BIOINFORMATICS AND FUNCTIONAL GENOMICS

Second Edition

Jonathan Pevsner

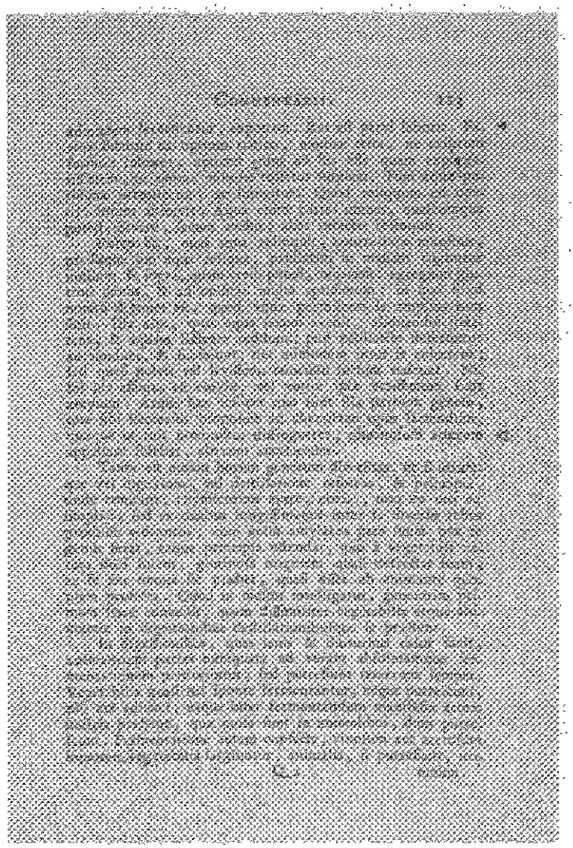
Department of Neurology, Kennedy Krieger Institute

and

Department of Neuroscience and Division of Health Sciences Informatics, The Johns Hopkins School of Medicine, Baltimore, Maryland

WILEY-BLACKWELL

A JOHN WILEY & SONS, INC., PUBLICATION



Chapter 2 introduces ways to access molecular data, including information about DNA and proteins. One of the first scientists to study proteins was Iacopo Bartolomeo Beccari (1682–1776), an Italian philosopher and physician who discovered protein as a component of vegetables. This image is from page 123 of the Bologna Commentaries, published in 1745 and written by a secretary on the basis of a 1728 lecture by Beccari. Beccari separated gluten (plant proteins) from wheaten flour. The passage beginning Res est parvi laboris *("it is a thing of little labor"; see solid arrowhead) is translated as follows (Beach, 1961, p. 362):*

"It is a thing of little labor. Flour is taken of the best wheat, moderately ground, the bran not passing though the sieve, for it is necessary that this be fully purged away, so that all traces of a mixture have been removed. Then it is mixed with pure water and kneaded. What is left by this procedure, washing clarifies. Water carries off with itself all it is able to dissolve, the rest remains untouched. After this, what the water leaves is worked with the hands, and pressed upon in the water that has stayed. Slowly it is drawn together in a doughy mass, and beyond what is possible to be believed, tenacious, a remarkable sort of glue, and suited to many uses; and what is especially worthy of note, it cannot any longer be mixed with water. The other particles, which water carries away with itself, for some time float and render the water milky; but after a while they are carried to the bottom and sink; nor in any way do they adhere to each other; but like powder they return upward on the lightest contact. Nothing is more like this than starch, or rather this truly is starch. And these are manifestly the two sorts of bodies which Beccari displayed through having done the work of a chemist and he distinguished them by their names, one being appropriately called glutinous (see open arrowhead) and the other amylaceous."

In addition to purifying gluten, Beccari identified it as an "animal substance" in contrast to starch, a "vegetable substance," based on differences on how they decomposed with heat or distillation. A century later Jons Jakob Berzelius proposed the word protein, and he also posited that plants form "animal materials" that are eaten by herbivorous animals.

2

Access to Sequence Data and Literature Information

Introduction to Biological Databases

All living organisms are characterized by the capacity to reproduce and evolve. The genome of an organism is defined as the collection of DNA within that organism, including the set of genes that encode proteins. In 1995 the complete genome of a free-living organism was sequenced for the first time, the bacterium *Haemophilus influenzae* (Fleischmann et al., 1995; Chapters 13 and 15). In the few years since then the genomes of thousands of organisms have been completely sequenced, ushering in a new era of biological data acquisition and information accessibility. Publicly available databanks now contain billions of nucleotides of DNA sequence data collected from over 260,000 different organisms (Kulikova et al., 2007). The goal of this chapter is to introduce the databases that store these data and strategies to extract information from them.

Three publicly accessible databases store large amounts of nucleotide and protein sequence data: GenBank at the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH) in Bethesda (Benson et al., 2009), the DNA Database of Japan (DDBJ) at the National Institute of

Bioinformatics and Functional Genomics, Second Edition. By Jonathan Pevsner

to a vast new influx of DNA sequence data (Fig. 2.1*b*). Next-generation sequencing involves the generation of massive amounts of sequence data, such as 1 billion bases (1 Gb) in a single experiment that is completed in a matter of days. In a single issue of the journal *Nature* in November 2008 Bentley et al. described the sequencing of an individual of Nigerian ancestry, Wang et al. reported the DNA sequence of an Asian individual, and Ley et al. analyzed the genome sequence of a tumor sample. Together, these three papers involved the generation and analysis of 492 gigabases (Gb) of DNA sequence. By the end of 2008 the 1000 Genomes Project generated several terabases of data. For major sequencing centers (such as those at the Wellcome Trust Sanger Institute, Beijing Genomics Institute Shenzhen, the Broad Institute of MIT and Harvard, Washington University School of Medicine's Genome Sequencing Center, and Baylor College of Medicine's Human Genome Sequencing Center) it is estimated that each will generate approximately 10 terabases in the year 2009. According to a Wellcome Trust Sanger Institute press release in 2008, that center now produces as much sequence data every 2 minutes as was generated in the first five years at GenBank. Thus the amount of DNA sequence generated by next-generation sequencing technologies has already dwarfed the amount of sequence in GenBank. Such data are available through the Trace Archive at NCBI and the Ensembl Trace Server at EBI, including the Short Read Archive that was initiated in 2007.

Organisms in GenBank

You can download all of the sequence data in GenBank at the website ▶ ftp://ftp.ncbi.nih.gov/genbank. For release 158.0 in February 2007, the total size of these files is about 250 gigabytes (250×10^9 bytes). By comparison, all the words in the United States Library of Congress add up to 20 terabytes (20×10^{12} bytes; 20 trillion bytes). And the particle accelerator used by physicists at CERN near Geneva (▶ http://public.web.cern.ch/Public/) collects petabytes of data each year (10^{15} bytes; 1 quadrillion bytes).

Over 260,000 different species are represented in GenBank, with over 1000 new species added per month (Benson et al., 2009). The number of organisms represented in GenBank is shown in Table 2.1. We will define the bacteria, archaea, and eukaryotes in detail in Chapters 13 to 18. Briefly, eukaryotes have a nucleus and are often multicellular, whereas bacteria do not have a nucleus. Archaea are single-celled organisms, distinct from eukaryotes and bacteria, which constitute a third major branch of life. Viruses, which contain nucleic acids (DNA or RNA) but can only replicate in a host cell, exist at the borderline of the definition of living organisms.

We have seen so far that GenBank is very large and growing rapidly. From Table 2.1 we see that the organisms in GenBank consist mostly of eukaryotes. Of the microbes, there are currently over 25 times more bacteria than archaea represented in GenBank.

TABLE 2.1 Taxa Represented in GenBank

Ranks:	Higher Taxa	Genus	Species	Lower Taxa	Total
Archaea	89	106	502	105	802
Bacteria	996	1,857	13,973	4,973	21,799
Eukaryota	15,205	45,066	167,764	13,200	241,235
Fungi	1,096	3,307	18,699	1,058	24,160
Metazoa	11,113	27,222	73,062	6,643	118,040
Viridiplantae	1,849	12,557	69,729	4,869	89,004
Viruses	445	294	5,054	33,909	39,702
All taxa	16,756	47,331	191,956	52,217	308,260

Source: From ▶ http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi (November 2008)

TABLE 2.2 Twenty Most Sequenced Organisms in GenBank

Entries	Bases	Species	Common Name
11,550,460	13,148,670,755	*Homo sapiens*	Human
7,255,650	8,361,230,436	*Mus musculus*	Mouse
1,757,685	6,060,823,765	*Rattus norvegicus*	Rat
2,086,880	5,235,078,866	*Bos taurus*	Cow
3,181,318	4,600,009,751	*Zea mays*	Corn
2,489,204	3,551,438,061	*Sus scrofa*	Pig
1,591,342	2,978,804,803	*Danio rerio*	Zebrafish
1,205,529	1,533,859,717	*Oryza sativa*	Rice
228,091	1,352,737,662	*Strongylocentrotus purpuratus*	Purple sea urchin
1,673,038	1,142,531,302	*Nicotiana tabacum*	Tobacco
1,413,112	1,088,892,839	*Xenopus (Silurana)*	Western clawed frog
212,967	996,533,885	*Pan troglodytes*	Chimpanzee
780,860	913,586,921	*Drosophila melanogaster*	Fruit fly
2,211,104	912,500,625	*Arabidopsis thaliana*	Thale cress
650,374	905,797,007	*Vitis vinifera*	Wine grape
804,246	871,336,795	*Gallus gallus*	Chicken
77,069	803,847,320	*Macaca mulatta*	Rhesus macaque
1,215,319	748,031,972	*Ciona intestinalis*	Sea squirt
1,224,224	744,373,069	*Canis lupus*	Dog
1,725,913	680,988,452	*Glycine max*	Soybean

Source: From ▶ ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt (GenBank release 168.0, October 2008).

The number of entries and bases of DNA/RNA for the 20 most sequenced organisms in GenBank is provided in Table 2.2 (excluding chloroplast and mitochondrial sequences). This list includes some of the most common model organisms that are studied in biology. Notably, the scientific community is studying a series of mammals (e.g., human, mouse, cow), other vertebrates (chicken, frog), and plants (corn, rice, bread wheat, wine grape). Different species are useful for a variety of different studies. Bacteria, archaea, and viruses are absent from the list in Table 2.2 because they have relatively small genomes.

We will discuss how genomes of various organisms are selected for complete sequencing in Chapter 13.

