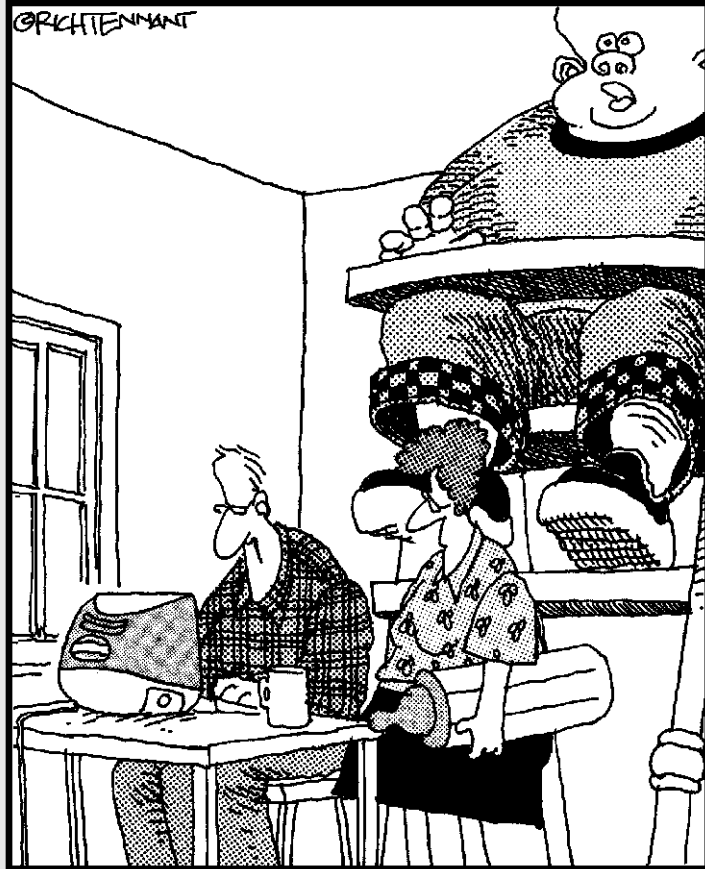


## The 5<sup>th</sup> Wave

By Rich Tennant



"Well, that proves it. He's definitely from your side of the family."

## *In this part . . .*

**B**ioinformatics is about finding and interpreting biological data online. In this part, we show you how to go shopping in the main bioinformatics databases. We've got lots of great tips for picking up the freshest, ripest data (and not just the low-hanging fruits!). Because all these databases are linked with one another, we also show you how you can travel across them and gather every piece of information you need. It's a fascinating journey across human knowledge — and a compulsory starting point for any research project.

With the right data in the fridge, it's time to start cooking. In this part, we show simple recipes for working with DNA and protein sequences in order to predict their most basic properties.

# Using Nucleotide Sequence Databases

---

## *In This Chapter*

- ▶ Nabbing a quick refresher on the structure of genes and genomes
  - ▶ Making use and sense of GenBank
  - ▶ Finding out about a specific gene
  - ▶ Working with complete microbial genomes
  - ▶ Browsing the human and other animal genomes
- 

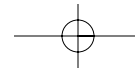
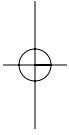
*The secret of success is to know something nobody else knows.*

— Aristotle Onassis (1906–1975)

**S**equences databases are great tools because they offer a unique window on the past. They make it possible to answer today's biological questions by enabling us to analyze sequences that may have been determined as many as 25 years ago, when the whole technology emerged. By doing this, they connect past and present molecular biology.

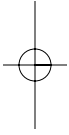
The first databases were in fact created as some sort of sequence museum, where sequences could be preserved for all eternity in pristine form, just as they were determined, interpreted, and published by their original authors. This *historical* (time capsule!) perspective pretty much remains in GenBank, the leading nucleotide sequence repository maintained as a consortium between the U.S. National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ). In this chapter, we show you how to use GenBank and decipher its entries.

Repository-type databases are great tools when you want to come up with a bibliography for a particular sequence, but they do not provide easy access to sequence data when your query deals with broader issues related to a gene or function rather than with a specific paper. For this reason, the



These days, nucleotide sequences are routinely determined at the whole-genome or chromosome scale — at least for microorganisms. We now have information not only about individual gene sequences, but also about their relative positions, strand orientation, and the presence or absence of biochemical functions within an entire organism. To take advantage of this more global information, researchers have had to design state-of-the-art *genome-centric* sequence-information management systems that can connect specialized sequence collections with browsing tools. As an added bonus, this chapter presents some of the great genome-centric resources dealing with viruses, bacteria, or (you guessed it) human beings.

However, before you start delving into these information-management systems in earnest, you may find it useful to read the next section. There, we quickly summarize the fundamentals of genes and genomes and introduce the vocabulary you need to read GenBank fluently. If you're an experienced biologist, you can skip this section completely — or read it word for word in hopes of finding some embarrassing mistake. Go ahead. We dare you! Conversely, if you want to learn much more, we advise you to go get *Genetics For Dummies*, by Tara Rodden Robinson, a great companion book for would-be bioinformaticians.

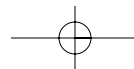


## ***Reading into Genes and Genomes***

True, nucleotide sequences are universal, but the structure of the genes they encode is markedly different between *prokaryotic* organisms (organisms lacking a true nucleus) and *eukaryotic* organisms (the kind lucky enough to have a true nucleus). You have to know the basic architecture of both types of genomes and genes to make sense of a simple GenBank entry. Remember that we need to define a *gene* as the contiguous genome segment encompassing all the nucleotide-sequence information necessary to bring about its successful *expression* — that is, the production of protein or RNA. This handy definition includes both the coding and regulatory parts of a gene.

### ***Prokaryotes: Small bugs, simple genes***

The three most basic classes of living organisms are the *prokaryotes* (usually bacteria), the *archaea* (bacteria-like organisms living in extreme conditions), and the *eukaryotes*. Eukaryotes go all the way from microscopic yeast to humans, animals, and plants.



- ✓ Their genome is a single, circular DNA molecule.
- ✓ Their genome size is in the order of a few million base pairs [0.6–8].
- ✓ Their *gene density* — the number of genes per base pairs in the genome — is approximately one gene per 1,000 base pairs.
- ✓ Their genome is lean and mean, containing few useless parts (70 percent is coding for proteins).
- ✓ Their genes do not overlap.
- ✓ Their genes are transcribed (copied into messenger RNA) right after a control region called a *promoter*.
- ✓ These messenger RNA (mRNA) are collinear with the genome sequence. In other words, genes are in a single piece, not interrupted by noncoding patches (called *introns*).
- ✓ Protein sequences are derived by translating the longest open reading frame (from ATG to STOP) spanning the gene-transcript sequence.

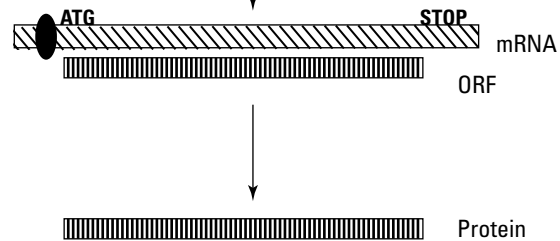
Figure 3-1 illustrates the simple collinear relationship between a bacterial genomic (gene) sequence, the transcript (mRNA), the open reading frame (ORF), and the final protein. (For a peek at a typical [small] bacterial genome, flip to Figure 1-10 in Chapter 1.) The mRNA sequence gets translated into a protein right after a special signal, called the Ribosome Binding Site (RBS in Figure 3-1). The ribosome is the main piece of machinery of the cell's protein-translation apparatus.

Accordingly, database entries describing a coding prokaryotic sequence should include three important features:

- ✓ The coordinates of some promoter elements
- ✓ The coordinates of the RBS
- ✓ The coordinates of the ORF boundaries

That's about all there is to say regarding the simple architecture of prokaryotic genes. Not all genes encode proteins; for some of them, the function is directly carried out by the transcribed RNA molecule. This includes transfer RNA (tRNA), ribosomal RNA (rRNA), and a few other fancy RNAs that shall remain nameless.

**Figure 3-1:**  
Relationship  
between  
gene,  
mRNA, and  
protein  
sequence  
for  
prokaryotes.



## *Eukaryotes: Bigger bugs, complex genes*

Eukaryotes encompass a wide variety of organisms, from microscopic yeasts and fungi to elephants, plants, and humans. Yet eukaryotes, too, have basic properties in common — properties that end up making them a bit more challenging when it comes to genomic and bioinformatic analysis:



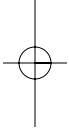
- ✓ Their genome consists of multiple linear pieces of DNA called *chromosomes* (up to a hundred million base pairs long).
- ✓ Their genome size (10–670,000 million base pairs), especially for animals, plants, and amoebae (!), is *much bigger* than in prokaryotes.
- ✓ Their gene density is much lower than that for prokaryotes (one human gene per 100,000 base pairs).
- ✓ Their genome is no model of efficiency, containing many useless parts (less than 5 percent of the human genome code for proteins).
- ✓ Genes on opposite DNA strands might overlap — although that's a relatively rare occurrence.
- ✓ Their genes are transcribed right after a control region called a *promoter*, but sequence elements located far away can have a strong influence on this process.
- ✓ Gene sequences are not collinear with the final messenger RNA (mRNA) and protein sequences. Only small bits (the *exons*) are retained in the mature mRNA that encodes the final product.
- ✓ Genes often exhibit more than one mRNA (and protein) form.

Genes of higher eukaryotes (animals) may span up to millions of base pairs — the human dystrophin gene (the mutation of which causes a dreadful disease),

fragments). Throughout the rest of this chapter, we show you several examples so that, after a bit of study, you can decipher the most intricate GenBank entries by yourself.

## *Making Use (and Sense) of GenBank*

For prokaryotes, the limited size of the genes involved — as well as the simple (linear) relationship among the gene DNA sequence, the mRNA, the ORFs, and the final protein sequence — make all that information relatively easy to annotate and store in database records. This is why database entries corresponding to bacterial genes are relatively easy to read and understand. In this section, we take you through the entry of the *Escherichia coli* dUTPase gene step by step (handy GenBank ID X01714).



### *Making sense of the GenBank entry of a prokaryotic gene*

First things first, we have to fetch our sample GenBank entry, as follows:

1. **Point your browser to** [www.ncbi.nlm.nih.gov/entrez/](http://www.ncbi.nlm.nih.gov/entrez/).

The NCBI PubMed home page appears.

2. **From the Search pull-down menu, choose Nucleotide (as in Figure 3-2).**

3. **Type the GenBank ID X01714 in the For field, and then click Go.**

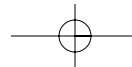
A Results page appears, displaying a short definition for the nucleotide sequence entry, preceded by its ID X01714 (underlined in blue).

4. **Click the [X01714](#) hyperlink. (Alternatively, you can change the display option from Summary to GenBank.)**

The X01714 GenBank entry appears in the default GenBank format, as shown in Figure 3-3.

Gurus refer to this as the GenBank *flat-file format* because you can read it in a linear fashion, and it doesn't involve indexes, pointers, or accessory information (which are customary in the structure of more sophisticated databases).

In fact, what you have here isn't totally "flat" because it contains a few hyperlinks.



## Choose Nucleotide.



**Figure 3-2:**  
Searching for  
nucleotide  
sequence  
X01714 at  
NCBI  
PubMed.

### *Reading the header of a GenBank prokaryotic entry*

The entry text saved in the preceding steps is divided into sections, with keywords introducing each section. Typical keywords include LOCUS, DEFINITION, ACCESSION, VERSION, KEYWORDS, and SOURCE. Refer to Figure 3-3 as we go through these keywords step by step:

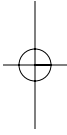
- ✓ **LOCUS** gives us the locus name (an arbitrary name; here it's ECDUT), the size of the nucleotide sequence in base pairs, the nature of the molecule (here it's DNA), and its topology (*linear* or *circular* molecule — in this particular case, linear).
- ✓ **DEFINITION** provides a short definition of the gene that corresponds to the entry sequence. Here, it's the *E. coli* dUTPase gene.
- ✓ **ACCESSION** lists the accession number — a unique identifier within and across various databases (such as the protein-sequence databases). Here, the accession number is X01714.
- ✓ **VERSION** fills you in on synonymous or past ID numbers.
- ✓ **KEYWORDS** introduces a list of terms that broadly characterize the entry. You can use these terms as keywords for certain database searches. (For more on using keywords in database searches, check out Chapter 2, where we discuss using PubMed with limits.)
- ✓ **SOURCE** divulges the common name of the relevant organism to which the sequence belongs.



to other keywords. Keywords starting at the same position are said to be at the same level. The leftmost position delimits sections of the GenBank entry that are independent, such as the preceding six sections.

In contrast, ORGANISM is indented to indicate that it's part of the SOURCE section. (The same thing happens in the REFERENCE section.)

- ✓ **REFERENCE** introduces a section where the credits for the sequence determination are given (different parts of the sequences can be credited to different authors). The REFERENCE section contains multiple parts as denoted by the indentation of the following self-explanatory keywords: AUTHORS, TITLE, JOURNAL, and PUBMED.
- ✓ **COMMENT** introduces the last line of the top of the X01714 GenBank entry. It contains free-formatted text, such as acknowledgments or info that doesn't fit in the previous sections.



Use this menu for FASTA format.

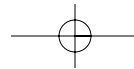
Click here for text format.

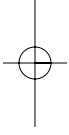
The screenshot shows the NCBI Nucleotide database interface. The main content area displays the GenBank entry for X01714.1. The entry includes the following information:

```

VERSION X01714.1 GI:41296
KEYWORDS dUTPase; unidentified reading frame.
SOURCE Escherichia coli
ORGANISM Escherichia coli
          Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
          Enterobacteriaceae; Escherichia.
REFERENCE 1. (bases 1 to 1609)
AUTHORS Lundberg,L.O., Thorsen,K.O., Karlstrom,O.H. and Nyman,P.O.
TITLE Nucleotide sequence of the structural gene for dUTPase of
      Escherichia coli K-12
JOURNAL ENDO J. 2 (6), 967-971 (1985)
PUBMED 639980
COMMENT Data kindly reviewed (25-NOV-1985) by L. Lundberg.
FEATURES
     source              Location/Qualifiers
                        1..1609
                        /organism="Escherichia coli"
                        /mol_type="genomic DNA"
                        /db_xref="taxon:562"
     promoter            286..291
                        /note="--35 region"
     promoter            310..316
                        /note="--10 region"
     misc_feature        322..324
                        /note="put. transcription start region"
     CDS                 390..393
  
```

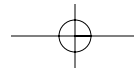
**Figure 3-3:**  
GenBank  
entry  
X01714.

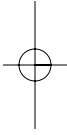




- ✓ **source** indicates the origin of specific regions of the sequence. This is useful when you want to distinguish cloning vectors from host sequences. In X01714, the whole sequence comes from *E. coli* genomic DNA.
- ✓ **promoter** shows the precise coordinates of a promoter element. In X01714, a *-35 region* is indicated from position 286 to 291 (indicated by *286..291*) in the nucleotide sequence.  
**promoter** introduces another line that contains the coordinates of another promoter element, this time a *-10 region* at positions 310 to 316.
- ✓ **misc feature** (miscellaneous feature) indicates the putative location of the transcription start (mRNA synthesis). For X01714, this is from positions 322 to 324.
- ✓ **RBS** (Ribosome Binding Site) indicates the location of the last upstream element. For X01714, this is at position 330 to 333.
- ✓ **CDS** (CoDing Segment) introduces a complex section that describes the gene's *open reading frame (ORF)*:
  - The first line indicates the coordinates of the ORF from its initial **ATG** to the last nucleotide of the first stop codon **TAA**. (For X01714, it's 343 to 798.) See Chapter 1 if you don't remember what a codon is.
  - Each of the following lines (indented at the same level) gives the name of a protein product, indicates the reading frame to use (here, 343 is the first base of the first codon), the genetic code to apply, and a number of IDs for the protein sequence.
  - */translation*, the final keyword of the CDS section, introduces the conceptual amino-acid sequence of the coding segment. This sequence is a computer translation that uses the coordinates, reading frame, and genetic code indicated in the preceding lines.
- ✓ Another **misc feature** follows the CDS section. It contains lines that point out putative stem-loop structures and repeats. These are potential regulatory elements of the dUTPase gene.

The FEATURES section of entry X01714 is typical of a simple, well-annotated bacterial nucleotide sequence, centered on a well-identified gene. However, this straightforward entry includes a complication: a putative additional reading frame. We selected this entry because we were interested in one gene: the dUTPase. However, the entry exhibits an extra putative gene, indicated by an additional RBS element and a second CDS section.





**Figure 3-4:**  
GenBank  
entry  
X01714: the  
FEATURES  
table part.

```
...
285..291
/note="-35 region"
310..316
/note="-10 region"
322..324
/note="put. transcription start region"
330..333
/note="put. rRNA binding site"
343..358
/note="GUTP-ase (aa 1-151)"
/codon_start=1
/transl_table=
/protein_id=" "
/db_xref="GI:41297"
/db_xref="SWISS-PROT:P06968"
/translation="EKKIDVKKILPFGVKEFPLPTIATSGSAGLDRACLMDAVELAP
GDTLVFTGLAIHIADPFLAANLPPGGLGHEHIVLGNLYGLIDSDYQQLMISVWV
PGQDSFTICPGERIAQHFVFPVVAEFNLVEDFDATRGEQGFHSGRQ"
831..851
/note="put.stem-loop structure"
831..838
/note="inverted repeat A"
844..851
/note="inverted repeat A'"
856..893
/note="put. stem-loop structure"
866..872
/note="imp. inverted repeat B"
888..893
/note="imp. inverted repeat B'"
899..899
/note="put. rRNA binding site"
905..1540
/note="unidentified reading frame"
/codon_start=1
```



GenBank entries containing more than one gene are frequent.

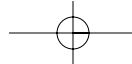
### *Using the Sequence section of a prokaryotic entry*

The last section of GenBank entry X01714 is the nucleotide sequence section. It starts with the *ORIGIN* keyword and finishes with the end-of-entry line introduced by two slash marks (/ /). Each line of nucleotide sequence starts with the position number of the first nucleotide in that line. Each line contains 60 nucleotides.



Because the sequence section of a GenBank entry mixes numbers and nucleotides, you *cannot* directly use it as an input on most sequence analysis servers. You first need to generate a FASTA-formatted sequence, as follows:

- 1. On your GenBank entry page, choose FASTA from the Display pull-down menu. (Refer to Figure 3-3.)**  
You'll find the Display pull-down menu at the top-left of the page.
- 2. Select the Text option from the Send To pull-down menu.**  
The nucleotide sequence appears now in a form suitable for most sequence-analysis programs.
- 3. Save this entry by choosing File⇨Save As from your browser's main menu or by copying/pasting it into an empty Word document.**



present in both prokaryotes and eukaryotes: Humans have it, too. Looking at the GenBank entries describing the human version of dUTPase clearly illustrates the increased complexity of higher eukaryotes compared to prokaryote genes. We start here with a simple one, GenBank entry U90223:

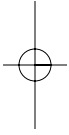
1. Point your browser to [www.ncbi.nlm.nih.gov/entrez/](http://www.ncbi.nlm.nih.gov/entrez/).

The NCBI PubMed home page appears on cue.

2. From the Search pull-down menu, choose Nucleotide.

3. Type in GenBank ID U90223 in the For window, and click Go.

A Results page appears, displaying a short definition for the nucleotide sequence entry, preceded by its ID U90223 (underlined in blue) as shown in Figure 3-5.



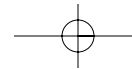
Jump to the corresponding gene.

Get related sequences.

**Figure 3-5:**  
Initial display for GenBank entry U90223.

4. Click the [U90223](#) hyperlink. (Alternatively, you can change the Display option from Summary to GenBank.)

You're now ready to read your first human GenBank entry.



very different from entry X01714, the one that describes its bacterial homologue. The top part of the entry follows the general information keywords order: LOCUS, ACCESSION, DEFINITION, and VERSION.



The KEYWORD line, which is supposed to list readily relevant and searchable terms (such as dUTPase), is empty for entry U90223. Unfortunately, this isn't a fluke. It illustrates a common problem in sequence databases — annotations may be incomplete. As a result, keyword-based database searches usually don't retrieve all the relevant sequences in GenBank. On a related note, the information slated for the SOURCE and REFERENCE sections may also be missing or incomplete. The U90223 entry, for example, says that the sequence here comes from an unpublished source. However, a simple PubMed search (see Chapter 2) using the author's names — Ladner RD and Caradonna SJ — brings us the relevant article in the *Journal of Biological Chemistry* — from 1997! A word to the wise: You should never expect GenBank (or any sequence database) annotations to be up to date; no wonder it's sometimes impossible to retrieve some sequences with PubMed searches.

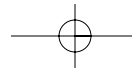
Figure 3-6 displays the FEATURES section of the U90223 entry. The CDS keyword indicates a coding region (63-821) sequence that corresponds to the mitochondrial form of human dUTPase. Following the conceptual amino-acid translation of the ORF, the *sig peptide* keyword indicates the location of a mitochondrial targeting sequence, and the *mat peptide* keyword provides the exact boundaries of the mature peptide.

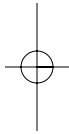
This human GenBank entry isn't really more complex than its bacterial homologue because we carefully selected an entry describing an *mRNA* sequence, not a *genomic* one. At the level of the mature mRNA, the relationship between a eukaryotic protein and its encoding nucleotide sequences is as collinear as its prokaryotic counterpart. As the next section shows, genuine genomic data can be a bit trickier.

## ***Making sense of a GenBank eukaryotic genomic entry***

The previous section described the GenBank entry corresponding to the mRNA sequence of the human dUTPase gene. In this case, we want to look at the GenBank entry AF018430 related to the *gene sequence* (as it is on the chromosome) from which this mRNA originated.

@Spy





**Figure 3-6:**  
FEATURES  
section of  
GenBank  
entry  
U90223.

```

/feature_id="ORIGIN"
/feature_start="1"
/feature_end="901"
/feature_type="text"
ORIGIN
1 ggtggaagcc tgggcacgt cggaggtgc cgaggacca accagccaa actctgggg
61 aatgactcc cctctgccc cgcgccggc tctgtacca ttctctaac tctctgttc
121 gctcaggat gcaaacccg cgaggcagg cagaggccg aagccgggt actctccgg
181 ccaggccgc cctcggccg cggcggggc agcacggat tcccggccg ctgtccacg
241 ctggccgct gagccaagg tgcggcgag ccagcacagt cggggccgt ggtggaag
301 gggagcttc taaggcggg ggaagccgg cgcggggcc ggagacacc gccattcac
361 ccagtaagc ggcggccct cggaggtgg cggcgtgca gctccgctt gccggctct
421 ccagcacgc cagccccc acccgggct cggcggcgc cggggctac gacctgaca
481 gtccctatg ttacacaata ccacctatg agaaagctg tgtgaaacg gacattcaga
541 tagcctccc ttctgggtg tatggaagag tggctccac gtcaggctg gctgcaaac
601 actttatga ttaggagct ggtgtcatg atgagatta tagagaaat gttgtgttg
661 tactgttaa tttggcaaa gaaaagtgg aagtcacaaa aggtgatga attgcacgc
721 tcaattgga accgattttt tatccagaaa tagaagaat tcaagcctg gatgacacc
781 aaaggggtc aggggtttt ggttccactg gaaagaata aaattatgc caagacaga
841 aacaagaag tcatacctt ttcttaaaa aaaaaaagt tttgttca agtgtttgg
901 tgtttgac ttctgtaac ttaactgct tacctctaa aagtactga tttttactt
//

```

mRNA (nucleotide) sequence

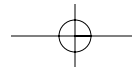
First, however, we have to fetch the entry by using the same protocol as before:

1. **Point your browser to** [www.ncbi.nlm.nih.gov/entrez/](http://www.ncbi.nlm.nih.gov/entrez/).
- The NCBI PubMed home page appears.
2. **From the Search pull-down menu, choose Nucleotide.**
3. **Type in AF018430 in the For window, and click Go.**
4. **Click the [AF018430](#) link.**

The AF018430 entry appears, as shown in Figure 3-7.

You're now ready to read your first human genomic GenBank entry. We've selected a simple one for you to start with, but it contains some of the specificities only encountered in eukaryotic entries.

@Spy



Features		Sequence	
LOCUS	HSDUT2	1177 bp	DNA linear PRI 26-SEP-1997
DEFINITION	Homo sapiens dUTPase (DUT) gene, exon 3.		
ACCESSION	AF018430		
VERSION	AF018430.1 GI:3441576		
KEYWORDS	.		
SEGMENT	2 of 4		
SOURCE	Homo sapiens (human)		
ORGANISM	Homo sapiens		
	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;		
	Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;		
	Catarrhini; Hominoidea; Homo.		
REFERENCE	1 [bases 1 to 1177]		
AUTHORS	Pearlman,R.E.		
TITLE	Human genomic nucleic acid and mitochondria dUTPase gene		
JOURNAL	Unpublished		
REFERENCE	2 [bases 1 to 1177]		
AUTHORS	Pearlman,R.E.		
TITLE	Direct Submission		
JOURNAL	Submitted (11-ADG-1997) Biology, York University, 4700 Keele St., North York, ONT M3J 1P3, Canada		
FEATURES	Location/Qualifiers		

**Figure 3-7:**  
GenBank  
entry  
AF018430.

The top part contains general information introduced by the usual keywords:

- ✓ **LOCUS:** The locus name is HSDUT2. The rest of the line tells us that we're dealing with 1177 bp of linear DNA.
- ✓ **DEFINITION:** This line indicates that the name of the gene is *DUT* and specifies that the entry encompasses *exon 3* of the gene. This reminds us that human genes generally spread their protein-coding regions across many disjointed pieces, the exons.
- ✓ The **ACCESSION**, **VERSION**, and **KEYWORDS** lines are standard. Note that for AF018430, the latter is empty, meaning you can't retrieve this entry via a keyword-field-restricted search.
- ✓ **SEGMENT:** This field relates to the mosaic structure of eukaryotic genes. It indicates that this current GenBank entry is the second segment of a super entry made of four. You need all four entries to reconstruct the complete mRNA sequence used as a template for producing the protein.
- ✓ The **ORGANISM**, **SOURCE**, and **REFERENCE** lines are standard.



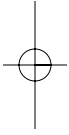
The FEATURES table is what really makes a eukaryotic genomic entry special — and, as such, is much longer than the ones for prokaryotic organisms. It contains the following elements (see Figure 3-8):

- ✓ The **source** section contains a */map* section. For AF018430, it indicates that the sequence belongs to chromosome 15, and was more precisely mapped on the long arm (q) of this chromosome, within the q21.1 cytogenetic band.

@Spy

This is nothing more than an *exon splicing* recipe telling us how to reconstruct the mRNA sequence. All it says is:

1. Take nucleotides from positions 1 to 1735 from entry AF018429.
2. Add nucleotides from positions 1 to 1177 from the current entry (AF018430).
3. Add nucleotides 1 to 45 from entry AF018431.
4. Add nucleotides 658 to 732 from entry AF018432.



**Figure 3-8:**  
GenBank  
entry  
AF018430:  
FEATURES  
table for a  
eukaryotic  
gene.

```
FEATURES             Location/Qualifiers
     source            1..1177
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /map="15q15-q21.1"
     order(AFO18429.1:<11735,1..1177,AF018431.1:1..45,
AF018432.1:658..732,AF018432.1:884..954,
AF018432.1:1391..>1447)
     /gene="DUT"
     join(AFO18429.1:<282..561,AF018429.1:1034..1172,
560..651,AF018431.1:1..45,AF018432.1:658..732,
AF018432.1:884..954,AF018432.1:1391..>1447)
     /product="dUTPase"
     /note="alternatively spliced; encodes mitochondrial form
of the protein"
     join(AFO18429.1:282..561,AF018429.1:1034..1172,560..651,
AF018431.1:1..45,AF018432.1:658..732,AF018432.1:884..954,
AF018432.1:1391..1447)
     /gene="DUT"
     /note="DUT-R; alternatively spliced; mitochondrial form of
the protein; similar to H. sapiens dUTPase encoded by
GenBank accession Number U90224"
     /codon_start=1
     /product="dUTPase"
     /protein_id="      "
     /db_xref="GI:2443580"
     /translation="MTPLCPFPALCYHFLTSLLESANQNARGTAEGRSRGTLRARPAP
RPPAACHGIPPLSSAGRLBQCCRGASTVGAAGKGLFPRAGGSPAPGHPETPAISPSK
RARAEVGKRLRFARLSEHATPTSCSARAAGTDLTSATDTLPPHEKAVETDIOI
ALPSCYGRVAFPSGLAAKHFIDVCAQVIDEYRQVGVVLFNFGREFFVKKGDRIA
QLICKRIFYPEIEYVALDDETERSGGGGSGTGM"
     join(AFO18429.1:<1018..1172,560..651,AF018431.1:1..45,
AF018432.1:658..732,AF018432.1:884..954,
AF018432.1:1391..>1447)
     /gene="DUT"
     /product="dUTPase"
     /note="alternatively spliced; encodes nuclear form of the
protein"
     join(AFO18429.1:1018..1172,560..651,AF018431.1:1..45,
AF018432.1:658..732,AF018432.1:884..954,
AF018432.1:1391..1447)
     /gene="DUT"
```

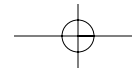
Protein form 1

mRNA form 1

mRNA form 2

Protein sequence form 1

@Spy





might actually continue beyond the indicated position.

- ✓ **The mRNA:** The AF018430 entry has two mRNA fields, which are interesting because they illustrate the way GenBank represents alternative splicing patterns. Alternative splicing is a common property of higher eukaryotic gene expression. If you use the same parsing logic as gene order (which we describe in the preceding bullet), you can produce the results we summarize in Table 3-1.

**Table 3-1 Mitochondrial or Nuclear dUTPase mRNAs in AF018430**

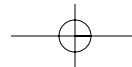
<i>Mrna</i>	<i>AF018429</i>	<i>AF018430</i>	<i>AF018431</i>	<i>AF018432</i>
Type 1 (Mitochondria)	282-561 1034-1172	560-651	1-45	658-732 884-954 1391-1447 ->
Type 2 (Nuclear)	<1018- 1172	560-651	1-45	658-732 884-954 1391-1447 ->

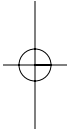
Table 3-1 makes it easy to understand what's going on here:

- There are two alternative mRNA: type 1 (for the mitochondria) and type 2 (for the nucleus).
  - The nuclear mRNA does not contain the first exon (stored in entry AF018429) although the mitochondrial mRNA uses this exon.
  - The nuclear mRNA uses a part of the second exon (stored in the AF018429) in a different reading frame from its mitochondrial counterpart.
  - The two protein sequences become identical on the third exon. You can see this in the two */translation* fields (*ETPAISPSK* and so on). See Figure 3-9 for the sequence of the nuclear form of the protein.
- ✓ The **exon** keyword indicates the position of the sole exon present in this sequence. This is near the end of the FEATURES part of the AF018430 entry (refer to Figure 3-9, positions 560–651). Look at AF018432 if you want to see multiple occurrences of the exon field in a single entry.

