7

Computational Analysis of DNA and Protein Sequences

ANDREAS D. BAXEVANIS, MARK S. BOGUSKI, AND B.F. FRANCIS OUELLETTE

Computer databases, networks, and software tools are essential resources for all aspects of genome analysis. The popular but imprecise term bioinformatics is used to describe a spectrum of methods and activities from laboratory information management systems through data analysis, interpretation, and integration; document preparation; and electronic publishing as well as submitting sequence and mapping data to public databases (Boguski 1994). This chapter touches on all of these topics, but it focuses on the analysis and annotation of DNA and protein sequence data.

The genomic sequences of several prokaryotic genomes and eukaryotic chromosomes have been completed. In addition, EST surveys continue to yield comprehensive sets of transcripts for humans and many model organisms such as the mouse. (ESTs are partial gene sequence data from cDNA clones. They provide sequence tags for genes, but the error rate can be high, perhaps approaching 5%, since there is usually only a single pass of sequencing.) Because of these developments, biologists will need to have more than a passing acquaintance with sequence analysis and annotation methods. Instead of reviewing the many excellent commercial sequence analysis programs available, this chapter concentrates on selected databases and software tools that are freely available in the public domain.

Since bioinformatics is a field in flux, with new techniques continuously being developed, the Internet addresses or Web sites listed in this chapter will no doubt change. Given this, the publishers have decided to make this chapter available in electronic format over the World Wide Web at:

`http://clio.cshl.org/books/g_a/bk1ch7`

If, in the course of using the printed version of this chapter, an invalid address or site is encountered, please refer to the electronic version for updated information.

INTERNET BASICS

Printed journals used to be the only "database and access system" that biologists had available to obtain published information. Although this information resource is still of utmost value, genome researchers must now be able to use electronic data stored in vast repositories spread out all over the world. This section therefore provides an introduction to basic Internet concepts and navigation tools and some pointers regarding where to look for relevant data and services. Selected information retrieval systems are discussed at length on pp. 560–573, and the submission of new or updated data to public data repositories is discussed on pp. 579–584. Be advised, however, that as the Internet expands, it also changes. In fact, the average half-life of an address on the Internet is about 4 years, and therefore some of the resources and addresses cited in this chapter may no longer be available or located at the same place they were at the time of this writing. All of the software discussed here can be found on the Internet free of charge, and most of it is available for use on Macintosh, PC, and UNIX systems.

More detailed descriptions of and tutorials on the Internet may be found in a wide variety of publications available in many bookstores. Two useful general reference books are *The Whole Internet User's Guide & Catalog, 2nd ed.* by Ed Krol and *The Internet Roadmap* by Bennett Falk. These and other similar titles provide more in-depth treatment of Internet issues such as communications protocols, networked applications, and connectivity.

Electronic Mail

The most familiar form of electronic communication to biologists is undoubtedly E-mail. The popularity of E-mail lies in its convenience in sending, receiving, and replying to messages. Communication tends to be direct and to the point, and the user has the ability to assess whether a message requires an immediate reply (or any reply at all). Additional advantages of E-mail are its obvious speed over traditional postal mail, or "snail mail," its ability to save and forward copies of messages in an orderly way, and the fact that it is a low-cost or no-cost alternative to postal mail. The major disadvantage of E-mail lies in security in that as the mail passes from node to node on its way to the intended recipient, there is the possibility that the message could be read or intercepted by a systems administrator or by someone else with similar access. In an academic environment, this is more than likely a nonissue, but in a corporate environment, E-mail systems may be treated as an asset of the company and be subject to monitoring. The advantages of E-mail far outweigh the disadvantages, and these advantages are what have made E-mail one of the primary forms of communication within the academic community.

In addition to its usefulness in sending messages to a single individual, E-mail can be used to communicate with large numbers of people all at the same time

through what is termed newsgroups. The members of these newsgroups are able to obtain information and exchange ideas on an extremely wide number of topics of interest. Subscribing to these groups is as simple as sending an E-mail message containing the word *subscribe* to a given E-mail address. The BIOSCI newsgroups are among the most highly subscribed to forums for biologists. Information on subscribing to individual BIOSCI newsgroups can be obtained by sending an E-mail message to:

`biosci-server@net.bio.net`

leaving the subject blank and placing the words *info faq* in the body of the message. A copy of frequently asked questions (FAQ) will then be returned in response.

Finally, E-mail can be used to perform computations, make predictions, or do database searches. By sending a formatted E-mail message to a remote computer (a server), users can ask the remote computer to perform a defined operation and return the result, again via E-mail. One advantage of this type of system is that it frees the user from both developing and maintaining software, since those functions are performed by the persons who maintain the server. Disadvantages include the lack of real-time interactivity and the limitation to strictly text-based output. Table 1 presents a list of E-mail servers of particular value to molecular biologists. In addition, an up-to-date list of E-mail servers is maintained by A. Bairoch of the University of Geneva; this list can be obtained via anonymous ftp (described below) at:

`expasy.hcuge.ch` (directory `/databases/info`, file `serv_ema.txt`)

With few exceptions, sending the message *help* to any E-mail server will return a detailed set of instructions for using that server. Practical examples on how to use E-mail to search for homology are provided on pp. 543–559.

File Transfer Protocol

E-mail provides an excellent mechanism for transmitting messages, but it is limited in its ability to transfer files. Even though most commercial mail packages have the ability to send files as an attachment to a message, it is not uncommon for the attached file to be unusable by the recipient: The file may be too big to be transferred or it may be in a format that is unrecognizable by the recipient's computer.