To help organize the available information, each sequence name in a GenBank record is followed by its data file division and primary accession number. (Accession numbers are defined below.) The following codes are used to designate the data file divisions:

1. PRI: primate sequences
2. ROD: rodent sequences
3. MAM: other mammalian sequences
4. VRT: other vertebrate sequences
5. INV: invertebrate sequences
6. PLN: plant, fungal, and algal sequences
7. BCT: bacterial sequences
8. VRL: viral sequences
9. PHG: bacteriophage sequences

The International Human Genome Sequencing Consortium adopted the Bermuda Principles in 1996, calling for the rapid release of raw genomic sequence data. You can read about recent versions of these principles at ▶ http://www.genome.gov/10506376.

The terms STS, GSS, EST, and HTGS are defined bellow.

10. SYN: synthetic sequences
11. UNA: unannotated sequences
12. EST: EST sequences (expressed sequence tags)
13. PAT: patent sequences
14. STS: STS sequences (sequence-tagged sites)
15. GSS: GSS sequences (genome survey sequences)
16. HTG: HTGS sequences (high throughput genomic sequences)
17. HTC: HTC sequences (high throughput cDNA sequences)
18. ENV: environmental sampling sequences

Beta globin is sometimes called hemoglobin-beta. In general, a gene does not always have the same name as the corresponding protein. Indeed there is no such thing as a "hemoglobin gene" because globin genes encode globin proteins, and the combination of these globins with heme forms the various types of hemoglobin. Often, multiple investigators study the same gene or protein and assign different names. The human genome organization (HUGO) Gene Nomenclature Committee (HGNC) has the critical task of assigning official names to genes and proteins. See ➧ http://www.gene.ucl.ac.uk/nomenclature/.

Types of Data in GenBank

There is an enormous number of molecular sequences in GenBank. We will next look at some of the basic kinds of data present in GenBank. Afterward, we will address strategies to extract the data you want from GenBank.

We start with an example. We want to find out the sequence of human beta globin. A fundamental distinction is that DNA, RNA-based, and protein sequences are stored in discrete databases. Furthermore, within each database, sequence data are represented in a variety of forms. For example, beta globin may be described at the DNA level (e.g. as a gene), at the RNA level (as a messenger RNA [mRNA] transcript), and at the protein level (see Fig. 2.2). Because RNA is relatively unstable, it is typically converted to complementary DNA (cDNA), and a variety

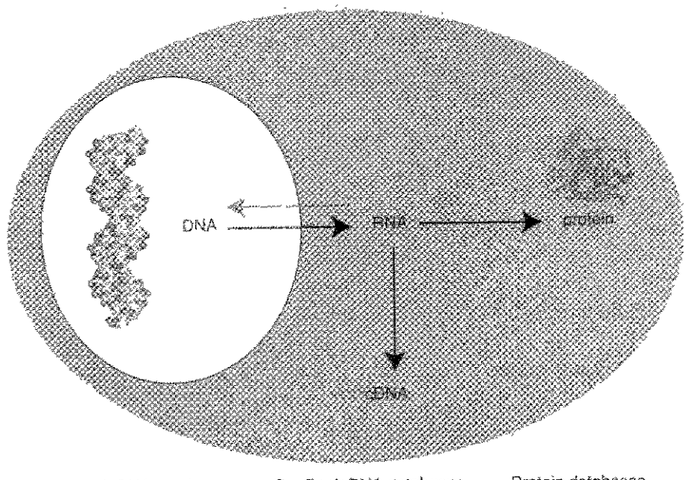


FIGURE 2.2. Types of sequence data in GenBank and other databases using human beta globin as an example. Note that "globin" may refer to a gene or other DNA feature, an RNA transcript (or its corresponding complementary DNA), or a protein. There are specialized databases corresponding to each of these three levels. See text for abbreviations. There are many other databases (not listed) that are not part of GenBank and NCBI; note that SwissProt, PDB, and PIR are protein databases that are independent of GenBank. The raw nucleotide sequence data in GenBank, DDBJ, and EBI are equivalent.

of databases contain cDNA sequences corresponding to RNA transcripts. Thus for our example of beta globin, the various forms of sequence data include the following.

Genomic DNA Databases

- Beta globin is part of a chromosome. In the case of human RBP we will see that its gene is situated on chromosome 11 (Chapter 16, on the eukaryotic chromosome).
- Beta globin may be a part of a large fragment of DNA such as a cosmid, bacterial artificial chromosome (BAC), or yeast artificial chromosome (YAC) that may contain several genes. A BAC is a large segment of DNA (typically about 200,000 base pairs [bp], or 200 kilobases [kb]) that is cloned into bacteria. Similarly, YACs are used to clone large amounts of DNA into yeast. BACs and YACs are useful vectors with which to sequence large portions of genomes.
- Beta globin is present in databases as a gene. The gene is the functional unit of heredity (further defined in Chapter 16), and it is a DNA sequence that typically consists of regulatory regions, protein-coding exons, and introns. Often, human genes are 10 to 100 kb in size.
- Beta globin is present as a sequence-tagged site (STS)—that is, as a small fragment of DNA (typically 500 bp long) that is used to link genetic and physical maps and which is part of a database of sequence-tagged sites (dbSTS).

Human chromosome 11, which is a mid-sized chromosome, contains about 1800 genes and is about 134,000 kilobases (kb) in length.

cDNA Databases Corresponding to Expressed Genes

Beta globin is represented in databases as an expressed sequence tag (EST), that is, a cDNA sequence derived from a particular cDNA library. If one obtains a tissue such as liver, purifies RNA, then converts the RNA to the more stable form of cDNA, some of the cDNA clones contained in that cDNA are likely to encode beta globin.

In GenBank, the convention is to use the four DNA nucleotides when referring to DNA derived from RNA.

Expressed Sequence Tags (ESTs)

The database of expressed sequence tags (dbEST) is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences from a number of organisms (Boguski et al., 1993). An EST is a partial DNA sequence of a cDNA clone. All cDNA clones, and thus all ESTs, are derived from some specific RNA source such as human brain or rat liver. The RNA is converted into a more stable form, cDNA, which may then be packaged into a cDNA library (refer to Fig. 2.2). ESTs are typically randomly selected cDNA clones that are sequenced on one strand (and thus may have a relatively high sequencing error rate). ESTs are often 300 to 800 bp in length. The earliest efforts to sequence ESTs resulted in the identification of many hundreds of genes that were novel at the time (Adams et al., 1991).

In November, 2008 GenBank had over 58,000,000 ESTs. We discuss ESTs further in Chapter 8.

Currently, GenBank divides ESTs into three major categories: human, mouse, and other. Table 2.3 shows the 10 organisms from which the greatest number of ESTs has been sequenced. Assuming that there are 22,000 human genes (see Chapter 19) and given that there are about 8.1 million human ESTs, there is currently an average of over 300 ESTs corresponding to each human gene.

TABLE 2-3 Top Ten Organisms for Which ESTs Have Been Sequenced

Organisms	Common Name	Number of ESTs
Homo sapiens	Human	8,138,094
Mus musculus – domesticus	Mouse	4,850,602
Zea mays	Maize	2,002,585
Arabidopsis thaliana	Thale cress	1,526,133
Bos taurus	Cattle	1,517,139
Sus scrofa	Pig	1,476,546
Danio rerio	Zebrafish	1,379,829
Glycine max	Soybean	1,351,356
Xenopus (Silurana) tropicalis	Western clawed frog	1,271,375
Oryza sativa	Rice	1,220,908

Many thousand of cDNA libraries have been generated from a variety of organism, and the total number of public entries is currently over 58 million.
Source: ▶ http://www.nebi.nlm.nin.gov/dbEST/dbEST_summary.html (dbEST release 022307, November 2008).

ESTs and UniGene

To find the entry for beta globin, go to ▶ http://www.ncbi.nlm.nih.gov, select All Databases then click UniGene, select human, then enter beta globin or HBB. The UniGene accession number is Hs.523443; note that Hs refers to *Homo sapiens*. To see the DNA sequence of a typical EST, click on an EST accession number from the UniGene page (e.g., AA970968.1), then follow the link to the GenBank entry in Entrez Nucleotide.

We are using beta globin as a specific example. If you want to type "globin" as a query, you will simply get more results from any database—in UniGene, you will find over 100 entries corresponding to a variety of globin genes in various species.

The UniGene project has become extremely important in the effort to identify protein-coding genes in newly sequenced genomes. We discuss this in Chapters 13 and 16.

The goal of the UniGene (unique gene) project is to create gene-oriented clusters by automatically partitioning ESTs into nonredundant sets. Ultimately there should be one UniGene cluster assigned to each gene of an organism. There may be as few as one EST in a cluster, reflecting a gene that is rarely expressed, to tens of thousands of ESTs, associated with a highly expressed gene. We discuss UniGene clusters further in Chapter 8 (on gene expression). There are over 100 organisms currently represented in UniGene, 71 of which are listed in Table 2.4.

For human beta globin, there is only a single UniGene entry. This entry currently has 2400 human ESTs that match the beta globin gene. This large number of ESTs reflects how abundantly the beta globin gene has been expressed in cDNA libraries that have been sequenced. A UniGene cluster is a database entry for a gene containing a group of corresponding ESTs (Fig. 2.3).

There are now thought to be approximately 22,000 human genes (see Chapter 19). One might expect an equal number of UniGene clusters. However, in practice, there are more UniGene clusters than there are genes—currently, there are about 120,000 human UniGene clusters. This discrepancy could occur for three reasons. (1) Clusters of ESTs could correspond to distinct regions of one gene. In that case there would be two (or more) UniGene entries corresponding to a single gene (see Fig. 2.3). Two UniGene clusters may properly cluster into one, and the number of UniGene clusters may collapse over time. (2) In the past several years it has become appreciated that much of the genome is transcribed at low levels (see Chapter 8). Currently, 40,000 human UniGene clusters consist of a single EST, and over 76,000 UniGene clusters consist of just one to four ESTs. These could reflect authentic genes that have not yet been appreciated by other means of gene identification. Alternatively they may represent rare transcription events of unknown biological relevance. (3) Some DNA may be transcribed during the creation of a cDNA library without corresponding to an authentic transcript. Thus it is a cloning artifact. We discuss the criteria for defining a eukaryotic gene in Chapter 16. Alternative splicing (Chapter 8) may introduce apparently new clusters of genes because the spliced exon is not homologous to the rest of the sequence.

