

CHAPTER 3

Archives and information retrieval

Chapter contents

Introduction 118

Database indexing and specification of search terms 118

Follow-up questions 120

Analysis of retrieved data 121

The archives 121

Nucleic acid sequence databases 122

Genome databases 124

Protein sequence databases 124

Databases of structures 128

Specialized, or 'boutique' databases 135

Expression and proteomics databases 136

Databases of metabolic pathways 138

Bibliographic databases 139

Surveys of molecular biology databases and servers 139

Gateways to archives 140

Access to databases in molecular biology 141

ENTREZ 141

The Sequence Retrieval System (SRS) 148

The Protein Identification Resource (PIR) 149

ExPASy—Expert Protein Analysis System 150

Ensembl 151

Where do we go from here? 152

Recommended reading 152

Exercises, Problems, and Weblems 153

Learning goals

1. To understand the general types of data about the molecules and processes of life that are assembled in support of research and applications in biology, medicine, agriculture and technology.
2. To know the basic infrastructure of bioinformatics, in terms of the sites and responsibilities of the major archival projects.
3. To understand the basic concepts of information retrieval, including how to frame queries.
4. To have facility with general search engines on the Web, and specific sites for bioinformatics.
5. To know how to search for specific information about sequences, structures, metabolic pathways, relationships to disease, and how to launch analyses of the data recovered.

Introduction

This chapter introduces the information retrieval skills that will allow you to make effective use of the databanks. The goal is to give you familiarity with basic operations. It will then be easy to improve and develop your technique. Indeed, embedded in many databanks are tutorials which make it easy to explore their facilities.

Database indexing and specification of search terms

An index is a set of pointers to information in a database. In searching the entire World Wide Web, or a specialized database in molecular biology, you propose one or more search terms, and a program checks for them in its tables of indices. The model is that the entire database is composed of **entries**—discrete coherent parcels of information. The information retrieval software identifies entries with contents relevant to your interest. An example of the simplest paradigm is that you submit the term ‘horse’ and the program returns a list of entries that contain the term horse.

A full search of the Web would turn up information about many different aspects of horses—molecular biology, breeding, racing, poems about horses—most of which you don’t want to see. For a successful search, it is not enough to mention what you *do* want—you must ensure that the desired responses don’t get buried in extraneous rubbish. (Of course rubbish is merely whatever *other* people are interested in.)

To focus the results, information retrieval engines accept multiple query terms or keywords. A search for 'horse liver alcohol dehydrogenase' would produce responses specialized to this enzyme. The search would identify entries that contain all four keywords that you submitted: horse AND liver AND alcohol AND dehydrogenase. It would not return poems about horses among its top hits (except in the unlikely event that a poem contained all four keywords).

It is possible to ask for other logical combinations of indexing terms. For instance, if a search engine didn't know about transatlantic spelling differences, it would be useful to be able to search for 'hemoglobin OR haemoglobin'. (Note that a search for 'hemoglobin haemoglobin' would probably be interpreted as 'hemoglobin AND haemoglobin' which would pick up only documents written by international committees or orthographically-challenged expatriates.)

If you wanted to know about other dehydrogenases, you could ask for 'dehydrogenase NOT alcohol'. This would retrieve entries that contain the term dehydrogenase but did NOT contain the word alcohol. You would find entries about lactate dehydrogenase, malate dehydrogenase, etc. You would miss references to review articles that compared alcohol dehydrogenases to other dehydrogenases, or alignments of the sequences of many dehydrogenases including alcohol dehydrogenase. You might regret missing these.

Many database search engines will allow complex logical expressions such as '(haemoglobin OR hemoglobin) AND (dehydrogenase NOT alcohol)'. Construction of such expressions is an exercise in set theory, and it is helped by drawing Venn diagrams. Although the logic of a search is independent of the software used to query a database, different programs demand different syntax to express the same conditions. For example the query for dehydrogenase NOT alcohol might have to be entered as DEHYDROGENASE -ALCOHOL OR DEHYDROGENASE !ALCOHOL.

Specialized databases, including those in molecular biology, impose a structure on the information, to separate different categories of information. This is essential. There are currently active biomedical scientists named E(lisabetta) Coli, (John D.) Yeast, (Patrice) Rat, and a large number of Rabbits, as well as several Crystals and Blots. If you wanted to find papers published by these investigators, it would be naive to perform a general search of a molecular biology database with any of their surnames. Many databases provide separate indexing and searching of different categories of information. They permit searching for papers of which E. Coli is an AUTHOR.

Some of the categories, such as taxonomy, have **controlled vocabularies**. Often these are presented to the user as pull-down menus. To do a search for 'globin NOT mammal', and pick out the relatively few entries about nonmammalian globins rather than the very many entries about globins, including human haemoglobins, that do not explicitly mention the term mammal, requires an information retrieval system that 'understands' the taxonomic hierarchy. Controlled

vocabularies—limited, explicit, and carefully defined sets of terms—are also important in distributing queries among several databases.

A technical problem that frequently creates difficulty is how to enter terms containing nonstandard characters such as accent marks or umlauts, Greek letters, and, as already mentioned, differences between US and British spelling. A specialized database such as NCBI's ENTREZ can handle the US-British spelling differences with a synonym dictionary. Programs that index the entire Web usually do not. Ignore the accent marks and hope for the best.

Follow-up questions

When searching in databases, it is rare that you will find exactly what you want on the first round of probing. Usually you have to modify the query, on the basis of the results initially returned. Most information retrieval software permits consecutive, cumulative searches, with altered sets of search terms and/or logical relationships. Conversely, once you find what you looked for, you will often want to extend your search to find related material. If you find a gene sequence, you might want to know about homologous genes in other organisms. Or whether a three-dimensional structure of the corresponding protein is available. Or you might want to know about papers published about the gene.

For these subsidiary queries you need links between entries in the same or different databases. This is a special example of the question of how one 'browses' in electronic libraries—a difficult problem, the subject of current research.

To find homologous genes you would like links to other items in the *same* database (a database of gene sequences). To find structures, or bibliographical references, related to a gene, you would like links *between* different databases (from the database of gene sequences to a database of three-dimensional structures, or to a bibliographical database). As the number of databases grows, intercommunication among them has become a high-priority goal. Indeed, the interactivity of the databases in molecular biology is growing more and more effective, so that these operations are fairly easy now—formerly one had to do separate searches on isolated databases. This is a generalization of the original model of a database as a closed set of independent entries that can be selected only by their indexed contents.

To some extent database activities in bioinformatics can be classified into *archiving*—with the major goals of conservation and curation—and *interpreting*—the compilation of biological information in a form most useful to support research. Different archives specialize in different kinds of data—nucleic acid sequences, protein sequences, structures—for reasons in part historical and in part because of the different specialized curatorial skills required. Interpretative databases are free to combine information from any available sources. In most cases, archival and interpretative projects are carried out at the same institution and even by the same people.

Two aspects of the development of bioinformatics databases are apparent. One is the very great growth of individual database projects that recombine the archived data in different ways. The other is the combination of many individual databases into ‘umbrella’ sites. There is really no paradox—both are going on. (Genes also both multiply and combine.)

Most database unifications are merely extensions, with greater appearance of intimacy, of the collaborations or competitive efforts that in most cases formerly existed. We shall see, for instance, that protein sequence databases are coordinating their activities as UniProt; and that the protein structure databases are coordinating as the Worldwide Protein Data Bank. Interpro is an umbrella database that integrates the contents, features, and annotation of individual databases of protein families, domains, and functional sites, and contains links to others, including the Gene Ontology ConsortiumTM functional classification. It currently subsumes the PROSITE, Pfam, PRINTS, SMART and ProDom databases, and intends to assimilate others. (Resistance is futile.)

Analysis of retrieved data

Sometimes as a result of a search you will want to launch a program using the results retrieved as input. For instance, if you identify a protein sequence of interest, you might want to perform a PSI-BLAST search. This is not strictly a database-lookup problem, and formerly you would have to run a separate job, and feed the retrieved sequence to the application program by hand. However, like searches in multiple databases, information retrieval systems in molecular biology often provide facilities for initiating such processes. This makes for very much improved fluency in your sessions at the computer.

The archives

Although our knowledge of biological sequence and structure data is very far from complete, it is of quite respectable size, and growing extremely rapidly. Many scientists are working to generate the data, or to carry out research projects analysing the results. Archiving and distribution are carried out by particular databanking organizations.

Archiving of bioinformatics data was originally carried out by individual research groups motivated by an interest in the associated science. As the requirements for equipment and personnel grew—and the nature of the skills required changed, to include much more emphasis on computing—they have been made the responsibility of special national and even international projects, on a very large scale indeed. Anyone who has followed the entire history of these projects cannot help being impressed by their growth from small, low-profile and ill-funded projects carried out by a few dedicated individuals, to a multinational

heavy industry subject to political takeovers and the scientific equivalent of leveraged buyouts.

Primary data collections related to biological macromolecules include:

- ◆ Nucleic acid sequences, including whole-genome projects
- ◆ Amino acid sequences of proteins
- ◆ Protein and nucleic acid structures
- ◆ Small-molecule crystal structures
- ◆ Protein functions
- ◆ Expression patterns of genes
- ◆ Metabolic pathways, and networks of interaction and control
- ◆ Publications

Nucleic acid sequence databases

The worldwide nucleic acid sequence archive is a triple partnership of The National Center for Biotechnology Information (USA), the EMBL Data Library (European Bioinformatics Institute, UK), and the DNA Data Bank of Japan (National Institute of Genetics, Japan). The groups exchange data daily. As a result the raw data are identical, although the format in which they are stored, and the nature of the annotation, vary among them. These databases curate, archive, and distribute DNA and RNA sequences collected from genome projects, scientific publications, and patent applications. To make sure that these fundamental data are freely available, scientific journals require deposition of new nucleotide sequences in a database as a condition for publication of an article. Similar conditions apply to amino acid sequences, and to nucleic acid and protein structures.

The nucleic acid sequence databases, as distributed, are collections of entries. Each entry has the form of a text file containing data and annotations for a single contiguous sequence. Many entries are assembled from several published papers reporting overlapping fragments of a complete sequence. Others are complete genomes.

Entries have a life cycle in the database. Because of the desire on the part of the user community for rapid access to data, new entries are made available before annotation is complete and checks are performed. Entries mature through the classes:

Unannotated → Preliminary → Unreviewed → Standard

Rarely, an entry 'dies'—a few have been removed when they were determined to be erroneous.

A sample DNA sequence entry from the EMBL data library, including annotations as well as sequence data, is the gene for bovine pancreatic trypsin inhibitor (the Box shows part of this entry, omitting most of the sequence itself).

The EMBL Data Library entry for the bovine pancreatic trypsin inhibitor gene

```

ID  BTBPTIG  standard; DNA; MAM; 3998 BP.
XX
AC  X03365; K00966;
XX
DT  18-NOV-1986 (Rel. 10, Created)
DT  20-MAY-1992 (Rel. 31, Last updated, Version 3)
XX
DE  Bovine pancreatic trypsin inhibitor (BPTI) gene
XX
KW  Alu-like repetitive sequence; protease inhibitor;
KW  trypsin inhibitor.
XX
OS  Bos taurus (cattle)
OC  Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
OC  Theria; Eutheria; Artiodactyla; Ruminantia; Pecora; Bovidae.
XX
RN  [1]
RP  1-3998
RA  Kingston I.B., Anderson S.;
RT  "Sequences encoding two trypsin inhibitors occur in strikingly
RT  similar genomic environments";
RL  Biochem. J. 233:443-450(1986).
XX
RN  [2]
RA  Anderson S., Kingston I.B.;
RT  "Isolation of a genomic clone for bovine pancreatic trypsin
RT  inhibitor by using a unique-sequence synthetic dna probe";
RL  Proc. Natl. Acad. Sci. U.S.A. 80:6838-6842(1983).
XX
DR  SWISS-PROT; P00974; BPT1_BOVIN.
XX
CC  Data kindly reviewed (08-DEC-1987) by Kingston I.B.
XX
FH  Key          Location/Qualifiers
FH
FT  misc_feature  795..800
FT                /note="pot. polyA signal"
FT  misc_feature  835..839
FT                /note="pot. polyA signal"
FT  repeat_region 837..847
FT                /note="direct repeat"
FT  misc_feature  930..945
FT                /note="sequence homologous to Alu-like
FT                consensus seq."
FT  repeat_region 1035..1045
FT                /note="direct repeat"
FT  misc_feature  2456..2461
FT                /note="pot. splice signal"
FT  CDS           2470..2736
FT                /note="put. precursor"
FT  misc_feature  2488..2489
FT                /note="pot. intron/exon splice junction"
FT  misc_feature  2506..2507
FT                /note="pot. intron/exon splice junction"
FT  CDS           2512..2685
FT                /note="trypsin inhibitor (aa 1-58)"
FT  misc_feature  2698..2699
FT                /note="pot. exon/intron splice junction"
FT  misc_feature  3690..3695
FT                /note="pot. polyA signal"
FT  misc_feature  3729..3733
FT                /note="pot. polyA signal"
XX
SQ  Sequence 3998 BP; 1053 A; 902 C; 892 G; 1151 T; 0 other;
aattctgata atgcagagaa ctggtaagga gttctgattg ttctgcttga ttaaattgggt
tgtaacagga tagtgctctg tctgatcct agcattcata tgggtgtgtg tctggggcaa
gtcattctgca gtttcttcac ctgaacaggg ggaccagggt acatgagttt cttaaaagat
taccagtcac gagtatgaag agtttacact ttctgatca atgacgtcca tttcccatca

                3720 nucleotides deleted ...

gccagggtcaa accttgggggt gtgttatttc cctgaatt
//

```

A **feature table** (lines beginning FT) is a component of the annotation of an entry that reports properties of specific regions, for instance coding sequences (CDS). Because these are designed to be readable by computer programs—for example, to translate a coding region to an amino acid sequence—they have a more carefully

controlled format and a more restricted vocabulary. Development of controlled vocabularies and a shared dictionary and thesaurus for keywords and feature tables is also important in establishing links between different databases.

