQGCGV
              130       140
```

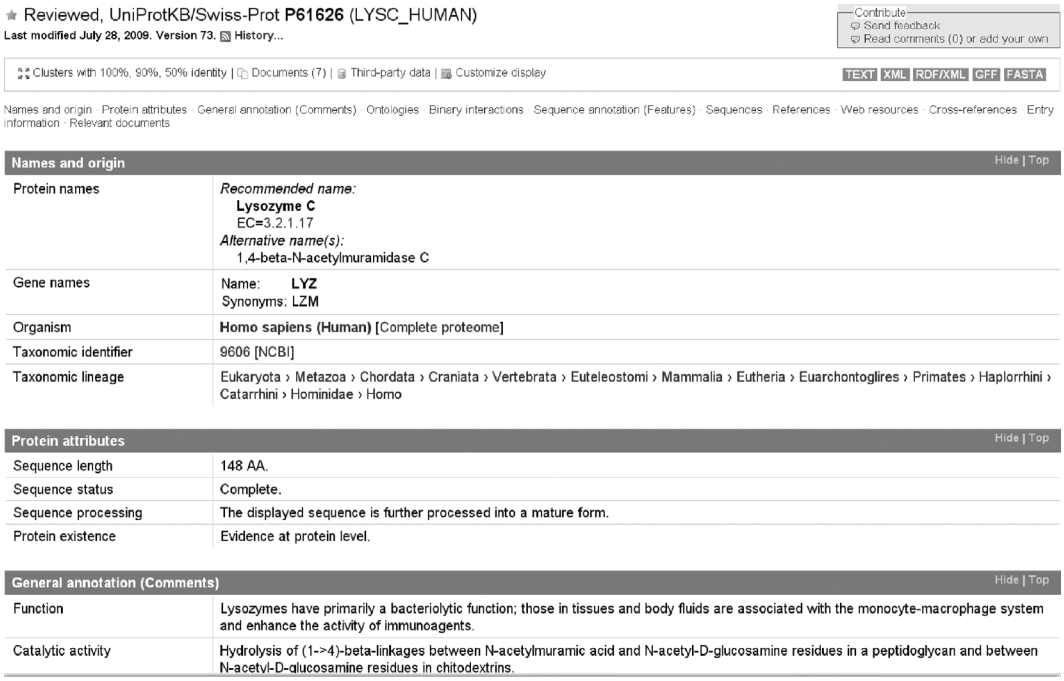
(b)

Figure 1.12 Results obtained with the search: (a) details of proteins that have high sequence identity and (b) alignment of residues (see **Chapter 2**) for the two proteins that have high sequence identity.

computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries, which are not yet integrated in SWISS-PROT. Currently, SWISS-PROT and TrEMBL have 0.5 and 7.6 million sequences, respectively. These databases are freely available at http://www.expasy.org/sprot/ and http://www.ebi.ac.uk/swissprot/.

SWISS-PROT contains the information about the name and origin of the protein, protein attributes, general information, ontologies, sequence annotation, amino acid sequence, bibliographic references, cross-references with sequence, structure and interaction databases, and entry information. An example for human lysozyme is shown in **Figure 1.13**. Furthermore, it has several search options, including sequence retrieval system (SRS), full-text search, advanced search, or by description or identification number.

In bioinformatics, developing a dataset plays a key role for any analysis or prediction. One can easily develop the dataset of amino acid sequences using SWISS-PROT. For example, the procedure for retrieving the data of "transcription factors" is shown in **Figure 1.14a**. Searching in UniProtKB with the keyword "transcription factors" will display all the relevant entries deposited in SWISS-PROT and TrEMBL.



★ Reviewed, UniProtKB/Swiss-Prot **P61626** (LYSC_HUMAN)
Last modified July 28, 2009. Version 73. History...

Contribute
Send feedback
Read comments (0) or add your own

Clusters with 100%, 90%, 50% identity | Documents (7) | Third-party data | Customize display

TEXT XML RDF/XML GFF FASTA

Names and origin · Protein attributes · General annotation (Comments) · Ontologies · Binary interactions · Sequence annotation (Features) · Sequences · References · Web resources · Cross-references · Entry information · Relevant documents

Names and origin	Hide | Top
Protein names	*Recommended name:* **Lysozyme C** EC=3.2.1.17 *Alternative name(s):* 1,4-beta-N-acetylmuramidase C
Gene names	Name: **LYZ** Synonyms: LZM
Organism	**Homo sapiens (Human)** [Complete proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo

Protein attributes	Hide | Top
Sequence length	148 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is further processed into a mature form.
Protein existence	Evidence at protein level.

General annotation (Comments)	Hide | Top
Function	Lysozymes have primarily a bacteriolytic function; those in tissues and body fluids are associated with the monocyte-macrophage system and enhance the activity of immunoagents.
Catalytic activity	Hydrolysis of (1->4)-beta-linkages between N-acetylmuramic acid and N-acetyl-D-glucosamine residues in a peptidoglycan and between N-acetyl-D-glucosamine residues in chitodextrins.

FIGURE 1.13 Sample entry for human lysozyme in UniProtKB/SWISS-PROT.