A simpler and more efficient method of transferring files is through an ftp. When an ftp is used, a connection is made between the user's computer and the remote computer, a connection that stays in place for the duration of the file transfer session. To make this connection, the user must have both an account and password on the remote computer.

Alternatively, users can perform what is termed anonymous ftp, which does not require a username or a password. This method is most often used in making public domain software freely available and accessible. Announcements of newly available software obtainable through this mechanism are often made

Name	Service	Server Address
BLAST	BLAST protein and DNA sequence homology searches	*blast@ncbi.nlm.nih.gov*
BLITZ	Protein sequence homology searches using MPsrch algorithm	*blitz@ebi.ac.uk*
BLOCKS	Protein motif searches against the database of protein blocks	*blocks@howard.fhcrc.org*
EBI File Server	Obtain EMBL and SWISS-PROT entries, public domain software	*netserv@ebi.ac.uk*
FASTA	FASTA protein and DNA sequence homology searches	*service@bchs.uh.edu*
GRAIL	Protein coding region determination using a neural network algorithm	*grail@ornl.gov*
HUGEMAP	Genethon human physical map data on clones, YACs, STSs	*hugemap@genethon.fr*
nnpredict	Protein secondary structure prediction using neural network algorithm	*nnpredict@celeste.ucsf.edu*
PredictProtein	Secondary structure prediction using profile network method (PHD)	*predictprotein@embl-heidelberg.de*
Query	Entrez query engine for protein, nucleotide, and MEDLINE records	*query@ncbi.nlm.nih.gov*
RETRIEVE	Retrieve DNA and protein sequences from most public databases	*retrieve@ncbi.nlm.nih.gov*
RTFM	Information and frequently-asked questions for FTP and UseNet	*mail-server@rtfm.mit.edu*

Table 1 Selected E-mail Servers for Molecular Biology. For most servers, sending a message to the server address with only the word *help* in the body of the message will return complete instructions on how to use that server. A comprehensive list of E-mail servers can be obtained via anonymous ftp at:

`expasy.hcuge.ch` (directory `/databases/info`, file `serve_ema.txt`)

or via the World Wide Web at:

`http://expasy.hcuge.ch/info/serv_ema.txt`

by E-mail newsgroups discussed above. With this method, users, when prompted for a username, reply with the word *anonymous* to signify that an anonymous ftp session is being requested. When prompted for a password, users then provide their E-mail address. Supplying the E-mail address allows the systems administrator at the remote site to maintain relevant access statistics, which is of use to those providing the public domain software. Once granted access, users can navigate through the public directories and download any software of interest.

Although ftp takes place in the context of a UNIX environment, programs for performing ftp are available with a graphical end-use interface that allow users to move through directories by pointing and clicking their computer mouse, in

Database Servers		
FTP Server	**Major Databases Available**	**FTP Server Address**
NCBI	GenBank, SWISS-PROT, PIR	`ncbi.nlm.nih.gov`
EBI	EMBL, SWISS-PROT	`ftp.ebi.ac.uk`
ExPASy	Enzyme, EPD, Prosite, SeqanalRef, SWISS-PROT, SWISS-2DPAGE, SWISS-3DIMAGE	`expasy.hcuge.ch`

Software Servers		
FTP Server	**Software Available**	**FTP Server Address**
NCBI	BLAST, Sequin, GenInfo Software Toolbox, MACAW	`ncbi.nlm.nih.gov`
EBI	Mac, VAX, DOS, UNIX molecular biology software	`ftp.ebi.ac.uk`
IuBio	Mac, VAX, DOS, Atari software; PHYLIP phylogeny inference package, ReadSeq alignment package	`ftp.bio.indiana.edu`

Table 2 Selected ftp Servers for Molecular Biology. A comprehensive list of ftp servers can be obtained via anonymous ftp at:

`expasy.hcuge/ch` (directory/databases/info, file `serv_ftp.txt`)

or via the World Wide Web at:

`http://expasy.hcuge.ch/info/serv_ftp.txt`

stead of by issuing UNIX commands. The most popular program with a graphical user interface for the Macintosh is called Fetch, and similar programs are also available for the PC. If a user knows the name of some public domain software of interest but does not have information on where it can be downloaded from E-mail, a search engine named ARCHIE

`archie@archie.rutgers.edu`

can be used to locate ftp sites containing that file. Related programs (XARCHIE for UNIX and ANARCHIE for Macintosh) can both perform the search and download the file in a single operation. Table 2 presents a selected list of relevant ftp sites. An example of how to download a file from an ftp site can be obtained by sending an E-mail message to:

`info@sunsite.unc.edu`

with the word *help* in the body of the message.

Gopher

Although an ftp allows documents or programs to be easily disseminated, it requires that a user actually download a file to examine its contents. One of the first-generation Internet tools, Gopher, addressed the need to distribute text documents to users without requiring an actual download. Gopher was devel-

Gopher Server	Gopher Server Address
Baylor Genome Center	gopher://gc.bcm.tmc.edu
Johns Hopkins Computational Biology	gopher://gopher.gdb.org
Genethon	gopher://gopher.genethon.fr
IUBio	gopher://ftp.bio.indiana.edu
Protein Data Bank	gopher://pdb.pdb.bnl.gov
Jackson Laboratory	gopher://hobbes.jax.org

Table 3 Selected Gopher Servers for Molecular Biology

oped at the University of Minnesota (hence its name, after the school mascot) and is a good example of what is termed a "distributed document delivery system." Gopher also falls into the client-server class of applications, since its use requires connection to a remote computer and is interactive.

