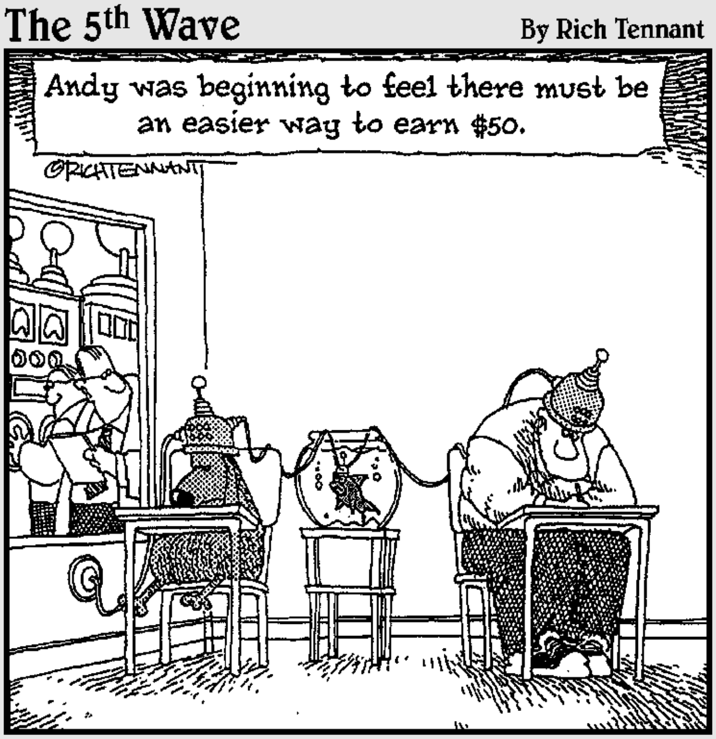
Bioinformatics


The 5th Wave
By Rich Tennant
Andy was beginning to feel there must be an easier way to earn $50.
©RICHTENNANT

In this part . . .

Bioinformatics is a new discipline, which means that nobody should feel ashamed if he or she doesn't have a clue what the excitement's all about. Don't worry; after finishing this book, you'll be speaking bioinformaticsspeak with the best of them.

We start you off in Part I with a quick reminder of what you need to know about DNA and proteins to make sense of this book. We also give you an overview of the main bioinformatics tools available on the Internet.

We don't give too many details here, but if all you need to know is which Internet page to open and which button to press, come on in, 'cuz we've got just what you need!

Finding Out What Bioinformatics Can Do for You

In This Chapter

- Defining bioinformatics
- Understanding the links between modern biology, genomics, and bioinformatics
- Determining which biological questions bioinformatics can help you answer quickly

> *Organic chemistry is the chemistry of carbon compounds. Biochemistry is the study of carbon compounds that crawl.*
>
> — Mike Adam

It looks like *bio*logists are colonizing the dictionary with all these *bio*-words: we have bio-chemistry, bio-metrics, bio-physics, bio-technology, bio-hazards, and even bio-terrorism. Now what's up with the new entry in the bio-sweepstakes, bio-informatics?

What Is Bioinformatics?

In today's world, computers are as likely to be used by biologists as by any other highly trained professionals — bankers or flight controllers, for example. Many of the tasks performed by such professionals are common to most of us: We all tend to write lots of memos and send lots of e-mails; many of us use spreadsheets, and we all store immense amounts of never-to-be-seen-again data in complicated file systems.

However, besides these general tasks, biologists also use computers to address problems that are very specific to biologists, which are of no interest to bankers or flight controllers. These specialized tasks, taken together, make up the field of *bioinformatics.* More specifically, we can define bioinformatics as the computational branch of molecular biology.

This new way of doing biology has certainly become very trendy, but don't think that "trendy" translates into "lightweight" or "flash-in-the-pan." Bioinformatics goes way beyond trendy — it's at the center of the most recent developments in biology, such as the deciphering of the human genome (another buzzword), "system biology" (trying to look at the global picture), new biotechnologies, new legal and forensic techniques, as well as the personalized medicine of the future.

Because of the centrality of bioinformatics to cutting-edge developments in molecular biology, people from many different fields have been stumbling across the term in a variety of different contexts. If you're a biology, medical, or computer science student, a professional in the pharmaceutical industry, a lawyer or a policeman worrying about DNA testing, a consumer concerned about GMOs (Genetically Modified Organisms), or even a NASDAQ investor interested in start-up companies, you'll already have come across the word *bioinformatics.* If you're good at what you do, you'll want to know what all the fuss is about. This chapter, then, is for you.

Instead of a formal definition that would take hours to cover all the ins and outs of the topic, the best way to get a quick feel for what bioinformatics — or swimming, for that matter — is all about is to jump right into the water; that's what we do next. Go ahead and get your feet wet with some basic molecular biology concepts — and the relevant questions intimately connected with such concepts — that all together define bioinformatics.

Analyzing Protein Sequences

If you eat steak, you're intimately acquainted with proteins. (Your taste buds know them intimately anyway, even if your rational mind was too busy with dinner to master the concept.) For you non-steak lovers out there, you'll be pleased to know that proteins abound in fish and vegetables, too. Moreover, all these proteins are made up of the same basic building blocks, called *amino acids.* Amino acids are already quite complex organic molecules, made of carbon, hydrogen, oxygen, nitrogen, and sulfur atoms. So the overall recipe for a protein (the one your rational mind will appreciate, even if your taste buds won't) is something like $C_{1200}H_{2400}O_{600}N_{300}S_{100}$.

and so on. Table 1-1 gives you the list of these 20 building blocks, with their full names, three-letter codes, and one-letter codes (the *IUPAC code,* after the *International Union of Pure and Applied Chemistry* committee that designed it).

Table 1-1 The 20 Amino Acids and Their Official Codes

#	*1-Letter Code*	*3-Letter Code*	*Name*
1	A	Ala	Alanine
2	R	Arg	Arginine
3	N	Asn	Asparagine
4	D	Asp	Aspartic acid
5	C	Cys	Cysteine
6	Q	Gln	Glutamine
7	E	Glu	Glutamic acid
8	G	Gly	Glycine
9	H	His	Histidine
10	I	Ile	Isoleucine
11	L	Leu	Leucine
12	K	Lys	Lysine
13	M	Met	Methionine
14	F	Phe	Phenylalanine
15	P	Pro	Proline
16	S	Ser	Serine
17	T	Thr	Threonine
18	W	Trp	Tryptophan
19	Y	Tyr	Tyrosine
20	V	Val	Valine

Finally, biochemists discovered that these amino acids are linked together as a chain — and that the true identity of a protein is derived not only from its composition, but also from the precise order of its constituent amino acids. The first amino-acid sequence of a protein — insulin — was determined in 1951. The actual recipe for human insulin, from which all its biological properties derive, is the following chain of 110 residues:

insulin = MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERG
FFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLY
QLENYCN

Now, more than 50 years later, analyzing protein sequences like these remains a central topic of bioinformatics in all laboratories throughout the world. (Check out Chapters 2, 4, and 6 through 11 to quickly figure out how to analyze your protein sequence and become a member of the club!)

A brief history of sequence analysis

Besides earning Alfred Sanger his first Nobel Prize, the sequencing of insulin inaugurated the modern era of molecular and structural biology. Traditionally a *soft science* (that is, more tolerant of fuzzy reasoning and hand-waving ambiguity than chemistry or physics), biology got a taste of its first fundamental dataset: molecular sequences. In the early 1960s, known protein sequences accumulated slowly — perhaps a blessing in disguise, given that the computers capable of analyzing them hadn't been developed! In this pre-computer era (from our present perspective, anyway), sequences were assembled, analyzed, and compared by (manually) writing them on pieces of paper, taping them side by side on laboratory walls, and/or moving them around for optimal alignment (now called *pattern matching*).

As soon as the early computers became available (as big as locomotives and just as fast, and with 8K of RAM!), the first computational biologists started to enter these manual algorithms into the memory banks. This practice was brand new — nobody before them had to manipulate and analyze molecular sequences as *texts*. Most methods had to be invented from scratch, and in the process, a new area of research — the analysis of protein sequences using computers — was generated. This was the genesis of bioinformatics.

When you work with databases or analysis programs, you're likely to have some unusual letters popping up now and then in your protein sequences. These letters are either used to designate exotic amino acids, or are used to denote various levels of ambiguity — that is, a total lack of information — about certain positions in the sequence. We've listed these particular letters in the following table.

Seven Codes for Ambiguity or Exceptional Amino Acids

1-Letter Code	***3-Letter Code***	***Meaning***
B	Asn or Asp	Asparagine or aspartic acid
J	Xle	Isoleucine or leucine
O (letter)	Pyl	Pyrrolysine
U	Sec	Selenocysteine
Z	Gln or Glu	Glutamine or glutamic acid
X	Xaa	Any residue
--	-----	No corresponding residue (gap)

The B and Z codes (which are now becoming obsolete) indicated how hard it was to distinguish between Asp and Asn (or Glu and Gln) in the early days of protein sequence determination. In contrast, the J code shows how difficult it is to distinguish between Ile and Leu using mass spectrometry, the latest sequencing technique. The Pyl and Sec exotic amino acids are specified by the UAG (Pyl) and UGA (Sec) stop codons read in a specific context. The *X* code is still very much used as a placeholder letter when you don't know the amino acid at a given position in the sequence. Alignment programs use "–" to denote positions apparently missing from the sequence.

Reading protein sequences from N to C

The twenty amino-acid molecules found in proteins have different *bodies* (their characteristic residues, listed in Table 1-1) — but all have the same pair of *hooks* — NH_2 and COOH. These groups of atoms are used to form the so-called *peptidic bonds* between the successive residues in the sequence. Figure 1-1 shows free individual amino acids floating about, displaying their hooks for all to see.

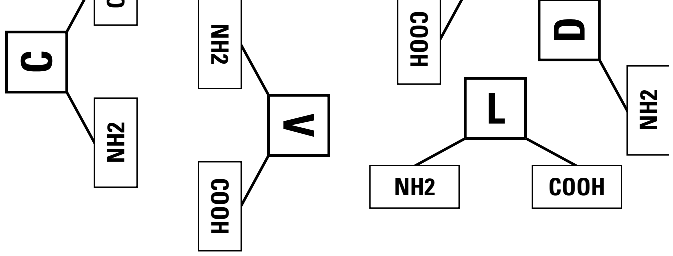


Figure 1-1: Free amino acids floating around.

The protein molecule itself is made when a free NH_2 group links chemically with a COOH group, forming the peptide bond CO-NH. Figure 1-2 shows a schematic picture of the resulting chain.

As a result of this chaining process, your protein molecule is going to be left with an unused NH_2 at one end and an unused COOH at the other end. These extremities are called (respectively) the *N-terminus* and *C-terminus* of the protein chain. This is important to know because scientific convention (in books, databases, and so on) defines the *sequence* of a protein — or of a protein fragment — as the succession of its constituent amino acids, listed in order from the N-terminus to the C-terminus. The sequence of our (short!) demo protein is then

```
MAVLD= Met-Ala-Val-Leu-Asp= Methionine-Alanine-
          Valine-Leucine-Aspartic
```

Working with protein 3-D structures

The precise succession of a protein's constituent amino acids is what defines a given protein molecule. This ribbon of amino acids, however, is not what

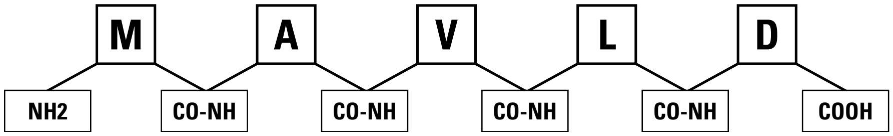


Figure 1-2: Amino acids chained together to constitute a protein molecule.

its sequence because some amino-acid types (for instance, *hydrophobic* residues L, V, I) have no desire whatsoever to be at the surface interacting with the surrounding water — while others (for instance, *hydrophilic* residues D, S, K) are actively looking for such an opportunity. The protein chain is also affected by other influences, such as the electric charges carried by some of the amino acids, or their capacity to fit with their immediate neighbors.

The first 3-D structure of a protein was determined in 1958 by Drs. Kendrew and Perutz, using the complicated technique of X-ray crystallography. (Not for the faint of heart. Don't grapple with how it works unless you want to turn professional!) Besides winning one more Nobel Prize for the nascent field of molecular biology, this feat made the doctors realize that proteins have precise and specific shapes, encoded in the sequence of amino acids. Hence, they predicted that proteins with similar sequences would fold into similar shapes — and, conversely, that proteins with similar structures would be encoded by similar sequences of amino acids. The function of a protein turned out to be a direct consequence of its 3-D structure (shape). The resulting logical linkage

```
SEQUENCE➪STRUCTURE➪FUNCTION
```

was established, and is now a central concept of molecular biology and bioinformatics.

Playing with protein structure models on a computer screen is, of course, much easier than carrying around a thousand-piece, 3-D plastic puzzle. As a consequence, an increasing proportion of the bioinformatics pie is now devoted to the development of cyber-tools to navigate between sequences and 3-D structures. (This specialized area is called *structural bioinformatics.*) Thanks to many free resources on the Internet, it is not difficult to display some beautiful protein pictures on your own computer — and start playing with them as in video games. (We show you how to do that in Chapter 11.)

Before you get a chance to read that chapter, Figure 1-3 gives you an idea of what a 400-amino-acid typical protein 3-D (schematic) structure looks like — when you don't have a color monitor and can't make it move and turn!

REMEMBER

Don't forget: Protein molecules, even in their wonderful complexity, are still pretty small. The one in Figure 1-3 would fit in a box whose sides measure 70/1,000,000 millimeters. There are thousands of different proteins in a single bacterium, each of them in thousands of copies — more than enough evidence that Life Is Not Simple!

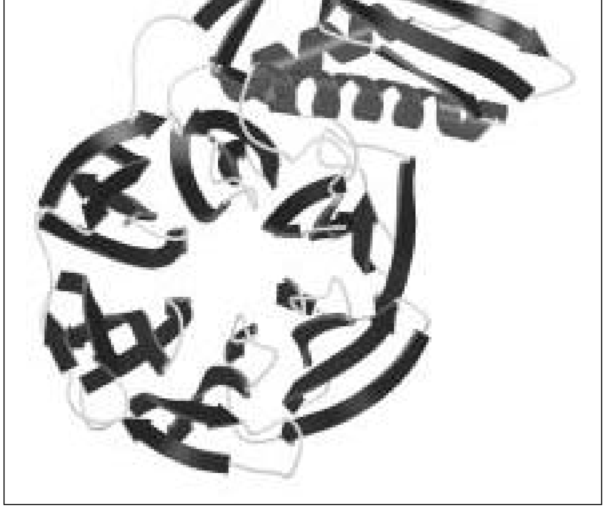


Figure 1-3: Example of protein 3-D structure (schematic).

Protein bioinformatics covered in this book

The study of protein sequences can get pretty complicated — so complicated, in fact, that it would take a pretty thick book to cover all aspects of the field. We'd like to take a more selective approach by focusing on those aspects of protein sequences where bioinformatic analyses can be most useful. The following list gives you a look at some topics where such an analysis is particularly relevant to protein sequences — and also tells which chapters of this book cover those topics in greater detail:

- Retrieving protein sequences from databases (Chapters 2, 3, and 4)
- Computing amino-acid composition, molecular weight, isoelectric point, and other parameters (Chapter 6)
- Computing how hydrophobic or hydrophilic a protein is, predicting antigenic sites, locating membrane-spanning segments (Chapter 6)
- Predicting elements of secondary structure (Chapters 6 and 11)
- Predicting the domain organization of proteins (Chapters 6, 7, 9, and 11)
- Visualizing protein structures in 3-D (Chapter 11)
- Predicting a protein's 3-D structure from its sequence (Chapter 11)

- Finding evolutionary relationships between proteins, drawing proteins family trees (Chapters 7, 9, 11, and 13)

Analyzing DNA Sequences

During the 1950s, while scientists such as Kendrew and Perutz were still struggling to determine the first 3-D structures of proteins, other biologists had already acquired a lot of indirect evidence (via extremely clever genetics experiments) that *deoxyribonucleic acid* (DNA) — the stuff that makes up our genes — was *also* a large macromolecule. It was a long, chainlike molecule twisted into a double helix, and each link in the chain was a pairing of two out of four constituents called *nucleotides.* (A nucleotide is made up of one phosphate group linked to a pentose sugar, which is itself linked to one of 4 types of nitrogenous organic bases symbolized by the four letters A, C, G, and T.)

However, molecular biologists had to wait until much later — the 1970s, to be more precise — before they could determine the sequence of DNA molecules and get direct access to the sequences of gene nucleotides.

This was a revolution (earning A. Sanger his second Nobel Prize!) because the small DNA sequence alphabet (4 nucleotides, as compared to 20 amino acids) allowed a much simpler and faster reading — and quickly lent itself to complete automation. Currently, the worldwide rate of determining DNA sequences is faster (by orders of magnitude) than the rate of protein sequencing.

Reading DNA sequences the right way

As was the case for the 20 amino acids found in proteins, the 4 nucleotides making DNA have different bodies but all have the same pair of hooks: 5' phosphoryl and 3' hydroxyl (pronounced *five prime* and *three prime*) by reference to their positions in the deoxyribose sugar molecule, which is part of the nucleotide chaining device. Figure 1-4 shows what free individual nucleotides look like.

Forming a bond between the 5' and 3' positions of the constituent nucleotides then makes the DNA molecule. Figure 1-5 shows a schematic representation of the resulting DNA strand.

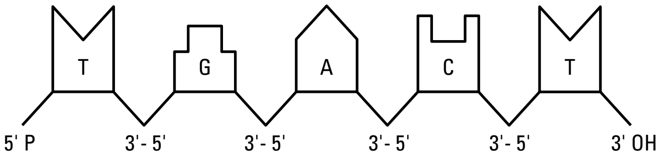


Figure 1-5: Chained nucleotides constituting a DNA strand.

After the nucleotides are linked, the resulting DNA strand exhibits an unused phosphoryl group (PO_4) at the 5' end, and an unused hydroxyl group (OH) at the 3' end. These extremities are respectively called the *5'-terminus* and the *3'-terminus* of the DNA strand.

A DNA sequence is always defined (in books, databases, articles, and programs) as the succession of its constituent nucleotides *listed from the 5'- to 3'- terminus* (that is, end). The sequence of the (short!) DNA strand shown in Figure 1-5 is then

```
TGACT = Thymine-Guanine-Adenine-Cytosine-Thymine
```

The two sides of a DNA sequence

In the same laboratory where Kendrew and Perutz were trying to figure out the first 3-D structure of a protein, Watson and Crick elucidated — in 1953 — the famous double-helical structure of the DNA molecule. These days everybody has a mental picture of this famous spiral-staircase molecule; the elegance of the DNA *double helix* probably helped make it the most popular notion to come out of molecular biology. But what made this discovery so important — earning one more Nobel Prize for molecular biology — was not the helical shape, but the discovery that the DNA molecule consists of two complementary strands, shown in Figure 1-6.

