CHAPTER 3

Introduction to the BLAST Suite and BLASTN

Key concepts

- Why and how to search a sequence database
- An introduction to nucleotide BLAST (BLASTN)
- Interpreting BLASTN results
- Cross-species searches: paralogs, orthologs, and homologs

3.1 INTRODUCTION

In Chapter 2 we learned how to search databases with text queries. All of these were exact matches—that is, we were expecting to find the exact accession number or exactly spelled words. In this chapter, a much harder database-searching problem is introduced. How do you find matches when your **query** is not a short accession number or a text term, but instead a DNA sequence that is 500 nucleotides long? In addition to finding all the exact matches, can you find those sequences with mismatches, clearly related to the query but not 100% identical? For all the hits that are not exact matches, can calculations generate statistics that help evaluate which hits are significant, and which should be ignored? On top of these challenges, can this search of a database, that contains millions of sequences, show the results in a reasonable time? These and other questions will be answered here. A computer program called BLAST is one of the most commonly used tools in bioinformatics and will be introduced in this chapter. The next three chapters will explore further uses of BLAST.

Why search a database?

Let's assume that you have an unknown sequence and you use it as a query to search a bioinformatics database:

- Is the query identical to something already in the database? Is the query a known gene? If so, you can learn a lot about this gene by looking at the annotation in the sequence records.
- Is your query just a small piece of a much larger gene? If so, you may have just found a way to obtain the rest of the gene. Or based on these results, your laboratory technique may need to be improved upon.
- Is the query similar to something already in the database? Has it already been found in another organism, or is it similar to something in the same organism? Did you just find members of a gene family? These **sequence similarities** may tell you something about the function of your sequence.

- Is the query unique? Never been seen before? Perhaps you have discovered a new gene!

You are asking a handful of questions every time you do a sequence similarity search.

3.2 WHAT IS BLAST?

BLAST, or the Basic Local Alignment Search Tool, was specifically designed to search nucleotide and protein databases. It takes your query (DNA or protein sequence) and searches either DNA or protein databases for levels of **identity** that range from perfect matches to very low similarity. Using statistics, it reports back to you what it finds, in order of decreasing significance, and in the form of graphics, tables, and alignments. There are multiple forms of BLAST, but in this chapter we concentrate on nucleotide BLAST (**BLASTN**, pronounced "blast en"). The query is a DNA sequence and the database you search is populated with DNA sequences, too. **Table 3.1** outlines the query and the subjects of the search.

How does BLAST work?

As mentioned previously, many millions of DNA sequences have been collected in databases and are available for searching at Websites such as those at the NCBI. To conduct sequence-based queries with BLAST, databases must be of a special format to optimize for this type of query. The annotation associated with these individual files must be removed leaving just the sequences. Links to the annotation are still maintained so the identities of these sequences are not lost. Each sequence in the database is then broken into **words** or short sequences for comparison to the query.

When a search is submitted, BLASTN first takes your query DNA sequence and breaks it into words that are quite short (11 nucleotides). It then compares these words to those in the database. As BLAST has to compare many millions of words in this manner, and subsequent steps can be time-consuming, BLAST looks for two adjacent word pairs and, if their similarities and distance between words are acceptable, only those advance to the next set of calculations.

Starting with this local similarity, BLAST then tries to extend the similarity in either direction (**Figure 3.1**). Using the sequence immediately upstream and downstream of the word in the original query, BLAST starts keeping track of the consequences of lengthening the alignment between the query and the sequence in the database. Still matching? The significance score increases. Mismatches encountered? Penalty points accumulate until the cost outweighs the benefit and BLAST stops extending. This approach is finely tuned; if the penalty threshold for extension is too low, BLAST would stop trying to extend similarity very quickly and distant but significant sequence similarities would be missed. If the penalty threshold were too high, BLAST would be given too much freedom to keep extending past real areas of similarity and start collecting many truly insignificant hits.

Finally, all the **alignments** between the query and the database subjects, or "high-scoring subject pairs" (HSPs), are ranked based on length and significance. The best hits are kept and shown to you in the forms of a graphic, a table, and alignments between the query and the hits.

Table 3.1 BLASTN definition

Type	Query	Database
BLASTN	Nucleotide	Nucleotide

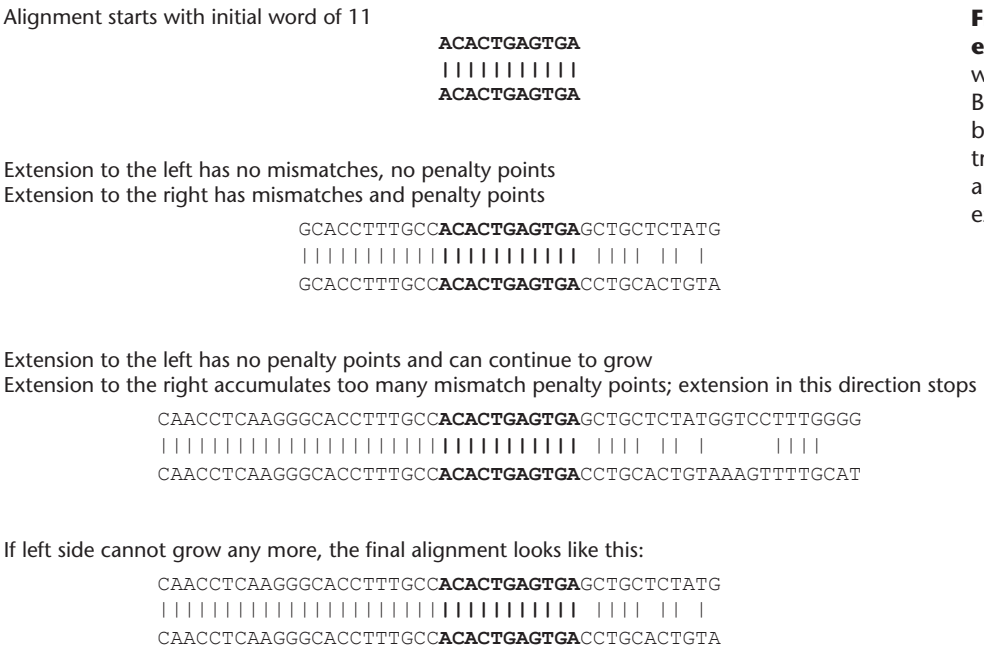


Figure 3.1 Simple extension example for BLASTN. Starting with an initial match of "words," BLAST extends the alignment between query and hit, keeping track of penalty points against, and increasing significance for, extending the alignment.

3.3 YOUR FIRST BLAST SEARCH

BLAST may be the most widely used sequence analysis program in the world. It is available as a tool from many Websites but is also downloadable as an application that works on your local personal computer or powerful server. It is free, but commercial parties have also created enhanced BLAST applications and charge a fee for these products. Please note that here only the Web form of BLAST will be discussed.

Below is a step-by-step description of your first BLAST search. In future use, these searches can take less than a minute, including your analysis.

Find the query sequence in GenBank

For your first BLASTN search, we'll use a sequence where information is very limited. This may be very typical of what you will face: you have an unknown sequence and you want to determine its identity, if possible.

1. Go to the NCBI Website, ncbi.nlm.nih.gov.
2. Near the top of the home page is the Entrez drop-down menu of database choices. Select "Nucleotide."
3. In the text field next to the menu, enter the GenBank accession number, DD148865. Accession numbers are unique identifiers for sequence records in the GenBank database. By searching with this accession number, you will find just one DNA sequence file.
4. Either press the return key on your keyboard or hit the Search button on the Web page. The Web page refreshes and you will see the nucleotide file.

There are two main sections to this file (**Figure 3.2**): on top, information about this sequence and, below it, the DNA sequence. The information of a sequence record is generally referred to as the annotation. Although the DNA sequence is the key information for all GenBank records, the annotation section can be a rich source of knowledge about the sequence and should not be overlooked. Databases such as GenBank have a specific structure to their annotation: fields

Figure 3.2
A GenBank file.

```
LOCUS       DD148865                 631 bp    DNA     linear   PAT 04-NOV-2005
DEFINITION  A group of genes which is differentially expressed in peripheral
            blood cells, and diagnostic methods and assay methods using the
            same.
ACCESSION   DD148865
VERSION     DD148865.1  GI:92839210
KEYWORDS    JP 2005102694-A/175.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 631)
  AUTHORS   Nojima,H.
  TITLE     A group of genes which is differentially expressed in peripheral
            blood cells, and diagnostic methods and assay methods using the
            same
  JOURNAL   Patent: JP 2005102694-A 175 21-APR-2005;
            Japan Science and Technology Agency,Hiroshi NOJIMA,GeneDesign Inc
COMMENT     OS   Homo sapiens
            PN   JP 2005102694-A/175
            PD   21-APR-2005
            PF   09-SEP-2004 JP 2004263092
            PI   hiroshi nojima
            CC
            FH   Key             Location/Qualifiers
            FT   misc_feature    (17)..(17)
            FT                   /note='n is a, c, g, or t'.
FEATURES             Location/Qualifiers
     source          1..631
                     /organism="Homo sapiens"
                     /mol_type="unassigned DNA"
                     /db_xref="taxon:9606"
ORIGIN
        1 gcaactgtgt tcactancaa cctcaaacag acaccatggt gcatctgact cctgaggaga
       61 agtctgccgt tactgccctg tggggcaagg tgaacgtgga tgaagttggt ggtgabgccc
      121 tgggcaggct gctggtggtc tacccttgga cccagaggtt ctttgagtcc tttggggatc
      181 tgtccactcc tgatgctgtt atgggcaacc ctaaggtgaa ggctcatggc aagaaagtgc
      241 tcggtgcctt tagtgatggc ctggctcacc tggacaacct caagggcacc tttgccacac
      301 tgagtgagct gcactgtgac aagctgcacg tggatcctga gaacttcagg ctcctgggca
      361 acgtgctggt ctgtgtgctg gcccatcact ttggcaaaga attcacccca ccagtgcagg
      421 ctgcctatca gaaagtggtg gctggtgtgg ctaatgccct ggcccacaag tatsactaag
      481 ctcgctttct tgctgtccaa tttctattaa aggttccttt gttccctaag tccaactact
      541 aaactggggg atattatgaa gggccttgag catctggatt ctgcctaata aaaaagcatt
      601 tattttcatt gcaaaaaaaa aaaaaaaaaa a
```

of information that are the same in each record. Examples include definition, accession number, and species of origin. We will spend a lot of time examining and utilizing the annotation of files so let's look at this record's annotation closely.

Down the left side of the annotation are section labels, all in uppercase. The LOCUS line contains some basic information: the name, usually synonymous with the accession number (DD148865); the length (631 bp); the type of molecule which was the source of this sequence (DNA); the topology of the source material (linear or circular); the division code (in this case, PAT, which stands for the Patent division of GenBank); and the date when the file was created or underwent revision (04-NOV-2005).

For a full description of the fields in a GenBank file, see the "Release Notes" associated with the database. One way to find these is by using the drop-down menu at the top of the NCBI home page: select "NCBI Web Site" and enter "GenBank release notes" as a query.

The DEFINITION field is usually a brief description of the sequence, its origin, and any additional information that may prove valuable to the reader. The definition line for this record is unusually long, clearly taken from the title of the patent: "A group of genes which is differentially expressed in peripheral blood cells, and diagnostic methods and assay methods using the same."