TABLE 2-4 Seventy-One Organisms Represented in UniGene

Group	No.	Species
Chordata: Mammalia	12	*Bos taurus* (cattle), *Canis familiaris* (dog), *Equus caballus* (horse), *Homo sapiens* (human), *Macaca fascicularis* (crab-eating macaque), *Macaca mulatta* (rhesus monkey), *Mus musculus* (mouse), *Oryctolagus cuniculus* (rabbit), *Ovis aries* (sheep), *Rattus norvegicus* (Norway rat), *Sus scrofa* (pig), *Trichosurus vulpecula* (silver-gray brushtail possum)
Chordata: Actinopterygii	8	*Danio rerio* (zebrafish), *Fundulus heteroclitus* (killifish), *Gasterosteus aculeatus* (three spined stickleback), *Oncorhynchus mykiss* (rainbow trout), *Oryzias latipes* (Japanese medaka), *Pimephales promelas* (fathead minnow), *Salmo salar* (Atlantic salmon), *Takifugu rubripes* (pufferfish)
Chordata: Amphibia	2	*Xenopus laevis* (African clawed frog), *Xenopus tropicalis* (western clawed frog)
Chordata: Ascidiacea	3	*Ciona intestinalis*, *Ciona savignyi*, *Molgula tectiformis*
Chordata: Aves	2	*Gallus gallus* (chicken), *Taeniopygia guttata* (zebra finch)
Chordata: Cephalochordata	1	*Branchiostoma floridae* (Florida lancelet)
Chordata: Hyperoartia	1	*Petromyzon marinus* (sea lamprey)
Echinodermata: Echinoidea	1	*Strongylocentrotus purpuratus* (purple sea urchin)
Arthopoda: Insecta	6	*Aedes aegypti* (yellow fever mosquito), *Anopheles gambiae* (African malaria mosquito), *Apis mellifera* (honey bee), *Bombyx mori* (domestic silkworm), *Drosophila melanogaster* (fruit fly), *Tribolium castaneum* (red flour beetle)
Nematoda: Chromadorea	1	*Caenorhabditis elegans* (nematode)
Platyhelminthes: Trematoda	2	*Schistosoma japonicum*, *Schistosoma mansoni*
Cnidaria: Hydrozoa	1	*Hydra magnipapillata*
Streptophyta: Bryopsida	1	*Physcomitrella patens*
Streptophyta: Coniferopsida	3	*Picea glauca* (white spruce), *Picea sitchensis* (Sitka spruce), *Pinus taeda* (loblolly pine)
Streptophyta: Eudicotyledons	18	*Aquilegia formosa* × *Aquilegia pubescens*, *Arabidopsis thaliana* (thale cress), *Brassica napus* (rape), *Citrus sinensis* (Valencia orange), *Glycine max* (soybean), *Gossypium hirsutum* (upland cotton), *Gossypium raimondii*, *Helianthus annuus* (sunflower), *Lactuca sativa* (garden lettuce), *Lotus japonicus*, *Malus* × *domestica* (apple), *Medicago truncatula* (barrel medic), *Nicotiana tabacum* (tobacco), *Populus tremula* × *Populus tremuloides*, *Populus trichocarpa* (western balsam poplar), *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), *Vitis vinifera* (wine grape)
Streptophyta: Liliopsida	6	*Hordeum vulgare* (barley), *Oryza sativa* (rice), *Saccharum officinarum* (sugarcane), *Sorghum bicolor* (sorghum), *Triticum aestivum* (wheat), *Zea mays* (maize)
Chlorophyta: Chlorophyceae	1	*Chlamydomonas reinhardtii*
Dictyosteliida: Dictyostelium	1	*Dictyostelium discoideum* (slime mold)
Apicomplexa: Coccidia	1	*Toxoplasma gondii*

Source: UniGene ▶ http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene (November 2008).

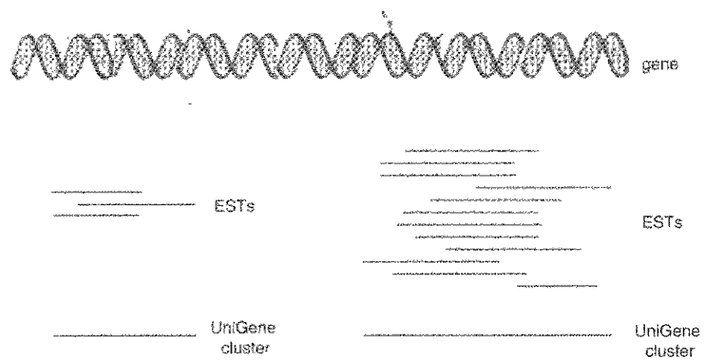


FIGURE 2.3. Schematic description of UniGene clusters. Expressed sequence tags (ESTs) are mapped to a particular gene and to each other. The number of ESTs that constitute a UniGene cluster ranges from 1 to tens of thousands; on average there are 300 human ESTs per cluster. Sometimes, as shown in the diagram, separate UniGene clusters correspond to distinct regions of a gene. Eventually, as genome sequencing increases our ability to define and annotate full-length genes, these two UniGene clusters would be collapsed into one single cluster. Ultimately, the number of UniGene clusters should equal the number of genes in the genome.

Sequence-Tagged Sites (STSs)

As of November 2008 there are 1.3 million STSs, derived from 300 organisms.

The dbSTS is an NCBI site containing STSs, which are short genomic landmark sequences for which both DNA sequence data and mapping data are available (Olson et al., 1989). STSs have been obtained from several hundred organisms, including primates and rodents (Table 2.5). A typical STS is approximately the size of an EST. Because they are sometimes polymorphic, containing short sequence repeats (Chapter 16), STSs can be useful for mapping studies.

Genome Survey Sequences (GSSs)

There are currently 24 million GSS entries from over 800 organisms (November 2008). The top four organisms (Table 2.6) account for about a third of all entries. This database is accessed via ▶ http://www.ncbi.nlm.nih.gov/projects/dbGSS/.

The GSS division of GenBank is similar to the EST division, except that its sequences are genomic in origin, rather than cDNA (mRNA). The GSS division contains the following types of data (see Chapters 13 and 16):

- Random "single-pass read" genome survey sequences
- Cosmid/BAC/YAC end sequences
- Exon-trapped genomic sequences
- The *Alu* polymerase chain reaction (PCR) sequences

TABLE 2.5 Organisms from Which STSs Have Been Obtained

Organism	Approximate Number of STSs
Homo sapiens	324,000
Pan troglodytes	161,000
Macaca mulatta	72,000
Mus musculus	56,000
Rattus norvegicus	50,000

These are the organisms with the most UniSTS entries.
Source: ▶ http://www.ncbi.nlm.nih.gov/genome/sts/unists_stats.html (November 2008).

TABLE 2.6 Selected Organisms from Which GSSs Have Been Obtained. For a discussion of Metagenomes see Chapter 13

Organism	Approximate Number of Sequences
Marine metagenome	2,643,000
Zea mays + subsp. *mays* (maize)	2,091,000
Mus musculus + *domesticus* (mouse)	1,864,000
Nicotiana tabacum (tobacco)	1,421,000
Homo sapiens (human)	1,214,000
Canis lupus familiaris (dog)	854,000

Source: ▶ http://www.ncbi.nlm.nih.gov/dbGSS/dbGSS_summary.html (November 2008).

All searches of the Entrez Nucleotide database provide results that are divided into three sections: GSS, ESTs, and "CoreNucleotide" (that is, the remaining nucleotide sequences). Recent holdings of the GSS database are listed in Table 2.6.

High Throughput Genomic Sequence (HTGS)

The HTGS home page is ▶ http://www.ncbi.nlm.nih.gov/HTGS/ and its sequences can be searched via BLAST (see Chapters 4 and 5).

The HTGS division was created to make "unfinished" genomic sequence data rapidly available to the scientific community. It was done in a coordinated effort between the three international nucleotide sequence databases: DDBJ, EMBL, and GenBank. The HTGS division contains unfinished DNA sequences generated by the high throughput sequencing centers.

Protein Databases

The name beta globin may refer to the DNA, the RNA, or the protein. As a protein, beta globin is present in databases such as the nonredundant (nr) database of GenBank (Benson et al., 2009), the SwissProt database (Boeckmann et al., 2003), UniProt (UniProt Consortium 2007), and the Protein Data Bank (Kouranov et al., 2006).

We have described some of the basic kinds of sequence data in GenBank. We will next turn our attention to Entrez and the other programs in NCBI and elsewhere, which allow you to access GenBank, EMBL, and DDBJ data and related literature information. In particular, we will introduce the NCBI website, one of the main web-based resources in the field of bioinformatics.

National Center for Biotechnology Information

Introduction to NCBI: Home Page

The NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information (Wheeler et al., 2007). The NCBI home page is shown in Fig. 2.4. Across the top bar of the website, there are seven categories: PubMed, Entrez, BLAST, OMIM, Books, Taxonomy, and Structure.

Extremely useful tutorials are available for Entrez, PubMed, and other NCBI resources at ▶ http://www.ncbi.nlm.nih.gov/Education/. You can also access this from the education link on the NCBI home page (▶ http://www.ncbi.nlm.nih.gov).

PubMed

PubMed is the search service from the National Library of Medicine (NLM) that provides access to over 18 million citations in MEDLINE (Medical Literature,

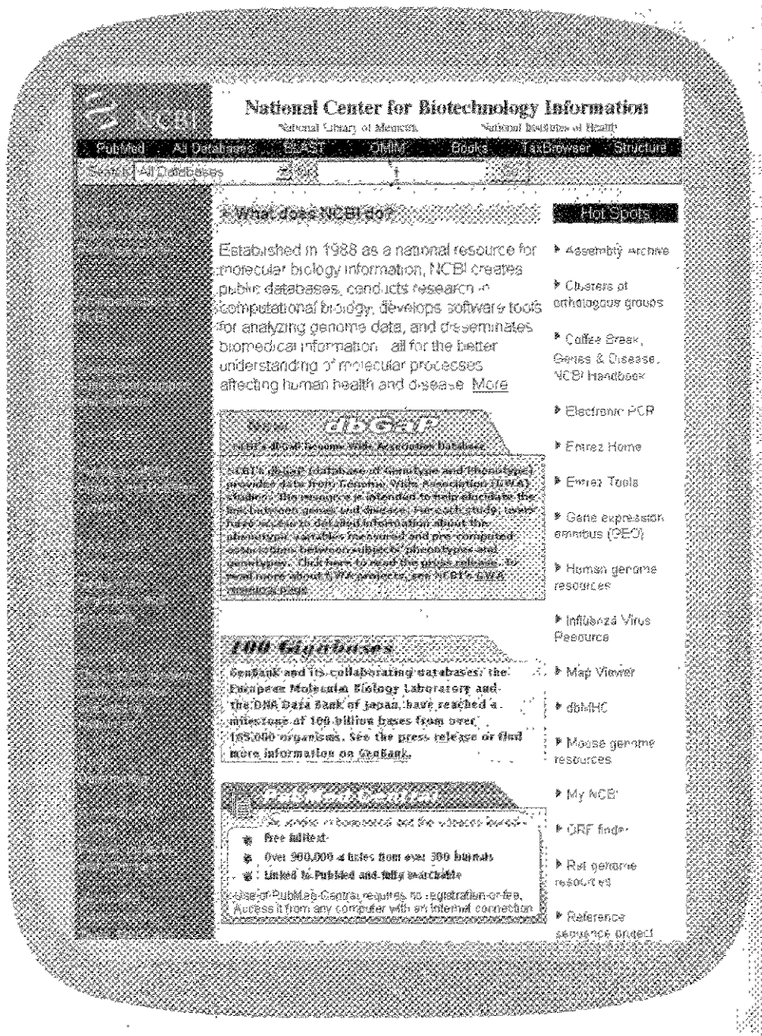


FIGURE 2.4. The main page of the National Center for Biotechnology Information (NCBI) website (➧ http://www.ncbi.nlm.nih.gov). Across the top bar, sections include PubMed, Entrez and Books (described in this chapter), BLAST (Chapters 3–5), Taxonomy (Chapters 13–19), Structure (Chapter 11), and Online Mendelian Inheritance in Man (OMIM, Chapter 20). Note that the left sidebar includes tutorials within the Education section.