The following table lists the one-letter codes (IUPAC codes) used to work with DNA sequences. Official IUPAC codes, from the International Union of Pure and Applied Chemistry, are defined for all possible two- and three-way ambiguities. The table shows only the ones most frequently used.

Most Common Letters Used for DNA Nucleotide Sequences

1-Letter Code	*Nucleotide Name*	*Category*
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
T	Thymine	Pyrimidine
N	Any nucleotide (any base)	(n/a)
R	A or G	Purine
Y	C or T	Pyrimidine
--	-----	None (gap)

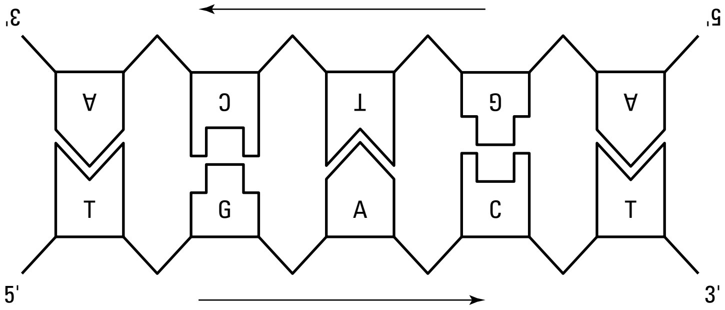


Figure 1-6: The two complementary strands of a complete DNA molecule.

By *complementarity,* we mean that a thymine (T) on one strand is always facing an adenine (A) (and vice versa) — and guanine (G) is always facing a cytosine (C). These couples, A-T and G-C, although not linked by a chemical bond, have a strict one-to-one reciprocal relationship. When you know the sequence of nucleotides along one strand, you can automatically deduce the sequence on the other one. This amazing property — and not the stylish helical structure — is the Rosetta Stone that explains everything about DNA

This double strand structure of DNA makes the definition of a DNA sequence ambiguous: Even with our convention of reading the nucleotides from the 5' end toward the 3' end, you may decide to write down the bottom or the top sequence. Convince yourself that they're both equally valid sequences by turning this book upside down! Thus, at each location, a DNA molecule corresponds to two — totally different — sequences, related by this reverse-and-complement operation. This isn't complicated; simply keep it in mind every time you work with DNA sequences.

Fortunately, most database mining programs, such as BLAST, know about this property, and take both strands into account when reporting their results. But some programs don't bother — and only analyze the sequence you gave them. In cases where both strands matter, always make sure that a complete analysis has been performed. (We discuss these details further in Chapters 3, 5, and 7.)

Palindromes in DNA sequences

Newcomers to DNA sequence analysis are usually confused by the notion of reverse complementary sequences. However, in due time you'll be able to recognize right away that the two sequences

```
ATGCTGATCTTGGCCATCAATG and CATTGATGGCCAAGATCAGCAT
```

correspond to facing strands of the same DNA molecule.

One fascinating property of DNA complementarity is the fact that regions of DNA may correspond to sequences that are identical when read from the two complementary strands. Figure 1-7 helps illustrate this magic trick.

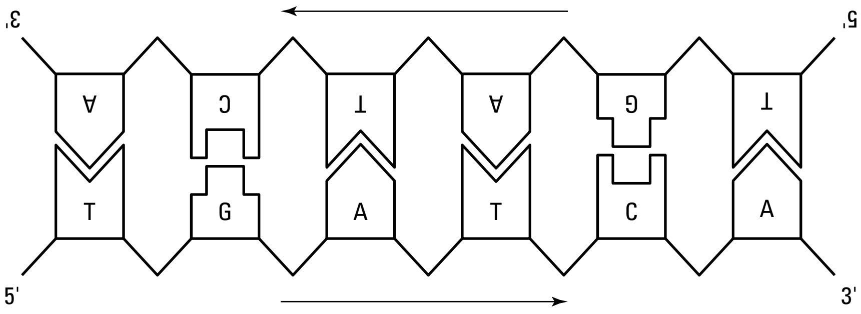


Figure 1-7: How two complementary strands can be read the same way.

tory proteins stick so they can turn genes on and off. Palindromic sequences also have a strong influence on the 3-D structure of DNA molecules. (And not just DNA. See the next section for more on palindromic sequences in RNA.) Looking for exact or approximate palindromes in DNA sequences is a classic bioinformatic exercise.

Analyzing RNA Sequences

DNA (deoxyribonucleic acid) is the most dignified member of the nucleic acid family of macromolecules. Its sole and only task is to ensure — forever — the conservation of the genetic information for its organism. It is thus very stable and resistant, and lies well-protected in the nucleus of each cell. *Ribonucleic acid* (RNA) is a much more active member of the nucleic acid family; it's synthesized and degraded constantly as it makes copies of genes available to the cell factory.

In the context of bioinformatics, there are only two important differences between RNA and DNA:

- RNA differs from DNA by one nucleotide.
- RNA comes as a single strand, not a helix.

The one-letter IUPAC codes for RNA sequences are shown in Table 1-2.

Table 1-2 Most Common Letters Used for RNA Nucleotide Sequences

1-Letter Code	*Nucleotide Base Name*	*Category*
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
U	Uracil	Pyrimidine
N	Any nucleotide	Purine or Pyrimidine

(continued)

Y	C or U	Pyrimidine
--	-------	None (gap)

Some programs automatically handle the U-instead-of-T conversion — and many don't even distinguish between the two classes of nucleic acids. So don't be surprised if a database entry displays RNA sequences (such as messenger RNA) with a T instead of a U. In fact, like proteins, RNA sequences are encoded in the DNA. For this reason, people have adopted the habit of working with the sequences of the RNA *genes* (written in DNA) rather than with RNA sequences.

RNA structures: Playing with sticky strands

Even though RNA molecules consist of single strands of nucleotides, their natural urge for pairing with complementary sequences is still there. Think of each such single strand as a free-floating piece of Scotch tape: You know that it won't take long for that tape to become a messy ball, until no sticky part remains exposed. This is exactly what happens to the single-stranded RNA molecule — more or less (for the sake of poetic license) — although Figure 1-8 shows more precisely how the stickiness works.

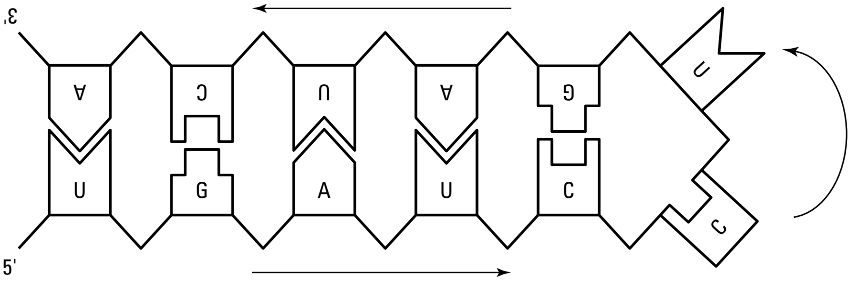


Figure 1-8: How RNA turns itself into a double-stranded structure.

Now you understand why we insisted on the notion of strand complementarity (refer to Figure 1-6). Single-stranded RNA molecules pair different regions of their sequences to form stable double-helical structures — admittedly less regular than (but quite similar to) the double-helical structure of DNA. Once synthesized, each RNA molecule quickly adopts a compact fold — trying to pair as many nucleotides as possible, while keeping the chain not only flexible but true to its own geometry. Hairpin shapes, as shown in Figure 1-8, are that's dictated by its sequences. As with proteins, the linear sequence of the building blocks dictates the final 3-D shape. The biological function of RNA molecules derives from their 3-D shapes or from their sequence complementarity with specific genes.

Computing (predicting) the final fold of an RNA molecule from its sequence is a challenging problem that drove many historical developments in bioinformatics. The recent discovery that small RNA molecules can switch off the activity of a number of genes is what triggered a renewed interest in these sticky sequences. (Go directly to Chapter 12 if your main interest is in RNA bioinformatics.)

More on nucleic acid nomenclature

Don't panic if you get the impression that books, courses, and the technical literature all use many different words and abbreviations to designate the building blocks of nucleic acids: That's actually true — for example, you'll find "base," "base pair," "nucleoside," and "nucleotide" — but note: These different names designate slightly different chemical entities, and those differences are irrelevant for us just now. So far we've used the term *nucleotide* — abbreviated *nt* (as in "a 400-nt-long sequence"). This way of labeling a sequence refers to the length of the DNA (or RNA) molecules in terms of the number of positions they have available for nucleotides. For instance, the sequence in Figure 1-5 is 5 nt long.

Notice that we say *number of positions* rather than *number of nucleotides.* A 400-nt long DNA molecule has 400 positions for nucleotides, but it actually contains twice that many (800) because every position contains a pair of nucleotides. To make this clearer, DNA sequence sizes are often given in *base pairs,* abbreviated *bp.* Thus the DNA sequence in Figure 1-5 is 5 bp long. Larger units, such as *kb* (1000 bp) or *Mb* (mega-bp) are also used.

DNA Coding Regions: Pretending to Work with Protein Sequences

Of the hundreds of thousands of protein sequences found in current databases, only a small percentage correspond to molecules that have actually been isolated by somebody or experimented upon. That's because determining been isolated in a test tube.

Turning DNA into proteins: The genetic code

When you know a DNA sequence, you can translate it into the corresponding protein sequence by using the genetic code, the very same way the cell itself generates a protein sequence. The genetic code is universal (with some exceptions — otherwise life would be too simple!), and it is nature's solution to the problem of how one uniquely relates a 4-nucleotide sequence (A, T, G, C) to a suite of 20 amino acids; we're using symbols (rather than actual chemicals) to do the same. Understanding how the cell does this was one of the most brilliant achievements of the biologists of the 1960s. Yet the final answer can be contained in a (miraculously small) table — as shown in Figure 1-9. Have a look, but feel free to indulge in awed silence as you enter the most sacred monument of modern biology.

Here's how to use the table shown in Figure 1-9: From a given starting point in your DNA sequence, start reading the sequence 3 nucleotides (one *triplet*) at a time. Then consult the genetic code table to read which amino acid corresponds to the current triplet (technically referred to as *codons*). For instance, the following DNA (or messenger RNA) sequence is decoded as follows:

1. **Read the DNA sequence:**

   ATGGAAGTATTTAAAGCGCCACCTATTGGGATATAAG

2. **Decompose it into successive triplets:**

   ATG GAA GTA TTT AAA GCG CCA CCT ATT GGG ATA TAA G . . .

3. **Translate each triplet into the corresponding amino acid:**

   M E V F K A P P I G I STOP

If your DNA sequence is correctly listed in the 5' to 3' orientation, you generate the protein sequence in the conventional N- to C-terminus as well. This approach has an advantage: You don't have to think about these orientation details ever again.

Thus, if you know where a protein-coding region starts in a DNA sequence, your computer can pretend to be a cell and generate the corresponding amino-acid sequence! This simple computer translation exercise is at the

More with coding DNA sequences

Using the example in the first paragraphs of the section “DNA Coding Regions: Pretending to Work with Protein Sequences,” you can see that the resulting protein sequence depends entirely on the way you converted your DNA sequence into triplets before using the genetic code. For instance, using the second position as starting point leads to

```
1- ATGGAAGTATTTAAAGCGCCACCTATTGGGATATAAG

2- A TGG AAG TAT TTA AAG CGC CAC CTA TTG GGA TAT AAG

3-  W K Y L K R H L L G Y K
```

Beginning with the third position (GGA-AGT- . . .) again leads to an entirely different translation.

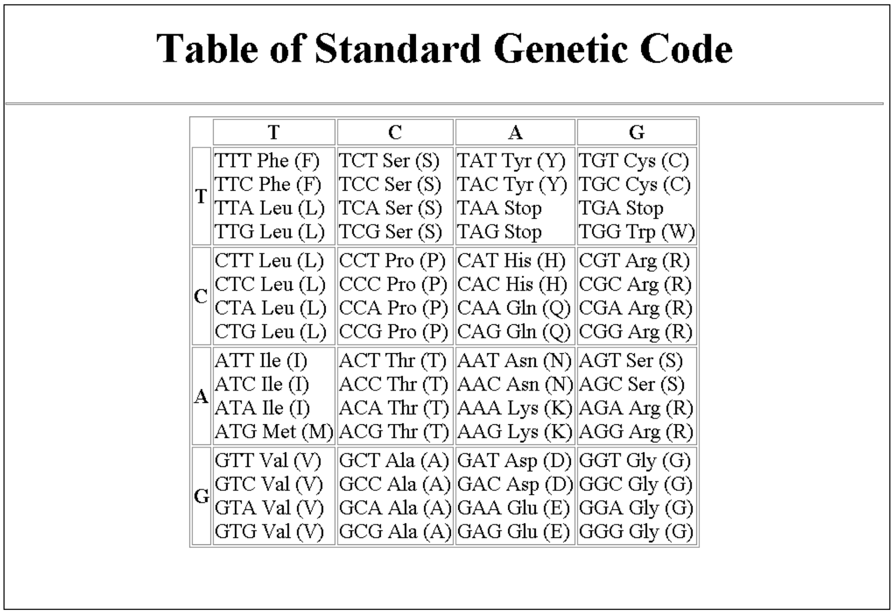


Table of Standard Genetic Code

	T	C	A	G
T	TTT Phe (F) TTC Phe (F) TTA Leu (L) TTG Leu (L)	TCT Ser (S) TCC Ser (S) TCA Ser (S) TCG Ser (S)	TAT Tyr (Y) TAC Tyr (Y) TAA Stop TAG Stop	TGT Cys (C) TGC Cys (C) TGA Stop TGG Trp (W)
C	CTT Leu (L) CTC Leu (L) CTA Leu (L) CTG Leu (L)	CCT Pro (P) CCC Pro (P) CCA Pro (P) CCG Pro (P)	CAT His (H) CAC His (H) CAA Gln (Q) CAG Gln (Q)	CGT Arg (R) CGC Arg (R) CGA Arg (R) CGG Arg (R)
A	ATT Ile (I) ATC Ile (I) ATA Ile (I) ATG Met (M)	ACT Thr (T) ACC Thr (T) ACA Thr (T) ACG Thr (T)	AAT Asn (N) AAC Asn (N) AAA Lys (K) AAG Lys (K)	AGT Ser (S) AGC Ser (S) AGA Arg (R) AGG Arg (R)
G	GTT Val (V) GTC Val (V) GTA Val (V) GTG Val (V)	GCT Ala (A) GCC Ala (A) GCA Ala (A) GCG Ala (A)	GAT Asp (D) GAC Asp (D) GAA Glu (E) GAG Glu (E)	GGT Gly (G) GGC Gly (G) GGA Gly (G) GGG Gly (G)

Figure 1-9: The universal genetic code.

DNA coding region. An interval of DNA sequence that remains free of STOP (the translation of TAA, TGA, or TAG) is called an *open reading frame* (ORF).

Additional complications arise from the fact that some DNA sequences are not encoding proteins at all — and that higher organisms have large pieces of noncoding DNA inserted within their genes. A large part of bioinformatics is devoted to the development of methods to locate protein-coding regions in DNA sequences, to delineate precisely where genes start and end, or where they are interrupted by the noncoding intervals (called *introns*).

DNA/RNA bioinformatics covered in this book

Need a road map to the bioinformatic analyses that are relevant to DNA/RNA sequences covered in this book? Here it is:

- Retrieving DNA sequences from databases (Chapters 2 and 3)
- Computing nucleotide compositions (Chapter 5)
- Identifying restriction sites (Chapter 5)
- Designing polymerase chain-reaction (PCR) primers (Chapter 5)
- Identifying open reading frames (ORFs) (Chapter 5)
- Predicting elements of DNA/RNA secondary structure (Chapter 12)
- Finding repeats (Chapter 5)
- Computing the optimal alignment between two or more DNA sequences (Chapters 7, 8, and 9)
- Finding polymorphic sites in genes (single nucleotide polymorphisms, SNPs) (Chapter 3)
- Assembling sequence fragments (Chapter 5)

Working with Entire Genomes

The first truly efficient technique to sequence DNA was discovered in 1977. In 1995, the first sequence of an entire genome (from the microbe *Hemophilus influenzae*) was determined. Between these two dates, DNA-

- All basic sequence-alignment programs
- Phylogenetic and classification methods
- Various display tools adapted to relatively small-sequence objects (such as protein sequences no more than a few thousand characters long)

Genomics: Getting all the genes at once

The determination of the first complete genome sequence terminated the gene-by-gene routine and initiated the era of *genomics,* the genetic mapping, physical mapping, and sequencing of entire genomes. As a consequence, the DNA sequences we have to work with now are much longer — close to a million-bp in length for microbes and up to several billion-bp in length for animals and humans. This revolution called for the design of new bioinformatic tools and databases capable to store, query, analyze, and display these huge objects in a user-friendly manner. Chapters 3, 5, and 7 present some of the questions that biologists address at the genome scale, and show the relevant bioinformatic tools in action.

In contrast to the early days of the gene-by-gene approach, DNA sequences are now often obtained (along with the presumed protein sequences derived from those DNA sequences) without any prior knowledge of what is actually there. In essence, genes are both sequenced *and discovered* at the same time. This development prompted the emergence of an entirely new branch of bioinformatics devoted to the *parsing* of large DNA sequences into their components (genes, transcription units, protein-coding regions, regulatory elements, and so forth). This first pass is then followed by a longer phase of genome *annotation,* where the biological functions of these various elements are (more or less tentatively) predicted. Part IV of this book presents you with some of these most advanced techniques.

Figure 1-10, representing the whole genome of the bacterium *Rickettsia conorii,* illustrates this new level of complexity. This circular DNA molecule is 1.3 million bp long, on the small side for a bacterium. Each little rectangle in the two most external circles of features (one circle per strand) corresponds to a protein-coding gene in the circular genome. Each rectangle corresponds to approximately 1000 bp. Nobody knew which genes — or which proteins — were in that bacterium before the sequencing started. Almost everything we know now about this bacterium (and many others we can describe as fairly inaccessible, such as those thriving on the ocean floor near volcanic vents at 100°C) has been derived from bioinformatic analyses.

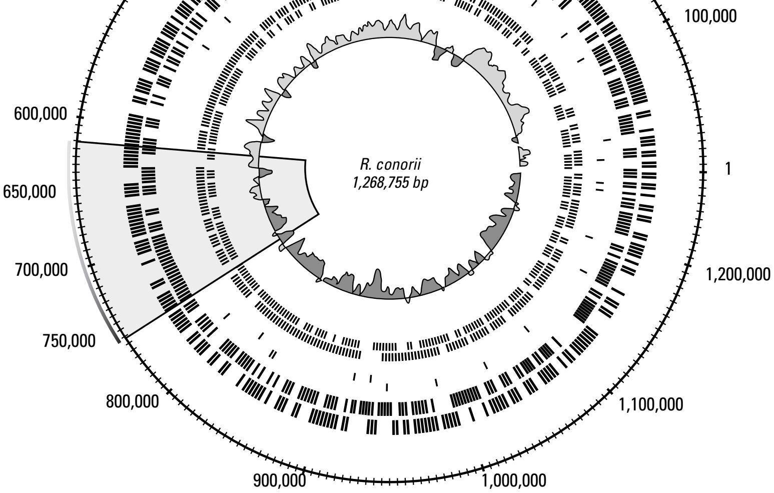


Figure 1-10: Representation of a bacterial genome.