The feature table may indicate regions that:

- ◆ perform or affect function
- ◆ interact with other molecules
- ◆ affect replication
- ◆ are involved in recombination
- ◆ are a repeated unit
- ◆ have secondary or tertiary structure
- ◆ are revised or corrected

Genome databases

Although genome sequences form entries in the standard nucleic acid sequence archives, many species have special databases that bring together the genome sequence and its annotation with other data related to the species.

Protein sequence databases

In 2002, three protein sequence databases—The Protein Information Resource, at the National Biomedical Research Foundation of the Georgetown University Medical Center in Washington, DC, USA; and SWISS-PROT and TrEMBL, from the Swiss Institute of Bioinformatics in Geneva and the European Bioinformatics Institute in Hinxton, UK—coordinated their efforts, to form the *UniProt* consortium. The partners in this enterprise share the database but continue to offer separate information retrieval tools for access.

The PIR grew out of the very first sequence database, developed by Margaret O. Dayhoff—the pioneer of the field of bioinformatics. SWISS-PROT was developed at the Swiss Institute of Bioinformatics. TrEMBL contains the translations of genes identified within DNA sequences in the EMBL Data Library. TrEMBL entries are regarded as preliminary, and are converted—after curation and extended annotation—to full SWISS-PROT entries.

Today, almost all amino acid sequence information arises from translation of nucleic acid sequences. Information about ligands, disulphide bridges, subunit associations, post-translational modifications, glycosylation, effects of mRNA editing, etc., are not available from gene sequences. For instance, from genetic information alone one would not know that human insulin is a dimer linked by disulphide bridges. Protein sequence databanks collect this additional information from the literature and provide suitable annotations.

From UniProt, the entry for the amino acid sequence of the protein bovine pancreatic trypsin inhibitor, in SWISS-PROT format, is shown in the Box, pages 126–127. (The comparison of SWISS-PROT format with ENTREZ and PIR formats is the subject of Weblem 3.12.)

Databases associated with SWISS-PROT

Two related databases closely associated with SWISS-PROT are the ENZYME DB, and PROSITE, a set of motifs.

The ENZYME DB stores the following information about enzymes:

- ◆ EC Number: a numerical identifier assigned by the Enzyme Commission (authorized by the International Union of Biochemistry and Molecular Biology; see <http://www.chem.qmw.ac.uk/iubmb/enzyme/>)
- ◆ Recommended name
- ◆ Alternative names, if any
- ◆ Catalytic activity
- ◆ Cofactors, if any
- ◆ Pointers to SWISS-PROT and other data banks
- ◆ Pointers to disease associated with enzyme deficiency if any known

A sample entry in ENZYME DB

```

ID 1.14.17.3
DE PEPTIDYLGLYCINE MONOOXYGENASE.
AN PEPTIDYL ALPHA-AMIDATING ENZYME.
AN PEPTIDYLGLYCINE 2-HYDROXYLASE.
CA PEPTIDYLGLYCINE + ASCORBATE + O(2) = PEPTIDYL(2-HYDROXYGLYCINE) +
CA DEHYDROASCORBATE + H(2)O.
CF COPPER.
CC -!- PEPTIDYLGLYCINES WITH A NEUTRAL AMINO ACID RESIDUE IN THE
CC PENULTIMATE POSITION ARE THE BEST SUBSTRATES FOR THE ENZYME.
CC -!- THE ENZYME ALSO CATALYZES THE DISMUTATION OF THE PRODUCT TO
CC GLYOXYLATE AND THE CORRESPONDING DESGLYCINE PEPTIDE AMIDE.
DR P10731, AMD.BOVIN ; P19021, AMD.HUMAN ; P14925, AMD.RAT ;
DR P08478, AMD1.XENLA; P12890, AMD2.XENLA;

```

The first two characters of each line identify the information that the line contains. For instance, ID = Identification, DE = Description = Official name, AN = Alternate name(s), CA = catalytic activity, CF = cofactor(s), CC = comments, DR = database reference (to SWISS-PROT).

PROSITE contains common patterns of residues of sets of proteins. Such a pattern (or motif, or signature, or fingerprint, or template) appears in a family of related proteins usually because of the requirements of binding sites that constrain the evolution of a protein family. Often they indicate distant relationships not otherwise detectable by comparing sequences. The consensus pattern for inorganic pyrophosphatase is: D- [SGN] -D- [PE] - [LIVM] -D- [LIVMGC]. The three conserved aspartates (D) bind divalent metal cations.

The PIR and associated databases

The PIR maintains several databases about proteins:

- ◆ PIR-PSD: the main protein sequence database
- ◆ iProClass: classification of proteins according to structure and function

Amino acid sequence entry for bovine pancreatic trypsin inhibitor**NiceProt View of Swiss-Prot: P00974****Entry information**

Entry name	BPT1_BOVIN
Primary accession number	P00974
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 01, July 1986
Sequence was last modified in	Release 10, March 1989
Annotations were last modified in	Release 44, June 2004

Name and origin of the protein

Protein name	Pancreatic trypsin inhibitor [Precursor]
Synonyms	Basic protease inhibitor BPI BPTI Aprotinin
Gene name	None
From	Bos taurus (Bovine) [TaxID: 9913]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos.

References

- [1] SEQUENCE FROM NUCLEIC ACID
MEDLINE=87283904; PubMed=2441071;
Creighton T.E., Charles I.G.;
"Sequences of the genes and polypeptide precursors for two bovine
protease inhibitors";
J. Mol. Biol. 194:11-22(1987).
ADDITIONAL REFERENCES DELETED

Comments

- ◆ **FUNCTION:** Inhibits trypsin, kallikrein, chymotrypsin, and plasmin.
- ◆ **SUBCELLULAR LOCATION:** Secreted.
- ◆ **PHARMACEUTICAL:** Available under the name Trasylol (Mile). Used for inhibiting coagulation so as to reduce blood loss during bypass surgery.
- ◆ **SIMILARITY:** Contains 1 BPTI/Kunitz inhibitor domain.
- ◆ **DATABASE:** Name=Trasylol; Note=Clinical information on Trasylol; www="http://www.trasylol.com/".
ADDITIONAL COMMENTS DELETED

Copyright

The Swiss-Prot entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation—the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.isb-sib.ch/announce/> or send an email to license@isb-sib.ch)



Cross-references

EMBL	M20934; AAD13685.1;- ADDITIONAL CROSS-REFERENCES TO EMBL DELETED
PIR	S00277; TIBO.
PDB	1K09; 10-JUL-02. ADDITIONAL CROSS-REFERENCES TO PDB DELETED
InterPro	IPR002223; Kunitz_BPTI.
Pfam	PF00014; Kunitz_BPTI; 1. Pfam graphical view of domain structure.
PRINTS	PR00759; BASICPTASE.
ProDom	PD000222; Kunitz_BPTI; 1. [Domain structure/List of seq. sharing at least 1 domain]
SMART	SM00131; KU; 1.
PROSITE	PS00280; BPTI_KUNITZ_1; 1. PS50279; BPTI_KUNITZ_1; 2. PROSITE graphical view of domain structure.
Implicit links to	HOVERGEN; BLOCKS; ProtoNet; ProtoMap; PRESAGE; DIP; ModBase; SMR; SWISS-2DPAGE; UniRef.

Keywords

Serine protease inhibitor; Signal; Pharmaceutical; 3D-structure.

Features

Key	From	To	Length	Description
SIGNAL	1	21	21	Potential.
PROPEP	22	35	14	
CHAIN	36	93	58	Pancreatic trypsin inhibitor.
PROPEP	94	100	7	
DOMAIN	40	90	51	BPTI/Kunitz inhibitor.
SITE	50	51	2	Reactive bond for trypsin.
DISULFID	40	90		
DISULFID	49	73		
DISULFID	65	86		
HELIX	38	41	4	
STRAND	53	59	7	
TURN	60	63	4	
STRAND	64	70	7	
STRAND	80	80	1	
HELIX	83	90	8	

Sequence information

Length: **100 AA** [This is the length of the unprocessed precursor]

Molecular weight: **10903 Da** [This is the MW of the unprocessed precursor]

CRC64: **6A778A4AD763FB19** [This is a checksum on the sequence]

```

      10           20           30           40           50           60
      |           |           |           |           |           |
MKMSRLCLSV  ALLVLLGTLA  ASTPGCDTSN  QAKAQRPDFC  LEPPYTGPCK  ARTIRYFYNA

      70           80           90           100
      |           |           |           |
KAGLCQTFVY  GGCRAKRNNF  KSAEDCMRTC  GGAIGPWENL

```

- ◆ ASDB: annotation and similarity database; each entry is linked to a list of similar sequences
- ◆ NRL_3D: a database of sequences and annotations of proteins of known structure deposited in the Protein Data Bank
- ◆ ALN: a database of protein sequence alignments
- ◆ RESID: a database of covalent protein structure modifications (recall that important structural features of proteins such as disulphide bridges are not inferrable from gene sequences, and will not appear in protein sequence databases derived solely by translation of genomic data)

The PIR has also created IESA: The Integrated Environment for Sequence Analysis, a site for information retrieval and launching of calculations.

The web server of PIR shows some of the richness of information retrieval tools available. It includes:

- ◆ FETCH DATABASE ENTRY
- ◆ PAIRWISE SEQUENCE ALIGNMENT
- ◆ PROT-FAM: classification by protein family of over 7000 multiple sequence alignments of protein families
- ◆ ATLAS: search text fields of databases, or scan sequences for short peptides
- ◆ ALERT: receive information about new database entries of interest by e-mail, automatically
- ◆ GATEWAY: pattern recognition, homology identification

Databases of structures

Structure databases archive, annotate and distribute sets of atomic coordinates. The major database for biological macromolecular structures is the Protein Data Bank (PDB). It contains structures of proteins, nucleic acids, and a few carbohydrates. Started by the late Walter Hamilton at Brookhaven National Laboratories, Long Island, New York, USA in 1971, the PDB is now managed by the Research Collaboratory for Structural Bioinformatics (RCSB), a distributed organization based at Rutgers University, in New Jersey; the San Diego Supercomputer Center, in California; and the National Institute of Standards and Technology, in Maryland, all in the USA. The parent web site of the Protein Data Bank is at <http://www.rcsb.org>.

The home page of the PDB contains links to the data files themselves, to expository and tutorial material including short news items and the PDB Newsletter, to facilities for deposition of new entries, and to specialized search software for retrieving structures.

Recently, the RCSB, the Molecular Structure Database and the European Bioinformatics Institute, and the Protein Data Bank Japan have formed the Worldwide Protein Data Bank (wwPDB), with the goal of producing a unified archive.