The ACCESSION and VERSION lines are related. GenBank uses two types of unique identifiers. *Accession numbers* are unique to a sequence record, but should that record be revised, then the accession number is given a new version number. When this sequence was submitted to GenBank, it was given an accession number of DD148865, and the version was DD148865.1. If this record is updated, for example if the annotation is revised by the author, then it will be given a new version number, DD148865.2. Both versions will be kept. Rather than force you to know which version is the latest, GenBank lets you retrieve the latest version by just entering the accession number (DD148865), without the version number extension (.1), as you did when you were asked to retrieve this sequence. *GI numbers* are also assigned as unique identifiers for each specific record, for practical reasons of constructing the GenBank database. Different versions of sequence records could have completely different GI numbers (for example, they won't be sequential or appended with numbers or letters). We will only be working with accession numbers in this book.

KEYWORDS is a field where authors can record any other information they feel might be useful. However, as this is an optional field, you cannot rely on it to comprehensively search a database for related records. In this case, the authors (inventors) put the patent number (JP 2005102694-A/175) in this field so a search of Nucleotide for "JP 2005102694" finds 305 sequences that are associated with this patent. Note that not all terms are indexed for searches. For example, using more of the original patent number (adding "-A/175") does not find these sequences.

The SOURCE and ORGANISM fields are related. The SOURCE is the genus and species according to the author of this record. This is a free text field, and so you might find variations and typographic mistakes; for example, *Homo sapien* instead of the correct *Homo sapiens*. The ORGANISM field is constrained and contains the full taxonomic classification for the source organism.

The REFERENCE field contains details about the sequence origin and can include publications. Also present under FEATURES is information found elsewhere in the annotation (for example, organism) but it may include additional facts or partial analysis results. We'll see this later when we examine sequence records more rich with information.

The COMMENT section may include information such as accession numbers of related sequences or references to other databases. In this patent record, it repeats basic details about the patent.

The last section of this file is the DNA sequence. It is one continuous sequence, broken up into groups of 10 nucleotides (with blank spaces in between), with 60 nucleotides per line. Each line is prefixed with the number of the first base it contains.

Convert the file to another format

The GenBank file described above contains annotation and features that are not used by BLAST and it is necessary to remove these components to run the search. This can be easily accomplished by converting this file into another file format called **FASTA** (pronounced "FAST-AY," the second syllable rhymes with "say").

1. Near the top of the sequence Web page there is a list of formats under "Display Settings" which includes GenBank, FASTA, Graphics, and more. Choose FASTA.
2. When the page refreshes you see the file in FASTA format (**Figure 3.3**).

```
>gi|92839210|dbj|DD148865.1| A group of genes which is differentially expressed in peripheral blood cells,
and diagnostic methods and assay methods using the same
GCAACTGTGTTCACTANCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGT
TACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGABGCCCTGGGCAGGCTGCTGGTGGTC
TACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACC
CTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCT
CAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGG
CTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGG
CTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATSACTAAGCTCGCTTTCT
TGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAA
GGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAAGCATTTATTTTCATTGCAAAAAAAAAAAAAAAAAA
A
```

Figure 3.3 GenBank file DD148865 in FASTA format.

Almost all the annotation is now missing. What description remains is on the top line which, in all FASTA files, begins with a ">" symbol. Next are fields, separated by vertical bars. The first field indicates the unique identifier for this file record (GI number 92839210). The next field shows the organization that first received this sequence (another sequence database called the DNA Data Bank of Japan or DBJ). This is followed by the accession number, and the definition. The DNA sequence begins on a new line. This signals "everything past this point is just DNA sequence": no numbers, spaces, or anything else among the sequence characters. This is the essence of the FASTA format: minimum annotation followed by sequence. In Figure 3.3 it appears that there are two lines of annotation because of the long definition. In fact, the annotation text has been "wrapped" to the next line due to page width.

This is a good opportunity to look at the sequence with your eyes. There are no numbers and spaces and with the simplicity of the format you might recognize details not visible in GenBank format. Do you see any pattern in the sequence? Any regions rich in As? Do you see any doublets or triplets that seem to be repeated throughout the sequence? Are there any bases that are not A, T, G, or C (**Box 3.1**)? This moment of "low-tech" examination will often reveal details that may help you interpret findings derived by other means. Throughout this book, take time to look at the sequences you encounter and your eyes will become trained at recognizing interesting features without the use of software. An obvious feature of this cDNA is the poly(A) tail.

Performing BLASTN searches

At the NCBI home page, the link to the BLAST forms is usually near the top of the list of Popular Resources. This hyperlink will take you to a page listing different types of BLAST. In this case we are using a DNA sequence query and will be searching a DNA database, so you will be using the "nucleotide blast" or BLASTN form. Navigate to this choice and we'll begin to populate the fields on the form.

- The Query Sequence field. The best way to provide sequence is the plain text found in FASTA format. Some Websites will take your DNA sequence and strip away any numbers, spaces, or otherwise non-DNA characters. This convenience is not found everywhere so it is best to get into the habit of using FASTA format. The NCBI also allows you to enter accession numbers, but in this case, paste the FASTA format of DD148865 into the Query Sequence field, including the annotation line. The BLAST program will grab the definition from this FASTA file and label your results. This will help you stay organized when conducting more than one BLAST search at a time. Depending on the Internet browser you are using, you may see the sequence underlined because the browser will interpret the 631 As, Ts, Gs, and Cs as a spelling error and underline the sequence, but you can ignore that warning.
- The Database field. Next we have to choose the database to search. There are many to choose from the drop-down menu; select the Reference RNA sequences, also known as **RefSeq** mRNA (refseq_rna) (**Figure 3.4**). This is a specialized, nonredundant database of sequences containing the definitive

reference sequence for RNAs. Note that RNA sequences are represented as DNA: you will see Ts instead of Us.

- The Organism field. This field allows you to limit your search to a specific organism's DNA. We could search all species, which is the default, but since the query is from *Homo sapiens*, and every human gene has (in theory) been sequenced, why not search a much smaller database and get a direct answer? Your results will also come back faster because of the smaller database size and the NCBI will use fewer computer resources to deliver your results. Enter "Homo sapiens" in this field (Figure 3.4). You'll see that as you begin to type this in the text box, choices will come up and you can finish your entry by clicking on "Homo sapiens (taxid:9606)."
- Program Selection. The default setting uses a version of BLASTN called megaBLAST. This version of BLAST is optimized to find nearly identical hits and is only found at the NCBI Website. We will often be looking for distantly related hits so BLASTN will be used in this book (**Figure 3.5**). Click on the radio button next to this choice.

Box 3.1 Uncertainty codes

As you work with nucleic acid sequence records and read scientific articles, you may come across representations of bases which are not A, T, G, or C. In fact there are several of these in Figure 3.3. These nonstandard letters are actually the uncertainty codes proposed by the International Union of Biochemistry and Molecular Biology (IUBMB) to represent certain groupings of nucleotides. On occasion, software or a scientist is unable to decide if a newly sequenced base is, for example, an A or a G. The three IUBMB symbols most frequently encountered are R for the puRines (A or G), Y for the pYrimidines (C or T), and N for aNy nucleotide. Here are the uncertainty codes:

IUBMB symbol	Definition
R	A or G
Y	C or T
K	G or T
M	A or C
W	A or T
S	C or G
B	C or G or T
D	A or G or T
V	A or C or G
H	A or C or T
N	G or A or T or C

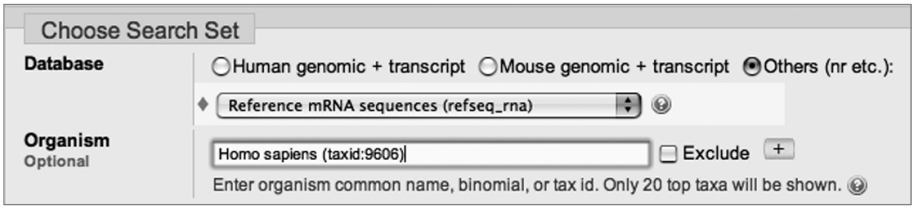


Figure 3.4 BLASTN database choices. The drop-down menu lists the databases and the species can be entered in the "Organism" field.

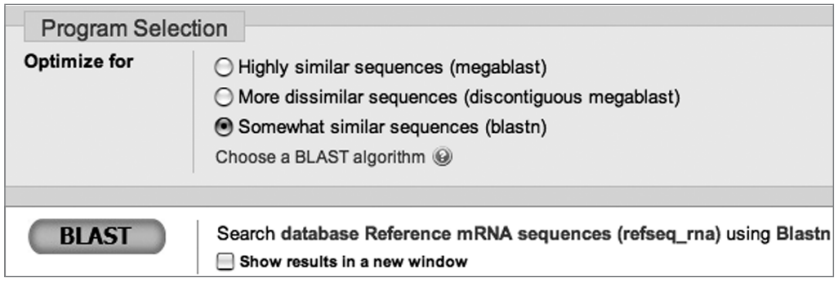


Figure 3.5 Select BLASTN as the program to use.

The search is now ready. Click on the "BLAST" button and wait for the results. Times will vary but this incredibly fast-working program will break your 631-nucleotide sequence into small "words," search a database measured in the many thousands of sequences, identify the best hits, try to extend and join the sequences of similarity, then generate the statistics so you can better evaluate the hits. By the time you finish reading this brief description, all of this may be performed and you are now looking at the results.

Depending on your browser settings, the next time you visit this BLASTN form, the defaults may now be changed to what you used in this search. Be sure to make it a habit to review all settings before initiating the search.

3.4 BLAST RESULTS

The results from a BLAST search are divided into three sections: the graphic pane, a results table, and the alignments between the query and the hits. Although some conclusions can be obtained based on interpretations of sections individually, it is best to consider all three sections and draw upon their complementary content during your analysis. These sections are described sequentially below, but when reviewing your results, you should move back and forth between sections as needed.

Graphic

The colorful graphic of the BLAST search results shows the query length across the top (see **Figure 3.6** and color plates). The line goes from 0 to over 600, corresponding to the 5P end and 3P end, respectively. The sequences found by BLAST, the "hits," appear below as horizontal bars in rows. If a hit lined up to every nucleotide of the query, then the bar goes across completely, as is almost the case in the first row. If other hits were similar to only the extreme 3P end of the query, then there would be short bars to the right, as there are in various rows of this graphic. To save space, several hits are often placed in the same row, as space and location allows. For example, the second row has a long red bar, and much shorter black and green bars (see Figure 3.6 and color plates). Bars in the same row do not represent different regions from the same sequence; they are different hits. Different regions of the same sequence are joined by a thin line, but there are none in this example. The color-coding within the graphic is generated by the statistics of each hit. As indicated by the key at the top of the graphic, hits with the highest score are red, the next highest are purple, then green, and so on. In this BLAST search, hits from all ranges of scores are found.