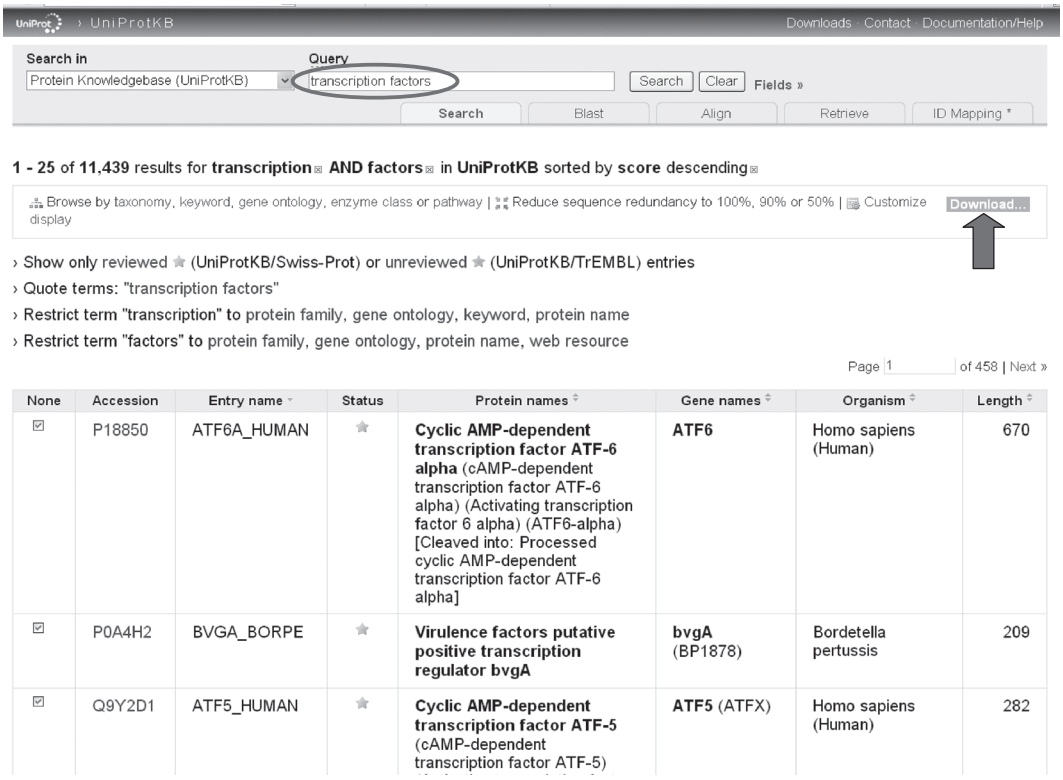
The retrieval system has the options to restrict the data deposited only in SWISS-PROT and the data at different sequence identities, 90% or 50%. Furthermore, the final data can be downloaded in different file formats, such as Tab delimited, Excel, FASTA, GFF, Flat text, XML, RDF/XML, and list. The result obtained for the sequences of transcriptions factors in FASTA format is shown in **Figure 1.14b**. This set of sequences can be used for the analysis of DNA-binding proteins. In a similar way, sequences for any kind of proteins can be easily obtained with SWISS-PROT.

1.4.3 UniProt: The Universal Protein Resource

Recently, the SWISS-PROT, TrEMBL, and PIR protein database activities have united to form the Universal Protein Knowledgebase (UniProt) consortium. It provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information (Bairoch et al. 2005). The UniProt produces three layers of protein sequence databases: UniProt Archive, Knowledgebase, and Reference database. The UniProt Knowledgebase is a comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase with extensive cross-references. It is freely available at: http://www.uniprot.org/

1.4.4 Other protein sequence databases

There are several other protein sequence databases, which aim for specific classes or functions. EXProt (http://www.cmbi.kun.nl/EXProt/) is a nonredundant protein database containing a selection of entries from genome annotation projects and



(a)

```
>sp|P18850|ATF6A_HUMAN Cyclic AMP-dependent transcription factor ATF-6 alpha OS=Homo s
MGEPAGVAGTMESPFSPGLFHRLDEDWDSALFAELGYFTDTDELQLEAANETYENNFDNL
DFDLDLMPWESDIWDINNQICTVKDIKAEPQPLSPASSSYSVSSPRSVDSYSSTQHVPEE
LDLSSSSQMSPLSLYGENSNSLSSAEPLKEDKPVTGPRNKTENGLTPKKKIQVNSKPSIQ
PKPLLLPAAPKTQTNSSVPAKTIIIQTVPTLMPLAKQQPIISLQPAPTKGQTVLLSQPTV
VQLQAPGVLPSAQPVLAVAGGVTQLPNHVVNVVPAPSANSPVNGKLSVTKPVLQSTMRNV
GSDIAVLRRQQRMIKNRESACQSRKKKKEYMLGLEARLKAALSENEQLKKENGTLKRQLD
EVVSENQRLKVPSPKRRVVCVMIVLAFIILNYGPMSMLEQDSRRMNPSVSPANQRRHLLG
FSAKEAQDTSDGIIQKNSYRYDHSVSNDKALMVLTEEPLLYIPPPPCQPLINTTESLRLN
HELRGWVHRHEVERTKSRRMTNNQQKTRILQGALEQGSNSQLMAVQYTETTSSISRNSGS
ELQVYYASPRSYQDFFEAIRRRGDTFYVVSFRRDHLLLPATTHNKTTRPKMSIVLPAINI
NENVINGQDYEVMMQIDCQVMDTRILHIKSSSVPPYLRDQQRNQTNTFFGSPPAATEATH
VVSTIPESLQ
>sp|P0A4H2|BVGA_BORPE Virulence factors putative positive transcription regulator bvgA
MYNKVLIIDDHPVLRFAVRVLMEKEGFEVIGETDNGIDGLKIAREKIPNLVVLDIGIPKL
DGLEVIARLQSLGLPLRVLVLTGQPPSLFARRCLNSGAAGFVCKHENLHEVINAAKAVMA
GYTYFPSTTLSEMRMGDNAKSDSTLISVLSNRELTVLQLLAQGMSNKDIADSMFLSNKTV
STYKTRLLQKLNATSLVELIDLAKRNNLA
>sp|Q9Y2D1|ATF5_HUMAN Cyclic AMP-dependent transcription factor ATF-5 OS=Homo sapiens
MSLLATLGLELDRALLPASGLGWLVDYGKLPPAPAPLAPYEVLGGALEGGLPVGGEPLAG
DGFSDWMTERVDFTALLPLEPPLPPGTLPQPSPTPPDLEAMASLLKKELEQMEDFFLDAP
LLPPPSPPPLPPPPLPPAPSLPLSLPSFDLPQPPVLDTLDLLAIYCRNEAGQEEVGMPPL
PPPQQPPPPSPPQPSRLAPYPHPATTRGDRKQKKRDQNKSAALRYRQRKRAEGEALEGEC
QGLEARNRELKERAESVEREIQYVKDLLIEVYKARSQRTRSC
```

Figure 1.14 Sequence retrieval in UniProtKB/SWISS-PROT: (a) The query and retrieved results and (b) sequences of "transcription factors" in FASTA format (the first line starts with ">" followed by amino acid sequences in single-letter code. Each line has 60 amino acid residues; see **Chapter 2**).

public databases, aiming at including only proteins with an experimentally verified function. The NCBI Entrez Protein database (http://www.ncbi.nlm.nih.gov/entrez) comprises sequences taken from a variety of sources, including SWISS-PROT, PIR, the Protein Research Foundation, the Protein Data Bank, and translations from annotated coding regions in the GenBank and RefSeq databases. Protein sequence records in Entrez have links to precomputed protein BLAST alignments, protein structures, conserved protein domains, nucleotide sequences, genomes, and genes. The Transport Classification Database (TCDB) is a curated, relational database containing sequence, classification, structural, functional, and evolutionary information about transport systems from a variety of living organisms (Busch and Saier, 2002; http://www.tcdb.org/).