The box shows part of a Protein Data Bank* entry for a structure of *E. coli* thioredoxin.† The information contained includes:

- What protein is the subject of the entry, and what species it came from
- Who solved the structure, and references to publications describing the structure determination
- Experimental details about the structure determination, including information related to the general quality of the result such as resolution of an X-ray structure determination and stereochemical statistics
- The amino acid sequence
- What additional molecules appear in the structure, including cofactors, inhibitors, and water molecules
- Assignments of secondary structure: helix, sheet
- Disulphide bridges
- The atomic coordinates

Protein Data Bank entry 2TRX, *E. Coli* thioredoxin

```

HEADER      ELECTRON TRANSPORT                19-MAR-90  2TRX
COMPND      THIOREDOXIN
SOURCE      (ESCHERICHIA $COLI)
AUTHOR      S.K.KATTI,D.M.LE*MASTER,H.EKLUND
REVDAT     2  15-JAN-93 2TRXA  1      HEADER COMPND
REVDAT     1  15-OCT-91 2TRX   0
JRNL       AUTH  S.K.KATTI,D.M.LE*MASTER,H.EKLUND
JRNL       TITL  CRYSTAL STRUCTURE OF THIOREDOXIN FROM ESCHERICHIA
JRNL       TITL 2 $COLI AT 1.68 ANGSTROMS RESOLUTION
JRNL       REF   J.MOL.BIOL.                V. 212  167 1990
JRNL       REFN  ASTM JMOBAK  UK ISSN 0022-2836                070
REMARK     1
REMARK     1 REFERENCE 1
REMARK     1 AUTH  A.HOLMGREN,B.-*O.SODERBERG,H.EKLUND,C.-*I.BRANDEN
REMARK     1 TITL  THREE-DIMENSIONAL STRUCTURE OF ESCHERICHIA COLI
REMARK     1 TITL 2 THIOREDOXIN-*S=2= TO 2.8 ANGSTROMS RESOLUTION
REMARK     1 REF   PROC.NAT.ACAD.SCI.USA      V. 72  2305 1975
REMARK     1 REFN  ASTM PNAS6  US ISSN 0027-8424                040
REMARK     1 REFERENCE 2
REMARK     1 AUTH  B.-*O.SODERBERG,A.HOLMGREN,C.-*I.BRANDEN
REMARK     1 TITL  STRUCTURE OF OXIDIZED THIOREDOXIN TO 4.5 ANGSTROMS
REMARK     1 TITL 2 RESOLUTION
REMARK     1 REF   J.MOL.BIOL.                V. 90   143 1974
REMARK     1 REFN  ASTM JMOBAK  UK ISSN 0022-2836                070
REMARK     1 REFERENCE 3
REMARK     1 AUTH  A.HOLMGREN,B.-*O.SODERBERG
REMARK     1 TITL  CRYSTALLIZATION AND PRELIMINARY CRYSTALLOGRAPHIC
REMARK     1 TITL 2 DATA FOR THIOREDOXIN FROM ESCHERICHIA $COLI B
REMARK     1 REF   J.MOL.BIOL.                V. 54   387 1970
REMARK     1 REFN  ASTM JMOBAK  UK ISSN 0022-2836                070
REMARK     2
REMARK     2 RESOLUTION. 1.68 ANGSTROMS.
REMARK     3
REMARK     3 REFINEMENT. BY THE RESTRAINED LEAST-SQUARES PROCEDURE OF J.
REMARK     3 KONNERT AND W. HENDRICKSON AS MODIFIED BY B. FINZEL
REMARK     3 (PROGRAM *PROFFT*). THE R VALUE IS 0.165 FOR 25969
REMARK     3 REFLECTIONS IN THE RESOLUTION RANGE 8.0 TO 1.68 ANGSTROMS
REMARK     3 WITH FOBS .GT. 3.0*SIGMA(FOBS)
REMARK     3

```

* Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000), The Protein Data Bank *Nucleic Acids Research*, **28**, 235-242.

† Katti, S. K., LeMaster, D. M. & Eklund, H. (1990), Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution, *J. Mol. Biol.*, **212**, 167-184.

Protein Data Bank entry 2TRX, *E. Coli* thioredoxin (continued)

REMARK 3 RMS DEVIATIONS FROM IDEAL VALUES (THE VALUES OF
 REMARK 3 SIGMA, IN PARENTHESES, ARE THE INPUT ESTIMATED
 REMARK 3 STANDARD DEVIATIONS THAT DETERMINE THE RELATIVE
 REMARK 3 WEIGHTS OF THE CORRESPONDING RESTRAINTS)
 REMARK 3 DISTANCE RESTRAINTS (ANGSTROMS)
 REMARK 3 BOND DISTANCE 0.015(0.020)
 REMARK 3 ANGLE DISTANCE 0.035(0.030)
 REMARK 3 PLANAR 1-4 DISTANCE 0.055(0.050)
 REMARK 3 PLANE RESTRAINT (ANGSTROMS) 0.021(0.020)
 REMARK 3 CHIRAL-CENTER RESTRAINT (ANGSTROMS**3) 0.131(0.150)
 REMARK 3 NON-BONDED CONTACT RESTRAINTS (ANGSTROMS)
 REMARK 3 SINGLE TORSION CONTACT 0.165(0.500)
 REMARK 3 MULTIPLE TORSION CONTACT 0.174(0.500)
 REMARK 3 POSSIBLE HYDROGEN BOND 0.180(0.500)
 REMARK 3 CONFORMATIONAL TORSION ANGLE RESTRAINT (DEGREES)
 REMARK 3 PLANAR (OMEGA) 4.0(3.0)
 REMARK 3 STAGGERED 16.3(15.0)
 REMARK 3 ORTHONORMAL 11.7(20.0)
 REMARK 3 ISOTROPIC THERMAL FACTOR RESTRAINTS (ANGSTROMS**2)
 REMARK 3 MAIN-CHAIN BOND 1.38(1.000)
 REMARK 3 MAIN-CHAIN ANGLE 2.28(1.000)
 REMARK 3 SIDE-CHAIN BOND 1.97(1.000)
 REMARK 3 SIDE-CHAIN ANGLE 3.27(1.500)
 REMARK 4
 REMARK 4 THERE ARE TWO MOLECULES IN THE ASYMMETRIC UNIT. THEY HAVE
 REMARK 4 BEEN ASSIGNED CHAIN INDICATORS *A* AND *B*. THEY HAVE BEEN
 REMARK 4 REFINED INDEPENDENTLY WITHOUT IMPOSING NON-CRYSTALLOGRAPHIC
 REMARK 4 SYMMETRY RESTRAINTS.
 REMARK 5
 REMARK 5 IN ADDITION TO THE METAL COORDINATION SPECIFIED ON CONECT
 REMARK 5 RECORDS BELOW, THERE ARE BONDS TO OD1 AND OD2 OF ASP 10 IN
 REMARK 5 A SYMMETRY-RELATED MOLECULE. DUE TO SOME LIMITATIONS OF
 REMARK 5 PROTEIN DATA BANK FORMAT, THESE BONDS CANNOT BE PRESENTED
 REMARK 5 ON CONECT RECORDS.
 REMARK 6
 REMARK 6 CORRECTION. CORRECT CLASSIFICATION ON HEADER RECORD AND
 REMARK 6 REMOVE E.C. CODE. 15-JAN-93.
 SEQRES 1 A 108 SER ASP LYS ILE ILE HIS LEU THR ASP ASP SER PHE ASP
 SEQRES 2 A 108 THR ASP VAL LEU LYS ALA ASP GLY ALA ILE LEU VAL ASP
 SEQRES 3 A 108 PHE TRP ALA GLU TRP CYS GLY PRO CYS LYS MET ILE ALA
 SEQRES 4 A 108 PRO ILE LEU ASP GLU ILE ALA ASP GLU TYR GLN GLY LYS
 SEQRES 5 A 108 LEU THR VAL ALA LYS LEU ASN ILE ASP GLN ASN PRO GLY
 SEQRES 6 A 108 THR ALA PRO LYS TYR GLY ILE ARG GLY ILE PRO THR LEU
 SEQRES 7 A 108 LEU LEU PHE LYS ASN GLY GLU VAL ALA ALA THR LYS VAL
 SEQRES 8 A 108 GLY ALA LEU SER LYS GLY GLN LEU LYS GLU PHE LEU ASP
 SEQRES 9 A 108 ALA ASN LEU ALA
 SEQRES 1 B 108 SER ASP LYS ILE ILE HIS LEU THR ASP ASP SER PHE ASP
 SEQRES 2 B 108 THR ASP VAL LEU LYS ALA ASP GLY ALA ILE LEU VAL ASP
 SEQRES 3 B 108 PHE TRP ALA GLU TRP CYS GLY PRO CYS LYS MET ILE ALA
 SEQRES 4 B 108 PRO ILE LEU ASP GLU ILE ALA ASP GLU TYR GLN GLY LYS
 SEQRES 5 B 108 LEU THR VAL ALA LYS LEU ASN ILE ASP GLN ASN PRO GLY
 SEQRES 6 B 108 THR ALA PRO LYS TYR GLY ILE ARG GLY ILE PRO THR LEU
 SEQRES 7 B 108 LEU LEU PHE LYS ASN GLY GLU VAL ALA ALA THR LYS VAL
 SEQRES 8 B 108 GLY ALA LEU SER LYS GLY GLN LEU LYS GLU PHE LEU ASP
 SEQRES 9 B 108 ALA ASN LEU ALA
 FTNOTE 1
 FTNOTE 1 RESIDUES PRO A 76 AND PRO B 76 ARE CIS PROLINES.
 FTNOTE 2
 FTNOTE 2 RESIDUES HIS A 6, LEU A 7, ILE A 23, ASP A 47, GLU A 48,
 FTNOTE 2 LEU A 58, LEU A 80, HIS B 6, ASP B 47, LEU B 58, AND
 FTNOTE 2 LEU B 80 HAVE BEEN MODELED AS TWO CONFORMERS.
 FTNOTE 3
 FTNOTE 3 RESIDUES 11 - 21 IN CHAIN B ARE DISORDERED.
 HET CU 109 1 COPPER ++ ION
 HET CU 109 1 COPPER ++ ION
 HET MPD 601 8 2-METHYL-2,4-PENTANEDIOL
 HET MPD 602 8 2-METHYL-2,4-PENTANEDIOL
 HET MPD 603 8 2-METHYL-2,4-PENTANEDIOL
 HET MPD 604 8 2-METHYL-2,4-PENTANEDIOL
 HET MPD 605 8 2-METHYL-2,4-PENTANEDIOL
 HET MPD 606 8 2-METHYL-2,4-PENTANEDIOL
 HET MPD 607 8 2-METHYL-2,4-PENTANEDIOL
 HET MPD 608 8 2-METHYL-2,4-PENTANEDIOL
 FORMUL 3 CU 2(CU1 ++)
 FORMUL 4 MPD 8(C6 H14 O2)
 FORMUL 5 HOH *140(H2 O1)
 HELIX 1 A1A SER A 11 LEU A 17 1 DISORDERED IN MOLECULE B
 HELIX 2 A2A CYS A 32 TYR A 49 1 BENT BY 30 DEGREES AT RES 39
 HELIX 3 A3A ASN A 59 ASN A 63 1
 HELIX 4 31A THR A 66 TYR A 70 5 DISTORTED H-BONDING C-TERMINUS
 HELIX 5 A4A SER A 95 LEU A 107 1
 HELIX 6 A1B SER B 11 LEU B 17 1 DISORDERED IN MOLECULE B
 HELIX 7 A2B CYS B 32 TYR B 49 1 BENT BY 30 DEGREES AT RES 39
 HELIX 8 A3B ASN B 59 ASN B 63 1
 HELIX 9 31B THR B 66 TYR B 70 5 DISTORTED H-BONDING C-TERMINUS
 HELIX 10 A4B SER B 95 LEU B 107 1

```

SHEET 1 B1A 5 LYS A 3 THR A 8 0
SHEET 2 B1A 5 LEU A 53 ASN A 59 1 0 VAL A 55 N ILE A 5
SHEET 3 B1A 5 GLY A 21 TRP A 28 1 N TRP A 28 O LEU A 58
SHEET 4 B1A 5 PRO A 76 LYS A 82 -1 0 THR A 77 N PHE A 27
SHEET 5 B1A 5 VAL A 86 GLY A 92 -1 N GLY A 92 O LYS A 82
SHEET 1 B1B 5 LYS B 3 THR B 8 0
SHEET 2 B1B 5 LEU B 53 ASN B 59 1 0 VAL B 55 N ILE B 5
SHEET 3 B1B 5 GLY B 21 TRP B 28 1 N TRP B 28 O LEU B 58
SHEET 4 B1B 5 PRO B 76 LYS B 82 -1 0 THR B 77 N PHE B 27
SHEET 5 B1B 5 VAL B 86 GLY B 92 -1 N GLY B 92 O LYS B 82
TURN 1 T1A THR A 8 SER A 11 III (TYPE I IN MOLECULE B)
TURN 2 T2A ALA A 29 CYS A 32 I
TURN 3 T3A TYR A 49 LYS A 52 II
TURN 4 T4A GLY A 74 THR A 77 VIB (INCLUDES CIS PRO 76)
TURN 5 T5A LYS A 82 GLU A 85 I'
TURN 6 T1B THR B 8 SER B 11 I (TYPE III IN MOLECULE A)
TURN 7 T2B ALA B 29 CYS B 32 I
TURN 8 T3B TYR B 49 LYS B 52 II
TURN 9 T4B GLY B 74 THR B 77 VIB (INCLUDES CIS PRO 76)
TURN 10 T5B LYS B 82 GLU B 85 I'
SSBOND 1 CYS A 32 CYS A 35
SSBOND 2 CYS B 32 CYS B 35
CRYST1 89.500 51.060 60.450 90.00 113.50 90.00 C 2 8
ORIGX1 1.000000 0.000000 0.000000 0.000000
ORIGX2 0.000000 1.000000 0.000000 0.000000
ORIGX3 0.000000 0.000000 1.000000 0.000000
SCALE1 0.011173 0.000000 0.004858 0.000000
SCALE2 0.000000 0.019585 0.000000 0.000000
SCALE3 0.000000 0.000000 0.018039 0.000000
ATOM 1 N SER A 1 21.389 25.406 -4.628 1.00 23.22
ATOM 2 CA SER A 1 21.628 26.691 -3.983 1.00 24.42
ATOM 3 C SER A 1 20.937 26.944 -2.679 1.00 24.21
ATOM 4 O SER A 1 21.072 28.079 -2.093 1.00 24.97
ATOM 5 CB SER A 1 21.117 27.770 -5.002 1.00 28.27
ATOM 6 OG SER A 1 22.276 27.925 -5.861 1.00 32.61
ATOM 7 N ASP A 2 20.173 26.028 -2.163 1.00 21.39
ATOM 8 CA ASP A 2 19.395 26.125 -0.949 1.00 21.57
ATOM 9 C ASP A 2 20.264 26.214 0.297 1.00 20.89
ATOM 10 O ASP A 2 19.760 26.575 1.371 1.00 21.49
ATOM 11 CB ASP A 2 18.439 24.914 -0.856 1.00 22.14
ATOM 12 CG ASP A 2 19.199 23.629 -0.576 1.00 23.23
ATOM 13 OD1 ASP A 2 20.107 23.371 -1.387 1.00 22.71
ATOM 14 OD2 ASP A 2 18.905 22.959 0.420 1.00 23.61