What we see in Figure 3.6 (see also color plates) is a single high-scoring hit which covers almost the entire length of the query (the red bar in the first row). Moving down the graph, there are five other high-scoring hits (red) that line up with (approximately) the first 475 nucleotides of the query. Two moderately scoring hits (purple) line up with about 200 nucleotides of the query, and the rest are short, low-scoring hits (green, blue, and black bars). These short bars appear in several stacks because members within the stack have something in common with the same place in the query. By floating your computer mouse over each

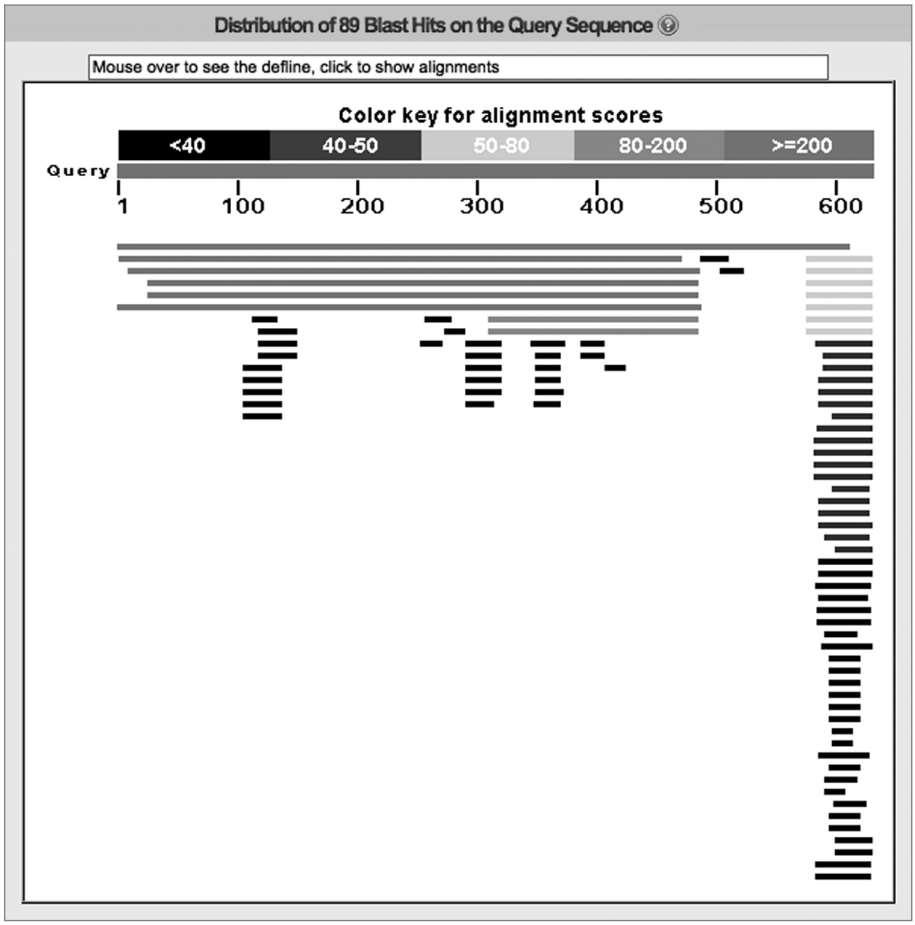


Figure 3.6 The graphic pane of the NCBI BLASTN results. The query coordinates and length correspond to the numbered scale across the top. Sequences found by BLAST, "hits," are represented as horizontal bars below this scale. These will vary in length, position, and color-coded scoring, shown here in gray shades. See color plates for a color version of this figure.

bar, the sequence definition line appears in the small window above the graphic. By clicking on the individual bars you can navigate to the alignments between the query and the hits.

It is important to realize that these short bars show the length of the similarity between the query and the hit and do not necessarily represent the entire length of the hit. For example, the hit might be one million bases long, but only contain 30 nucleotides in common with the query. This identity will be represented by a bar that is 30 nucleotides long.

Interpretation of the graphic

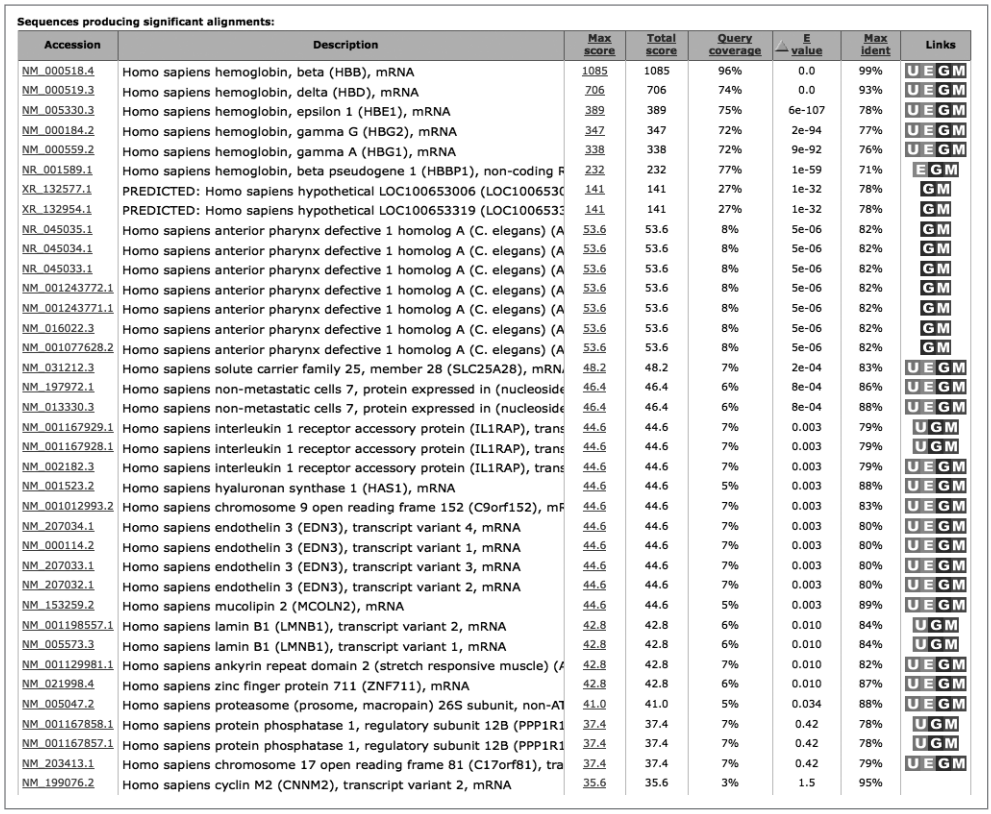
Even without knowing the identities of the hits shown in the graphic, you can still conclude several things. First, there appears to be only one *Homo sapiens* reference mRNA sequence that aligns with almost the entire query. That is, BLAST was able to align almost every section of the query, in a continuous fashion, across a similar length of sequence within the hit.

Considering the other high-scoring hits, it appears that the 3P end of the query does not align with anything else in the database that is high scoring (ignoring the low-scoring hits for now). The endpoint of the five short red bars is around nucleotide 475, and two other purple bars align between nucleotides 300 and 475. Does this reflect some kind of feature about the query? More information is needed to find the answer and this is discussed later.

Results table

Below the graphic is the BLAST results table that provides basic information about the hits along with the statistics of each hit. This table is long but the top hits are shown in **Figure 3.7**.

Figure 3.7 Tabular display of BLASTN hits. BLASTN lists the significant hits in a table that has both identifiers and descriptions of the hits, as well as statistical measures of the significance.



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_000518.4	Homo sapiens hemoglobin, beta (HBB), mRNA	1085	1085	96%	0.0	99%	UEGM
NM_000519.3	Homo sapiens hemoglobin, delta (HBD), mRNA	706	706	74%	0.0	93%	UEGM
NM_005330.3	Homo sapiens hemoglobin, epsilon 1 (HBE1), mRNA	389	389	75%	6e-107	78%	UEGM
NM_000184.2	Homo sapiens hemoglobin, gamma G (HBG2), mRNA	347	347	72%	2e-94	77%	UEGM
NM_000559.2	Homo sapiens hemoglobin, gamma A (HBG1), mRNA	338	338	72%	9e-92	76%	UEGM
NR_001589.1	Homo sapiens hemoglobin, beta pseudogene 1 (HBBP1), non-coding R	232	232	77%	1e-59	71%	EGM
XR_132577.1	PREDICTED: Homo sapiens hypothetical LOC100653006 (LOC100653(	141	141	27%	1e-32	78%	GM
XR_132954.1	PREDICTED: Homo sapiens hypothetical LOC100653319 (LOC100653:	141	141	27%	1e-32	78%	GM
NR_045035.1	Homo sapiens anterior pharynx defective 1 homolog A (C. elegans) (A	53.6	53.6	8%	5e-06	82%	GM
NR_045034.1	Homo sapiens anterior pharynx defective 1 homolog A (C. elegans) (A	53.6	53.6	8%	5e-06	82%	GM
NR_045033.1	Homo sapiens anterior pharynx defective 1 homolog A (C. elegans) (A	53.6	53.6	8%	5e-06	82%	GM
NM_001243772.1	Homo sapiens anterior pharynx defective 1 homolog A (C. elegans) (A	53.6	53.6	8%	5e-06	82%	GM
NM_001243771.1	Homo sapiens anterior pharynx defective 1 homolog A (C. elegans) (A	53.6	53.6	8%	5e-06	82%	GM
NM_016022.3	Homo sapiens anterior pharynx defective 1 homolog A (C. elegans) (A	53.6	53.6	8%	5e-06	82%	GM
NM_001077628.2	Homo sapiens anterior pharynx defective 1 homolog A (C. elegans) (A	53.6	53.6	8%	5e-06	82%	GM
NM_031212.3	Homo sapiens solute carrier family 25, member 28 (SLC25A28), mRN	48.2	48.2	7%	2e-04	83%	UEGM
NM_197972.1	Homo sapiens non-metastatic cells 7, protein expressed in (nucleoside	46.4	46.4	6%	8e-04	86%	UEGM
NM_013330.3	Homo sapiens non-metastatic cells 7, protein expressed in (nucleoside	46.4	46.4	6%	8e-04	88%	UEGM
NM_001167929.1	Homo sapiens interleukin 1 receptor accessory protein (IL1RAP), trans	44.6	44.6	7%	0.003	79%	UGM
NM_001167928.1	Homo sapiens interleukin 1 receptor accessory protein (IL1RAP), trans	44.6	44.6	7%	0.003	79%	UGM
NM_002182.3	Homo sapiens interleukin 1 receptor accessory protein (IL1RAP), trans	44.6	44.6	7%	0.003	79%	UEGM
NM_001523.2	Homo sapiens hyaluronan synthase 1 (HAS1), mRNA	44.6	44.6	5%	0.003	88%	UEGM
NM_001012993.2	Homo sapiens chromosome 9 open reading frame 152 (C9orf152), mR	44.6	44.6	7%	0.003	83%	UEGM
NM_207034.1	Homo sapiens endothelin 3 (EDN3), transcript variant 4, mRNA	44.6	44.6	7%	0.003	80%	UEGM
NM_000114.2	Homo sapiens endothelin 3 (EDN3), transcript variant 1, mRNA	44.6	44.6	7%	0.003	80%	UEGM
NM_207033.1	Homo sapiens endothelin 3 (EDN3), transcript variant 3, mRNA	44.6	44.6	7%	0.003	80%	UEGM
NM_207032.1	Homo sapiens endothelin 3 (EDN3), transcript variant 2, mRNA	44.6	44.6	7%	0.003	80%	UEGM
NM_153259.2	Homo sapiens mucolipin 2 (MCOLN2), mRNA	44.6	44.6	5%	0.003	89%	UEGM
NM_001198557.1	Homo sapiens lamin B1 (LMNB1), transcript variant 2, mRNA	42.8	42.8	6%	0.010	84%	UGM
NM_005573.3	Homo sapiens lamin B1 (LMNB1), transcript variant 1, mRNA	42.8	42.8	6%	0.010	84%	UGM
NM_001129981.1	Homo sapiens ankyrin repeat domain 2 (stretch responsive muscle) (A	42.8	42.8	7%	0.010	82%	UEGM
NM_021998.4	Homo sapiens zinc finger protein 711 (ZNF711), mRNA	42.8	42.8	6%	0.010	87%	UEGM
NM_005047.2	Homo sapiens proteasome (prosome, macropain) 26S subunit, non-AT	41.0	41.0	5%	0.034	88%	UEGM
NM_001167858.1	Homo sapiens protein phosphatase 1, regulatory subunit 12B (PPP1R1	37.4	37.4	7%	0.42	78%	UGM
NM_001167857.1	Homo sapiens protein phosphatase 1, regulatory subunit 12B (PPP1R1	37.4	37.4	7%	0.42	78%	UGM
NM_203413.1	Homo sapiens chromosome 17 open reading frame 81 (C17orf81), tra	37.4	37.4	7%	0.42	79%	UEGM
NM_199076.2	Homo sapiens cyclin M2 (CNNM2), transcript variant 2, mRNA	35.6	35.6	3%	1.5	95%	

The first column in this table lists the accession numbers of the hits in the database. These accession numbers are hypertext so you can follow these to the RefSeq records where you can find the annotation on these sequences.