Genome bioinformatics covered in this book

The following list lets you know where in this book you'll find more in-depth coverage of specific topics (some of them bristling with scary, mouth-filling terms) related to genome bioinformatics:

- Finding which genomes are available (Chapter 3)
- Analyzing sequences in relation to specific genomes (Chapters 3 and 7)
- Displaying genomes (Chapter 3)
- Parsing a microbial genome sequence: ORFing (Chapter 5)
- Parsing a eukaryotic genome sequence: GenScan (Chapter 5)
- Finding orthologous and paralogous genes (Chapter 3)
- Finding repeats (Chapter 5)

How Most People Use Bioinformatics

In This Chapter

- Using PubMed to become quickly knowledgeable on any biological subject
- Retrieving protein and DNA sequences relevant to your work
- Running your first sequence/database comparison with BLAST
- Performing your first multiple sequence alignment with ClustalW

> *Personally, I'm always ready to learn, although I do not always like being taught.*
>
> — Sir Winston Churchill

In this chapter, we show you enough bioinformatics to perform the routine — but nonetheless incredibly useful — tasks that experienced molecular biologists once spent months mastering by wandering helplessly on the Net. Chances are that you can fulfill 90 percent of your professional needs without knowing more than the content of this chapter. If you've already got a pretty good chunk of bioinformatics under your belt — enough to qualify as a power user — you may still find useful reminders and tricks here. If your background is in computer science, we show you the kinds of tools that biologists demand. So get on your favorite PC, follow the guide, and enjoy the ride!

Becoming an Instant Expert with PubMed/Medline

Think about the following situation: You just got the sequence of the gene you've been working on for years, and you gave it to the grumpy local specialist in bioinformatics to analyze it for you. (This is, of course, *before* you had an opportunity to read this book!) The specialist's diagnosis comes back, and here it is: "Jim, your sequence looks pretty much like a dUTPase."

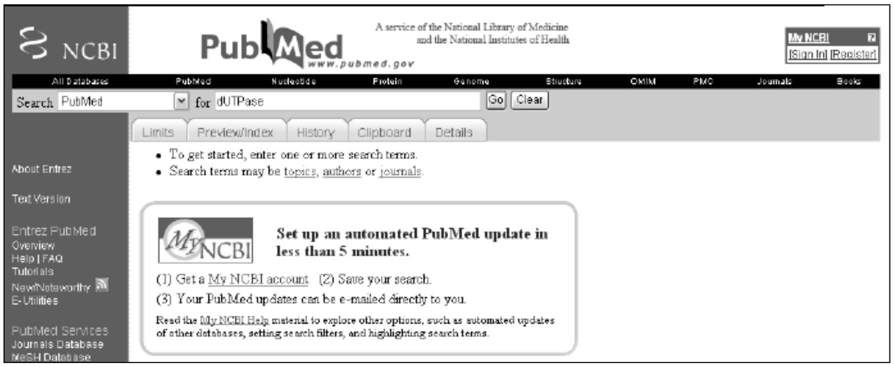
out more about this seemingly obscure term with the strange spelling? How do you quickly decide whether this story makes any sense in the context of what you already know about your gene?

Searching PubMed

The answer is to stay in your chair and go to PubMed! To do so, follow these steps:

1. **Using your favorite Internet browser, navigate to** `www.ncbi.nlm.nih.gov/entrez/` **on the World Wide Web.**

   Your screen now looks like Figure 2-1, with a Search window already preset to PubMed and a For text box next to it, ready for you to type in a few words relevant to your question.



- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.

Set up an automated PubMed update in less than 5 minutes.

(1) Get a My NCBI account (2) Save your search.

(3) Your PubMed updates can be e-mailed directly to you.

Read the My NCBI Help material to explore other options, such as automated updates of other databases, setting search filters, and highlighting search terms.

Figure 2-1: The initial PubMed search window.

2. **Type in** dUTPase **in the For window, and click the Go button.**

   Soon after, a Results list like the one in Figure 2-2 appears on-screen.

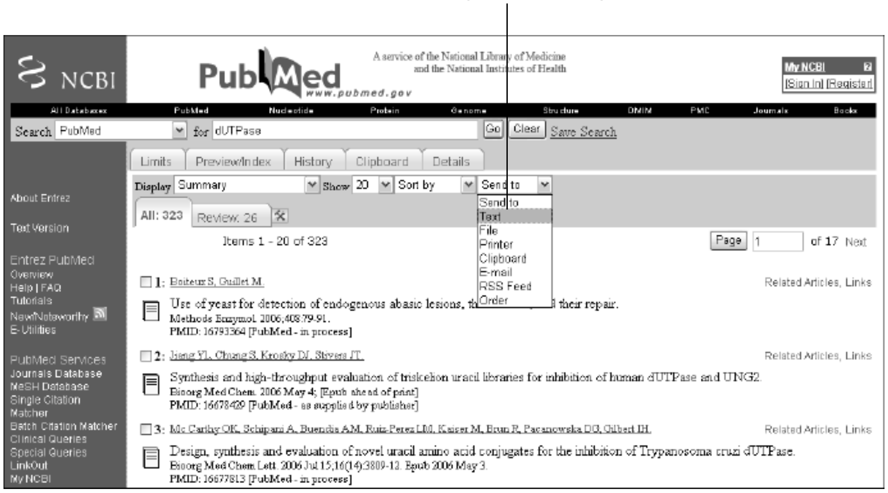
   For our dUTPase example, we now have more than 200 references at our fingertips, more than enough to start unraveling the mysteries of dUTPase — including its relevance to your gene. But don't feel you have to rush to the library with your printed list of references before it closes: You may not even have to go.

3. **For any entry in the Results list, click the associated author names.**

   Author names stand out because they appear in a blue font, underlined — two signs of a clickable hyperlink.

File➪Save As option.

Alternatively, you can transform the display into a simpler (more printer-friendly) format by selecting Text from the Send To drop-down menu and then choosing File➪Save As.



Search PubMed for dUTPase

Display Summary Show 20 Sort by Send to

Items 1 - 20 of 323

Page 1 of 17

□ 1: Boiteux S, Guillet M.

Use of yeast for detection of endogenous abasic lesions, th their repair.
Methods Enzymol. 2006;408:79-91.
PMID: 16793364 [PubMed - in process]

□ 2: Jiang YL, Chung S, Krosky DJ, Stivers JT.

Synthesis and high-throughput evaluation of triskelion uracil libraries for inhibition of human dUTPase and UNG2.
Bioorg Med Chem. 2006 May 4; [Epub ahead of print]
PMID: 16678429 [PubMed - as supplied by publisher]

□ 3: Mc Carthy OK, Schipani A, Buendia AM, Ruiz-Perez LM, Kaiser M, Brun R, Pacanowska DG, Gilbert IH.

Design, synthesis and evaluation of novel uracil amino acid conjugates for the inhibition of Trypanosoma cruzi dUTPase.
Bioorg Med Chem Lett. 2006 Jul 15;16(14):3809-12. Epub 2006 May 3.
PMID: 16677813 [PubMed - in process]

Click here for printer-friendly format.

Figure 2-2: The initial result of a standard PubMed search for dUTPase.

Saving multiple summaries

When your search yields many references, the best move is to start scanning a few pages — and select the most promising papers for future use by checking the corresponding boxes. To move through the Results pages, just use the Next link (top right).

When you think you have selected enough references, just do the following:

1. **Choose Abstract from the Display drop-down menu.**

   The abstracts of all the papers you selected in the various pages appear for you to read.

2. **If you want to print this display as is, choose File➪Print from your browser menu.**

File➪Save As from your browser menu, enter a new filename for the file, and then choose the file format you want to save to.

Although this particular drop-down menu — the one you see fully extended back in Figure 2-2 — provides a File option, it sometimes conflicts with security settings. We advise you to stay away from it and use the standard File➪Save As command on your browser menu.

Searching PubMed using author's names

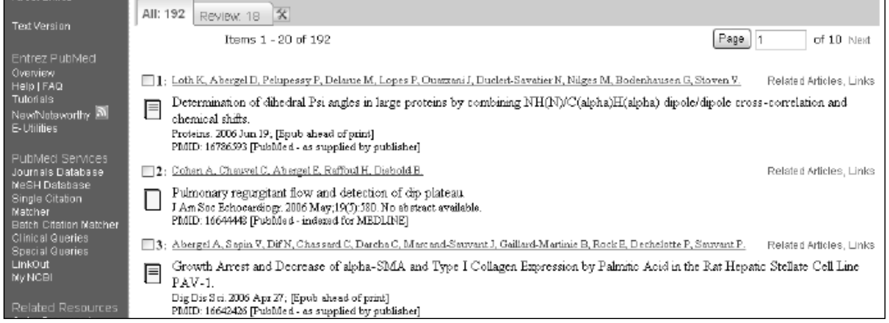
Although you can certainly search PubMed by using topic keywords (such as protein names), you do have other search options available to you. For example, we're sure you've found yourself in the situation where one of your colleagues (or perhaps your advisor) has told you something like the following: "I read a paper by so and so, published not too long ago, and I think it was on dUTPase; you should check it out." In the old-style academic world of card catalogs and stacks of periodicals, this isn't much of a lead. But on the World Wide Web, with PubMed at your side, you're almost home free. Suppose you only remember one of these author's names, such as Abergel. Using PubMed, take the following steps to track down the elusive reference:

1. **Point your browser to** `www.ncbi.nlm.nih.gov/entrez/`**.**
2. **Type** Abergel **— the name of your prospective author — in the For window, and then click the Go button.**

   The Results list appears, as shown in Figure 2-3.

   We now have a list of many papers, all of them with "Abergel" as an author. Again, you can browse through this list, select some of them, and bring out their abstracts. However, the bad news is that none of them appear to have anything to do with dUTPase. To solve this problem, you simply combine the protein name and the author information in your query — and you don't even need to memorize a complicated syntax or Boolean symbols.
3. **Type** dUTPase **next to** Abergel **in the For window — don't forget to put a single space between the search terms — and click Go.**

   By default, PubMed assumes that you want to use both search terms — `Abergel` AND `dUTPase` — when looking for the appropriate research papers. By default, PubMed also assumes that you want to look for these words in the titles OR in the abstracts of the papers. Figure 2-4 shows the results of your first combined search.



All: 192 Review: 18

Items 1 - 20 of 192

Page 1 of 10 Next

□ 1: Loth K, Abergel D, Pelupessy P, Delarue M, Lopes P, Ouazzani J, Duclert-Savatier N, Nilges M, Bodenhausen G, Stoven V. Related Articles, Links

Determination of dihedral Psi angles in large proteins by combining NH(N)/C(alpha)H(alpha) dipole/dipole cross-correlation and chemical shifts.
Proteins. 2006 Jun 19; [Epub ahead of print]
PMID: 16786393 [PubMed - as supplied by publisher]

□ 2: Cohen A, Chauvel C, Abergel E, Raffoul H, Diebold B. Related Articles, Links

Pulmonary regurgitant flow and detection of dip plateau.
J Am Soc Echocardiogr. 2006 May;19(5):580. No abstract available.
PMID: 16644448 [PubMed - indexed for MEDLINE]

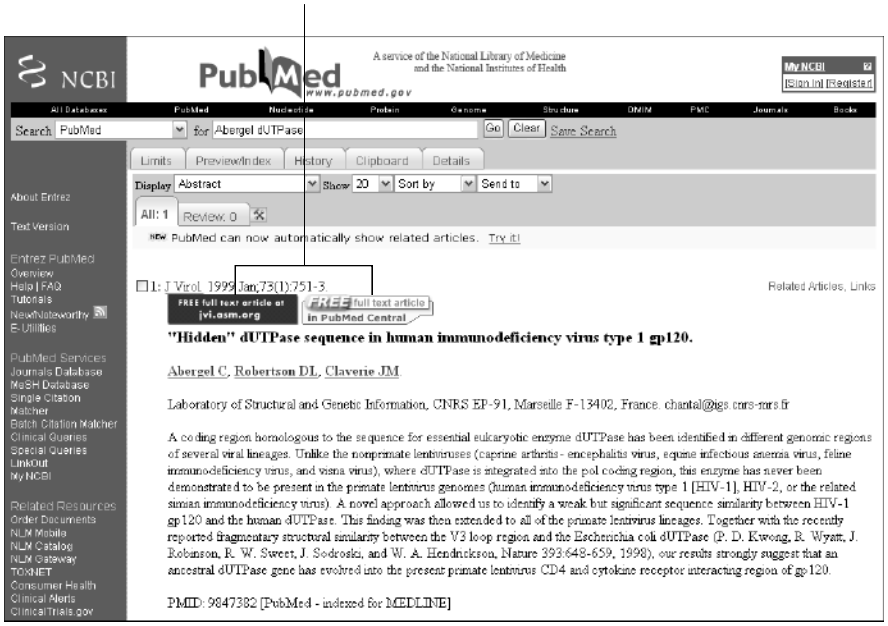
□ 3: Abergel A, Sapin V, Dif N, Chassard C, Darcha C, Marcand-Sauvant J, Gaillard-Martinie B, Rock E, Dechelotte P, Sauvant P. Related Articles, Links

Growth Arrest and Decrease of alpha-SMA and Type I Collagen Expression by Palmitic Acid in the Rat Hepatic Stellate Cell Line PAV-1.
Dig Dis Sci. 2006 Apr 27; [Epub ahead of print]
PMID: 16642426 [PubMed - as supplied by publisher]

Figure 2-3: The results of a standard PubMed search for Abergel.

Because only one paper corresponds to your query, PubMed automatically switches from the "summary" to the "abstract" mode, and you can now read what this unique paper is all about.

This must be the paper your boss was talking about. Congratulations! You've just accomplished your first bibliographic search with PubMed!


Click here for full text.

Search PubMed for Abergel dUTPase

Display Abstract Show 20

All: 1 Review: 0

NEW PubMed can now automatically show related articles. Try it!

□ 1: J Virol. 1999 Jan;73(1):751-3. Related Articles, Links

"Hidden" dUTPase sequence in human immunodeficiency virus type 1 gp120.

Abergel C, Robertson DL, Claverie JM.

Laboratory of Structural and Genetic Information, CNRS EP-91, Marseille F-13402, France. chantal@igs.cnrs-mrs.fr

A coding region homologous to the sequence for essential eukaryotic enzyme dUTPase has been identified in different genomic regions of several viral lineages. Unlike the nonprimate lentiviruses (caprine arthritis- encephalitis virus, equine infectious anemia virus, feline immunodeficiency virus, and visna virus), where dUTPase is integrated into the pol coding region, this enzyme has never been demonstrated to be present in the primate lentivirus genomes (human immunodeficiency virus type 1 [HIV-1], HIV-2, or the related simian immunodeficiency virus). A novel approach allowed us to identify a weak but significant sequence similarity between HIV-1 gp120 and the human dUTPase. This finding was then extended to all of the primate lentivirus lineages. Together with the recently reported fragmentary structural similarity between the V3 loop region and the Escherichia coli dUTPase (P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson, Nature 393:648-659, 1998), our results strongly suggest that an ancestral dUTPase gene has evolved into the present primate lentivirus CD4 and cytokine receptor interacting region of gp120.

PMID: 9847382 [PubMed - indexed for MEDLINE]

Figure 2-4: The result of a standard combined search for Abergel and dUTPase.

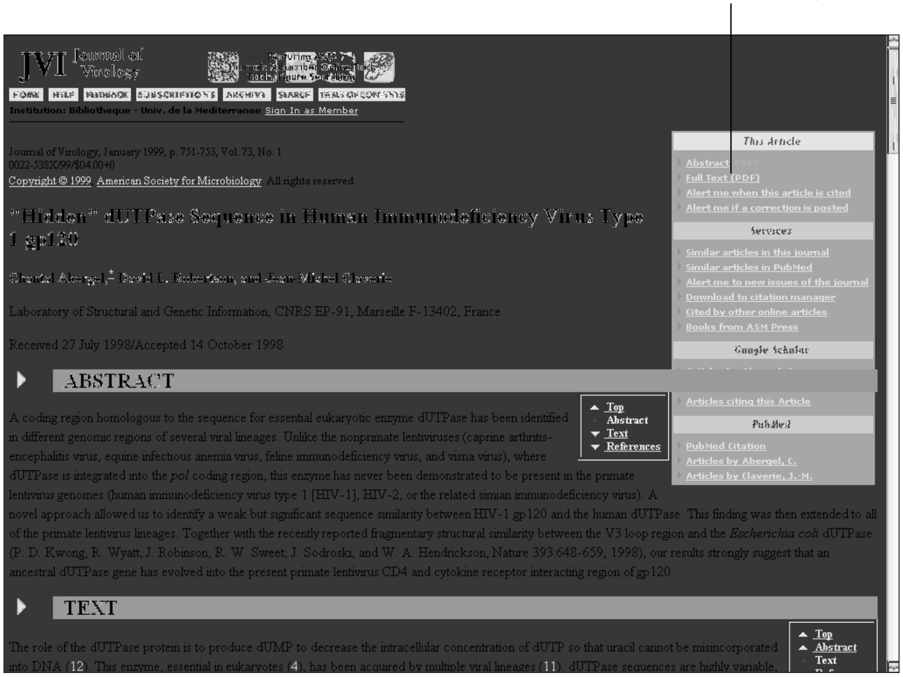
4. **Click the blue rectangle on the left.**

   This brings you directly to the publisher's site, where you read an interactive online copy (HTML) or download a reprint in .PDF format, among other choices (Figure 2-5). An increasing number of scientific journals — including some leading journals in their respective fields — are offering full-text access to articles. However, most of them will still ask you to pay a fee, to be a registered client (for instance, through your university), or to hold a personal subscription. For some journals (such as the one in Figure 2-5), full-text access becomes free six months after print publication.

   By the way, clicking on the rectangle next to the blue one gives you access to the same article, this time extracted from the public, open-access bibliographic repository PubMed Central.

   Thus, if you're lucky, you can go through the entire bibliographic process in a flash — and find out all you need about some mysterious biological subject — without having to leave the comfort of your chair, thanks to PubMed and the cooperating publishers.

Click here for a PDF reprint.