```

...protein atoms deleted

```

ATOM 844 N ALA A 108 41.357 21.341 9.676 1.00 42.93
ATOM 845 CA ALA A 108 42.151 20.619 10.674 1.00 46.31
ATOM 846 C ALA A 108 42.632 19.312 10.013 1.00 48.21
ATOM 847 O ALA A 108 41.703 18.483 9.767 1.00 49.54
ATOM 848 CB ALA A 108 41.441 20.369 11.988 1.00 46.65
ATOM 849 OXT ALA A 108 43.857 19.249 9.766 1.00 49.19
TER 850 ALA A 108

```

...second chain, and methane, pentane-diol molecules deleted

```

HETATM 1749 O HOH 401 30.339 33.478 16.727 1.00 17.61
HETATM 1750 O HOH 402 29.396 44.583 6.834 0.95 17.71

```

...72 additional water molecules deleted

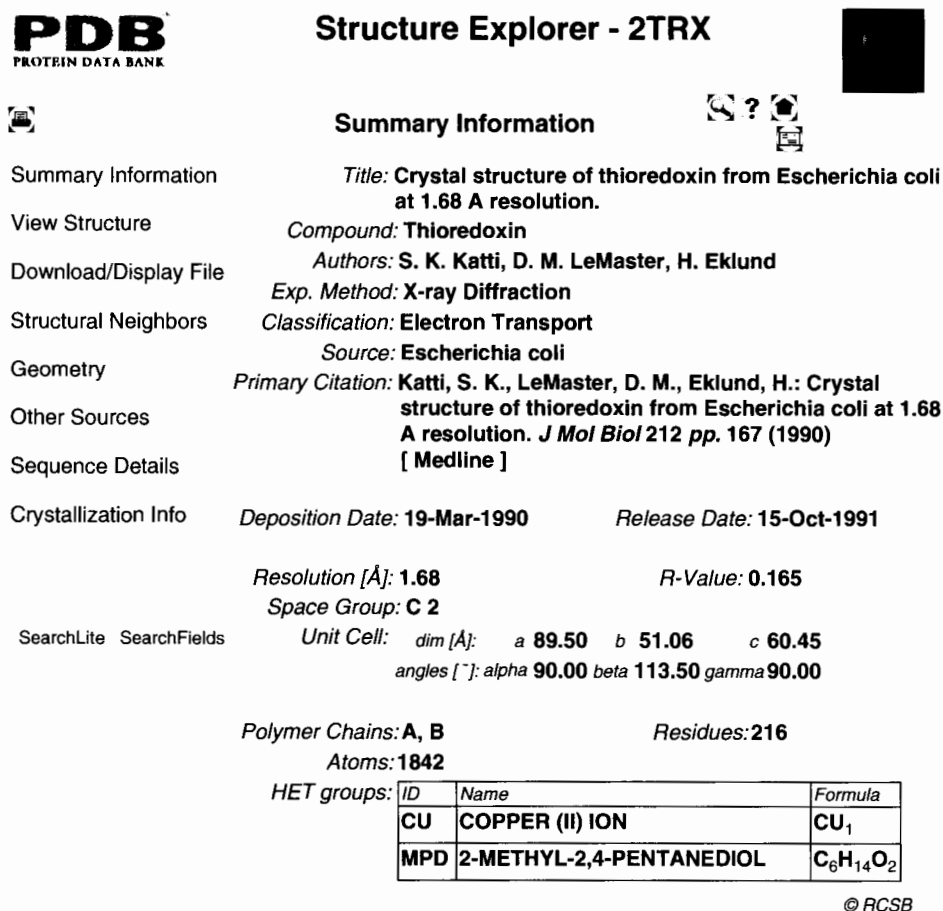
The PDB overlaps in scope with several other databases. The Cambridge Crystallographic Data Centre archives the structures of small molecules; oligonucleotides appear in both the CCDC and PDB. This information is extremely useful in studies of conformations of the component units of biological macromolecules, and for investigations of macromolecule-ligand interactions, including but not limited to applications to drug design. The Nucleic Acid Database (NDB) at Rutgers University, New Brunswick, New Jersey, USA complements the PDB. The BioMagResBank, at the Department of Biochemistry, University of Wisconsin, Madison, Wisconsin, USA, archives protein structures determined by Nuclear Magnetic Resonance.

The archives collect not only the results of structure determinations, but also the measurements on which they are based. The PDB keeps the new data from X-ray structure determinations, and the BioMagRes Bank those from NMR.

The PDB assigns a four-character identifier to each structure deposited. The first character is a number from 1-9. Do not expect mnemonic significance. In many cases several entries correspond to one protein—solved in different states of ligation, or in different crystal forms, or re-solved using better crystals or more accurate data collection techniques. For instance, there have been at least four generations of sperm whale myoglobin crystal structures.

It is easy to retrieve a structure if you know its identifier. From the RCSB home page, entering a PDB ID and selecting Explore gives a 1-page summary of the entry. Figure 3.1 shows the summary page for the thioredoxin structure, identifier 2TRX. Links from this page take you to:

- ◆ The publication in which the entry was described, via the bibliographic database PubMed



PDB
PROTEIN DATA BANK

Structure Explorer - 2TRX

Summary Information

Title: Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution.

Compound: Thioredoxin

Authors: S. K. Katti, D. M. LeMaster, H. Eklund

Exp. Method: X-ray Diffraction

Classification: Electron Transport

Source: *Escherichia coli*

Primary Citation: Katti, S. K., LeMaster, D. M., Eklund, H.: Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J Mol Biol* 212 pp. 167 (1990) [Medline]

Crystallization Info *Deposition Date:* 19-Mar-1990 *Release Date:* 15-Oct-1991

Resolution [Å]: 1.68 *R-Value:* 0.165

Space Group: C 2

Unit Cell: *dim [Å]:* a 89.50 b 51.06 c 60.45

angles [°]: alpha 90.00 beta 113.50 gamma 90.00

Polymer Chains: A, B **Residues:** 216

Atoms: 1842

HET groups:

ID	Name	Formula
CU	COPPER (II) ION	CU ₁
MPD	2-METHYL-2,4-PENTANEDIOL	C ₆ H ₁₄ O ₂

© RCSB

Fig. 3.1 The summary page for the PDB entry 2TRX, *E. coli* thioredoxin.

- ◆ Pictures of the structure (some of these may require that you install a viewing program on your computer)
- ◆ Access to the file containing the entry itself
- ◆ Lists of related structures, according to several different classifications of protein structures
- ◆ Stereochemical analysis—the distribution of bond lengths and angles, and conformational angles
- ◆ Sources of other information about this entry
- ◆ The sequence and secondary structure assignment
- ◆ Details about the crystal form and methods by which the crystals were produced

Fine, if you know the identifier. If not, how do you find it? A simple tool accessible from the PDB home page, called SearchLite, permits a search for keywords. Entering `coli` and `thioredoxin` returns 145 entries, including 2TRX and other crystal structures of the same molecule or mutants, but also several structures of Staphylococcal nuclease, because embedded in the nuclease structure entries is a reference to an article that contains the word `thioredoxin` in the title. The information returned would easily permit you to choose structures to look at or analyse, according to your particular interest in this family of molecules.

The PDB also offers more complex browsers. The Macromolecular Structure Database at the European Bioinformatics Institute (EBI) offers a useful list of facilities for searching and browsing the PDB including a search tool called OCA. OCA is a browser database for protein structure and function, integrating information from numerous databanks. Developed originally by J. Prilusky, OCA is supported by the EBI and is available there and at numerous mirror sites. (The name OCA, in addition to being the Spanish word for goose, has the same relationship to PDB as A. C. Clarke's computer HAL in the movie 2001 has to IBM.)

Another useful information source available at the EBI is the database of Probable Quaternary Structures (PQS) of the biologically active forms of proteins. Often the asymmetric unit of the crystal structure, as deposited in the PDB entry, contains only part of the active unit, or alternatively multiple copies of the active unit. In many cases it is not obvious how to go from the deposited entry to the active form, and this information is available in PQS.

Indicators of structure quality

X-ray crystal structure analysis produces estimates of the positions of the atoms in a molecule and of their effective sizes, known as **B-factors**. An important feature of the experimental data (the absolute values of the Fourier coefficients of the electron density) is that all atoms contribute to all observations. It is difficult to estimate errors in individual atomic positions.

Crystal structure determinations are at the mercy of the degree of order in different parts of the molecule. (Order is the extent to which different unit cells of the crystal are exact copies of one another.)

The degree of order governs the available **resolution** of the experimental data. Resolution is an index of potential quality of an X-ray structure determination. It measures the ratio of the number of parameters to be determined to the number of observations. In structure determinations of small organic molecules or of minerals, this ratio is usually generous: ~ 10 . But for a typical protein crystal:

	Low resolution			...	High	
Resolution in Å	4.0	3.5	3.0	2.5	2.0	1.5
Ratio of observations to parameters	0.3	0.4	0.6	1.1	2.2	3.8

(Resolution measures the fineness of the details that can be distinguished, hence the lower the number, the higher the resolution.)

In addition to disorder, errors in crystal structures reflect both errors in data and errors in solving the structure. A comparison of four independently-solved structures of interleukin-1 β showed an average variation in atomic position of 0.84 Å, higher than the expected experimental error.

Many crystallographers deposit their experimental data along with the solved structures. This permits detailed checks on the results. But in many cases the experimental data are not available. How can one then assess the quality of a structure? B-factors provide important clues; high B-factors in an entire region suggest that the region has not been well-determined. This usually reflects imperfect order in the crystal. Programs can flag stereochemical outliers—exceptions to regularities common to well-determined protein structures. The entries corresponding to the PDB entries in www.cmbi.kun.nl/gv/pdbreport describe diagnostic analysis and identification of problems and outliers.

But although outliers are relatively easy to *detect*, it is difficult to decide whether they are correct but unusual features of the structure, or the result of errors in building the model, or the inevitable result of crystal disorder. Proper assessment requires access to the experimental data; and fixing real errors may well require the attention of an experienced crystallographer. The conclusion seems inescapable that structure factors should be archived and available.