One of the key features of Gopher is that it provides relatively effortless travel around the Internet. The information stored at Gopher sites is organized in a series of hierarchical menus, and movement through these menus is accomplished either by using arrow keys or by clicking a mouse. Movement through this hierarchy is not restricted to a single site; users can traverse the Internet, visiting other Gopher sites that are interconnected through "Gopher holes." This is one of Gopher's strengths, since users do not need to know the exact location of the information and they can sequentially follow the hierarchy until the desired information is found.

A short list of relevant Gopher sites is provided in Table 3. An extensive manual on Gopher developed by the University of Minnesota is available via ftp at:

`boombox.micro.umn.edu` (directory/pub/gopher/docs/)

The World Wide Web

The logical next step in the development of Internet tools was to provide an interface through which information could be accessed directly and to provide non-text-based media, such as images, sounds, and video. This need resulted in the development of the World Wide Web by CERN.

The programs used to traverse the Web are client-server applications, as are Gopher programs, but the similarities between Gopher and the Web in the overall presentation of information end soon thereafter. The information distributed on the Web is not strictly text-based. The Web can display images, produce sounds, or play back video and other types of animation. Navigation on the Web is accomplished by clicking the computer mouse on hyperlinks (specific text, buttons, or pictures within a document). These hyperlinks can transport the user to another Web location at the same site or across the globe. Locations on

the Web are called Web sites, and the individual files that are stored and displayed at these sites are called Web pages. A process nicknamed "Web surfing" allows users to follow hyperlinks from page to page until the information of interest is found. The links are not organized in a rigid hierarchy, which is different from Gopher.

In addition, users can access a specific site directly by typing in its Web address. Web addresses are also called URLs. The "uniform" part of URL refers to the fact that the software used to look at documents on the Web (browsers) is capable of visiting not only Web sites, but also Gopher and ftp sites by specifying the appropriate protocol. To accomplish this, a uniform method for specifying sites was introduced (the URL) that would indicate to the Web browser both the address of the remote site and the type of site it is. URLs take the general form `protocol://somewhere.domain`, where `protocol` specifies the type of site and `somewhere.domain` specifies the remote location. The following are examples of URLs.

ftp site	`ftp://ftp.bio.indiana.edu`
Gopher site	`gopher://hobbes.jax.org`
Web site	`http://www.ncbi.nlm.nih.gov`

The `http://` in the Web address is the protocol used to transfer Web files from the server to the client.

As mentioned above, browsers are used to look at documents on the Web. These browsers interpret the code underlying a Web page (the HTML) and display it in the correct format, regardless of whether the user is on a Macintosh, PC, or UNIX system. By far, the most widely used browser software is Netscape Navigator, with estimates by two different market research firms placing Netscape Navigator's share of the Web browser market at 75–85%. Netscape Navigator became the de facto standard by offering users an interface that displayed Web pages much faster than previously available products, an important factor for users using either dial-up connections or commercial Internet providers. Netscape Navigator also provides special features that enable Web developers to take advantage of HTML enhancements not available in other browsers. Other browsers available include the America Online browser and Microsoft's Internet Explorer, both of which are estimated to have about a 10% market share based on the surveys above.

Although the most prevalent way of finding information on the Web is by verbal or printed sources (e.g., the tables that accompany this text), users may consult compiled lists of Web sites, known as virtual libraries, to locate Web sites that are most likely to have the information being seached for. Three popular virtual libraries of interest to biologists are the *WWW Virtual Library* maintained by Keith Robison at Harvard; *Pedro's Biomolecular Research Tools*, compiled by Pedro Coutinho at Iowa State; and the *EBI BioCatalog*, a collaborative project based at the European Bioinformatics Institute. Table 4 provides the addresses for these and other useful Web sites.

A Web surfer can also find Web sites of interest by using special programs called search engines. Search engines use a variety of methods to perform either

Compendia of WWW Resources	
BIOSCI Newsgroups	http://www.bio.net/
EBI	http://www.ebi.ac.uk/
Pedro's Biomolecular Research Tools	http://www.public.iastate.edu/~pedro/research_tools.html
WWW Virtual Library	http://golgi.harvard.edu/htbin/biopages
Sequence Retrieval and Analysis	
ExPASy Molecular Biology Server	http://expasy.hcuge.ch/
NCBI	http://www.ncbi.nlm.nih.gov
NCBI BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
NCBI Entrez	http://www3.ncbi.nlm.nih.gov/Entrez/
PDB	http://www.pdb.bnl.gov
Organism-Specific Web Resources	
Arabidopsis (AtDB)	http://genome-www.stanford.edu/Arabidopsis/
C. elegans	http://eatworms.swmed.edu/
Flybase	http://morgan.harvard.edu/
Mouse Genome Database	http://www.informatics.jax.org/mgd.html
Saccharomyces (SacchDB)	http://genome-www.stanford.edu/Saccharomyces/
Electronic Journals	
Cell	http://www.cell.com/
Genome Research	http://www.cshl.org:80/journals/gr/
Journal of Biological Chemistry	http://www-jbc.stanford.edu/jbc/
Journal of Molecular Biology	http://www.hbuk.co.uk/jmb
Nature	http://www.nature.com/
Science	http://science-mag.aaas.org/science/home/
Pedro's List of Bio/Chemical Journals and Newsletters	http://www.public.iastate.edu/~pedro/rt_journals.html