The next column is the description. BLAST takes the information from the DEFINITION line in the database record and places it here, but because of space limitations, many of the descriptions will appear truncated.

The next two columns are associated with the statistics of the database search. As mentioned earlier, BLAST uses statistics to sort through all the hits, shows you only the best, and then tells you why (statistically) they are the best. The first of these numbers is called the **Max Score**. Although the average user of BLAST often overlooks it, the change in this score is often important. If you see a sudden drop in the Max Score, expect to see a change in the query–hit alignment length, quality, or both.

The next column, "Total Score," becomes important when BLAST finds multiple, but not joined, sections of similarity between the query and the hit. For each area of similarity, BLAST generates an alignment and a score. If the Max Score is equal to the Total Score, then only a single alignment is present. If the Total Score is larger than the Max Score, then multiple alignments must be present and their individual scores have contributed to the Total Score. For this BLASTN with DD148865, the values in Max Score and Total Score are identical, indicating that only single alignments were generated for this BLASTN search.

"Query Coverage" is the next column. The original query, DD148865, is 631 nucleotides long. If BLAST can align all 631 nucleotides of this query against a hit, then that would be 100% coverage. Remember, Query Coverage does not take into account the length of the hit, only the percentage of the query that aligns with the hit.

Next is the **E value** or **Expect value**, which represents the number of hits you would expect to find by chance given the quality of the alignment and the size of the database. If a database of only As, Ts, Gs, and Cs gets sufficiently large, you

start finding sequence similarities by chance, particularly with short queries. The E value in BLAST takes into account both the length and composition of the alignment along with the percentage identity found. A number close to zero means that the hit has to be significant and not due to chance. BLAST results tables are sorted by E value, the most significant hits appearing at the top. When there are two or more identical E values, the Max Score is then used to sort the hits.

The next column is called "Maximum Identity." BLAST calculates the percentage identity between the query and the hit in a nucleotide-to-nucleotide alignment. If there are multiple alignments with a single hit, then only the highest percent identity is shown. The last column, "Links," contains links to databases for that hit and these will not be discussed here.

Interpretation of the table

The top two hits are very significant, both having E values of 0.0. Even the eighth hit has a very small E value (1e–32) so this search has found a number of significant hits.

Based on the first line of the table, it appears that the unknown query used in this BLASTN is nearly identical to the human beta hemoglobin mRNA sequence. Over 96% of the query's length is 99% identical to the first hit. Just concentrating on the Maximum Identity column, the other top hits, although strong, show 93% or less identity when considering large portions of the query length. Note that the 6% drop in Maximum Identity and 22% drop in Query Coverage translated to a significant drop in the Max Score between the first and second hit. The descriptions of the next four hits reveal that the query has found other members of the hemoglobin family: delta, epsilon 1, gamma G, and gamma A. The next three hits, starting with accession number NR_001589, are a hemoglobin pseudogene and two predicted genes. The pseudogene may be both divergent and missing sequence found in the functional family members. Note that the query coverage is higher than the hit above it, but the percent identity is lower and the E value is greater. Gene predictions are often generated automatically, without human supervision, and can be based on incomplete experimental evidence. These may be missing portions, such as exons, of the real gene. The Maximum Identity of these two predictions to the query is identical to that seen with the epsilon 1 transcript (the third hit), but the Query Coverage is significantly lower. It would be interesting to analyze these sequences in detail to support or refute these predictions.

After the two predicted genes, there is a dramatic drop in the Query Coverage, and the E value makes a huge jump. Although there are multiple hits that have high Maximum Identity, the large E value (approaching or greater than 1) indicates that the identities seen can be due to chance. Exploring the annotation of these hits would find that they have significantly different functions from hemoglobin. Not knowing anything else about the identities that are seen for these low-scoring hits, it is easy to conclude that these are not significant.

The alignments

Below the table are the alignments (also called high-scoring subject pairs, HSPs) between the query and the hits. The statistics seen in the BLAST table are repeated here, along with additional important numbers. Within the alignments, the E value is called "Expect." The description line is now shown in its entirety and the length of the hit is shown. The alignments clearly show the relationships that BLAST has found between your query and the hits.

Examine the data in **Figure 3.8** above the alignment between the query (DD148865) and the first hit, NM_000518. The identity is 99%, with 607 nucleotides out of 611 nucleotides aligned. In the graphic above, this hit was shown as a solid red line almost all the way across the length of the query. Here, BLAST shows this by displaying horizontal pairs of sequence; the query is shown above the hit, now labeled the **Subject** (**Sbjct**). There are vertical lines between the two sequences wherever they have the same nucleotide. Continuous alignments are

Figure 3.8 BLASTN alignment between Query DD148865 and Sbjct, NM_000518. The Query and Sbjct lines are labeled and aligned bases have a vertical bar "|" between identical bases. Notice that nucleotide 2 of DD148865 aligns with nucleotide 17 of NM_000518, and nucleotide 61 aligns with nucleotide 76.

```
>ref|NM_000518.4| Homo sapiens hemoglobin, beta (HBB), mRNA
Length=626

 Score = 1085 bits (1202),  Expect = 0.0
 Identities = 607/611 (99%), Gaps = 1/611 (0%)
 Strand=Plus/Plus

Query  2    CAACTGTGTTCACTANCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAA  61
            ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  17   CAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAA  76

Query  62   GTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGABGCCCT  121
            ||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  77   GTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCT  136

Query  122  GGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCT  181
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  137  GGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCT  196

Query  182  GTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCT  241
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  197  GTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCT  256

Query  242  CGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACT  301
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  257  CGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACT  316

Query  302  GAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAA  361
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  317  GAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAA  376

Query  362  CGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGC  421
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  377  CGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGC  436

Query  422  TGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATSACTAAGC  481
            |||||||||||||||||||||||||||||||||||||||||||||||||| |||||||
Sbjct  437  TGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGC  496

Query  482  TCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTA  541
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  497  TCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTA  556

Query  542  AACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAAGCATTT  601
            |||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  557  AACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAA-CATTT  615

Query  602  ATTTTCATTGC  612
            ||||||||||
Sbjct  616  ATTTTCATTGC  626
```

displayed in lines of 60 nucleotides, with both query and subject wrapping to new lines until the alignment ends. Each line is numbered at the beginning and the end, allowing you to see where the alignment begins and ends.

Looking at this first alignment you can easily conclude that your "unknown" query, DD148865, is *Homo sapiens* beta hemoglobin. The query is nearly fully aligned to a reference sequence for human beta hemoglobin, NM_000518. What do the statistics tell you? This first hit is not a chance finding. In fact, the chance that you would encounter this randomly is so low that the calculated E value is 0.0: there is a zero chance you would find this by accident. The unknown must be *Homo sapiens* beta hemoglobin mRNA. Look back at Figure 3.2 and read the title of the patent associated with this sequence: "A group of genes which is differentially expressed in peripheral blood cells, and diagnostic methods and assay methods using the same." Our conclusion is consistent with the subject matter

of the patent. Would it have been easier to just read the patent? It is important to remember that many sequences are published in patents before their identities are known, and so reading the patent may not reveal the identity. In this case, a simple BLAST search gave us the identity.

The graphic (see Figure 3.6 and color plates) and the alignment (Figure 3.8) might suggest that DD148865 is a complete copy of beta hemoglobin, at least compared to this reference sequence, but it is not so. Notice the numbering of the lines at the beginning, or the 5P end, of the sequence. These coordinates indicate that the second nucleotide of the query aligns to nucleotide 17 of the reference sequence. At the 5P end, DD148865 is missing 16 nucleotides possessed by the reference sequence, NM_000518. In DD148865, the first base is a G while the 16th base of NM_000518 is an A. Rather than start the alignment with a mismatch, BLAST just trimmed off the first base of DD148865. At the other end of this first line, see that nucleotide 61 of the Query lines up with nucleotide 76 of the Sbjct.

The statistics just above this alignment indicate that there are 607 nucleotides aligned out of 611 in this alignment. Where are the four mismatches? The one discussed above at the beginning of the query is not included in these numbers because it is not shown. The first is in the first row of the alignment at coordinate 17 of the query. The authors knew there was a base here but could not tell which base. So the IUBMB symbol N for "any base" was used. The second mismatch is in line two of the alignment, at coordinate 116 of the query. The authors were not sure which nucleotide was here but they knew it was a C, G, or T. Thus, the IUBMB symbol for these possibilities (B) appears in this position. In the reference sequence NM_000518, the nucleotide in this position is a G, consistent with the nucleotide code (B) in DD148865. The third mismatch is at query coordinate 474 where there is an S (C or G) in the query and a C in the reference position, again consistent with the correct base.

The final mismatch seen at coordinate 596 of the query is different (Figure 3.8). Here, the query has an extra base, a G, so a gap (–) is inserted in the reference sequence at this position. This maintains the alignment for the remainder of both sequences. Which is correct? The reference sequence should definitely be given serious consideration; however, this single base may be key to the author's reason for submitting this sequence. Is this an interesting mutation, either natural or synthetic? Or is the sequence just incorrect at this position? Further reading and analysis would be required to answer this question. For example, can you find any other human beta globin mRNA sequences with this extra base?

Look at the last lines of the alignment between this query and the hit: nucleotide 612 of DD148865 aligns with nucleotide 626 of NM_000518. What is the length of the entire query sequence? Looking back at the annotation for this record, the DD148865 sequence is 631 nucleotides long (top line of Figure 3.2). The alignment stops short of the poly(A) stretch, visible earlier at the end of the DD148865 sequence file (see Figure 3.3). Notice that the length of the reference sequence NM_000518 is given in Figure 3.8 as "Length=626." BLAST took the alignment as far as it could go and then stopped when it ran out of reference sequence at the 3P end.