Analysis, and Retrieval System Online) and other related databases, with links to participating online journals.

Entrez

Entrez integrates the scientific literature, DNA and protein sequence databases, three-dimensional protein structure data, population study data sets, and assemblies of complete genomes into a tightly coupled system. PubMed is the literature component of Entrez.

BLAST

BLAST (Basic Local Alignment Search Tool) is NCBI's sequence similarity search tool designed to support analysis of nucleotide and protein databases (Altschul et al., 1990, 1997). BLAST is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. We explore BLAST in Chapters 3 to 5.

OMIM

Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders. It was created by Victor McKusick and his colleagues and developed for the World Wide Web by NCBI (Hamosh et al., 2005). The database contains detailed reference information. It also contains links to PubMed articles and sequence information. We describe OMIM in Chapter 20 (on human disease).

Books

NCBI offers several dozen books online. These books are searchable, and are linked to PubMed.

Taxonomy

The NCBI taxonomy website includes a taxonomy browser for the major divisions of living organisms (archaea, bacteria, eukaryota, and viruses). The site features taxonomy information such as genetic codes and taxonomy resources and additional information such as molecular data on extinct organisms and recent changes to classification schemes. We will visit this site in Chapters 7 (on evolution) and 13 to 18 (on genomes and the tree of life).

Structure

The NCBI structure site maintains the Molecular Modelling Database (MMDB), a database of macromolecular three-dimensional structures, as well as tools for their visualization and comparative analysis. MMDB contains experimentally determined biopolymer structures obtained from the Protein Data Bank (PDB). Structure resources at NCBI include PDBeast (a taxonomy site within MMDB), Cn3D (a three-dimensional structure viewer), and a vector alignment search tool (VAST) which allows comparison of structures. (See Chapter 11, on protein structure.)

The Protein Data Bank (➧ http://www.rcsb.org/pdb/) is the single worldwide repository for the processing and distribution of biological macromolecular structure data. We explore the PDB in Chapter 11.

The European Bioinformatics Institute (EBI)

The EBI website is comparable to NCBI in its scope and mission, and it represents a complementary, independent resource. EBI features six core molecular databases (Brooksbank et al., 2003), as follows. (1) EMBL-Bank is the repository of DNA and RNA sequences that is complementary to GenBank and DDBJ (Kulikova et al., 2007). (2) SWISS-PROT and (3) TrEMBL are two protein databases that are described further below. (4) MSD is a protein structure database (see Chapter 11). (5) Ensembl is one of the three main genome browsers (described below). (6) ArrayExpress is one of the two main worldwide repositories for gene expression

You can access EBI at ➧EBI at http://www.ebi.ac.uk/.

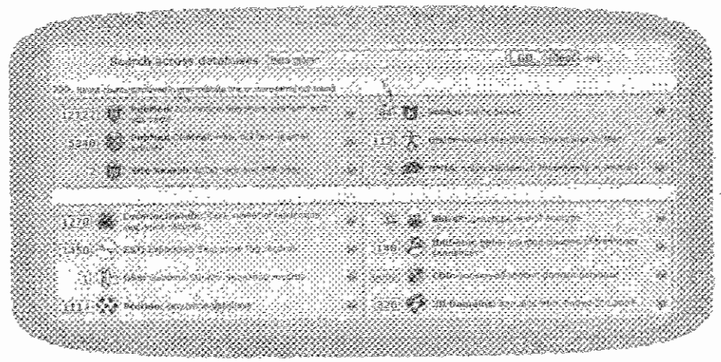


FIGURE 2.5. There are thousands of accession numbers corresponding to many genes and proteins. A search with the query "beta globin" from the main page of NCBI shows the results across the databases of the Entrez search engine. There are over 1000 each of core nucleotide sequences, expressed sequence tags (ESTs), and proteins. The RefSeq project is particularly important in trying to provide the best representative sequence of each normal (nonmutated) transcript produced by a gene and of each distinct, normal protein sequence.

A GenBank or RefSeq accession number refers to the most recent version of a given sequence. For example NM_000558.3 is currently a RefSeq identifier for human alpha globin. The suffix ".3" is the version number. By default, if you do not specify a version number then the most recent version is provided. Try doing an Entrez nucleotide search for NM_000558.1 and you can learn about the revision history of that accession number. In Chapter 3 we will learn how to compare two sequences; you can blast NM_000558.1 against NM_000558.3 to see the differences, or view the results in web document 2.2 at ➧ http://www.bioinfbook.org/chapter2.

RefSeq entries are curated by the staff at NCBI, and are nearly nonredundant. However, there can be two proteins encoded by distinct genes sharing 100% amino acid identity. Each is assigned its own unique RefSeq identifier. For example, the alpha-1 globin and alpha-2 globin genes in human are physically separate genes that encode proteins with identical sequences. The encoded alpha-1 globin and alpha-2 globin proteins are assigned the RefSeq identifiers NP_000549 and NP_000508.

Refseq entries have different status levels (predicted, provisional, and reviewed), but in each case the RefSeq entry is intended to unify the sequence records. You can recognize a RefSeq accession by its format, such as NP_000509 (P stands for beta globin protein) or NM_006744 (for beta globin mRNA). A variety of RefSeq identifiers are shown in Table 2.7, and examples of beta globin identifiers are given in Table 2.8.

TABLE 2.7 Formats of Accession Numbers for RefSeq Entries

Molecule	Accession Format	Genome
Complete genome	NC_123456	Complete genomic molecules, including genomes, chromosomes, organelles, and plasmids
Genomic DNA	NW_123456 NW_123456789	Intermediate genomic assemblies
Genomic DNA	NZ_ABCD12345678	Collection of whole genome shotgun sequence data
Genomic DNA	NT_123456	Intermediate genomic assemblies (BAC and/or WGS sequence data)
mRNA	NM_123456 or NM_123456789	Transcript products: mature mRNA protein-coding transcripts
Protein	NP_123456 or NM_123456789	Protein products (primarily full-length)
RNA	NR_123456	Noncoding transcripts (e.g. structural RNAs, transcribed pseudogenes)

There are currently 21 different RefSeq accession formats. The methods include expert manual curation, automated curation, or a combination. Abbreviations: BAC, bacterial artificial chromosome; WGS, whole genome shotgun (see Chapter 13).

Source: Adapted from ➧ http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions (March 2007).

TABLE 2.8 RefSeq Accession Numbers Corresponding to Human Beta Globin

Category	Accession	Size	Description
DNA	NC_000011	134,452,384 bp	Genomic contig
DNA	NM_000518.4	626 bp	DNA corresponding to mRNA
DNA	NG_000007.3	81,706 bp	Genomic reference
DNA	NW_925006.1	1,606 bp	Alternate assembly
Protein	NP_000509.1	147 amino acids	Protein

The Consensus Coding Sequence (CCDS) Project

The Consensus Coding Sequence (CCDS) project was established to identify a core set of protein coding sequences that provide a basis for a standard set of gene annotations. The CCDS project is a collaboration between four groups (EBI, NCBI, the Wellcome Trust Sanger Institute, and the University of California, Santa Cruz [UCSC]). Currently, the CCDS project has been applied to the human and mouse genomes, and thus its scope is considerably more limited than RefSeq.

You can learn about the CCDS project at ➧ http://www.ncbi.nlm.nih.gov/projects/CCDS/.

Access to Information via Entrez Gene at NCBI

How can one navigate through the bewildering number of protein and DNA sequences in the various databases? An emerging feature is that the various databases are increasingly interconnected, providing a variety of convenient links to each other and to algorithms that are useful for DNA, RNA, and protein analysis. Entrez Gene (formerly LocusLink) is particularly useful as a major portal. It is a curated database containing descriptive information about genetic loci (Maglott et al., 2007). You can obtain information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, OMIM numbers, UniGene clusters, HomoloGene (a database that reports eukaryotic orthologs), map locations, and related websites.

To illustrate the use of Entrez Gene we will search for human myoglobin. The result of entering an Entrez Gene search is shown in Fig. 2.6. Note that in performing this search, it can be convenient to restrict the search to a particular organism of interest. (This can be done using the "limits" tab on the Entrez Gene page.) The "Links" button (Fig. 2.6, top right) provides access to various other database entries on myoglobin. Clicking on the main link to the human myoglobin entry results in the following information (Fig. 2.7):

- At the top right, there is a table of contents for the Entrez Gene myoglobin entry. Below it are further links to myoglobin entries in NCBI databases (e.g. protein and nucleotide databases and PubMed), as well as external databases (e.g. Ensembl and UCSC; see below and Chapter 16).
- Entrez Gene provides the official symbol and name for human myoglobin, MB.
- A schematic overview of the gene structure is provided, hyperlinked to the Map Viewer (see below).
- There is a brief description of the function of MB, defining it as a carrier protein of the globin family.

Entrez Gene is accessed from the main NCBI web page (by clicking All Databases). Currently (November 2008), Entrez Gene encompasses about 5,700 taxa and 4.6 million genes. We will explore many of the resources within Entrez Gene in later chapters.