Nuclear Magnetic Resonance (NMR)

NMR is the second major technique for determining macromolecular structure. It produces structures that are generally correct in topology but not as precise as a good X-ray structure determination and therefore less useful for the study of fine structural details. Crystallographers report a single structure, or only a small number. NMR spectroscopists usually produce a family of ~ 10 – 20 related structures or even more, calculated from the same experimental data. Comparison across such an ensemble indicates precision; regions in which the local variation in structure is small are well defined by the data. This is a rough equivalent of the crystallographer's B-factor.

**Web resources: Protein and nucleic acid structures****Home page of Protein Data Bank:**

<http://www.rcsb.org>

Home page of EBI Macromolecular Structure Database:

<http://msd.ebi.ac.uk/>

Home page of BioMagResBank:

<http://www.bmrb.wisc.edu/>

Searching the Protein Data Bank:**Home page of SCOP (Structural Classification of Proteins):**

<http://scop.mrc-lmb.cam.ac.uk/scop/>

List of browsers: http://pdb-browsers.ebi.ac.uk/browse_it.shtml

OCA: <http://oca.ebi.ac.uk/oca-bin/ocamain>

Database of Protein Quaternary Structure:

<http://pqs.ebi.ac.uk/>

Reports of structure quality:

<http://www.cmbi.kun.nl/gv/pdbreport>

Classifications of protein structures

Several web sites offer hierarchical classifications of the entire Protein Data Bank according to the folding patterns of the proteins (see Chapter 5):

- SCOP: Structural Classification of Proteins
- CATH: Class/Architecture/Topology/Homologous superfamily?
- DALI: based on extraction of similar structures from distance matrices
- CE: a database of structural alignments

These sites are useful general entry points to protein structural data. For instance, SCOP offers facilities for searching on keywords to identify structures, navigation up and down the hierarchy, generation of pictures, access to the annotation records in the PDB entries, and links to related databases.

Specialized, or 'boutique' databases

Many individuals or groups select, annotate, and recombine data focused on particular topics, and include links affording streamlined access to information about subjects of interest.

For instance, the protein kinase resource is a specialized compilation that includes sequences, structures, functional information, laboratory procedures, lists of interested scientists, tools for analysis, a bulletin board, and links.

The HIV protease database archives structures of Human Immunodeficiency Virus 1 proteinases, Human Immunodeficiency Virus 2 proteinases, and Simian Immunodeficiency Virus Proteinases, and their complexes; and provides tools for their analysis and links to other sites with AIDS-related information. This database contains some crystal structures not deposited in the PDB.

In the field of immunology:

- ◆ IMGT, the international ImMunoGeneTics database, is a high-quality integrated database specializing in Immunoglobulins, T-cell receptors and Major Histocompatibility Complex (MHC) molecules of all vertebrate species. The IMGT server provides a common access to all Immunogenetics data. At present, it includes two databases: IMGT/LIGM-DB, a comprehensive database of immunoglobulin and T-cell receptor gene sequences from human and other vertebrates, with translation for fully annotated sequences; and IMGT/HLA-DB, a database of the human MHC referred to as HLA (Human Leucocyte Antigens).
- ◆ KABAT—Database of Sequences of Proteins of Immunological Interest—North-Western University (USA)
- ◆ MHCPEP—Major Histocompatibility Complex Binding Peptides Database—WEHI (Melbourne, Australia)



Web resources: Databases for specific protein families

Protein kinases:

<http://www.sdsc.edu/kinases/>

HIV proteases:

<http://www-fbnc.ncifcrf.gov/HIVdb/>

Immunology:

IGMT: <http://imgt.cines.fr>

KABAT: <http://immuno.bme.nwu.edu/>

MHCPEP: <http://wehih.wehi.edu.au/mhcpep/>

Expression and proteomics databases

Recall the central dogma: DNA makes RNA makes protein. Genomic databases contain DNA sequences. Expression databases record measurements of *mRNA* levels, usually via ESTs (expressed sequence tags—short terminal sequences of cDNA synthesized from mRNA) describing patterns of gene transcription. Proteomics databases record measurements on *proteins*, describing patterns of gene translation.

Comparisons of expression patterns give clues to: (1) the function and mechanism of action of gene products, (2) how organisms coordinate their control over metabolic processes in different conditions—for instance yeast under aerobic or anaerobic conditions, (3) the variations in mobilization of genes at different stages of the cell cycle, or of the development of an organism, (4) mechanisms of antibiotic resistance in bacteria, and consequent suggestion of targets for drug development (5) the response to challenge by a parasite, (6) the response to medications of different types and dosages, to guide effective therapy.

There are many databases of ESTs. In most, the entries contain fields indicating tissue of origin and/or subcellular location, state of development, conditions of

growth, and quantitation of expression level. Within GenBank the dbEST collection currently contains almost 23 million entries, from 719 species, led by:

Species with largest number of entries in dbEST

Species	Number of entries
<i>Homo sapiens</i> (human)	5 654 825
<i>Mus musculus</i> + <i>domesticus</i> (mouse)	4 235 142
<i>Ciona intestinalis</i> (primitive chordate)	684 280
<i>Rattus</i> sp. (rat)	636 658
<i>Triticum aestivum</i> (wheat)	559 149
<i>Danio rerio</i> (zebrafish)	532 545
<i>Gallus gallus</i> (chicken)	494 605
<i>Bos taurus</i> (cattle)	465 743
<i>Zea mays</i> (maize)	415 211
<i>Xenopus tropicalis</i>	392 901
<i>Xenopus laevis</i> (African clawed frog)	385 714
<i>Drosophila melanogaster</i> (fruit fly)	382 439
<i>Hordeum vulgare</i> + subsp. <i>vulgare</i> (barley)	356 856
<i>Glycine max</i> (soybean)	334 668
<i>Sus scrofa</i> (pig)	328 573
<i>Arabidopsis thaliana</i> (thale cress)	322 641
<i>Caenorhabditis elegans</i> (nematode)	298 805
<i>Oryza sativa</i> (rice)	284 007

Some EST collections are specialized to particular tissues (e.g. muscle, teeth) or to species. In many cases there is an effort to link expression patterns to other knowledge of the organism. For instance, the Jackson Lab Gene Expression Information Resource Project for Mouse Development coordinates data on gene expression and developmental anatomy.

Many databases provide connections between ESTs in different species, for instance, linking human and mouse homologues, or relationships between human disease genes and yeast proteins. Other EST collections are specialized to a type of protein, for instance, cytokines. A large effort is focussed on cancer: integrating information on mutations, chromosomal rearrangements, and changes in expression patterns, to identify genetic changes during tumour formation and progression.

Although of course there is a close relationship between patterns of transcription and patterns of translation, direct measurements of protein contents of cells and tissues—proteomics—provides additional valuable information. Because of differential rates of translation of different mRNAs, measurements of proteins directly give a more accurate description of patterns of gene expression than measurements of transcription. Post-translational modifications can be detected *only* by examining the proteins.

Proteome analysis involves separation, identification, and determination of the quantitative amounts of proteins in a sample (see Chapter 6). Proteome databases store images of gels, and their interpretation in terms of protein patterns. For each protein, an entry typically records (see Weblem 3.21):

- ◆ identification of protein
- ◆ relative amount
- ◆ function
- ◆ mechanism of action
- ◆ expression pattern
- ◆ subcellular localization
- ◆ related proteins
- ◆ post-translational modifications
- ◆ interactions with other proteins
- ◆ links to other databases

Bioinformatics is contributing to the development of these databases, and also to the development of algorithms for comparing and analysing the patterns they contain.

Databases of metabolic pathways

The Kyoto Encyclopedia of Genes and Genomes (KEGG) collects individual genomes, gene products and their functions, but its special strengths lie in its integration of biochemical and genetic information. KEGG focuses on interactions: molecular assemblies, and metabolic and regulatory networks. It has been developed under the direction of M. Kanehisa.

KEGG organizes five types of data into a comprehensive system:

1. Catalogues of chemical compounds in living cells
2. Gene catalogues
3. Genome maps
4. Pathway maps
5. Orthologue tables

The catalogues of chemical compounds and genes—items 1 and 2—contain information about particular molecules or sequences. Item 3, genome maps, integrates the genes themselves according to their appearance on chromosomes. In some cases knowing that a gene appears in an operon can provide clues to its function.

Item 4, the pathway maps, describe potential networks of molecular activities, both metabolic and regulatory. A metabolic pathway in KEGG is an idealization corresponding to a large number of possible metabolic cascades. It can generate a real metabolic pathway of a particular organism, by matching the proteins of that organism to enzymes within the reference pathways.

One enzyme in one organism would be referred to in KEGG in its orthologue tables, item 5, which link the enzyme to related ones in other organisms. This permits analysis of relationships between the metabolic pathways of different organisms.

KEGG derives its power from the very dense network of links among these categories of information, and additional links to many other databases to which the system maintains access. Two examples of the kinds of questions that can be treated by KEGG are:

- ◆ It has been suggested that simple metabolic pathways evolve into more complex ones by gene duplication and subsequent divergence. Searching the pathway catalogue for sets of enzymes that share a folding pattern will reveal clusters of paralogues.
- ◆ KEGG can take the set of enzymes from some organism and check whether they can be integrated into known metabolic pathways. A gap in a pathway suggests a missing enzyme or an unexpected alternative pathway.

Bibliographic databases

MEDLINE (based at the US National Library of Medicine) integrates the biomedical literature, including very many papers dealing with subjects in molecular biology not overtly clinical in content. It is included in PubMed, a bibliographical database offering abstracts of scientific articles, integrated with other information retrieval tools of the National Center for Biotechnology Information (NCBI) within the National Library of Medicine (<http://www.ncbi.nlm.nih.gov/PubMed/>).

One very effective feature of PubMed is the option to retrieve *related articles*. This is a very quick way to 'get into' the literature of a topic. Combined with the use of a general search engine for web sites that do not correspond to articles published in journals, fairly comprehensive information is readily available about most subjects. Here's a tip: if you are trying to start to learn about an unfamiliar subject, try adding the keyword *tutorial* to your search in a general search engine, or the keyword *review* to your search in PubMed.

Almost all scientific journals now place their tables of contents, and in many cases their entire issues, on web sites. The US National Institutes of Health have established a centralized web-based library of scientific articles, called PubMed Central (<http://www.pubmedcentral.nih.gov/>). In collaboration with scientific journals, the NCBI is organizing the electronic distribution of the full texts of published articles.

A new organization, the Public Library of Science, has the goal of making the scientific (including medical) literature publicly and freely accessible. A non-profit organization, the Public Library of Science has received support from foundations for its efforts in distributing literature published by others, and to start its own publications, which will permit exploration of different relationships—including but not limited to economic ones—between authors, publishers and readers.

Surveys of molecular biology databases and servers

It is difficult to explore any topic in molecular biology on the web without quickly bumping into a list of this nature. Lists of web resources in molecular biology are very common. They contain, to a large extent, the same information,

but vary widely in their 'look and feel' aspects. The real problem is that unless they are curated they tend to degenerate into lists of dead links. (A draft of this section contained a reference to a web site that contained a reasonable survey. Returning to it two months later, the name of the site had changed, and over half of the sites listed had disappeared.)

This book does not contain a long annotated list of relevant and recommended sites, for the following reasons: (1) You don't want a long list, you need a short one. (2) The Web is too volatile for such a list to stay useful for very long. *It is much more effective to use a general search engine to find what you want at the moment you want it.* Each year the January issue of the journal *Nucleic Acids Research* contains a set of articles on databases in molecular biology. This is an invaluable reference.

Moreover, the content of the databases is expanding all the time. If you try the searches described in examples in this chapter you will obtain more 'hits' than the results printed here. (Indeed, I have not hesitated to use older sets of results if, because they contain more variety than the latest results, they seem more informative. The problem of suppressing extensive redundancy in responses to websearches is a challenge for research in the field of information retrieval.)

My advice is: spend some time browsing; it won't take you long to find a site that appears reasonably stable and has a style compatible with your methods of work. Alternatively, here's a site that is comprehensive and shows signs of a commitment to keeping it up to date: <http://www.expasy.org/alinks.html>. It is a suitable site for starting a browsing session.