An added complication to the alignment interpretation is the poly(A) tail in DD148865. The poly(A) tails of many sequences are trimmed before the sequence is submitted to GenBank and other databases. Others are not trimmed, as seen in this example. It is possible that the original poly(A) tail for NM_000518 started in the same location as DD148865. But the poly(A) tail for sequences is variable in both length and location. Furthermore, RefSeq sequences are compiled from one or more sequences and the curators of RefSeq may have trimmed this sequence here for other reasons. Regardless, we can't be sure what's going on at the 3P end, but the rest of the alignment is excellent and requires little consideration.

As you work through this book, it is highly recommended that you generate simple drawings (pencil and paper are fine!) to understand the relationships that BLAST finds for you. **Figure 3.9** is an example "sketch" which shows the relationship

Figure 3.9 A rough "sketch" outlining the relationship between the query and the hit. Coordinates are taken directly from the BLAST alignment. Although this is an "electronic" sketch, pencil and paper work fine.

between the sequences described above, based on the coordinates. It shows that although the query is missing the 5P end, it has more nucleotides at the 3P end, although they are a stretch of As.

Does this mean that DD148865 is only a fragment? This is a strong possibility, but it will take more analysis to confirm this. Remember that cDNA synthesis is an imperfect process and truncations, particularly at the 5P end of cDNAs, are very common. The poly(A) tail is an obvious difference but database records are inconsistent about this feature. There are many sequences in RefSeq that have poly(A) tails, but NM_000518 does not. Remember that a poly(A) tail is added to transcripts after transcription takes place, so if you look at the genomic DNA, you would not see a stretch of As at the end of the last exon.

Other BLASTN hits from this query

It is a good habit to look at additional alignments to confirm any conclusions you make from the first alignment and to learn more about your query. The second hit, NM_000519, is delta hemoglobin (**Figure 3.10**). If you knew nothing about the beta hemoglobin gene before, you now know that it is a member of a gene family and it is quite similar to another member, delta. The identity between the query and this second hit is 93%, 438 nucleotides out of 470.

Looking at the alignment closely, you can now see nucleotides that do not align. For example, in the second line you can see that seven nucleotides are mismatched and missing the vertical line of identity between the query and reference sequences. Still, 93% identity is high and visually the alignment still appears strong. If these nucleotide differences fell within the coding region of the delta hemoglobin transcript, these could translate into a different amino acid sequence, but delta and beta hemoglobin would look quite similar in protein sequence nevertheless.

The alignment length is shorter than that seen with the first hit. Only 438 nucleotides out of 470 nucleotides from the query align with delta hemoglobin (remember, the total length of the query is 631 nucleotides). The decrease in this length came from the 3P end. This can be seen with the 3P coordinates of the query (472 versus 612, above). However, the 5P end and 3P end of delta hemoglobin mRNA appear to be longer than that of beta hemoglobin. Draw a picture to see the difference.

The annotation for the delta hemoglobin sequence indicates that the coding region is from nucleotides 196 to 639 of NM_000519. These numbers are in good (but not exact) agreement with the alignment. That is, the coding region between these two globins is relatively conserved and aligns well with BLAST, but the noncoding regions have diverged enough to not align. It appears that sequence immediately upstream of the coding regions (approximately nucleotides 163–195) is still quite conserved, reflecting the importance of this sequence.

Comparing the first two hits, the E value stays the same (0.0), but the Score (Max Score in the results table) drops from 1085 to 706. The drop in score reflects the decrease in alignment length between the query and the hit, from 611 to 470 nucleotides, as well as the drop in percent identity. This shows how the Max Score can sometimes tell you more than the E value.

Let's now skip to the seventh alignment between the query and XR_132577 (**Figure 3.11**). This is a predicted mRNA. It has never been isolated in the laboratory, but instead it was stitched together by a gene prediction tool that took cDNAs

```
>ref|NM_000519.3 Homo sapiens hemoglobin, delta (HBD), mRNA
Length=774

 Score =  706 bits (782),  Expect = 0.0
 Identities = 438/470 (93%), Gaps = 0/470 (0%)
 Strand=Plus/Plus

Query  3    AACTGTGTTCACTANCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAG  62
            ||| |||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  163  AACAGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAG  222

Query  63   TCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGABGCCCTG  122
             |||| || | ||||||||||||||| |||||||||||| |||||||||||| ||||||
Sbjct  223  ACTGCTGTCAATGCCCTGTGGGGCAAAGTGAACGTGGATGCAGTTGGTGGTGAGGCCCTG  282

Query  123  GGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTG  182
             |||||  | ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  283  GGCAGATTACTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTG  342

Query  183  TCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTC  242
             ||| |||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  343  TCCTCTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAGGTGCTA  402

Query  243  GGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTG  302
             |||||||||||||||||||||||||||||||||||||||||||||||| ||| |    |||
Sbjct  403  GGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACTTTTTCTCAGCTG  462

Query  303  AGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAAC  362
             ||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||
Sbjct  463  AGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCTTGGGCAAT  522

Query  363  GTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCT  422
             |||||||| ||||||||||||   |||||||||| |||||||||||| | ||||||||
Sbjct  523  GTGCTGGTGTGTGTGCTGGCCCGCAACTTTGGCAAGGAATTCACCCCACAAATGCAGGCT  582

Query  423  GCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTA  472
             |||||||||| |||||||||||||||||||||||||| ||||||||
Sbjct  583  GCCTATCAGAAGGTGGTGGCTGGTGTGGCTAATGCCCTGGCTCACAAGTA  632
```

Figure 3.10 Alignment between query DD148865 and the second BLAST hit, delta hemoglobin NM_000519. Nonidentical nucleotides lack the vertical bar "|" seen between identical nucleotides.

```
>ref|XR_132577.1| PREDICTED: Homo sapiens hypothetical LOC100653006
(LOC100653006), miscRNA
Length=255

 Score =  141 bits (156),  Expect = 1e-32
 Identities = 136/175 (78%), Gaps = 0/175 (0%)
 Strand=Plus/Minus

Query  312  CACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTC  371
             |||||||||||||||| |||||||||||||||| | ||||||||| || ||||||||
Sbjct  255  CACTGTGACAAGCTGCATGTGGATCCTGAGAACTTAAAGCTCCTGGGAAATGTGCTGGTG  196

Query  372  TGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAG  431
                ||  ||||    || || ||||||||||||||||    ||||||||| |||  |||
Sbjct  195  ACCGTTTTGGCAATCCATTTCGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCCTGGCAG  136

Query  432  AAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATSACTAAGCTCGCT  486
             ||  ||||| |||| ||||| | ||||||| ||  ||  ||  ||| ||||| ||
Sbjct  135  AAGATGGTGACTGGAGTGGCCAGTGCCCTGTCCTCCAGATACCACTGAGCTCACT  81
```

Figure 3.11 The seventh hit (XR_132577) from the BLASTN results with query DD148865. This gene prediction is very divergent compared to the top hits.

originating from this genomic region as evidence for its existence. This predicted transcript shows enough similarity to our query to appear on the list of hits with a very significant E value: 1e–32. There is 78% identity over a 175-nucleotide alignment, much smaller than the values seen with beta and delta hemoglobin. The sequence seems to be related to the query, but is clearly not a close member of the family. Maybe an exon was copied and moved to a distant location.

Let's return to the results table for a moment. Below the two predicted sequences (XR_132577 and XR_132954) the table shows a number of hits with increasing E values, well above the very low numbers seen at the top of the table. The biological functions of these low-scoring hits are also very varied, with no common theme for primary function. There is a cluster of hits to **transcription variants** of "Homo sapiens anterior pharynx defective 1 homolog A," and the alignment of one of these is shown in **Figure 3.12A**. Notice that the poly(A) tail of the query contributed a large number of the aligned bases. BLAST tries to emphasize this to you by changing these nucleotides (a simple sequence) to lowercase. There may be some significance to these two 3P ends being similar but the statistics say that care should be taken. Without the poly(A) contribution, only 36 nucleotides of the query's remaining 612 nucleotides show some similarity to this transcript, which is hardly significant.

Hit number 44 (NM_000794) of this BLASTN result has a Max Score of 35.6 and an E value of 1.4. By comparison, the first hit has a Max Score of 1085 and E value of 0.0. Alignments with numbers like those for hit number 44 can be expected to be very short and insignificant (Figure 3.12B).

Now go back to the graphic. By floating your mouse over the bars in the graphic, find the bar that represents the alignment with the dopamine receptor D1 (NM_000794) seen in Figure 3.12B (hint: use the coordinates of the query and the score-expected color to narrow your region of searching). There are other hits that align to the same region, based on the stack of bars in the graphic. What does this mean? Dopamine is a neurotransmitter and this receptor is expressed in the brain. This function seems very different to that of the hemoglobins, which is to bind oxygen in the blood. A quick look at the dopamine receptor annotation shows that the region of alignment, between nucleotides 1053 and 1081, is

Figure 3.12 Low-scoring alignments from BLASTN with query DD148865. Compared to the top hit, these alignments have a very low Score, a high Expect value, and a very short alignment length. (A) Hit number 9, NR_045035. (B) Hit number 44, NM_000794.

(A)

```
>ref|NR_045035.1| Homo sapiens anterior pharynx defective 1 homolog A (C. elegans)
(APH1A), transcript variant 7, non-coding RNA
Length=2216

 Score = 53.6 bits (58),  Expect = 5e-06
 Identities = 45/55 (82%), Gaps = 3/55 (5%)
 Strand=Plus/Plus

Query  577   GATTCTGCCTAATAAAAAAGCATTTATTTTCATTGCaaaaaaaaaaaaaaaaaaa  631
              |||| || |||||||||||| ||   || | ||||  ||||||||||||||||||
Sbjct  2159  GATTTTGACTAATAAAAAAGAAT---TTGTAATTGTGAAAAAAAAAAAAAAAAAA  2210
```

(B)

```
>ref|NM_000794.3| Homo sapiens dopamine receptor D1 (DRD1), mRNA
Length=3373

 Score = 35.6 bits (38),  Expect = 1.4
 Identities = 25/29 (86%), Gaps = 0/29 (0%)
 Strand=Plus/Plus

Query  347   CAGGCTCCTGGGCAACGTGCTGGTCTGTG  375
              || ||||||||| |||  ||||||||||
Sbjct  1053  CACGCTCCTGGGGAACACGCTGGTCTGTG  1081
```

in the coding region of the receptor. Remembering that codons are in groups of three nucleotides, a quick calculation says that this small alignment of 29 nucleotides encodes about 10 amino acids which appear to be similar between beta hemoglobin and the dopamine D1 receptor. Considering that these two proteins are 147 and 477 amino acids long, respectively, such a short similarity does not suggest similarity in protein function. But, nevertheless, it would be fun to investigate this small stretch of amino acids, especially since many other proteins share something in common with hemoglobin. You will need additional skills to study this, and you will receive them later on in this book.

Simultaneous review of the graphic, table, and alignments

Although we reviewed the results sections independently, we did move back and forth between the sections to get a better view of the independent pieces of data. When getting your first look at BLAST results, it is often helpful to do a quick review of all the sections to get a general understanding of the results, and then examine the sections individually and more slowly for details. While looking at the table, you can visualize the hits across the query because you saw them in the graphics panel. When looking at the alignments, you can visualize the trend in the scores because you saw them in the table. And when you look at the graphics, you can visualize the identities because you have scanned the alignments. Now that you understand the components of the BLAST output, let's look at the results again more quickly and introduce some additional results. Below is a fast-paced narrative to better demonstrate the quick review. In each case, look back at the section figure and follow along.