Pongor's group (Pongor et al. 1993; Vlahovicek et al. 2005) developed a database of annotated protein domain sequences, SBASE. It facilitates the detection of domain homologies based on direct sequence database search using BLAST or a high-speed Smith Waterman algorithm, and returns a predicted domain architecture of the query sequence. Unlike traditional consensus representations, such as HMM (hidden Markov models), profiles and regular expressions, the SBASE domain library approach gives equal weights to all representatives, and a search against this library will detect both the typical and atypical (rare) representatives. It is to be noted that the domain library approach does not require either multiple alignment or learning algorithms to achieve accuracy. SBASE is freely available at http://www.icgeb.trieste.it/sbase.

1.5 Protein structure databases

Kendrew et al. (1958) solved the first three-dimensional structure of the protein myoglobin using X-ray crystallography. Subsequently, several structures have been determined with this technique. After two decades, NMR spectroscopy has been used to determine the protein structures in solution. Recently, neutron diffraction and electron microscopy have been used to determine protein structures. The number of protein structures has been increased every year, and they are deposited in a database, Protein Data Bank (PDB). On the basis of the structures available in PDB, several other databases have been established for the structural classification of proteins (SCOP), topology and architecture, and amino acid properties, etc.

1.5.1 Protein Data Bank

The PDB was established at Brookhaven National Laboratories, USA, in 1971 as an archive for biological macromolecular crystal structures (Bernstein et al. 1977). Recently, the management of PDB became the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB), and it has six mirror sites at San Diego Supercomputer Center and Rutgers University in the USA; Cambridge Crystallographic Data Center, UK; National University of Singapore; Osaka University, Japan; and Max Delbruck Center for Molecular Medicine, Germany (Berman et al. 2000). PDB is available at http://www.rcsb.org/.

PDB stores the data in a uniform format atomic coordinates and partial bond connectivities, as derived from crystallographic studies. Text included in each data entry gives pertinent information for the structure at hand (e.g., species from which

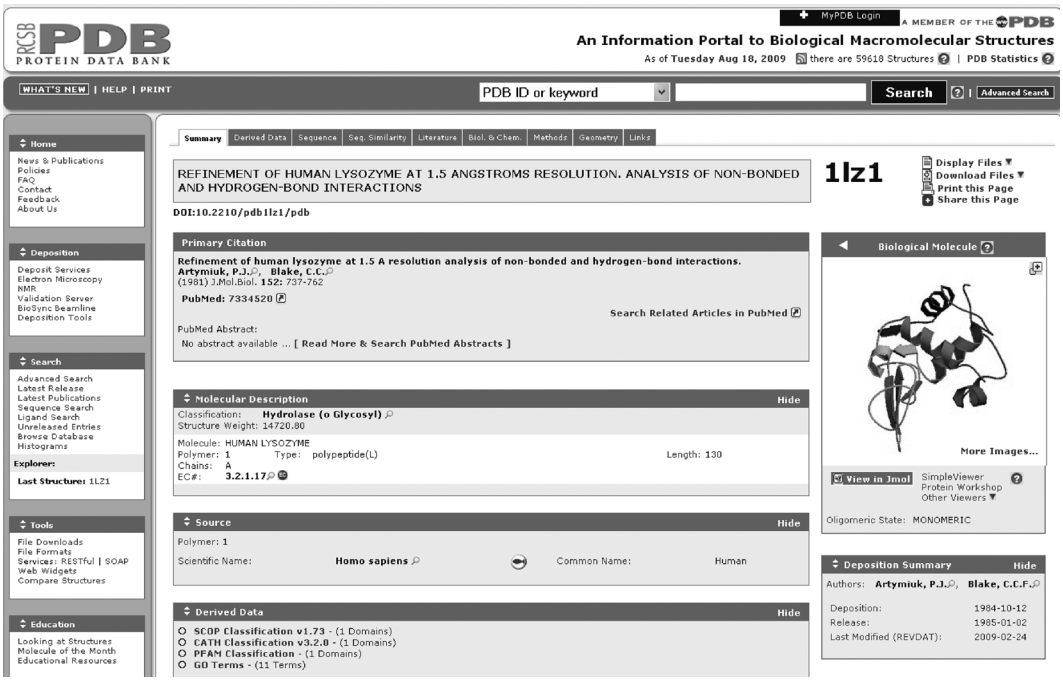


FIGURE 1.15 Snapshot showing the details of Protein Data Bank. The data for human lysozyme are shown.

the molecule has been obtained, resolution of diffraction data, literature citations and specifications of secondary structure). In addition to atomic coordinates and connectivities, the PDB stores the temperature factor for each atom. The PDB has been widely used in structural analysis on various aspects, such as atomic and residue contacts, amino acid clusters, developing potentials, and amino acid properties.

Currently, PDB has more than 57,000 structures, and 53,000 of them are proteins and their complexes (May 12, 2009). PDB has the search option with its code, authors, or full-text search. It contains the summary information about the name and source of the protein, experimental method, authors, and references. In addition, the details about crystallographic conditions are also provided. An example to human lysozyme is shown in **Figure 1.15**. Furthermore, it has the options to view the structure, display and download the files with/without three-dimensional coordinates, and it provides information about structural neighbors, geometry, sequence details, and other sources.

The sequence provides the detailed information about the number of amino acid residues in the protein, chain information, and amino acid sequence for each chain along with their secondary structures derived using DSSP (Kabsch and Sander, 1983), and number and percentage of each secondary structure through external links. For example, human lysozyme has four α-helices and three β-strands, and the content of α-helices is about 31%. Furthermore, it has the facility to save the sequence in FASTA format, which can be used in other programs.

PDB has the option to see the structural coordinates and view the structure on the screen or to save the structural information in a disk. It also has the possibility to download the data by ftp (ftp://ftp.wwpdb.org/pub/pdb/) or request the structures in CD. The structures can also be visualized using other programs, SWISS-PDB viewer (http://spdbv.vital-it.ch/), RASMOL (Sayle and Milner-White, 1995; http://www.umass.edu/microbio/rasmol/), Jmol (http://jmol.sourceforge.net/), KiNG (http://kinemage.biochem.duke.edu/software/king.php), PyMOL (DeLano, 2002; http://www.pymol.org), etc. Structural data have the information about the protein, references, refinement details, sequence and secondary structure information, translation and rotation matrices for oligomers, disulfide bonds, and the atomic positions. The atomic position has the X, Y, and Z coordinates and temperature factors (**Figure 1.16**). The atomic coordinates have been extensively used to analyze the principles governing the folding and stability of protein structures, mechanism of protein folding and its interactions with other molecules, etc.