Gateways to archives

Databases of nucleic acid and protein sequences maintain facilities for a very wide variety of information retrieval and analysis operations. Categories of these operations include:

1. **Retrieval of sequences from the database** Sequences can be 'called up' either on the basis of features of the annotations, or by patterns found within the sequences themselves.
2. **Sequence comparison** This is not a facility, this is a heavy industry! It was introduced in Chapter 1 and will be discussed in detail in Chapter 4. It includes the very important searches for relatives.
3. **Translation of DNA sequences to protein sequences**
4. **Simple types of structure analysis and prediction** For example, statistical methods for predicting the secondary structure of proteins from sequences alone, including hydrophobicity profiles—from which the transmembrane proteins can generally be identified (see page 193).
5. **Pattern recognition** It is possible to search for all sequences containing a pattern or combination of patterns, expressed as probabilities for finding certain sets of residues at consecutive positions. In DNA sequences, these may be

recognition sites for enzymes such as those responsible for splicing interrupted genes. In proteins, short and localized patterns sometimes identify molecules that share a common function even if there is no obvious overall relationship between their sequences. PROSITE is a collection of these protein 'signature' patterns.

6. **Molecular graphics** is necessary to provide intelligible depictions of very complicated systems. Typical applications of molecular graphics include:
- ◆ Mapping residues believed to be involved in function, onto the three-dimensional framework of a protein. Often this will isolate an active site.
 - ◆ Classifying and comparing the folding patterns of proteins.
 - ◆ Analysing changes between closely-related structures, or between two conformational states of a single molecule,
 - ◆ Studying the interaction of a small molecule with a protein, in order to attempt to assign function, or for drug development,
 - ◆ Interactive fitting of a model to the noisy and fuzzy image of the molecule that arises initially from the measurements in solving protein structures by X-ray crystallography.
 - ◆ Design and modelling of new structures.

Access to databases in molecular biology

How to learn web skills

It would be difficult to learn to ride a bicycle by reading a book describing the sets of movements required, much less one about the theory of the gyroscope. Similarly, the place to learn web skills is at a terminal, running a browser. True enough, but there is always a certain initial period of difficulty and imbalance. Here the goal is only to provide some temporary assistance to get you started. Then, off you go!

This section contains introductions to some of the major databanks and information retrieval systems in molecular biology. In each case we show relatively simple searches and applications. When appropriate, unique features of each system will be emphasized.

ENTREZ

The National Center for Biotechnology Information, a component of the United States National Library of Medicine, maintains databases and avenues of access to them. ENTREZ offers access via the following database divisions:

- ◆ Protein
- ◆ Peptide
- ◆ Nucleotide
- ◆ Structure
- ◆ Genome

- ◆ Popset—information about populations
- ◆ OMIM—Online Mendelian Inheritance in Man

Links between various databases are a strong point of NCBI's system. The starting point for retrieval of sequences and structures is called ENTREZ: <http://www.ncbi.nlm.nih.gov/Entrez/>.

Let us pick a molecule—human neutrophil elastase—and search for relevant entries in the different sections of ENTREZ.

Search in ENTREZ protein database

Go to <http://www.ncbi.nlm.nih.gov/Entrez/>. Select Protein: sequence database, enter the search terms HUMAN ELASTASE and click on GO.

The Box shows fifteen answers returned by the program. (In a browser, you will also find links to the sequence databank entries.) The top hit is ELASTASE 1 PRECURSOR [HOMO SAPIENS]; other responses include elastases from other species, inhibitors from human and from leech, and tyrosyl-tRNA synthetase. (Why should a leech protein and tRNA synthetase show up in a search for human elastase? See Weblem 3.9.) Later we shall see how to tune the query to eliminate these extraneous responses.

ENTREZ responses to *human elastase* in PROTEIN database

1. elastase 1 precursor [Homo sapiens]
gi—4731318—gb—AAD28441.1—AF120493_1[4731318]
2. ALPHA-1-ANTITRYPSIN PRECURSOR (ALPHA-1 PROTEASE INHIBITOR)
(ALPHA-1-ANTIPROTEINASE)
gi—1703025—sp—P01009—A1AT_HUMAN[1703025]
3. elastase [Mus musculus]
gi—7657060—ref—NP_056594.1—[7657060]
4. proteinase 3 [Mus musculus]
gi—6755184—ref—NP_035308.1—[6755184]
5. ANTIMICROBIAL PEPTIDE ENAP-2
gi—7674025—sp—P56928—ENA2_HORSE[7674025]
6. AMBP PROTEIN PRECURSOR [CONTAINS: ALPHA-1-MICROGLOBULIN
(PROTEIN HC)
(COMPLEX-FORMING GLYCOPROTEIN HETEROGENEOUS IN CHARGE);
INTER-ALPHA-TRYPSIN INHIBITOR LIGHT CHAIN (ITI-LC) (BIKUNIN) (HI-30)]
gi—122801—sp—P02760—AMBP_HUMAN[122801]
7. ELAFIN PRECURSOR (ELASTASE-SPECIFIC INHIBITOR) (ESI) (SKIN-DERIVED
ANTILEUKOPROTEINASE) (SKALP)
gi—119262—sp—P19957—ELAF_HUMAN[119262]
8. ANTILEUKOPROTEINASE
gi—113637—sp—P22298—ALK1_PIG[113637]





9. ANTILEUKOPROTEINASE 1 PRECURSOR (ALP) (HUSI-1) (SEMINAL PROTEINASE INHIBITOR) (SECRETORY LEUKOCYTE PROTEASE INHIBITOR) (BLPI) (MUCUS PROTEINASE INHIBITOR) (MPI)
gi-113636-sp-P03973-ALK1_HUMAN[113636]
10. ALPHA-2-MACROGLOBULIN PRECURSOR (ALPHA-2-M)
gi-112911-sp-P01023-A2MG_HUMAN[112911]
11. tyrosyl-tRNA synthetase [Homo sapiens]
gi-4507947-ref-NP_003671.1-[4507947]
12. pancreatic elastase IIB [Homo sapiens]
gi-7705648-ref-NP_056933.1-[7705648]
13. protease inhibitor 3, skin-derived (SKALP) [Homo sapiens]
gi-4505787-ref-NP_002629.1-[4505787]
14. pancreatic elastase I (allele HEL1-36)—human (fragment)
gi-7513237-pir-S70441[7513237]
15. guamerin—Korean leech

The format of the responses is as follows. In each case, the first line gives the name and synonyms of the molecule, and the species of origin. Note that Greek letters are spelt out. The last line gives references to the source databanks: gi = GenInfo Identifier, (see page 25), gb = GenBank accession number, sp = Swiss-Prot, pir = Protein Identification Resource, ref = the Reference Sequence project of NCBI. The entries retrieved include elastases from human and other species, and also inhibitors of elastase.

Opening the entry corresponding to the first hit retrieves the file shown in the next Box. The first lines are mostly database housekeeping—accession numbers, molecule name, date of deposition, etc. Then descriptive material such as the source, this case human, with the full taxonomic classification, credit to the scientists who deposited the entry, and literature references. Finally the particular scientific information: the location of the gene, and its product (CDS = coding sequence), and the sequence itself (see Exercise 3.2).

Searches in ENTREZ nucleotide database

We next look again for HUMAN ELASTASE, this time in the Nucleotide database. Let us try to tune the search, to eliminate the responses that refer to elastase inhibitors.

1. Select NUCLEOTIDE at the ENTREZ site.
2. Click on LIMITS, select ORGANISM from the pulldown menu, type HOMO SAPIENS in the search box.
3. Next select SUBSTANCE NAME from the pulldown menu, and then type AND ELASTASE in the search box.

Top result of search for human elastase in ENTREZ Protein database

```

LOCUS          AF120493_1    258 aa                      PRI          03-AUG-2000
DEFINITION    elastase 1 precursor [Homo sapiens].
ACCESSION    AAD28441
PID          g4731318
VERSION      AAD28441.1    GI:4731318
DBSOURCE     locus AF120493 accession AF120493.1
KEYWORDS     .
SOURCE       human.
  ORGANISM   Homo sapiens
             Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
             Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE    1 (residues 1 to 258)
  AUTHORS    Talas,U., Dunlop,J., Khalaf,S., Leigh,I.M. and Kelsell,D.P.
  TITLE      Human elastase 1: evidence for expression in the skin and the
             identification of a frequent frameshift polymorphism
  JOURNAL    J. Invest. Dermatol. 114 (1), 165-170 (2000)
  MEDLINE   20087075
  PUBMED    10620133
REFERENCE    2 (residues 1 to 258)
  AUTHORS    Talas,U., Dunlop,J., Leigh,I.M. and Kelsell,D.P.
  TITLE      Direct Submission
  JOURNAL    Submitted (15-JAN-1999) Centre for Cutaneous Research, Queen Mary
             and Westfield College, 2 Newark Street, London E1 2AT, UK
COMMENT      Method: conceptual translation supplied by author.
FEATURES     Location/Qualifiers
   source    1..258
             /organism="Homo sapiens"
             /db_xref="taxon:9606"
             /chromosome="12"
             /map="12q13"
             /cell_type="keratinocyte"
   Protein   1..258
             /product="elastase 1 precursor"
   CDS       1..258
             /gene="ELA1"
             /coded_by="AF120493.1:42..818"
ORIGIN
   1 mlvlyghstq dlpetnarvv ggteagrns w psqislqyrs ggsryhtcgg tllrqnvwmt
   61 aahcvdyqkt frvvagdhn l sqndgteqyv svqkivvhpy wnsdnvaagy diallrlaqs
  121 vtlnsyvqlg vlpqegaila nnspeyitgw gkktngqla qtlqqaylps vdyaicssss
  181 ywgstvkntm vcaggdgvr s gcqgdsggpl hclvngkysl hgvtsfvssr gcnvsrktv
  241 ftqvsayisw innviasn
//

```

4. Finally select TEXT WORD from the pulldown menu, and then type NOT INHIBITOR in the search box. Now click on GO.

If you click on Details, you will find:

```

HOMO SAPIENS[ORGANISM] AND ELASTASE[SUBSTANCE NAME] NOT INHIBITOR
[TEXT WORD]

```

The search returns over 400 hits, including many individual clones. The top hit (see Box) is: HOMO SAPIENS ELASTASE 1 PRECURSOR (ELA1) MRNA, COMPLETE CDS. The term 'complete cds' means complete coding sequence.

Compare this file with the result of searching in the Protein database (see Exercise 3.5).

Top result of search for human elastase in ENTREZ Nucleotide database

LOCUS AF120493 952 bp mRNA PRI 03-AUG-2000
 DEFINITION Homo sapiens elastase 1 precursor (ELA1) mRNA, complete cds.
 ACCESSION AF120493
 VERSION AF120493.1 GI:4731317
 KEYWORDS .
 SOURCE human.
 ORGANISM Homo sapiens
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
 Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 952)
 AUTHORS Talas,U., Dunlop,J., Khalaf,S., Leigh,I.M. and Kelsell,D.P.
 TITLE Human elastase 1: evidence for expression in the skin and the
 identification of a frequent frameshift polymorphism
 JOURNAL J. Invest. Dermatol. 114 (1), 165-170 (2000)
 MEDLINE 20087075
 PUBMED 10620133

REFERENCE 2 (bases 1 to 952)
 AUTHORS Talas,U., Dunlop,J., Leigh,I.M. and Kelsell,D.P.
 TITLE Direct Submission
 JOURNAL Submitted (15-JAN-1999) Centre for Cutaneous Research, Queen Mary
 and Westfield College, 2 Newark Street, London E1 2AT, UK

FEATURES Location/Qualifiers
 source 1..952
 /organism="Homo sapiens"
 /db_xref="taxon:9606"
 /chromosome="12"
 /map="12q13"
 /cell_type="keratinocyte"
 gene 1..952
 /gene="ELA1"
 CDS 42..818
 /gene="ELA1"
 /codon_start=1
 /product="elastase 1 precursor"
 /protein_id="AAD28441.1"
 /db_xref="GI:4731318"
 /translation="MLVLYGHSTQDLPETNARVVGGTEAGRNSWPSQISLQYRSGGSR
 YHTCGGTLIRQNWVMTAAHCVDYQKTRFRVAVGDHNSQNDGTEQYVSVQKIVVHPYWN
 SDNVAAGYDIALLRLAQSRTLNSYVQLGVLPQEGAILANNSPCYITGWGKTKTNGQLA
 QTLQQAYLPSVDYAISSSSYWGSTVKNTMVCAGGDGVRSGCQGDSSGGPLHCLVNGKY
 SLHGVTSTFVSSRGCNVSRKPTVFTQVSAYISWINNVIASN"