- The graphic shows a single high-scoring hit which stretches from end to end of the query. Five other high-scoring hits appear to end around query coordinate 475. Another pair of hits aligns from approximately nucleotide 300 to 475. Finally, there are multiple groups of low-scoring hits that are quite short in length. Some of these are at the 3P end of the query—maybe the poly(A) stretch is responsible for most of these hits. Floating the mouse over the top red bar shows (in the small text window above the graphics pane) that the best hit is beta hemoglobin. The next five red bars represent the four members of the hemoglobin family (delta, epsilon 1, gamma G, and gamma A) and a pseudogene. Their bars are shorter than that of beta hemoglobin, consistent with the drop in Query Coverage seen in the table. The black bars are short and correspond to the low-scoring hits in the table. There is some vertical stacking of the black bars so this will have to be analyzed further by looking at the alignments.
- The table descriptions show multiple hits to members of the hemoglobin family. The Max Scores reflect a drop in quality and length of alignment (Query Coverage) after the best hit, beta hemoglobin. Four family members and a pseudogene show good alignment length and E values, but none are as good as the first hit, so it is safe to say that our query is closest to beta hemoglobin. Two predicted genes appear high on the list, but have shorter query coverage than the other family members. There are many other hits listed but their scores, query coverage, and E values are all very poor. Their descriptions are also varied, with names that look very different than the globins that bind oxygen.
- The alignments reflect what we have seen in the graphic and table. The best hit is an almost base-for-base alignment between the query and human beta hemoglobin. There are some missing bases at the ends of the alignments, but these should have little impact on deciding the identity of our unknown query. The next few alignments show that the sequences of the hemoglobin family are quite similar but some significant differences at the 3P ends prevent full alignment. Except for some predictions, the rest of the hits appear to be insignificant.

3.5 BLASTN ACROSS SPECIES

Now let's perform another BLASTN search. Rather than try to identify an unknown, the goal of this search is to find hemoglobin genes in other animals.

BLASTN of the reference sequence for human beta hemoglobin against nonhuman transcripts

In the first search of this chapter, we identified the reference mRNA sequence for human beta hemoglobin. Go back to those results and, using the accession number, retrieve this sequence from the database in FASTA format. Keep this beta hemoglobin browser window open for later reference.

In another window, navigate to the NCBI BLASTN form and paste in the FASTA format for the reference sequence for human beta hemoglobin. When trying to identify our unknown, above, we restricted the BLASTN search to just reference sequences annotated as coming from *Homo sapiens*. This time, enter "vertebrates" in the Organism box, broadening the search but still restricting it to organisms that are likely to have hemoglobin (**Figure 3.13**). Again, select BLASTN (not megaBLAST) and launch the BLAST search.

When the screen refreshes and the results appear, perform your quick review of the results and then look at the details. Let's first look at the table (**Figure 3.14**).

The first hit is the reference human beta hemoglobin cDNA: as expected, this BLAST query found itself in the database. Below the human sequence are hits from a variety of vertebrates. Should these Latin names look unfamiliar to you, use the NCBI Taxonomy database to look up the common name for these species.

The leading hits are from primates, for example chimpanzee (*Pan troglodytes*), gibbon (*Nomascus leucogenys*), orangutan (*Pongo abelii*), marmoset (*Callithrix*

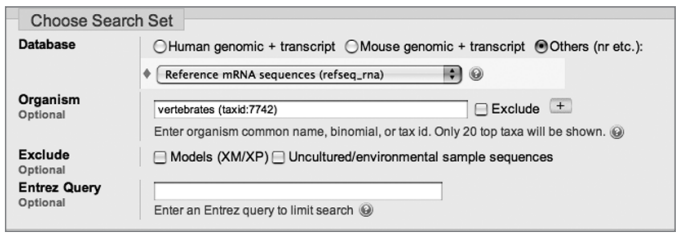
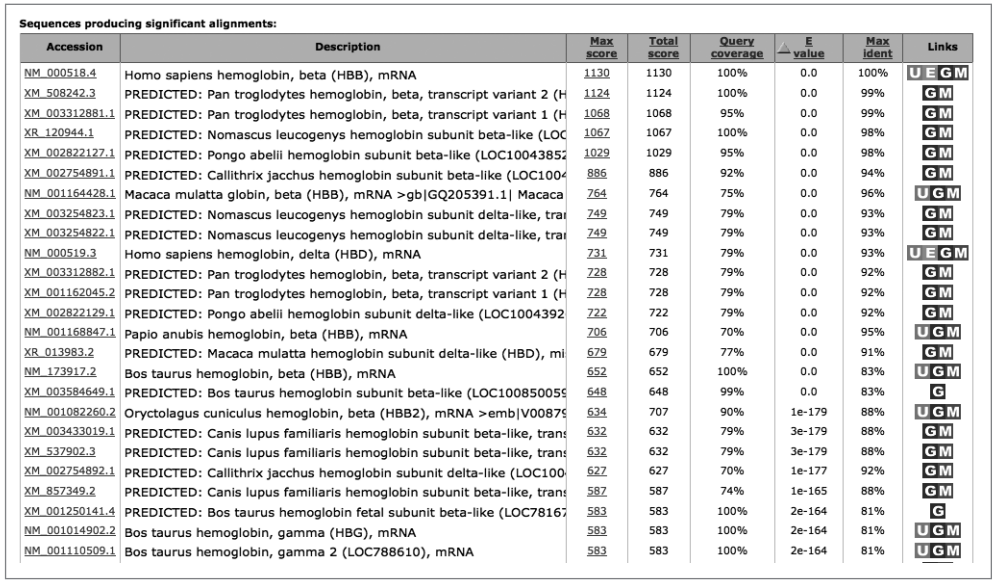


Figure 3.13 Configuring the BLASTN form to search reference mRNAs from other species.



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_000518.4	Homo sapiens hemoglobin, beta (HBB), mRNA	1130	1130	100%	0.0	100%	UEGM
XM_508242.3	PREDICTED: Pan troglodytes hemoglobin, beta, transcript variant 2 (H	1124	1124	100%	0.0	99%	GM
XM_003312881.1	PREDICTED: Pan troglodytes hemoglobin, beta, transcript variant 1 (H	1068	1068	95%	0.0	99%	GM
XR_120944.1	PREDICTED: Nomascus leucogenys hemoglobin subunit beta-like (LOC	1067	1067	100%	0.0	98%	GM
XM_002822127.1	PREDICTED: Pongo abelii hemoglobin subunit beta-like (LOC1004385	1029	1029	95%	0.0	98%	GM
XM_002754891.1	PREDICTED: Callithrix jacchus hemoglobin subunit beta-like (LOC100	886	886	92%	0.0	94%	GM
NM_001164428.1	Macaca mulatta globin, beta (HBB), mRNA >gb|GQ205391.1| Macaca	764	764	75%	0.0	96%	UGM
XM_003254823.1	PREDICTED: Nomascus leucogenys hemoglobin subunit delta-like, tra	749	749	79%	0.0	93%	GM
XM_003254822.1	PREDICTED: Nomascus leucogenys hemoglobin subunit delta-like, tra	749	749	79%	0.0	93%	GM
NM_000519.3	Homo sapiens hemoglobin, delta (HBD), mRNA	731	731	79%	0.0	93%	UEGM
XM_003312882.1	PREDICTED: Pan troglodytes hemoglobin, beta, transcript variant 2 (H	728	728	79%	0.0	92%	GM
XM_001162045.2	PREDICTED: Pan troglodytes hemoglobin, beta, transcript variant 1 (H	728	728	79%	0.0	92%	GM
XM_002822129.1	PREDICTED: Pongo abelii hemoglobin subunit delta-like (LOC1004392	722	722	79%	0.0	92%	GM
NM_001168847.1	Papio anubis hemoglobin, beta (HBB), mRNA	706	706	70%	0.0	95%	UGM
XR_013983.2	PREDICTED: Macaca mulatta hemoglobin subunit delta-like (HBD), mi	679	679	77%	0.0	91%	GM
NM_173917.2	Bos taurus hemoglobin, beta (HBB), mRNA	652	652	100%	0.0	83%	UGM
XM_003584649.1	PREDICTED: Bos taurus hemoglobin subunit beta-like (LOC100850059	648	648	99%	0.0	83%	G
NM_001082260.2	Oryctolagus cuniculus hemoglobin, beta (HBB2), mRNA >emb|V00879	634	707	90%	1e-179	88%	UGM
XM_003433019.1	PREDICTED: Canis lupus familiaris hemoglobin subunit beta-like, trans	632	632	79%	3e-179	88%	GM
XM_537902.3	PREDICTED: Canis lupus familiaris hemoglobin subunit beta-like, trans	632	632	79%	3e-179	88%	GM
XM_002754892.1	PREDICTED: Callithrix jacchus hemoglobin subunit delta-like (LOC100	627	627	70%	1e-177	92%	GM
XM_857349.2	PREDICTED: Canis lupus familiaris hemoglobin subunit beta-like, trans	587	587	74%	1e-165	88%	GM
XM_001250141.4	PREDICTED: Bos taurus hemoglobin fetal subunit beta-like (LOC78167	583	583	100%	2e-164	81%	G
NM_001014902.2	Bos taurus hemoglobin, gamma (HBG), mRNA	583	583	100%	2e-164	81%	UGM
NM_001110509.1	Bos taurus hemoglobin, gamma 2 (LOC788610), mRNA	583	583	100%	2e-164	81%	UGM

Figure 3.14 Human beta hemoglobin BLASTN results table, showing hits across many species.

jacchus), and Rhesus monkey (*Macaca mulatta*). Note that for many of these and other species, the mRNA is annotated as "Predicted" reflecting, in most cases, that the sequence is derived from genomic sequencing, not sequencing of cDNA. This indicates the amount of attention these other animals have received in studying their genes. Once you get outside of mainstream model organisms, relatively little mRNA/cDNA cloning of globins and many other genes has taken place. These genes were predicted based on very strong similarity to known mRNA sequences from human, mouse, rat, and other well-studied organisms.

Figure 3.14 shows that the E value for many of the top hits is 0.0. For these sequences, there is almost a twofold drop in Max Scores. The Query Coverage drops throughout the list until it gets to the cow (*Bos taurus*) where it is 100%. However, the percent identity between the human and cow sequence is only 81%, which pushed it down the list of hits.

Looking at the alignment to the predicted chimpanzee beta hemoglobin it is easy to see strong similarity between the human and chimpanzee sequences. The identities measurement indicates that the human and chimpanzee sequences are 99% identical. The alignment of 625 out of 626 nucleotides is nearly 100%, but this field uses whole numbers and rounds this value down to 99%. With this and many other human–chimpanzee sequence alignments, you can clearly see that at the DNA level, chimpanzees are our closest relatives.

Moving down the table or the alignments, certain trends are seen. Unlike the first BLASTN, which found all the members of the human hemoglobin family among the top hits, this search shows that most of the other human hemoglobin family members are not seen before many beta hemoglobin sequences in other species are found. *Homo sapiens* delta hemoglobin is the next human hit in the table, but no other human sequences are seen until much further down on the list. This indicates that, at least at the mRNA sequence level, the beta hemoglobins in these top species are more similar than human hemoglobin family members are to each other.