In addition, PDB has plenty of external links to other software for structural analysis and verification, modeling and simulation, molecular graphics, and so on.

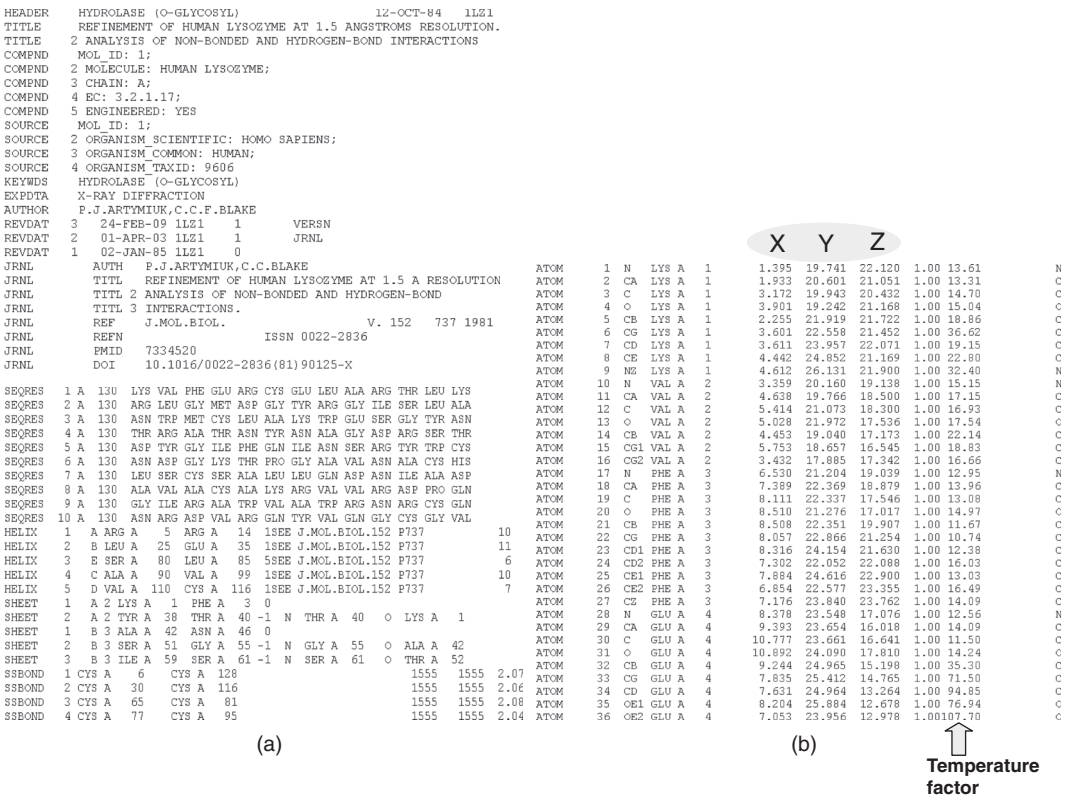


Figure 1.16 Information available in the PDB file: (a) It has the header information along with compound, source, authors, references, complete amino acid sequence, secondary structural details, and three-dimensional coordinate; (b) coordinate has the information about the residue name and number, atom name and number, chain and X, Y, Z coordinates. It also has the temperature factor for each atom.

1.5.2 Database for nonredundant PDB structures

A dataset with several similar structures influences the characteristic features of such data, which cause a bias on the result. Hence, a dataset of nonredundant proteins is necessary for analyzing the features of protein structures. The level of redundancy may vary depending upon the nature of the problem and the number of proteins in the database. The ASTRAL database provides the PDB codes and amino acid sequences for two cutoff values, i.e., less than 40% and 95% sequence identities. Furthermore, it has the option of selecting the cutoff values from 10% to 100%. The description about the selection of nonredundant dataset of proteins with 25% sequence identity is shown in **Figure 1.17**. The PDB identifiers for this cutoff have been selected to display the results, and it is also possible to get the sequences for these protein structures. Furthermore, ASTRAL has the options to get the sequences with e-values, fold, superfamily, etc.

For a general analysis, one can get the information about the nonredundant structures from ASTRAL. Furthermore, the nonredundant dataset of any

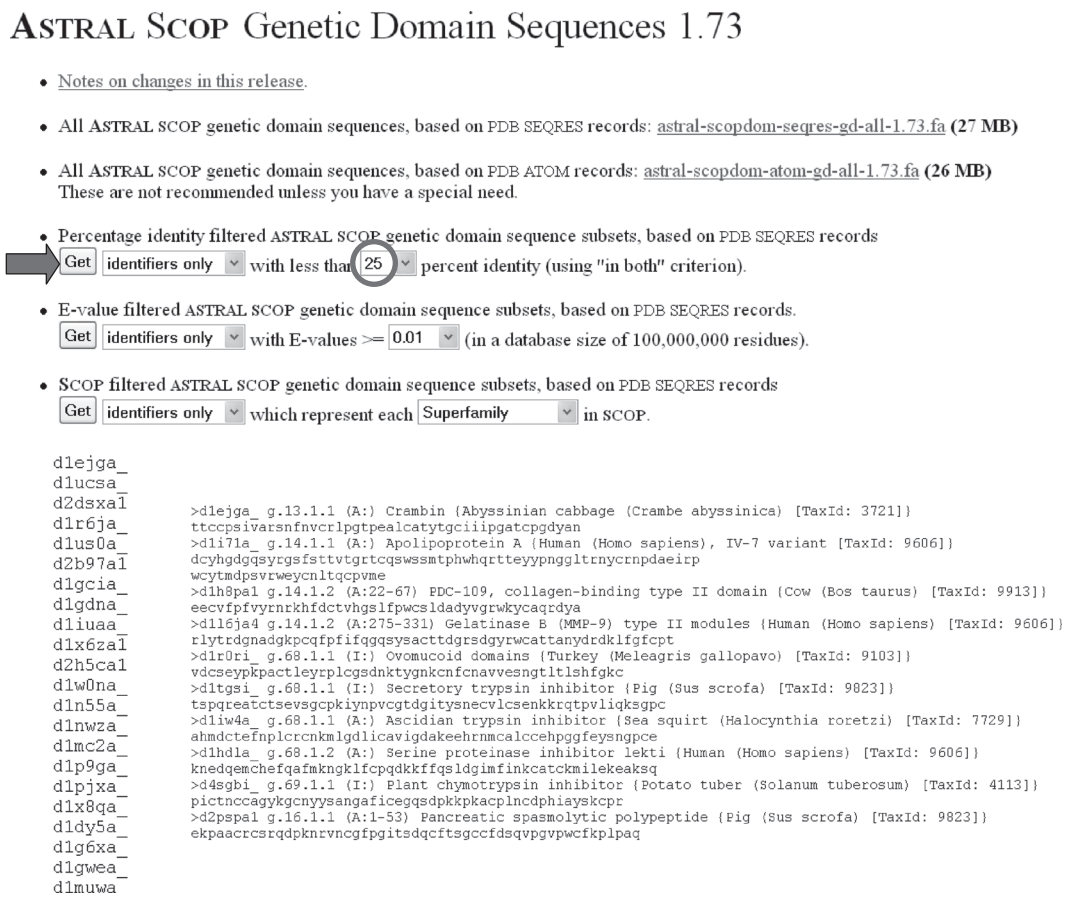


FIGURE 1.17 Retrieval of nonredundant protein sequences from PDB using ASTRAL. It uses SCOP classification for the domains, and the data obtained with the sequence identity of less than 25% are shown.