BASE COUNT 226 a 261 c 250 g 215 t
 ORIGIN
 1 ttggtccaag caagaaggca gttgtctact ccatcgcaa catgctggtc ctttatggac
 61 acagcaccca ggaccttcg gaaaccaatg cccgcgtagt cggagggatc gaggccggga
 121 ggaattcctg gccctctcag attccctcc agtaccggtc tggaggttcc cggtatcaca
 181 cctgtggagg gacccttacc agacagaact gggatgatgac agctgctcac tgcgtggatt
 241 accagaagac tttccgcgtg gttgctggag accataacct gagccagaat gatggcactg
 301 agcagtacgt gagtgtgacg aagatcgtgg tgcattccata ctggaacagc gataacgtgg
 361 ctgcccggcta tgacatgcc cttgctgccc tggcccagag cgttaccctc aatagctatg
 421 tccagctggg tgttctgccc caggagggag ccatcctggc taacaacagt cctgctaca
 481 tcacaggtct gggcaagacc aagaccaatg ggcagctggc ccagaccctg cagcaggctt
 541 acctgccctc tgtgactat gccatctgct ccagctcctc ctactggggc tccactgtga
 601 agaacacccat ggtgtgtgct ggtggagatg gatttcgctc tggatgccag ggtgactctg
 661 gggggccccct acctgtcttg gtgaatggca agtattctct ccatggagtg accagcttgg
 721 tgtccagccg cggctgtaat gtcctccagga agcctacagt cttcaccag gtctctgctt
 781 acatctcctg gataaataat gtcattgcct ccaactgaac attttctga gtccaacgac
 841 cttcccaaaa tgggtcttag atctgcaata ggacttgcca tcaaaaagta aaacacattc
 901 tgaaagacta ttgagccatt gatagaaaag caataaaac tagatataca tt

//

Searches in ENTREZ genome database

A search for HUMAN ELASTASE returns:

1. NC_000967 CAENORHABDITIS ELEGANS CHROMOSOME III[64] LCL—WORM_CHR_III
2. NC_001099 HOMO SAPIENS CHROMOSOME 19[19] REF—NC_001099—HSAP-19
3. NC_001065 HOMO SAPIENS CHROMOSOME 14[14] REF—NC_001065—HSAP-14
4. NC_001044 HOMO SAPIENS CHROMOSOME 11[11] REF—NC_001044—HSAP-11
5. NC_001008 HOMO SAPIENS CHROMOSOME 6[6] REF—NC_001008—HSAP-6

Why should a *C. elegans* protein appear in a search for human elastase? The entry NC_000967 is chromosome III of *C. elegans* in its entirety. Comments on one of the genes detected include:

```
gene="T07A5.1" /note="weak similarity with elastase (PIR accession number A406659)"
```

Many other genes in *C. elegans* are annotated with similarities to human proteins. However, although *C. elegans* does contain an elastase, this is *not* flagged as similar to human elastase, although it is a homologue.

Searches in ENTREZ structure database

Is the three-dimensional structure of human elastase known? Select the STRUCTURE database, from the choices to the left of the query box, and rerun the search. The program returns at least five answers:

- | | |
|------|---|
| 1JK3 | CRYSTAL STRUCTURE OF HUMAN MMP-12 (MACROPHAGE ELASTASE) AT TRUE ATOMIC RESOLUTION |
| 1HAZ | SNAPSHOTS OF SERINE PROTEASE CATALYSIS: (C) ACYL-ENZYME INTERMEDIATE BETWEEN PORCINE PANCREATIC ELASTASE AND HUMAN BETA-CASOMORPHIN-7 JUMPED TO PH 9 FOR 1 MINUTE |
| 1HAX | SNAPSHOTS OF SERINE PROTEASE CATALYSIS: (A) ACYL-ENZYME INTERMEDIATE BETWEEN PORCINE PANCREATIC ELASTASE AND HUMAN BETA-CASOMORPHIN-7 AT PH 5 |
| 1BOF | CRYSTAL STRUCTURE OF HUMAN NEUTROPHIL ELASTASE WITH MDL 101, 146 |
| 1QIX | PORCINE PANCREATIC ELASTASE COMPLEXED WITH HUMAN BETA- CASOMORPHIN-7 |

The designations 1JK3, 1HAZ, 1HAZ, 1BOF, and 1QIX are entry codes from the Protein Data Bank.

OOPS!—we may not realize it, but we have missed many useful entries. There are many elastase structures solved in complex with inhibitors, which we have asked the system to reject. Deleting NOT INHIBITORS and rerunning the query returns several more structures.

Searches in the bibliographic database PubMed

Perhaps it is time to look at what people have had to say about our molecule. Of course the literature on elastase is huge. A search in PubMed for HUMAN ELASTASE returns over 7500 entries. To prune the results, let us try to find citations to articles describing the role of elastase in disease. A search for HUMAN ELASTASE DISEASE returns over 1600 entries. What about specific elastase **mutants** related to human disease? A search for HUMAN ELASTASE DISEASE MUTATION returns more than 40 articles, in reverse chronological order. Here are 10 of them:

Hermans MH, Touw IP. Significance of neutrophil elastase mutations versus G-CSF receptor mutations for leukemic progression of congenital neutropenia. *Blood*. 2001 Apr 1;97(7):2185-6. No abstract available.

Li FQ, Horwitz M. Characterization of mutant neutrophil elastase in severe congenital neutropenia. *J Biol Chem*. 2001 Apr 27;276(17):14230-41.

Ye S. Polymorphism in matrix metalloproteinase gene promoters: implication in regulation of gene expression and susceptibility of various diseases. *Matrix Biol*. 2000 Dec;19(7):623-9. Review.

Dale DC, Person RE, Bolyard AA, Aprikyan AG, Bos C, Bonilla MA, Boxer LA, Kannourakis G, Zeidler C, Welte K, Benson KF, Horwitz M. Mutations in the gene encoding neutrophil elastase in congenital and cyclic neutropenia. *Blood*. 2000 Oct 1;96(7):2317-22.

McGettrick AJ, Knott V, Willis A, Handford PA. Molecular effects of calcium binding mutations in Marfan syndrome depend on domain context. *Hum Mol Genet*. 2000 Aug 12;9(13):1987-94.

Rashid MH, Rumbaugh K, Free in PMC, Passador L, Davies DG, Hamood AN, Iglewski BH, Kornberg A. Polyphosphate kinase is essential for biofilm development, quorum sensing, and virulence of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A*. 2000 Aug 15;97(17):9636-41.

Jormsjo S, Ye S, Moritz J, Walter DH, Dimmeler S, Zeiher AM, Henney A, Hamsten A, Eriksson P. Allele-specific regulation of matrix metalloproteinase-12 gene activity is associated with coronary artery luminal dimensions in diabetic patients with manifest coronary artery disease. *Circ Res*. 2000 May 12;86(9):998-1003.

Talas U, Dunlop J, Khalaf S, Leigh IM, Kelsell DP. Human elastase 1: evidence for expression in the skin and the identification of a frequent frameshift polymorphism. *J Invest Dermatol*. 2000 Jan;114(1):165-70.

Horwitz M, Benson KF, Person RE, Aprikyan AG, Dale DC. Mutations in ELA2, encoding neutrophil elastase, define a 21-day biological clock in cyclic haematopoiesis. *Nat Genet*. 1999 Dec;23(4):433-6.

Griffin MD, Torres VE, Grande JP, Kumar R. Vascular expression of polycystin. *J Am Soc Nephrol*. 1997 Apr;8(4):616-26.

There are references to a relation between mutations in neutrophil elastase and neutropenia—a low level of a type of white blood cells called neutrophils. To pursue this, we can look for elastase in the database of human genetic disease:

Online Mendelian Inheritance in Man (OMIM™)

OMIM is a database of human genes and genetic disorders. It was originally compiled by V. A. McKusick, M. Smith and colleagues and published on paper. The National Center for Biotechnology Information (NCBI) of the US National Library of Medicine has developed it into a database accessible from the Web, and introduced links to other archives of related information, including sequence databanks and the medical literature. OMIM is now well integrated with the NCBI information retrieval system ENTREZ. A related database, the OMIM Morbid Map, treats genetic diseases and their chromosomal locations.

The response to ELASTASE in a search of OMIM describes the results linking mutations in the gene to cyclic neutropenia.

The collection of results on elastase that we have assembled would support research on the system; for instance, we could map elastase mutants onto the structure of the molecule to see whether we could derive clues to the cause of cyclic neutropenia.

The Sequence Retrieval System (SRS)

SRS, originally developed by T. Etzold, is an integrated system for information retrieval from many different sequence databases, and for feeding the sequences retrieved into analytic tools such as sequence comparison and alignment programs.

SRS can search a total of 141 databases of protein and nucleotide sequences, metabolic pathways, 3D structures and functions, genomes, and disease and phenotype information (see Box). These include many small databases such as the Prosite and Blocks databases of protein structural motifs, transcription factor databases, and databases specialized to certain pathogens.

Some categories of databases searchable from SRS

Nucleotide Sequence	Literature
Uniprot	Mapping
Protein Function	Protein Structure
Enzymes	Metabolic Pathways
Mutation, SNP	Gene Ontology

In addition to the number and variety of databases to which it offers access, SRS offers tight links among the databases, and fluency in launching applications. A search in a single database component can be extended to a search in the complete network; that is, entries in all databases pertaining to a given protein can be found easily. Similarity searches and alignments can be launched directly, without saving the responses in an intermediate file.

In an SRS session, you begin by selecting one or more of the databases in which to search. The databases are grouped by category: nucleotide sequence-related, protein-related, etc. Then you can enter a set of query terms. As with ENTREZ, you may search for them either in all fields, or assign terms to categories. The program will respond with a set of entries containing your terms. As follow-on queries one might:

1. Examine one of the sequences identified by linking to the file retrieved.
2. Select one or more of the sequences identified and search other databases for related entries.
3. Launch an application, such as a secondary structure prediction or a multiple sequence alignment.

Other options on the search results page allow you to create and download reports on the selected matches. This might be simply a listing of the sequences, or the result of a more complex analysis of the results. Applying the multiple sequence alignment program CLUSTAL-W to the results produces an alignment such as appears in Plate III.

The Protein Identification Resource (PIR)

The PIR is an effective combination of a carefully curated database, information retrieval access software, and a workbench for investigations of sequences. The PIR also produces the Integrated Environment for Sequence Analysis (IESA). Think of this as an analysis package sitting on top of a retrieval system. Its functionality includes browsing, searching and similarity analysis, and links to other databases. Users may:

- ◆ Browse by annotations.
- ◆ Search selected text fields for different annotations, such as Superfamily, Family, Title, Species, Taxonomy group, Keywords and Domains.
- ◆ Analyse sequences using BLAST or FASTA Searches, Pattern Match, Multiple alignment.
- ◆ Global and Domain Search, and Annotation-sorted Search.
- ◆ View Statistics for Superfamily, Family, Title, Species, Taxonomy group, Keywords, Domains, Features.
- ◆ View Links to other databases, including PDB, COG, KEGG, WIT, and BRENDA.
- ◆ Select Specialized Sequence Groups such as Human, Mouse, Yeast and *E. coli* genomes.

The URLs for search of PIR by Text terms are:

In the US: <http://www-nbrf.georgetown.edu/pirwww/search/textpsd.html>

In Europe: <http://www.mips.gsf.de>

One feature of the PIR International system is the search for a specific peptide. Looking at the alignment of mammalian elastases in Plate III, we note at positions 220–228 a conserved motif: most of the sequences contain CNGDSGGPLN.

In the PIR, we can select PATTERN/PEPTIDE MATCH and search for exact matches for the subsequence CNGDSGGPLN giving 63 results.

Returning to the alignment table (Plate III), variations in the pattern appear in some molecules. The more general search for C[RNQF]GDSG[GS]PL[HNV], in which [XYZ] means a position containing either X or Y or Z, would pull out all the mammalian elastases in the alignment, plus a total of 82 sequences in all. Even these are not all the elastase homologues in the databank, as one could find by running a PSI-BLAST search for any of the sequences, or, remaining strictly within PIR, by looking up elastase in the PROT-FAM database. The pattern matches 20 families, all serine proteinases.

We are well on the way to generating a complete list of homologues.

ExpASy—Expert Protein Analysis System

ExpASy is the information retrieval and analysis system of the Swiss Institute of Bioinformatics, which (in collaboration with the European Institute of Bioinformatics) also produces the protein sequence databases SWISS-PROT and TrEMBL. TrEMBL contains translations of nucleotide sequences from the EMBL Data Library not yet fully integrated into SWISS-PROT.