The sequence section, located in the bottom part of the entry, is formatted as usual. (Refer again to Figure 3-9.)

@Spy





**Figure 3-9:**  
GenBank  
entry  
AF018430:  
Bottom and  
sequence  
part.

```

/product="dUTPase"
/protein_id=" "
/db_xref="GI:2443581"
/translation="MPCSEETPAISPSKRARPARVGGNLRPARLSEHATPTGSRAR
AAGYDLYSADYTIIPREKAVVKTDLQIALPSGCGYRVAPRSGLAARHFIDVAGVID
EDYRGNVGVVLFNFGREKFEVKKGRDIAQLICERIFYPIREEVQALDDTERSGGFGS
TGRN"
550..651
/genes="dUT"
/number=3
ORIGIN
1  tccttaaac  aacacagatc  atgtggaaga  ataaaatggg  gttaatatat  gtaaaaccaa
61  ttaggaaact  gttctgggg  caaacacagta  aagggcttat  tcaatggata  ggotagtatt
121  attagttagt  aatggggccc  tttttttctt  tctttctttt  ctccattttt  ttccttttca
181  aactatgggt  tgraaagatc  ccaccttttg  aaagtgtgac  tttctggcct  ttaacgttga
241  taagtaactc  agttctaac  aaacttttgg  tcaagggaca  acattttaca  tgttgactc
301  tcttaacacc  accaaaata  tccatggaga  attattttat  ctaaagctgt  ctttttata
361  ataaatagc  cactctaac  tttctcaaa  actttaaga  tgaattgga  attacata
421  gcaagttga  ttttagaac  taagtgtca  ttaattcatt  aaatcacctg  aaagtattt
481  tgtatgcttg  gtcacaaaga  aaatataaaa  acaattttat  aaatagatt  gcagttattt
541  tttttcaata  ttttttagt  gctatgatt  acacaatacc  acctatggag  aaagctgtg
601  tgaasaacga  caticagata  ggcctcctt  ctgggtgta  tggaaagatg  gtaagctat
661  ttaagaacaa  gtaactatt  tgcacagtc  tcccttgca  tagattcttc  agtttcatt
721  tgggtata  agaggcac  atctgttgg  ctgtgcta  aaagaagac  catttgcat
781  agcaaatgca  ctcttgaaa  gctttactt  acactctgc  ttgctctt  ttgacctt
841  ttattttct  cctctctac  tggagcttt  agctcacac  tggcttaca  ggotctctc
901  agaacatgc  attttat  atgagatga  aactctgac  ctgttgctc  cagaatggt
961  aagctactt  aactttttt  tgtttgcca  tgggtttag  gtaagggat  actttcagt
1021  gttttagag  gcactggag  gaagttaga  caaatggag  ttacactca  acagttga
1081  tttttctg  aagcaattc  agtttacc  agacattcc  ttgcagagc  gttagtctt
1141  tttgactac  ctcaagta  acttaaggag  gaatgga

```

Genomic sequence



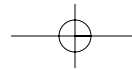
Understanding how to splice back the various nucleotide sequence fragments to form an mRNA and the associated coding regions from segmented GenBank entries was the main difficulty of this chapter. When you've done that, congratulations — you can sit back and relax while we continue our tour of GenBank.

## *Working with related GenBank entries*

In every example that we've used so far, we knew the accession numbers for the entries we wanted to look up — X01714, U90223, or AF018430, for example. Normally you get these accession numbers by reading articles that explicitly mention them when reporting about the corresponding sequence.

That's a good strategy to start off with, but after you've accessed the first GenBank entry relevant to your work, you can retrieve other related genes by selecting the Related Sequences option in the pull-down menu that appears when clicking the [Links](#) link on the right side of the screen for each GenBank entry retrieved by its accession number (refer back to Figure 3-5). If you click this link for entry U90223, for example, it returns 37 human dUTPase-related GenBank entries. They include various mRNA forms and partial sequences,

@Spy



# Retrieving GenBank entries without accession numbers

Although GenBank isn't the best database for keyword-based searches (see "Using a Gene-Centric Database," a bit later in this chapter, for a better method), querying GenBank by using gene or protein keywords rather than accession numbers is still possible. Suppose that you want to find the nucleotide sequence encoding the human dUTPase, but you don't have any handy accession numbers lying around. The following shows you how you should proceed, step by step:

1. **Point your browser to** [www.ncbi.nlm.nih.gov/entrez/](http://www.ncbi.nlm.nih.gov/entrez/).

The NCBI PubMed home page appears.

2. **From the Search pull-down menu, choose Nucleotide.**

3. **Type human [organism] AND dUTPase [Protein name] in the Search window, and then click Go.**

This search retrieves five GenBank entries: AH005568, AF018429, AF018430, AF018431, AF018432. They encompass exons 1 and 2 (AH005568), exon 3 (AF018430), exon 4 (AF018431), and exons 5, 6, and 7 (AF018432). The four last entries indicate the full amino-acid sequence of the two forms (nuclear and mitochondrial) of the dUTPase protein, as well as the alternative exon usage pattern. Not a bad start!

4. **Click the [Links](#) link to the right of the AF018432 entry ID line, and then choose Related Sequences from the pull-down menu that appears.**

This retrieves a total of 20 entries. Among these entries, some contain mRNA sequences such as U90223.

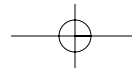
**Note:** We do not yet have all the entries related to human dUTPase in GenBank. By looking at some of the new entries we just pulled out, we realize that *dUTPase* is a nickname for "dUTP pyrophosphatase". Let's try to use these new terms in our next search:

5. **Type human [organism] AND "dUTP pyrophosphatase" [Title] (the "quotes" are important here) in the Search window, and then click the Go button.**

Notice that we restricted the search to the [Title] field here — not as a protein name, which is what we did in Step 3.



@Spy





To remove these ESTs from the list, you can use the Search-within-Limits trick we used with PubMed in Chapter 2:

1. **Open the Limits setting menu by clicking the Limits link (below the Search window).**
2. **Select the Exclude ESTs check box.**
3. **Scroll to the top of the form, and click the Go button.**

Only eleven GenBank entries survive this particular limit-setting.

Feel free to use this handy Search-within-Limits protocol to try other field-restricted searches in GenBank.

## Using a Gene-Centric Database

Besides the traditional GenBank, NCBI recently developed other databases (or database interfaces) that are more adapted to gene-centric queries.

Making a gene-centered query involves asking a question that relates directly to a specific gene, rather than going through all known pieces of sequences related to that gene. The main advantage of gene-centric databases is that they return results that are more synthetic than a long list of GenBank entries, and make much more sense to the biologists. Basically, you get the whole story at once.

In this section, we take you through the Entrez/Gene resource available on the NCBI server. This resource makes it possible to gather important information related to a genetic *locus*, a specific place on a chromosome where a given gene has been identified. Thanks to this service, you can rapidly find out everything that is known about your favorite gene, or its genomic surrounding.

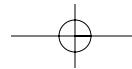
1. **Point your browser to [www.ncbi.nlm.nih.gov/entrez/](http://www.ncbi.nlm.nih.gov/entrez/).**

The NCBI/Entrez home page appears, eager to serve.

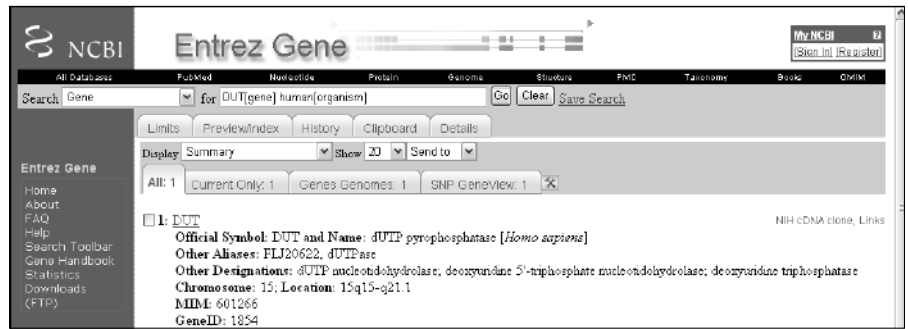
2. **From the Search pull-down menu, choose Gene.**
3. **Type DUT [gene] human[organism] in the For window, and then click Go.**

You get a screen that looks like Figure 3-10.

@Spy



**Figure 3-10:**  
Entrez Gene  
entry page  
for human  
gene DUT.



The Results page you see in Figure 3-10 is your doorway to a wealth of information. By clicking the DUT link — or by changing the display option into Full Report — you can now get to a large body of information concerning this particular gene and its genomic environment.

The top of the DUT entry (see Figure 3-11) provides a general description of what this gene is all about — and what function its products are known to perform, as well as a large variety of links (right-side menu) to other databases or NCBI files.

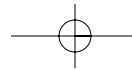
Figure 3-12 displays the next part of the entry: a schematic view of the Human DUT gene structure, with its seven exons used differentially and spread over 11,909 base pairs of genomic DNA on Chromosome 15.

The long entry then continues — additional sections provide information on potential interactions with other gene products, homologous sequences in other organisms, protein functions, and relevant metabolic pathways — as well as a list of all corresponding sequence entries in GenBank.

Each section, in turn, provides multiple links to help you get a complete (and sometimes redundant) picture of what's known about your gene. This one-stop shopping capacity illustrates the useful concept of a gene-centric database.

What you have here are all the types of mRNA sequences that have been observed and recorded in GenBank for this gene. You can see that variations mainly involve the two first exons. These variants (alternative transcripts) include the mitochondrial and nuclear forms of dUTPase (Table 3-1).

@Spy



Entrez Gene Home

Table of Contents

Summary

Official Symbol: **DUT** and Name: **dUTP pyrophosphatase** provided by HUGO Gene Nomenclature Committee

See related: HPRD:05165, MIM:601266

Gene type: protein coding

Gene name: DUT

Gene description: dUTP pyrophosphatase

RefSeq status: Reviewed

Organism: *Homo sapiens*

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhina; Catarrhini; Hominoidea; Homo

Gene aliases: dUTPase, FLJ20622

Summary: This gene encodes an essential enzyme of nucleotide metabolism. The encoded protein forms a ubiquitous, homotrimeric enzyme that hydrolyzes dUTP to dUMP and pyrophosphate. This reaction serves two cellular purposes: providing a precursor (dUMP) for the synthesis of thymine nucleotides needed for DNA replication, and limiting intracellular pools of dUTP. Elevated levels of dUTP lead to increased incorporation of uracil into DNA, which induces extensive excision repair mediated by uracil glycosylase. This repair process, resulting in the removal and reincorporation of dUTP, is self-defeating and leads to DNA fragmentation and cell death. Alternative splicing of this gene leads to different isoforms that localize to either the mitochondrion or nucleus. A related pseudogene is located on chromosome 19.

**Figure 3-11:**  
Top of the  
entry for  
human gene  
DUT.

Click here for a detailed map and experimental evidence.

Genomic regions, transcripts, and products

RefSeq below

NC\_000015.8

Genomic context

See DUT in MapViewer

chromosome 15, Location: 15q15-q21.1

44201441 | 44725570

SLC12A8 | DUT | PDK1 | LOC45161

**Figure 3-12:**  
Genomic  
context and  
detailed  
alternative  
mRNAs for  
the Human  
DUT gene.

Click the [See DUT in MapViewer](#) link to go to a detailed display of the gene structure. You can review for yourself the experimental arguments in favor of the various mRNA models presented, right down to the nucleotide level.

## Working with Whole-Genome Databases

The most recent genome-centric databases are the modern bioinformatic response to the proliferation of complete genome sequencing projects. The goal of these new types of resources is to gather all the information you need on a given organism, clearly separated from all the others, so you can more easily target your analyses on all genes from a specific genome. These new resources also promote comparisons of whole genomes with whole genomes, a new field of endeavor called *comparative genomics*.

@Spy

Genome-centric databases also provide the easiest way to gather every gene/protein sequence of a given organism at once, with only a few mouse clicks.

We start our exploration of whole genome databases by taking a peek at the Viral Genome section available on the NCBI server.

## *Working with complete viral genomes*

Viruses are fascinating objects, on the edge of the living world. They function as minimal molecular bits of machinery, cleverly designed to ensure the multiplication of nucleic-acid molecules (the *viral genome*) at the expense of cellular hosts (eukaryotic, bacterial, or archaeobacterial). While going about their business, viruses might go unnoticed — or trigger dreadful diseases and epidemics, such as smallpox, poliomyelitis, or AIDS. Viral genomes come in all varieties of biochemical forms (RNA/DNA, circular/linear, single/double stranded) and size (a few kb to a million bp).

Visiting the NCBI Viral Genome resource is a great way to find out more about them. To demonstrate, we show you what you can find out about the dreadful AIDS-causing, type-1 human immunodeficiency virus, or HIV-1.

**1. Point your browser to** [www.ncbi.nlm.nih.gov/entrez/](http://www.ncbi.nlm.nih.gov/entrez/).

You've probably memorized this address by now. The NCBI PubMed home page appears.

**2. On the black menu bar at the top of the form, click Genome.**

This takes you to the Entrez Genome page.

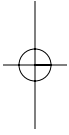
**3. Click the Viruses link (on the right side of the form).**

The Viral Genomes reference page appears.

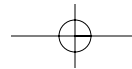
**4. Scroll down the Viral Reference Genomes page until you reach the table of available viral-genome sequences grouped by class (Deltavirus, Retroid viruses, and so on), as shown in Figure 3-13.**

**5. Type HIV1 in the Search window, and then click the Find button.**

Your browser returns a nice global summary of the HIV-1 genome, as shown in Figure 3-14. At the bottom, a clickable picture indicates the identity and respective positions of all the genes.



@Spy



**Figure 3-13:**  
The table of  
all available  
viral  
genome  
sequences.

Entrez Genomes currently contains 2425 Reference Sequences for 1658 viral genomes and 35 Reference Sequences for viroids.

<a href="#">1</a> <a href="#">Deltavirus</a>	<a href="#">91</a> <a href="#">Retro-transcribing viruses</a>	<a href="#">94</a> <a href="#">Satellites</a>
<a href="#">458</a> <a href="#">dsDNA viruses, no RNA stage</a>	<a href="#">95</a> <a href="#">dsRNA viruses</a>	<a href="#">295</a> <a href="#">ssDNA viruses</a>
<a href="#">103</a> <a href="#">ssRNA negative-strand viruses</a>	<a href="#">478</a> <a href="#">ssRNA positive-strand viruses, no DNA stage</a>	<a href="#">36</a> <a href="#">unclassified bacteriophages</a>
<a href="#">7</a> <a href="#">unclassified viruses</a>		

To obtain a list of viral genomes belonging to a particular group, family or floating genus, enter the taxonomy node name in the search textbox. Then click "Find".

HIV1

Enter the name of your virus of interest.

Click here for all proteins.

**Genome > Viruses > Human immunodeficiency virus 1, complete genome**

Lineage: [Viruses](#) : [Retro-transcribing viruses](#) : [Retroviridae](#) : [Orthoretrovirinae](#) : [Lentivirus](#) : [Primate lentivirus group](#) : [Human immunodeficiency virus 1](#)

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: <a href="#">NC_01802</a>	Genes: <a href="#">9</a>	<a href="#">COG</a>	<a href="#">Genome Project</a>	Publications: <a href="#">11</a>
GenBank: <a href="#">AF133318</a>	Protein coding: <a href="#">9</a>	<a href="#">3D Structure</a>	<a href="#">Refseq FTP</a>	Refseq Status: <b>Reviewed</b>
Length: <b>9,181 nt</b>	Structural RNAs: <b>None</b>	<a href="#">TaxMap</a>	<a href="#">GenBank FTP</a>	Seq Status: <b>Completed</b>
GC Content: <b>42%</b>	Pseudo genes: <b>None</b>	<a href="#">TaxPlot</a>	<a href="#">BLAST</a>	Sequencing center: <b>NLM, NIH, USA, Bethesda</b>
% Coding: <b>93%</b>	Others: <b>7</b>	<a href="#">GenePlot</a>	<a href="#">TraceAssembly</a>	Completed: <b>1998-01-22</b>
Topology: <b>linear</b>	Contigs: <b>1</b>	<a href="#">gMap</a>	<a href="#">COG</a>	<a href="#">Organism Group</a>
Molecule: <b>ssRNA</b>			Other genomes for species: <a href="#">Sg2</a>	

Gene Classification based on COG functional categories

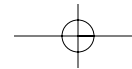
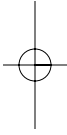
Search gene, GeneID or locus\_tag:

Click here for Sequence Viewer presentation, Click sequence and

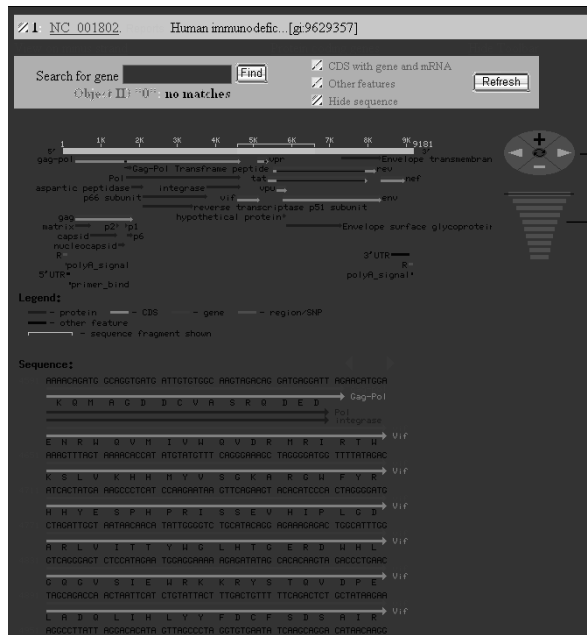
Protein: [Nef](#) [Human immunodeficiency virus 1]  
Gene: [nef](#)  
Locus: [tag:HV1tag9](#)  
Location: [\(8343-8963\)](#)  
Click to open

Click here for a live map.

Click here to get gene details.







**Figure 3-15:**  
Zooming in on the HIV-1 genome: the Pol/Vif overlap.

Figure 3-15 shows a position in the genome where the end of the Pol gene slightly overlaps with the beginning of the Vif gene. Viruses commonly have the same nucleotide sequence involved in the making of two different amino-acid sequences.

**6. Click your browser's Back button.**

This takes you back to the HIV-1 Genome entry page. (Refer to Figure 3-14.)

**7. Click the number (9) following Protein coding (second column and row of the table).**

The Protein List page appears, as shown in Figure 3-16. Here, you can retrieve the DNA and protein sequences in different formats — GenBank format, or FASTA format — by clicking one of the three lozenges in the rightmost column.

**Figure 3-16:**  
Download-  
ing HIV-1  
gene and  
protein  
sequences.

Click on a bar to select length range.

9 proteins(s) shown

DNA region in flatfile format
  DNA region in FASTA format
  Protein in FASTA format

Product Name	Start	End	Strand	Length	Gi	GeneID	Locus	Locus_tag	COG(s)	Links
Pr55(Gag)	336	1838	+	500	9629360	155030	gag	HIV1_gp2	-	◇◇◇
Gag-Pol	336	4642	+	1435	28872618	155348	gag-pol	HIV1_gp1	-	◇◇◇
Vif	4587	5165	+	192	9629361	155459	vif	HIV1_gp3	-	◇◇◇
Vpr	6105	6208	+	96	28872612	155302	vpr	HIV1_gp4	-	◇◇◇
Tat	5377	7970	+	86	9629358	155671	tat	HIV1_gp5	-	◇◇◇
Rev	5616	8193	+	116	9629359	155503	rev	HIV1_gp6	-	◇◇◇
Vif	5908	5956	+	42	9629366	155945	vif	HIV1_gp7	-	◇◇◇
Envelope surface glycoprotein gp180, precursor	6771	8341	+	856	9629363	155671	env	HIV1_gp8	-	◇◇◇
Nef	8343	8963	+	206	28872618	156110	nef	HIV1_gp9	-	◇◇◇

Display  Show  Send to

Select Protein FASTA here to download all sequences.

## Working with complete bacterial genomes

NCBI Entrez offers a nice interface to all publicly available complete bacterial genome sequences. For a quick tour, do the following:

- 1. Point your browser to [www.ncbi.nlm.nih.gov/entrez/](http://www.ncbi.nlm.nih.gov/entrez/).**  
The NCBI PubMed home page dutifully appears.
- 2. Click Genome on the black menu bar near the top of the form.**  
This takes you to the Entrez Genome page.
- 3. In the left dark-blue margin, click the Chromosome link located under the Bacteria heading.**

This step returns a listing of all bacteria whose chromosomes have been fully sequenced. (Directly clicking Bacteria would have given you a longer list, including parasitic DNA segment called plasmids.) Figure 3-17 shows the top of the list.

Bacteria Complete Chromosome		Accession / List		357
Acidobacterium bacterium Fliu341		NC_008008	563038 bp	May 4 2006
Acinetobacter sp. ACP1		NC_005986	2398621 bp	Jul 9 2004
Agrobacterium tumefaciens str. C58	circular	NC_003067	2841381 bp	Oct 3 2001
Agrobacterium tumefaciens str. C58	linear	NC_003063	2074782 bp	Oct 3 2001
Agrobacterium tumefaciens str. C58	circular	NC_003304	2841480 bp	Dec 14 2001
Agrobacterium tumefaciens str. C58	linear	NC_003305	2075560 bp	Dec 14 2001
Anabaena variabilis ATCC 29412		NC_007412	6363727 bp	Sep 20 2005
Anaerotruncobacter delhalogenans ZCP-2		NC_007760	3013479 bp	Jan 27 2006
Apicomplexa marginalis str. St. Maria		NC_004841	1159687 bp	Dec 8 2004
Asiaticum phagocytophilum H2		NC_007792	1471282 bp	Feb 21 2006
Aquifex anolicus VFS		NC_000918	1551335 bp	Sep 7 2001
Atletia yellow enticed-broom phytoplasm AYWBE		NC_007716	706569 bp	Jan 18 2006
Azarcus sp. EB01		NC_005512	426223 bp	Dec 9 2004
Bacillus anthracis str. Ames Ancestor		NC_007530	5227419 bp	May 20 2004
Bacillus anthracis str. Ames		NC_003927	5227292 bp	Apr 30 2003
Bacillus anthracis str. Steinhilber		NC_003945	5228663 bp	Jun 24 2004
Bacillus cereus ATCC 10827		NC_003000	5224283 bp	Feb 24 2004
Bacillus cereus ATCC 14929		NC_004722	5411839 bp	Apr 17 2003
Bacillus cereus F331		NC_006274	5300915 bp	Sep 16 2004
Bacillus thuringiensis KSM-K16		NC_006582	4903871 bp	Jan 3 2005
Bacillus thuringiensis C-125		NC_002570	4202352 bp	Sep 10 2001
Bacillus thuringiensis ATCC 14580		NC_003270	4222334 bp	Sep 15 2004
Bacillus thuringiensis ATCC 14580		NC_003271	4222643 bp	Sep 28 2004
Bacillus subtilis subsp. subtilis str. 168		NC_000964	4214630 bp	Nov 20 1997
Bacillus thuringiensis anovorae kowalukian str. 97-27		NC_003957	5237682 bp	Jun 30 2004
Bacteroides fragilis NCTC 9242		NC_003328	5205143 bp	Mar 10 2005
Bacteroides fragilis NCTC 9242		NC_003329	5227724 bp	Oct 1 2004

**Figure 3-17:**  
The list of all available bacterial genome sequences (top part).

**4. Click the [NC\\_007530](#) link corresponding to the dangerous bioterrorism agent *Bacillus anthracis*.**

Your browser displays a table that summarizes the content and features of the bacterial genome. It's similar to what we already got for viruses (see Figure 3-18).



Clicking the [Here](#) link near the bottom gets you the same type of live map we saw with the HIV-1 virus (see the previous section), although the genome is now much larger. On this map, you can click to zoom into a particular region of the genome. The genome summary table contains numerous links to analyses pre-computed for all the genes (function, similarity, evolutionary relationships) that we leave you to try out for yourself.

**5. Back in the Summary table (refer to Figure 3-18) click the [Genome Project](#) link.**

Doing so gets you a quick description of the bacterium, along with a picture, details about its habitat, the disease it causes, and so on — as well as the reasons why it was important to decipher its genome sequence in the first place. (See Figure 3-19.)

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_007530	Genes: 6635	CQG	Genome Project	Publications
GenBank: AE017334	Protein coding: 5309	3D Structure	Rafseq FTP	Rafseq Status: <b>Provisional</b>
Length: 5,227,419 nt	Structural RNAs: 128	TaxMap	GenBank FTP	Seq Status: <b>Completed</b>
GC Content: 35%	Pseudo genes: 1	TaxPlot	BLAST	Sequencing center: TIGR
% Coding: 80%	Others: 6	GenePlot	TraceAssembly	Completed: 2004-05-20
Topology: circular	Contigs: 1	gMap	CGD	Organism Group
Molecule: DNA			Other genomes for species	

Gene Classification, based on COG functional categories

Search gene, GeneID or locus\_tag:  Find Gene

227410 nt

Click here for Sequence Viewer presentation (base sequence and aligned amino acids) of selected region

**Figure 3-18:**  
*Bacillus anthracis*  
(strain Ames ancestor)  
genome summary.

Click here for a live map.

NCBI ENTREZ Genome Project

Search: Genome Project for  Go Clear

Display: Overview Show: 20 Sort by: Send to:

All: 1 Prokaryotes: 1

Genome Project > *Bacillus anthracis* (anthrax bacterium) > *Bacillus anthracis* str. 'Ames Ancestor' project at TIGR

Resource Links

Endospore-forming bacteria that causes anthrax and is gold standard for comparative genomics

Lineage: Bacteria; Firmicutes; Bacillales; Bacillaceae; Bacillus; Bacillus cereus group; Bacillus anthracis str. 'Ames Ancestor'

Photo: ? Frederick C. Michel, ASM MicrobeLibrary

**Figure 3-19:**  
Genome Project page for *Bacillus anthracis* (strain Ames ancestor).

## More bacterial genomics at TIGR

The Institute for Genome Research (better known by its acronym TIGR) is home to a team of scientists who pioneered the field of bacterial genomics. In 1995, these scientists produced the two first complete bacterial genomes:

tools for use, and the possibility to run similarity searches on its genome sequence data well before the actual completion of the projects themselves. In addition, TIGR's scientists offer a specific perspective on the genomes they have sequenced through their own genome viewing software.

To pay them a quick visit, do the following:

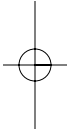
**1. Point your browser to [www.tigr.org/tdb/](http://www.tigr.org/tdb/).**

The TIGR home page appears, as shown in Figure 3-20.

**2. Click the Comprehensive Microbial Resource (CMR) link.**

A long list of micro-organisms appears, from which you get to select the one you are interested in.

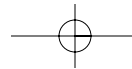
**3. Select one microbe in the list, and then explore the various analyses tools by clicking the Genome Searches, Genome Toolbox, and Genome Analyses links at the top of each microbe-specific page.**



Click here.

The screenshot shows the TIGR website interface. At the top, there is a logo for TIGR (The Institute for Genomic Research) and a navigation bar with links for Home, What's New, Search, About TIGR, Careers, Site Map, Contact Us, and FAQ. Below the navigation bar, there is a sidebar menu on the left with the following items: Comprehensive Microbial Resource, Patherna, Unfinished Microbial Genomes, Eukaryotic Resources, Gene Indices, Parasites Databases, TIGRFAMS, Fungal Databases, and Human Sequencing Projects. The main content area is titled "Genome Projects" and contains the following text: "TIGR's Genome Projects are a collection of curated databases containing DNA and protein sequence, gene expression, cellular role, protein function, and taxonomic data for microbes, plants and humans. Anonymous FTP access to sequence data is also provided. Please read the disclaimer regarding use of data. The TIGR clone distribution policy is available for viewing." Below this text are several project links, each with a small circular icon: "Bacillus anthracis" (with sub-links for Press Release, Questions and Answers, and Access to TIGR Data), "Benchmark Data for Genome Assembly" (described as complete genome sequence files for download), "Comprehensive Microbial Resource (CMR)" (described as containing analysis on world-wide completed microbial genome sequencing), "New! View predicted operons in completed microbial genomes based on conserved gene clusters", "Eukaryotic Projects" (described as containing analysis on TIGR's Eukaryotic Projects), and "Monterey Bay Coastal Ocean Microbial Observatory" (described as analysis of genetic variability, gene content, and genomic potential in uncultured marine picoplankton).

**Figure 3-20:**  
The  
Genome  
Projects site  
at The  
Institute for  
Genome  
Research  
(TIGR).



ment, or (b) offering some perspective in solving the incoming worldwide energy crisis (such as cheap ways of producing hydrogen).

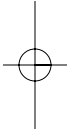
To take a look at its data, follow these steps:

1. **Point your browser to** `img.jgi.doe.gov/`.

The Integrated Microbial Genomes resources home page appears, as shown in Figure 3-21. As was the case with the NCBI and TIGR sites, specific sets of tools are proposed. The next step looks at the IMG genome display — one of the more useful IMG resources.

2. **To access the IMG Genome display, first click the Find Genomes link.**

A table of the available organisms appears.



Click here to start.

Dark Genome Search:

**img** INTEGRATED MICROBIAL GENOMES

IMG Home Find Genes Find Genomes Find Functions Compare Genomes MyIMG Analysis Cuts About IMG Using IMG News

IMG Genomes		
Genetic diversity	JGI	Total
Bacteria	39/96	664/1071
Archaea	0/1	22/5
Eukarya	0/0	12/5
Viruses	0/0	258/0
All Genomes	62/100	692/1079
Overall Total	162	741

Version 1.5 June 1, 2006  
Questions/Comments  
©2006 The Regents of the University of California  
Disclaimer

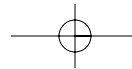
The Integrated Microbial Genomes (IMG) system (Nucleic Acids Research, 2006, Vol. 34, Database Issue D344-D349) provides a framework for comparative analysis of the genomes sequenced by the Joint Genome Institute. Its goal is to facilitate the visualization and exploration of genomes from a functional and evolutionary perspective.

The IMG user interface (see User Interface Map) allows navigating the microbial genome data space along its three key dimensions (genes, genomes, and functions), and groups together the main comparative analysis tools. Microbial genome data analysis in IMG usually starts with the definition of an analysis context in terms of selected genomes, functional annotations, and/or genes, followed by the individual or comparative analysis of genomes, functional annotations, or genes.

**Figure 3-21:**  
The Integrated Microbial Genomes site at the DoE Joint Genome Institute.

3. **Select an available organism by checking the corresponding box, then clicking Save Selections.**

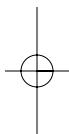
The first organism, *Aeropyrum pernix*, would be a good choice for now.