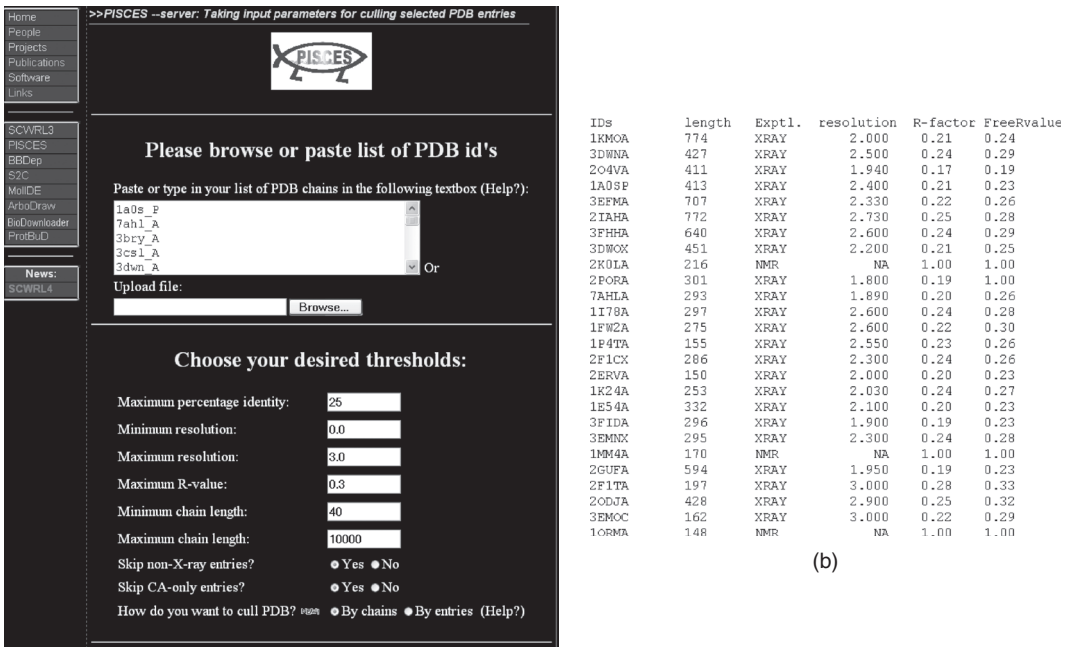


Figure 1.18 Retrieval of nonredundant protein structures using PISCES. The users have the feasibility of selecting the identity, resolution, etc. and provide own dataset of proteins. (a) The search options and (b) results.

specific class of proteins (e.g., all proteins, membrane proteins, DNA-binding proteins, protein–protein complexes, etc.) can be derived by matching the dataset of interest with ASTRAL nonredundant dataset. ASTRAL is available at http://astral.berkeley.edu/.

Wang and Dunbrack (2005) developed a sequence-culling database server, PISCES, for producing lists of nonredundant proteins from the PDB using entry- and chain-specific criteria and mutual sequence identity. It uses a combination of PSI-BLAST and structure-based alignments to determine sequence identities. An example is shown in **Figure 1.18**. It takes the PDB codes with chain information along with other conditions, such as % sequence identity, resolution, R-factor, inclusion of non–X-ray structures, number of amino acid residues, etc (**Figure 1.18a**), and sends the results via e-mail. The output has four files: (i) original id, (ii) culled id, (iii) FASTA format sequence, and (iv) similarity log. The result obtained for the culled sequences from the list of 34 β-barrel membrane proteins is shown in **Figure 1.18b**. PISCES is available at http://dunbrack.fccc.edu/pisces/.

Noguchi and Akiyama (2003) developed a database of representative chains from PDB. The search options to retrieve nonredundant set of proteins are shown in **Figure 1.19a**. It has several features, including the limits with resolution, R-factor, number of residues, fragments, complex structures, and membrane proteins. Furthermore, the users have the options to use any cutoff for sequence identity and root-mean-square deviation between the sequences and structures, respectively as shown in **Figure 1.19b**. The results obtained with the search are displayed in

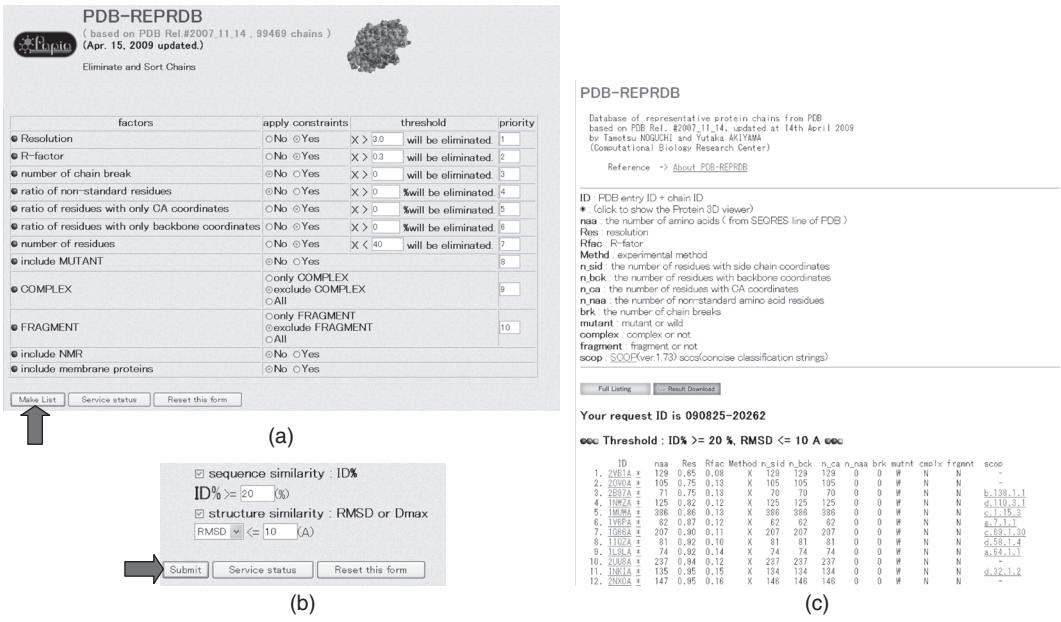


FIGURE 1.19 Retrieval of nonredundant proteins with PDB-REPRDB. This server will have several options to retrieve protein sequences, including the sequence identity and RMSD between two structures. (a) The initial parameters, (b) similarities, and (c) results are shown in the figure.