Opening the main web page of ExpASy (<http://www.expasy.org>) and selecting SWISS-PROT and TrEMBL gives access to a set of information retrieval tools, including a link to SRS. There is also the option of searching SWISS-PROT directly. If we select FULL TEXT SEARCH and probe SWISS-PROT with the single term ELASTASE, we find ELNE_HUMAN, the real goal of our search, and around 150 other hits, including many inhibitors. One elastase homologue found is from the blood fluke: CERC_SCHMA. Both sequences are precursors; in the following alignment of these two sequences, upper case letters indicate the mature enzyme:

```

CERC_SCHMA  --msnrwrfvvvvtlftycltfervstwlIRSGEPVQHPAEFFPIAFLTTER-TMCTGSL  57
ELNE_HUMAN  mtlgrrlaclflacvlpalllggtalaseIVGGR-RARPHAWPFMVSLLQRRGGHFCGATL  59
           :...*   :... :.  *   :  : * . * .   : * : * * . *   : * : *
CERC_SCHMA  VSTRAVLTAGHCVCSPLPVIRVSFLTLRNGDQQGIHHQPSGVKVA PGYMPCMSARQRRP  117
ELNE_HUMAN  IAPNFVMSAAHCVAN—VNVRAVRVVLGAHNLSRREP—TRQVFAVQRIFENGYDP  111
           :... * : * . * * * . . : * : : * : : : * . . : . . *
CERC_SCHMA  IAQTLSGFDIAIVMLAQMVNLQSGIRVISLPPQPSDIPPPGTGVFIVGYGRDDNDRDPSRK  177
ELNE_HUMAN  VNLLN--DIVILQLNGSATINANVQVAQLPAQGRRLGNGVQCLAMGWLLGRNRG—  164
           :      * * . * : * . : : : * . * * . * . : : * * . . : *
CERC_SCHMA  NGGILKKGRATIMECRHATNGNPICVKAGQNFQQLPAPGDSGGPLLPV-LQGPVLGVVSH  236
ELNE_HUMAN  IASVLQELNVTVVTS-LCRRSNVCTLVRGRQAG-VCFGDSGSPVLCNGLIHGIA SFVRG  221
           : : * * : : . . . . * : * : * . * * * . * * . * : . . *
CERC_SCHMA  GVTLPNLPDIIVEYASVARMMLDFVRSNI-----  264
ELNE_HUMAN  GCASGLYPDAFAPVAQFVNWIDSTIIQRSEDNPCPHPRDPDPASRTH  267
           * : * * : . * . . . : * : . .

```

The structure of human neutrophil elastase is known from X-ray crystallography, but that of the blood fluke elastase is not.

One of the unique facilities of the ExpASy server is the link to SWISS-MODEL, an automatic web server for building homology models. Opening SWISS-MODEL and choosing FIRST APPROACH MODE (the simplest), we can simply enter the

SWISS-PROT code CERC_SCHMA, and launch the application. Model building is not a trivial operation, so the job is done off-line and the results sent by e-mail.

We shall discuss SWISS-MODEL further in Chapter 5.

Ensembl

Ensembl (<http://www.ensembl.org>) is intended to be the universal information source for the human genome. The goals are to collect and annotate all available information about human DNA sequences, link it to the master genome sequence, and make it accessible to the many scientists who will approach the data with many different points of view and different requirements. To this end, in addition to collecting and organizing the information, very serious effort has gone into developing computational infrastructure. Suitable conventions of nomenclature are established: it is not trivial to devise a scheme for maintaining stable identifiers in the face of data that will be undergoing not only growth but revision. The most visible result of these efforts is the web site, very rich in facilities both for browsing and for focussing in on details.

Ensembl is a joint project of the European Bioinformatics Institute and The Sanger Centre; participants include E. Birney, M. Clamp, T. Cox and T. J. P. Hubbard. However, Ensembl is organized as an open project, encouraging outside contributions. All but the most naive of readers must recognize the great demands this will place on quality control procedures.

Data collected in Ensembl includes genes, SNPs, repeats, and homologies. Genes may either be known experimentally, or deduced from the sequence. Because the experimental support for annotation of the human genome is so variable, Ensembl presents the supporting evidence for identification of every gene. Very extensive linking to other databases containing related information, such as Online Mendelian Inheritance in Man (OMIM), or expression databases, extend the accessible information.

Ensembl is structured around the human genome sequence. Users may identify regions via several types of lookups or searches:

- ◆ BLAST searches on a sequence or fragment
- ◆ Browsing—starting at the chromosome level then zooming in
- ◆ Gene name
- ◆ Relation to diseases, via OMIM
- ◆ ENSEMBL ID if the user knows it
- ◆ General text search

A text search in Ensembl for BRCA1 produced the page displayed, showing the region around the BRCA1 locus. The upper frame shows a megabase, mapped to the q21.2 and q21.31 bands of chromosome 17. It reports markers, and assigned genes. The bottom frame shows a more detailed view. Note the control panels between the two frames that permit navigation and 'zooming'. The bottom frame shows a 0.1 megabase region, reporting many more details, including the detailed structure of the BRCA1 gene, and the SNPs observed.

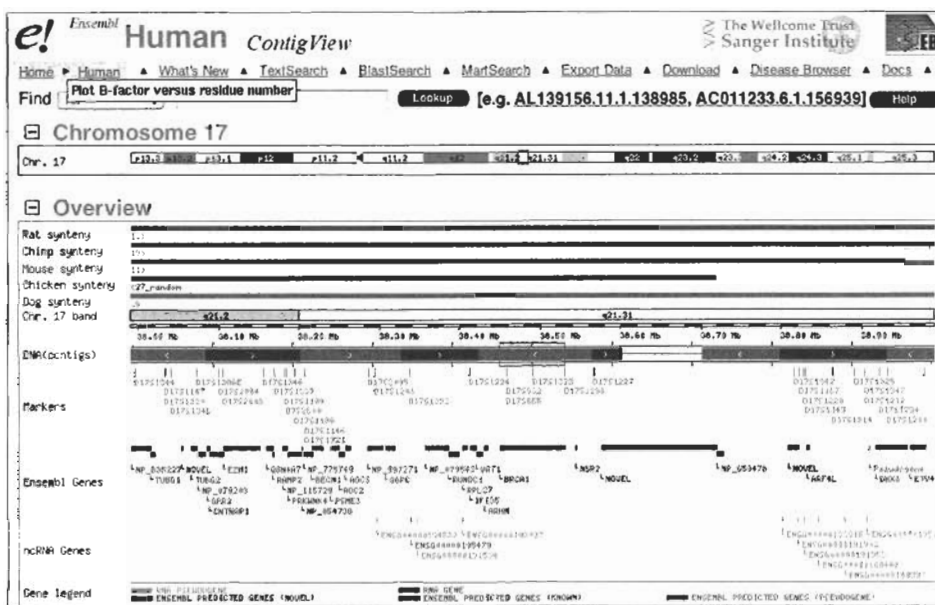


Fig. 3.4 A portion of the ENSEMBL page showing the region of the human genome surrounding BRAC1.

Where do we go from here?

We have visited only a few of the many databanks in molecular biology accessible on the Web. In the short term, readers will explore these sites and others, and become familiar with not only with the contents of the Web but its dynamics—the appearance and disappearance of sites and links. There are various biological metaphors for the Web—as an ecosystem that is evolving, or that is growing polluted by dead sites and links to dead sites. Unfortunately, there is no effective mechanism for decay and recycling as in the organic world!

Databanks are developing more effective avenues of intercommunication, to the point where ever more intimate links shade into apparent coalescence. The time is not far off when there will be one molecular biology databank, with many avenues of access. Scientists will be able to configure their own access to selected slices of the information, creating 'virtual databases' tailored to their own needs.

Recommended Reading

Each year the January issues of the journal *Nucleic Acid Research* contains a set of articles on databases in molecular biology. This is an invaluable reference.

Bishop, M. J., *Genetics databases* (London: Academic, 1999). [A compendium of databases, access and analysis.]

Zdobnov, E. M., Lopez, R., Apweiler, R. & Etzold, T. (2002), The EBI SRS server—new features. *Bioinformatics*, 18, 1149–1150.

Exercises, Problems, and Weblems

Exercises

3.1 A database of vehicles has entries for the following: bicycle, tricycle, motorcycle, car. It stores only the following information about each entry: (1) how many wheels (a number), and (2) source of propulsion = human or engine. For every possible pair of vehicles, devise a logical combination of query terms referring to either the exact value or the range in the number of wheels, and to the source of propulsion, that will return the two selected vehicles and no others.

3.2 The Box on page 144 showed the NCBI protein entry for human elastase 1 precursor. On a photocopy of this page, indicate which items are (a) purely database housekeeping, (b) peripheral data such as literature references, (c) the results of experimental measurements, (d) information inferred from experimental measurements.

3.3 Write a PERL script to extract the amino acid sequence from an entry in the ENTREZ protein sequence database as shown in the Box, page 144, and convert it to FASTA format.

3.4 Compare the file retrieved by a search in NCBI for human elastase under Protein (page 144) and Nucleotide (page 145). On photocopies of these two pages, mark with a highlighter all items that the two files have in common.

Weblems

3.1 Retrieve the complete SWISS-PROT entry for bovine pancreatic trypsin inhibitor (*not* pancreatic secretory trypsin inhibitor) and the complete PIR entry for this protein. What information does each have that the other does not?

3.2 Find a list of official and unofficial mirror sites of the Protein Data Bank. Which is closest to you?

3.3 Find all structures of sperm whale myoglobin in the Protein Data Bank and draw a histogram of their dates of deposition.

3.4 Find protein structures determined by Peter Hudson, alone or with colleagues.

3.5 Design a search string for use with the Protein Data Bank tool SearchLite that would return *E. coli* thioredoxin structures but *not* Staphylococcal nuclease structures.

3.6 For what fraction of structures determined by X-ray crystallography deposited in the Protein Data Bank have structure factor files also been deposited?

3.7 Protein Data Bank entry 8XIA contains the structure of one monomer of D-Xylose isomerase from *Streptomyces rubiginosus*. What is the probable quaternary structure? How was the geometry of the assembly corresponding to the probable quaternary structure derived from the coordinates in the entry?

3.8 Find structural neighbours of Protein Data Bank entry 2TRX (*E. coli* thioredoxin), according to SCOP, CATH, FSSP, and CE. Which, if any, structures do *all* these classifications consider structural neighbours of 2TRX? Which structures are considered structural neighbours in some but not all classifications?

3.9 Why did an ENTREZ search in the protein category for HUMAN ELASTASE return a tRNA synthetase?

3.10 The Box on page 154 contains the amino acid sequence of human elastase 1 precursor. What sequence differences are there between this and the mature protein?

3.11 What is the relation between the elastase sequences recovered from searching the NCBI and the PIR?

3.12 Using SWISS-PROT directly, or SRS, recover the SWISS-PROT entry for human elastase. What information does this file contain that does not appear in (a) the corresponding entry in ENTREZ (protein) and (b) the corresponding entry in PIR?

3.13 What homologues of human neutrophil elastase can be identified by PSI-BLAST?

3.14 Search for structures of elastases using the Protein Data Bank search facilities. Compare the results with those from ENTREZ, described in the text.

3.15 Which gene in *C. elegans* encodes a protein similar in sequence to human elastase?

3.16 What is the chromosomal location of the human gene for glucose-6-phosphate dehydrogenase?

3.17 Pseudogenes in eukaryotes can be classified into those that arose by gene duplication and divergence, and those reinserted into the genome from mRNA by a retrovirus, called *processed* pseudogenes. Processed pseudogenes can be identified by the absence of introns. Which if any of the pseudogenes in the human globin gene clusters are processed pseudogenes?

3.18 Preliminary genetic analysis on the way to isolating the gene associated with cystic fibrosis bracketed it between the MET oncogene and RFLP D7S8. It was then estimated that this region contained 1–2 million bp, and might contain 100–200 genes. (a) How many base pairs long did this region actually turn out to be? (b) How many expressed genes is this region now believed to contain?

3.19 The gene for Berardinelli-Seip syndrome was initially localized between two markers on chromosome band 11q13—D11S4191 and D11S987. How many base pairs are there in the interval between these two markers?

3.20 Is there a database available on the Web that specifically collects structural and thermodynamic information on protein-nucleic acid interactions?

3.21 The Yeast proteome database contains an entry for *cdc6*, the protein that regulates initiation of DNA replication. (a) On what chromosome is the gene for

yeast cdc6? (b) What post-translational modification does this protein undergo to reach its mature active state? (c) What are the closest known relatives of this protein in other species? (d) With what other proteins is yeast cdc6 known to interact? (e) What is the effect of distamycin A on the activity of yeast cdc6? (f) What is the effect of actinomycin D on the the activity of yeast cdc6?