JVI Journal of Virology
Featuring ASM Journals Subscriber Online Tools Tables/Figure Searching
HOME
HELP
FEEDBACK
SUBSCRIPTIONS
ARCHIVE
SEARCH
TABLE OF CONTENTS
This Article
Abstract FREE
Full Text (PDF)
Alert me when this article is cited
Alert me if a correction is posted
Services
Similar articles in this journal
Similar articles in PubMed
Alert me to new issues of the journal
Download to citation manager
Cited by other online articles
Books from ASM Press
Google Scholar
Articles citing this Article
PubMed
PubMed Citation
Articles by Abergel, C.
Articles by Claverie, J.-M.
Top
Abstract
Text
References

Institution: Bibliotheque - Univ. de la Mediterranee Sign In as Member

Journal of Virology, January 1999, p. 751-753, Vol. 73, No. 1
0022-538X/99/$04.00+0
Copyright © 1999, American Society for Microbiology. All rights reserved.

"Hidden" dUTPase Sequence in Human Immunodeficiency Virus Type 1 gp120

Chantal Abergel,* David L. Robertson, and Jean-Michel Claverie

Laboratory of Structural and Genetic Information, CNRS EP-91, Marseille F-13402, France

Received 27 July 1998/Accepted 14 October 1998

ABSTRACT

A coding region homologous to the sequence for essential eukaryotic enzyme dUTPase has been identified in different genomic regions of several viral lineages. Unlike the nonprimate lentiviruses (caprine arthritis-encephalitis virus, equine infectious anemia virus, feline immunodeficiency virus, and visna virus), where dUTPase is integrated into the *pol* coding region, this enzyme has never been demonstrated to be present in the primate lentivirus genomes (human immunodeficiency virus type 1 [HIV-1], HIV-2, or the related simian immunodeficiency virus). A novel approach allowed us to identify a weak but significant sequence similarity between HIV-1 gp120 and the human dUTPase. This finding was then extended to all of the primate lentivirus lineages. Together with the recently reported fragmentary structural similarity between the V3 loop region and the *Escherichia coli* dUTPase (P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson, Nature 393:648-659, 1998), our results strongly suggest that an ancestral dUTPase gene has evolved into the present primate lentivirus CD4 and cytokine receptor interacting region of gp120.

TEXT

The role of the dUTPase protein is to produce dUMP to decrease the intracellular concentration of dUTP so that uracil cannot be misincorporated into DNA (12). This enzyme, essential in eukaryotes (4), has been acquired by multiple viral lineages (11). dUTPase sequences are highly variable.

Figure 2-5: The Journal of Virology grants free full-text access.

Searching PubMed using fields

In addition to searching by author and topic, PubMed offers you many more ways to narrow down your bibliographic searches to more specific topics. To get a handle on these different ways, however, we need to know more about the internal structure of Medline entries — the pool of entries from which PubMed gathers its search results.

To gain some insight into this internal structure, refer to Figure 2-4. We had just used the PubMed site (`www.ncbi.nlm.nih.gov/entrez/`) to track down a single article after inputting *Abergel* and *dUTPase* as a query. With that Results list on your computer screen, you can now do the following:

1. **Click the small arrow to the right of the Display drop-down menu.**

   The contents of the drop-down menu appear: These options are different ways to display information related to the current article, or are links to related information.

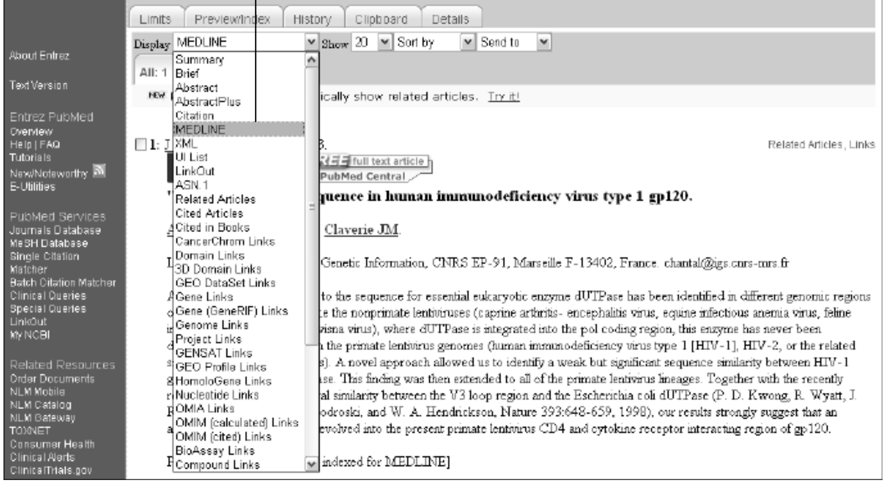
2. **Select the MEDLINE option (Figure 2-6).**

   The Medline page appears, as shown in Figure 2-7.

In Figure 2-7 you can see the internal structure of a Medline database record. The information is spread out over separate sections, called *fields,* each one preceded by a specific abbreviation — `TI` for title fields, `AB` for abstracts, `AD` for the laboratory address, `AU` for the authors, `SO` for the journal abbreviation, and so forth. This structure applies to all Medline records.

Putting fields to good use

Unfortunately, not all PubMed searches are as easy and productive as the one we used in the dUTPase example. You can get flooded with an overwhelming number of *hits* (the standard term for search results) if you formulate queries that contain common names (such as *Smith* or *Cohen*) or use search terms (such as *Down*) that can occur in different contexts — such as titles (for example, *Down Syndrome*), abstracts, author names, or even addresses (*955 Down Street*).



1:

quence in human immunodeficiency virus type 1 gp120.

Claverie JM.

Genetic Information, CNRS EP-91, Marseille F-13402, France. chantal@igs.cnrs-mrs.fr

to the sequence for essential eukaryotic enzyme dUTPase has been identified in different genomic regions
e the nonprimate lentiviruses (caprine arthritis- encephalitis virus, equine infectious anemia virus, feline
visna virus), where dUTPase is integrated into the pol coding region, this enzyme has never been
the primate lentivirus genomes (human immunodeficiency virus type 1 [HIV-1], HIV-2, or the related
s). A novel approach allowed us to identify a weak but significant sequence similarity between HIV-1
se. This finding was then extended to all of the primate lentivirus lineages. Together with the recently
al similarity between the V3 loop region and the Escherichia coli dUTPase (P. D. Kwong, R. Wyatt, J.
odroski, and W. A. Hendrickson, Nature 393:648-659, 1998), our results strongly suggest that an
evolved into the present primate lentivirus CD4 and cytokine receptor interacting region of gp120.

indexed for MEDLINE]

Figure 2-6: Changing the default abstract format to MEDLINE.

To alleviate this problem, you can query PubMed by restricting the search for each word of your query to a given field. Thus papers containing the word at the wrong place are not selected. To do so, you simply have to follow each term with the code (in brackets) that identifies the field in which you want to find the search term. Changing the field can totally modify the result of the searches. For instance, try entering three different queries — `Down [AU]`, `Down [TI]`, and `Down [AD]` into the For text box at the PubMed site.

When we ran those queries, we got (respectively) 275, 16,318, and 1,213 totally unrelated references. At least that narrowed the search a bit.

Using fields to find experts near you

Field-restricted queries are great for dealing with author's names or subject terms that can be found in different fields (which would otherwise inflate and confuse the search results). But field-restricted queries have many other applications, as well — they're especially useful for locating experts in particular research fields in given locations.

```
PUBM  - Print
IS    - 0022-538X (Print)
VI    - 73
IP    - 1
DP    - 1999 Jan
TI    - "Hidden" dUTPase sequence in human immunodeficiency virus type 1 gp120.
PG    - 751-3
AB    - A coding region homologous to the sequence for essential eukaryotic enzyme
        dUTPase has been identified in different genomic regions of several viral
        lineages. -------------------------------------------------------------------------------
        -----, our results strongly suggest that an ancestral dUTPase
        gene has evolved into the present primate lentivirus CD4 and cytokine
        receptor interacting region of gp120.
AD    - Laboratory of Structural and Genetic Information, CNRS EP-91, Marseille
        F-12302, France. chantal@igs.cnrs-mrs.fr
FAU   - Abergel, C
AU    - Abergel C
FAU   - Robertson, D L
AU    - Robertson DL
FAU   - Claverie, J M
AU    - Claverie JM
LA    - eng
PT    - Journal Article
PL    - UNITED STATES
TA    - J Virol
JT    - Journal of virology.
JID   - 0113724
RN    - 0 (HIV Envelope Protein gp120)
RN    - EC 3.6.1. - (Pyrophosphatases)
RN    - EC 3.6.1.23 (dUTP pyrophosphatase)
SB    - IM
SB    - X
MH    - Amino Acid Sequence
MH    - HIV Envelope Protein gp120/*chemistry
MH    - HIV-1/* chemistry
MH    - Humans
MH    - Molecular Sequence Data
MH    - Pyrophosphatases/*chemistry/genetics
MH    - Research Support, Non-U.S. Gov't
EDAT  - 1998/12/16
MHDA  - 1998/12/16 00:01
PST   - ppublish
SO    - J Virol. 1999 Jan;73(1):751-3.
```

Figure 2-7: The internal structure of a Medline record.

2. **In the For window, enter** dUTPase [TIAB] Chicago [AD]**, and click Go.**

   By specifying the `[TIAB]` field, you'll be scanning the titles and abstracts of potential articles; the `[AD]` field specifies the address of the main laboratory associated with these articles.

   A couple of papers show up (Figure 2-8).

3. **Click the list of authors (in blue, underlined).**

   You get the abstract of the corresponding article, together with the main laboratory address. Alternatively, to get all the abstracts at once, follow Steps 4 and 5.

4. **Go back to the previous URL (after Step 2) using the Go Back button of your browser.**

5. **In the Display drop-down menu, change the display option from *Summary* to *Abstract.***

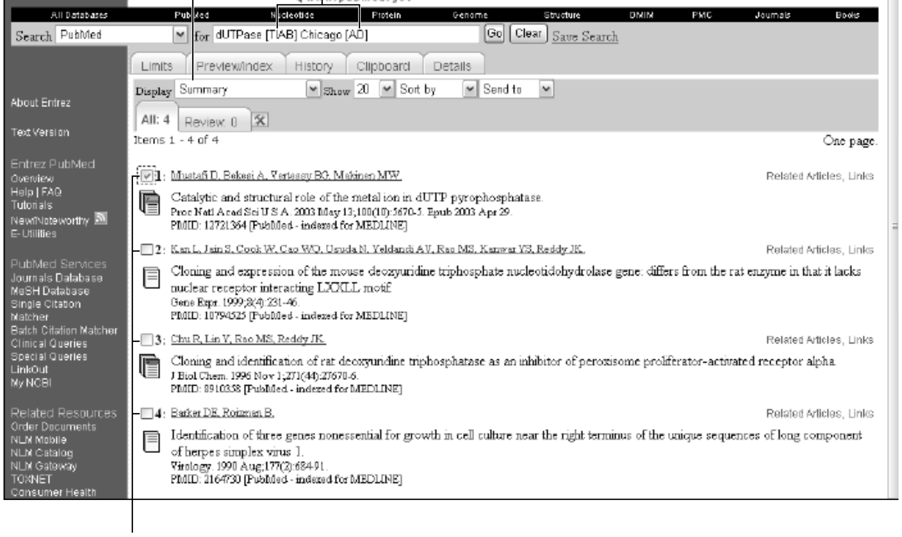
   A list of abstracts appears. At the top of each abstract, you get the name of the relevant laboratory in Chicago.

   With the information contained in the Medline records, you have names, street addresses, and sometimes e-mail addresses. If necessary, you can use a telephone book (or a Web search engine) to supplement this information and find out how to contact these experts — the same day. (But please don't call these people on our behalf!)

Searching PubMed using limits

PubMed offers yet another way of fine-tuning your queries: You can pre-define ranges for different attributes in different fields before you run your search. Said like this, the process sounds awfully technical and complex — but here's where we show you how to do it (again using our dUTPase example).

When you want to quickly find out the basics about a subject you know nothing about, it's best not to read any single highly specialized article. A better bet is to stick with review articles, where one expert summarizes the state of the art for you. Unless you're an historian, you won't have much need to know about the state of the art as it was 20 years ago (which seems more like three centuries ago in the field of molecular biology). Thus, we want to limit our PubMed search to *recent* review articles about dUTPase. Here's how it's done:



Search PubMed for dUTPase [TIAB] Chicago [AD]

Display Summary Show 20 Sort by Send to

All: 4 Review: 0

Items 1 - 4 of 4 One page.

☑ 1: Mustafi D, Bekesi A, Vertessy BG, Makinen MW. Related Articles, Links

Catalytic and structural role of the metal ion in dUTP pyrophosphatase.
Proc Natl Acad Sci U S A. 2003 May 13;100(10):5670-5. Epub 2003 Apr 29.
PMID: 12721364 [PubMed - indexed for MEDLINE]

□ 2: Kan L, Jain S, Cook W, Cao WQ, Usuda N, Yeldandi AV, Rao MS, Kanwar YS, Reddy JK. Related Articles, Links

Cloning and expression of the mouse deoxyuridine triphosphate nucleotidohydrolase gene: differs from the rat enzyme in that it lacks nuclear receptor interacting LXXLL motif
Gene Expr. 1999;8(4):231-46.
PMID: 10794525 [PubMed - indexed for MEDLINE]

□ 3: Chu R, Lin Y, Rao MS, Reddy JK. Related Articles, Links

Cloning and identification of rat deoxyuridine triphosphatase as an inhibitor of peroxisome proliferator-activated receptor alpha.
J Biol Chem. 1996 Nov 1;271(44):27670-6.
PMID: 8910358 [PubMed - indexed for MEDLINE]

□ 4: Barker DE, Roizman B. Related Articles, Links

Identification of three genes nonessential for growth in cell culture near the right terminus of the unique sequences of long component of herpes simplex virus 1.
Virology. 1990 Aug;177(2):684-91.
PMID: 2164730 [PubMed - indexed for MEDLINE]

Figure 2-8: Searching for experts on dUTPase in Chicago.

Check these boxes to select papers.

1. **Point your browser to the PubMed site at**

   `www.ncbi.nlm.nih.gov/entrez/`

2. **Type** dUTPase **in the For text box.**

   If you clicked the Go button at this point, you'd get a list of more than 320 articles — a bit too much reading for a nonspecialist! Restricting this list to the most recent review articles written in English would be practical.

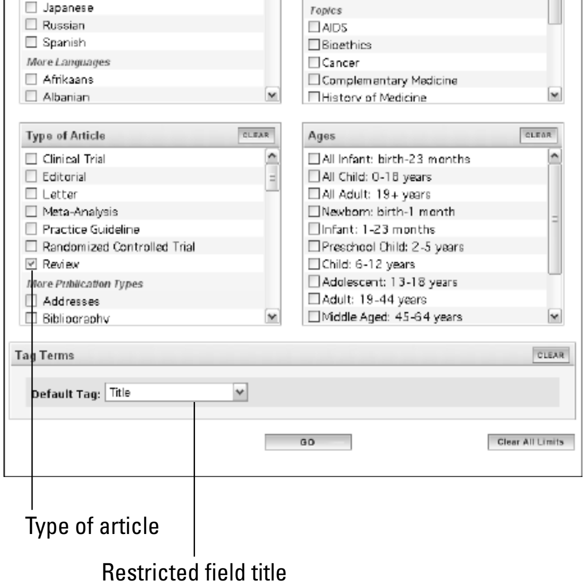
3. **Click the Limits tab, located just beneath the little arrow for the pull-down menu of the Search window.**

   The Limits screen appears, as shown in Figure 2-9. Here, below the Limited To line, you now have plenty of fields and attributes you can use for setting limits. You can go back later to this page and explore these various options.

4. **Check the English box in the Language section.**

   Unless you're fluent in French, *bien sûr.*

5. **Check the Review box in the Type of Article section.**



Default Tag: Title

Figure 2-9: Limiting a search for dUTPase to the titles of review articles in English.

6. **Choose Title from the Default Tag drop-down menu.**

   Choosing Title restricts the search to articles for which the topic of dUTPase is central.

7. **Finally, click the Go button.**

   Your (new and more concise) Results page appears.

In our hands, searching for dUTPase using these limits resulted in five articles — much better than the 200 hits we would have gotten without the limits. More important, we've limited the search to *review* articles, which hopefully contain — in concise form — all we'd ever want to know about this subject.

A common use of the Limit menu is also to restrict the search for papers published within a given date range (such as the most recent).

buttons — especially the ones we haven't talked about — might do. Here are a few last-minute tips:

- **How to get the most out of your query:**
  - Quoted queries (for example, `"down syndrome"`) behave as a single word, and are a great way to improve the relevance of your search.
  - Impress your colleagues by starting using logical connectors (`AND`, `OR`, `NOT`) in your queries, as in

    dUTPase[TI] OR pyrophosphatase[TI] NOT Smith[AU]
  - Adding initials to proper names (for example, `"Abergel C"`) can greatly reduce the number of hits.
  - Write down the PubMed Identifier (the number in the PMID field) of that interesting paper you just found. It can be very useful in any subsequent searches for related items, such as associated gene and protein sequences.
  - Don't forget to deselect the Limit box when starting a new search.
  - Don't put too much initial faith in a search that produces no results. Spelling mistakes, wrong field restrictions, or improper limits settings can all throw off an effective search.
  - As a beginner researching a new subject, read through a couple of abstracts to enlarge your initial "jargon" vocabulary — and look for synonyms. For example, if you don't know that some papers on dUTPase might use the term "dUTP pyrophosphatase" instead, you may miss out on some interesting papers.
  - Try the Related Articles link — the one to the extreme right of the PubMed output — to enlarge a search that isn't giving you enough references.
- **Things you (unfortunately) *won't* find in PubMed:**
  - **Names ranking beyond the 10th place in the author's list for older papers (before 1995).** This is a significant problem for many pioneering articles in genomics.
  - **Papers recorded before 1965.** PubMed doesn't have any. Don't rely on PubMed as a primary source if you're writing an historical article in your field.
  - **Abstracts for most references recorded before 1976.** Don't expect great results from PubMed searches involving these.

the subject at the molecular level. Although PubMed provides you with a direct link to the protein world, we find that it's sometimes confusing to use. So here we introduce you to another site, where finding protein sequences is really easy.