At the bottom of the table in Figure 3.14, among all the beta hemoglobin hits, notice that the cow (*Bos taurus*) hits include "gamma" and "gamma 2," which is a family member not seen in humans. As you explore genes from other species, you will see variation in names that will often reflect differences in physiology between organisms, differences in naming conventions, history, and even mistakes.

Looking at the graphic display for this BLASTN search (**Figure 3.15**), notice that many hits do not align with the 5P or the 3P ends of the query. Look at the alignment between the query and the *Equus caballus* (horse) beta hemoglobin (**Figure 3.16**). The alignment with the query starts at nucleotide 51 and ends with nucleotide 492. Why are so many sequences failing to align from end to end? The answer can be gathered from the alignment and the annotation of the sequence records.

Go to the GenBank record (NM_000518) and look at the annotation for the human beta hemoglobin and you see that the coding region sequence, abbreviated as **CDS**, starts at nucleotide 51, the A of the ATG start codon. The horse sequence, NM_001164018, starts at this exact base (horse nucleotide number 1 aligns with human nucleotide 51). The 5P untranslated region (UTR) is not even present in this horse reference sequence. Based on this BLAST search, it appears that the 5P untranslated regions of many vertebrate mRNAs are not present in the sequence records or cannot align well with the human untranslated region; hence the common start site for many of the alignments is in the vicinity of human nucleotide 51.

The 3P ends of many of these alignments terminate around query nucleotide 494 (seen in the graphic, Figure 3.15), which is the 3P boundary of the human beta hemoglobin coding region. Certainly, divergence of sequence in the 3P UTR explains some of these terminations. In the case of the horse sequence, there is

Latin species names

Throughout this book you will often encounter the Latin species names for organisms. Some will be easily recognized (for example, *Rattus*) but others can be guessed if you have knowledge of other topics, such as constellations of the evening sky, that also use Latin names. *Bos taurus* is the cow and the constellation is Taurus the bull. Canis major is a hunting dog of Orion, and the Latin name for the dog is *Canis familiaris*. Sheep is *Ovis aries*, and the zodiac sign is Aries.

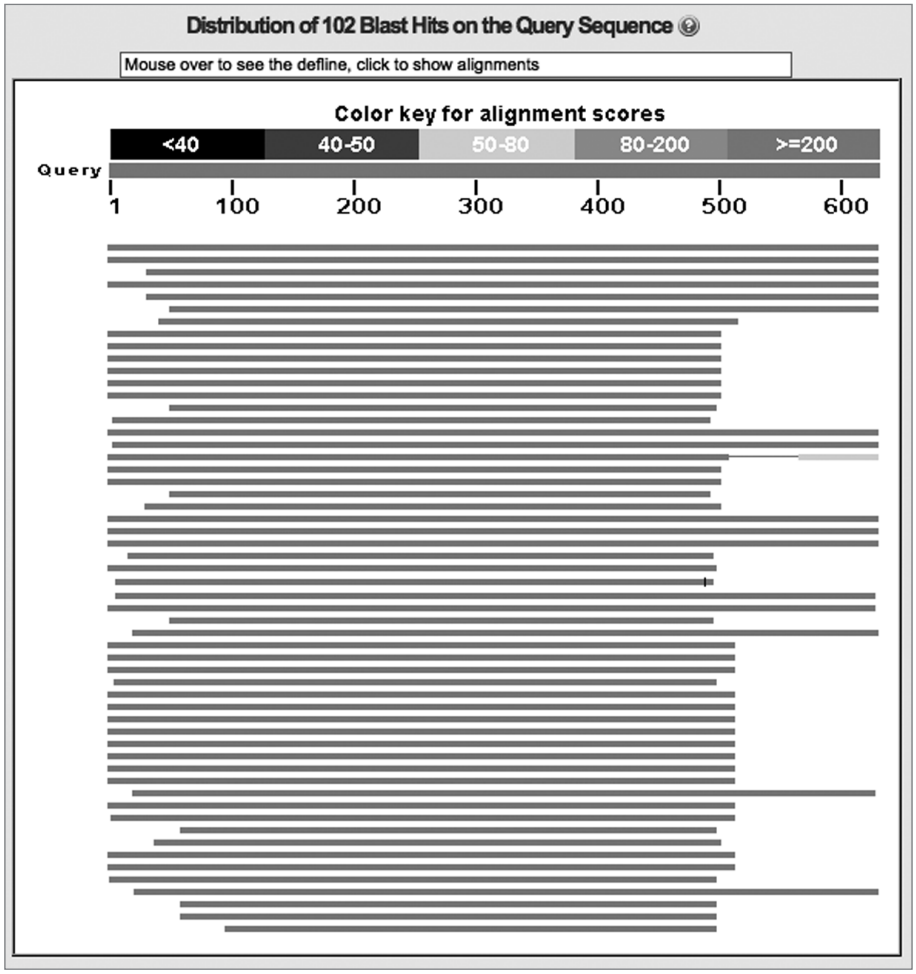


Figure 3.15 The BLASTN graphic of human beta hemoglobin mRNA against many species. Note the truncations at the 5P (left) end of the graphic as well as the distinct boundary around nucleotide 500 at the 3P end.

a simple reason to explain the sudden stop of alignment at horse nucleotide 442: the horse sequence comes to an end. Look at the description line of this alignment and it says "Length=444." Scroll through the alignments and you'll notice that many sequences do not include the 3P UTR of the mRNA. Many genomes are annotated in an automated fashion and genes are predicted based on similarity to known, well-annotated genes from other organisms. Like the 5P UTR, the 3P UTRs of many predicted transcripts are underrepresented in the database.

In general, gene coding regions are more conserved than the noncoding regions. A single nucleotide change in the coding region can change the amino acid sequence, possibly alter the structure and function of the protein, or introduce a stop codon and truncate the translation product. There are fewer constraints for sequence and function on untranslated regions. However, there are very important regulatory elements at work in untranslated regions. As long as regulatory elements, if present, are not disrupted, many nucleotide substitutions, insertions, and deletions are tolerated and lead to sequence differences in untranslated regions so extensive that they fail to align using BLAST.

Paralogs, orthologs, and homologs

In the first BLASTN search of this chapter, we were able to identify members of the human hemoglobin family: in order, beta (the unknown found the reference sequence for itself), delta, epsilon 1, gamma G, and gamma A hemoglobins. Based on the identities between our beta hemoglobin query and these hits, it is clear the family members are still closely related to each other, the lowest identity

```
>ref|NM_001164018.1|  Equus caballus hemoglobin, beta (HBB), mRNA
Length=444

Score =  522 bits (578),  Expect = 3e-146
 Identities = 381/442 (86%), Gaps = 0/442 (0%)
 Strand=Plus/Plus

Query  51   ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAAC  110
            ||||||||  ||||  |  ||| ||||||  | ||  ||    ||||||||||  |||||||||
Sbjct  1    ATGGTGCAACTGAGTGGTGAAGAGAAGGCAGCTGTCTTGGCCCTGTGGGACAAGGTGAAT  60

Query  111  GTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAG  170
            | |||  ||||||||||||||  ||||||||||||||||||||  ||||||||  |||||  |||
Sbjct  61   GAGGAAGAAGTTGGTGGTGAAGCCCTGGGCAGGCTGCTGGTTGTCTACCCATGGACTCAG  120

Query  171  AGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAG  230
            ||||||||||  ||||||||||||||||||  |||||  ||||||  ||||||||||  |||
Sbjct  121  AGGTTCTTTGACTCCTTTGGGGATCTGTCCAATCCTGGTGCTGTGATGGGCAACCCCAAG  180

Query  231  GTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC  290
            ||||||||  ||  ||||||||||||||     |||||  ||||  |||  ||  |||  ||  |||
Sbjct  181  GTGAAGGCCCACGGCAAGAAAGTGCTACACTCCTTTGGTGAGGGCGTGCATCATCTTGAC  240

Query  291  AACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGAT  350
            ||||||||||||||||||||  | ||||||||||||||||||||||||||||||||||||
Sbjct  241  AACCTCAAGGGCACCTTTGCTGCGCTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGAT  300

Query  351  CCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGC  410
            ||||||||||||||||||||||||||||||||||   |||||||||  |  |||||||||
Sbjct  301  CCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTTGTTGTGCTGGCTCGCCACTTTGGC  360

Query  411  AAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT  470
            ||  ||  |||||||||    ||||||||  |||||||  ||  ||||||||||||||||  |||
Sbjct  361  AAGGATTTCACCCCAGAGTTGCAGGCTTCCTATCAAAAGGTGGTGGCTGGTGTGGCCAAT  420

Query  471  GCCCTGGCCCACAAGTATCACT  492
            ||  ||||||||||  ||  ||||
Sbjct  421  GCACTGGCCCACAAATACCACT  442
```

Figure 3.16 BLASTN alignment between human (NM_000518) and horse beta hemoglobin (NM_001164018). Note that the alignment starts at the "ATG" of the coding region (underlined), base number one of the horse sequence. There is no horse 5P untranslated region.

being 76% between beta and gamma A. The simplest explanation for these very strong identities is that they all share a common ancestor. Some time in the distant past, there was a first hemoglobin gene. Then, through gene duplication events, other members arose, diverged in sequence, and became specialized in function. We can find them today as we did in the above BLASTN search. This stands as a model of what has happened throughout evolution; one gene gave rise to other family members.

Gene family members within the same organism are referred to as **paralogs**. Paralogs share a common ancestor and reside in the same genome. They are clearly related to each other but are usually specialized and have different functions. Beta hemoglobin is expressed in adults, while gamma A hemoglobin protein is only found in the fetus. This specialization reflects the distinctly different oxygen-binding needs between an air-breathing adult and a fetus growing in a womb and getting oxygen from its mother's blood. This is further illustrated by the human diseases called thalassemias, where a globin gene does not function and the other globin family members cannot adequately substitute for the lost function.

As seen in the last BLASTN search, many other animals also have hemoglobins. Genes that perform identical functions in different organisms are called **orthologs**. The human beta hemoglobin gene is orthologous to the horse beta hemoglobin gene. Both are expressed in adult animals. Did they evolve

independently? No. The simplest explanation is that there was a common animal ancestor who had the first beta hemoglobin gene, and the evolutionary descendants of that ancestor inherited this gene.

Homolog is a term that describes both paralogs and orthologs. When comparing genes between organisms, and it is not clear if they are orthologous, then the genes are described as homologs. When describing genes that show some identity but it is not clear if they are family members, then it is safer to describe them as sharing homology. The human alpha hemoglobin and mouse beta hemoglobin clearly have a common ancestor but perform different functions. They are homologs, not orthologs, and certainly not paralogs.

3.6 BLAST OUTPUT FORMAT

The output of BLAST described above is the HTML or Web format of the results. This format allows easy navigation between the graphic, table, and alignments, as well as instant access to sequence files through hypertext. The NCBI and other Websites provide this to you because of these obvious advantages. However, you may encounter Websites or an instance of BLAST you run from a command-line interface like UNIX where the output is raw text. In this case, the results may look like **Figure 3.17**. The advantage of this format is the simplicity; copying from this output, or parsing using a simple programming script, is uncomplicated by hidden formatting. The NCBI gives you the option to output your results as "plain text" by clicking on "Formatting options" near the top of the results page. In fact, if you are copying your own BLAST results and pasting them into reports, you may wish to use this format. This book will often show you the raw text format for simplicity. Note that in this raw form, the only columns after the description are the Score and E Value.