6. In the new page that appears, select a coordinate range (for instance 1 . . . 500000).

This results in a live display of the microbe gene content in that range, as shown in Figure 3-22. Putting the mouse pointer over the gene symbol gets you its name — and clicking the symbol provides you with a wealth of information on the corresponding protein.

Click on a gene symbol to find out more about its function.

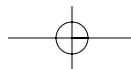


**Figure 3-22:**  
The chromosome viewer at the Integrated Microbial Genomes DoE resource.



## Exploring the Human Genome

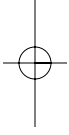
The sequencing of the human genome is probably one of the greatest scientific accomplishments of modern times. With 3000 million (and then some) nucleotides spread over 23 chromosomes, this genome is definitely a complex object to deal with. The major task ahead for bioinformatics is to integrate all the past, present, and future information that human genes contain in a maintainable, user-friendly resource.



almost-finished — human genome are periodically released. This state of flux is true for all large animal genomes.

- ✓ This sequence was obtained in raw format; the next challenge is the annotation of the raw data — creating a detailed and accurate FEATURES table of the human genome.
- ✓ Throughout the world, new information is generated daily on human gene properties and functions, using a wide array of techniques. Ideally, someone has to gather it all, package it nicely, and offer it to the whole research community for free!

Having doubts that the last goal will ever be attainable in this less-than-perfect world? Well, think again, because this is where we're taking you right now.



## *Finding out about the Ensembl project*

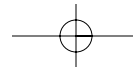
The Internet home page of Ensembl ([www.ensembl.org](http://www.ensembl.org)) says it all: Ensembl is a joint project between the European Bioinformatics Institute (EBI) and the Sanger Institute, both located near Cambridge (U.K.). Together they've developed an integrated database and software system to produce and maintain automatic annotations for the genomes of animals, with a special attention to our closest relatives: the vertebrates. This project — like the others we describe in this chapter — relies heavily on the collaboration of a large number of individual laboratories. It all started as part of the International Human Genome Project, continued for the Mouse Genome project, and is now being pursued for other animals. Data and software are also freely flowing among numerous national database and bioinformatics centers from all over the world, allowing a complex cross-linking to take place.

### *Getting started on the Ensembl site*

Considering how complex the human genome is, you may not be surprised to find that you can attack the Ensembl resources from many different angles. To quickly find out about your options, the best thing to do is to jump on the guided tour that Ensembl proposes on its home page. Here's how it's done:

- 1. Point your browser to** [www.ensembl.org/](http://www.ensembl.org/).

The impressive Ensembl home page appears, as shown in Figure 3-23.





(*Dasybus novemcinctus*!). However, the main reason we're here is to take a look at the human genome, so let's click the DaVinci-inspired *Homo sapiens* icon.

Use BioMart for serious data mining.

Click here to start navigating the human genome.

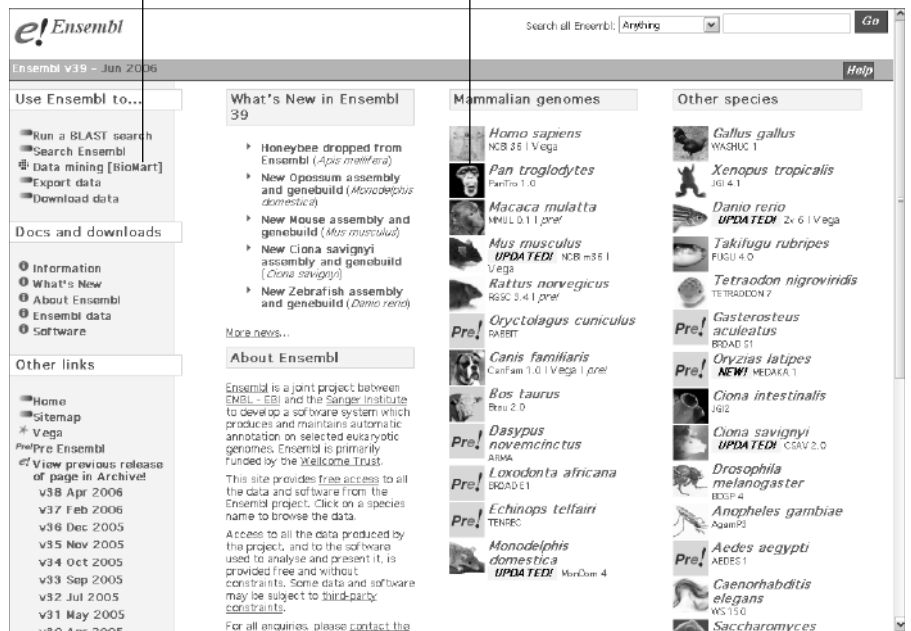
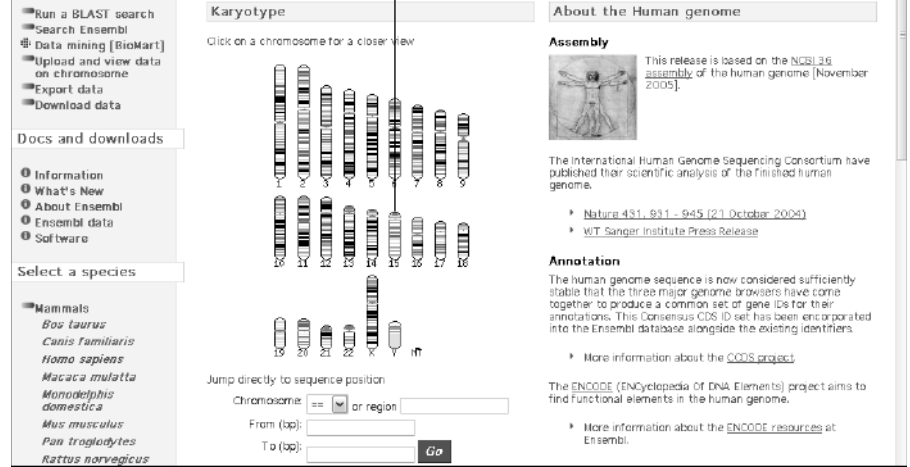


Figure 3-23:  
The  
Ensembl  
project  
home page.

## 2. Click the *Homo sapiens* icon.

The page that appears (see Figure 3-24) presents a schematic image of the various human chromosomes (numbered according to their size), including the sex chromosomes X and Y, as well as the DNA molecule of the human mitochondria (a small energy-producing organelle present in all human cells). This schematic image is “hot,” meaning that clicking a certain area brings up a new window associated with that area.

**Figure 3-24:**  
Starting  
your journey  
in the  
human  
genome.



### 3. Click anywhere on the chromosome 15 picture.

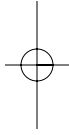
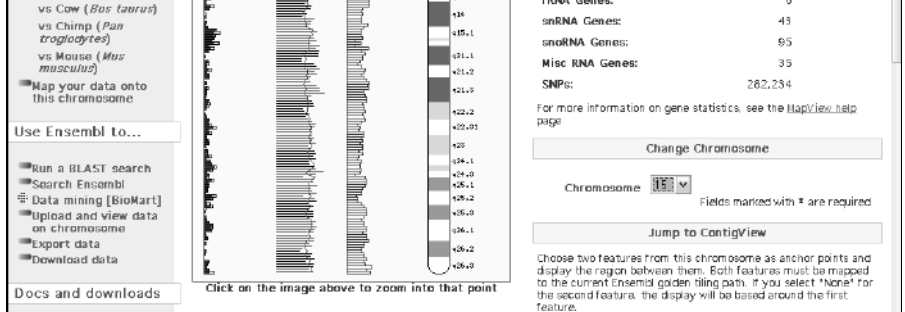
This lands you in the Chromosome 15 data subset (Figure 3-25), within which we can now ask specific questions, such as: Where is the dUTPase gene? We can do this the hard way by clicking in the region of the q21.1 band and then manually scanning the region. (We see this information earlier in this chapter — remember Figure 3-8?) An easier way is to use the Feature Locator at the bottom right corner of the page, below the Jump to ContigView yellow banner, as shown in Figure 3-26. To do so, follow these steps:

- a. Select Gene in the From pull-down menu.
- b. Enter DUT in the search window, then click the red Go button.

You are now looking at a complex output, giving you the structure and the genomic context of the DUT gene at increasing resolution from top to bottom: a schematic overview (1 Mbp range) (see Figure 3-27), a detailed view (10-kbp range) (see Figure 3-28), and even a base-pair view (100-bp range).

Mousing over the display at various levels lets you examine any chromosome location (down to the level of a gene and its neighbors), study the internal structure and the alternative forms of expression of this specific gene, and assess the eventual consequences of single-nucleotide polymorphisms within that gene (SNP).

**Figure 3-25:**  
Focusing on  
a given  
genome  
data subset:  
Human  
Chromosome  
15.

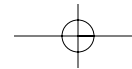


Select gene.      Input gene name.

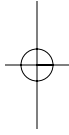
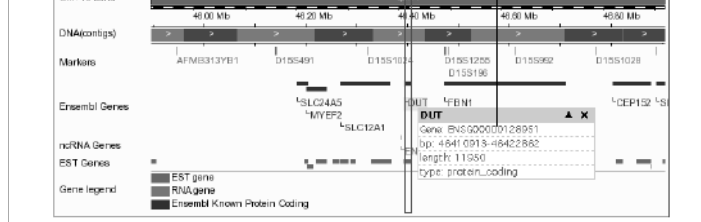
**Figure 3-26:**  
Looking for  
the Human  
dUTPase  
gene: DUT.

### ***Getting a complete ID card on the Human DUT gene***

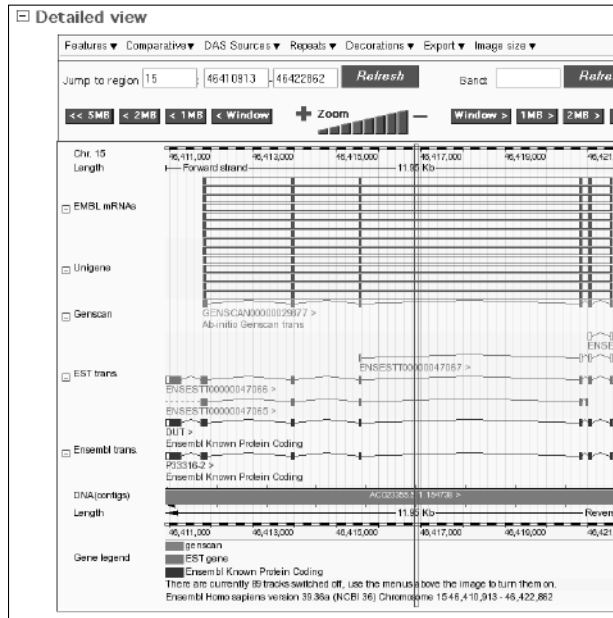
Notice that in the overview display (refer to Figure 3-27), the DUT gene name is highlighted. Clicking the gene name here reveals a label with its Ensembl gene number, its location, length, and type (protein coding). In turn, clicking the gene number returns an exhaustive ID card where everything you ever wanted to know about this gene can be found, either explicitly, or as one of the zillions of links to relevant entries in other databases. Check out Figure 3-29 to see how the Human DUT ID card looks like.



**Figure 3-27:**  
The Human  
DUT gene  
location:  
1-Mbp  
range  
overview.

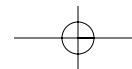


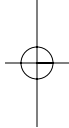
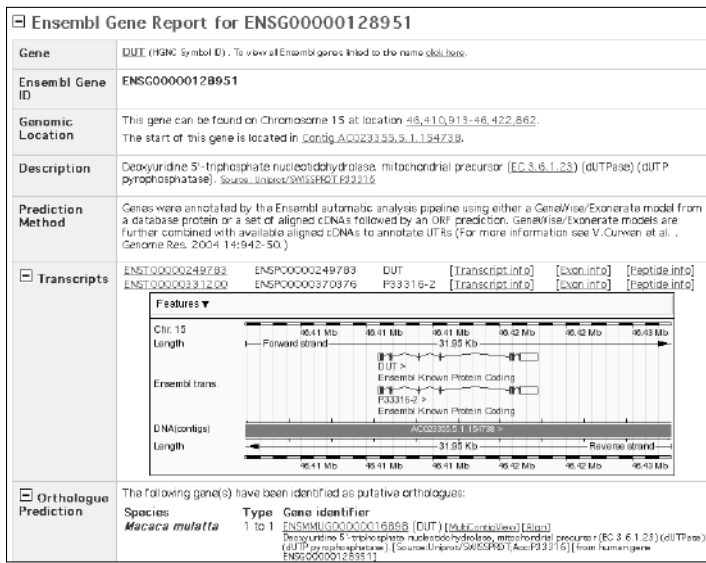
**Figure 3-28:**  
The Human  
DUT gene  
location:  
10-kb range  
detailed  
view.



### *Finding disease genes with coding SNPs using BioMart*

There's more to Ensembl fun than just random browsing. If you're a serious scientist, it can answer meaningful questions very easily, very precisely, and very quickly, even if the output isn't always as pretty as what we showed you in the preceding section. The data-mining system that the Ensembl people designed for this purpose is called BioMart. (Refer to Figure 3-23, upper-left corner.)





**Figure 3-29:**  
Top part of the Ensembl ID card for the Human DUT gene.

If you're just dying to find a candidate gene for a disease you've just genetically mapped on chromosome 15, the first question makes a lot of sense.

To make things even more interesting, we can look for *coding SNPs* — the impetus for our second question. (SNPs involve nucleotide changes that may alter the sequence and possibly the shape and/or function of the corresponding protein product.)

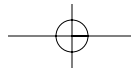
Now, both of these questions are real doozies. Believe it or not, you can answer them in less than a minute with Ensembl! Follow these steps to see how:

1. Point your browser to [www.ensembl.org/](http://www.ensembl.org/).

The Project Ensembl home page appears.

2. Click the **Data Mining** link in the upper-left corner of the page. (Refer to Figure 3-23.)

Your browser displays the starting page MartView page, which is already set up for working with the last version of the Human genome data (otherwise select the dataset).



- In the following section (Gene set up), check the top box, and select With Disease Association from the accompanying pull-down menu. Then check the Gene Type box, and select protein\_coding from the associated pull-down menu.
- Scroll down the form until you reach the SNP set up area; check the Coding box and then click the Next button at the bottom of the page.
- In the Output form that appears, check the Chromosome Name and Band boxes, and then check the Ensembl Gene ID, Description and MIM Disease ID boxes in the next section.

MIM stands for *Mendelian Inheritance in Man*, and is the leading human-genetic-disease database.

- Proceed to the bottom of the form and click the Export button.

In no time at all, Ensembl returns a table that looks like Figure 3-30. Our BioMart search identified many different genes known to be associated with various diseases, and for which sequence polymorphisms inducing protein changes have been documented. These will be our prime candidates for further research. As before (refer to Figure 3-29), clicking the Ensembl Gene ID will produce a comprehensive information card for these genes, allowing you to quickly figure out whether their functions are related to the disease symptoms.

**Figure 3-30:**  
Output of  
your first  
BioMart  
query.

Chromosome Name	Band	Ensembl Gene ID	Description	MIM Disease ID
15	q11.2	ENSG00000182638	Neodn. [Source:Uniprot/SWISSPROT,Acc:Q99608]	
15	q12	ENSG00000104044	P protein (Melanocyte-specific transporter protein) (Pink-eyed dilution protein homolog). [Source:Uniprot/SWISSPROT,Acc:Q04671]	176270 203200
15	q12	ENSG00000104044	P protein (Melanocyte-specific transporter protein) (Pink-eyed dilution protein homolog). [Source:Uniprot/SWISSPROT,Acc:Q04671]	203310
15	q13.3	ENSG00000175944	Neuronal acetylcholine receptor protein, alpha-7 subunit precursor. [Source:Uniprot/SWISSPROT,Acc:P36944]	118511
15	q21.1	ENSG00000171766	Glycine amidinotransferase, mitochondrial precursor (EC 2.1.4.1) (L-arginine glycine amidinotransferase) (Transamidinase) (AT). [Source:Uniprot/SWISSPROT,Acc:P60442]	602360
15	q21.2	ENSG00000137569	Cytochrome P450 19A1 (EC 1.14.14.1) (Aromatase) (CYP19) (Estrogen synthetase) (P-450AROM). [Source:Uniprot/SWISSPROT,Acc:P11511]	107910
15	q21.3	ENSG00000103569	Aquaporin-9 (AQP-9) (Small solute channel 1). [Source:Uniprot/SWISSPROT,Acc:O43316]	602914
15	q21.3	ENSG00000166039	Hepatic triacylglycerol lipase precursor (EC 3.1.1.3) (Hepatic lipase) (HL). [Source:Uniprot/SWISSPROT,Acc:P11150]	151670
15	q24.1	ENSG00000138623	Semaphorin-7A precursor (Semaphorin L) (Sema L) (Semaphorin K1) (Sema K1) (John-Milton-Hargen human blood group Ag) (JMh blood group antigen) (CD108 antigen) (CDw108). [Source:Uniprot/SWISSPROT,Acc:O76326]	607961
15	q24.3	ENSG00000140369	Proline-serine-threonine phosphatase-interacting protein 1 (PEST phosphatase-interacting protein 1) (CD2-binding protein 1) (H-PIP). [Source:Uniprot/SWISSPROT,Acc:O43586]	604416
15	q26.1	ENSG00000103676	Fumarylacetoacetase (EC 3.7.1.2) (Fumarylacetoacetate hydrolase) (Beta-diketonase) (FAA). [Source:Uniprot/SWISSPROT,Acc:P16930]	276700

The range and complexity of the questions you can address through the Ensembl BioMart resource is truly impressive. We really encourage you to spend some time playing with it, even though it isn't as visual as the zooming tools we showed you in previous sections.

# Using Protein and Specialized Sequence Databases

---

## *In This Chapter*

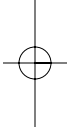
- ▶ Exploring protein maturation
  - ▶ Understanding a UniProtKB/Swiss-Prot entry from the top down
  - ▶ Finding out about detailed protein biochemistry
  - ▶ Discovering the function of your protein
  - ▶ Hunting up useful protein-related information resources
- 

*Mankind is a catalyzing enzyme for the transition from a carbon-based to a silicon-based intelligence.*

— Gerard Bricogne

**W**e've focused this chapter to finding information on protein sequences. The databases on proteins don't limit themselves to providing sequences. In this chapter, we show you that by mastering the — sometimes tangled — network of Web links available to you on the Internet, you can relate proteins to their gene sequences, to their functions, and even to their 3-D structures.

Don't let that abundance and diversity of information fool you. As it is for the nucleotide sequence world, where everything revolves around one central resource (that is, GenBank), most of the links between genes, proteins, and functions rely on UniProtKB/Swiss-Prot, the central resource on annotated protein sequences co-funded by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). In this chapter, we show you how to extract every bit of information you need from this database.



nity of biologists, their overall philosophies aren't quite the same. GenBank, as a primary sequence repository, obeys a relatively strict historical point of view. In GenBank, the authors have full authority over the content of the entries they submit. GenBank annotators are only responsible for the recently introduced RefSeq entries — the ones they derive from their own expert analysis of the community-submitted entries. However, the RefSeq entries never replace the original GenBank entries, and all these layers of information are maintained side by side.

In contrast, Swiss-Prot is not a repository database but a derived information resource, conveying the vision of its head and founder, Amos Bairoch (helped out by a group of experts). As a consequence, the only truly current Swiss-Prot version is the one on Amos' portable computer. This idea of "personal vision" also means that Amos Bairoch doesn't need anybody's permission to correct or change a Swiss-Prot entry on the spot. To do this, Amos simply needs to be convinced by an expert, or by his own evaluation of the literature, that a change is necessary. Believe us, we have seen this happening on our own kitchen table!

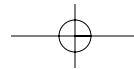
The benefit of such a philosophy is obviously great when it comes to flexibility; blatant errors can be removed very quickly while hot discoveries are incorporated immediately. This is the reason why Swiss-Prot is considered the best-annotated protein database, and why it occupies

upper limit of what the best researcher in the world can do (or supervise), even when helped by a large team of annotators — at the European Bioinformatics Institute (EBI) in Hinxton as well as the Swiss Institute of Bioinformatics (SIB) in Geneva — and a circle of experts. These days, Swiss-Prot has troubles coping with the present rate of new (nucleotide) sequence determination and is falling behind in terms of completeness.

To alleviate this problem, a Swiss-Prot buffer has been created that is called TrEMBL (for automatic *T*Ranslation of *E*uropean *M*olecular *B*iology *L*aboratory nucleotide sequences). TrEMBL entries are generated at the EBI from GenBank submissions and annotated mostly automatically, using sequence similarity as a main criterion. Upon visual inspection, manual correction, and final approval by Amos Bairoch, TrEMBL entries are then converted into *bona fide* Swiss-Prot entries.

The idea that such a key worldwide information resource rests on a single's man shoulders may come as a shock to you. However, this sort of thing has been quite common since the early days of bioinformatics. Among other famous examples of one-man shows, we can cite Elvin Kabat's Immunoglobulin sequence database, Richard J. Roberts' Restriction Enzyme Database, or Victor McKusick's database of human genetic diseases (Online Mendelian Inheritance in Man, OMIM). To paraphrase Sir Winston Churchill: "In the field of bioinformatics, rarely have so many owed so much to so few!"

You have many ways of landing on the Swiss-Prot database (or the UniProt knowledge base, as they also call it now!). Most genomic databases, genome browsers, or sites running a sequence retrieval system (SRS) link you directly to the relevant Swiss-Prot entry. If you want to know how to query Swiss-Prot to find an entry by keyword, go to Chapter 2, where we explain how to do this in some detail.





We start this chapter by giving you some general ideas on how the cell turns a native protein sequence into a mature protein. These post-translational modifications are very important for the protein function. Finding out about them is one of the main reasons you want to use a protein database. (See Chapter 6 if you want to find out how to predict these modifications.)

## ***From Translated ORFs to Mature Proteins***

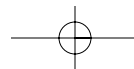
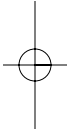
Because sequencing DNA molecules is so much easier and cheaper than sequencing proteins, most of the amino-acid sequences we know do in fact correspond to computer translations of open reading frames (ORFs) detected through the analysis of genomic data. Most of the corresponding proteins have never actually been isolated by anybody. This is why protein specialists hate us molecular biologists and bioinformaticists when we keep talking about translated ORFs as if they were genuine proteins. For a while, at the beginning of the genomic era, the easy thing to do was to consider these protein specialists a bunch of grumpy old men and simply ignore them.

However, after genomes were sequenced, people's interest in proteins came back in a large-scale format known as proteomics. *Proteomics* is the scientific field dealing with the visualization and quantification of the set of protein molecules present in a given tissue or organism. Proteomics analysis brought a rapid accumulation of data — and quickly demonstrated that protein specialists had very good reasons to be angry with us.

### ***ORFs: What you see is NOT what you get***

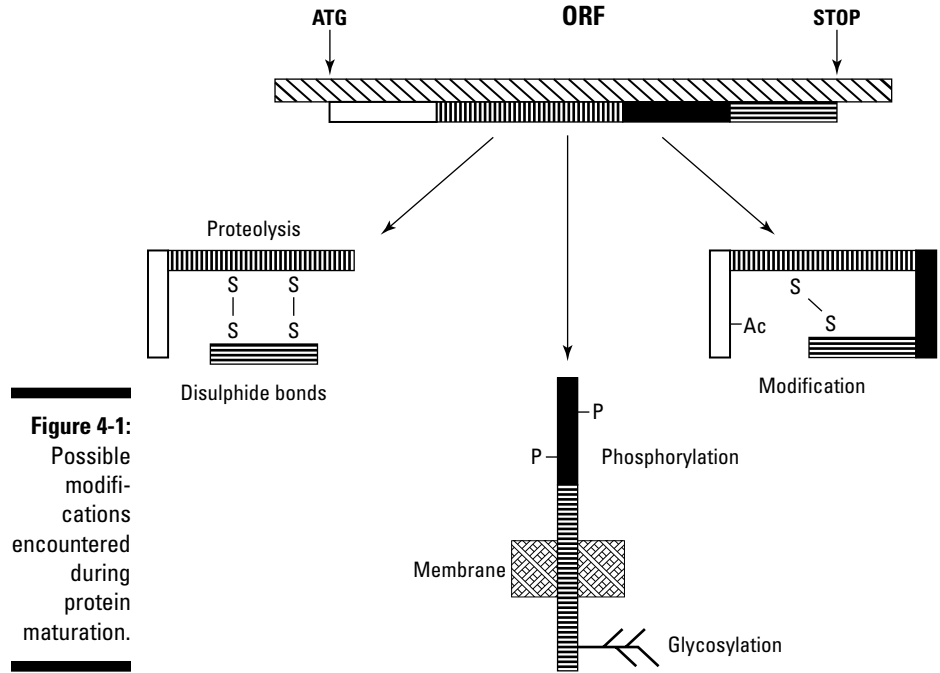
When biologists perform a 2-D gel electrophoresis (a tricky experiment separating protein molecules according to their mass in one direction and their charge in the other), real proteins almost never behave as you'd expect given the computer translation. In the world of ORFs and proteins, *what you see is NOT what you get*.

The reason is that, when translated from RNA, the nascent amino-acid chain can be heavily modified on its way to becoming a mature protein. Even simple physico-chemical properties of a mature protein — such as size, molecular weight, or isoelectric point — are hard to predict if you know only the computer-generated amino-acid sequence. This complex process of protein



- ✓ Removal of fragments of the amino-acid chain (this is the case for insulin)
- ✓ Chemical modifications of specific amino acids (methylation, for example)
- ✓ Addition of lipid molecules (myristoylation, for example)
- ✓ Addition of glycosidic (sugar) molecules (glycosylation, for example)

A major role for a *protein* database (in contrast with an ORF sequence collection) is to display this type of information, when it is available from experimental data or is predicted using various computational techniques. (For more on such techniques, see Chapter 6). Hence post-translational modifications make up a sizeable portion of the protein features recorded in Swiss-Prot entries.



**Figure 4-1:**  
Possible modifications encountered during protein maturation.

right destination in the organism or within the cell. As it is translated, the peptide chain may expose a variety of highly specific sequence signals — “ZIP codes,” if you will — which the cell then uses to direct the protein to the appropriate compartment (in or out of the cell). This sorting always involves the transport of the protein across one or several membranes and is also referred to as *translocation*. The final activities and destinations of a protein include

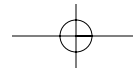
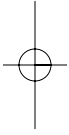
- ✓ Getting attached to the cell membrane
- ✓ Being secreted outside the cell
- ✓ Being transported into the periplasm (for bacteria)
- ✓ Being transported to the mitochondria or any other organelle
- ✓ Being transported into the cell nucleus

Because knowing the final compartment where a protein ends up is important in understanding its function, this information (proven or predicted) is one of the important features recorded in protein databases like Swiss-Prot.

## *A combinatorial diversity of folds and functions*

The most important step in turning a newly synthesized peptide chain into a functional protein is the folding of this chain into a compact and stable 3-D structure. Except for small proteins (fewer than 100 amino acids), the final protein structure generally consists of several relatively independent domains. You can imagine these domains as a basic set of fairly rigid LEGO bricks. Nature can assemble these bricks to produce the immense variety of existing proteins. For instance, given a basic set of three domains (A, B, C), you can end up with Protein AAA or AB or BCC or BAC, and so on.

Most natural proteins are made of combinations of one to ten domains picked from a set of a few thousands. Despite significant sequence variations, the domains are identifiable by their *scaffold sequence signatures* — the motifs in the protein amino-acid texts that remain recognizable despite a zillion years of divergent evolution. The recognition and the definition of protein domains is a major research topic of bioinformatics. (See Chapters 6, 7, 9 and 11.) The domain architecture underlying a particular protein sequence is important to know because it gives hints about the protein’s possible 3-D structure — and suggests its potential biochemical or cellular function.



# Reading a Swiss-Prot Entry

Despite the complicated concepts that their descriptions rely on, proteins are much simpler objects than genes:

- ✓ Proteins correspond to relatively small sequences (350 amino acids long, on the average).
- ✓ Unlike genes, proteins have clear beginnings and clear ends.
- ✓ Proteins are defined on a single strand.
- ✓ Whatever modifications occur between the ORF sequence and the mature protein, the amino acids they contain remain in the same order.

Given all these reasons, you'd probably expect Swiss-Prot entries to be easier to read and understand than GenBank's — and you'd be right!

Like a GenBank entry — which we cover with great flair in Chapter 3 — a Swiss-Prot entry is divided into several main sections:

- ✓ The general information part
- ✓ The bibliographic information part
- ✓ The functional information part
- ✓ The feature table
- ✓ The sequence part

Using the human *epidermal growth factor receptor* (EGFR) as an example, the following sections show you how to decipher a Swiss-Prot entry from top to bottom, step by step (or line by line).

## Deciphering the EGFR Swiss-Prot entry

The EGF receptor is a rather complex molecule. We use it here because it illustrates remarkably well the wealth of information you can glean from a Swiss-Prot entry.

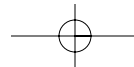
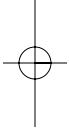


Figure 2-10, Chapter 2, to remember what it looks like).

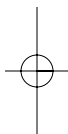
## 2. Type the Swiss-Prot ID P00533 in the Search window (top of the page).

This is the entry corresponding to the human receptor of the epidermal growth factor. We chose it because this protein is hot! It is an oncogene (a trigger for cancer), and it exhibits a fairly detailed annotation. Because more people work on hot proteins, they're usually better annotated than the more mundane ones.

## 3. Click the Go button.

The UniProtKB/Swiss-Prot P00533 entry page appears. If you want to print this page, click the Printer-Friendly View button at the top right.

Because of the large amount of information available that deals with this particular protein, this entry is probably among the most complicated you'll ever come across when interacting with Swiss-Prot. Most other entries don't exhibit a third of the keywords and fields you can find here.



## General information about the entry

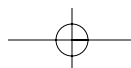
You're now ready to go through this fairly complex entry step by step. Blue banners separate the main sections across the screen. The first one is titled Entry Information and contains the following six fields:

✓ **Entry Name: *EGFR\_HUMAN***. This Swiss-Prot identifier gives a quick idea of what the entry is all about. Here, the name is quite explicit, but don't count on that — such clarity isn't always the case! For instance, the fly lysozyme entry name is *LYSA\_DROME*, certainly more ambiguous and more difficult to remember (DROME stands for *DRO*sophila *ME*lanogaster).



As a general rule, the entry name is not supposed to be meaningful, nor stable. As a consequence, you're strongly advised *not* to use it for cross-reference purposes, or as a sequence identifier in an article. Instead, you should use the primary accession number that appears on the next line.

✓ **Primary Accession Number: *P00533***: This is the truly unique and stable identifier of this sequence. You must use this number when referencing the entry in your work.