Table 4 Selected World Wide Web Servers for Molecular Biology

keyword or full-text searches across the Web, returning a hyperlinked list of results that the user can then scan and click on to visit any or all of the found sites. Since each search engine uses a different method to search the Web, and

in some cases to search only subsets of the Web, the resulting hit lists can vary tremendously. Below are two searches performed using three different search engines.

search engine	human genome	positional cloning
Web Crawler	752	16
Infoseek	1252	44
Inktomi	18713	794

These results should not be interpreted as "more is better," since a single page could conceivably produce multiple hits if the phrase appears more than once, or the search parameters may be loose in the sense that the algorithm allows the words to be slightly separated within the same document. Instead of performing search after search using each different engine, the user can perform searches with meta-search engines, which have been developed to poll different search engines automatically, collect the results, filter out any duplicates, and return a single hit list. Although these searches necessarily take longer to perform, a user can have confidence in having found most, if not all, of the sites that would fit a given search query. Two such meta-search engines are

SavvySearch
`http://guaraldi.cs.colostate.edu:2000/form`

MetaCrawler
`http://metacrawler.cs.washington.edu:8080/`

Electronic Publishing

Given the increasing reliance of scientists on the Internet for communication and the changing nature of experimental data, many scientific journals are now establishing a World Wide Web presence. The advantages of having traditional print journals available on the Web include the following:

- Information is disseminated much more quickly, with no lag time from the time of printing to the time the actual physical copy is received by a reader.
- Data sets too large to be accommodated within a journal article are provided as an electronic supplement, which can be updated as new information becomes available.
- Alternative means of presentation are now available, such as narration or animation.
- Full-text searches are possible, with hyperlinks to biological sequences and structural or genomic information.

Among the journals that have established Web sites are *Science*, *Nature*, *Cell*, and *The Journal of Biological Chemistry*. The content on these journal Web sites will vary. For example, *Cell* presents tables of contents and abstracts for all articles,

whereas *The Journal of Biological Chemistry* and *Science* present full text and figures for all published articles, as well as links to the relevant sequence and bibliographic databases. Table 4 provides URLs for these and other journal sites.

Specialized Client-Server Applications

Gopher and Netscape Navigator are excellent examples of Internet navigation tools that operate as client-server programs, but their universality may also be a limitation under certain conditions. More powerful client-server systems have been developed to take advantage of scientific knowledge and data interrelationships in specialized fields. One such system is Entrez, an integrated information retrieval system designed for seamless traversing of sequence, structure, genomic, and bibliographic databases. Entrez is discussed on pp. 561–566.

SEQUENCE ANALYSIS

Searching for homology is the process of comparing a new sequence with all other known sequences and then attempting to infer the function of the new sequence by assessing the matches and their biological annotations as described in the database itself and in the literature. In this section, sequence analysis is described from the point of view of positional cloners whose primary goal is to find the exons rapidly in a megabase expanse of genomic DNA; this usually involves a combination of exon trapping, cDNA selection, and genomic sequencing. The goals here are considerably different from those of investigators whose primary aim is to sequence a genome or large contigs thereof. In the former case, every single-pass sequence must be carefully examined for homologies and coding potential at the earliest possible time both to find genes and to prioritize mutation detection. In the latter case, detailed analysis is usually preceded by contig assembly and "finishing," with the goal of analysis being to construct complete gene models.

There are basically two types of sequence analysis: analysis by similarity (homology) and analysis by intrinsic sequence properties (Fickett 1994). These methodologies are often used in concert (Claverie 1994). Similarity analysis includes database searching and alignment (including multiple alignment), whereas intrinsic sequence analysis is broader in scope and ranges from predicting exons on the basis of statistical properties of sequence composition to ab initio predictions of protein structure. An important type of intrinsic sequence analysis is the segmentation of sequences by local compositional complexity (Wootton 1994a; Wootton and Federhen 1996), which is of great utility in increasing the signal-to-noise ratio in database searching and is essential for high-throughput analysis of sequence data. Another important application of sequence analysis is prediction of gene structure by programs such as GRAIL. These programs have recently been reviewed (Burset and Guigo 1996; Fickett 1996).

Searching Databases for Similarities

SEQUENCE DATABASES

There are a number of important issues in searching DNA and protein sequence databases (Altschul et al. 1994), but the most important is access to a comprehensive and up-to-date data repository. GenBank, the EMBL nucleotide sequence database, and DDBJ are three partners in a long-standing collaboration to collect and distribute all publicly available sequence data. Sites in Bethesda, Maryland (United States), Hinxton (United Kingdom), and Mishima (Japan) exchange new sequence data and updates over the Internet on a daily basis and make this information immediately available to the public by a variety of means, including E-mail, anonymous ftp, and the World Wide Web (Harper 1994).

Details on how to submit new or updated sequence information and annotation to the databases are provided on pp. 579–584. The major providers of protein sequence data are GenBank, EMBL (translations of coding sequences), DDBJ PIR International, and SWISS-PROT. More detailed information on all of these data sources may be found in the annual database issue of *Nucleic Acids Research.*

Operationally, there are two forms of sequence data. The first is the complete database record that contains the names of authors/submitters, literature citations, and biological annotations as well as the sequence itself and a table of sequence features such as locations of introns, exons, and start and stop codons. The second form of sequence data consists simply of an accession number and a short descriptive header followed by the sequence itself; this is the form usually used for rapid searches for similarities. Examples of such sequences in FASTA format are shown below. Accession numbers are unique identifiers for a particular sequence. They are assigned to data when first submitted to a database and should always be referred to in any description or publication concerning sequence data.