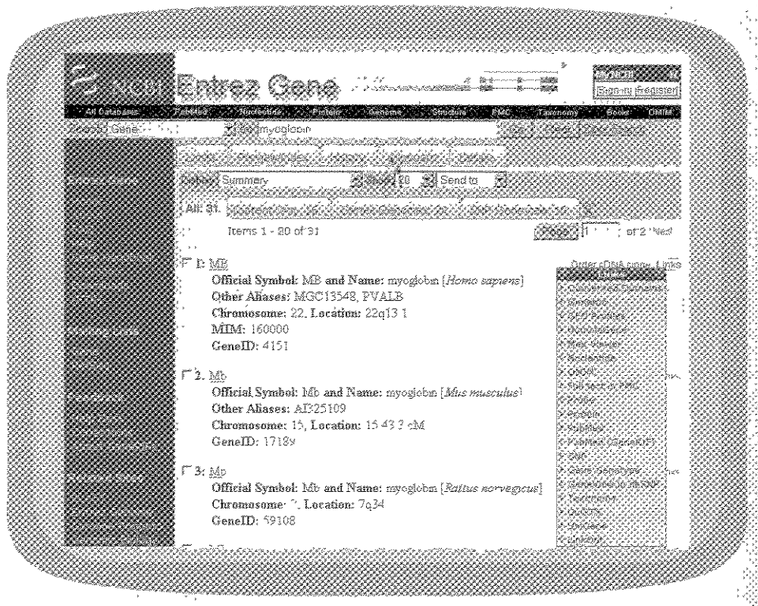


FIGURE 2.6. Result of a search for "myoglobin" in Entrez Gene. Information is provided for a variety of organisms, including Homo sapiens, Mus musculus, *and* Rattus norvegicus. *The links button (top right) provides access to information on myoglobin from a variety of other databases.*

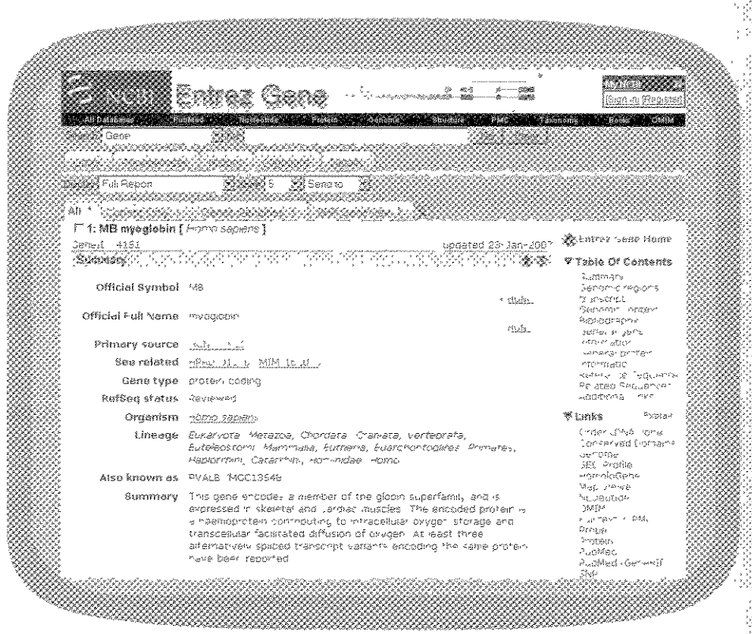


FIGURE 2.7. Portion of the Entrez Gene entry for human myoglobin. Information is provided on the gene structure, chromosomal location, as well as a summary of the protein's function. RefSeq accession numbers are also provided (not shown); you can access them by clicking "Reference sequences" in the table of contents (top right). The menu (right sidebar) provides extensive links to additional databases, including PubMed, OMIM, UniGene, a variation database (dbSNP), HomoloGene (with information on homologs), a gene ontology database, and Ensembl viewers at EBI. We will describe these resources in later chapters.

- The Reference Sequence (RefSeq) accession numbers are provided: NM_005368 for the DNA sequence encoding the longest myoglobin transcript and NP_005359 for the protein entry. GenBank accession numbers corresponding to myoglobin (both nucleotide and protein) are also provided.

Figure 2.8 shows the standard, default form of a typical Entrez Protein record (for myoglobin). It is simple to obtain a variety of formats by changing the Entrez display options. By using the Display pulldown menu (Fig. 2.8a) one can obtain

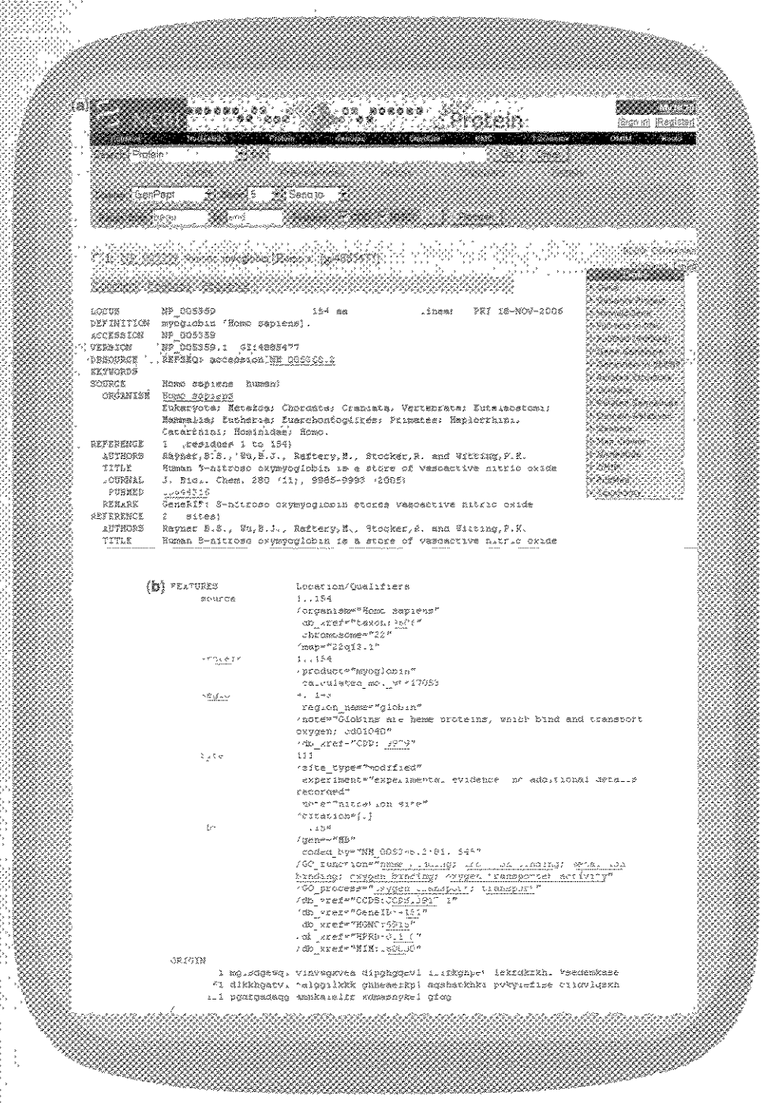


*FIGURE 2.8. Display of an Entrez Protein record for human myoglobin. This is a typical entry for any protein. (a) Top portion of the record. Key information includes the length of the protein (154 amino acids), the division (PRI, or primate), the accession number (NP_005359), the organism (*H. sapiens*), literature references, comments on the function of globins, and many links to other databases (right side). At the top of the page, the display option allows you to obtain this record in a variety of formats, such as FASTA (Figure 2.9). (b) Bottom portion of the record. This includes features such as the coding sequence (CDS). The amino acid sequence is provided at the bottom in the single letter amino acid code.*

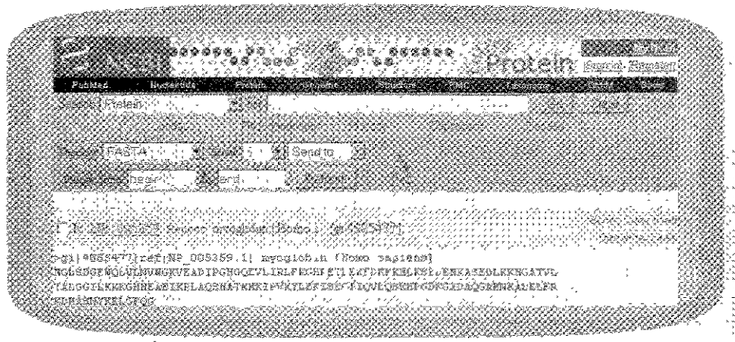


FIGURE 2.9 The protein entry for human myoglobin can be displayed in the FASTA format. This is easily accomplished by adjusting the "Display" pull-down menu from an Entrez protein record. The FASTA format is used in a variety of software programs that we will use in later chapters.

FASTA is both an alignment program (described in Chapter 3) and a commonly used sequence format (further described in Chapter 4).

the commonly used FASTA format for protein (or DNA) sequences, as shown in Fig. 2.9. Note also that by clicking the CDS (coding sequence) link of an Entrez Protein or Entrez Nucleotide record (shown in Fig. 2.8b), you can obtain the nucleotides that encode a particular protein, typically beginning with a start methionine (ATG) and ending with a stop codon (TAG, TAA, or TGA). This can be useful for a variety of applications including multiple sequence alignment (Chapter 6) and molecular phylogeny (Chapter 7).

Relationship of Entrez Gene, Entrez Nucleotide, and Entrez Protein

If you are interested in obtaining information about a particular DNA or protein sequence, it is reasonable to visit Entrez Nucleotide or Entrez Protein and do a search. A variety of search strategies are available, such as limiting the output to a particular organism or taxonomic group of interest, or limiting the output to RefSeq entries.

There are also many advantages to beginning your search through Entrez Gene. There, you can identify the official gene name, and you can be assured of the chromosomal location of the gene (thus providing unambigous information about which particular gene you are studying). Furthermore, each Entrez Gene entry includes a section of reference sequences that provides all the DNA and protein variants that are assigned RefSeq accession numbers.

Comparison of Entrez Gene and UniGene

As described above, the UniGene project assigns one cluster of sequences to one gene. For example, for *RBP4* there is one UniGene entry with the UniGene accession number Hs.50223. This UniGene entry includes a list of all the GenBank entries, including ESTs, that correspond to the *RBP4* gene. The UniGene entry also includes mapping information, homologies, and expression information (i.e., a list of the tissues from which cDNA libraries were generated that contain ESTs corresponding to the RBP gene).

Entrez Gene now has about 40,000 human gene entries (as of November 2008).

UniGene and Entrez Gene have features in common, such as links to OMIM, homologs, and mapping information. They both show RefSeq accession numbers. There are four main differences between UniGene and Entrez Gene:

1. UniGene has detailed expression information; the regional distributions of cDNA libraries from which particular ESTs have been sequenced are listed.
2. UniGene lists ESTs corresponding to a gene, allowing one to study them in detail.
3. Entrez Gene may provide a more stable description of a particular gene; as described above, UniGene entries may be collapsed as genome-sequencing efforts proceed.
4. Entrez Gene has fewer entries than UniGene, but these entries are better curated.

Entrez Gene and HomoloGene

The HomoloGene database provides groups of annotated proteins from a set of completely sequenced eukaryotic genomes. Proteins are compared (by blastp; see Chapter 4), placed in groups of homologs, and then the protein alignments are matched to the corresponding DNA sequences. This allows distance metrics to be calculated such as Ka/Ks, the ratio of nonsynonymous to synonymous mutations (see Chapter 7). You can find a HomoloGene entry for a gene/protein of interest by following a link on the Entrez Gene page.