Figure 1.19c. It provides several details, including the links to SCOP database (see **Section 1.5.4**). It is available at http://www.cbrc.jp/pdbreprdb/.

1.5.3 PDB-related databases

On the basis of PDB, several other databases have been developed for different sets of proteins, such as membrane proteins, protein–protein complexes, protein–nucleic acid complexes, and ligand binding proteins. Tusnady et al. (2005) developed a database of transmembrane proteins, PDBTM (http://pdbtm.enzim.hu/), which includes the sequences and structures of redundant and nonredundant α-helical and β-barrel membrane proteins along with their membrane-spanning segments. Jayasinghe et al. (2001) compiled the structures of known membrane proteins, MPtopo (http://blanco.biomol.uci.edu/mptopo), and classified them into several groups, such as monotopic, GPCRs, rhodopsins, and β-barrel membrane proteins. They have included the PDB codes, structures, and their respective references. Ikeda et al. (2003) developed a database of transmembrane protein topologies, TMPDB (http://bioinfo.si.hirosaki-u.ac.jp/~TMPDB/), which is based on the experimental evidences from X-ray crystallography, NMR spectroscopy, etc. Sarai and colleagues (An et al. 1998) developed a database for protein–nucleic acid complex structures, and these structures have been classified into different groups based on the recognition motif of proteins and DNA involved in the complex. Puvanendrampillai and Mitchell (2003) developed the Protein Ligand Database (PLD), which has the PDB codes for protein-ligand complexes along with binding information.

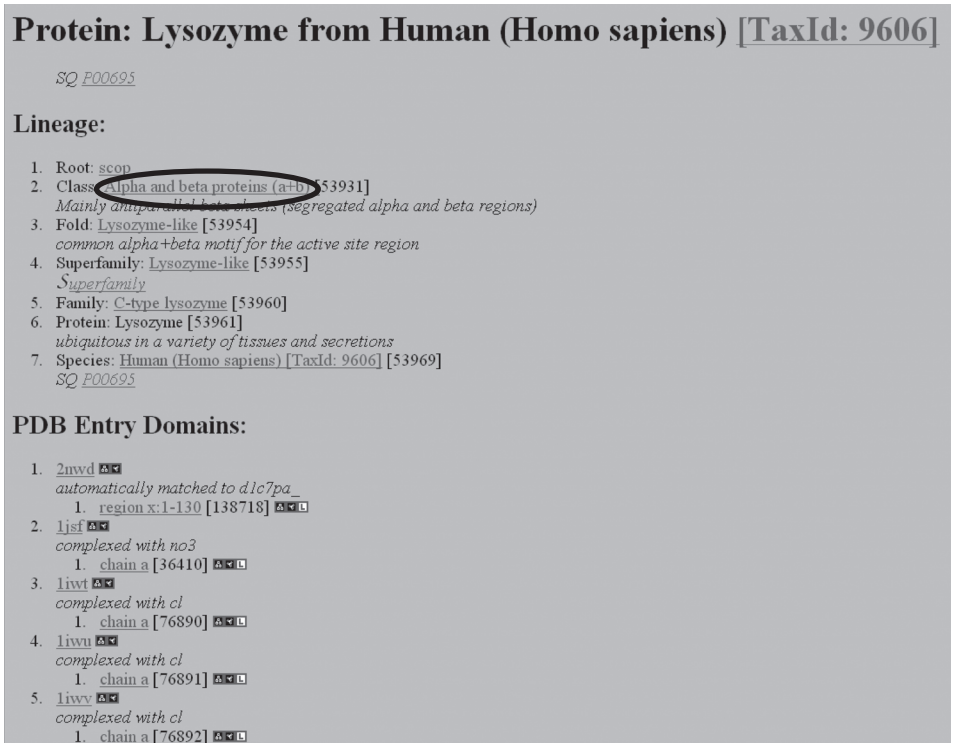


Figure 1.20 Sample entry for human lysozyme in SCOP database. It is classified as $\alpha + \beta$ protein. Furthermore, links are available for the different folds and families within this class.

1.5.4 Databases for structural classification of proteins

Murzin et al. (1995) constructed the SCOP database, which provides a detailed and comprehensive description of the structural and evolutionary relationships of the proteins of known structures. For each protein, the classification has the hierarchical levels, family, superfamily, fold, and structural class. An example for human lysozyme is shown in **Figure 1.20**. It belongs to the family of C-type lysozyme and the fold of lysozyme-like under $\alpha + \beta$ structural class. The structure can be identified with a six-letter code (1lz1__). The first four letters of the code (1lz1) represent the PDB code followed by the chain name (_; it indicates there are no multiple chains in this protein) and domain name (_). SCOP database has been linked from the PDB to obtain the structural class information of each protein directly. On the other hand, one can search the SCOP database for obtaining the structural class, fold, and domain information.