(perhaps in the process correcting earlier mistakes). Unlike GenBank, Swiss-Prot keeps only one version of the EGF receptor and summarizes our knowledge up to this day.

Ancient accession numbers never die; they just get stored in the secondary accession number field.

- ✓ **Secondary Accession Numbers: O00688, P06268, Q14225, et al:** Secondary accession numbers make sure that when you read an old paper mentioning Q12225 or P06268 as accession numbers for EGFR, you can still use the old reference to find the most recent one. See for yourself and try to use one of these older numbers in the Search window.

Whatever secondary accession number you type in, you always fall back on the current one: P00533. This is a clean and simple way to remove old entries while keeping them unambiguously linked to the state-of-the-art version.

The next lines in the General information are self-explanatory.

- ✓ **Integrated into Swiss-Prot on: July 21, 1986.**
- ✓ **Sequence was last modified on: November 1, 1997.**

This suggests that it took ten years to get the whole sequence right!

- ✓ **Annotations were last modified on: July 25, 2006.**

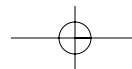
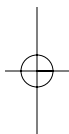
Clearly, a lot of work is still going on about this important protein.



Notice that, for the whole entry, the field definition (the left part of each line) is an active link that directs you to the relevant part of the Swiss-Prot manual. Reading a few of those might earn you some respect from the local bioinformatics gurus!

## *Name and origin of the protein*

At this point, we encounter a new section (introduced by the blue banner): Name and Origin of the Protein. It contains the following five fields:



cleaved to form the mature form of the protein.

- ✓ **Synonyms:** *Receptor tyrosine-protein kinase ErbB-1, EC 2.7.10.1*: The E.C. number (2.7.10.1) encodes the biochemical reaction (tyrosine-protein kinase) that this protein performs. E.C. stands for Enzyme Nomenclature Committee, and we give you a bit more information about it in Table 4-1.

This little line doesn't look very impressive, but it can be the starting point of a fascinating journey toward a complete understanding of this protein enzymatic function. For example, click the [2.7.10.1](#) link to see the NiceZyme view of the enzyme, as shown in Figure 4-2. The links on this page lead to various types of information; you can

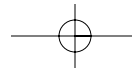
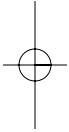
- Locate your protein within a chart of all cellular metabolic pathways
- Get a description and sequence profile of the family it belongs to
- Get the chemical formulas of the biochemical compounds it can interact with

- ✓ **Gene name:** *EGFR*: This field is useful when you need to interact with nucleotide sequence and genome databases. It can help keep queries clear and unambiguous. (For more on gene names, see Chapter 3.)

- ✓ **From:** *Homo sapiens (Human) [TaxID:9606]*: The From field defines precisely the origin of the protein. The species designation consists, in most cases, of the Latin genus and species designation followed by the English name (in parentheses). For viruses, only the common English name is given. Clicking the Taxonomic ID [9606](#) link gets you to the *H.sapiens* page of the NEWT taxonomy database maintained by the UniProtKB group ([www.ebi.ac.uk/newt/index.html](http://www.ebi.ac.uk/newt/index.html)).

- ✓ **Taxonomy:** *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; ..., etc*: This field lists the taxonomic classification of the source organism. The taxonomic classification used is that maintained at the National Center for Biotechnology Information (NCBI). Each group name is a link to a list of all the Swiss-Prot entries from that group.

This field is handy if you want to quickly gather all the protein sequences of a given species for your research projects.



ATP + a [protein]-L-tyrosine ↔ ADP + a [protein]-L-tyrosine phosphate

**Comment(s)**

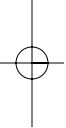
- The receptor protein-tyrosine kinases, which can be defined as having a transmembrane domain, are a large and diverse multigene family found only in metazoans.
- In the human genome, 58 receptor-type protein-tyrosine kinases have been identified and these are distributed into 20 subfamilies.
- Formerly EC 2.7.1.112.

**Cross-references**

PROSITE	PDOC00100
BRENDA	2.7.10.1
PUMA2	2.7.10.1
PRISM enzyme-specific profiles	2.7.10.1
Kyoto University LIGAND chemical database	2.7.10.1
IUBMB Enzyme Nomenclature	2.7.10.1
IntEnz	2.7.10.1
MEDLINE	Find literature relating to 2.7.10.1
MetaCyc	2.7.10.1

P13368, TLESS\_DROME; P20806, TLESS\_DROME; Q5UN73, ALK\_HUMAN;  
P97793, ALK\_MOUSE; Q9VEK3, CAD96\_DROME; P18460, CEK1\_CHICK;  
P18461, CEK1\_CHICK; F13349, CSF1R\_FELCA; P07333, CSF1R\_HUMAN;  
P09551, CSF2E\_MOUSE; Q06495, CSF1R\_RAT; Q06495, CSF1R\_HUMAN

**Figure 4-2:**  
NiceZyme  
view of  
Swiss-Prot  
entry  
P00533.



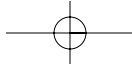
## The References

The next section of the P00533 entry is the References section — and it really is as simple as it looks: a list of the bibliographical references used to build the current entry. In this example, it includes the various studies that have contributed to the sequencing, to the definition of several isoforms of the protein as expressed in different tissues or tumors, and to the mapping of various post-translational modifications, disulphide bonds, or ligand-binding sites. This protein's importance can be measured by the number of references here — 34 is pretty impressive.

For each of the published citations, Swiss-Prot provides the relevant NCBI/PubMed identifier and for most of them a direct link to the article abstract.

## The Comments

The Comments section holds any useful information that just doesn't fit anywhere else. The comments are free texts grouped together in comment blocks; a block contains one or more comment lines and is introduced by a topic keyword. The topic keywords themselves are pretty wide ranging, as the listing in Figure 4-3 makes clear. (You can get to this listing — part of the Swiss-Prot manual — by clicking the Comments heading.)





CAUTION	Warning about possible errors and/or grounds for confusion
COFACTOR	Description of any non-protein substance required by an enzyme for its catalytic activity
DEVELOPMENTAL STAGE	Description of the developmentally-specific expression of mRNA or protein
DISEASE	Description of the disease(s) associated with a deficiency of a protein
DOMAIN	Description of the domain structure of a protein
ENZYME REGULATION	Description of an enzyme regulatory mechanism
FUNCTION	General description of the function(s) of a protein
INDUCTION	Description of the compound(s) or condition(s) that regulate gene expression
INTERACTION	Conveys information relevant to binary protein-protein interaction 3.21.12
MASS SPECTROMETRY	Reports the exact molecular weight of a protein or part of a protein as determined by mass spectrometric methods, see 3.21.23
MISCELLANEOUS	Any comment which does not belong to any of the other defined topics
PATHWAY	Description of the metabolic pathway(s) with which a protein is associated
PHARMACEUTICAL	Description of the use of a protein as a pharmaceutical drug
POLYMORPHISM	Description of polymorphism(s)
PTM	Description of any chemical alteration of a polypeptide (proteolytic cleavage, amino acid modifications including crosslinks). This topic complements information given in the feature table or indicates polypeptide modifications for which position-specific data is not available
RNA EDITING	Description of any type of RNA editing that leads to one or more amino acid changes
SIMILARITY	Description of the similarity(s) (sequence or structural) of a protein with other proteins
SUBCELLULAR LOCATION	Description of the subcellular location of the mature protein
SUBUNIT	Description of the quaternary structure of a protein and any kind of interactions with other proteins or protein complexes, except for receptor-ligand interactions, which are described in the topic FUNCTION.

**Figure 4-3:**  
Key topics  
in the  
Comments  
section of  
Swiss-Prot.

The Comments section of entry P00533 is particularly rich, as it includes information about the following topics:

- ✓ **FUNCTION:** Receptor for EGF, involved in the control of cell growth
- ✓ **CATALYTIC ACTIVITY:** Tyrosine-protein kinase
- ✓ **SUBUNIT:** Other proteins it forms stable complexes with
- ✓ **INTERACTION:** Other proteins it interacts with
- ✓ **SUBCELLULAR LOCATION:** Cellular-membrane protein
- ✓ **ALTERNATIVE PRODUCTS:** The isoforms from splice variants
- ✓ **TISSUE SPECIFICITY:** Placenta, ovarian cancer
- ✓ **PTM:** List some of the Post Translational Modifications
- ✓ **DISEASE:** Defects in EGFR are associated with lung cancer
- ✓ **MISCELLANEOUS:** Used here to describe the molecular mechanism (dimerization) underlying the function of the protein
- ✓ **SIMILARITY:** With other sequences of the EGF receptor subfamily
- ✓ **WEB RESOURCE:** A link to GeneTests, an NIH-funded database on clinically available genetic tests

# The Cross-References

The Cross-References section contains links to entries in other databases that contain some information about our protein.



Most of the links in this section take you to new databases. With all this Web jumping about from database to database, it can be easy to get lost and lose your original Swiss-Prot entry. To avoid this confusion, when you browse the Cross-References section, open the links in a new window: Click the link with the *right* button of your mouse (not the left, as you usually do) and choose Open in a New Window from the context menu that appears.

There are bunches of fields in the Cross-References section. The following list will help you keep at least the major ones straight:



- ✓ **EMBL:** This field contains all necessary links with the nucleotide sequences world. As we point out in Chapter 3, numerous GenBank, EMBL, and DDBJ entries can be related to a single protein sequence.  
Click the [CoDingSequence](#) link to send a query to the EBI SRS server for finding the CDS (*CoDing* Segment, the part of the nucleotide sequence that precisely encodes the protein from start to stop) of these entries.
- ✓ **PIR:** This historical field contains the accession numbers of the corresponding entries in the late Protein Information Resource (PIR), now discontinued, but incorporated in the UniProtKB consortium.
- ✓ **UniGene:** A link to NCBI's gene expression database (see also the CleanEx field for linking to DNA chip experimental data).
- ✓ **PDB:** This field contains a link to a sequence homologue of the current query, for which 3-D structural information (X-ray crystal structures) is available. Here, this is Swiss-Prot entry P11362, the basic fibroblast growth factor 1, for which several crystal structures exist in the protein Databank (PDB). Click the [PDB](#) link to see the relevant page. (The PDB ID here is 1FGK.)
- ✓ **ModBase:** This field links you to ModBase, a database of theoretically calculated models, not experimentally determined structures. The models may contain significant errors, as pointed out by the authors themselves.

stands for *Database of Interacting Proteins*, a database maintained by UCLA. This is an interesting resource in that it lists all the proteins that have been experimentally shown to interact with our EGF-receptor protein. IntAct (a project from EBI) is built from the information found in the literature.

- ✓ **GlycoSuiteDb:** This field is only there when your protein has a known glycosylation pattern. It provides a link to the relevant chemical composition of the sugar moiety. This is a nice site, but the free service is provided only to academics and requires a personal registration, a significant drawback to other researchers.
- ✓ **SWISS-2DPAGE:** This field contains a link to the proteomics world. Clicking [SWISS-2DPAGE](#) brings you to a collection of 2-D gel electrophoreses for different human tissues or tumors. If your protein has been identified on a gel, the spots are highlighted. This is the case for P00533. Otherwise, all that appears on-screen is a box of its expected location (drawn after its theoretical isoelectric point pI and predicted molecular weight). This is great if you're ready for real experimental work!
- ✓ **HGNC:** This field (for human proteins only) provides a link to the official human gene names provided by the HUGO (*HU*man Genome Organization) Gene Nomenclature Committee.
- ✓ **GeneCards:** This field provides a link to the relevant entry in the GeneCards database maintained by the Weizmann Institute of Science, based in Israel.
- GeneLynx:** This field provides a link to the relevant entry of GeneLynx, a collection of hyperlinks to the publicly available resources associated with each human gene.
- ✓ **MIM:** This field (mostly for human/mouse proteins) provides a link to the relevant entry in OMIM, the human genetic disease database Online Mendelian Inheritance in Man.
- ✓ **Ontologies:** This subsection attempts to list, in a hierarchical manner (called an *ontology*) all the functions the EGFR gene product is known to be associated with. Clicking on the QuickGo view link (at the send of this subsection) gets you a nice summary of all that information.

recurrent structural/sequence domains, as well as for the clustering of the proteins in families. The principle behind these classifications and the use you can make of them are well described in Chapters 7, 9, and 11. The relevant keywords are

- InterPro (summarizing most of the following ones)
- Pfam
- Prodom
- PRINTS
- SMART
- PROSITE
- BLOCKs

If you have time for only one click, use the [InterPro](#) link; it is a summary of all the others (except for BLOCKs). Clicking the [Graphical view of domain structure](#) link in the InterPro row shows you at once that many different recurrent domains have been identified in the EGF receptor protein P00533.

- ✓ **Ensembl:** This field is only here for human and mouse proteins. It provides a link to the relevant gene report in the Ensembl database, a major human genome resource. (For more on Ensembl, see Chapter 3.)

With the NCBI/PubMed and the PDB links, we believe this is one of the most useful hyperlinks provided in the Swiss-Prot entries for mammalian genes.

- ✓ **RZPD-ProtExp and SOURCE:** These useful fields (for the experimentalists) let you identify sources of publicly available starting material (like cDNA clones) in case you decide to work on the *EGFR* gene, or on its protein product.

This ends the very long Cross-References section for entry P00533. If you've been clicking many of these links, you've pretty much found out everything there is to know about your protein.



## The Keywords

The following section of the entry is the Keywords field (on the top of Figure 4-4). It is a simple list of terms relevant to your current protein, such as transmembrane, receptor, glycoprotein, and so on. Clicking any one of



server's SRS search (see Chapter 2) allows you to query Swiss-Prot by using keywords in combination. This technique gives you more specific searches.

## The Features

Farther down the P00533 entry, we enter the Features section, where information on the protein is precisely mapped onto the sequence. This concept will be familiar if you've worked with GenBank (as described in Chapter 3) — fortunately, it's much simpler for proteins.

The main part of Figure 4-4 shows the beginning of the Features section for P00533.

The whole section is organized in a table with the following self-explanatory headings:

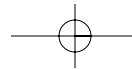
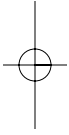
Key	From	To	Length	Description
-----	------	----	--------	-------------

The **Key** field indicates the kind of feature you're looking at. The numbers in the **From** and **To** fields correspond to the N-terminus-through-C-terminus numbering of the sequence, whereas the **Length** field does the math for you and gives the length of the sequence. The **Description** field gives you a brief written summary of the feature.

Going down the **Key** field column for the P00533 entry (refer to Figure 4-4), we can see numbers associated with the following key terms:

- ✓ **SIGNAL:** The numbers associated with this Key term indicate that residues 1 to 24 are a signal peptide, as expected for a transmembrane receptor protein.

Notice that clicking the [SIGNAL](#) link (like any other keyword link in this field) brings out the relevant part of the Swiss-Prot user manual (refer to Figure 4-4) where you can read about the precise meaning of this term. Clicking the underlined sequence range (for example, [1-24](#)) highlights this segment in the display of the sequence.



domain) or remains inside (the intracellular domain).

✓ **TRANSMEM:** Here we see that the following 23 residues (646–668) are the transmembrane segment of the protein.

✓ **DOMAIN:** These numbers indicate another domain. Residues 669 to 1210 are all inside the cell, forming the intracellular domain of the protein.

Notice that what Swiss-Prot calls DOMAIN here, corresponds to a broader topological definition than what specialists usually mean with this word in databases like InterPro or Pfam. For instance, the extracellular domain of our EGFR protein contains several InterPro domains. (Click the links to see for yourself.)



Graphic display

On line manual

Key	From	To	Length	Description	FTID
SIGNAL	1	24	24		
CHAIN	25	1210	1186	Epidermal growth factor receptor.	PRO_000001665
TOPO_DOM	25	645	621	Extracellular (Potential).	
TRANSMEM	646	668	23	Potential.	
TOPO_DOM	669	1210	542	Cytoplasmic (Potential).	
REPEAT	75	300	226	Approximate.	
REPEAT	390	600	211	Approximate.	
DOMAIN	712	979	268	Protein kinase.	
NP_BIND	718	725	9	ATP (By similarity).	
COMPELIS	1025	1071	47	Sec-rich.	
ACT_SITE	837	837		By similarity.	
BINDING	745	745		ATP (By similarity).	
SITE	1016	1016	1	Important for interaction with P39328.	
MOD_RES	678	678		Phosphochreonine (by PKC).	
MOD_RES	693	693		Phosphochreonine.	
MOD_RES	695	695		Phosphoserine (partial).	
MOD_RES	978	978		Phosphotyrosine.	
MOD_RES	991	991		Phosphoserine.	
MOD_RES	1026	1026		Phosphoserine.	
MOD_RES	1070	1070		Phosphoserine.	
MOD_RES	1071	1071		Phosphoserine.	
MOD_RES	1092	1092		Phosphotyrosine (by autocatalysis).	
MOD_RES	1110	1110		Phosphotyrosine (by autocatalysis).	
MOD_RES	1172	1172		Phosphotyrosine (by autocatalysis).	
MOD_RES	1197	1197		Phosphotyrosine (by autocatalysis).	
CARBOHYD	56	56		N-linked (GlcNAc...)	CAR_000227
CARBOHYD	128	128		N-linked (GlcNAc...)	
CARBOHYD	175	175		N-linked (GlcNAc...)	

**Figure 4-4:**  
Top of the  
Features  
section for  
Swiss-Prot  
Entry  
P00533.

Key fields

phate binding region — from residue 718 to 726 — for binding ATP.

- ✓ **BINDING:** Here we can see the precise binding site for ATP — on lysine at position 745.
- ✓ **ACT\_SITE:** The numbers here indicate amino acids involved in the activity of an enzyme.  
To the nonspecialist, the three Key field terms listed above address fairly overlapping concepts.
- ✓ **COMPBIAS:** Extent of a compositionally biased (in this case, a serine rich) region of the protein sequence.

Next comes some information about residues that are chemically modified after translation (as shown in Figure 4-5). The corresponding keywords are:

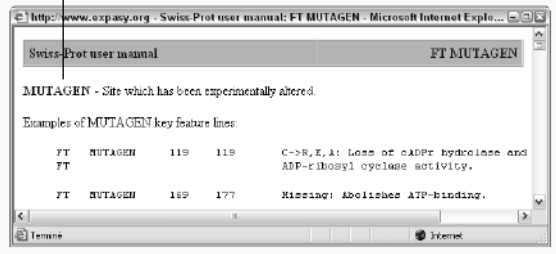
- ✓ **MOD\_RES:** The numbers here indicate residues undergoing phosphorylation. Click [MOD\\_RES](#) to find out the list of chemical modifications associated with that keyword.
- ✓ **CARBOHYD:** Here we can see the numerous residues on which carbohydrate molecules have been attached as well as the type of link (C-, N-, or O-). Note that they're all in the extracellular domain, as one would expect. Clicking the [CARBOHYD](#) keyword can tell you much more about the intricacies of glycosylation.
- ✓ **DISULPHID:** The numbers here indicate cysteine pairs forming a bond in the mature protein; there are plenty such pairs here.

Finally, the Features section ends up with a series of specialized features:

- ✓ **VAR\_SEQ:** This field is used to indicate amino-acid changes from the translation of different mRNAs (isoforms) generated by alternative splicing.
- ✓ **VARIANT:** This field identifies natural variation of the EGFR protein sequence, most of which have been associated with lung cancer.
- ✓ **MUTAGEN:** In contrast with the previous field, this field is used to record sequence changes that have been experimentally (voluntarily) introduced in the protein.
- ✓ **CONFLICT:** This feature is used to indicate discrepancies between different sources of the same protein sequence — basically errors or unrecognized polymorphisms.

VARIANT	873	873	1	G -> E (in lung cancer).	VAR_Q26101
VARIANT	962	962	1	R -> G (in dbSNP:17337451) [NCBI/Ensembl].	VAR_Q19299
VARIANT	988	988	1	M -> P (in dbSNP:173390699) [NCBI/Ensembl].	VAR_Q19300
MUTAGEN	1016	1016			
MUTAGEN	1092	1092		V->F: 50% decrease in interaction with PIK3C2B. 65% decrease in interaction with PIK3C2B; when associated with F-1197. Abolishes interaction with PIK3C2B; when associated with F-1197 and F-1092.	
MUTAGEN	1110	1110		V->F: No change in interaction with PIK3C2B. Abolishes interaction with PIK3C2B; when associated with F-1197 and F-1016.	
MUTAGEN	1172	1172		T->F: No change in interaction with PIK3C2B.	
MUTAGEN	1197	1197		V->F: No change in interaction with PIK3C2B. 65% decrease in interaction with PIK3C2B; when associated with F-1016. Abolishes interaction with PIK3C2B; when associated with F-1092 and F-1016.	
CONFLICT	540	540		N -> K (in Ref: 1).	
STRAND	30	31	2		
STRAND	34	34	1		
TURN	37	38	2		
STRAND	40	40	1		
STRAND	42	43	2		
HELIX	44	55	12		
TURN	56	57	2		
STRAND	58	68	10		
TURN	72	73	2		
HELIX	77	81	5		
STRAND	84	87	4		
STRAND	89	93	5		
STRAND	97	98	2		
TURN	102	103	2		
STRAND	106	107	2		
STRAND	110	111	2		
STRAND	113	113	1		
TURN	114	116	3		
STRAND	117	122	6		
STRAND	125	129	5		

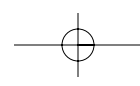
**Figure 4-5:**  
The bottom  
of the  
Features  
section for  
entry  
P00533.



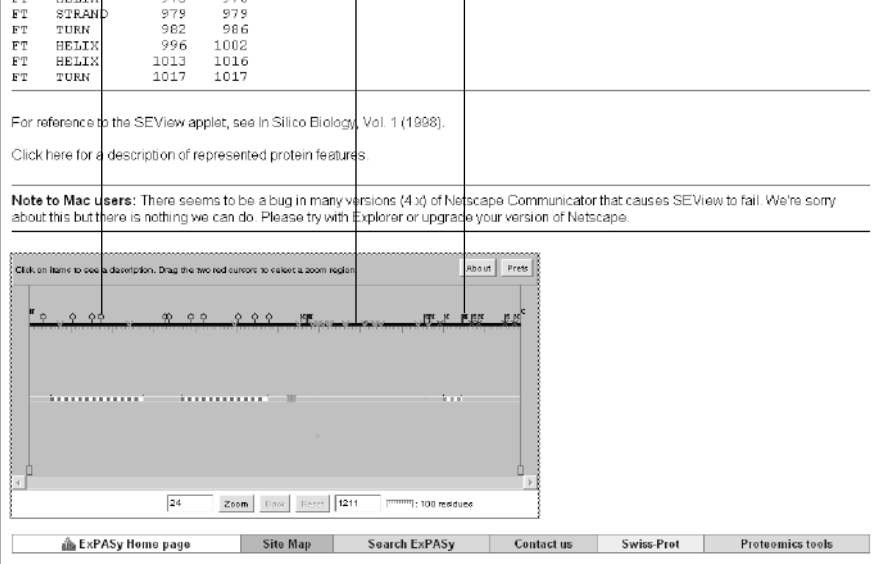
3-D structural information

Following these features lines, we enter into a detailed linear description of the local shapes (called the *secondary structure*) of the backbone of the protein. Secondary structures come in three flavors: STRAND (an extended, stringy shape), TURN (as it says!), and HELIX (a twisted, springlike shape).

Before leaving this long Features section, look along the left side for the Feature Table Viewer icon (refer to Figure 4-4, top left). Click the icon and scroll down the new page that appears to see a little graphic display, as shown in Figure 4-6. It gives you a global picture of the EGFR protein primary structure. Mousing over this picture tells you what feature corresponds to each symbol. With P00533, this shows you at a glance that all the phosphorylation events take place in the intracellular domain — while all the other modifications (disulphide bridges and glycosylations) occur in the extracellular domain. If you've done a bit of biochemistry, you know that this makes sense and is as it should be!







**Figure 4-6:**  
The Feature  
table viewer  
for Swiss-  
Prot entry  
P00533.

## *Finally, the sequence itself*

There's nothing special to say about this straightforward section. Remember that you can get the sequence in the more convenient FASTA format version by clicking the [P00533 in FASTA format](#) link, located at the bottom right of the entry. Use the File→Save As option of your browser to save it, or cut and paste it into a word processor.

## *Finding Out More about Your Protein*

After carefully reading the Swiss-Prot entry on your protein of interest (or its homologue) and using all the hyperlinks it provides, you probably know 90 percent of what the whole world knows about this protein.

In brief, there may come a time when you need additional information to better understand what you've read in Swiss-Prot. In the following sections, we point out some extremely valuable resources for our knowledge-craving readers!

## *Finding out more about modified amino acids*

If you want to find out everything you always wanted to know about the modification of amino acids during protein maturation, have a look at RESID<sup>(7)</sup>, the post-translational modification database maintained by John Garavelli at the European Bioinformatics Institute. For instance, you can use this resource to quickly find out what myristylation is exactly. Here's how:

**1. Point your browser to [www.ebi.ac.uk/RESID/](http://www.ebi.ac.uk/RESID/).**

The RESID Database of Protein Modifications page appears.

**2. Click the Search RESID link.**

A search form appears.

**3. Enter myristoylation in the search window (refer to Figure 4-7).**

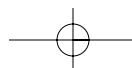
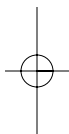
**4. Click the Search button.**

The query returns a table with 3 entries: AA0059 (N-myristoyl-glycine), AA0078 (N6-myristoyl-L-lysine), and AA0307 (S-myristoyl-L-cysteine)

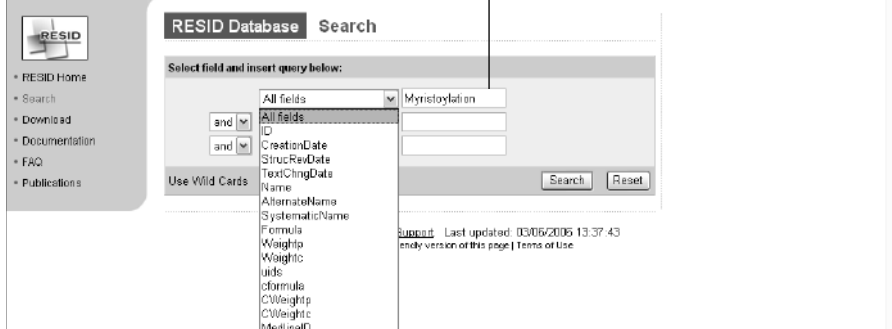
**5. Click the AA0078 link in the RESID ID leftmost column.**

You obtain a complete report on this modified amino acid, ending with its detailed chemical formula.

Not impressed with this one? Why not have a go at the modified cysteine — L-cysteinyl molybdopterin guanine dinucleotide (AA0281)? Now, *there's* a compound name to conjure with; reading it aloud several times can only help the processing of your query!



**Figure 4-7:**  
Querying  
the RESID  
database for  
myristoylation.



## *Some advanced biochemistry sites*

If you want to brush up your biochemistry and chemistry skills, visiting the following sites could help:

- ✓ **The Glycan Structure Database** at [www.glycosuite.com/](http://www.glycosuite.com/)  
This one is free for academics only — and registration is required.
- ✓ **The Lipid Bank** at [lipidbank.jp](http://lipidbank.jp)
- ✓ **ChemIDplus** at [chem.sis.nlm.nih.gov/chemidplus/](http://chem.sis.nlm.nih.gov/chemidplus/)  
This one is fun; you can query the databank by drawing the molecule of your dreams!



Be aware that using these databases is tricky for nonspecialists. We only introduce them here so you know they exist. There is some brushing up on biochemical paths to do before you can search them in a sensible way.

## *Finding out more about biochemical pathways*

Reading that your protein is involved in the biosynthesis of di-amino-pymelate probably rings no bell at all! To find out which biochemical pathway (a succession of elementary reactions leading to a compound central to the cell's well being) your protein belongs to, pay a visit to the Boehringer site.

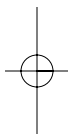
**2. Type pimelate in the Keyword Search window and then click the Submit Query button.**

Your browser displays the Pimelate page. Pimelate — a central molecule for the creation of bacterial cell walls — is just an example here. If you like, you can substitute the name of the compound you're working on at the moment.

**3. Click J3 below 6-DIAMINOPIMELATE, the first entry in the list.**

This brings up a local chart where you can see that this compound is central to the biosynthesis of bacterial cell walls.

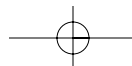
Table 4-1 lists more great biochemical-pathway resources.



<b>Table 4-1 Biochemical Pathways Resources Available Online</b>	
<i>Address</i>	<i>Description</i>
www.genome.ad.jp/kegg/	The famous Kyoto Encyclopedia of Genes and Genomes (KEGG). E.C. numbers or gene names are the best starting points for this resource.
brenda.bc.uni-koeln.de	The Comprehensive Enzyme Information System BRENDA is a must if you plan to get into real experiments!
www.chem.qmul.ac.uk/iubmb/	The official site for Enzyme Nomenclature of the International Union of Biochemistry and Molecular Biology (IUBMB).
www.ecocyc.org	The Encyclopedia of E. coli Genes and Metabolism. Now extending progressively to other bacteria.

## *Finding out more about protein structures*

If you have analyzed your protein at the sequence level by all available means (motif search in Chapter 6, database search in Chapter 7, multiple alignment



**Table 4-2 Databases Dedicated to Structural Information**

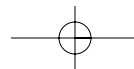
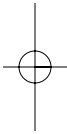
<i>Address</i>	<i>Description</i>
www.rcsb.org/pdb/	PDB, the official repository database for protein 3-D structures.
www.ncbi.nlm.nih.gov/Structure/	with tools for their visualization and comparative analysis.
scop.mrc-lmb.cam.ac.uk/scop/	SCOP, a Structural Classification Of Proteins, aims to provide descriptions of the structural and evolutionary relationships among all proteins whose structure is known.
www.cathdb.info	CATH ( <i>C</i> lass, <i>A</i> rchitecture, <i>T</i> opology, <i>H</i> omologous superfamily) also provides a hierarchical classification of protein-domain structures.
swissmodel.expasy.org	Swiss-Model is a fully automated server that models protein-structure homology, accessible via the ExPASy server.

## *Finding out more about major protein families*

Some protein families are so hot they've become whole worlds of their own. There is always a time when general protein databases just can't manage the sheer amount of detailed information required by the specialized research community. Highly specialized information resources exist to fulfill that niche. If your favorite protein belongs to one of the following categories, you may want to try some of the resources that we list in Table 4-3.

histocompatibility complex molecules of all vertebrate species.