A search of HomoloGene with the term hemoglobin results in dozens of matches for myoglobin, alpha globin, and beta globin. By clicking on the beta globin group one gains access to a list of proteins with RefSeq accession numbers from human, chimpanzee, dog, mouse, and chicken. The pairwise alignment scores (see Chapter 3) are summarized and linked to, and the sequences can be displayed as a multiple sequence alignment (Chapter 6), or in the FASTA format.

HomoloGene is available by clicking All Databases from the NCBI home page, or at ▶ http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene. Release 53 (March 2007) has over 170,000 groups. We will define homologs in Chapter 3.

Access to Information: Protein Databases

In many cases you are interested in obtaining protein sequences. The Entrez Protein database at NCBI consists of translated coding regions from GenBank as well as sequences from external databases (the Protein Information Resource [PIR], SWISS-PROT, Protein Research Foundation [PRF], and the Protein Data Bank [PDB]). The EBI also provides information on proteins via these major databases. We will next explore ways to obtain protein data through UniProt, an authoritative and comprehensive protein database.

EBI offers access to over a dozen different protein databases, listed at ▶ http://www.ebi.ac.uk/Databases/protein.html.

UniProt

The Universal Protein Resource (UniProt) is the most comprehensive, centralized protein sequence catalog (UniProt Consortium, 2009). Formed as a collaborative effort in 2002, it consists of a combination of three key databases. (1) Swiss-Prot is considered the best-annotated protein database, with descriptions of protein structure and function added by expert curators. (2) The translated EMBL (TrEMBL) Nucleotide Sequence Database Library provides automated (rather than manual) annotations of proteins not in Swiss-Prot. It was created because of the vast number of protein sequences that have become available through genome sequencing projects. (3) PIR maintains the Protein Sequence Database, another protein database curated by experts.

UniProt is organized in three database layers. (1) The UniProt Knowledgebase (UniProtKB) is the central database that is divided into the manually annotated UniProtKB/Swiss-Prot and the computationally annotated UniProtKB/TrEMBL.

The European Bioinformatics Institute (EBI) in Hinxton and the Swiss Institute of Bioinformatics (SIB) in Geneva created Swiss-Prot and TrEMBL. PIR is a division of the National Biomedical Research Foundation (▶ http://pir.georgetown.edu/) in Washington, D.C. PIR was founded by Margaret Dayhoff, whose work is described in Chapter 3. The UniProt web site is ▶ http://www.uniprot.org. It contains over 7 million entries (release 14.4, November 2008).

To access UniProt from EBI, visit ▶ http://www.ebi.ac.uk/uniprot/. To access UniProt from ExPASy, visit ▶ http://www.expasy.org/sprot/.

(2) The UniProt Reference Clusters (UniRef) offer nonredundant reference clusters based on UniProtKB. UniRef clusters are available with members sharing at least 50%, 90%, or 100% identity. (3) The UniProt Archive, UniParc, consists of a stable, nonredundant archive of protein sequences from a wide variety of sources (including model organism databases, patent offices, RefSeq, and Ensembl).

You can access UniProt directly from its website, or from EBI or ExPASy.

The Sequence Retrieval System at ExPASy

One of the most useful resources available to obtain protein sequences and associated data is provided by ExPASy, the Expert Protein Analysis System. The ExPASy server is a major resource for proteomics-related analysis tools, software, and databases. In addition to providing access to the UniProt database, ExPASy serves as a portal for the Sequence Retrieval System (SRS). The query page has four rectangular boxes (Fig. 2.10). Each has an associated pull-down menu, and as a default condition each says "AllText." In the first box, type "retinol-binding." (Note that queries should consist of one word.) In the second box, type "human," change the corresponding pull-down menu to "organism," then click "do query." You see 10 entries listed. Click the link in which we are interested (SWISS_PROT: RETB_HUMAN P02753).

An output consists of a SwissProt record. This provides very useful, well-organized information, including alternative names and accession numbers; literature links; functional data and information about cellular localization; links to GenBank and other database records for both the RBP protein and gene; and links to many databases such as OMIM, InterPro, Pfam, Prints, GeneCards, PROSITE, and two-dimensional protein gel databases. We will describe these resources later (Chapters 6 and 10). The record includes features; note that by clicking on any of the linked features, you can see the protein sequence with that feature highlighted in color. While we have mentioned several key ways to acquire sequence data, there are dozens of other useful servers. As an example, the Protein Information Resource (PIR) provides access to sequences (Wu et al., 2002). PIR is especially useful for its efforts to annotate functional information on proteins.

ExPASy is a proteomics server of the Swiss Institute of Bioinformatics (▶ http://www.expasy.ch/), another portal from which the Sequence Retrieval System (SRS) is accessed. From ▶ http://www.expasy.ch/srs5/, click "Start a new SRS session," then click "continue." SRS was created by Lion Biosciences, and a list of several dozen publicly available SRS servers is at ▶ http://downloads.lionbio.co.uk/publicsrs.html.

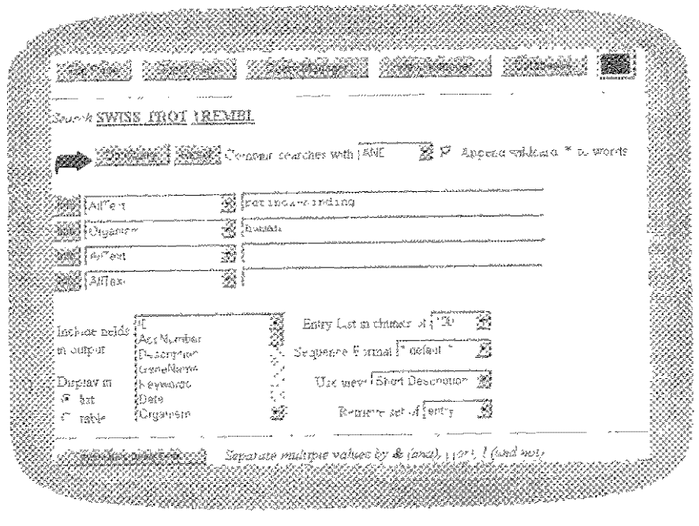


FIGURE 2.10. Format of a query at the Sequence Retrieval System (SRS) of the Expert Protein Analysis System (ExPASy) (▶ http://www.expasy.ch/srs5/). This website provides one of the most useful resources for protein analysis. You can also access the SRS through other sites such as the European Bioinformatics Institute (▶ http://srs6.ebi.ac.uk/).

Access to Information: The Three Main Genome Browsers

Genome browsers are databases with a graphical interface that presents a representation of sequence information and other data as a function of position across the chromosomes. We will focus on viral, prokaryotic, and eukaryotic chromosomes in Chapters 14 to 19. Genome browsers have emerged as an essential tool for organizing information about genomes. We will now briefly introduce the three principal genome browsers and describe how they may be used to acquire information about a gene or protein of interest.

Genomes are analyzed over time in assemblies (see Chapter 13). The main human genome browsers share the same underlying assemblies, and differ in the ways they annotate and present information. NCBI Build 36 (November, 2005) is an example of a human assembly.

The Map Viewer at NCBI

The NCBI Map Viewer includes chromosomal maps (both physical maps and genetic maps; see Chapter 16) for a variety of organisms, including metazoans (animals), fungi, and plants. Map Viewer allows text-based queries (e.g., "beta globin") or sequence-based queries (e.g., BLAST; see Chapter 4). For each genome, four levels of detail are available: (1) the home page of an organism; (2) the genome view, showing ideograms (representations of the chromosomes); (3) the map view, allowing you to view regions at various levels of resolution; and (4) the sequence view, displaying sequence data as well as annotation of interest such as the location of genes.

The Map Viewer is accessed from the main page of NCBI or via ▶ http://www.ncbi.nlm.nih.gov/mapview/. Records in Entrez Gene, Entrez Nucleotide, and Entrez Protein also provide direct links to the Map Viewer.

The University of California, Santa Cruz (UCSC) Genome Browser

The UCSC browser currently supports the analysis of three dozen vertebrate and invertebrate genomes, and it is perhaps the most widely used genome browser for human and other prominent organisms such as mouse. The Genome Browser provides graphical views of chromosomal locations at various levels of resolution (from several base pairs up to hundreds of millions of base pairs spanning an entire chromosome). Each chromosomal view is accompanied by horizontally oriented annotation tracks. There are hundreds of available tracks in categories such as mapping and sequencing, phenotype and disease associations, genes, expression, comparative genomics, and genomic variation. These annotation tracks offer the Genome Browser tremendous depth and flexibility. The Genome Browser has a complementary, interconnected Table Browser that provides tabular output of information.

The UCSC genome browser is available from the UCSC bioinformatics site at ▶ http://genome.ucsc.edu. You can see examples of it in Figs. 5.17, 5.20, 6.10, 8.8, 12.8, 16.4, and 9.20.

As an example of how to use the browser, go the UCSC bioinformatics site, click Genome Browser, set the clade (group) to Vertebrate, the genome to human, the assembly to March 2006 (or any other build date), and under "position or search term" type beta globin (Fig. 2.11a). Click submit and you will see a list of known genes and a RefSeq gene entry for beta globin on chromosome 11 (Fig. 2.11b). By following this RefSeq link you will view the beta globin gene (spanning about 1600 base pairs) on chromosome 11, and can perform detailed analyses of the beta globin gene (including neighboring regulatory elements), the messenger RNA (see Chapter 8), and the protein (Fig. 2.11c).

The Ensembl Genome Browser

The Ensembl project offers a series of comprehensive websites for a variety of eukaryotic organisms (Hubbard et al., 2007). The project's goals are to automatically analyze and annotate genome data (see Chapter 13) and to present genomic data via its

In a separate approach, one can obtain the HIV-1 reverse transcriptase sequence from SRS. Select the SwissProt database to search. In the four available dialog boxes, set one row to "organism" and "HIV-1," then set another row to "AllText" and "reverse." Upon clicking "Do query," a list of several dozen entries is returned; many of these are identified as fragments and may be ignored. One entry is SWISS_PROT:POL_HV1A2 (SWISS-PROT accession P03369), a protein of 1437 amino acids. Following the SwissProt link, one finds the "NiceProt" for this database entry. This information includes entry and modification dates, names of this protein and synonyms, references (with PubMed links), comments (including a brief functional description), cross-references to over a dozen other useful databases, a keyword listing, features such as predicted secondary structure, and finally, the amino acid sequence in the single-letter amino acid code and the predicted molecular weight of the protein. For this case, the gene encodes a protein as an unprocessed precursor that is further cleaved to generate many smaller proteins, including matrix protein p17, capsid protein p24, nucleocapsid protein p7, a viral protease, a reverse transcriptase/ribonuclease H multifunctional protein, and an integrase. These features are clearly described in the UniProtKB/Swiss-Prot entry for P03369.