Since sequence data come from a variety of sources (including the United States and European Patent Offices), NCBI provides comprehensive, quasi-nonredundant data sets (designated nr) for both nucleotide and protein sequences. These data sets are updated daily and made available for searching for homology. Further details on the rationale and construction of nr have been described by Altschul et al. (1994). The component data sources making up nr are always reported in the search output (see below), and further details are available through the NCBI home page:

```
http://www.ncbi.nlm.nih.gov
```

Finally, it must be noted that sequence data entering the public databases are growing exponentially, with a doubling time of approximately 18 months. GenBank (release 95.0, June 15, 1996) contains more than 551 megabases in 835,000 sequences. NCBI's protein nr database currently contains 312,250 protein sequences.

THE BLAST FAMILY OF PROGRAMS: USES AND EXAMPLES

The second most important issue in searching databases is the computer program used to search the sequence database. A number of different search algorithms have been developed over the years (for further information, see Altschul et al. 1994; Schuler et al. 1994; and references therein). Only the BLAST programs, which offer a good combination of speed, sensitivity, flexibility, and statistical rigor, are discussed here. BLAST can be used via an E-mail server or as a network client-server application. Alternatively, two World Wide Web interfaces (basic and advanced) are available and are very easy to use. The detailed examples below illustrate use of the E-mail server, but the same analyses can also be performed by using any of the other BLAST implementations available. Since the databases are constantly being updated, the search results obtained by repeating these exercises may differ slightly from the output shown below.

Program	Query Sequence	Database Sequences	Comments
BLASTN	Nucleotide, both strands	Nucleotide	• Parameters optimized for speed, not sensitivity. • Not intended for finding distantly-related coding sequences. • Automatically checks for complementary strand of query. • Low-complexity filter option (dust)
BLASTX	Nucleotide, six-frame translation	Protein	• Very useful for preliminary data containing potential frameshift errors, *i.e.*, ESTs and other "single-pass" sequences. • 12 different genetic codes available. • 65 different scoring matrices available. • Low-complexity filter options with SEG or XNU algorithms (highly recommended).
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation	• Very useful for preliminary data containing potential frameshift errors, *i.e.*, ESTs and other "single-pass" sequences. • 12 different genetic codes available. • 65 different scoring matrices available. • Low-complexity filter options with SEG or XNU algorithms (highly recommended).
BLASTP	Protein	Protein	• 65 different scoring matrices available. • Low-complexity filter options.
TBLASTN	Protein	Nucleotide, six-frame translation	• Essential for searching protein queries against EST database. • Often useful for finding undocumented open reading frames or frameshift errors in database sequences. • Same genetic code and scoring matrix options as BLASTX. • Low-complexity filter options (highly recommended).

Table 5 The BLAST Family of Programs

The first step in setting up a database search is to select the most appropriate BLAST program (Table 5). There are five implementations of BLAST, three designed for nucleotide sequence queries (BLASTN, BLASTX, and TBLASTX) and two designed for protein sequence queries (BLASTP and TBLASTN). The former are used for the analysis of genomic sequence (including putative exons) and single-pass cDNA data, whereas the latter usually come into play when discrete gene products from the finished sequence have been identified. Once a BLAST program has been selected, an appropriate sequence database must also be selected. Table 6 presents a list of databases available for use with the BLAST programs.

Databases	Database Sequences	Used with Program	Comments
NR	Nucleotide or protein	BLASTN BLASTP BLASTX	The non-redundant dataset contains nucleic or protein sequences from all known data sources.
MONTH	Nucleotide or protein	BLASTN BLASTP BLASTX	A 'rolling-month' dataset which represents the new records introduced to the database in the last month. Very useful for periodical verification of the databases.
SWISSPROT	Protein	BLASTP BLASTX	The Swiss-Prot protein databases.
DBEST	Nucleotide	BLASTN TBLASTN TBLASTX	The database of Expressed Sequence Tags.
DBSTS	Nucleotide	BLASTN TBLASTN TBLASTX	The database of Sequenced Tag Sites.
PDB	Protein	BLASTP BLASTX	The Brookhaven Protein database
VECTOR	Nucleotide	BLASTN	A databse of vector sequences: a useful check to do before database submissions.
KABAT	Protein	BLASTP BLASTX	Proteins of Immunological Interest.
MITO	Nucleotide	BLASTN	A database of mitochondrial DNA sequences, also a useful check to do before database submissions.
ALU	Nucleotide	BLASTN	A database of repetitive DNA sequences, also a useful check to do before database submissions. (mostly mamalian sequences)
EPD	Nucleotide	BLASTN	Eukaroytic Promoter Database

Table 6 Databases for Use with the BLAST Family of Programs

DNA Query Sequences

BLASTN takes a nucleotide sequence (the query sequence) and its reverse complement and searches them against a nucleotide sequence database. BLASTN was designed for speed, not maximum sensitivity, and is not intended for finding distantly related coding sequences. BLASTX takes a nucleotide sequence, translates it in three forward reading frames and three reverse complement reading frames, and then compares the six translations against a protein sequence database. BLASTX is extremely useful for sensitive analysis of preliminary (single-pass) sequence data and is quite tolerant of sequencing errors (Gish and States 1993). BLASTN and BLASTX are used in concert for analyzing EST data, the products of exon trapping, and low-pass sequence data (sequences in which each base is read an average of one to two times).