<a href="http://rebase.neb.com">rebase.neb.com</a>	REBASE is a restriction/modification enzyme database.
<a href="http://www.cazy.org/CAZY/">www.cazy.org/CAZY/</a>	CAZy is an information resource on enzymes that degrade, modify, or create glycosidic bonds.
<a href="http://merops.sanger.ac.uk">merops.sanger.ac.uk</a>	MEROPS is a database providing a wealth of information on proteases.
<a href="http://www.kinaset.org/pkr/">www.kinaset.org/pkr/</a>	PKR, the Protein Kinase Resource, is a Web-accessible compendium of information on the protein kinase family of enzymes.
<a href="http://www.nursa.org">www.nursa.org</a>	Nursa, the Nuclear Receptor Signaling Atlas, is a great information resource on nuclear (for example, steroid) receptors together with their coactivators, corepressors, and ligands.
<a href="http://senselab.med.yale.edu/senselab">senselab.med.yale.edu/senselab</a>	The Human Brain Database provides information on the proteins involved in neural processes, such as ion channels, membrane receptors of neurotransmitters and neuromodulators, and olfactory receptors.
<a href="http://www.ncbi.nlm.nih.gov/COG">www.ncbi.nlm.nih.gov/COG</a>	The COG (Clusters of Orthologous Groups) database regroups proteins shared by at least three major phylogenetic lineages, thus corresponding to ancient conserved domains.



# Working with a Single DNA Sequence

---

## *In This Chapter*

- ▶ Identifying problems with your sequence
  - ▶ Computing and displaying a restriction map
  - ▶ Profiling your sequence for a variety of properties
  - ▶ Finding ORFs, exons, and genes
  - ▶ Assembling sequence fragments
- 

*Not everything that can be counted counts, and not everything that counts can be counted.*

— Albert Einstein (1879–1955)

**p**roteins perform most biological functions, so biologists tend to consider DNA sequences kind of dull. They're mostly right. The truth of the matter is, 90 percent of the DNA you may use in your work is nothing more than an information-storing device, sort of the organic equivalent of a floppy disk or a DVD. Still, we probably shouldn't turn up our noses at information-storing devices. DNA contains information that the cell knows how to put to effective use, as quickly as your computer can turn a dull-looking DVD into the latest destroy-all-at-any-cost network game.

If you're interested in a protein, directly working out its amino-acid sequence is hell. Sequencing its gene and then deducing the protein sequence from the genetic code is much easier. This is why DNA sequences are so important. These days, working with DNA sequences is an obligatory intermediate step; you simply must know how to handle a nucleotide sequence in order to work with protein or RNA sequences. DNA sequences are the mothers of all sequences!

copy of a messenger RNA — is the name of the game in this chapter. This is the kind of sequence you can routinely determine in a well-equipped molecular biology lab or get done for a reasonable price by a sequencing company.

When you get an experimental sequence back, your first task is to make sure it's okay so you don't waste a month trying to make sense of something completely botched and end up nowhere. This chapter begins with a section showing you how to check for the most common problems encountered in sequences fresh from the bench. But remember Murphy's Law: If something can go wrong, *and has not been discussed in this section*, it will happen to you!

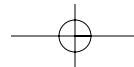
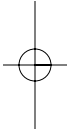
## *Catching Errors Before It's Too Late*

Sequencing a DNA fragment involves purifying it, cloning it into a vector (such as a plasmid), amplifying it into a biological host (most often a bacterium like *E. coli*), and finally submitting it to one of the various sequencing protocols, such as primer extension or dye-termination. During this process, accessory DNA segments are deliberately linked to your target DNA. In addition, many unintended events can occur, resulting in a sequence that doesn't faithfully correspond to the genetic information you intended to study.

The most common problem is that the sequence you get back has been contaminated with some of the sequence of the vector you used (or that somebody else used in the laboratory). Recognizing this problem — and knowing how to deal with it — is so important that we devote the next section to precisely this topic.

### *Removing vector sequences*

The DNA (or cDNA) you send out for sequencing is inserted into a cloning vector — plasmid, phage, cosmid, BAC, PAC, or YAC — so that it can be manipulated. The sequences you get from such constructs inevitably include segments derived from these vectors, at least at one extremity or the other. You can detect and remove them from your sequence by simply running a search for similarity against the sequence of the vector you have been using. (As a responsible scientist, you're expected to have this information!)





1. **Point your browser to** [www.ncbi.nlm.nih.gov/VecScreen/VecScreen\\_docs.html](http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen_docs.html).

The entry page for VecScreen appears. It contains both a nice tutorial on sequence contamination (click the [Contamination](#) link to read it) as well as an explanation of how VecScreen works. Basically, it performs a blastn search (see Chapter 7) of your sequence against a special database of all vectors and cloning adapter sequences known to man. This database is known as UniVec, and a simple click the [UniVec](#) link on this page provides you with more details about this unique database.

When you believe you're ready for the real thing, rather than just reading about the real thing, go grab your new sequence and proceed to Step 2.

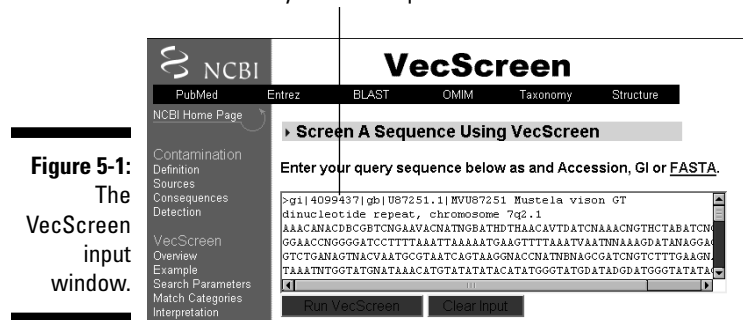
2. **Point your browser to** [www.ncbi.nlm.nih.gov/VecScreen/](http://www.ncbi.nlm.nih.gov/VecScreen/).

A typical sequence-input window appears, ready for your input.

3. **Paste in your newly determined sequence in the input window, as shown in Figure 5-1.**


You can use FASTA or raw format.

Paste in your new sequence.



4. **Click the Run VecScreen button below the input window.**

You immediately get back the BLAST formatting page, showing you that your VecScreen request was in fact a BLAST search in disguise. (For more on BLAST, see Chapter 7.)



At this point — and depending on the workload of the BLAST server — you'll either get the standard waiting page (This page will be automatically updated in xx seconds), or the results of your search pop up immediately.

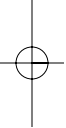
## 6. The two possible outcomes are

- **Non-significant similarity found.**

*This is good news!* This message indicates that the sequence you submitted does not resemble any known vector/adaptor. You can proceed to the next stage of its analysis, following the rest of this chapter.

- **An output listing matches of some kind.**

*This is bad news!* Your query does contain some vector/query sequence.



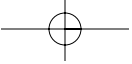
A VecScreen contamination report output contains two parts. The top part displays the distribution of the contaminating sequences along the length of your query sequence, using a color-coded graph (Figure 5-2 — you'll have to go there online to see it in color). The second part is a standard blastn output, providing you with the identity of the matched vector/adaptor as well as the detailed alignment.

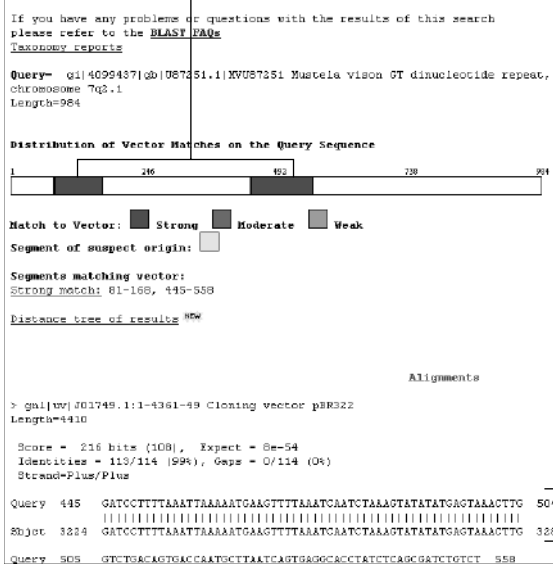
The different colors on the graph correspond to the strong (red), moderate (purple), or weak (green) matches in the UniVec database. A yellow code is used to delineate segments of suspect origin such as short segments in between two vector matches. The U87251 exhibits two regions of 100 percent identity with the popular PBR322 cloning vector. Most likely, it is a contaminated sequence that was submitted to GenBank before vector-sequence filtering was systematically included in the submission process.

What to do next with contaminated sequences is fully explained in a complete tutorial that you can follow by clicking the [Interpretation of VecScreen Results](#) link provided at the top of any VecScreen positive output.



In general, you have two options when finding contaminations. If the contaminated parts are at the extremity of your sequence and correspond to the vector you used, all is fine. Just remove these parts and proceed. If, on the other hand, contaminated parts are found elsewhere (as in Figure 5-2), or if the contaminating vector is not yours, you'd better consider this sequence to be the result of a chimeric clone or a DNA cross-contamination. In that case, throwing it away is the safest thing to do!





**Figure 5-2:**  
VecScreen  
output  
display of a  
contaminated  
sequence.

Blast N sequence alignment

## Cases when you shouldn't discard your sequence

In case of a VecScreen match, do not merely rely on the name of the UniVec database entry to decide that your experiment was contaminated by somebody else's vector. You may have used pUC19 as a cloning vector and see that VecScreen is reporting a strong match with pUR222 or lactose operon genes from *E. coli*. This is okay. Most commercial vectors are derived from the same initial natural plasmid and *E. coli* constructs. Their sequences must thus be 100 percent identical. The UniVec matches are simply reported in the order that they occur in the database. If your vector is not first, another entry (corresponding to a locally 100-percent-identical sequence) will be reported instead.

Also, if you're working on genes that resemble the kinds of genes used to build vectors (such as antibiotic resistance genes, repressor/activator genes, or some biosynthetic genes), you should actually *expect* matches in the UniVec database.

# Computing/Verifying a Restriction Map

Here's another way of checking whether the sequence you get back from the sequencing lab or company actually corresponds to the piece of DNA you sent out: Compare the pattern of restriction enzyme digestion (also called a *restriction map*) that you experimentally observe with the pattern (*theoretical restriction map*) that you can compute from the sequence.

Computing a theoretical restriction map is easy. All you do is look for exact matches of a given restriction-enzyme site within your sequence. All the known enzymes and sites are described in the REBASE database ([rebase.neb.com](http://rebase.neb.com)). After locating the matches in the sequence, it's simply a matter of counting the number of resulting fragments (DNA segments between two successive sites) and computing their sizes. You can then compare the computer prediction with your lab experiment. The more restriction enzymes you test, the more certain you'll be that your sequence is correct.

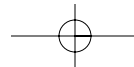
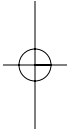
You can use this approach to verify long sequences — such as bacterial genome sequences — that were assembled from many shorter ones, as happens with the Shotgun DNA sequencing protocol. If the sequence and its assembly are correct, the number and size of the predicted fragments should perfectly correspond to the experimental restriction map.

A couple of good Internet sites spare you all the work of having to collect the information about restriction-enzyme sites and doing the matching yourself. These sites maintain their own current versions of the authoritative Restriction Enzyme Database, REBASE, so you don't have to worry about it. They scan your sequence for all the specific sites corresponding to an enzyme list that you can select according to various criteria.

One such Internet site we find especially useful offers a nice little interactive tool called Webcutter, developed by Max Heiman. Here's how to put it to work for you:

1. **Point your browser to** [www.firstmarket.com/cutter/cut2.html](http://www.firstmarket.com/cutter/cut2.html).

This gives you access to the most current version of Webcutter.





- Upload the sequence from your computer (Netscape users only). The sequence has to be in a text format, without its name or any FASTA-style header.
- Directly fetch a GenBank entry by using its accession number or other identification methods. (The latter case applies if you're designing a cloning experiment for a gene already known.)

### 3. Select the options that correspond to your problem.

Options include

- Choosing whether to treat the input sequence as linear or circular
- Selecting various output formats for the results (maps, table of enzymes)
- Filtering the results based on whether enzymes cut or don't cut your sequence — or cut your sequence a given number of times
- Specifying a subset of enzymes — based on various criteria — to be included in the analysis

### 4. Click the Analyze Sequence button at the bottom of the form.

The result usually arrives very quickly and is — thankfully — self-explanatory.

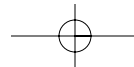
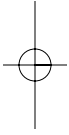
The University of Massachusetts Medical School offers a similar service with a different interface at [biotools.umassmed.edu/tacg/WWWtacg.php](http://biotools.umassmed.edu/tacg/WWWtacg.php).



For a list of other restriction mapping servers, check out the [REBASE Related Sites](#) link on the REBASE home page ([rebase.neb.com](http://rebase.neb.com)). All these servers are equivalent in terms of accuracy. Shop around until you find the one whose input/output format you like best.

## *Designing PCR Primers*

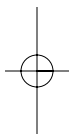
The polymerase chain reaction (PCR) is a very common experimental laboratory technique people use to amplify segments of DNA they find interesting. To make a PCR, you must



sequence that you mean to amplify), the primers, a cocktail of nucleotides and other biochemical compounds, and a heat resistant enzyme called DNA polymerase — all in a single little plastic tube. You then put that tube in a benchtop expensive machine called a *thermal cycler*. According to a preset program, this machine will make the tube go up and down in temperature. For each temperature cycle, the

30 cycles, you have  $2^{30}$  (1 billion) more of it. This is why PCR is such a great technique in forensic science: Lick a stamp, and scientists will be able to sequence your DNA!

If you want to know more about PCR, visit [nature.umesci.maine.edu/forensics/p\\_intro.htm](http://nature.umesci.maine.edu/forensics/p_intro.htm) for a great animated introductory course.



**1. Identify the DNA sequence you want to amplify.**

**2. Order primers from a DNA synthesis company.**

*Primers* are small pieces of DNA (20–30 bases) that match the extremities of the complete sequence you're interested in. Designing good primers is the most delicate step in a PCR.

**3. Run your PCR experiment.**

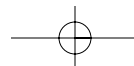
These days, people use PCR for almost anything that has to do with wet-lab sequence analysis — subcloning a gene in an expression vector, verifying its assembly and sequence, inducing or detecting mutations, or detecting its relatives in other organisms.

The trickiest step when you do a PCR is the design of the *primers* — the two small DNA fragments capable of firmly hybridizing on each side of your gene in a highly specific manner. Primer design programs are here to help you decide which portion of your large sequence makes the best primers, thus avoiding a series of potential problems too numerous to be listed here.

The Internet site of the University of Massachusetts Medical School ([biotools.umassmed.edu/](http://biotools.umassmed.edu/)) provides a link to a very complete and easy-to-use tool for primer design. It is a Web implementation of Primer3, the well-known program developed by Steve Rosen and Helen Skaletsky. To put it to (hopefully good) use, do the following:

**1. Point your browser to [biotools.umassmed.edu/](http://biotools.umassmed.edu/).**

The Bioinformatics Resources page of the University of Massachusetts Medical School appears, offering access to several sequence-analysis programs that you may want to try by yourself later.



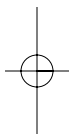
### 3. Paste your sequence in the sequence window.

### 4. Click the Pick Primers button.

The results return very quickly. They include a map of the best oligonucleotide pair (left and right primers) according to the constraints listed in the form. In addition, four alternative solutions (by default) are proposed. The position, length, predicted melting temperature, and G+C percentage (the fraction of guanine and cytosine nucleotides) are given for each listed oligonucleotide.

If you're ready to modify the default setting, the form allows all reasonable experimental situations and constraints to be imposed upon the primer-selection process, including

- ✓ Searching for only a left or right primer, or a single hybridization probe
- ✓ Proposing your own left or right primer
- ✓ Selecting sequence positions to be flanked or excluded
- ✓ Selecting a range of product sizes



Paste your sequence here.

Primer3: WWW primer tool

pick primers from a DNA sequence [cautions](#)

Paste source sequence below (5' to 3', string of ACGTNaactm -- other letters treated as N -- numbers and blanks ignored). FASTA format ok. Please N-out undesirable sequence (vector, ALUs, LINEs, etc.) or use a [Mismatching Library \(repeat library\)](#). NONE

TGTATTCTTTTATCATGATGAAGGAAAAGTCCACTCTCCGGCAAGCTTTTATATTCGTCGCTGTTATT  
CCATATTAGAGGGTCTTGGCTTGAATTTAGAATTCGACGGTAAGGATGTGATTTATTCAGGATTGATAGAA  
AAAGAAAGCTTTATACTACTACTTTGCTTAGAGCTATAGGTATGAGTACTGAAGAAATTATAAAATTTTA  
TTATAATTCAGTAACTTATAAGCTTGTAAAAATAAAGTTGGGCGGTTAAATTTATACCTCAGCATATT  
ACTGCTCATCGTTTAAACAAGTGTATTAGTAGATGCAGATACCGGAAATATTCCTACTGAAAGCAGGACAAA  
AAATTAACCCGGCTTTGGCTAAAAAATATTCTGGGGAAGGGCTTAAATAATTTTGTAGTCAATGAAAC  
TTTAATCGGCAATATCTATCCGARGATTTAAGAGATCCTGCAAGCGATGAAGTATTAGCAAAAATCGGT

Pick left primer or use left primer below  Pick hybridization probe (internal oligo) or use oligo below  Pick right primer or use right primer below (5' to 3' on opposite strand)

Pick Primers Reset Form

A string to identify your output

E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the with [ and ]  
e.g. ...ATCT[CCCC]TCAT... means that primers must flank the central CCCC

E.g. 401,7 68,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 68. Or mark the  
with < and > e.g. ...ATCT<CCCC>TCAT... forbids primers in the central CCCC

Min 100 Opt 200 Max 1000

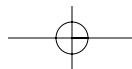
5 3.0

12.00 24.00

Pick Primers Reset Form

Figure 5-3:  
Primer3  
input form.

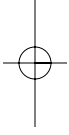
Click here to get your primers.



they will not hybridize anywhere — except where you *intend* them to hybridize. For instance, you want to make sure that the selected oligonucleotide sequences aren't found outside the gene you're interested in — or (worse) resemble a frequent repeat in the DNA you're going to amplify.

The techniques for avoiding these problems involve BLAST searches against the vector sequences, the relevant genomes as well as their most common repeats. You can use the NCBI BLAST server for this purpose. (For more on BLAST searches, see Chapter 7.)

## Analyzing DNA Composition

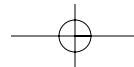


After these analyses — with all the going back and forth from computer to bench they entail — you're now convinced that the sequence in your test tube is indeed the right stuff. The time has now come to analyze it! Whenever you have a new piece of a new genome, the first question people will ask you is: *What is the percentage of G+C?* Because the pairing between the guanosine (G) and cytosine (C) nucleotides is more stable than the pairing between adenosine (A) and thymidine (T) in the DNA ladder, this percentage determines how the DNA will behave in your experiments. The first level of analysis you can perform is, thus, simply to count the number of A (adenosine), G (guanosine), C (cytosine), and T (thymidine) in your sequence.

### *Establishing the G+C content of your sequence*

Most often, people establish G+C content by using a program installed on their own computers. Many programs that do this are commercial packages, available from an open source consortium (such as EMBOSS), or are written up by the computer science student you keep handy and pay with pizza. If you want to save the cost of the pizza, you can even do the coding yourself if you've learned the basics of one of the basic computer-programming languages out there.

However, if you want to use an online service to establish the G+C content of your sequence for you, check out the Pasteur Institute's EMBOSS server (at





## Counting words in DNA sequences

After establishing the G+C content, the next step in complexity is to consider your DNA text as made up of overlapping “words.” For instance, consider the following DNA sequence:

```
ATGGCTGACTGACTGACTGACTGAC . . . . .
```

You can read it as

```
A, T, G, G, C, T, G, A, C, T, G, A, C, T, G, A, . . . . .
```

where counting the various components A, C, G, and T leads to the simple *nucleotide* statistics. But you can also read the sequence as

```
AT, TG, GG, GC, CT, TG, GA, AC, CT, TG, . . . .
```

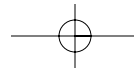
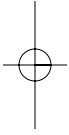
and compute the frequency of the 16 different *dinucleotides* (2-letter words). Then again, you can read the same sequence as made of triplets — like this —

```
ATG, TGG, GGC, GCT, CTG, TGA, GAC, ACT, CTG, . . .
```

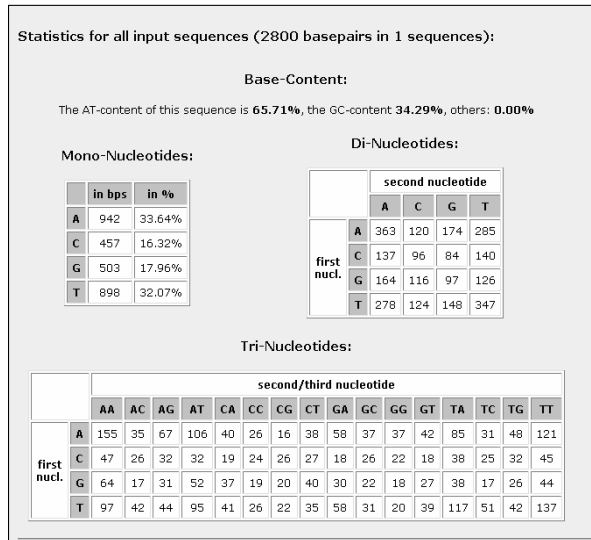
and compute the frequency of the 64 different *trinucleotides* (3-letter words), and so on.

The German-based company Genomatix offers a nice free Web site where you can compute the nucleotide, dinucleotide, and trinucleotide frequencies of your DNA sequence. Here’s how it’s done:

- 1. Prepare your DNA sequence in a file, using a text format.**
- 2. Point your browser to** [www.genomatix.de/cgi-bin/tools/tools.pl](http://www.genomatix.de/cgi-bin/tools/tools.pl).
- 3. Select the Create Sequence Statistics button.**
- 4. Click the Start Selected Task button at the bottom left of the page.**  
A Sequence Input window appears.
- 5. Cut and paste your sequence into the Sequence Input window.**  
Use the format of your choice. (Here we use the *plain* — that is, raw — format.)



the nucleotides, dinucleotides, and trinucleotides frequencies and presents the results, as shown in Figure 5-4.

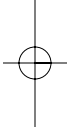


**Figure 5-4:**  
 Typical  
 compositional  
 analysis  
 result.

## Counting long words in DNA sequences

When analyzing DNA, it can be interesting to count words that are longer than three nucleotides, although it is rare to use words with sizes above six or eight. When their size is larger than 3, words are often referred to as k-tuples (6-letter words = hexamers = 6-tuple). There are 64 different 3-tuples, 256 4-tuples, 1,024 5-tuples, and 4,096 6-tuples. It makes no sense to compute the hexamer statistics on a short sequence where most hexamers occur, at most, only once. On the other hand, identifying hexamers (6-tuples) with unexpected high frequencies in a set of sequences (such as promoter regions) is often the starting point for discovering regulatory sequence motifs.

The EMBOSS server at the Pasteur Institute, based in Paris, offers an online version of the program wordcount that allows you to compute the word frequency in your DNA sequence for any size (we tested it up to 20 letters). Here's how you get it to work for you:



These computations are always done by reading the DNA sequence on one strand, in an overlapping manner. This should not be confused with the computation of codon usage. (Remember that a *codon* is a triplet of nucleotides used in the context of the genetic code, when translating a DNA sequence into a protein; see Chapter 1.) The concept of codon usage thus only applies to DNA sequences corresponding to protein coding regions (for example, open reading frames or ORFs). It consists in computing the frequency of trinucleotides in a non-overlapping manner, mimicking the protein translation process. Hence, for a DNA sequence such as

```
ATGTTTAGTGATGGACGCCAGCATGC
GAC . . . .
```

```
ATG, TTT, AGT, GAT, GGA, CGC,
CCA, GCA, TGC, GAC, . . .
```

The resulting table of codon usage tells you which codons are preferentially used to code for a given amino acid — and, more importantly, to determine which ones are rare. Codon usage varies among species. For instance, the codon usage in human genes is quite different from the one used by *E. coli*. For this reason, some human genes don't work well in these bacteria. Take a virtual trip to Japan at

```
www.kazusa.or.jp/codon/
countcodon.html
```

to test a simple codon usage program on your DNA sequence.

**1. Point your browser to [bioweb.pasteur.fr](http://bioweb.pasteur.fr).**

Click the English/American flag (top right) for the English version.

**2. Click the [EMBOSS](#) link at the bottom-left of the page.**

The list of EMBOSS modules appears.

**3. Click the [wordcount](#) link.**

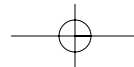
The wordcount page appears in all its glory.

**4. Enter your e-mail address and paste your sequence (raw format is okay) in the appropriate fields. Be sure to enter the word size you're interested in as well.**

**5. Click the Run wordcount button at the top-left of the page.**

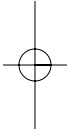
After a few seconds, a result page appears.

**6. Click the [wordcount.out](#) link to get your results as a simple list.**



## ***Finding internal repeats in your sequence***

Another useful type of composition analysis involves locating segments that occur more than once within your sequence. Such segments are called *repeats*. There is no real difference between long words (6-tuple, 8-tuple, and larger) and repeats. Here's a common-sense rule: *A repeat is a word long enough so that it's unlikely to occur very often by chance, given a random sequence.* For instance, a GTC triplet found 4 times within a 500-nucleotide long sequence doesn't qualify as a repeat.



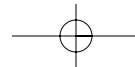
Another difference between word-counting and repeat analysis is that repeats can be imperfect. Unlike words, similar repeats don't need to be identical. Finally, biologists like to distinguish *tandem repeats* (similar subsequences along the same DNA strand) from *inverted repeats* (similar sequences occurring on the direct and reverse strands). Biologists are interested in repeats because they are often involved in genome rearrangements or regulatory mechanisms of gene expression.

There are many different algorithms for finding repeats within a DNA (or protein) sequence. They all try to identify segments more similar to each other than would be expected by mere chance alone. The tricky part is in the scoring and ranking of the similar subsequence segments. Is the exact matching of five consecutive nucleotides good enough to be considered a repeat? And is it better than 9 out of 10, or 123 over 160? Which one do you want reported first? How far down the list of possible repeats do you want to go?

### ***Finding repeats is a tricky business***

Because there are no universal answers to the questions surrounding the precise nature of repeats, repeat-finding programs ask you to fix thresholds related to their scoring algorithms, repeat size, copy number, periodicity (distance between repeats), and other things you don't always understand. This makes them difficult to use. The default settings they provide may or may not work for your particular sequence.

In that respect, our quick survey of Web-based repeat finders — using a repeat-containing sequence we made just for this purpose — was amazingly



all these difficulties. It is the *dot-plot* approach, which we show you in the following steps list. (For more on dot plots, see Chapter 8.)

For long nucleic acid sequences, you can use the dot-plot program provided by the Molecular Toolkit Web service at Colorado State University. Just do the following:

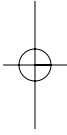
1. **Point your browser to** [arbl.cvmbs.colostate.edu/molkit/](http://arbl.cvmbs.colostate.edu/molkit/).

The Molecular Toolkit is a group of programs for analysis and manipulation of nucleic acid and protein sequence data. The programs are written in Java (1.0), and require that your browser support this language.

This site is very useful if you need to perform some simple DNA sequence manipulation, such as reverse complementation (changing strands), protein translation, or if you want to display a restriction map. For now, continue through these steps.

2. **Click the [Dot Plots](#) link at the top of the list.**

You obtain a straightforward input form, with two side-by-side boxes titled DNA Number 1 and DNA Number 2, as shown in Figure 5-5.



**Figure 5-5:**  
Input form  
of the DNA  
dot-plot  
program.

3481	TTAATAGTGCGAATTTTACTTCTAGAGAATT	3481	TTAATAGTGCGAATTTTACTTCTAGAGAATT
3541	GCAGAGAAAATTTTCAGAGAATTAGACAT	3541	GCAGAGAAAATTTTCAGAGAATTAGACAT
3601	TATTTGTGTGAAGGTATTAAGCCACAGGATA	3601	TATTTGTGTGAAGGTATTAAGCCACAGGATA
3661	GATAAAGAAAATCTAATAGACGATGAAAAAT	3661	GATAAAGAAAATCTAATAGACGATGAAAAAT
3721	GATTTTGACTATACAGATCTTTTTAAATCT	3721	GATTTTGACTATACAGATCTTTTTAAATCT
3781	TGTATTTATGGCCATAAGCAAGATAAGGATAG	3781	TGTATTTATGGCCATAAGCAAGATAAGGATAG
3841	AACCGTCATTGCGAGGAGCGAAGCGACGTGGC	3841	AACCGTCATTGCGAGGAGCGAAGCGACGTGGC
3901	CGCTCTTTCAGTCCCTGGCAATGACGATTT	3901	CGCTCTTTCAGTCCCTGGCAATGACGATTT
3961	ATTTCTSAATGAATATTTTAAAGCCAAA	3961	ATTTCTSAATGAATATTTTAAAGCCAAA
4021	TATACGGATTTTAAATATAGTACAAAAGAGTGC	4021	TATACGGATTTTAAATATAGTACAAAAGAGTGC

Copy DNA1 -> DNA2      Clear DNAs & Plot      Demo DNA -> DNA1

Make Plot      Window Size 9      Mismatch Limit 0

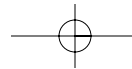
3. **Copy your sequence from a .txt file or a Word document, by using Ctrl+C or the Copy button on the Word toolbar.**

4. **Paste your sequence in the DNA Number 1 window by using Ctrl+V or Word's Paste button.**

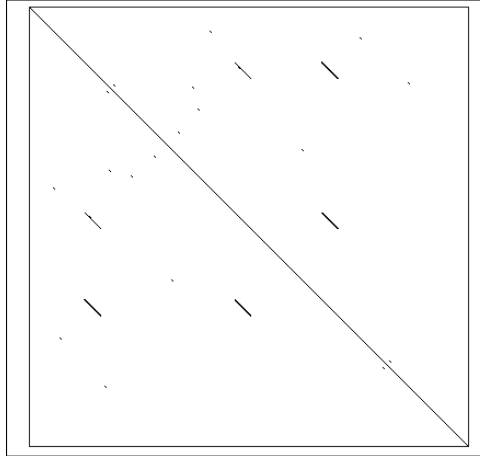
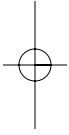
The program is happy with the raw sequence format (nucleotide only).

5. **Copy your sequence into the DNA Number 2 window as well.**

For this, you can use the special Copy DNA1 -> DNA2 button provided below the DNA Number 1 window.



In this case, you can get the dot plot quickly — in real time. It appears in the square below the input windows. For the example in Figure 5-6, we made up a sequence in which a large segment is imperfectly repeated three times. These repeats show up as three lines both in the upper and lower diagonal parts of the plot. The locations of the repeated sequence segments are found by projecting them onto the main diagonal (see Figure 5-6). The long repeats stand out in comparison to non-significant, shorter, random repeats of length 9 to 15 (which show up as mere dots). Using your own sequence, you can experiment with the Window Size parameters (minimal size of the repeat) as well as the Mismatch Limit parameters (number of non-identical nucleotides) to see how they influence the graph appearance.