Orengo et al. (1997) developed a semiautomatic procedure for deriving a novel hierarchical classification of protein domain structures (CATH) and created a database in providing the class information for all the structures in PDB. The four main levels of CATH classification are protein class (C), architecture (A), topology (T), and homologous superfamily (H). Class is the simplest level, and it essentially describes the secondary structure composition of each domain. Architecture summarizes the shape revealed by the orientations of the secondary structure units,

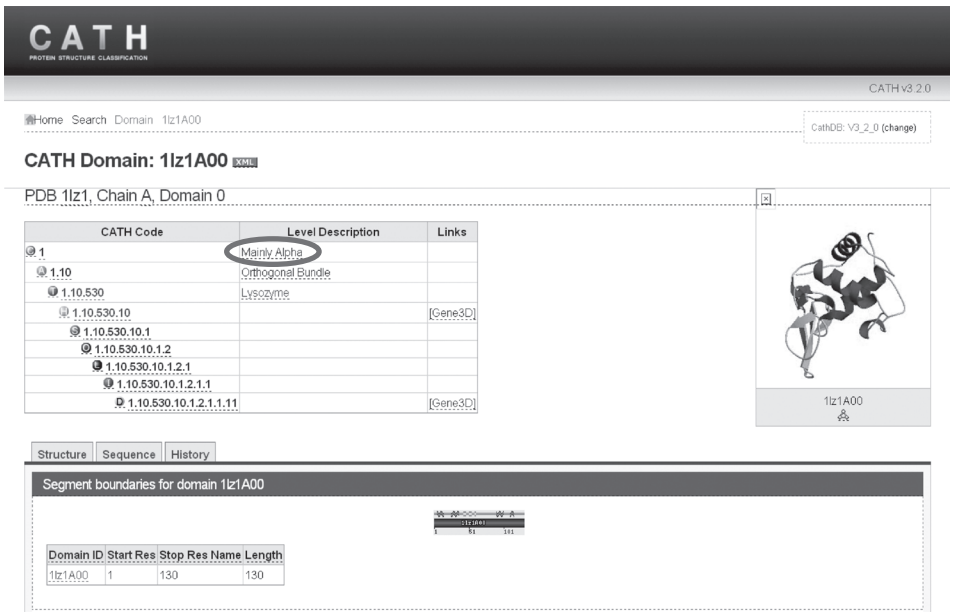


Figure 1.21 Structural classification of human lysozyme with CATH.

such as barrels and sandwiches. At the topology level, sequential connectivity is considered, such that members of the same architecture might have quite different topologies. The homologous superfamilies contain proteins with highly similar structures and functions. The CATH classification for human lysozyme is shown in **Figure 1.21**. It has the architecture of an orthogonal bundle, and it is designated an all-α class proteins in CATH. It may be possible to have different assignments in SCOP and CATH for some proteins, especially if the percentage of helical content is high and strand content is less or vice versa. In human lysozyme, the content of α-helical structures is 31% and that of β-strands is 8%. On the basis of the high content of α-helix it was classified as all-α proteins in CATH, and due to the presence of α-helices and β-strands SCOP classified it as $\alpha + \beta$ protein.

1.6 Literature databases

Literature databases play a vital role to understand the current status of research on various fields of interest. At present, several databases provide the information with/without restrictions. These databases include PubMed, Google Scholar, Scopus, Science Citation Index, Chemical abstracts, and so on. The PubMed literature database covers most of the protein-related works published in international journals, and it has different features to search the database and display the results. It can be searched with author's name, year of publication, name of the journal, and key words. As an example, a search for the articles published in *Nature* about protein structure is shown in **Figure 1.22**. The journal name is specified with *Nature*. The search retrieved 3240 records, including 45 reviews. **Figure 1.22** shows the summary of results, and it is possible to display abstract, abstract plus with

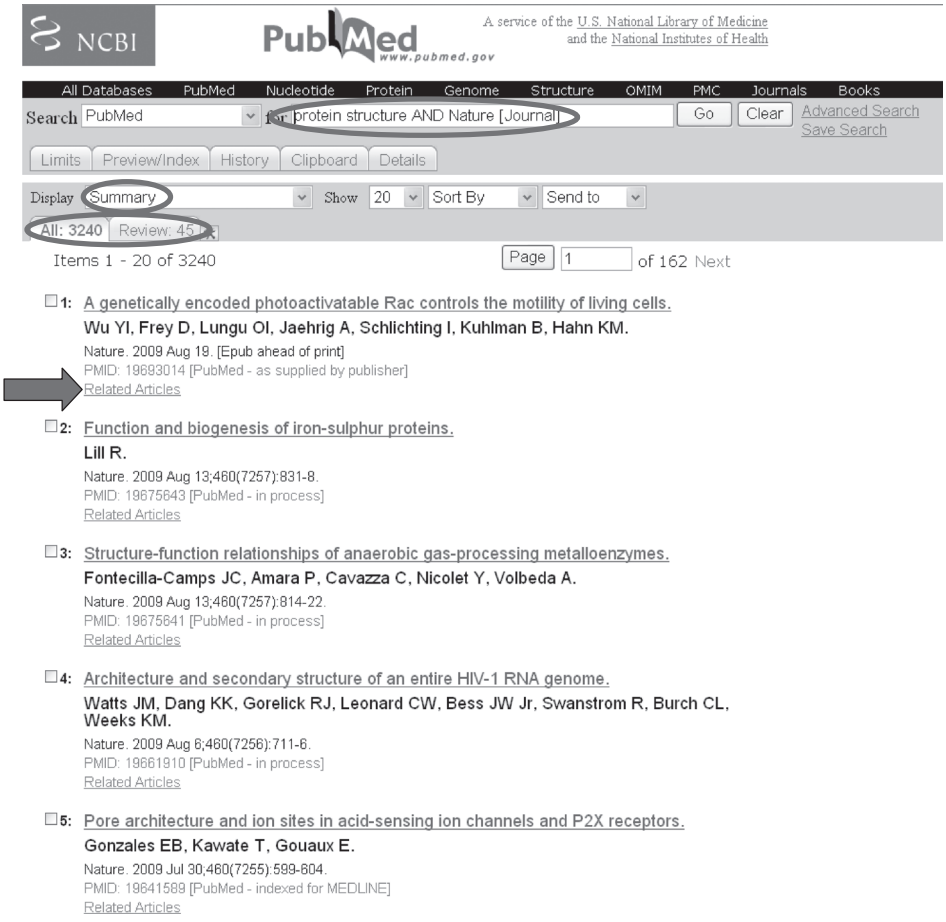


Figure 1.22 Retrieval of results using PubMed literature database. The search with keywords and journal name, display options, and number of retrieved records are indicated. The link to related articles is shown with an arrow.

key words, etc. It is also possible to see the related articles for all the references listed in the result page. Each hit is linked with its respective journal site, and one can get the complete article if the journal has open access options or the users have the license to access the full text of the journal. PubMed can be freely accessed at http://www.ncbi.nlm.nih.gov/pubmed/.

1.7 Exercises

1. Retrieve the sequence of E. coli Triose phosphate isomerase in FASTA format. *Hint:* Use the text search in iProClass, and search with Triose phosphate isomerase and E. coli; select the correct one among the displayed results; "show" the selected one; and click on "save the result" as FASTA.