Example 1a

To find the breast cancer susceptibility gene by exon trapping from clones spanning the *BRCA1* region on chromosome 17, the E-mail BLAST search request would be structured as follows:

```
mail blast@ncbi.nlm.nih.gov
Subject:

PROGRAM blastn
DATALIB nr
BEGIN
>putative exon from BRCA1 region
TCTGGAGTTGATCAAGGAACCTGTCTCCACAAAGTGTGACC
ACATATTTTGCAA
```

This message specifies that the BLAST program to be used is BLASTN and that the database to be searched is nr. The word BEGIN signals that the query sequence follows and that the query sequence is in FASTA format. Although the BLAST programs have a number of optional parameters, some of which are illustrated in the examples below, this example constitutes the minimum information needed to carry out a basic BLAST search.

The results of this search (Figure 1a) include a number of matches to human and mouse *BRCA1* sequences. (Pretend that this gene has not already been cloned and ignore these matches.) The sixth match is to a sequence described as Mouse mRNA for estrogen-responsive and shows a statistically significant p value (5.3e–05, which is 5.3×10^{-5}). The next two matches represent alignments to Sequence 1 from Patent WO 8907652 and to Mouse down regulatory protein, with p values of only 0.1 and 0.11, respectively. These p values are unimpressive, and it is difficult to assess the meanings of the actual sequence alignments. It now becomes important to repeat the search using BLASTX to translate the query sequence in all six reading frames to detect potential protein similarities.

Example 1b

```
mail blast@ncbi.nlm.nih.gov
Subject:

PROGRAM blastx
DATALIB nr
MATRIX pam40
BEGIN
>putative exon from BRCA1 region
TCTGGAGTTGATCAAGGAACCTGTCTCCACAAAGTGTGACC
ACATATTTTGCAA
```

In Example 1b above, an optional parameter specifying a difference scoring matrix has been included. The matrix statement indicates that the PAM40 matrix should be used instead of the default matrix BLOSUM62 (Henikoff and Henikoff 1993). In this case, the PAM40 matrix was chosen because it is more suited to the short ORFs that will be generated from the 54-nucleotide query sequence. More details and advice on choosing appropriate scoring matrices are given on p. 555. Note that the database name, nr, is the same for both nucleotide (BLASTN) and protein (BLASTX) database searches; since there are both nucleotide and protein versions of nr, the appropriate version of nr will be used automatically.

The hit list here is much longer than the one in Example 1a and includes some new and more significant matches (Figure 1b), underscoring the greater sensitivity of protein sequence searches. Among the statistically significant matches, both BLAST programs produced a match to rpt-1, described as a Mouse down regulatory protein (score of 82 and p value of 0.00020). Consistency (or