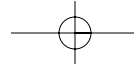


**Figure 5-6:**  
Dot plot  
showing a  
segment  
repeated  
three times.

### *Using dot plots to identify inverted repeats*



It's important to remember that you can also use a dot-plot program to search for inverted repeats — local similarities between the *direct* and *reverse* strand of a DNA sequence. To do this, you simply have to use the sequence along the X-axis (DNA Number 1), and its *reverse-complement* along the Y-axis (DNA Number 2). Use the Manipulate Sequences option at [arbl.cvmbs.colostate.edu/molkit/](http://arbl.cvmbs.colostate.edu/molkit/) to generate a sequence's *reverse-complement*. The only difference with the previous protocol is that the main diagonal disappears. Inverted repeats will show up again as off-diagonal segments.

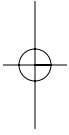




identification of a repeat from a pre-established list of repeats, like the Alu repeat family in the human genome. Discovering a new repeat has to do with the internal structure of your sequence; identifying an established one is simply a matter of recognizing some local similarities between your sequence and a predefined reference database of repeats.

Repeated elements mostly occur in multicellular organisms. Vertebrate genomes are particularly rich in many families of repeated elements of various sizes. You can get a fairly complete picture of this topic by visiting the Rebase Web site of the Genetic Information Research Institute at [www.girinst.org](http://www.girinst.org). This site gives you access to the CENSOR software tool, which screens query sequences against a reference collection of repeats, “censors” (that is, masks) matching portions with special symbols, and also generates a report on found repeats.

A similar service is offered at [www.repeatmasker.org](http://www.repeatmasker.org) on a server based at the Institute for Systems Biology. Note that the program will mask, on average, a whopping 50% of the human genome sequence!

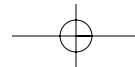


## ***Finding Protein-Coding Regions***

Get ready — this is where the fun starts! We’ve checked our DNA sequence for contamination, verified its restriction maps, computed various composition statistics, and identified its repeats. It’s time to move to the really exciting stuff! We can search to find whether and where a protein is encoded in that obscure ATGCTACG gobbledygook.

To understand what is known as *protein discovery*, you must remember that protein-coding genes have vastly different structures in microbes and multicellular organisms. In microbes, each protein is encoded by a simple DNA segment — from start to end — called an open reading frame (ORF). In animals and plant genes (vertebrates are the worst), proteins are encoded in several pieces called *exons*, separated by non-coding DNA segments called *introns*. That explains why the methods and programs used for finding proteins in microbes and higher eukaryotes are rather different.

This section starts with a simple strategy called *ORFing*, which shows you how to find protein-coding regions in microbial DNA sequences or eukaryotic mRNA sequences.



tion of an open reading frame. Given that proteins have an average length of about 350 residues (and that very few of them are smaller than 100 residues), you can use an additional criterion — minimal size — for example, requesting that at least 300 nucleotides separate the Start from the Stop. This is what ORFing is all about. You can get some practice by using ORF Finder at NCBI (the National Center for Biotechnology Information), as we describe in the following steps list:

**1. Point your browser to** [www.ncbi.nlm.nih.gov/gorf/gorf.html](http://www.ncbi.nlm.nih.gov/gorf/gorf.html).

The Input form is made up of a simple box, where you must paste your sequence (raw format), and a pull-down menu of genetic code choices. (We're ready to bet good money that you never knew there were 16 genetic code choices. So much for the universal code!) For now, let's stick with the standard genetic code (the default option).

**2. Copy your sequence from a .txt file or a Word document.**

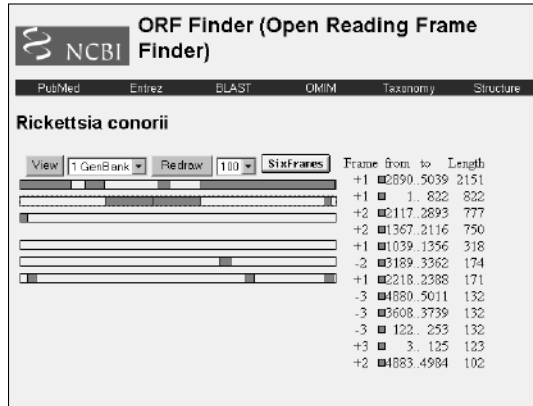
We used the first 5,000 bp of GenBank entry AE008569 (*Rickettsia conorii* genome). Alternatively, you can enter this accession number in the relevant input box, and indicate from 1 to 5,000 below the sequence input box. Then directly go to Step 4.

**3. Paste your sequence in the input box.**

The program is happy with the raw sequence format (nucleotide only).

**4. Click the OrfFind button.**

Figure 5-7 shows a typical output.



**Figure 5-7:**  
Typical  
output of the  
ORF Finder  
program.





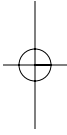
Always click the Redraw button to refresh the output.



To examine each of the ORFs more closely, click the corresponding rectangle in the graphical display or in the list to the side. This expands the output to include the predicted amino-acid sequence and some supplementary buttons, as shown in Figure 5-8. These buttons let you screen some of the NCBI databases for sequences similar to this ORF, *right on the spot!* This is a great, simple tool for ORFing your sequence.



Before getting into more complicated programs, we want to remind you that this simple ORFing program is also good for finding protein-coding regions for higher organisms *if your sequence is a cDNA*. cDNAs (the image of mRNAs) don't include introns — and they have a simple, microbe-like ORF structure. So you don't need to use another, more sophisticated program if you only want to delineate the protein-coding region within a human cDNA/mRNA sequence.



**Rickettsia conorii**

Program  Database    with parameters

View

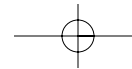
Frame	from	to	Length
+1	2890	5039	2151
+1	1	822	822
+2	2117	2893	777
+2	1367	2116	750
+1	1039	1356	318
-2	3189	3362	174
+1	2218	2388	171
-3	4880	5011	132
-3	3608	3739	132
-3	122	253	132
+3	3	125	123
+2	4883	4984	102

Length: 105 aa

```

1039 atggcaaatacgtaatcgatagctcttttaaaagggaagtatta
   H A N N V M D S S F K K E V L
1084 gaatcggattaccctgtattgggtgattttggcagaatggtgc
   E S D L P V L V D F W A E W C
1129 ggaacggttaaaatttaaacaccgataatagatgaatcagtaaa
   G P C R H L T P I I D E I S K
1174 gaattacaagcacaagttaaagcttaaaatgaatattgatgaa
   E L Q G K V K V L K H N I D E
1219 aatcctaacactcttcagaatcggatattcgtagtattccaacg
   N P N T P S E Y G I R S I P T
1264 ataatttatttaaaatggtgaacaaaaagatactaaaataggt
   I H L F K N G E Q K D T K I G
1309 ttgcaacaaaaaattctcttttagattggattaaatctatt
   L Q Q K N S L L D W I N K S I
1354 taa 1356
  
```

**Figure 5-8:** Secondary output of the NCBI ORF Finder.



the protein-coding regions you may be interested in. However, there are a variety of situations that frequently occur where you may need to use a more sophisticated approach — the approach taken by GeneMark, for example. Such situations include

- ✓ Finding very short proteins
- ✓ Resolving ambiguous cases where overlapping ORFs are predicted in different reading frames — on the direct and reverse strand, for instance
- ✓ Wanting to pinpoint the exact Start codon (the most distal ATG isn't always the correct one)



GeneMark searches for coding regions using a criterion that's a bit more sophisticated than "it has to be an uninterrupted reading frame longer than a certain length." This program also takes into account the statistical properties (very similar to word/codon usage) of your sequence and associates some sort of a probabilistic quality index to each candidate's ORFs. In the process, some small ORFs may get promoted, and the precise Start location may be redefined. The quality index measures the similarity between the candidate ORFs and an ideal gene model. This gene model is often implemented as a Markov model, a mathematical concept beyond the scope of this book. The good news for you here is that these programs are easier to use than to understand. This is what we show you in the upcoming steps. Here goes:

**1. Point your browser to** [exon.gatech.edu/GeneMark/](http://exon.gatech.edu/GeneMark/).

This is the main GeneMark home page, coming to us from Georgia Tech in Atlanta. It offers different specialized versions of the program (each corresponds to a different gene model) for working on prokaryotic (microbes), eukaryotic (animals), cDNA (the DNA version of mRNA), or virus sequences. (For the purpose of the example, let's say you want to analyze a sequence from a microbe that is little known.)

**2. In the Bacteria/Archaea section, click the Heuristic models link.**

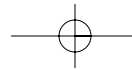
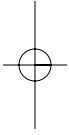
This selects the program version corresponding to the analysis of sequences from a new organism. It brings you to a simple form with the usual sequence input box.

**3. Copy your sequence from a .txt file or a Word document.**

We recommend using a microbial sequence about 5,000 bp in length.

**4. Paste the sequence in the input box.**

The program is happy with the raw sequence format (nucleotides only) as well as many others.



```

Gene Prediction Results

Information on input sequence

Sequence title: Wed Sep 25 13:56:18 EDT 2002
Length: 5040 bp
G+C percentage: 31.41 %

Parse predicted by GeneMark.hmm 2.0

GeneMark.hmm PROKARYOTIC (Version 2.1)
Sequence file name: sequence,  FBS: N
Model file name: heuristic_no_rbs.mat
Model organism: Heuristic_model
Wed Sep 25 13:56:28 2002

Predicted genes


| Gene # | Strand | LeftEnd | RightEnd | Gene Length | Class |
|--------|--------|---------|----------|-------------|-------|
| 1      | +      | 1       | 822      | 822         | 1     |
| 2      | +      | 1039    | 1356     | 318         | 1     |
| 3      | +      | 1367    | 2116     | 750         | 1     |
| 4      | +      | 2117    | 2893     | 777         | 1     |
| 5      | +      | 2890    | >5040    | 2151        | 1     |


```

**Figure 5-9:**  
Output of  
GeneMark.

If you're wondering what difference it makes to use a Markov model rather than a simple ORFing, the answer is in the differences between Figure 5-8 and Figure 5-9 — both of which relate to the same query sequence. GeneMark only retained the top five genes of the ORF Finder list!

Note that you can also run GeneMark (in the known model version) from its site at the European Bioinformatics Institute. The URL is

[www.ebi.ac.uk/genemark/](http://www.ebi.ac.uk/genemark/)

## ***Finding internal exons in vertebrate genomic sequences***

When it comes to predicting proteins from DNA, the most challenging problem around is analyzing a piece of human genomic DNA sequence. We already mentioned that vertebrate genes have a mosaic structure, where small bits of the protein are encoded by exons separated by non-coding introns. If you're looking at a human genomic sequence, your first question should be: *Do I have a protein-coding exon somewhere in there?*

According to what molecular biologists have worked out, a protein coding exon is an ORF flanked by two specific signals known as *splice sites*. Several programs exist that are meant to recognize these exons. They all work on the same principle: On a first pass, they identify candidate ORFs with good compositional properties (such as word frequency and codon usage). This is

program to work nearly as well as something like ORF Finder or GeneMark. (After all, those programs have much larger targets to shoot at.) Still, we'd like to show you what an exon detection program can do. Check out the following steps, where we use Michael Zhang's program MZEF:

**1. Point your browser to `rulai.cshl.edu/`.**

This is the Zhang Laboratory home page at Cold Spring Harbor Laboratory, on beautiful Long Island.

**2. On the next page, click the Gene-Finding link in the Software Tools section.**

**3. In the next Gene Finder page, select Human.**

This selects the program version calibrated for human coding-region statistics. A simple input form is then displayed.

**4. Copy your sequence from a .txt file or a Word document.**

If you don't have a sequence handy, you can fetch the sequence AF018429 from GenBank at NCBI ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)). This entry contains the exons 1 and 2 of the dUTPase gene.

Make sure you get the sequence in FASTA format.

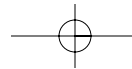
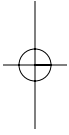
**5. Paste the sequence you copied into the input box.**

**6. Click the Submit button (below the sequence-input box) to start your analysis.**

The program quickly returns a minimally formatted output, as shown in Figure 5-10. MZEF correctly identified one exon between positions 1056–1172. The rest of the line consists of various quality values associated with the overall prediction, the various reading frames (FR1, FR2, FR3), the splice site, and the coding potential. The prediction is half correct because it missed one exon. Furthermore, the predicted exon actually starts at position 1018. Such a success rate — about half the attempts, which isn't entirely perfect — isn't so unusual for exon-finding programs.

An exon predictor that combines MZEF with other approaches (which enhances its performances) is available at this Michigan Tech site:

`genome.cs.mtu.edu/aat/aat.html`



AF018429  
sequence  
entry.

Coordinates	P	Ex1	Ex2	Ex3	Orf	3ss	Cds	5ss
1056 - 1172	0.873	0.663	0.471	0.459	112	0.467	0.581	0.668

For more information about MZEF click [here](#)

## Complete gene parsing for eukaryotic genomes

If you ever get into serious genome sequencing, you'll need the highest level of sophistication in computer-assisted annotation: genome-parsing software. These programs are designed to take large (100,000 to several million bp) pieces of a genomic sequence at once and predict the detailed exon/intron structure of entire genes.



Like the MZEF program, these programs have a modular structure, where each module has been designed to recognize a given gene component: coding exons, first/last exons, promoter regions, poly-adenylation sites, and so on. The results of these independent modules are then combined into coherent gene-structure predictions (limiting exon splicing to compatible reading frames, for instance) — and these are finally scored according to their similarity with an ideal gene model. Markov models and dynamic programming optimization are what make these programs so effective.

The most recent genome parsers, such as GenomeScan, also take into account information on protein-sequence similarity.

Despite their increasing algorithmic complexity, these programs remain easy to use; all you really have to do is paste your sequence into an input window, click a button, and voilà — you have parsed genes!

## Analyzing your sequence with GenomeScan

To show you how simple it is, we're going to show you how to use the GenomeScan parsing software program. To get things ready, prepare your

You can find such proteins by doing a blastx comparison of your sequence to all known proteins. (For more on blastx, see Chapter 7.) Alternatively, you can use the demonstration set available on the GenomeScan site.

1. **Point your browser to** [genes.mit.edu/genomescan/](http://genes.mit.edu/genomescan/).

The GenomeScan home page at the Massachusetts Institute of Technology duly appears. GenomeScan is the successor to Chris Burge's GenScan. The main difference is the use of protein homology information.

2. **Click the [GenomeScan Webservice](#) link to call up the page with the input form.**

3. **Choose Vertebrate from the Organism pull-down menu.**

This selects the program version calibrated for human coding regions statistics.

4. **Copy your DNA sequence from a .txt file or a Word document.**

Don't forget the FASTA header.

5. **Paste the DNA sequence you copied into the DNA Sequence input box — the upper of the two larger input boxes.**

Alternatively, you may click the [DNA testfile](#) link at the top of the page to use the demo sequence.

6. **Copy your homologous protein sequence set from a .txt file or a Word document.**

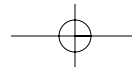
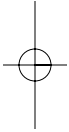
Alternatively you may click the [protein file](#) link at the top of the page to use the demo protein sequences.

Each sequence must have a FASTA header. The set may contain multiple proteins so long as each is separated by a header on its own line. Files should contain less than 1 million bases.

7. **Paste your protein sequence into the Protein Sequence input box — the lower of the two larger input boxes.**

8. **Click the Run GenomeScan button at the bottom of the form to start the analysis.**

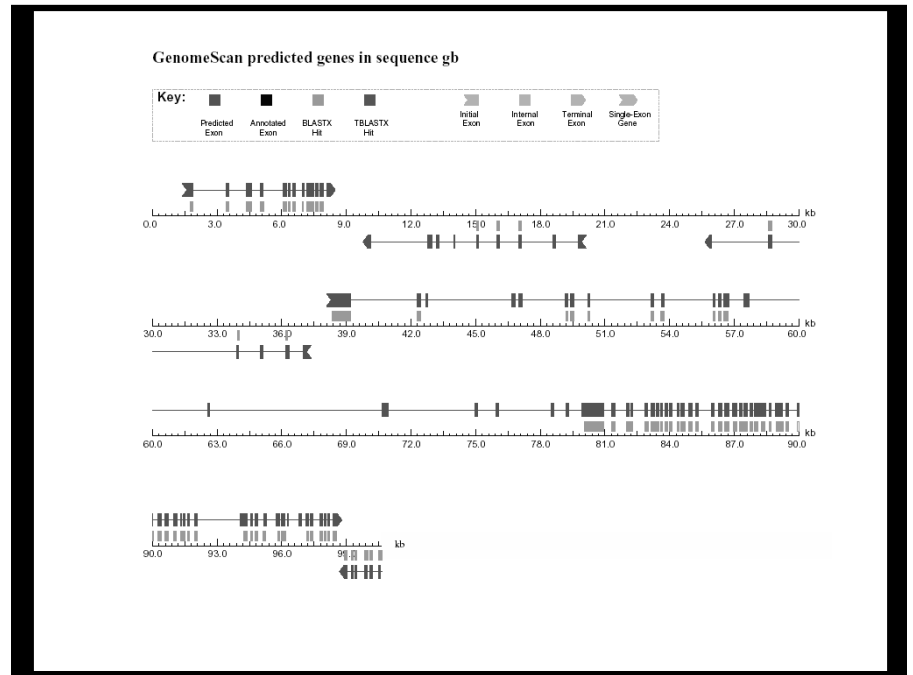
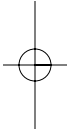
The output is a very long table listing all the components of the various predicted genes, with their coordinates and associated quality indices. This output is meant to be read by computer programs in the context of large-scale



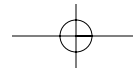
arrows) and their predicted exons (red rectangles) are clearly indicated, together with their supportive evidence from blastx similarity (green rectangle below). A total of 5 genes are predicted along this 100,000-bp long vertebrate genomic sequence.

## Assembling Sequence Fragments

Depending on their technology, current sequencing machines can only produce sequences at lengths of 500 to a 1,000 nucleotides at a time. Thus, if your DNA sequence is longer than this, it has been obtained by assembling shorter, overlapping fragments. Research into the various assembly algorithms — as well as research devoted to the development of better assembly software — is vital to bioinformatics.



**Figure 5-11:**  
Example of  
GenomeScan  
graphical  
output.



together a typical microbial genome sequence (4MB) using over 50,000 individual fragments known as *reads*. The programs used for such genome sequencing — or for bigger projects involving plant or animal genomes — take their input directly from the sequencing machine's *chromatograms* or *traces*. These traces are complex fluorescent profiles with peaks and valleys revealing the order and nature (A, T, G, C) of each nucleotide in the reads. Such programs include a base-calling system that associates a quality score to each nucleotide at each position of each read. They also include a data-management system, as well as interactive editing and display tools. Because of their complexity, these complete genome-sequencing software packages are not interactively available on Web sites; you have to install them locally on a dedicated powerful machine. Using them efficiently also requires some serious study. The most popular genome-sequencing packages that are publicly available (for academics) are these:

✓ **The Staden Package:** [staden.sourceforge.net/](http://staden.sourceforge.net/)

✓ **Phred/Phrap/Consed:** [www.phrap.org/](http://www.phrap.org/)

✓ **The Acembly engine included in the popular AceDB suite:**  
[www.acedb.org/](http://www.acedb.org/)

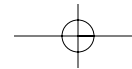
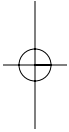
✓ **The Assembler from The Institute for Genome Research (TIGR):**  
[www.tigr.org/software/](http://www.tigr.org/software/)

Commercial genome-sequencing software packages are also available from several companies:

✓ **Sequencher (from Gene Codes):** [www.genecodes.com/](http://www.genecodes.com/)

✓ **Lasergene (from DNASTAR):** [www.dnastar.com/](http://www.dnastar.com/)

We believe that showing you how to manage a large genome-sequencing project is beyond the scope of this introductory book. However, you may still encounter the need to assemble a handful of sequences in your everyday research life. For instance, you might be working on a cDNA that's just a few thousand bp in length, represented by a dozen PCR fragments. You may have generated a number of *expression sequence tags* (ESTs — that is, pass-sequenced partial cDNAs) and want to see whether you can deduce a complete cDNA from the tags. Alternatively, you may want to assemble ESTs you extracted from a database. In the latter case, you may not have the corresponding traces and base-call quality scores.





## Assembling your sequences with CAP3

Prepare the set of sequences that you want to assemble in FASTA format as a .txt file or a Word document. Each individual sequence must be in FASTA format and have its own header. For the sake of simplicity, we're going to use the small demo dataset provided with the CAP3 documentation at [genome.cs.mtu.edu/cap/data/](http://genome.cs.mtu.edu/cap/data/). Simply open the seq file and copy the six FASTA-formatted sequences (R1 to R6) using the Select All/ Copy options from your browser. We are now ready to assemble them with CAP3 on the Lyon bioinformatics platform.

- 1. Point your browser to [pbil.univ-lyon1.fr/cap3.php](http://pbil.univ-lyon1.fr/cap3.php).**

The simplest possible form appears, made of one input sequence window above a Submit button and a Clear button. The Lyon server allows a total of 50 kb of nucleotide sequences to be processed, which is usually plenty for the purpose of assembling a cDNA.

- 2. Paste the set of sequence into the Sequence input box.**

We used the CAP3 demo sequences R1 to R6.

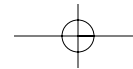
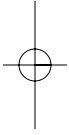
Don't forget the FASTA headers! (This is the only way for the program to distinguish a fragment sequence from the next.)

It doesn't matter whether you provide a mixture of sequences in different orientations. The program tries them in both orientations automatically.

- 3. Click the SUBMIT button to run the assembly.**

The server runs CAP3 with the default setting. A full implementation would allow you to play with the length threshold and level of similarity of the overlaps requested to assemble fragments, and to associate a file of sequence quality to the sequence data. (See the CAP3 documentation for details.)

If you provided the demo sequences, the program response will be almost immediate, and the output will look like Figure 5-12, displaying four links:

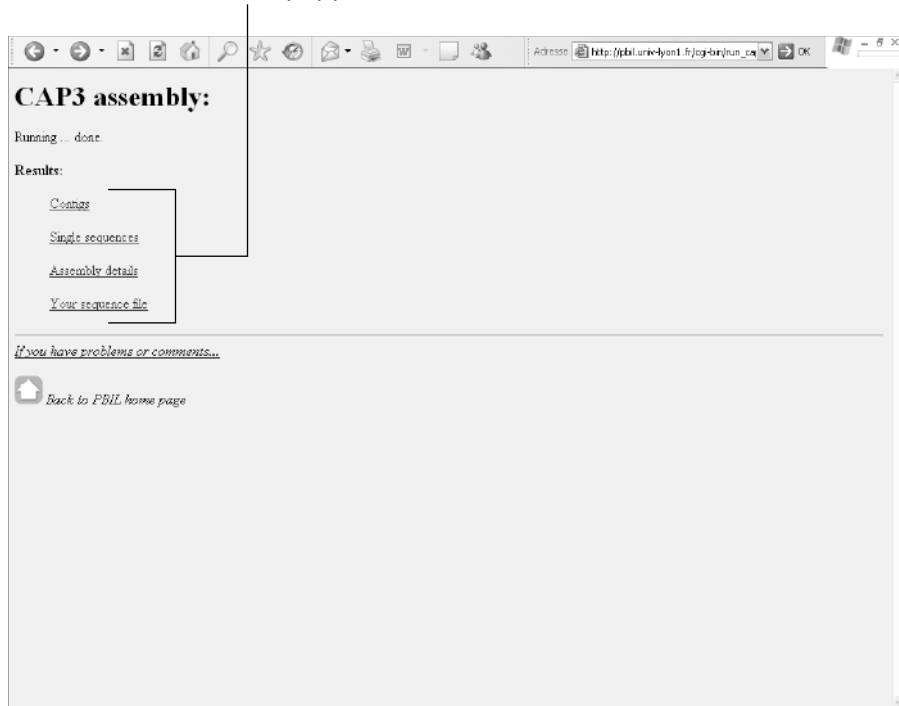
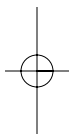


✓ **Your Sequence File:** A summary of the input sequence fragments

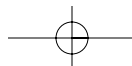
Click each of these links in turn to see their content.

For instance, clicking the [Assembly Details](#) link displays the overlaps and contig structure. (See Figure 5-13.) One can see that R5 is not part of the assembly. It is in fact found in the Single Sequences file. The resulting contig consensus sequence is displayed by clicking the [Contigs](#) link. (Refer to Figure 5-12.)

Click on these links to display your results.



**Figure 5-12:**  
Output files  
of the CAP3  
assembly  
program.





Spring Harbor, USA

and various conserved  
elements

Center for Information  
Technology, NIH, USA;  
WWW Signal Scan

bimas.dcr.t.nih.  
gov/molbio/  
signal/

Search for transcrip-  
tional elements

Center for Information  
Technology, NIH, USA;  
WWW Promoter Scan

bimas.dcr.t.nih.  
gov/molbio/  
proscan/

Predict eukaryotic  
promoter regions

Munich Bioinformatics  
Center, GER, CREDO

mips.gsf.de/  
services/  
analysis

*Cis*-Regulatory Element  
Detection

San Diego Supercomputer  
Center, USA; MEME  
& MAST

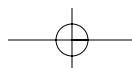
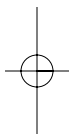
meme.sdsc.edu/  
meme/intro.html

Discover motifs in groups  
of related DNA or protein  
sequences

Université Libre de  
Bruxelles, Belgium

rsat.ulb.ac.be/  
rsat/

Analyze regulatory  
sequences



# Working with a Single Protein Sequence

---

## *In This Chapter*

- ▶ Knowing what you must know about domains, HMMs, profiles, and the Pfam domain collection
  - ▶ Visiting the three most popular sites for finding domains in your protein
  - ▶ Predicting simple physical properties of your sequences
  - ▶ Predicting protease digestion patterns
  - ▶ Predicting coiled-coil domains
  - ▶ Predicting post-translational modifications
- 

*Once you eliminate the impossible, whatever remains, no matter how improbable, must be the truth.*

— Sherlock Holmes, as told to Sir Arthur Conan Doyle (1859–1930)

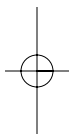
**W**hen you're studying a protein, you turn yourself into an investigator. Basically, you want to find out everything you can before designing a new experiment in the lab. For instance, in order to make sure that the protein you have in your test tube really is what you think it is, you must carry out a simple physical analysis. Many online tools exist that help you predict the results of these analyses (and make sure you're on the right track). In this chapter, we show you where you can find these tools and how to use them. Among other things, we show you how to predict molecular weights, measure isoelectric points, count residues, or predict the outcome of a protease digestion.

Of course, there's more to a protein than its weight or its charge. While these properties do help you get a sense of what's what with your particular protein, uncovering a nice biological story takes more than that. An important question to ask yourself is what your protein looks like when it's active. Does it need to be modified after being translated, does it contain coiled-coil

don't talk much about structures in this chapter.

Knowing what your protein looks like is one thing, but knowing what it actually does is another. For instance, you may want to know if your protein binds calcium or if it contains an enzymatic site. If it's an enzyme, there's no doubt that you need to know about its substrate (the kind of molecule it binds). To answer these types of questions, we show you three powerful methods to check whether your protein contains a domain with a known function.

The other powerful technique that can help you when it comes to guessing the function of a protein involves *similarity searches*. If, in a database somewhere, you find a well-characterized protein whose sequence is very similar to your protein, then you're allowed to say that most of what is true for this sequence is true for your sequence as well. Although we don't wade into this type of analysis in this chapter, don't worry — we won't leave you high and dry. We cover everything you need to know about similarity searches in Chapter 7.



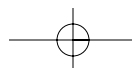
## Doing Biochemistry on a Computer

If you want to do some biochemistry using a computer — we're guessing you do — two terrific places to go are

- ✓ The ExPASy (*Expert Protein Analysis System*) server at [www.expasy.org](http://www.expasy.org), with a specific page dedicated to protein analysis methods
- ✓ The Swiss EMBnet at [www.ch.embnet.org](http://www.ch.embnet.org)

The people who created and maintain a big chunk of Swiss-Prot run these two sites. When it comes to looking closely at proteins, these folks know their jobs.

In this section, we assume that your heart is set on a protein you've never seen before. The only thing you know about this beauty is its sequence. Maybe this protein comes from a sequencing project or something similar. Before you open it up and check out the insides, you want to find out — and guess — as much about it as you can. That's the reason you're here reading this very informative chapter.





well-characterized protein, simply to check and see whether the program does what it says it does (and whether it does that well). To find a suitable example, use one of the SRS servers as we describe in Chapter 2. This can give you a sense of how good the program you're using really is.

## *Predicting the main physico-chemical properties of a protein*

*ProtParam*, a program you can use online on the ExPASy server, is a convenient way to estimate every simple physico-chemical property (that is, each *protein parameter*) that can be deduced from your protein sequence. It makes no complex and adventurous assumptions about your little protein: It simply counts for you.

Here's how it works:

### 1. Point your favorite browser to

```
www.expasy.org/tools/#primary
```

The Primary Structure Analysis section of the ExPASy Proteomics Tools page appears, as shown in Figure 6-1.

At the end of this chapter, we give you an exhaustive list of all the available ExPASy mirror sites you can use if the main ExPASy site is down.

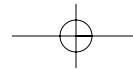
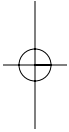
### 2. Click the **ProtParam** link near the top of the page.

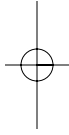
The ProtParam Tool page appears.

### 3. Enter your sequence in the search boxes provided.

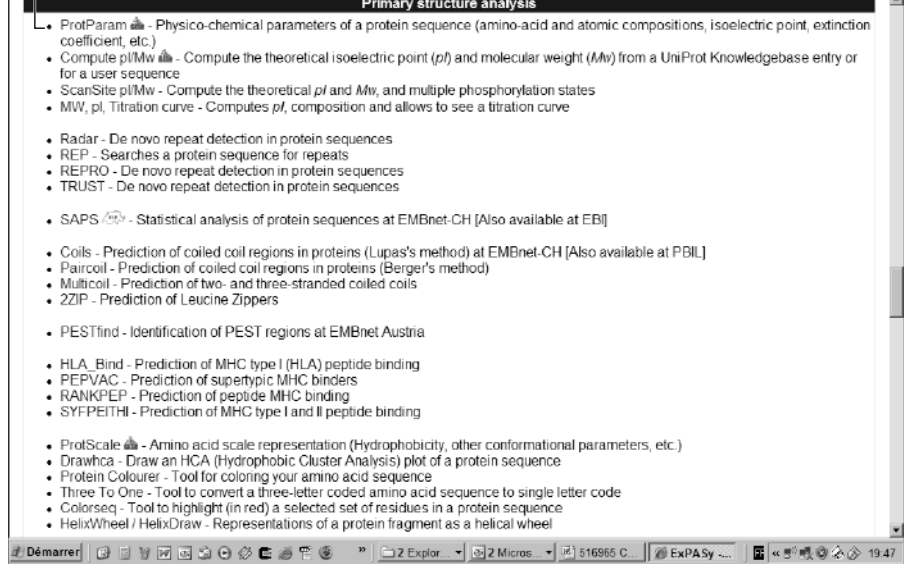
Provide a sequence in one of two ways:

- Enter the accession number, as shown in Figure 6-2. You can do this if your sequence exists in the Swiss-Prot sequence database. Swiss-Prot accession numbers usually start with a letter followed by five digits. In this example, we use a Swiss-Prot protein: rat Syntaxin 1A (P32851).