2. Obtain the amino acid sequences of membrane proteins located in outer membrane using SWISS-PROT/UniProtKB database.
   Hint: use the keyword, "membrane protein" and location, "outer membrane." Click on Download and select the format. Number of hits is about 500.
3. Find the three-dimensional coordinates of porin from Rhodobactor capsulatus and store the data.
   Hint: Search at the top of the page of PDB using the key words. Select the best match (e.g., 3POR or 2POR). Click on Display files and select PDB text. Click on Download files to save the coordinates.
4. Get the list of PDB codes for RNA-binding proteins.
   Hint: Search with RNA-binding proteins. In the result page, click on "Results ID List" to get the list of codes.
5. Check the interacting partners (protein and DNA) in 1RXW.
   Hint: Use Jmol to view the protein–DNA complex, and find the chain information.
6. Obtain the PDB codes that have ligands and the sequence identity is less than 30%.
   Hint: Go to Advanced Search and search with "ligand" in Structure title. Select the sequence identity as 30%.
7. How many helices are there in 2LZM?
   Hint: Check the PDB file of 2LZM and the presence of helices and strands.
8. What is the average fluctuation for the residue Phe at position 4 in 2LZM?
   Hint: Identify the temperature factors of all atoms in Phe4, and get the average. The information is shown in **Figure 1.16**.
9. List the major transport systems in Transport classification database?
   Hint: Check the TCDB database and click on TC system
10. What is the major class of transporters?
    Hint: Find the number of proteins in each transporter, and find the one with the maximum number of proteins.
11. Get the dataset of protein structures with less than 40% sequence identity.
    Hint: Click on ASTRAL 1.73 with less than 40% sequence identity.
12. Get the dataset of protein structures with less than 20% sequence identity.
    Hint: Select the sequences and follow **Figure 1.17**.
13. Get the PDBs with less than 20% sequence identity using the IDs obtained for RNA-binding proteins (question 4).
    Hint: Follow **Figure 1.18** using the list of PDB codes obtained as the answer to question 4.
14. Obtain the list of nonmembrane proteins without mutation and not complexed with other molecules and the sequence identity of less than 25%.
    Hint: Use PDB-REPRDB and provide the necessary conditions.
15. Get the nonredundant structures of β-barrel membrane proteins.
    Hint: Go to PDBTM and click on download. The codes are listed with the link at the end of the page.
16. Retrieve the proteins with seven α-helical segments in PDBTM.
    Hint: Search with type of transmembrane protein, and select the number of segments.

17. Compare the folding types and number of proteins belonging to different structural classes of proteins.
    Hint: SCOP has the information.
18. What are the major folding types in four structural classes?
    Hint: Compare the folding types with number of entries.
19. Compare the heme-binding proteins (family: globin) and contents in SCOP and CATH databases.
    Hint: SCOP: All-Globin like-Globin; CATH: 1.10.490.10.
20. Find the domain information for outer membrane protein TolC.
    Hint: search CATH database with FASTA sequence, which can be found from UniProtKB/SWISS-PROT (P02930) or PDB.
21. Find the articles published about "protein interactions" and "shape complimentarity."
    Hint: Search PubMed with relevant keywords.
22. Identify the papers published in the journal *Cell* about mitochondrial β-barrel membrane proteins.
    Hint: Search with mitochondrial β-barrel membrane proteins AND *Cell* [Journal].

References

An J, Nakama T, Kubota Y, Sarai A. 3DinSight: an integrated relational database and search tool for the structure, function and properties of biomolecules. Bioinformatics. 1998;14(2):188–195.

Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181(96): 223–230.

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2005;33(Database issue):D154–159.

Bairoch A, Apweiler R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res. 1996;24(1):21–25.

Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao C, Yeh LS, Ledley RS, Janda JF, Pfeiffer F, Mewes HW, Tsugita A, Wu C. The protein information resource (PIR). Nucleic Acids Res. 2000;28(1):41–44.

Berman HM, Westbrook JZ, Feng G, Gilliland TN, Bhat H, Weissig IN, Shindyalov, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–242.

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977;112(3):535–542.

Busch W, Saier MH Jr. The transporter classification (TC) system, 2002. Crit Rev Biochem Mol Biol. 2002;37:287–337.

DeLano WL. The PyMOL Molecular Graphics System. San Carlos, CA: DeLano Scientific; 2002. Available online at http://www.pymol.org.

Dayhoff MO, Eck RV, Chang MA, Sochard MR. Atlas of Protein Sequence and Structure, Vol. 1. Silver Spring, MD: National Biomedical Research Foundation; 1965.

Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol. 2004;86(2):235–277.

Ikeda M, Arai M, Okuno T, Shimizu T. TMPDB: a database of experimentally-characterized transmembrane topologies. Nucleic Acids Res. 2003;31(1):406–409.

Jayasinghe S, Hristova K, White SH. MPtopo: A database of membrane protein topology. Protein Sci. 2001;10(2):455–458.
Kabsch W, Sander C, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–2637.
Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature. 1958;181(4610):662–666.
Levitt M, Chothia C. Structural patterns in globular proteins. Nature. 1976;261(5561): 552–558.
Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247(4):536–540.
Nelson DL, Cox MM. Lehninger Principles of Biochemistry. New York: W.H. Freeman and Company; 2005.
Noguchi T, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res. 2003;31(1):492–493.
Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH: a hierarchic classification of protein domain structures. Structure. 1997;5(8):1093–1108.
Pauling L, Corey RB, Branson HR. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A. 1951;37(4):205–211.
Pongor S, Skerl V, Cserzö M, Hátsági Z, Simon G, Bevilacqua V. The SBASE domain library: a collection of annotated protein segments. Protein Eng. 1993;6(4):391–395.
Puvanendrampillai D, Mitchell JB. Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein–ligand complexes. Bioinformatics. 2003;19(14):1856–1857.
Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963;7:95–99.
Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci. 1995;20(9):374.
Schulz GE. Transmembrane beta-barrel proteins. Adv Protein Chem. 2003;63:47–70.
Tusnady GE, Dosztanyi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res. 2005;33(Database issue):D275–278.
Vlahovicek K, Kaján L, Agoston V, Pongor S. The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines. Nucleic Acids Res. 2005;33(Database issue):D223–D225.
Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003;19:1589–1591.
White SH, Wimley WC. Membrane protein folding and stability: physical principles. Annu Rev Biophys Biomol Struct. 1999;28:319–365.