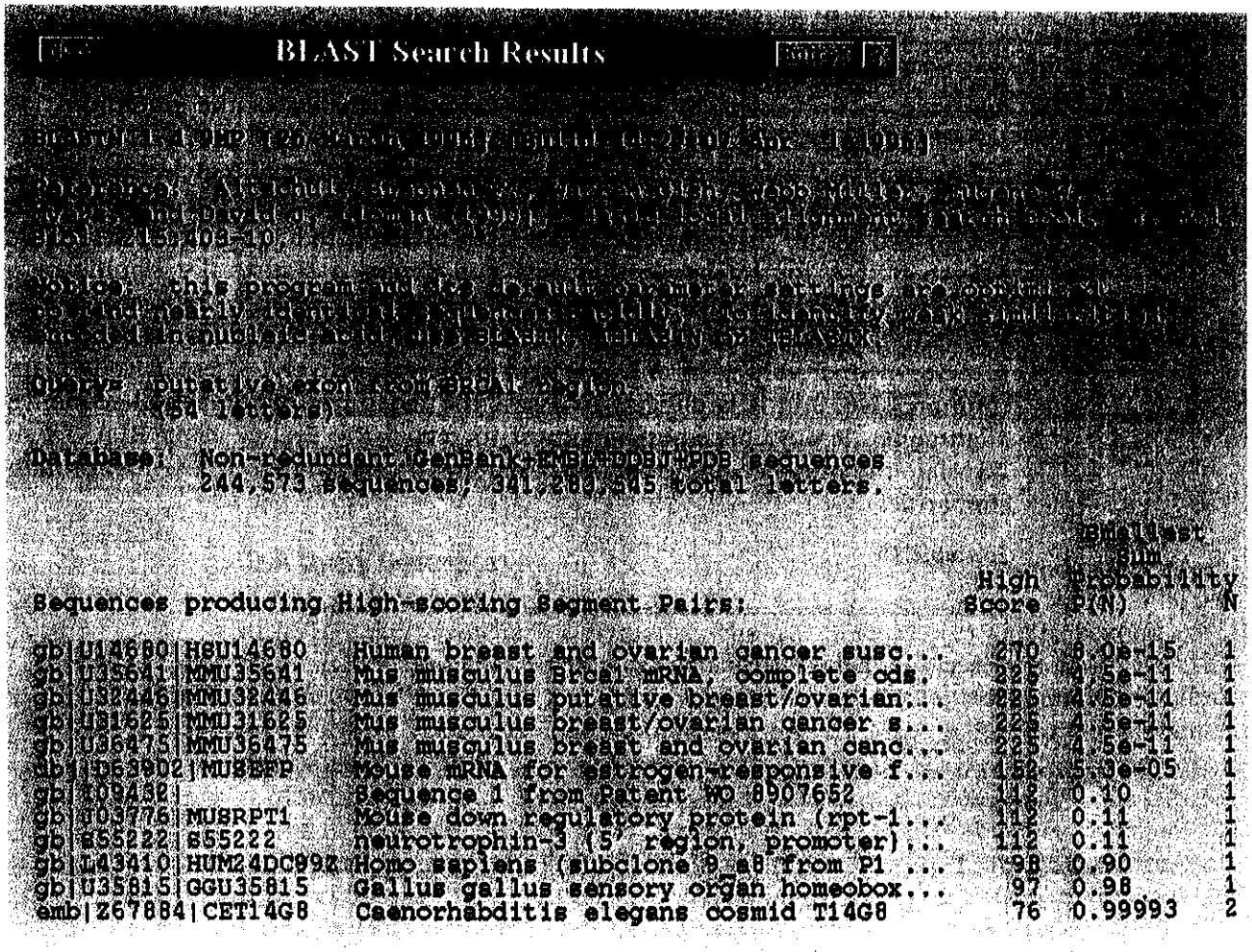


Figure 1a Results of a BLASTN query against nr using a putative exon from the *BRCA1* region as the query sequence.

lack thereof) is often a useful clue to the significance of database search results. In the BLASTN search (Example 1a), it is difficult to infer much from the nucleotide alignments. However, the alignments produced by the BLASTX search show an interesting pattern of conserved cysteine and histidine residues, possibly indicating some type of common metal-binding motif. Inspection of the SWISS-PROT database for this sequence indicates that it is a DNA-binding zinc finger protein. The putative exon in this example indeed turns out to be exon 3 of the *BRCA1* gene, encoding a portion of the zinc finger domain as described by Miki et al. (1994). On pp. 561–566, the Entrez system is used to assess the results of the database search rapidly by examining related sequences, map locations, all relevant literature, and even three-dimensional structures.

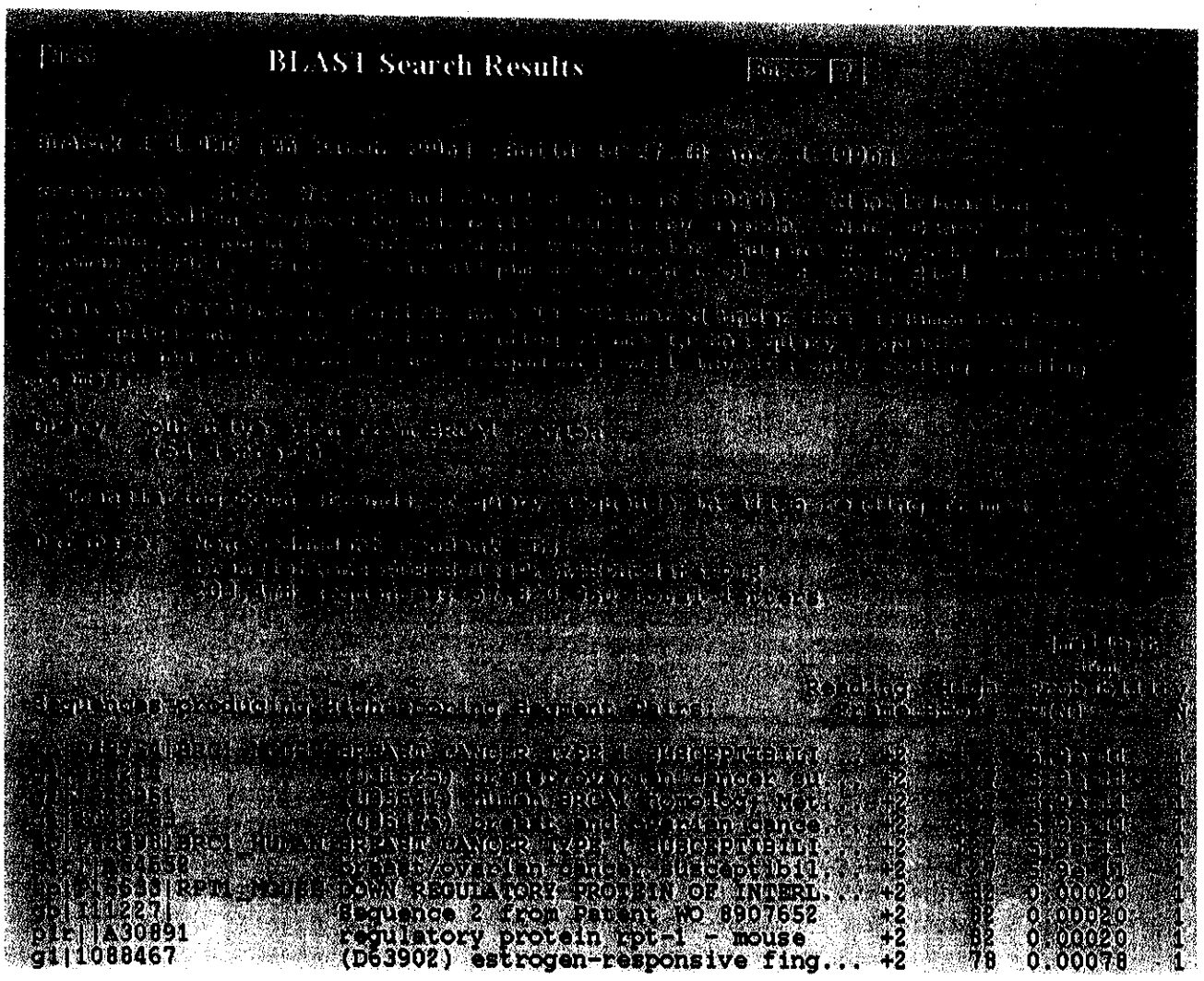


Figure 1b Results of a BLASTX query against nr using a putative exon from the *BRCA1* region as the query sequence. The PAM40 scoring matrix was used instead of the default BLOSUM62 matrix.