ExPASy: A prime Internet site for protein information

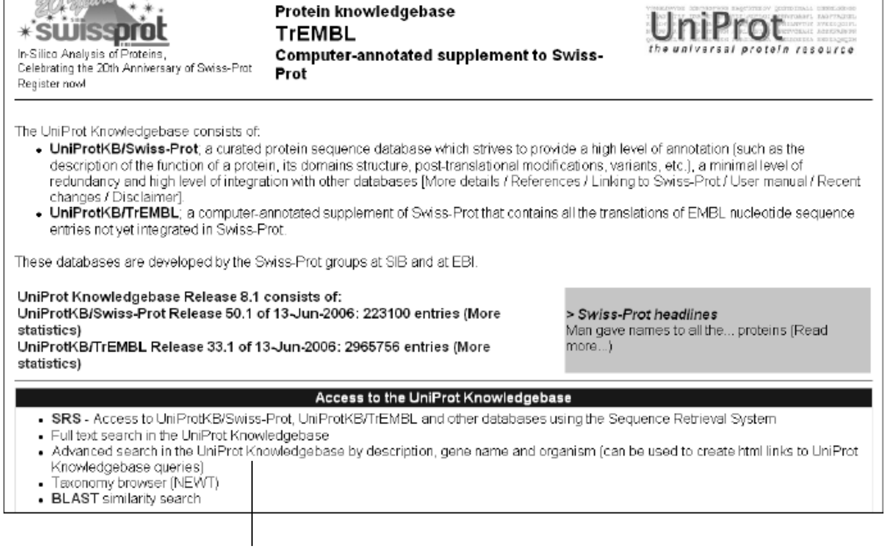
Created and managed by one of the pioneers of protein bioinformatics, Prof. Amos Bairoch, the ExPASy server is a world-leading resource for protein information. In addition to the Swiss-Prot protein sequence database, the ExPASy site provides numerous analysis tools that we use throughout this book. It also provides a wealth of outside links for more specialized analyses on other servers.

After using PubMed to acquire some preliminary information about a particular function that you're interested in — dUTPase, for example — go ahead and find out more about it by retrieving a few examples of protein sequences that perform this function. (Don't worry — we show you how.) Because *Escherichia coli* (abbreviated *E. coli*) is one of the most studied organisms, we use it in our example to show you how to retrieve the sequence of the protein performing the dUTPase function in this bacterium.

1. **Point your browser to** `www.expasy.org/sprot/`**, the Swiss-Prot database home page (Figure 2-10).**
2. **Type dUTPase coli in the Search window, and then click the Search button.**

   A list of three relevant protein sequences (as shown in Figure 2-11) appears.

   Now, when you click the last DUT_ECOLI (P06968) link, a full page of information about this dUTPase protein *of E. coli* appears on-screen, as shown in Figures 2-12 and 2-13. (No single browser window can hold such a wealth of information, so we've broken up the Results page into two figures.)



In-Silico Analysis of Proteins,
Celebrating the 20th Anniversary of Swiss-Prot
Register now!

Protein knowledgebase
TrEMBL
Computer-annotated supplement to Swiss-Prot

The UniProt Knowledgebase consists of:

- **UniProtKB/Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [More details / References / Linking to Swiss-Prot / User manual / Recent changes / Disclaimer].
- **UniProtKB/TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups at SIB and at EBI.

UniProt Knowledgebase Release 8.1 consists of:
UniProtKB/Swiss-Prot Release 50.1 of 13-Jun-2006: 223100 entries (More statistics)
UniProtKB/TrEMBL Release 33.1 of 13-Jun-2006: 2965756 entries (More statistics)

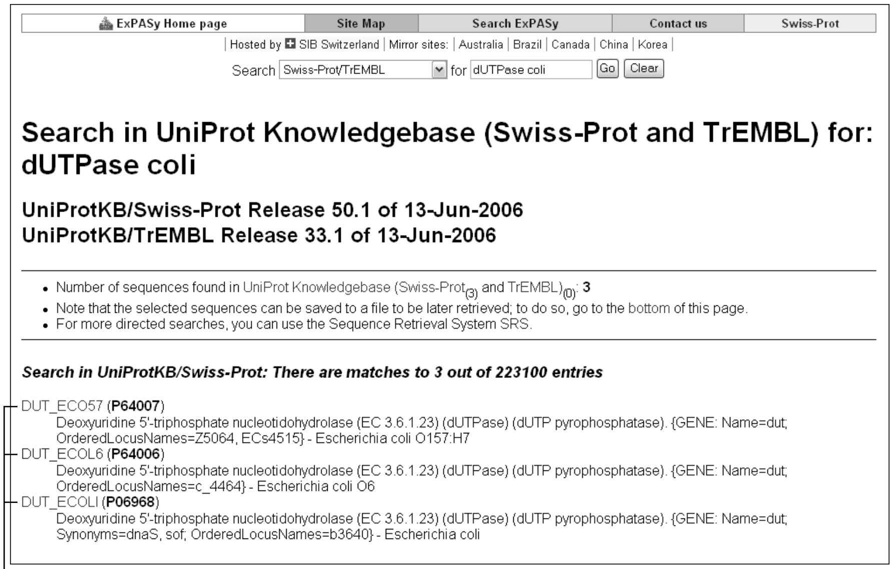
> Swiss-Prot headlines
Man gave names to all the... proteins (Read more...)

Access to the UniProt Knowledgebase

- **SRS** - Access to UniProtKB/Swiss-Prot, UniProtKB/TrEMBL and other databases using the Sequence Retrieval System
- Full text search in the UniProt Knowledgebase
- Advanced search in the UniProt Knowledgebase by description, gene name and organism (can be used to create html links to UniProt Knowledgebase queries)
- Taxonomy browser (NEWT)
- **BLAST** similarity search

Figure 2-10: The Search window at the top of the Swiss-Prot home page.

Click here for Advanced Search.



Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: dUTPase coli

UniProtKB/Swiss-Prot Release 50.1 of 13-Jun-2006
UniProtKB/TrEMBL Release 33.1 of 13-Jun-2006

- Number of sequences found in UniProt Knowledgebase (Swiss-Prot(3) and TrEMBL)(0): **3**
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more directed searches, you can use the Sequence Retrieval System SRS.

Search in UniProtKB/Swiss-Prot: There are matches to 3 out of 223100 entries

DUT_ECO57 (**P64007**)
Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase). {GENE: Name=dut; OrderedLocusNames=Z5064, ECs4515} - Escherichia coli O157:H7

DUT_ECOL6 (**P64006**)
Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase). {GENE: Name=dut; OrderedLocusNames=c_4464} - Escherichia coli O6

DUT_ECOLI (**P06968**)
Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase). {GENE: Name=dut; Synonyms=dnaS, sof; OrderedLocusNames=b3640} - Escherichia coli

Figure 2-11: Three dUTPase sequences from the Escherichia coli bacterium.

3 relevant dUTPase entries found.

codes are worth writing down; they're used to cross-reference related entries in other databases.

- The top section offers a biochemical description of the protein, including its standard name, its international Enzyme Committee number ("E.C." does not mean "E. coli"), as well as a couple of synonyms. These synonyms can be very useful to broaden your search for relevant articles in Medline. Then comes a list of bibliographic references relevant to the sequencing of the protein or some functional studies. Note that the PubMed ID is given for all cited articles to allow for easy cross-referencing.

- The middle section offers a whole series of links to various functional classification schemes — including relevant protein domains, 3-D structures, and functional signatures. (We describe this section in more detail in Chapter 4.)

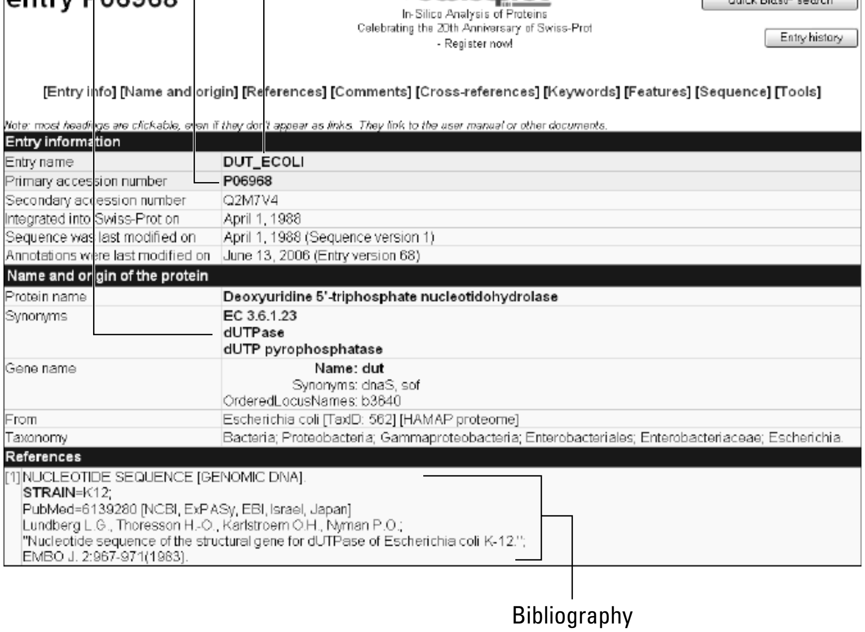
- The bottom section — the *sequence* section — provides you with the actual amino-acid sequence of the protein. In Figure 2-14, you can see that the *E. coli* dUTPase is made up of 151 amino acids, corresponding to a predicted molecular weight of 16,155 Da. (The sequence shown in Figure 2-14 uses the one-letter code format for describing amino acids. For more on the different formats used, see Chapter 1.)

For further studies, you need this amino-acid sequence in the FASTA format — a much simpler format that doesn't bother with punctuation. See the sidebar "The FASTA (and RAW) format," elsewhere in this chapter, for more information.

3. **To get the FASTA format, click the FASTA Format link, on the extreme right of the very bottom of the entry.**

   The FASTA format for your sequence is shown in Figure 2-15.

In just three easy steps, you've gone all the way from an informal definition such as dUTPase to the complete description of the protein that performs this function in everybody's "favorite" bacterium. (And you thought bioinformatics was complicated?)



entry P06968

In-Silico Analysis of Proteins
Celebrating the 20th Anniversary of Swiss-Prot
- Register now!

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**DUT_ECOLI**
Primary accession number	**P06968**
Secondary accession number	Q2M7V4
Integrated into Swiss-Prot on	April 1, 1988
Sequence was last modified on	April 1, 1988 (Sequence version 1)
Annotations were last modified on	June 13, 2006 (Entry version 68)
Name and origin of the protein	
Protein name	**Deoxyuridine 5'-triphosphate nucleotidohydrolase**
Synonyms	**EC 3.6.1.23** **dUTPase** **dUTP pyrophosphatase**
Gene name	**Name: dut** Synonyms: dnaS, sof OrderedLocusNames: b3640
From	Escherichia coli [TaxID: 562] [HAMAP proteome]
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.

References

[1] NUCLEOTIDE SEQUENCE [GENOMIC DNA].
STRAIN=K12;
PubMed=6139280 [NCBI, ExPASy, EBI, Israel, Japan]
Lundberg L.G., Thoresson H.-O., Karlstroem O.H., Nyman P.O.;
"Nucleotide sequence of the structural gene for dUTPase of Escherichia coli K-12.";
EMBO J. 2:967-971(1983).

Figure 2-12: Swiss-Prot entry for the E. coli dUTPase protein (upper section).

More advanced ways to retrieve protein sequences

Identifying the right protein in Swiss-Prot isn't always so easy because the information you start with may not be specific enough. For instance, the Search feature we outline in the previous section may not work particularly well if the same term is found in different sections of the Swiss-Prot entry. To get better results, you have to use some field restriction (which we describe in the context of PubMed searches).

This type of restricted search always depends on the organization of each particular database, as well as on the idiosyncrasies of search software. For Swiss-Prot, the steps are as follows:

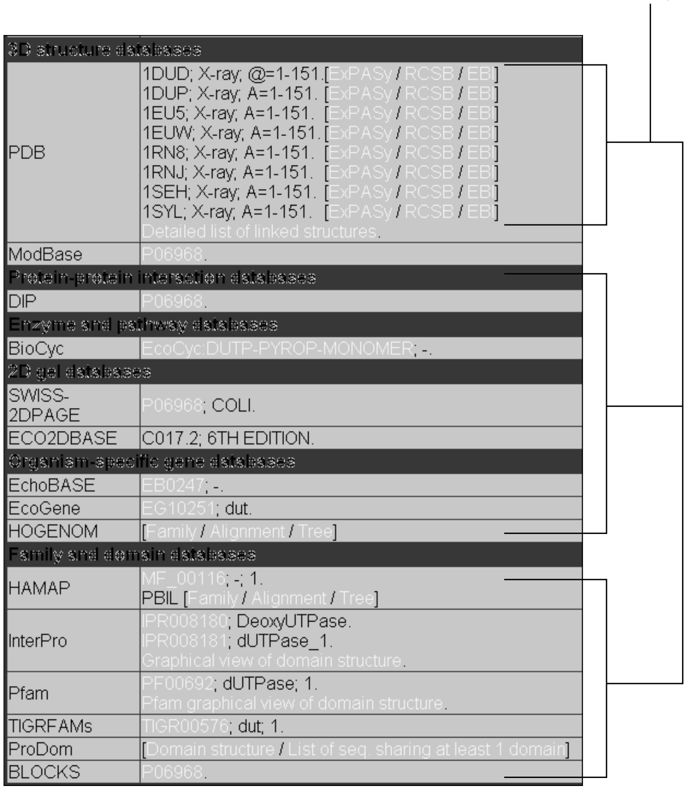
1. **Point your browser to** `www.expasy.org/sprot/`.

   You'll see the by-now-familiar Swiss-Prot database home page (refer to Figure 2-10).

with three specific fields: Description, Gene name, and Organism.

3. **For the purposes of our example, type** dUTPase **in the Description field, choose baker's yeast from the Organism drop-down menu, and click the Submit Query button. (Refer to Figure 2-16.)**

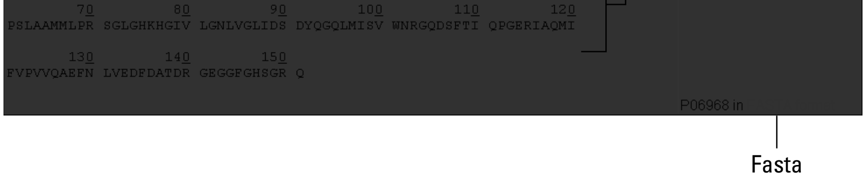
   Your results appear on a new page. If you use the dUTPase example, you'll successfully retrieve the yeast dUTPase protein, no ifs, ands, or buts — that's to say, exactly the protein you were looking for. Its Swiss-Prot entry name is `DUT_YEAST`, and its accession number is `P33317`.



Links to relevant entries in other databases (structure, biochemistry, signatures, . . . , etc.)

3D structure databases	
PDB	1DUD; X-ray; @=1-151.[ExPASy / RCSB / EBI] 1DUP; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1EU5; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1EUW; X-ray; A=1-151.[ExPASy / RCSB / EBI] 1RN8; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1RNJ; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1SEH; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1SYL; X-ray; A=1-151. [ExPASy / RCSB / EBI] Detailed list of linked structures.
ModBase	P06968.
Protein-protein interaction databases	
DIP	P06968.
Enzyme and pathway databases	
BioCyc	EcoCyc:DUTP-PYROP-MONOMER; -.
2D gel databases	
SWISS-2DPAGE	P06968; COLI.
ECO2DBASE	C017.2; 6TH EDITION.
Organism-specific gene databases	
EchoBASE	EB0247; -.
EcoGene	EG10251; dut.
HOGENOM	[Family / Alignment / Tree]
Family and domain databases	
HAMAP	MF_00116; -; 1. PBIL [Family / Alignment / Tree]
InterPro	IPR008180; DeoxyUTPase. IPR008181; dUTPase_1. Graphical view of domain structure.
Pfam	PF00692; dUTPase; 1. Pfam graphical view of domain structure.
TIGRFAMs	TIGR00576; dut; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain]
BLOCKS	P06968.

Figure 2-13: Swiss-Prot entry for the E. coli dUTPase protein (middle section).



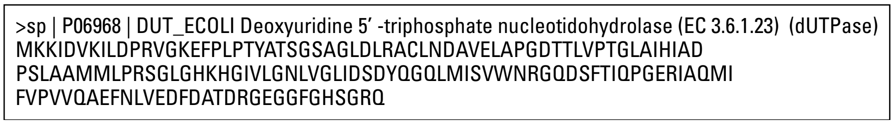
```
       70         80         90        100        110        120
PSLAAMMLPR SGLGHKHGIV LGNLVGLIDS DYQGQLMISV WNRGQDSFTI QPGERIAQMI

       130        140        150
FVPVVQAEFN LVEDFDATDR GEGGFGHSGR Q
```

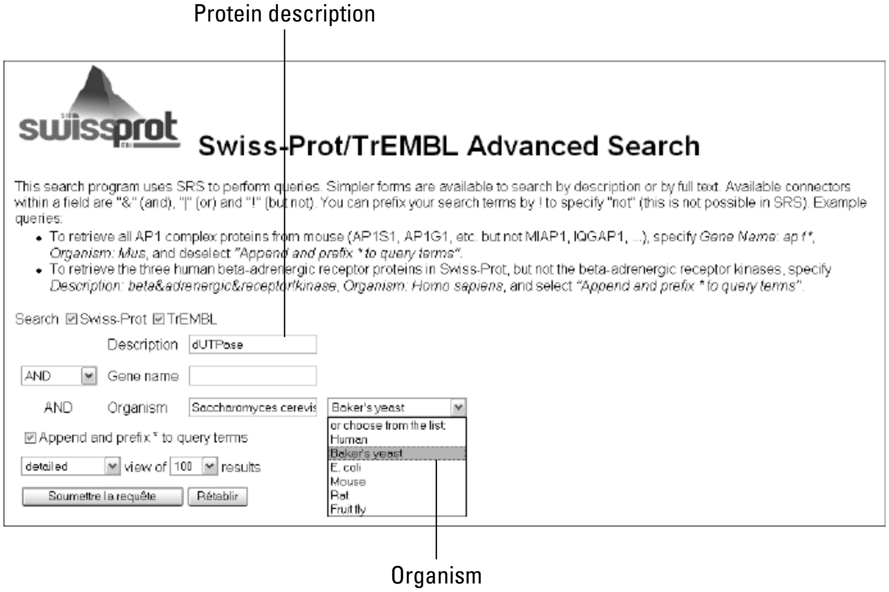
P06968 in

protein (bottom section).

Figure 2-15: E. coli dUTPase protein sequence in FASTA format.



```
>sp | P06968 | DUT_ECOLI Deoxyuridine 5' -triphosphate nucleotidohydrolase (EC 3.6.1.23)  (dUTPase)
MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAPGDTTLVPTGLAIHIAD
PSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWNRGQDSFTIQPGERIAQMI
FVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ
```



Swiss-Prot/TrEMBL Advanced Search

This search program uses SRS to perform queries. Simpler forms are available to search by description or by full text. Available connectors within a field are "&" (and), "|" (or) and "!" (but not). You can prefix your search terms by ! to specify "not" (this is not possible in SRS). Example queries:

- To retrieve all AP1 complex proteins from mouse (AP1S1, AP1G1, etc. but not MIAP1, IQGAP1, ...), specify *Gene Name: ap1**, *Organism: Mus*, and deselect *"Append and prefix * to query terms"*.
- To retrieve the three human beta-adrenergic receptor proteins in Swiss-Prot, but not the beta-adrenergic receptor kinases, specify *Description: beta&adrenergic&receptor!kinase*, *Organism: Homo sapiens*, and select *"Append and prefix * to query terms"*.