**Figure 6-1:**  
Working  
with the  
Primary  
Structure  
Analysis  
page.

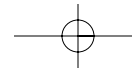


- Paste the sequence — in raw format — into the window provided (see Figure 6-2 again). If your sequence isn't a Swiss-Prot sequence, or if you don't know the sequence name, you can paste the sequence itself, using this other window. Note that the sequence must be entered in *raw format* — it should not include anything but the standard abbreviations for amino acids (white space and numbers are tolerated but automatically removed). If your sequence is in FASTA format, DO NOT include the first line that starts with a greater-than sign (>).

**4. Click the Compute Parameters button.**

If, in Step 2, you provided the accession number of the sequence, an intermediate page (like the one shown in Figure 6-3) appears.

- 5. If you have entered a Swiss-Prot accession number — and Figure 6-3 has dutifully appeared — enter the range of the analysis in the N-Terminal and C-Terminal fields. (The N-terminus is the left side of your protein sequence and the C-terminus is the right side.)**







## ProtParam tool

**ProtParam** (References / Documentation) is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) (Disclaimer).

Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P06130**) or a sequence identifier (ID) (for example **KPC1\_DROME**):

P32651

Or you can paste your own sequence in the box below:

Paste a raw sequence here.

Enter accession number here.

**Figure 6-2:**  
Entering an  
accession  
number into  
ProtParam.

At the top of the page (partially) shown in Figure 6-3, you can find a list of the features contained in your sequence. This isn't a prediction — rather, it's a display of the information contained in the Swiss-Prot entry. Below are two boxes (N-terminal and C-Terminal) that you can fill with the coordinates of the segment you're interested in. If you leave these two boxes blank, ProtParam analyzes the complete sequence.

Using coordinates such as those in Figure 6-3, ProtParam analyzes the segment of your protein that starts on the 266th amino acid and ends on the 288th.

### 6. Press the Submit button to proceed with the analysis.

The results — including molecular weight, extinction coefficient, and estimated half-life — appear on-screen.

### 7. In your Web browser, choose File⇨Save As to save your results.

FT	COMPBIAS	13-19	Asp-rich (acidic).
FT	HELIX	28-63	
FT	STRAND	65-66	
FT	HELIX	69-104	
FT	TURN	105-106	
FT	STRAND	107-107	
FT	HELIX	111-146	
FT	TURN	147-147	
FT	TURN	156-157	
FT	HELIX	162-170	
FT	TURN	171-171	
FT	STRAND	173-174	
FT	TURN	176-180	
FT	HELIX	192-254	
FT	TURN	255-255	

Or, if you wish to select a different sequence fragment (at least 5 amino acids long), you can enter the desired endpoints on the sequence here (by default, the computation will be carried out for the complete sequence).

N-terminal:

C-terminal:

The sequence STX1A\_RAT consists of 288 amino acids.

ExPASy Home page   Site Map   Search ExPASy   Contact us   Swiss-Prot   Proteomics tools

**Figure 6-3:**  
Using  
ranges with  
ProtParam.

## Interpreting ProtParam results

If you've used Syntaxin as your sample protein in your trial run of ProtParam, you can now proudly pass on the knowledge that Syntaxin's theoretical molecular weight is 33067,4 daltons, that its theoretical pI is 5.14, and that its aliphatic index is 83.96 — and this is only a sample of the large amount of information ProtParam can provide!

You can find detailed explanations on how ProtParam computes these parameters at [www.expasy.org/tools/protparam-ref.html](http://www.expasy.org/tools/protparam-ref.html). Among other issues, the following sections draw your attention to a few points you may be interested in.

### Molecular weight

The computer program is simply summing the weight of all the residues in the sequence. Remember that the program doesn't consider the following:

- ✓ It does NOT consider post-translational modifications such as glycosylations or phosphorylations.

Of course, these problems can eventually add up and lead to experimental results significantly different from the theory. In a way, that's where the fun begins . . . so keep your eyes open!

### ***Extinction coefficients***

An *extinction coefficient* tells you how much light (visible and invisible) your protein absorbs at a certain wavelength. Estimates of these values are useful if you need to follow your protein with a spectrophotometer when you purify it.



When you use this result, remember that the ProtParam value is only an indication. ProtParam predicts the extinction coefficient by summing the contribution of every amino acid as if each were independent. This calculation ignores the fact that the behavior of amino acids can be dramatically altered depending on their immediate surroundings. That behavior can't be predicted, so this estimate isn't exactly the last word (so to speak) on your protein's extinction coefficient.

If you need the *exact* coefficient, you must measure it experimentally. On the other hand, for most proteins, the experimental coefficient is (fortunately) rarely very different from the theoretical one.

### ***Instability***

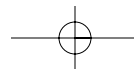
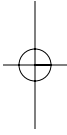
This parameter is only just a crude estimate of your protein's stability in a test tube. When the index is below 40, the protein is usually stable. Above 40, it may not be stable.

### ***Half-life***

The *half-life* is a crude prediction of the time it takes for half of the amount of your protein present in a cell to disappear completely after its synthesis in the cell. This prediction is given for three types of organisms. You can safely extrapolate it to similar organisms.



The half-life value given here is meaningless if the degradation of your protein is part of a regulatory process. For instance, if the cell specifically directs a *protease* (a protein that digests other proteins) at your protein, your protein may disappear much faster than ProtParam tells you.



- ✓ Separate the domains in your protein
- ✓ Identify potential post-translational modification by mass spectrometry
- ✓ Remove a tag protein when you express a fusion protein
- ✓ Make sure that the protein you're cloning isn't sensitive to some endogenous proteases

Fortunately, these days you can snip models of protein sequences in a computer — and see what happens *before* you start any hands-on lab procedures. PeptideCutter is a very useful tool for this kind of purpose — and is available from the ExPASy Web site at [www.expasy.org/tools/#proteome](http://www.expasy.org/tools/#proteome). Check it out; you'll find it easy to use. Paste in your sequence, choose an option or two, click a button, and a model of your protein is sliced, diced, and chopped to your specifications.

## Doing Primary Structure Analysis

The *primary structure* is the amino acid sequence of your protein. When you analyze the primary structure, you ignore the potential interactions between amino acids. Primary structure analysis of your protein doesn't tell you about the potential interactions between the amino acids in your sequence; you predict those when you predict the secondary or tertiary structure of your protein. (See Chapter 11 for more on such predictions.)

You conduct a primary structure analysis to try to find segments in your protein that display a special composition. These segments can reveal some interesting properties of your protein, such as

- ✓ *Hydrophobic regions* that could be membrane-spanning segments in proteins that anchor themselves into a membrane
- ✓ *Coiled-coil regions* that indicate potential protein-protein interaction
- ✓ *Hydrophilic stretches* that could be looping out at the surface of the protein

The methods we show you in this section, though less powerful than state-of-the-art methods of predicting secondary structure (which we cover in Chapter 11), still have a lot going for them. For one, they're simple to use and simple to understand. When you use them, you can easily see what's going on — and easily avoid making mistakes.

principle is very simple. What you need is a chemical property and a list of values associated with each of the 20 amino acids. This property can be any measurable physico-chemical parameter, such as size, polarity, hydrophobicity, or even the propensity of amino acids to be in a specific structural state. The values in this table are the amino acids' *scale values*. Many such tables exist that have been determined experimentally for almost any characteristic you can think of.

After you have this table, you choose a window size and slide it along your sequence. When the window is centered on an amino acid, the scale values associated with the amino acids it contains are summed up and averaged. The resulting value is associated with the central amino acid, and then the window is shifted by one amino acid. This process goes on to the end of the sequence. The following example illustrates

the window is centered.

Sequence

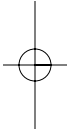
AGVCFGTRESALPTFREDCYGHZPLI  
KJFDESAQZ

<---A---> Window 1

<---G---> Window 2

<---V---> Window 3

When the sliding operation is finished, the values associated with every amino acid are plotted against the sequence. Biologists name this display a property profile (do not confuse it with a domain profile, which is a formulation of a multiple sequence alignment). If you're lucky, you may be able to identify transmembrane segments, loops, or coiled-coil regions by using sliding windows. Hydrophobicity is the most popular analysis because it's a good indicator of transmembrane segments or core regions within a protein.



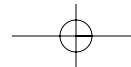
Many of the methods we show you here rely on the *sliding windows* technique. (For more on these guys, check out the "Sliding windows" sidebar.) Predictions based on sliding windows aren't very sensitive or very precise, but they are very robust. If you see a strong signal when using a method based on sliding windows, chances are you're looking at something that's a genuine biological signal.

Ironically, the main advantage of the methods that use sliding windows is also one of their main shortcomings: They don't interpret the results for you; they provide only a raw signal.



Because you have to do the interpretation yourself, two simple rules apply when interpreting the results of an analysis based on sliding windows:

- ✓ Be very strict and consider only strong signals.
- ✓ Check the robustness of your signal. Good signals aren't shy. They don't go away simply because you increase or decrease the window size by one amino acid or replace a property table with another similar table.





looking for.

For instance, if you're looking for transmembrane domains that are normally 21 amino acids long, you want to use a window size that is close to this value — in practice, 19 is the best window value for finding transmembrane regions. On the other hand, if you're interested in the structural features of globular proteins, a range between 7 and 11 would be more appropriate.

## Looking for transmembrane segments

Predicting that your protein has transmembrane segments tells you much more about its function than almost any other simple prediction you can do.

When you know your protein is a transmembrane protein, you know that you can't work on it with the same techniques you'd use if it were a globular protein. If you find one transmembrane segment at the N-terminus of your sequence, you can predict that the protein is secreted. On the other hand, many transmembrane domains in one protein may indicate a channel.

Because these predictions are so important, we show you two methods for testing for transmembrane segments: ProtScale (a very simple one) and TMHMM (one of the most complete). TMHMM is not part of the ExPASy server; it is a service offered to the community by the Technical University of Denmark.

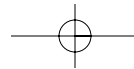
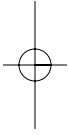
- ✓ *ProtScale* uses a sliding-window technique and one of many amino-acid scale values. In this example, we use the hydrophobicity to identify the groups of hydrophobic segments that characterize transmembrane proteins. ProtScale doesn't predict anything for you; it returns a hydrophobicity profile and lets you do the interpretation.
- ✓ *TMHMM* is a state-of-the-art program that predicts transmembrane segments in your protein. TMHMM also tells you about the portions of your protein that are probably inside the cell and those that are probably outside.

### Running ProtScale

Let's start with ProtScale:

1. **Point your browser to** [www.expasy.org/cgi-bin/protscale.pl](http://www.expasy.org/cgi-bin/protscale.pl).

The ProtScale page duly appears.



**3. Scroll down the same page and then select the radio button next to Hphob. / Kyte & Doolittle, as shown in Figure 6-4.**

ProtScale gives you a large range of properties that you can choose and test on your protein. The one we've chosen here is the most appropriate for predicting transmembrane helices.

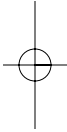
**4. From the Window Size pull-down menu, choose 19.**

This value is appropriate when you're looking for transmembrane domains.

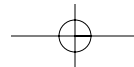
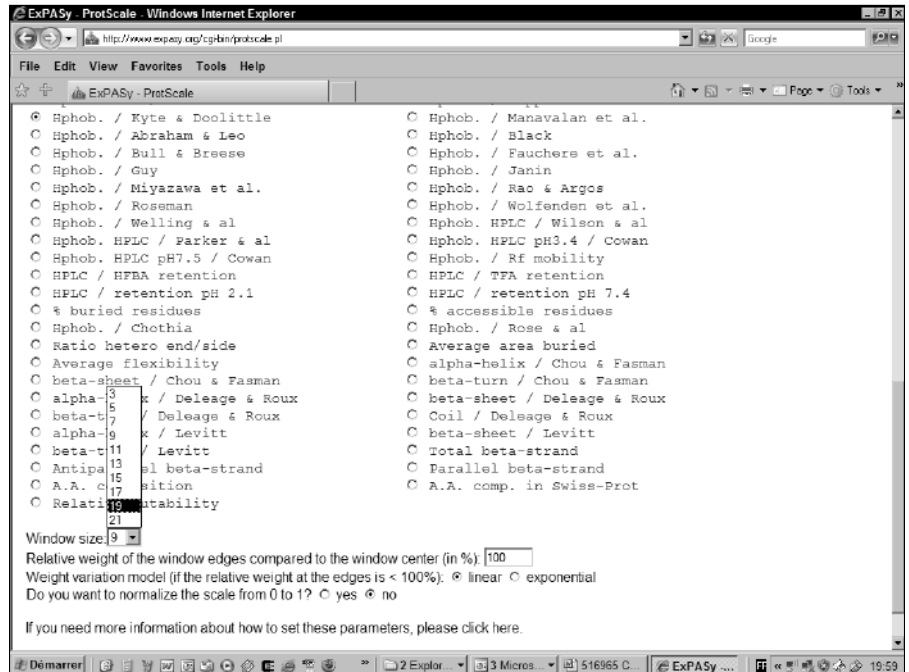
**5. Press the Submit button at the bottom of the page.**

**6. If you have entered a Swiss-Prot accession number, enter the range of the analysis.**

If, in Step 2, you have provided the accession number of the sequence, an intermediate page appears like the one you saw in Figure 6-3.



**Figure 6-4:**  
Choosing  
a property  
on the  
ProtScale  
server.



7. Click the **Submit** button to proceed with the analysis, and then wait for your browser to return the results.

8. When the **Results** page appears, click *Image in GIF format* at the bottom of the page.

A new page should appear on your browser, containing only the graph.

Of the three formats proposed here, the GIF (*Graphic Interchange Format*) is the most convenient if you want to include your graph in a presentation.

9. Choose **File** → **Save As** from your browser's main menu to save your results.

### *Interpreting ProtScale results*

A good signal is strong and not very sensitive to the parameters. For instance, Figure 6-5 shows the (strong) results you obtain on the Swiss-Prot protein P78588 when using the Kyte and Doolittle Hydrophobicity Scale.

In order to convince yourself that this result is meaningful, redo the comparison with another hydrophobicity scale — like the one from Eisenberg, with the results you see in Figure 6-6. The details differ very little between these two graphs, but you can clearly see that the main features are well conserved and that the strongest peaks come at roughly the same positions.



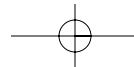
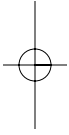
In Figure 6-5, we've drawn a bold line that gives you a hint on how to interpret your results.



The recommended threshold value when using Kyte and Doolittle is 1.6 — but if you're like us and you keep forgetting this magic value, here is a simple recipe:

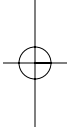
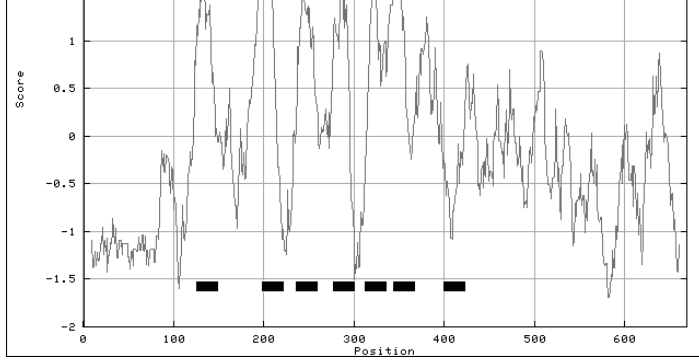
1. **Place a piece of paper over your results.**
2. **Lower this piece of paper until the tips of the strongest peaks appear.**
3. **Keep lowering this threshold as long as you can see nice sharp peaks.**

With this simple method, you can identify unambiguously five out of seven transmembrane regions. On a sixth one, we could make an adventurous guess — It's easy to guess like this when you know the answer! — but the seventh transmembrane domain is impossible to find.

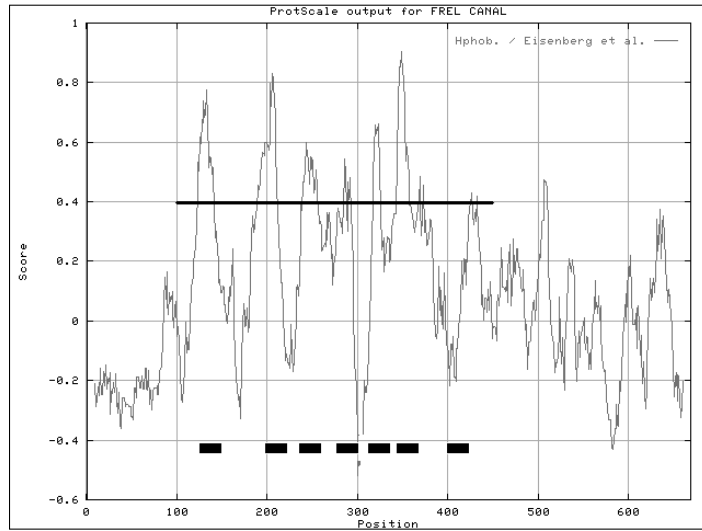




Hydrophobicity profile returned by ProtScale by using the Kyte & Doolittle Scale. The boxes mark known transmembrane segments.



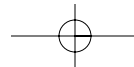
**Figure 6-6:** Hydrophobicity profile returned by ProtScale using the Eisenberg Scale.



That isn't so surprising: Many proteins contain these seven transmembrane regions — in which six are easy to find and the seventh domain is notoriously difficult to predict.

### *Running TMHMM*

TMHMM is maintained by the Center for Biological Sequence Analysis (CBS) at the Technical University of Denmark. The CBS ([www.cbs.dtu.dk/services/](http://www.cbs.dtu.dk/services/))



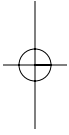
1. Point your browser to [www.cbs.dtu.dk/services/TMHMM-2.0](http://www.cbs.dtu.dk/services/TMHMM-2.0).

The TMHMM page of the CBS site appears (Figure 6-7).

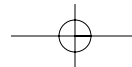
2. Scroll down the page a bit to enter your sequence in the TMHMM search box.

TMHMM only recognizes the sequence in FASTA format. For the following example, we use the protein sequence FREL\_CANAL. To obtain this sequence from Swiss-Prot:

- a. Open a new browser window.
- b. Open the page [www.expasy.ch/cgi-bin/get-sprot-fasta?FREL\\_CANAL](http://www.expasy.ch/cgi-bin/get-sprot-fasta?FREL_CANAL), which contains REL\_CANAL in FASTA format.
- c. Copy the sequence onto the Clipboard.
- d. Paste the sequence in the TMHMM window (refer to Figure 6-7).



**Figure 6-7:**  
The bottom  
half of the  
TMHMM  
page.



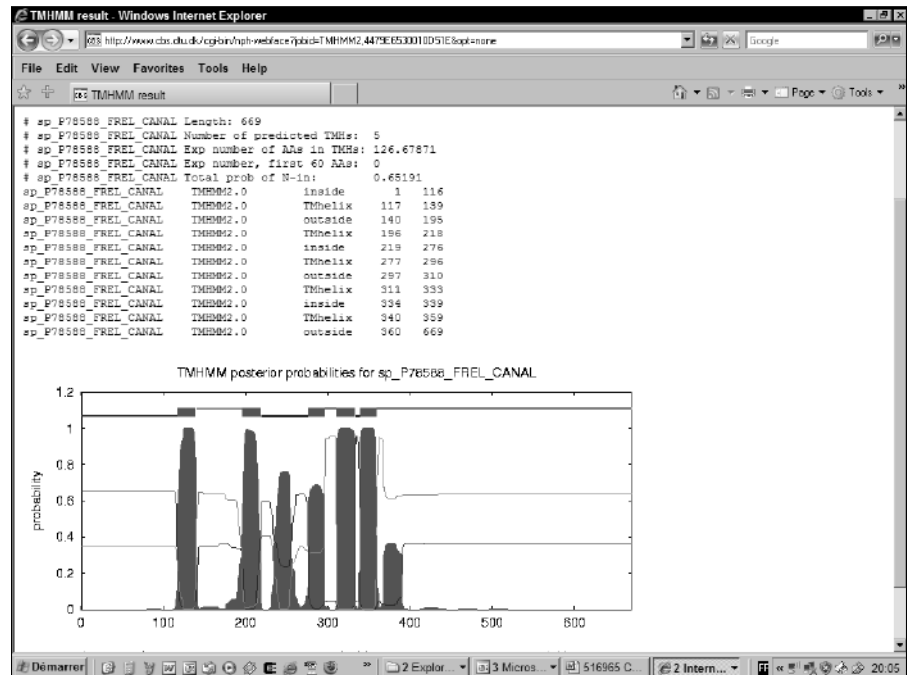
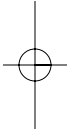
If you are using Internet Explorer, the Save As Type pull-down menu appears; choose Web Page, Complete (\*.htm, \*.html). Your file is saved along with a directory that has the same name. Don't separate your file from this directory.

### Interpreting results from TMHMM

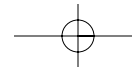
Although results from TMHMM are slightly different from those you get from ProtScale, they are nevertheless in rough agreement. This is good news because the two methods use very different principles.

There are two major differences between ProtScale and TMHMM:

- ✓ TMHMM returns a precise prediction. This prediction is at the top of the output — just before the plotted graph — as shown in Figure 6-8.
- ✓ TMHMM predicts the segments that are inside the cell and the segments that are outside the cell.



**Figure 6-8:**  
TMHMM  
results on  
the Swiss-  
Prot Protein  
FREL\_  
CANAL





the cell.

This probability is only a prediction; it's not necessarily correct.



If you really need an accurate prediction because you're about to design an experiment, we recommend running many predictions using different methods. Very different methods that give you the same results are usually a good indication that you're on the right track.

## Looking for coiled-coil regions

*Coiled-coil regions* are portions of a protein formed by the intertwining of two or three alpha-helices. One reason it's considered interesting to find coiled-coil regions is that they're often involved in protein-protein interactions. Another (less glorious) reason is that these coiled-coil regions can give false matches when you do a database search (see Chapter 7) — and it can be a good thing to filter them out. If you want to predict these regions in your protein of interest, you can use the conveniently named COILS server at EMBnet. Check them out at

[www.ch.embnet.org/software/COILS\\_form.html](http://www.ch.embnet.org/software/COILS_form.html)

## Predicting Post-Translational Modifications in Your Protein

Proteins often need to be modified before they become active in the cell. Biologists call such operations *post-translational modifications* because they occur after the translation (synthesis) of the protein. These modifications may involve adding sugars, modifying amino acids, or removing pieces of the newly synthesized protein. If you're studying a new protein, you probably want to know about such matters. This may be very important if you want to clone and express a human protein in bacteria — because, in order to be active, your protein may require some post-translational modifications that the bacterium itself cannot make.

One of the most useful tools for analyzing post-translational modifications is PROSITE — a database you can find on the ExPASy site. It contains a list of short sequence motifs (also some named patterns) that experiments have



When you do sequence analysis, there is something you must ALWAYS remember: Similar short sequences (such as those with less than 20 amino acids) don't ALWAYS have the same function. Thus, if a small sequence has been shown to function as an ATP binding in a protein — and if the protein you're analyzing contains a short segment with EXACTLY the same sequence — it doesn't NECESSARILY mean that your protein is also an ATP binding protein. What you have is an indication that it *may* be an ATP binding protein. Of course, the longer the segment, the stronger the indication.

This warning on the meaning of short similarity regions also applies to PROSITE patterns — patterns listed there are often quite short. That said, we can show you how to check your sequence to see whether it contains any known PROSITE pattern.

## Looking for PROSITE patterns

ScanProsite is a server that allows you to compare your protein with the list of patterns contained in the PROSITE database. Highly trained specialists designed each pattern in this database. If you find that your protein contains a PROSITE pattern, this one fact can (often) give you a pretty clear indication of its function.

### The PROSITE patterns

When biologists first started having access to protein sequences, one of their first discoveries was that some small conserved sequences are often associated with important properties — such as cellular localization, ligand binding, and so on.

Amos Bairoch, the creator of Swiss-Prot, started exploiting this discovery while building his protein database. To help organize and annotate the proteins, he created a collection of small, well-conserved segments that he could use to classify and analyze new proteins.

These special segments are known as *patterns* — and they are still widely used today as a means of characterizing new proteins. PROSITE is the name Amos gave to this particular pattern database. These days, PROSITE no longer contains only patterns; it also includes a new, more sophisticated type of model: the *profile*. Profiles describe every position of an entire protein family — not just a few highly conserved positions, as patterns do. (These domain profiles are very different from the property profiles we also describe in this chapter.)

- The right section in Figure 6-9 is right up your alley if you already have a pattern and you want to check whether other proteins in Swiss-Prot contain this pattern.
- The left section in Figure 6-9 is right up your alley if you have a protein sequence and you want to find out which PROSITE pattern it contains.

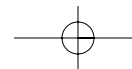
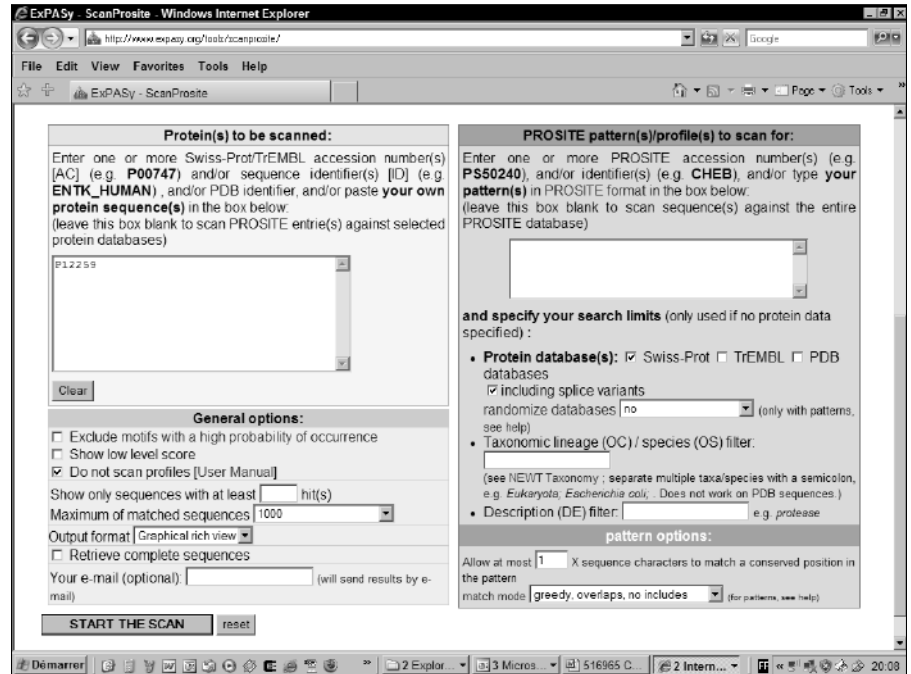
**2. Paste your sequence or enter its accession number into the search text box on the left (in the Protein[s] to be Scanned section).**

We want to go from a protein sequence to a pattern, so we're staying with the search box on the left. You can use

- Raw format (amino acids only, space and numbers are tolerated but removed automatically).
- An accession number. For this example we use the Swiss-Prot protein P12259, which is the precursor of the Human Coagulation factor V.



**Figure 6-9:**  
The  
ScanProsite  
Interface on  
the ExPASy  
Server.



For the purposes of this steps list, it's okay not to select to scan the profiles; that takes more time. Furthermore, when it comes to post-translational modifications, patterns are usually more interesting than profiles.

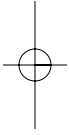
**5. Press the Start the Scan button and wait.**

An intermediate page appears. Keep waiting until your browser displays a Results page like the one shown in Figure 6-10.

Wait for the results! Don't click any of the hyperlinks that appear on the intermediate page. The ScanProsite server can be very slow at some times of the day.

If your results never arrive, you have the alternative of using any ExPASy site that is closer to you or running in a part of the world that is sleeping. (See the listing at the end of this chapter for alternative ExPASy addresses.)

**6. Choose File → Save As from your browser's main menu to save the information on the Results page.**



## *Interpreting ScanProsite results*

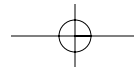
The main problem with most post-translational patterns is that they are short. The short sequences these patterns match may occur by chance, and could have no real biological significance.

The rule is that similarities between short sequences must always be treated with suspicion. A good way to rule out many unlikely modifications is to read carefully the documentation associated with the pattern (the PDOC).

### *Understanding the ScanProsite output*

The first section of the output is named "Hits by Patterns" and shows the location of every type of pattern found on your protein. It is basically a summary where each pattern family has its own color code. Pass the mouse pointer over one of the colored rectangles to see its name displayed.

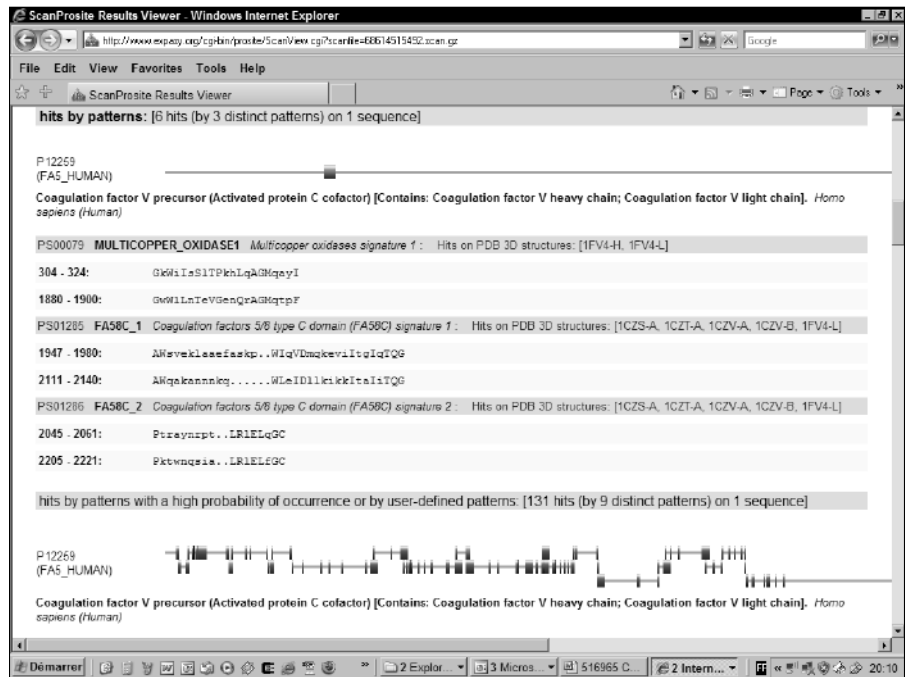
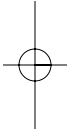
In moving down through the Hits by Patterns section, you can see the details of each pattern found in your sequence. Documentation is available via the hyperlink displayed in the output. A portion of this output is shown in Figure 6-10. For each pattern found in PROSITE, ScanProsite reports the following information:



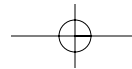
3-D structure. If this case with the pattern you're looking at, you can find the PDB name of the structure(s) here — as well as a hyperlink to the 3-D representation of this structure. PDB stands for the *Protein Data Base*, a database that contains all the known protein 3-D structures. (See Chapter 11 for more information on the PDB.) If you click one of these hyperlinks, a static GIF image of the corresponding structure appears in your browser. Figure 6-11 shows the image you can obtain by clicking the [1czv](#) hyperlink (or use the address [www.expasy.org/cgi-bin/view-pdb?pdb=1czv&ps=PS01285](http://www.expasy.org/cgi-bin/view-pdb?pdb=1czv&ps=PS01285)). The image contains

- A 3-D structure of your protein using a ribbon representation
- The location of this particular pattern (marked in green) in the structure as a whole

✓ **The list:** A list of the segments that contain the patterns within your protein. The numbers indicate the position of the match within your sequence. Capital letters indicate residues that were specified by the pattern; lowercase letters indicate residues that weren't specified by the pattern.