Histones

By clicking the Details tab on an Entrez Protein search, you can see that the command is interpreted as "txid9606[Organism:exp] AND histone[All Fields]". The Boolean operator AND is included between search terms by default.

The biological complexity of proteins can be astonishing, and accessing information about some proteins can be extraordinarily challenging. Histones are among the most familiar proteins by name. They are small proteins (12 to 20 kilodaltons) that are localized to the nucleus where they interact with DNA. There are five major histone subtypes as well as additional variant forms; the major forms serve as core histones (the H2A, H2B, H3, and H4 families) which ~147 base pairs of DNA wrap around, and linker histones (the H1 family). Suppose you want to inspect a typical human histone for the purpose of understanding the properties of a representative gene and its corresponding protein. A challenge is that there are currently 80,000 histone entries in Entrez Protein (November 2008). Restricting the output to human histone proteins (using the command "txid9606[Organism:exp] histone") there are currently 5000 human histone proteins, of which 1200 have RefSeq accession numbers. Some of these are histone deacetylases and histone acetyltransferases; by expanding the query to "txid9606[Organism:exp] AND histone[All Fields] NOT deacetylase NOT acetyltransferase" there are 800 proteins with RefSeq accession numbers. There are many additional strategies for limiting Entrez searches (Box 2.2).

The Histone Sequence Database is available at ➤ http://research.nhgri.nih.gov/histones/. (Sullivan et al., 2002). It was created by David Landsman, Andy Baxevanis, and colleagues at the National Human Genome Research Institute.

You can find links to a large collection of specialized databases at ➤ http://www.expasy.org/links.html, the Life Science Directory at the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB).

How can the search be further pursued? (1) You may select a histone at random and study it although you may not know whether it is representative. (2) There are specialized, expert-curated databases available online for many genes, proteins, diseases, and other molecular features of interest. The Histone Sequence Database (Sullivan et al., 2002) shows that the human genome has about 80 histone genes, including a cluster of 68 adjacent genes on chromosome 6p. This information is useful to understand the scope of the family. (3) There are databases of protein families, including Pfam and InterPro. We will introduce these in Chapters 6 (multiple sequence alignment) and 10 (proteomics). Such databases offer succinct descriptions of protein and gene families and can orient you toward identifying representative members.

Access to Biomedical Literature

The NLM website is ➤ http://www.nlm.nih.gov/.

The NLM is the world's largest medical library. In 1971 the NLM created MEDLINE (Medical Literature, Analysis, and Retrieval System Online), a bibliographic database. MEDLINE currently contains over 18 million references to journal articles in the life sciences with citations from over 4300 biomedical journals in 70 countries. Free access to MEDLINE is provided on the World Wide Web through PubMed (➤ http://www.ncbi.nlm.nih.gov/PubMed/), which is developed by NCBI. While MEDLINE and PubMed both provide bibliographic citations, PubMed also contains links to online full-text journal articles. PubMed also provides access and links to the integrated molecular biology databases maintained by NCBI. These databases contain DNA and protein sequences, genome-mapping data, and three-dimensional protein structures.

MEDLINE is also accessible through the SRS at the European Bioinformatics Institute via ➤ http://srs.ebi.ac.uk/. A PubMed tutorial is offered at ➤ http://www.nlm.nih.gov/bsd/pubmed_tutorial/m1001.html. The growth of MEDLINE is described at ➤ http://www.nlm.nih.gov/bsd/medline_growth.html. Despite the multinational contributions to MEDLINE, the percentage of articles written in English has risen from 59% at its inception in 1966 to 92% in the year 2008 (➤ http://www.nlm.nih.gov/bsd/medline_lang_distr.html).

PubMed Central and Movement toward Free Journal Access

The biomedical research community has steadily increased access to literature information. Groups such as the Association of Research Libraries (ARL) monitor the migration of publications to an electronic form. Thousands of journals are currently available online. Increasingly, online versions of articles include supplementary material such as molecular data (e.g., the sequence of complete

BOX 2-2
Tips for Using Entrez Databases

- The Boolean operators AND, OR, and NOT must be capitalized. By default, AND is assumed to connect two terms; subject terms are automatically combined.
- You can perform a search of a specific phrase by adding quotation marks. This may potentially restrict the output, so it is a good idea to repeat a search with and without quotation marks.
- Boolean operators are processed from left to right. If you add parentheses, the enclosed terms will be processed as a unit rather than sequentially. A search of Entrez Gene with the query "globin AND promoter OR enhancer" yields 4800 results; however, by adding parentheses, the query "globin AND (promoter or enhancer)" yields just 70 results.
- If you are interested in obtaining results from a particular organism (or from any taxonomic group such as the primates or viruses), try beginning with TaxBrowser to select the organism first. See Fig. 2–11 for a detailed explanation. Adding the search term human[ORGN] will restrict the output to human. Alternatively, you can use the taxonomy identifier for human, 9606, as follows: txid9606[Organism:exp]
- A variety of limiters can be added. In Entrez Protein, the search 500000:999999[Molecular weight] will return proteins having a molecular weight from 500,000 to 1 million daltons. If you would like to see proteins between 10,000 and 50,000 daltons that I have worked on, enter 010000:050000[Molecular weight] pevsner j (or, equivalently, 010000[MOLWT]: 050000[MOLWT] AND pevsner j[Author]).
- By truncating a query with an asterisk, you can search for all records that begin with a particular text string. For example, a search of Entrez Nucleotide with the query "globin" returns 5800 results; querying with "glob*" returns 8.2 million results. These include entries with the species *Chaetomium globosum* or the word global.
- Keep in mind that any Entrez query can be applied to a BLAST search to restrict its output (Chapter 4).

The National Library of Medicine also offers access to PubMed through NLM Gateway (▶ http://gateway.nlm.nih.gov). This comprehensive service includes access to a variety of NLM databases not offered through PubMed, such as meeting abstracts and a medical encyclopedia.

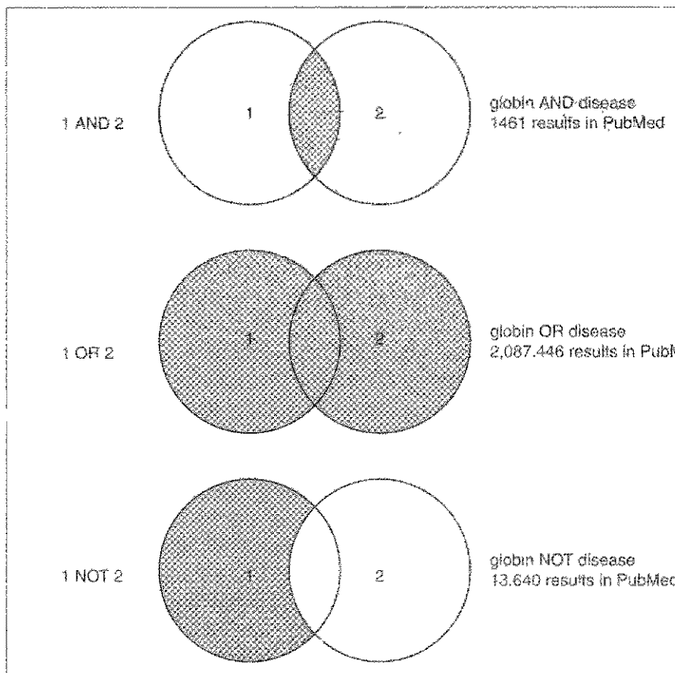
genomes, or gene expression data) or videotapes illustrating an article. PubMed Central provides a central repository for biological literature (Roberts, 2001). All these articles have been peer reviewed and published simultaneously in another journal. As of 2008, publications resulting from research funded by the NIH, Wellcome Trust, and Medical Research Council must be made freely available in PubMed Central.

The ARL website is ▶ http://www.arl.org/index.shtml.

Example of PubMed Search: RBP

A search of PubMed for information about "RBP" yields 1700 entries. Box 2.3 describes the basics of using Boolean operators in PubMed. There are many additional ways to limit this search. Press "limits" and try applying features such as restricting the output to articles that are freely available through PubMed Central.

BOX 2-3
Venn Diagrams of Boolean Operators AND, OR, and NOT for Hypothetical Search Terms 1 and 2



The AND command restricts the search to entries that are both present in a query. The OR command allows either one or both of the terms to be present. The NOT command excludes query results. The shaded areas represent search queries that are retrieved. Examples are provided for the queries "lipocalin" or "retinol-binding protein" in PubMed. The Boolean operators affect the searches as indicated.

The Medical Subject Headings (MeSH) browser provides a convenient way to focus or expand a search. MeSH is a controlled vocabulary thesaurus containing 25,000 descriptors (headings). From PubMed, click "MeSH Database" on the left sidebar and enter "retinol-binding protein." The result suggests a series of possibly related topics. By adding MeSH terms, a search can be focused and structured according to the specific information you seek. Lewitter (1998) and Fielding and Powell (2002) discuss strategies for effective MEDLINE searches, such as avoiding inconsistencies in MeSH terminology and finding a balance between sensitivity (i.e., finding relevant articles) and specificity (i.e., excluding irrelevant citations). For example, for a subject that is not well indexed, it is helpful to combine a text keyword with a MeSH term. It can also be helpful to use truncations; for example, the search "therap*" introduces a wildcard that will retrieve variations such as therapy, therapist, and therapeutic. Figure 2.13 provides an example of sensitivity and specificity in a PubMed search for articles on hemoglobin.

The MeSH website at NLM is ▶ http://www.nlm.nih.gov/mesh/meshhome.html; you can also access MeSH via the NCBI website including its PubMed page.