Search ☑ Swiss-Prot ☑ TrEMBL

Description: dUTPase

AND Gene name:

AND Organism: Saccharomyces cerevis — Baker's yeast

☑ Append and prefix * to query terms

detailed view of 100 results

Figure 2-16: Using Swiss-Prot Advanced search.

FASTA is the name of a popular sequence-alignment-and-database-scanning program created by W.R. Pearson and D.J. Lipman in 1988 (you can use your brand new PubMed skills to find the original article). The sequences used by FASTA have to obey the following format:

```
>My_Sequence_Name

ARCGTCRGCKINTANDRGCKINTAND

CKINTANDARCGTCRGCKINTANDRG

CKINTAND
```

The line starting with > (the *definition line*) contains a unique identifier followed by an optional short definition. The lines that follow it contain the DNA or protein sequence (in one-letter code) until the next > character in the file indicates the beginning of a new sequence.

Because FASTA is easy to parse, this format has become hugely popular — and is now the default input format for much sequence analysis software, including BLAST and CLUSTALW.

Be aware, though, that programs using FASTA-formatted sequences as input are sometimes case-sensitive. Here are some pointers:

- Always use CAPITAL letters for the one-letter codes.
- When using FASTA-formatted sequences on a PC, always use the TEXT option of your preferred word-processing software (that is, skip the formatting and use nothing but ASCII characters).
- When displaying these sequences as a word-processing document, use the Courier font for easy alignment.

Some programs analyzing one sequence at a time work with the *RAW* format. This is simply the sequence part of the FASTA format, without the definition line — but machines can be finicky. Using the FASTA format when the RAW format is required may cause an error — or some of the definition line may end up included in the protein or DNA sequence (!).

Retrieving a list of related protein sequences

Many questions in molecular biology (your dissertation topic, your own research, or your personal interests) require downloading a large collection of similar protein sequences, all related to the same function, rather than just one sequence. These biological questions typically include the detection of conserved functional motifs (segments of sequences that look the same in proteins with the same function), the simultaneous alignment of multiple sequences, the assessment of their variability, or *phylogenetic* studies — how sequences relate to each other through evolution.

First, go back to the Advanced Search page of the ExPASy server:

1. **Point your browser to** `www.expasy.org/sprot/` **and click the** Advanced Search in the UniProt Knowledgebase **link.**
2. **In the Search line — directly above the Description window — keep the Swiss-Prot box checked but deselect the TrEMBL box.**

   TrEMBL is a database made up of unsupervised computer translations of new DNA sequences, Swiss-Prot only includes entries validated by expert curators. Restricting the search to Swiss-Prot thus ensures — to the best of our knowledge — that all returned proteins are *actual* dUTPases.
3. **Type** dUTPase **in the Description window.**

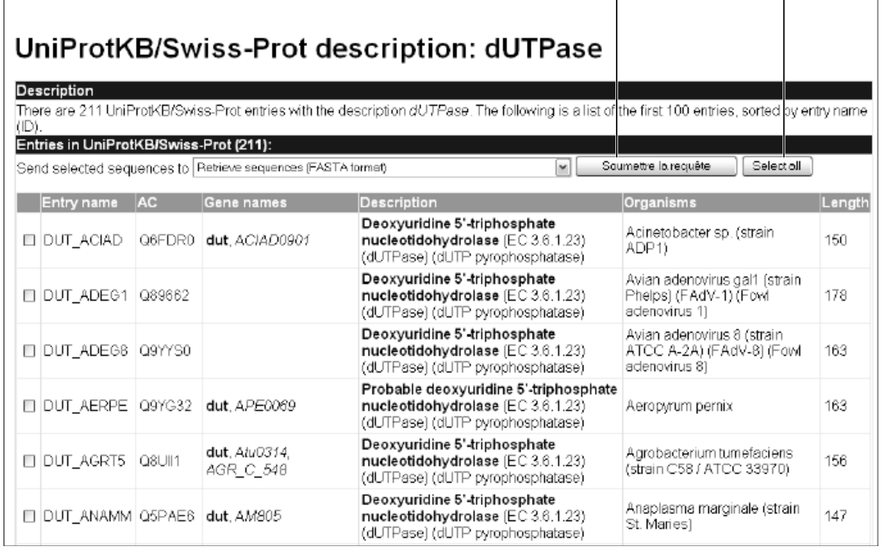
   Be sure that you don't type in anything else. In particular, don't put anything in the Organism window.
4. **Click the Submit Query button.**

   A Results page appears, as shown in Figure 2-17. This particular search (at the time of writing) yielded 211 Swiss-Prot entries. (You may get more entries when you get around to doing your own search.) Scrolling down the list, you can see that all the entries look like fine and upstanding dUTPase protein sequences. At this point, you can click each of the individual names (such as `DUT_ADEG1`) to access a complete ID card. (For a sense of what you can find on such an ID card, refer to Figures 2-12, 2-13 and 2-14.)

But to complete your assignment on dUTPases, for instance, you now have to gather all these sequences into a single file. This is easy; just do the following:

1. **Perform Steps 1 through 4 in the preceding list, and then click the Select All button at the top-right of the Results screen (refer to Figure 2-17) or check the boxes for some specific sequences you are interested in.**
2. **Don't waste any time here and don't change anything; simply click the Send query button (that's *soumettre la requête* in French).**

   The Download page appears, as shown in Figure 2-18. Your sequences are ready to be saved to your hard drive.



UniProtKB/Swiss-Prot description: dUTPase

Description

There are 211 UniProtKB/Swiss-Prot entries with the description *dUTPase*. The following is a list of the first 100 entries, sorted by entry name (ID).

Entries in UniProtKB/Swiss-Prot (211):

Send selected sequences to Retrieve sequences (FASTA format)

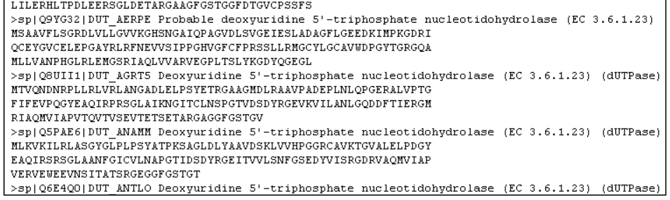
	Entry name	AC	Gene names	Description	Organisms	Length
□	DUT_ACIAD	Q6FDR0	**dut**, *ACIAD0901*	**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Acinetobacter sp. (strain ADP1)	150
□	DUT_ADEG1	Q89662		**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Avian adenovirus gal1 (strain Phelps) (FAdV-1) (Fowl adenovirus 1)	178
□	DUT_ADEG8	Q9YYS0		**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Avian adenovirus 8 (strain ATCC A-2A) (FAdV-8) (Fowl adenovirus 8)	163
□	DUT_AERPE	Q9YG32	**dut**, *APE0069*	**Probable deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Aeropyrum pernix	163
□	DUT_AGRT5	Q8UII1	**dut**, *Atu0314*, *AGR_C_548*	**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Agrobacterium tumefaciens (strain C58 / ATCC 33970)	156
□	DUT_ANAMM	Q5PAE6	**dut**, *AM805*	**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Anaplasma marginale (strain St. Maries)	147

Figure 2-17: Top of the list of all dUTPase proteins found in Swiss-Prot.

You can save the content of this page in various ways, depending on the Internet browser version you're using. We recommend the following procedure because it always works — and it doesn't add mysterious extra characters to the resulting file!

1. **Choose the Select All feature from your browser menu (in Internet Explorer, choose Edit➪Select All) and then choose Copy. (Again, in Internet Explorer you choose Edit➪Copy.)**
2. **Open a new document in your word-processing program (Microsoft Word, for example).**
3. **Choose Paste from the Edit menu of your word-processing program.**
4. **Reformat the whole document with a Courier font (8 or 10 point) to realign the sequences.**
5. **Finally, save your document as** `dUTPaseDB.txt`**, using the Save As type option *text only*** (`*.txt`).

You now have, on your own personal computer, all the well-characterized dUTPase protein sequences known to man. Keep this file on your hard drive; you need it later in this chapter.



```
LILERHLTPDLEERSGLDETARGAAGFGSTGGFDTGVCPSSFS
>sp|Q9YG32|DUT_AERPE Probable deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23)
MSAAVFLSGRDLVLLGVVKGHSNGAIQPAGVDLSVGEIESLADAGFLGEEDKIMPKGDRI
QCEYGVCELEPGAYRLRFNEVVSIPPGHVGFCFPRSSLLRMGCYLGCAVWDPGYTGRGQA
MLLVANPHGLRLEMGSRIAQLVVARVEGPLTSLYKGDYQGEGL
>sp|Q8UII1|DUT_AGRT5 Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
MTVQNDNRPLLRLVRLANGADLELPSYETRGAAGMDLRAAVPADEPLNLQPGERALVPTG
FIFEVPQGYEAQIRPRSGLAIKNGITCLNSPGTVDSDYRGEVKVILANLGQDDFTIERGM
RIAQMVIAPVTQVTVSEVTETSETARGAGGFGSTGV
>sp|Q5PAE6|DUT_ANAMM Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
MLKVKILRLASGYGLPLPSYATPKSAGLDLYAAVDSKLVVHPGGRCAVKTGVALELPDGY
EAQIRSRSGLAANFGICVLNAPGTIDSDYRGEITVVLSNFGSEDYVISRGDRVAQMVIAP
VERVEWEEVNSITATSRGEGGFGSTGT
>sp|Q6E4Q0|DUT_ANTLO Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
```

FASTA sequences ready to be saved to your hard drive.

Retrieving DNA Sequences

Protein sequences are simple objects with a relatively narrow range of sizes (they're about 300 amino acids long, plus or minus 200, except for a few giant ones), clearly defined boundaries, and specific functional attributes. Furthermore, proteins of microbes or higher *eukaryotes* (animal and plants) have roughly the same properties.

As you might expect, the corresponding gene (DNA) sequences get more varied and complex in higher animals. Gene sizes in humans may vary from microbe-like lengths (a few thousand bp) to several hundred thousand bp.

Not all DNA is coding for protein

Various types of DNA sequences are involved in defining a gene:

- Regulatory regions (usually preceding the coding region)
- Untranslated regions that precede and follow the coding regions
- The protein-coding region

In eukaryotes (yeast, plants, animals), the protein-coding region is divided into a variable number of *exons* — gene segments that contribute to the final protein — interspersed with *introns* — gene segments that do not.

As a consequence, working with DNA sequences is always trickier than working with protein sequences. Go to Chapters 3 and 5 to find out more about the architecture of genes.

in downloaded sequence files

In the dUTPAse example, we didn't use the File➪Save As command on the Internet browser main menu to download the content of the window because files saved using the File➪Save As command have some problems:

- They may contain some hidden *parasite characters* — Control/Alt/Shift + something — often displayed as **<PRE>** at the beginning of the file — which corrupts the FASTA format.
- They're trickier to reopen by your PC word-processing software, because they don't have the right extension (for instance `.doc`), or ask you to choose an obscure encoding scheme (about which we know nothing, so we can't advise you) to load the file.

It is our experience that using a browser's File➪Save As option produces unpredictable results, depending on the browser type, version, or implementation. In general, it's a good and wise practice to inspect the sequence data files that you download from the Internet for unexpected leading or trailing signs. For most sequence-analysis programs, FASTA-formatted sequence files must begin with a definition line, such as

```
>P0343456
    My_Sequence_definition
```

and nothing else! Any leading character (even blank ones) that differs from > may produce an error. (For instance, the definition line might be considered part of the protein sequence.) Sequence-data files also have to end with a final `<New Line>` character (showing as a blank line).

The good news is that usually no constraints restrict the length of the definition and sequence lines (but use reasonable numbers, no more than about 60 to 100 characters long, to be safe). Except for the constraint of having > as the first character of any definition line, you can freely use blank characters, such as `<Space>`, `<Tab>`, and line delimiters without interfering with the parsing of the sequences.

Finally, do not use characters that are not among the standard amino-acid codes such as `<->` or `<*>`. They aren't treated in a consistent way by different analysis programs. They're skipped (deleting a position), replaced by `X`, or may simply cause an error. (For more on standard amino-acid codes, see Chapter 1.)

Going from protein sequences to DNA sequences

In databases, the correspondence between protein and DNA sequences is *not* one-to-one. Many different — even non-overlapping — DNA sequences can be linked to the same protein or gene name, as the following list makes clear (flip ahead to Figures 3-1 and 3-2):

- Numerous types of partial sequences

Life is a lot simpler if you can stick to working with just protein sequences, but we realize that situations will arise that will require you to go back to the DNA sequence. The most common situation of this kind is when you want to amplify a gene to transfer it to another organism — using a technique called Polymerase Chain Reaction or PCR — and then either synthesize the corresponding protein or induce specific mutations. Your problem, succinctly stated, is this: *Given a protein sequence, how can I retrieve the DNA sequence encompassing its coding region?*

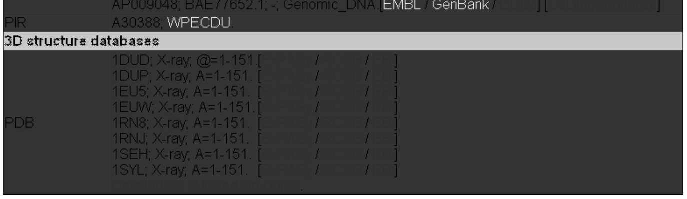
Retrieving the DNA sequence relevant to my protein

Imagine that you want to retrieve enough DNA sequence to clone the dUTPase gene of *E. coli.* Sounds tricky, right? We show you how it's done in the following steps:

1. **Point your browser to** `www.expasy.org/sprot/`**.**
2. **To access the *E. coli* dUTPase entry quickly, simply enter the accession number (**P06968**) in the Search window at the top of the page and then click the Search button.**

   Again, we have the information page devoted to protein P06968 at hand (refer to Figure 2-12.).
3. **Click the** Cross-References **link near the top of the form.**

   Your browser jumps to the relevant section, as shown in Figure 2-19. The Cross-References section consists of successive groups of lines introduced by keywords such as `EMBL`, `PIR`, `PDB`, and so on. All these keywords correspond to databases of various kinds that can give you additional information about your protein sequence. By clicking the different underlined words, you jump right to the corresponding entry in those different databases. This web of links between different sites that concern the same object is, in fact, a web of cross-references. Take the EMBL group of lines, for instance: There are four of them, starting with obscure numbers and followed by a set of four underlined links to EMBL, GenBank, DDBJ (those are databases), and CoDingSequence (the precise DNA segment coding for the protein). Let's see what happens when we use them.

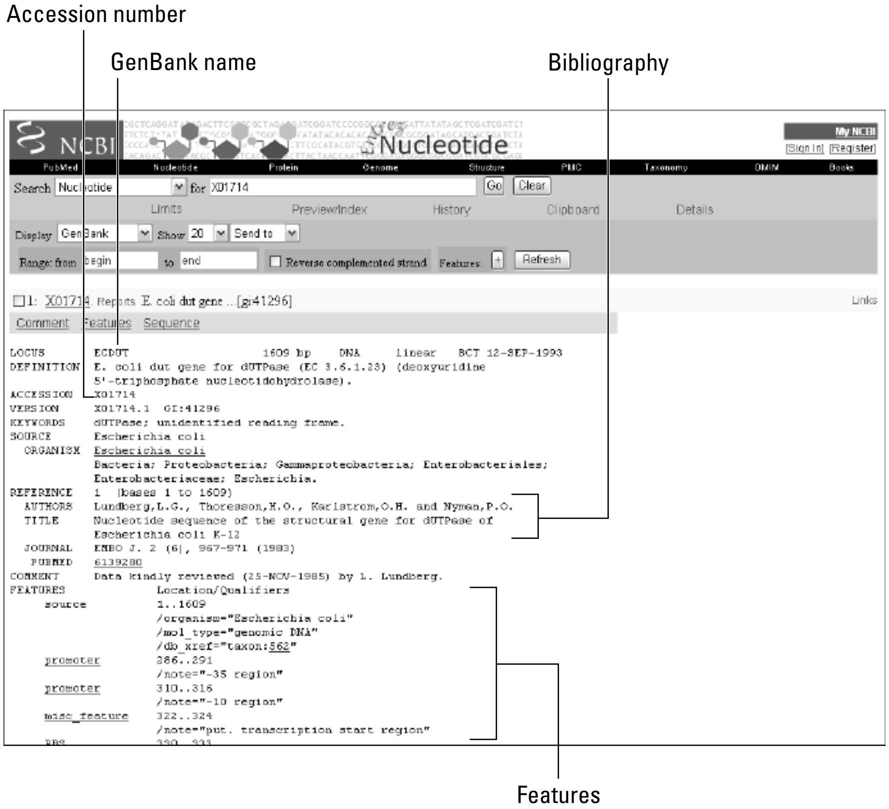


```
              AP009048; BAE77652.1; -; Genomic_DNA.[EMBL / GenBank /      ] [                ]
PIR           A30388; WPECDU.
3D structure databases
              1DUD; X-ray; @=1-151.[EXPASy / RCSB / EBI]
              1DUP; X-ray; A=1-151. [EXPASy / RCSB / EBI]
              1EU5; X-ray; A=1-151.  [EXPASy / RCSB / EBI]
              1EUW; X-ray; A=1-151. [EXPASy / RCSB / EBI]
PDB           1RN8; X-ray; A=1-151.  [EXPASy / RCSB / EBI]
              1RNJ; X-ray; A=1-151.  [EXPASy / RCSB / EBI]
              1SEH; X-ray; A=1-151.  [EXPASy / RCSB / EBI]
              1SYL; X-ray; A=1-151.  [EXPASy / RCSB / EBI]
              Detailed list of linked structures.
```

Cross-References section of the P06968 information page.

4. **Click the GenBank link in the EMBL row's first line.**

   The GenBank record (Accession # X01714) corresponding to protein P06968 appears, as shown in Figure 2-20. It is several pages long.