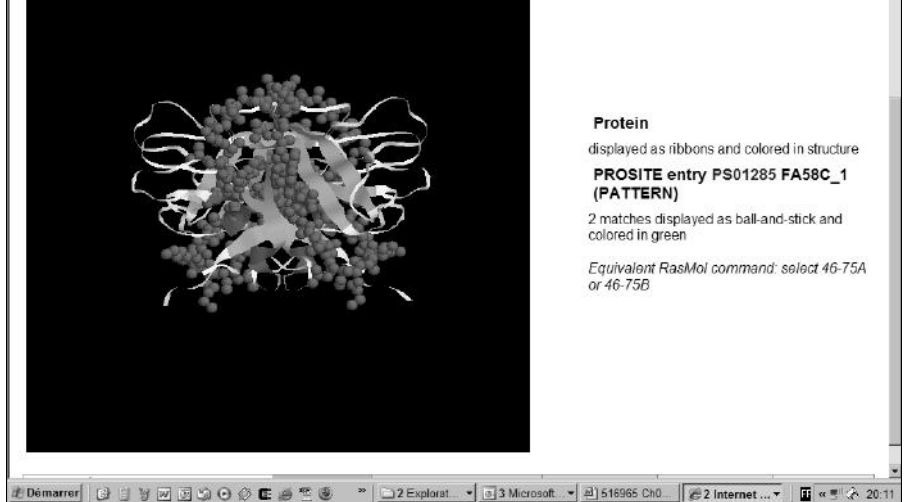


**Figure 6-10:**  
The  
ScanProsite  
Output.





**Figure 6-11:**  
Localization  
of a  
PROSITE  
pattern on  
the 3-D  
structure of  
a protein.



The next section of the output (titled “Hits by Patterns with a High Probability of Occurrence”) is similar to this one, and shows you the location of short common patterns in your sequence.

### *Being careful with short patterns*

The complete output obtained from P12259 gives us a good illustration of the irrelevance of most short patterns: This output indicates that 19 sites of myristylation exist in the Human Coagulation Factor V. Is this information you can trust or not?

If you click on the link PS0008 near the end of the “Hits by Patterns with a High Probability of Occurrence” section, it will take you to the documentation where you will discover that myristate is a fatty acid attached to the N-terminus (left end of the sequence) of a protein, anchoring it into the membrane. None of the hits reported here are on the N-terminus. So you can safely conclude that none of these matches are genuine.

### *Using the species information*

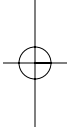


When you find a post-translational modification, make sure that this modification is consistent with the species the sequence comes from. The modifications that occur in prokaryotes are usually different from those that occur in eukaryotes. You must look for this information in the PDOC corresponding to

Further along the output that ScanProsite returns, the pattern list contains two matches that are specific of coagulation factors: FA58C\_1 and FA58C\_2.

If you look at them individually, neither pattern is really significant. On the other hand, the fact that you find two related patterns at so close a distance is very exciting: It's a good indication that both patterns are probably genuine.

This is a good example of the fact that two weak indications can make up for strong evidence when they are consistent (and independent!). It's a bit like two shortsighted witnesses telling you exactly the same thing about what they saw. For an investigator, this is even more interesting than the report of one eagle-eye witness, as long as the investigator can be sure that the two shortsighted witnesses never had a chance to talk with each other.



### ***Eliminating weak patterns***

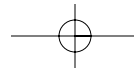
When you have too many weak patterns, a good strategy is to build a multiple sequence alignment with related sequences. A pattern that corresponds to a genuine post-translational modification is normally well conserved. We provide you with a brief session on multiple alignments in Chapter 2 and most of what you need to know to build high-quality multiple sequence alignments is in Chapter 9.

### ***Everything is not in PROSITE***

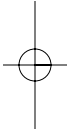
Also important to know is that the PROSITE pattern collection doesn't contain a description for every known post-translational modification. If you can't find what you need in PROSITE, the ExPASy server gives you links to other tools that are specialized for this type of analysis. (The links are located in the Post-Translational Modification Prediction section of PROSITE at [www.expasy.org/tools/#ptm](http://www.expasy.org/tools/#ptm).)

## ***Finding Known Domains in Your Protein***

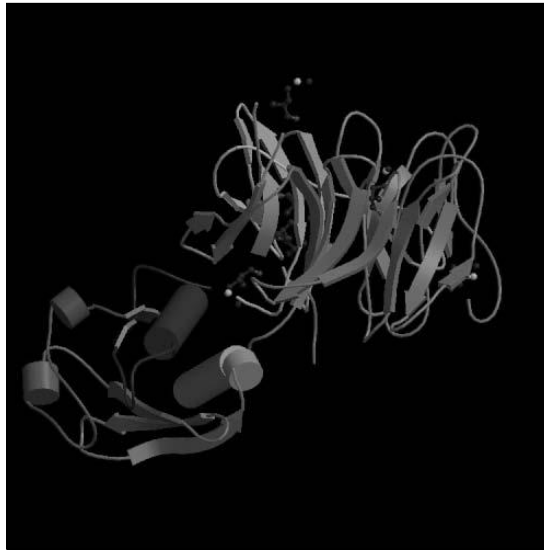
Gurus define domains as "independent globular folding units." For the rest of us, a *domain* is a portion of protein that can keep its shape if you remove it from the rest of the protein. A domain consists of at least 50 amino acids. Domains are like the various components of your kitchen: the oven, the microwave, the fridge, and so on. All together they constitute the complete



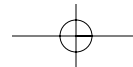
domain plays a specific role in the function of the protein. It may interact with other proteins, it may bind an ion like calcium or zinc, or it may contain an active site. It is common to have a catalytic domain associated with a binding domain and a regulatory domain. Imagine, if you will, a toaster, where you have the grill (catalytic), the toast holder (binding), and the switch (regulation).



**Figure 6-12:**  
The TolB  
protein is  
made of two  
domains.



In principle and in practice, little difference lies between searching patterns and searching domains in your sequence, and once a domain is characterized, it is easy to check whether your protein contains it or not. A dozen or so domain collections exist around the world (Table 6-1). For some of these collections, a bunch of specialists handcrafted the domains one by one, like Swiss watches. These collections give very precise results but are sometimes a bit incomplete. (See the entries marked Manual in Table 6-1.) In other collections, a computer crunches all known sequences and splits them into a bunch of domains. These collections are more extensive but less informative. (See the entries marked Automatic in Table 6-1.)



and (b) gurus often find it hard to agree with each other. (So now you know why various authors have described most of the important domains in slightly different ways.)

For us, the consequences are that there's a large redundancy among the eight major collections of domains available today (Table 6-1). Life would be simpler if we could tell you that TastyDom (r) is the best and that you must not use PooPooDom (r)! Unfortunately, this isn't the way it works. Each collection has its pros and cons — and eventually, if you really want to understand your protein, you need to look at each collection! It takes time, but in the end, it can be truly worth the effort.

The InterPro project has made it possible to search many domain databases simultaneously. In Table 6-1, (IP) indicates the domain databases that can be accessed by InterPro. InterPro helps you make sure that you're not missing even the tiniest bit of information. You can also find out easily when some domain collections disagree on your sequence.

Unfortunately, InterPro is not exhaustive. The only way to make complete analyses of the domains contained in your sequence is to use the three major domain servers:

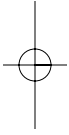
- ✓ InterProScan
- ✓ CD-Search
- ✓ Motif-Scan

It is highly probable that one, and *only* one, of these servers contains *the* bit of information that helps you make sense of your protein — meaning that you'll have to check out each server in turn. The rest of this chapter shows you how to use each of these servers and get the best out of each.

**Table 6-1**                      **The Main Domain Collections**

<b>Name</b>	<b>Web Address</b>	<b>Number of Domains</b>	<b>Generation</b>
PROSITE-Profile (IP)	<a href="http://www.expasy.org/prosite">www.expasy.org/prosite</a>	616	Manual
PfamA (IP)	<a href="http://www.sanger.ac.uk/Software/Pfam">www.sanger.ac.uk/Software/Pfam</a>	7973	Manual

PRODOM (IP)	protein.toulouse.inra.fr/prodom/current/html/home.php	736000	Automatic
SMART (IP)	smart.embl-heidelberg.de	685	Manual
COGs	www.ncbi.nlm.nih.gov/COG/new/	4852	Manual
TIGRFAM (IP)	www.tigr.org/TIGRFAMs	2453	Manual
BLOCKs	blocks.fhcrc.org/	12542	Automatic



## *Finding domains with InterProScan*

You want to use the InterProScan server if you have a protein sequence and you want to know which domain this sequence may contain. InterProScan allows you to compare your sequence with InterPro, a domain database that includes most of the major domain collections available online. In a way, InterProScan is for protein sequences and domains what metasearch engines are for the World Wide Web!

Here's how you can get InterProScan working for you:

- 1. Point your browser to** [www.ebi.ac.uk/InterProScan/](http://www.ebi.ac.uk/InterProScan/).

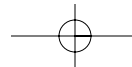
The InterProScan home page appears.

- 2. Enter your sequence in the search text box.**

Most legal formats are recognized here (including the raw format), but the accession numbers don't work.

In the following example, we use the protein sequence FOSB\_HUMAN. Here's how to obtain this sequence from Swiss-Prot:

- Open a new Internet browser window.
- Open the page [www.expasy.ch/cgi-bin/get-sprot-fasta?FOSB\\_HUMAN](http://www.expasy.ch/cgi-bin/get-sprot-fasta?FOSB_HUMAN), which contains FOSB\_HUMAN in format FASTA.



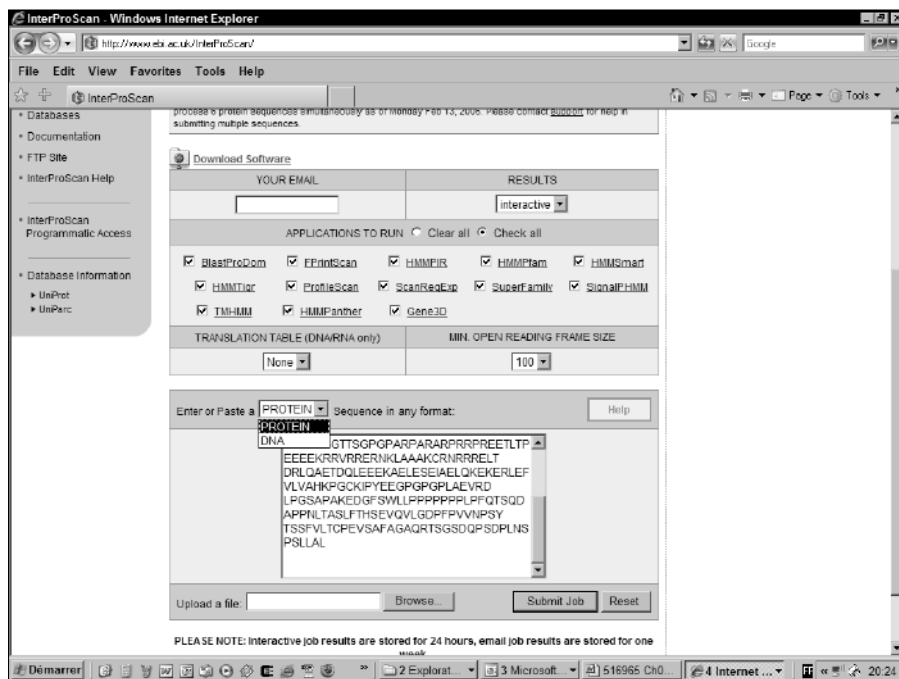
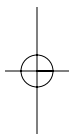
**3. In the Applications to Run section, choose the domain databases you are interested in.**

InterPro lets you search domain databases, but it also lets you run some specific prediction applications, such as *TMHMM* (for trans-membrane predictions) and *SignalPHMM*, which runs a prediction to determine whether your sequence contains a *signal peptide* (a short N-terminus segment that causes your protein to reach its destination in the cell, as precisely as a Zip code.)

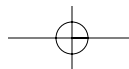
You can obtain your results faster if you remove some of the largest databases (ProDom, for example) from your search.

**4. Click the Submit Job button.**

An intermediate page pops up and informs you of how long your search has been running.



**Figure 6-13:**  
The Inter  
ProScan  
home page.



your favorite application (Powerpoint, for instance). To keep your exact results for further reference, the simplest thing to do is to save the raw output:

- a. Click the handy Raw Output button.  
A new window appears, displaying just the raw data.
- b. Use your browser's File→Save As command.

## Interpreting InterProScan results

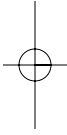
Because InterProScan returns a large amount of information, interpreting its results correctly is something of an art form. Before clicking any hyperlink, make sure that you understand exactly what the result page contains.

### Understanding the InterProScan output

Figure 6-14 shows the results you obtain when scanning the protein FOSB\_HUMAN with the InterProScan server. It illustrates well the variety of results one can expect when you compare a sequence with domain databases.

In the output, each line represents a match with some type of InterPro domain.

- ✓ **The first entry in each column indicates the type of diagnosis provided: Family or Domain.** Some domains or signatures are specific of a complete protein family while others are simply specific of a domain. When several domains from different domain databases describe the same thing, the InterPro database groups them in the same box, like the IPR 004827 domain.
- ✓ **The IPR#### hyperlink points to the InterPro documentation.** Here InterPro attempts to summarize the information spread in the documentations of various domain databases. It's always a good idea to read the InterPro documentation, but don't stop there — be sure to read the documentation for the individual sequences as well.
- ✓ **In front of each line is a hyperlink that can take you to the domain entry in database.** For instance, if you click the [PS00138 link](#), your browser takes you to the entry 138 of the PROSITE database, where you can find the individual PROSITE documentation.
- ✓ **On every line, little colored boxes show you where the match occurred on your sequence.** Their size is proportional to the size of the domain. If you pass the mouse pointer over these rectangles, the coordinates of the match (in your sequence) appear on-screen.



**Figure 6-14:**  
Output of the InterProScan Server.

SEQUENCE: FOSE_HUMAN CRC64: DDF827C5047850F LENGTH: 338 aa	
InterPro IPR000209 Domain	Peptidase S8 and S53, subtilisin, kexin, sedolisin PS00138 SUBTILASE_SER
InterPro IPR000937 Family	Fos transforming protein PR00042 LEUZIPPRFOS
InterPro IPR004827 Domain	Basic leucine zipper (bZIP) transcription factor SM00338 PS00038 PS50212 BRLZ BZIP_BASIC BZIP
InterPro IPR006917 Domain	Eukaryotic transcription factor, DNA-binding SSF47454 Euk_transcr_DNA
InterPro IPR011700 Domain	Basic leucine zipper PF07716 bZIP_2

Table View Raw Output XML Output Original Sequences SUBMIT ANOTHER JOB

### *Making sense of the inconsistencies in the InterProScan output*

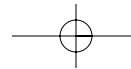
You can see in Figure 6-14 that all domain databases agree on the presence of a series of Leucine zippers that are roughly in the middle of the sequence.

On the other hand, the domain databases do *not* agree *exactly* on the boundaries of these matches. Deciding who is right and who is wrong is always a bit of an art, but in this case a majority vote could be a good indication. (Now you know why using all the domain collections — and not just one — is so important.)



Another reason why seven databases are better than one is that there's always a good chance that one of the databases may contain a domain that isn't in the others. This is especially true with exotic domains that are not yet included in all domain collections.

Of course, the fact that a domain is or is not reported isn't 100 percent proof of anything. (See the sidebar "A common mistake when scanning domains.") To make sure that a reported hit is genuine, we recommend taking a close





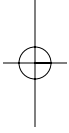
# Finding domains with the CD server

The CD (Conserved Domain) server of the NCBI follows the same principle as the InterProScan. The main advantage of the CD server is that reported hits come with a score that helps you discriminate the good from the spurious matches. On the downside, the CD server doesn't integrate as many databases as InterProScan, although it contains quite a few domains contributed by the NCBI that you can't find anywhere else.

If you're interested in giving the CD server a try, here's how you'd go about it:

1. **Point your browser to** [www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi](http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

You can also access this server from the main NCBI BLAST page. Point your browser to [www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/) and click the [Search the Conserved Domain Database using RPS-BLAST](#) link. This hyperlink is the fourth one in the second column.



## A common mistake when scanning domains

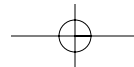
Don't forget that when a domain server reports some hits between your sequence and a domain, this information always results from an alignment between the domain (profile) you've found and the sequence you're using for comparison. If the program believes the alignment is good enough, it will report the match. Otherwise it will not.

Unfortunately, neither the profile nor the programs are perfect — and mistakes occur. The server can tell you that your protein contains a specific domain when that isn't true, or it may fail to report a domain that *is* present in your protein. If you trust the results blindly, there's always a chance you may get it wrong.

The InterProScan server isn't very helpful when it comes to avoiding that type of mistake: It

doesn't report the score of the hits and does not even display the alignments. The other servers we show you in this chapter (CD-Search and Motif Scan) give you a score along with the hits. This score has a statistical meaning, and it informs you how likely it is that your match may have occurred by chance only and is devoid of a biological meaning.

Conservative interpretations (that is, only believing very high scores) are almost always correct. On the other hand, if the score isn't so good — or if only a portion of the domain matches your sequence — you need additional evidence to make sure that what you see is real.



**the Advanced Search Options heading.**

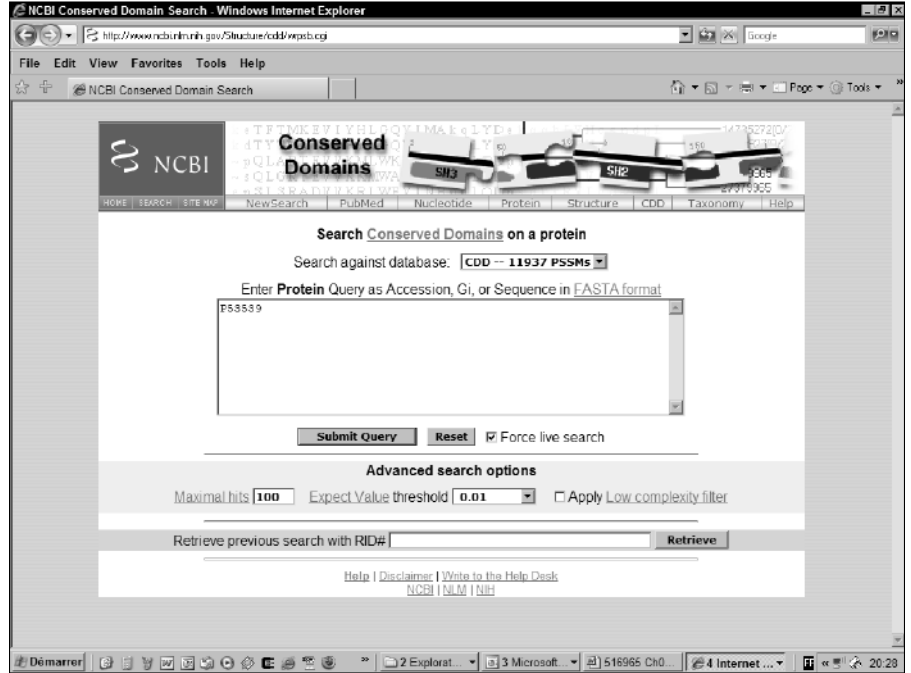
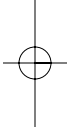
In many interesting domains, a particular type of amino acid is over-represented. For instance, there are more leucines than expected by chance in the leucine zippers, or more glycines than expected by chance in the glycine-rich domains — and so on — for many domains.

Because repeated residues make a sequence simpler, sequences that contain them are described as *low-complexity*. If you keep the Apply Low Complexity Filter box checked, you may lose these domains.

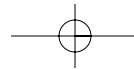
**4. Set the Expect Value Threshold to 1.**

The default value (0.01) can be too stringent, especially when considering low complexity domains.

**5. Click the Submit Query button.**



**Figure 6-15:**  
The CD  
search  
server at  
NCBI.



# Interpreting and understanding CD server results

The principle of the output from the NCBI CD server is very similar to that of InterPro, and is divided into two sections:

✓ **The Graphic:** The Graphic display, as shown in Figure 6-16, shows the regions of your protein that match a domain:

- Red domains are from SMART.
- Ragged ends indicate partial matches.

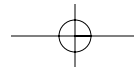
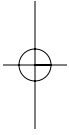
If you pass the mouse pointer over a domain, its complete name appears in the little window along with its score. The score is an *E-value*: It tells you how many times you can expect a hit that good by sheer chance only, given your sequence and the database you scanned.

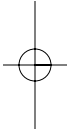
You can interpret your score by using the following rules:

- The lower the E-value, the better the score — and the more likely your hit is to be genuine.
- E-values need to be below 0.01 to mean something.
- The E-value may need to be much lower than 0.01 if you remove the low-complexity filtering (Step 3).
- Treat ragged-end matches with suspicion, especially if they occur in low-complexity regions. In our example, none of the ragged-end matches is significant.

✓ **The Hit List:** The hit list (as shown in Figure 6-17) reports the domains that match your sequence, sorted by E-value. The best hits come first, shown at the top. If you click the hyperlink, a page pops up that contains the documentation associated with this entry.

Clicking the (+) sign next to the hit will extend it into an alignment between your query and some estimated consensus of the domain sequences. This consensus is something like the average domain sequence.





**Figure 6-16:**  
Graphic  
output from  
the NCBI CD  
server.

Name	Title	Pseqid	Accession	E-value
BRLZ	smart00338, BRLZ, basic region leucin zipper...	47665	smart00338	2e-15

CD Search Reference:  
Machler-Bauer A, Bryant SH (2004). "CD-Search: protein domain annotations on the fly." *Nucleic Acids Res.* 32(W):327-331.  
[Help](#) | [Disclaimer](#) | [Write to the Help Desk](#)  
[NCBI](#) | [NLM](#) | [NIH](#)



If you click the Search For Similar Domain Architectures button at the bottom of the hit list, the server will fetch all the protein sequences that contain the same domains as your query sequence. These could be very remote homologues, impossible to detect in another way.

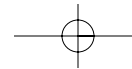
**Figure 6-17:**  
The Hit  
List and  
Alignments  
output from  
the NCBI CD  
server.

CD Length: 65, Pct. Aligned: 96.923077, Bit Score: 77.187452, E-value: 2e-15

```
query 153 EEEKRVVRSFVHKLAKKCNHRELTFRIGATYQLSEAELESEIAEIQGKRELFVY 215
consensus 1 EDEKRRRRRREHRAARRSREKKGAVTEELERKVEQLKACNERLKKQTEQLRRELEKIKSEL 63
```

## Finding domains with Motif Scan

The people who developed Motif Scan are the same folks who maintain PROSITE. The Motif Scan server provides you with the most powerful interface available



1. **Point your browser to** `myhits.isb-sib.ch/cgi-bin/motif_scan`.

The Motif Scan home page appears.

2. **Paste your sequence or its ID number into the Protein Sequence Input text box, as shown in Figure 6-18.**

You have the choice among the following formats:

- Sequence ID
- Raw format (Residues only, space and numbers tolerated)
- FASTA

Motif Scan automatically recognizes the chosen format.

For this example, you can use the FOSB\_HUMAN. To do so, type **P53539** into the Protein Sequence Input box.

3. **In the Database of Motifs section, select the domain collection you want.**

PROSITE is much smaller than Pfam. Selecting PROSITE gives you a much quicker answer than if you also select Pfam.

4. **Click the Search Button.**

Doing so gets you a Results page (eventually).

5. **Print your results.**

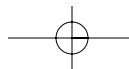
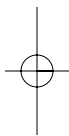
- **Saving the graphic display:** Do a screen dump by pressing the Prt Sc key, then cut and paste it into a PowerPoint presentation.
- **Saving your results:** Cut and paste the List of Matches section.

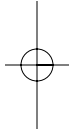
### *Interpreting and understanding the Motif Scan results*

Motif Scan provides one of the richest outputs available for domain analysis. This comes at the cost of an interface that is slightly more difficult to interpret than that provided by CD or InterProScan. (See Figure 6-20)

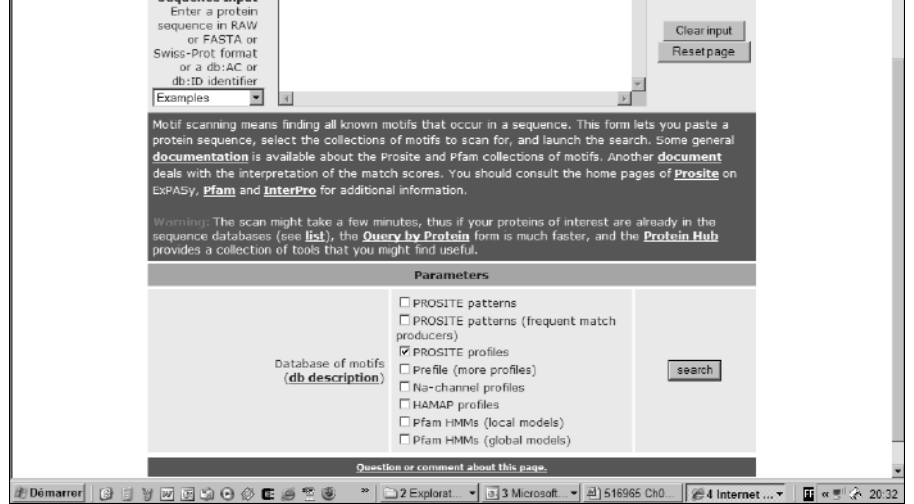
Here are some things to keep in mind when interpreting Motif Scan output:

- ✓ **Match Map:** This match map indicates the location of every domain match on your sequence. Each match is numbered, and a legend is given at the bottom of the match.





**Figure 6-18:**  
Motif Scan  
server at  
the Swiss  
Institute  
of Bioinform-  
atics.



- ✓ **List of Matches:** This is a text version of the match map. Copy it for further reference.
- ✓ **The Match Details Section:** Motif Scan uses the normalized score:
  - High score means a good match.
  - Only scores above 7 are considered good.



Motif Scan doesn't sort the hits according to their scores, so the best matches aren't necessarily at the top of the list. Relevant hits are indicated with an exclamation mark (!).

**Figure 6-19:**  
Output of  
Motif Scan  
(top).

Reference: Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K & Bairoch A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.* 30:235-238

searching PROSITE profiles  
postprocessing

**Summary**

**Original output**

prf

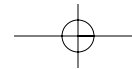
VIRENT 6 -> S

Matches map  
(features from query are above the ruler, matches of the motif scan are below the ruler)

Legenda: 1, CONFLICT L -> R (in Ref. 2); 2, Leucine=aspar; 3, DNA\_BIND Basic motif; 4, prf:ARG\_RICH (!); 5, prf:BZIP (!); 6, prf:GLU\_RICH (?); 7, prf:PRO\_RICH (!); 8, prf:SER\_RICH (!).

**List of matches**

FT	MYHIT	136	182	prf:ARG_RICH (!)
FT	MYHIT	155	218	prf:BZIP (!)
FT	MYHIT	147	212	prf:GLU_RICH (?)
FT	MYHIT	220	274	prf:PRO_RICH (!)
FT	MYHIT	11	34	prf:SER_RICH (?)





Looking for new domains is a bit of an art, mastered by only a few highly trained biologists around the world. But the good news is that everything you want to know is at hand — if you fancy giving it a try. The simplest way is to use BLAST, and turn your database search into what BLAST gurus call a PSSM (and simple folks call a domain). Read the BLAST chapter (Chapter 8) if you are not familiar with this tool, and you will find everything you need to build and use PSSMs at the following online address:

[www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm](http://www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm)

If you go this route, you can do everything online, but don't expect miracles. Domains are like diamonds, scattered here and there in the protein world. While you can expect to occasionally stumble by chance on a nice gem, it will take more than that to uncover the Crown Jewels! If you want to go down that road, you shouldn't mind running a few programs on Unix, digging for sequences in odd DNA sequence databases (like EST databases for instance, see Chapters 3 and 4), gathering your sequences with BLAST, aligning them with a Multiple Sequence Alignment program (see Chapter 9), turning your alignment into a Hidden Markov Model (HMM) with the Hmmer program ([hmmer.wustl.edu/](http://hmmer.wustl.edu/)), and using Hmmer to search protein databases. This is what folks at Pfam do everyday.



They say the Eskimos have 40 words for snow, and can describe even the tiniest differences. It's similar with biologists; you won't be surprised to hear that they have more than one word for protein domains. In the context of a biological paper, you can assume that the words *HMM*, *PSSM*, *profile*, *domain*, *MSA*, *weight matrix*, and *extended profile* mean roughly the same thing.

## More Protein Analysis for Free over the Internet

The Internet offers an extremely large number of resources for doing sequence analysis online — and they're free; we've listed a few in Table 6-2. The following links are only a sample of the most stable sites available to you.



In general, if your work depends on one of these sites, we suggest that you choose a very stable resource. As a rule, avoid sites that run from a personal home page ([www.something.somewhere/~somebody](http://www.something.somewhere/~somebody)) as they're generally less reliable.



Pdb	<a href="http://npsa-pbil.ibcp.fr">npsa-pbil.ibcp.fr</a>	Proteins
PIR	<a href="http://pir.georgetown.edu">pir.georgetown.edu</a>	Proteins
CBS	<a href="http://www.cbs.dtu.dk/services">www.cbs.dtu.dk/services</a>	Proteins
Hits	<a href="http://hits.isb-sib.ch/">hits.isb-sib.ch/</a>	Proteins
InterPro	<a href="http://www.ebi.ac.uk/interpro/scan.html">www.ebi.ac.uk/interpro/ scan.html</a>	Domains
CD search	<a href="http://www.ebi.ac.uk/InterProScan/">http://www.ebi.ac.uk/ InterProScan/</a>	Domains