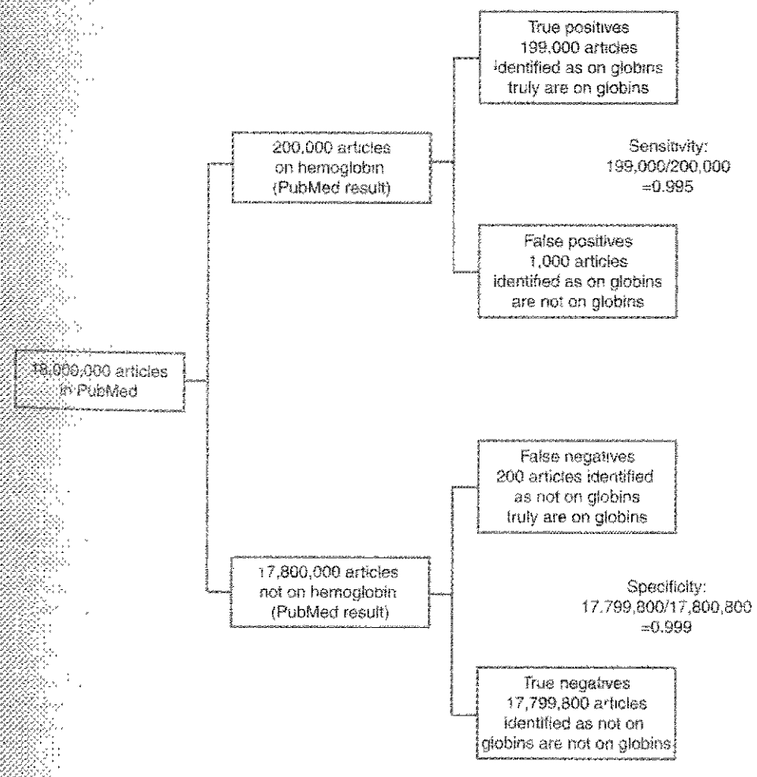


FIGURE 2.13. Sensitivity and specificity in a database search. We will describe sensitivity and specificity in Chapter 3 (see Fig. 3.27) but can begin thinking about those concepts in terms of a hypothetical search of PubMed for hemoglobin. Each search of a database yields results that are reported (positives) or not (negatives). According to some "gold standard" or objective measure of the truth, these results may be true positives (e.g., a search for globins does return literature citations on globins) or false positives (e.g., a search for glob returns information about the species C. globosum but those citations are irrelevant to globins). The sensitivity is defined as the proportion of true positives relative to true plus false positives. There also will be many negative results (lower portion of figure). These may include true negatives (e.g., articles that do not describe globins and are not included in the search results) and false negatives (e.g., articles that do discuss globins but are not part of the search results; this might occur if the title and abstract do not mention globins but the body of the article does). Specificity may be defined as the proportion of true negative results divided by the sum of true negative and false positive results.*

Perspective

Bioinformatics is a young, emerging field whose defining feature is the accumulation of biological information in databases. The three major DNA databases—GenBank, EMBL, and DDBJ—are adding several million new sequences each year as well as billions of nucleotides. Beginning in 2008, terabases (thousands of gigabases) of DNA sequence are arriving.

In this chapter, we described ways to find information on the DNA and/or protein sequence of globins, RBP4, and the HIV *pol* gene. In addition to the three major databases, a variety of additional resources are available on the web. Increasingly, there is no single correct way to find information; many approaches are possible. Moreover, resources such as those described in this chapter—NCBI, ExPASy, EBI/EMBL, and Ensembl—are closely interrelated, providing links between the databases.

Pitfalls

There are many pitfalls associated with the acquisition of both sequence and literature information. In any search, the most important first step is to define your goal; for example, decide whether you want protein or DNA sequence data. A common difficulty that is encountered in database searches is receiving too much information; this problem can be addressed by learning how to generate specific searches with appropriate limits.

Web Resources

You can visit the website for this book (▶ http://www.bioinfbook.org) to find many of the URLs, organized by chapter. The Wiley-Blackwell website for this book is http://www.wiley.com/go/pevsnerbioinformatics.

Discussion Questions

[2-1] What categories of errors occur in databases? How are these errors assessed?

[2-2] How is quality control maintained in GenBank, given that thousands of individual investigators submit data?

Problems

[2-1] In this chapter we explored histones as an example of a protein that can be challenging to study because it is part of a large gene family. Another challenging example is ubiquitin. How many ubiquitins are there in the human genome, and what is the sequence of a prototypical (that is, representative) ubiquitin?

[2-2] How many human proteins are bigger than 300,000 daltons? *Hints:* Try to first limit your search to human by using TaxBrowser. Then follow the link to Entrez Protein, where all the results will be limited to human. Enter a command in the format xxxxxx:yyyyyy[molwt] to restrict the output to a certain number of daltons; for example, 002000:019000[molwt] will select proteins of molecular weight 2,000 to 10,000.

[2-3] You are interested in learning about genes involved in breast cancer. Which genes have been implicated? What are the DNA and protein accession numbers for several of these genes? Try all of these approaches: PubMed, Entrez, OMIM, and SRS at ExPASy.

[2-4] An ATP (adenosine triphosphate) binding cassette (ABC) is an example of a common protein domain that is found in many so-called ABC transporter proteins. However, you are not familiar with this motif and would like to learn more. Approximately how many human proteins have ABC domains? Approximately how many bacterial proteins have ABC domains? Which of the resources you used in problem 2.3 is most useful in providing you a clear definition of an ABC motif? (We will discuss additional resources to solve this problem in Chapter 16.)

[2-5] Find the accession number of a lipocalin protein (e.g., retinol-binding protein, lactoglobulin, any bacterial lipocalin, glycodelin, or odorant-binding protein). First, use Entrez, then UniGene, then OMIM. Which approach is most effective? What is the function of this protein?

[2-6] Three prominent tools for *text*-based searching of molecular information are:

- the National Center for Biotechnology Information's PubMed, Entrez, and OMIM tools (▶ http://www.ncbi.nlm.nih.gov),
- the European Bioinformatics Institute (EBI) Sequence Retrieval System (SRS) (▶ http://srs.ebi.ac.uk) or its related SRS site (▶ http://www.expasy.ch/srs5/), and
- DBGET, the GenomeNet tool of Kyoto University, and the University of Tokyo (▶ http://www.genome.ad.jp/dbget/dbget.2.html) literature database LitDB.

You are interested in learning more about West Nile virus. What happens when you use that query to search each of these three resources?

[2-7] You would like to know what articles about viruses have been published in the journal *BMC Bioinformatics*. Do this search using PubMed.

Self-Test Quiz

[2-1] Which of the following is a RefSeq accession number corresponding to an mRNA?
(a) J01536
(b) NM_15392
(c) NP_52280
(d) AAB134506

[2-2] Approximately how many human clusters are currently in UniGene?
(a) About 8,000
(b) About 25,000
(c) About 100,000
(d) About 300,000

[2-3] You have a favorite gene, and you want to determine in what tissues it is expressed. Which one of the following resources is likely the most direct route to this information?
(a) UniGene
(b) Entrez
(c) PubMed
(d) PCR

[2-4] Is it possible for a single gene to have more than one UniGene cluster?
(a) Yes
(b) No

[2-5] Which of the following databases is derived from mRNA information?
(a) dbEST
(b) PBD
(c) OMIM
(d) HTGS

[2-6] Which of the following databases can be used to access text information about human diseases?
(a) EST
(b) PBD
(c) OMIM
(d) HTGS

[2-7] What is the difference between RefSeq and GenBank?
(a) RefSeq includes publicly available DNA sequences submitted from individual laboratories and sequencing projects.
(b) GenBank provides nonredundant curated data.
(c) GenBank sequences are derived from RefSeq.
(d) RefSeq sequences are derived from GenBank and provide nonredundant curated data.

[2-8] If you want literature information, what is the best website to visit?
(a) OMIM
(b) Entrez
(c) PubMed
(d) PROSITE

[2-9] Compare the use of Entrez and ExPASy to retrieve information about a protein sequence.
(a) Entrez is likely to yield a more comprehensive search because GenBank has more data than EMBL.
(b) The search results are likely to be identical because the underlying raw data from GenBank and EMBL are the same.
(c) The search results are likely to be comparable, but the SwissProt record from ExPASy will offer a different output format with distinct kinds of information.

Suggested Reading

Bioinformatics databases are evolving extremely rapidly. Each January, the first issue of the journal *Nucleic Acids Research* includes nearly 100 brief articles on databases. These include descriptions of NCBI (Wheeler et al., 2007), GenBank (Benson et al., 2009), and EMBL (Cochrane et al., 2008).

References

Adams, M. D., et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).

Altschul, S. E., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Altschul, S. E., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

Beach, E. F. Beccari of Bologna. The discoverer of vegetable protein. *J. Hist. Med.* **16**, 354–373 (1961).

Bentley, D. R., et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. GenBank. *Nucl. Acids Res.* **37**, D26–D31 (2009).

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. dbEST—database for "expressed sequence tags." *Nat. Genet.* **4**, 332–333 (1993).

Brooksbank, C., Camon, E., Harris, M. A., Magrane, M., Martin, M. J., Mulder, N., O'Donovan, C., Parkinson, H., Tuli, M. A., Apweiler, R., Birney, E., Brazma, A., Henrick, K., Lopez, R., Stoesser, G., Stoehr, P., and Cameron, G. The European Bioinformatics Institute's data resources. *Nucleic Acids Res.* **31**, 43–50 (2003).

Cochrane, G., et al. Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **36**, D5–D12 (2008).

Fielding, A. M., and Powell, A. Using Medline to achieve an evidence-based approach to diagnostic clinical biochemistry. *Ann. Clin. Biochem.* **39**, 345–350 (2002).

Fleischmann, R. D., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).

Frankel, A. D., and Young, J. A. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67**, 1–25 (1998).

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).

Hubbard, T. J., et al. Ensembl 2007. *Nucleic Acids Res.* **35**, D610–D617 (2007).

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P. E., and Berman, H. M. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.* **34**, D302–D305 (2006).

Kulikova, T., et al. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.* **35**, D16–D20 (2007).

Lewitter, F. Text-based database searching. *Bioinformatics: A Trends Guide* **19**, 3–5 (1998).

Ley, T. J., et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126–128 (2000).

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res.* **35**, D26–D31 (2007).

Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T., and Tateno, Y. DDBJ in the stream of various biological data. *Nucleic Acids Res.* **32**, D31–D34 (2004).

Olson, M., Hood, L., Cantor, C., and Botstein, D. A common language for physical mapping of the human genome. *Science* **245**, 1434–1435 (1989).

Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucl. Acids Res.* **37**, D32–D36 (2009).

Roberts, R. J. PubMed Central: The GenBank of the published literature. *Proc. Natl. Acad. Sci. USA* **98**, 381–382 (2001).

Sullivan, S., Sink, D. W., Trout, K. L., Makalowska, I., Taylor, P. M., Baxevanis, A. D., and Landsman, D. The Histone Database. *Nucleic Acids Res.* **30**, 341–342 (2002).

UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **35** (Database issue), D193–D197 (2007).

UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucl. Acids Res.* **37**, D169–D174 (2009).

Wang, J., et al. The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).

Wheeler, D. L., et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35** (Database issue), D5–D12 (2007).

Wu, C. H., et al. The Protein Information Resource: An integrated public resource of functional annotation of proteins. *Nucleic Acids Res.* **30**, 35–37 (2002).