□1: X01714. Reports E. coli dut gene ...[gi41296]

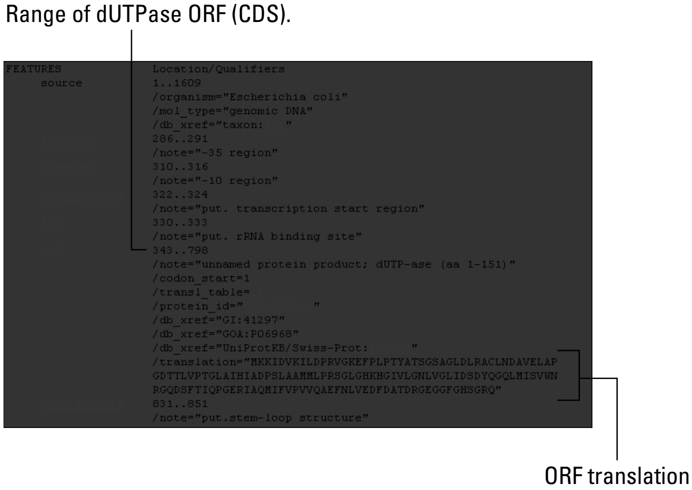
```
LOCUS       ECDUT                   1609 bp    DNA     linear   BCT 12-SEP-1993
DEFINITION  E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
            5'-triphosphate nucleotidohydrolase).
ACCESSION   X01714
VERSION     X01714.1  GI:41296
KEYWORDS    dUTPase; unidentified reading frame.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1609)
  AUTHORS   Lundberg,L.G., Thoresson,H.O., Karlstrom,O.H. and Nyman,P.O.
  TITLE     Nucleotide sequence of the structural gene for dUTPase of
            Escherichia coli K-12
  JOURNAL   EMBO J. 2 (6), 967-971 (1983)
   PUBMED   6139280
COMMENT     Data kindly reviewed (25-NOV-1985) by L. Lundberg.
FEATURES             Location/Qualifiers
     source          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"
     RBS             330..333
```

Figure 2-20: Top of GenBank entry ECDUT/ X01714.

mend that you write it down because it's used to cross-reference related entries in other databases (which is what we did with Swiss-Prot). A few lines below, the `SOURCE` and `ORGANISM` fields describe the biological origin of the DNA sequence.

- The *Reference section* lists article(s) relevant to the sequence determination. This list can be quite long for large sequences.
- The *Features section* lists the definitions and exact ranges of multiple types of elements that have been recognized in the sequence. Our example — the X01714 entry — includes promoter elements, ribosome binding sites (RBS), and protein coding segments (CDS). Figure 2-21 shows — right next to the keyword CDS — the limits of the dUTPase ORF as `343..798`. The CDS range is followed by the name and the amino-acid translation of the corresponding protein.
- The *Sequence section* rounds out the GenBank entry, where the nucleotides are listed between the Origin keyword and the final `//` that signals the very end of the entry. Numbering is provided to help relate the location of the dUTPase ORF (343-798) to the actual nucleotide sequence. (See Figure 2-22.)

When you've read through the entry, you can save the whole thing to your hard drive by using the steps we defined in the preceding list for saving Swiss-Prot sequences.

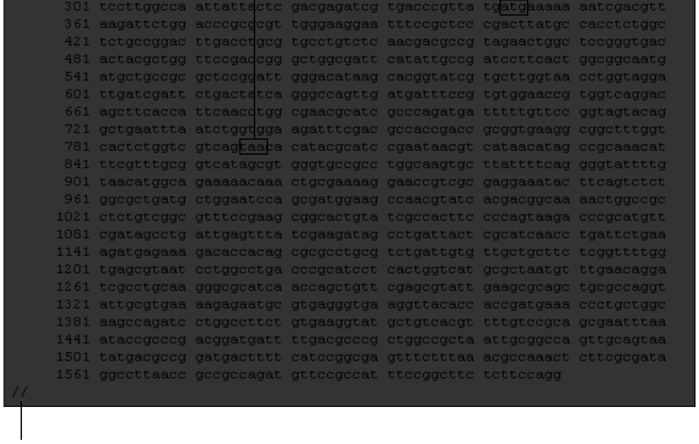


Range of dUTPase ORF (CDS).

```
FEATURES             Location/Qualifiers
     source          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"
     RBS             330..333
                     /note="put. rRNA binding site"
     CDS             343..798
                     /note="unnamed protein product; dUTP-ase (aa 1-151)"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="CAA25859.1"
                     /db_xref="GI:41297"
                     /db_xref="GOA:P06968"
                     /db_xref="UniProtKB/Swiss-Prot:P06968"
                     /translation="MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAP
                     GDTTLVPTGLAIHIADPSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWN
                     RGQDSFTIQPGERIAQMIFVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ"
     misc_feature    831..851
                     /note="put.stem-loop structure"
```

ORF translation

Figure 2-21: The Features section of GenBank entry ECDUT/X01714.



```
 301 tccttggcca attattactc gacgagatcg tgacccgtta tgatgaaaaa aatcgacgtt
 361 aagattctgg acccgcgcgt tgggaaggaa tttccgctcc cgacttatgc cacctctggc
 421 tctgccggac ttgacctgcg tgcctgtctc aacgacgccg tagaactggc tccgggtgac
 481 actacgctgg ttccgadcgg gctggcgatt catattgccg atccttcact ggcggcaatg
 541 atgctgccgc gctccggatt gggacataag cacggtatcg tgcttggtaa cctggtagga
 601 ttgatcgatt ctgactatca gggccagttg atgatttccg tgtggaaccg tggtcaggac
 661 agcttcacca ttcaacctgg cgaacgcatc gcccagatga tttttgttcc ggtagtacag
 721 gctgaattta atctggtgga agatttcgac gccaccgacc gcggtgaagg cggctttggt
 781 cactctggtc gtcagtaaca catacgcatc cgaataacgt cataacatag ccgcaaacat
 841 ttcgtttgcg gtcatagcgt gggtgccgcc tggcaagtgc ttattttcag ggtattttg
 901 taacatggca gaaaaacaaa ctgcgaaaag gaaccgtcgc gaggaaatac ttcagtctct
 961 ggcgctgatg ctggaatcca gcgatggaag ccaacgtatc acgacggcaa aactggccgc
1021 ctctgtcggc gtttccgaag cggcactgta tcgccacttc cccagtaaga cccgcatgtt
1081 cgatagcctg attgagttta tcgaagatag cctgattact cgcatcaacc tgattctgaa
1141 agatgagaaa gacaccacag cgcgcctgcg tctgattgtg ttgctgcttc tcggttttgg
1201 tgagcgtaat cctggcctga cccgcatcct cactggtcat gcgctaatgt ttgaacagga
1261 tcgcctgcaa gggcgcatca accagctgtt cgagcgtatt gaagcgcagc tgcgccaggt
1321 attgcgtgaa aagagaatgc gtgagggtga aggttacacc accgatgaaa ccctgctggc
1381 aagccagatc ctggccttct gtgaaggtat gctgtcacgt tttgtccgca gcgaatttaa
1441 ataccgcccg acggatgatt ttgacgcccg ctggccgcta attgcggcca gttgcagtaa
1501 tatgacgccg gatgactttt catccggcga gtttctttaa acgccaaact cttcgcgata
1561 ggccttaacc gccgccagat gttccgccat ttccggcttc tcttccagg
//
```

End of the X01714 entry

Figure 2-22: The Sequence section of GenBank entry ECDUT/X01714.

However, to use the nucleotide sequence as input for other programs (for instance, to design primers), you may need to isolate it from the text part of the entry and convert it in FASTA format. Here's how that's done:

1. **Scroll back to the top of the page for the ECDUT/X01714 entry.**

   Refer to Figure 2-20 for what your screen should look like.

2. **Choose FASTA from the Display drop-down menu, as shown in Figure 2-23.**

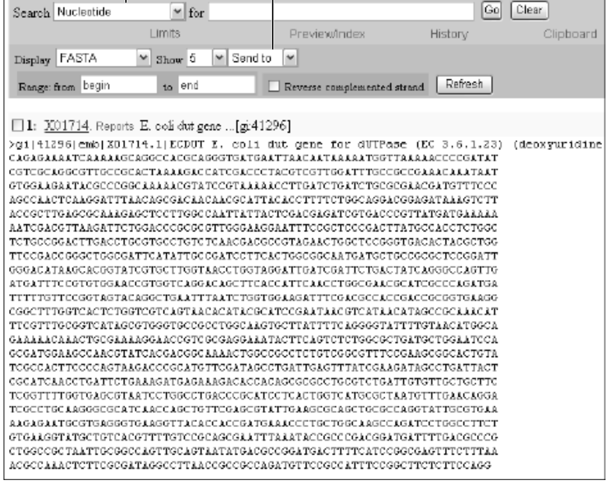
3. **Transform the content of this window into plain text by choosing Text from the drop-down menu located on the far right of the menu bar.**

4. **Save the FASTA sequence by using the following protocol:**

   a. In the Edit menu of your Web browser, click Select All and then click Copy.

   b. Open a default Word document and, in the Edit menu of Word, click Paste. Then select a Courier font (8 or 10).

   c. Finally, save your document as dUTPaseDNA.txt by choosing the Save as type option text only (`*.txt`).



```
>gi|41296|emb|X01714.1|ECDUT E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
CAGAGAAAATCAAAAAGCAGGCCACGCAGGGTGATGAATTAACAATAAAAATGGTTAAAAACCCCGATAT
CGTCGCAGGCGTTGCCGCACTAAAAGACCATCGACCCTACGTCGTTGGATTTGCCGCCGAAACAAATAAT
GTGGAAGAATACGCCCGGCAAAAACGTATCCGTAAAAACCTTGATCTGATCTGCGCGAACGATGTTTCCC
AGCCAACTCAAGGATTTAACAGCGACAACAACGCATTACACCTTTTCTGGCAGGACGGAGATAAAGTCTT
ACCGCTTGAGCGCAAAGAGCTCCTTGGCCAATTATTACTCGACGAGATCGTGACCCGTTATGATGAAAAA
AATCGACGTTAAGATTCTGGACCCGCGCGTTGGGAAGGAATTTCCGCTCCCGACTTATGCCACCTCTGGC
TCTGCCGGACTTGACCTGCGTGCCTGTCTCAACGACGCCGTAGAACTGGCTCCGGGTGACACTACGCTGG
TTCCGACCGGGCTGGCGATTCATATTGCCGATCCTTCACTGGCGGCAATGATGCTGCCGCGCTCCGGATT
GGGACATAAGCACGGTATCGTGCTTGGTAACCTGGTAGGATTGATCGATTCTGACTATCAGGGCCAGTTG
ATGATTTCCGTGTGGAACCGTGGTCAGGACAGCTTCACCATTCAACCTGGCGAACGCATCGCCCAGATGA
TTTTTGTTCCGGTAGTACAGGCTGAATTTAATCTGGTGGAAGATTTCGACGCCACCGACCGCGGTGAAGG
CGGCTTTGGTCACTCTGGTCGTCAGTAACACATACGCATCCGAATAACGTCATAACATAGCCGCAAACAT
TTCGTTTGCGGTCATAGCGTGGGTGCCGCCTGGCAAGTGCTTATTTTCAGGGGTATTTTGTAACATGGCA
GAAAAACAAACTGCGAAAAGGAACCGTCGCGAGGAAATACTTCAGTCTCTGGCGCTGATGCTGGAATCCA
GCGATGGAAGCCAACGTATCACGACGGCAAAACTGGCCGCCTCTGTCGGCGTTTCCGAAGCGGCACTGTA
TCGCCACTTCCCCAGTAAGACCCGCATGTTCGATAGCCTGATTGAGTTTATCGAAGATAGCCTGATTACT
CGCATCAACCTGATTCTGAAAGATGAGAAAGACACCACAGCGCGCCTGCGTCTGATTGTGTTGCTGCTTC
TCGGTTTTGGTGAGCGTAATCCTGGCCTGACCCGCATCCTCACTGGTCATGCGCTAATGTTTGAACAGGA
TCGCCTGCAAGGGCGCATCAACCAGCTGTTCGAGCGTATTGAAGCGCAGCTGCGCCAGGTATTGCGTGAA
AAGAGAATGCGTGAGGGTGAAGGTTACACCACCGATGAAACCCTGCTGGCAAGCCAGATCCTGGCCTTCT
GTGAAGGTATGCTGTCACGTTTTGTCCGCAGCGAATTTAAATACCGCCCGACGGATGATTTTGACGCCCG
CTGGCCGCTAATTGCGGCCAGTTGCAGTAATATGACGCCGGATGACTTTTCATCCGGCGAGTTTCTTTAA
ACGCCAAACTCTTCGCGATAGGCCTTAACCGCCGCCAGATGTTCCGCCATTTCCGGCTTCTCTTCCAGG
```

Figure 2-23: Nucleotide sequence of GenBank entry ECDUT/X01714 in FASTA format.

Using BLAST to Compare My Protein Sequence to Other Protein Sequences

After you know the basics of retrieving a protein sequence from a database (see the earlier, aptly named "Retrieving Protein Sequences" section), you're ready to perform your first analysis with it. For most people, the next step is to perform a BLAST search.

BLAST (short for *B*asic *L*ocal *A*lignment *S*earch *T*ool — with a name like that, no wonder they shortened it to BLAST) is a great sequence-comparison tool that quickly tells you which of the other known proteins out there has a sequence similar to yours. You can then use this information for a variety of purposes — including the prediction of protein function, 3-D structure and domain organization, or the identification of homologues (similar proteins) in other organisms. (In Chapter 7, we show you how to use BLAST in great detail. At this point, we only want to get you started and show you how easy it is to use.)

1. **Point your favorite Internet browser to**

   `www.ncbi.nlm.nih.gov/BLAST/`

   The BLAST home page — probably the most frequented bioinformatic Web page in the world — appears, as shown in Figure 2-24. Because this is your first time here, keeping things simple is best.

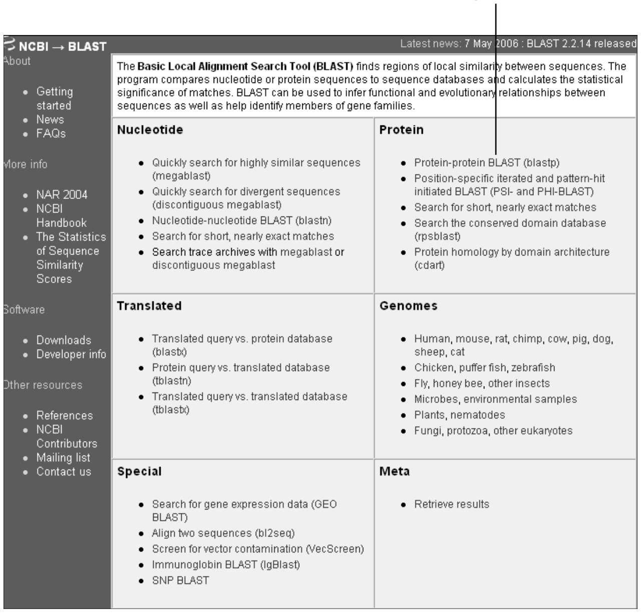
2. **Click the** Protein-Protein BLAST (blastp) **link in the top right.**

   A Query screen appears, as shown in Figure 2-25. At this point, you need a FASTA-formatted protein sequence.

3. **Open the file that contains your dUTPase FASTA-formatted protein sequence.**

   This is the file that you (hopefully) created on your PC by using the steps shown earlier in the "Retrieving a list of related protein sequences" section of this chapter.

Standard protein BLAST



NCBI → BLAST

Latest news: 7 May 2006 : BLAST 2.2.14 released

About

- Getting started
- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

- References
- NCBI Contributors
- Mailing list
- Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Fly, honey bee, other insects
- Microbes, environmental samples
- Plants, nematodes
- Fungi, protozoa, other eukaryotes

Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- SNP BLAST

Meta

- Retrieve results

Figure 2-24: NCBI BLAST entry page.

a. Point your browser to `www.expasy.org/sprot/`.

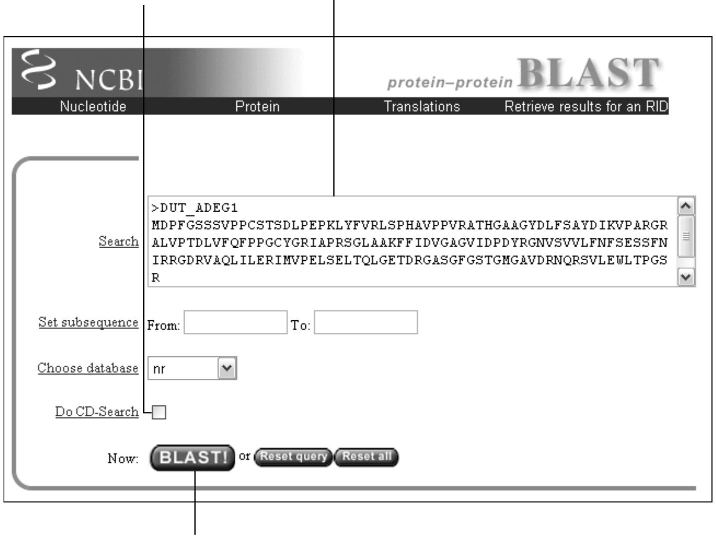
b. Enter **DUT_ADEG1** in the Search window at the top, and then click Search.

d. Copy the protein sequence with its header line.

e. Close that navigator window, and paste the sequence in the BLAST Search window.

Paste your query sequence in here.

Uncheck the CD search box.



Search

```
>DUT_ADEG1
MDPFGSSSVPPCSTSDLPEPKLYFVRLSPHAVPPVRATHGAAGYDLFSAYDIKVPARGR
ALVPTDLVFQFPPGCYGRIAPRSGLAAKFFIDVGAGVIDPDYRGNVSVVLFNFSESSFN
IRRGDRVAQLILERIMVPELSELTQLGETDRGASGFGSTGMGAVDRNQRSVLEWLTPGS
R
```

Set subsequence From: To:

Choose database nr

Do CD-Search □

Now: BLAST! or Reset query Reset all

Click here to run the search.

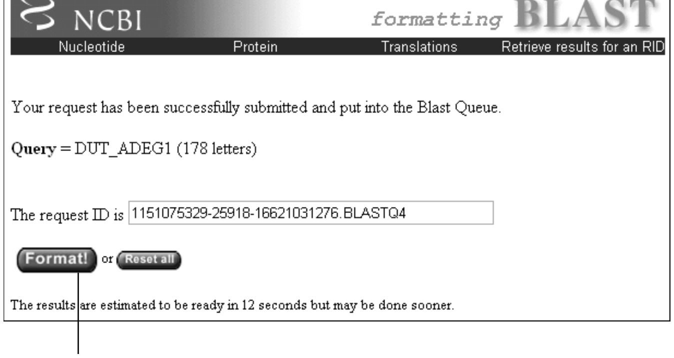
Figure 2-25: NCBI BLAST query form.

5. **Deselect the Do CD-Search box, but don't change the Choose Database setting.**

   The ***nr*** (for *nonredundant*) default database includes all protein sequences known to man — so you don't have to bother about selecting a specific organism or subset.

6. **Click the BLAST! button.**

   You just launched your first BLAST search. Welcome to the club!



Your request has been successfully submitted and put into the Blast Queue.

Query = DUT_ADEG1 (178 letters)

The request ID is 1151075329-25918-16621031276.BLASTQ4

The results are estimated to be ready in 12 seconds but may be done sooner.

Click here once and wait.

Figure 2-26: NCBI BLAST "format-while-you-wait" form.

7. **Click the Format! button.**

   BLAST now gives you an estimated running time for your query. If the time of the day is right, this shouldn't take more than a few seconds.