3.7 SUMMARY

In this chapter, the focus was BLASTN, a Web application that allows you to search nucleotide databases with nucleotide sequence queries. You learned how to paste a sequence into the query window, select a database to search, sometimes narrow your search to sequences of a certain species, and then launch the search. When the results came back you were able to quickly review the search by looking at the graphic and the table. You could look at the graphic to get a visual idea of how your query lined up to the hits, and sometimes saw that the query found smaller sequences. In the BLASTN results table, you saw the hits along with the statistics relating to them. Besides obvious criteria such as percent identity, there was the E value that showed you the probability that hits were found by chance. An E value of 0.0 meant that these hits were very real and could not be explained by random occurrences. Finally, you looked at the alignments where you could see, base by base, how your query lined up with the hits.

EXERCISES

Exercise 1: Biofilm analysis

Public water supply lines are immersed in water for decades and a community of microorganisms thrives on these wet surfaces. These slippery coatings are referred to as biofilms and the bacterial makeup is generally unknown because scientists are unable to culture and study the vast majority of these organisms in the laboratory. In 2003, Schmeisser and colleagues published a study where they collected and sequenced the DNA from bacteria growing on pipe valves of a drinking water network in Northern Germany. Through sequence similarity, they were able to classify a large number of these organisms as belonging to certain species or groups. In this process they identified many new species. In this

(A)

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

ref|NM_000518.4|  Homo sapiens hemoglobin, beta (HBB), mRNA         1034     0.0
ref|NM_000519.3|  Homo sapiens hemoglobin, delta (HBD), mRNA         654     0.0
ref|NM_005330.3|  Homo sapiens hemoglobin, epsilon 1 (HBE1), mRNA    381     8e-105
ref|NM_000184.2|  Homo sapiens hemoglobin, gamma G (HBG2), mRNA      343     2e-93
ref|NM_000559.2|  Homo sapiens hemoglobin, gamma A (HBG1), mRNA      334     1e-90
ref|XM_002344540.1|  PREDICTED: Homo sapiens similar to PRO298...    241     2e-62
ref|XM_002347218.1|  PREDICTED: Homo sapiens similar to PRO298...    241     2e-62
ref|XM_002343046.1|  PREDICTED: Homo sapiens similar to PRO298...    241     2e-62
ref|NR_001589.1|  Homo sapiens hemoglobin, beta pseudogene 1 (...    233     2e-60
ref|NM_001128602.1|  Homo sapiens RAS guanyl releasing protein...   35.6     1.3
ref|NM_005739.3|  Homo sapiens RAS guanyl releasing protein 1 ...   35.6     1.3
ref|NM_080723.4|  Homo sapiens neurensin 1 (NRSN1), mRNA            35.6     1.3
ref|NM_016642.2|  Homo sapiens spectrin, beta, non-erythrocyti...   35.6     1.3
ref|NM_000794.3|  Homo sapiens dopamine receptor D1 (DRD1), mRNA    35.6     1.3
ref|NM_199077.1|  Homo sapiens cyclin M2 (CNNM2), transcript v...   35.6     1.3
ref|NM_199076.1|  Homo sapiens cyclin M2 (CNNM2), transcript v...   35.6     1.3
ref|NM_017649.3|  Homo sapiens cyclin M2 (CNNM2), transcript v...   35.6     1.3
ref|NM_144666.2|  Homo sapiens dynein heavy chain domain 1 (DN...   33.7     4.5
ref|NM_021020.2|  Homo sapiens leucine zipper, putative tumor ...   33.7     4.5
ref|NM_000798.4|  Homo sapiens dopamine receptor D5 (DRD5), mRNA    33.7     4.5
ref|NM_015221.2|  Homo sapiens dynamin binding protein (DNMBP)...   33.7     4.5
ref|NM_080539.3|  Homo sapiens collagen-like tail subunit (sin...   33.7     4.5
ref|NM_182515.2|  Homo sapiens zinc finger protein 714 (ZNF714...   33.7     4.5
ref|NM_080538.2|  Homo sapiens collagen-like tail subunit (sin...   33.7     4.5
ref|NM_005677.3|  Homo sapiens collagen-like tail subunit (sin...   33.7     4.5
ref|NM_016315.2|  Homo sapiens GULP, engulfment adaptor PTB do...   33.7     4.5
ref|NM_024686.4|  Homo sapiens tubulin tyrosine ligase-like fa...   33.7     4.5
ref|NM_015540.2|  Homo sapiens RNA polymerase II associated pr...   33.7     4.5
ref|NM_032444.2|  Homo sapiens BTB (POZ) domain containing 12 ...   33.7     4.5
ref|NM_014234.3|  Homo sapiens hydroxysteroid (17-beta) dehydr...   33.7     4.5
ref|NM_148414.1|  Homo sapiens ataxin 2-like (ATXN2L), transcr...   33.7     4.5
ref|NM_007245.2|  Homo sapiens ataxin 2-like (ATXN2L), transcr...   33.7     4.5
ref|NM_145714.1|  Homo sapiens ataxin 2-like (ATXN2L), transcr...   33.7     4.5
ref|NM_148415.1|  Homo sapiens ataxin 2-like (ATXN2L), transcr...   33.7     4.5
ref|NM_148416.1|  Homo sapiens ataxin 2-like (ATXN2L), transcr...   33.7     4.5
ref|NM_016261.2|  Homo sapiens tubulin, delta 1 (TUBD1), mRNA       33.7     4.5
ref|NM_000911.3|  Homo sapiens opioid receptor, delta 1 (OPRD1...   33.7     4.5
ref|NM_007261.2|  Homo sapiens CD300a molecule (CD300A), mRNA      33.7     4.5
ref|NM_020857.2|  Homo sapiens vacuolar protein sorting 18 hom...   33.7     4.5
```

(B)

```
>ref|NM_000518.4| Homo sapiens hemoglobin, beta (HBB), mRNA
Length=626

 Score = 1034 bits (1146),  Expect = 0.0
 Identities = 573/573 (100%), Gaps = 0/573 (0%)
 Strand=Plus/Plus

Query  1    GTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  54   GTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTG  113

Query  61   GATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGG  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  114  GATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGG  173

Query  121  TTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  174  TTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTG  233
```

Figure 3.17 The "plain text" format of NCBI BLAST results. (A) The BLASTN results table. (B) The top of a BLASTN alignment.

exercise, you are to use BLASTN to repeat some of their analysis and identify the makeup of these biofilms.

Below is a list of 10 sequence accession numbers from their study. You are to use the NCBI BLASTN Web form to search for sequence similarities to try to identify the bacteria growing within these biofilms.

AY187314

AY187315

AY187316

AY187317

AY187318

AY187325

AY187326

AY187330

AY187332

AY187333

1. Retrieve each sequence from the NCBI GenBank and, based on the annotation of these sequence records, identify what gene was used in their analysis.
2. For each sequence, convert the file format to FASTA using the "Display Settings."
3. Navigate to the NCBI BLASTN Web form and paste the FASTA format of each DNA sequence into the Query window.
4. Choose the "Nucleotide collection (nr/nt)" as the database to be searched.
5. To save lots of time for your searches, restrict your search to "bacteria (taxid:2)" in the Organism field.
6. Pick "Somewhat similar sequences (BLASTN)" as the program to be used in the search.
7. When ready, launch the search by clicking on the "BLAST" button.
8. Open up additional Internet browser windows and launch the other searches.
9. Ten individual windows of results will be returned within a few minutes. Be sure to stay organized and record your conclusions for each accession number.
10. For each BLASTN search, survey the results graphic, table, and alignments to assign each unknown sequence to an organism. You may not find 100% identity between your query and the hits, except for the self-hit. Note that the first hit may also be an unknown so you should examine all the hits before drawing any conclusions as to what kind of bacteria the sequence came from.
11. Using the NCBI PubMed database or other Internet resources, try to find basic information about the genus and/or species; for example, habitats where these bacteria grow, and if they are associated with any diseases or environmental pollutants.

Exercise 2: RuBisCO

It is often said that ribulose bisphosphate carboxylase (RuBisCO) is the most abundant protein on the planet. This enzyme is part of the Calvin cycle and is the key enzyme in the incorporation of carbon from carbon dioxide into living organisms. It is part of an enzyme complex found in plants, terrestrial or aquatic, and most probably played an important role in the development of our atmosphere and life on Earth.

Arabidopsis thaliana, a member of the mustard family, is an important model system for higher plants. It is easily cultivated in the laboratory, undergoes rapid development, and produces a large number of seeds, making it amenable to genetic studies. Although not important agronomically, *Arabidopsis* has provided fundamental knowledge of plant biology and it was the first plant genome to be sequenced (in 2000).

In this exercise, you will use BLASTN to identify members of the RuBisCO gene family in *Arabidopsis*.

1. Retrieve the reference mRNA for the *Arabidopsis* RuBisCO small chain subunit 1b, NM_123204, at the NCBI Website.
2. Change the format to FASTA and paste the sequence into the NCBI BLASTN Web form Query window.
3. Set the database to "Reference RNA sequences (refseq_rna)" and restrict the organism to "Arabidopsis thaliana (taxid:3702)."
4. Set the program selection to "Somewhat similar sequences (BLASTN)" and click on the "BLAST" button to launch the search.
5. When the results are returned, you should now utilize the graphic, table, and alignments to identify the family members.
6. The Reference RNA database should not have any redundancy but two family members have alternatively spliced mRNAs. Compare the alignments carefully and examine the annotation (especially the coordinates of the coding regions) of all the relevant sequence records to describe and understand the major differences between these family member transcripts.
7. Create a table with a listing of the names of family member transcripts and their accession numbers, their mRNA length, the coordinates of the coding regions (CDS), and a brief description of what is observed in the alignments.

FURTHER READING

Altschul SF, Gish W, Miller W et al. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. The first BLAST paper, a classic in the field of bioinformatics.

Altschul SF, Madden TL, Schäffer AA et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. This is another foundation paper by the creators of the BLAST suite of programs, describing significant improvements to the algorithm.

Schmeisser C, Stöckigt C, Raasch C et al. (2003) Metagenome survey of biofilms in drinking-water networks. *Appl. Environ. Microbiol.* 69, 7298–7309. This paper is relevant to the biofilms exercise.

Internet resources

To learn more about *Arabidopsis* and the *Arabidopsis* genome sequencing project, go to www.arabidopsis.org and click on their Education and Outreach portal. This is an extensive Website with many resources ranging from fundamental learning about *Arabidopsis* to detailed workings of plant research and how the genome was sequenced.

To learn more about the globin family of genes, go to the NCBI Bookshelf of electronic books, choose "Human Molecular Genetics" from the extensive list of textbooks, and in the search window for this text, enter "globin." This gene was discussed or illustrated 74 times in this book, within a variety of topics. For example, there is a figure showing the "Evolution of the globin superfamily." Due to their history, biology, biochemistry, genetics, diseases, size, and genomic structure, globin genes are frequent subjects in textbooks.