8. **When the Results page appears, scroll down the page until you reach a long list of sequences (Figure 2-27).**

   What you have here are all the sequences that have a significant similarity with your query sequence, ranked by decreasing score values. If you used `DUT_ADEG1` or another protein sequence taken from a database, the best matching protein is probably the one you started with. (This is what happened in Figure 2-27, with the *Fowl adenovirus* dUTPase.) Here the highest similarity score is 361. The score value depends on the length of the most-similar segments that occur between two sequences — as well as on the quality of the match. It is not normalized to 100 percent. (See Chapter 7 to find out more about the significance of BLAST scores.)

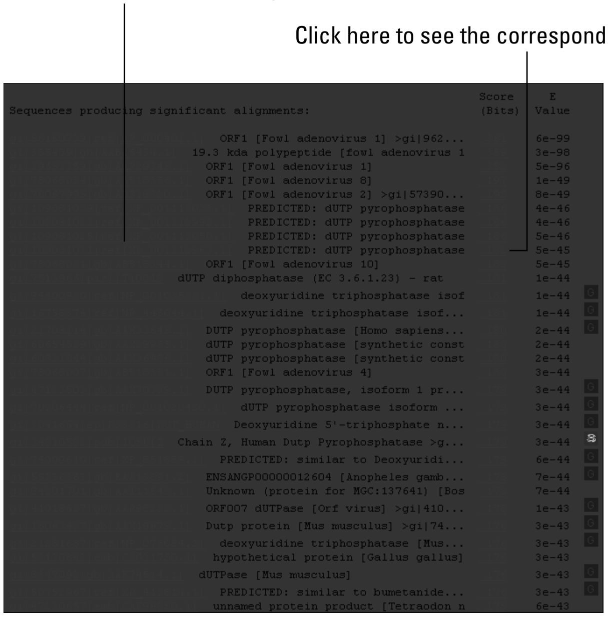
   In this list of sequences, you would simply click an underlined identifier to link to the corresponding *nr* (nonredundant) database entry.

9. **Click one of the underlined scores in the second column from the right.**

   This brings you further down in the page (Figure 2-28), to see the alignment between your query sequence and the matching nr sequence of the protein that corresponds to this score.

In our example, the best match is (not surprisingly) an identical one. By going down the list, you can see less-than-perfect matches, slowly degrading as the corresponding score decreases and the E-value increases. The *E-value* is an

Click here to see the corresponding alignment.



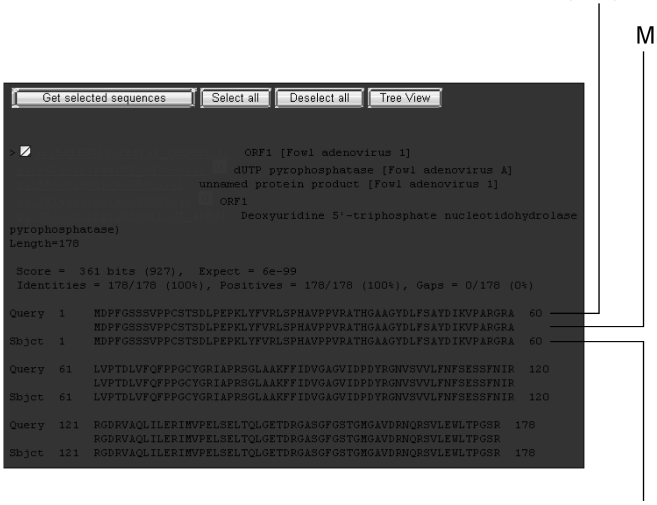
```
                                                            Score    E
Sequences producing significant alignments:                 (Bits)  Value

                        ORF1 [Fowl adenovirus 1] >gi|962...          6e-99
                    19.3 kda polypeptide [fowl adenovirus 1          3e-98
                      ORF1 [Fowl adenovirus 1]                       5e-96
                      ORF1 [Fowl adenovirus 8]                       1e-49
                      ORF1 [Fowl adenovirus 2] >gi|57390...          8e-49
                            PREDICTED: dUTP pyrophosphatase          4e-46
                            PREDICTED: dUTP pyrophosphatase          4e-46
                            PREDICTED: dUTP pyrophosphatase          5e-46
                            PREDICTED: dUTP pyrophosphatase          5e-45
                      ORF1 [Fowl adenovirus 10]                      5e-45
                  dUTP diphosphatase (EC 3.6.1.23) - rat             1e-44
                           deoxyuridine triphosphatase isof          1e-44  G
                        deoxyuridine triphosphatase isof...          1e-44  G
                      DUTP pyrophosphatase [Homo sapiens...          2e-44  G
                      dUTP pyrophosphatase [synthetic const          2e-44
                      dUTP pyrophosphatase [synthetic const          2e-44
                      ORF1 [Fowl adenovirus 4]                       3e-44
                      DUTP pyrophosphatase, isoform 1 pr...          3e-44  G
                           dUTP pyrophosphatase isoform ...          3e-44  G
                          Deoxyuridine 5'-triphosphate n...          3e-44  G
                  Chain Z, Human Dutp Pyrophosphatase >g...          3e-44
                        PREDICTED: similar to Deoxyuridi...          6e-44  G
                      ENSANGP00000012604 [Anopheles gamb...          7e-44  G
                      Unknown (protein for MGC:137641) [Bos          7e-44
                      ORF007 dUTPase [Orf virus] >gi|410...          1e-43  G
                      Dutp protein [Mus musculus] >gi|74...          3e-43  G
                        deoxyuridine triphosphatase [Mus...          3e-43  G
                       hypothetical protein [Gallus gallus]          3e-43
                     dUTPase [Mus musculus]                          3e-43  G
                        PREDICTED: similar to bumetanide...          3e-43  G
                       unnamed protein product [Tetraodon n          6e-43
```

Figure 2-27: NCBI BLAST output form: best match listing.

Query sequence

Matching sequence



```
> ☐                ORF1 [Fowl adenovirus 1]
                     dUTP pyrophosphatase [Fowl adenovirus A]
                 unnamed protein product [Fowl adenovirus 1]
                   ORF1
                      Deoxyuridine 5'-triphosphate nucleotidohydrolase
pyrophosphatase)
Length=178

 Score =  361 bits (927),  Expect = 6e-99
 Identities = 178/178 (100%), Positives = 178/178 (100%), Gaps = 0/178 (0%)

Query  1    MDPFGSSSVPPCSTSDLPEPKLYFVRLSPHAVPPVRATHGAAGYDLFSAYDIKVPARGRA  60
            MDPFGSSSVPPCSTSDLPEPKLYFVRLSPHAVPPVRATHGAAGYDLFSAYDIKVPARGRA
Sbjct  1    MDPFGSSSVPPCSTSDLPEPKLYFVRLSPHAVPPVRATHGAAGYDLFSAYDIKVPARGRA  60

Query  61   LVPTDLVFQFPPGCYGRIAPRSGLAAKFFIDVGAGVIDPDYRGNVSVVLFNFSESSFNIR  120
            LVPTDLVFQFPPGCYGRIAPRSGLAAKFFIDVGAGVIDPDYRGNVSVVLFNFSESSFNIR
Sbjct  61   LVPTDLVFQFPPGCYGRIAPRSGLAAKFFIDVGAGVIDPDYRGNVSVVLFNFSESSFNIR  120

Query  121  RGDRVAQLILERIMVPELSELTQLGETDRGASGFGSTGMGAVDRNQRSVLEWLTPGSR  178
            RGDRVAQLILERIMVPELSELTQLGETDRGASGFGSTGMGAVDRNQRSVLEWLTPGSR
Sbjct  121  RGDRVAQLILERIMVPELSELTQLGETDRGASGFGSTGMGAVDRNQRSVLEWLTPGSR  178
```

Database sequence

Figure 2-28: NCBI BLAST output form: local sequence alignments.

Besides running database searches to identify similar proteins pair-by-pair, the second most common bioinformatic task that biologists like to perform with protein sequences is a multiple alignment. *Multiple alignments* consist in lining up many similar proteins side by side for the sake of comparison. Multiple alignments are used to

- Identify sequence positions where specific amino acids really matter for the structural integrity or the function of a given protein
- Define specific sequence signatures for protein families
- Classify sequences and build evolutionary trees

We detail all the intricacies of making good multiple alignments in Chapter 9. At this point, however, we simply want to show you that performing a multiple alignment is really easy — especially when you have the pleasure of using some nice Internet server, such as the one maintained by the Protein Information Resource (PIR) people at Washington, D.C.'s Georgetown University.

The PIR actually originated from the *Atlas of Protein Sequences,* the first protein-sequence collection (which was built by the late Prof. M. Dayhoff in the late 1970s). The PIR site offers some useful protein analysis tools and databases that we invite you to explore by yourself. Among these tools, it offers a multiple-alignment server (running the standard ClustalW program) that is really easy to use for beginners.

Do the following to get your feet wet using ClustalW:

1. **Point your browser to** `pir.georgetown.edu`**.**

   The PIR home page appears, as shown in Figure 2-29.

2. **Under the Search/Analysis heading, choose Multiple Alignment from the drop-down menu to display the input form.**

   The input form appears. At this point, you need a few FASTA-formatted protein sequences.

3. **Open the dUTPase FASTA-formatted sequence file that you created on your PC in the previous section of this chapter.**

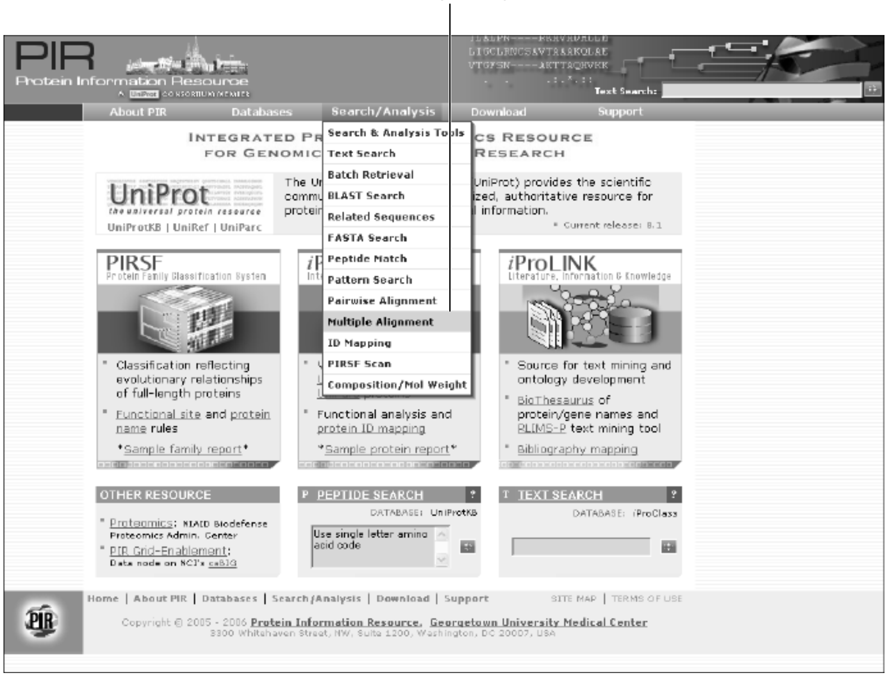
chapter.

4. **Copy/paste five of these sequences (all at once) into the input window, as shown in Figure 2-30.**

   Be careful to copy the definition line and the *entire* amino-acid sequence for each of them.

5. **Click the Submit button.**

   Within a few seconds, your screen (Figure 2-31) is proudly displaying its first successful multiple-sequence alignment — and a tree-like representation of the pair-wise similarity of these sequences.



The Ur commu protein UniProt) provides the scientific ized, authoritative resource for l information.

- Current release: 8.1

- Classification reflecting evolutionary relationships of full-length proteins
- Functional site and protein name rules

Sample family report

- Functional analysis and protein ID mapping

Sample protein report

- Source for text mining and ontology development
- BioThesaurus of protein/gene names and PLIMS-P text mining tool
- Bibliography mapping

- Proteomics: NIAID Biodefense Proteomics Admin. Center
- PIR Grid-Enablement: Data node on NCI's caBIG

Use single letter amino acid code

Copyright © 2005 - 2006 Protein Information Resource, Georgetown University Medical Center
3300 Whitehaven Street, NW, Suite 1200, Washington, DC 20007, USA

Select Multiple Alignment.

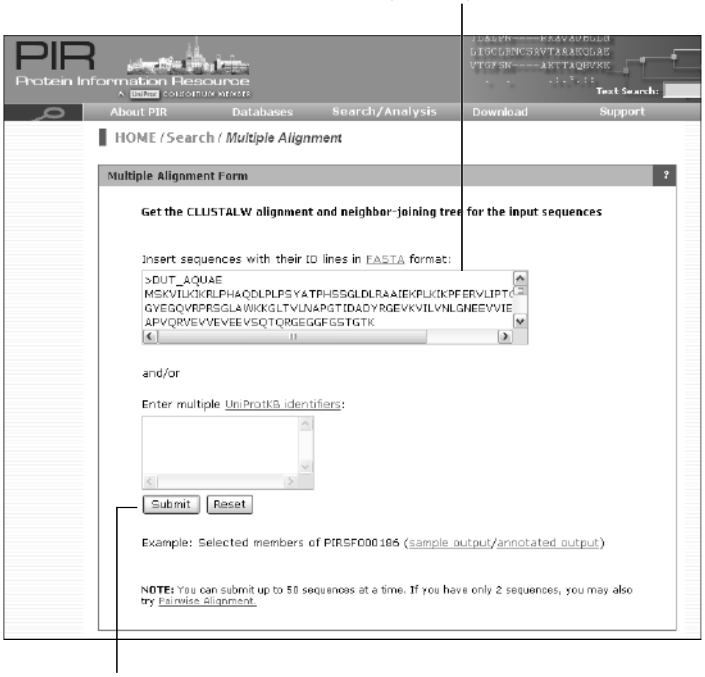
Figure 2-29: Another useful protein analysis center: PIR.

Positions 100 percent identical (highlighted by <*>) or occupied by chemically similar amino-acids (highlighted by <:> or <.>) tend to occur in a clustered fashion along the sequence, defining regions that are probably directly involved in shaping the catalytic site or performing the pyrophosphatase biochemical reaction.

- Various types of *functional signatures* (small sequence segments associated to a given biochemical function) can be extracted from such a multiple alignment.
- Evolutionary relationships between proteins are inferred from phylogenetic trees.

In Chapters 9, 11, and 13, we tell you more about these key applications of bioinformatics.

Paste your sequences here.



HOME / Search / *Multiple Alignment*

Multiple Alignment Form

Get the CLUSTALW alignment and neighbor-joining tree for the input sequences

Insert sequences with their ID lines in FASTA format:

```
>DUT_AQUAE
MSKVILKIKRLPHAQDLPLPSYATPHSSGLDLRAAIEKPLKIKPFERVLIPT
GYEGQVRPRSGLAWKKGLTVLNAPGTIDADYRGEVKVILVNLGNEEVVIE
APVQRVEVVEVEEVSQTQRGEGGFGSTGTK
```

and/or

Enter multiple UniProtKB identifiers:

Submit Reset

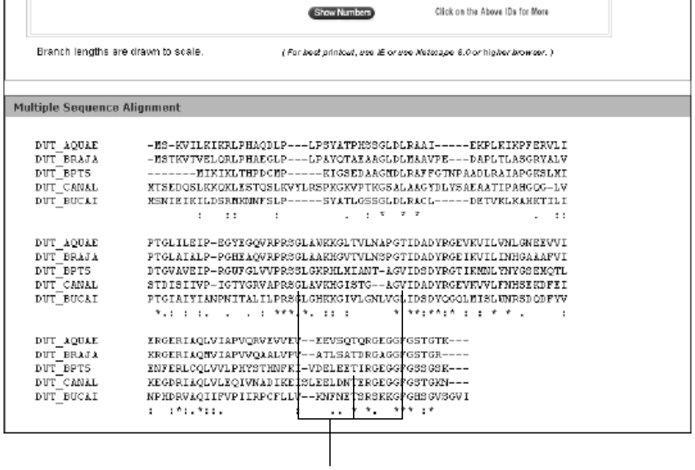
Example: Selected members of PIRSF000186 (sample output/annotated output)

NOTE: You can submit up to 50 sequences at a time. If you have only 2 sequences, you may also try Pairwise Alignment.

Click here to start the multiple alignment.

Figure 2-30: Filling up the multiple alignment input window at PIR.

Figure 2-31: Multiple alignment of five bacterial dUTPase sequences and the corresponding tree built from their pair-wise similarity.



Branch lengths are drawn to scale.

(For best printout, use IE or use Netscape 6.0 or higher browser.)

Multiple Sequence Alignment

```
DUT_AQUAE       -MS-KVILKIKRLPHAQDLP---LPSYATPHSSGLDLRAAI-----EKPLKIKPFERVLI
DUT_BRAJA       -MSTKVTVELQRLPHAEGLP---LPAYQTAEAAGLDLMAAVPE---DAPLTLASGRYALV
DUT_BPT5        -------MIKIKLTHPDCMP-----KIGSEDAAGMDLRAFFGTNPAADLRAIAPGKSLMI
DUT_CANAL       MTSEDQSLKKQKLESTQSLKVYLRSPKGKVPTKGSALAAGYDLYSAEAATIPAHGQG-LV
DUT_BUCAI       MSNIEIKILDSRMKDNNFSLP-----SVATLGSSGLDLRACL-----DETVKLKAHKTLI
                       :   ::     :             .  : *  * *                .  ::

DUT_AQUAE       PTGLILEIP-EGYEGQVRPRSGLANKKGLTVLNAPGTIDADYRGEVKVILVNLGNEEVVI
DUT_BRAJA       PTGLAIALP-PGHEAQVRPRSGLAAKHGVTVLNSPGTIDADYRGEIKVILINHGAAAFVI
DUT_BPT5        DTGVAVEIP-RGVFGLVVPRSSLGKRHLMIANT-AGVIDSDYRGTIKMNLYNYGSEMQTL
DUT_CANAL       STDISIIVP-IGTYGRVAPRSGLAVKHGISTG--AGVIDADYRGEVKVVLFNHSEKDFEI
DUT_BUCAI       PTGIAIYIANPNITALILPRSGLGHKKGIVLGNLVGLIDSDYQGQLMISLWNRSDQDFVV
                 *.: : :.  .  . : ***.*. :: :        * **:**:* : : * * .     :

DUT_AQUAE       ERGERIAQLVIAPVQRVEVVEV--EEVSQTQRGEGGFGSTGTK---
DUT_BRAJA       KRGERIAQMVIAPVVQAALVPV--ATLSATDRGAGGFGSTGR----
DUT_BPT5        ENFERLCQLVVLPHYSTHNFKI-VDELEETIRGEGGFGSSGSK---
DUT_CANAL       KEGDRIAQLVLEQIVNADIKEISLEELDNTERGEGGFGSTGKN---
DUT_BUCAI       NPHDRVAQIIFVPIIRPCFLLV--KNFNETSRSKKGFGHSGVSGVI
                :  :*:.*::.           :    .. * *.   *** :*
```

Invariant regions