Example 2a

To illustrate another important aspect of database searching, another piece of genomic DNA isolated by exon trapping is analyzed.

```
mail blast@ncbi.nlm.nih.gov
Subject:

PROGRAM blastn
DATALIB nr
BEGIN
>another putative exon from BRCA1 region
GGTCTTACTCTGTTGTCCCAGCTGGAGTACAGTGGTGCGATCA
TGAGGCTTACTGTTGCCTTGACCTCCTAGGCTCAAGCGATCCT
ATCACCTCAGTCTCCCAAGTAGCTGGGACT
```

The results of this BLASTN search (Figure 2a) include many very interesting and statistically significant matches. Such a result seems too good to be true, especially since there are so very many apparently significant hits, yet there is no consistency at all among the matching genes, which include a serotonin transporter, a retinoblastoma susceptibility gene, and antithrombin III. Nevertheless, an appropriate follow-up would be to perform a BLASTX search.

Example 2b

```
mail blast@ncbi.nlm.nih.gov
Subject:

PROGRAM blastx
DATALIB swissprot
MATRIX pam40
BEGIN
>another putative exon from BRCA1 region
GGTCTTACTCTGTTGTCCCAGCTGGAGTACAGTGGTGCGATC
ATGAGGCTTACTGTTGCCTTGACCTCCTAGGCTCAAGCGATC
CTATCACCTCAGTCTCCCAAGTAGCTGGGACT
```

This time, the SWISS-PROT protein database is searched instead of nr; the results are presented in Figure 2b. It is obvious why so many matches were obtained in the previous search: The query sequence is a fragment of an *Alu* repetitive element, and what the BLASTN search produced was simply matches to other *Alu* elements that are present in thousands of sequences in the nucleotide databases. Although most genome researchers are aware of this pitfall in sequence analysis, many other biologists who search databases are not, and many mistaken or misleading results are still being published (and deposited in the databases) as a consequence (see, e.g., Tugendreich et al. 1994). Therefore, the curator of SWISS-PROT, A. Bairoch, has wisely included "dummy" entries (representing the translation of *Alu* family consensus sequences) in the database as warning flags.

The lesson here is that any DNA sequence should be checked for the presence of repetitive elements before any further analysis is performed. This is conveniently done by specifying DATALIB alu in the BLASTN search request. It is also imperative to check all sequences for residual vector-derived sequences that are often present on the ends of individual sequence readings and that sometimes go unrecognized with DATALIB vector. Such "filtering" procedures are built into the new PowerBLAST program (see p. 559). Query masking is an important general tool for sequence analysis and is discussed in greater detail on pp. 555–558.

Example 3

TBLASTX is the newest member of the BLAST family of programs. It performs six-frame translations of both the query and all database nucleotide sequences and then looks for matches at the protein level. In terms of computational resources, this is very expensive, so TBLASTX searches are restricted to queries on the EST division of GenBank. If the preceding BLASTN and BLASTX searches against nr found nothing (i.e., there are no statistically significant matches with sequences in the nucleotide or protein databases), one more procedure remains to be performed. Whereas most coding sequences in GenBank have their cognate proteins represented in the GenPept database (and thus in nr), there is one important exception: Because single-pass cDNA sequences (ESTs) seldom have coding sequence annotation and are subject to frameshift errors, conceptual translations (ORFs) from these sequences are not included in GenPept or nr and ESTs must therefore be translated by using programs such as TBLASTX or TBLASTN. It is always best to use TBLASTN if the protein sequence is available for use as a query. However, in the case of putative exons, single-pass genomic DNA, or cDNA, the query as well as the cDNA/EST database should be translated to maximize the chances of finding any significant but subtle (particularly cross-phylum) sequence homologies where there may be no similarity between nucleotide sequences at all.

Suppose that some cDNA clones that hybridize to a YAC believed to contain a gene of interest have been isolated and that some partial cDNA sequence data from one of the clones reveal no matches based on BLASTN or BLASTX searches, a TBLASTX search of the EST division of GenBank would be formulated as follows:

```
mail blast@ncbi.nlm.nih.gov
Subject:

PROGRAM tblastx
DATALIB dbest
BEGIN
>OCRL-selected mRNA, partial sequence
TTGAACATCATGAAACATGAGGTTGTCATTTGGTTGGGAGATTTGAATTATAGACTTTGC
ATGCCTGATGCCAATGAGGTGAAAAGTCTTATTAATAAGAAAGACCTTCAGAGACTCTTG
AAATTCGACCAGCTAAATATTCAGCGCACACAGAAAAAAGCTTTTGTTGACTTCAATGAA
GGGGAAATCAAGTTCATCCCCACTTATAAGTATGACTCTAA
```

NCBI **BLAST Search Results** Entrez ?

BLASTN 1.4.9MP [26-March-1996] [Build 14:27:07 Apr 1 1996]

Reference: Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-10.

Notice: this program and its default parameter settings are optimized to find nearly identical sequences rapidly. To identify weak similarities encoded in nucleic acid, use BLASTX, TBLASTN or TBLASTX.

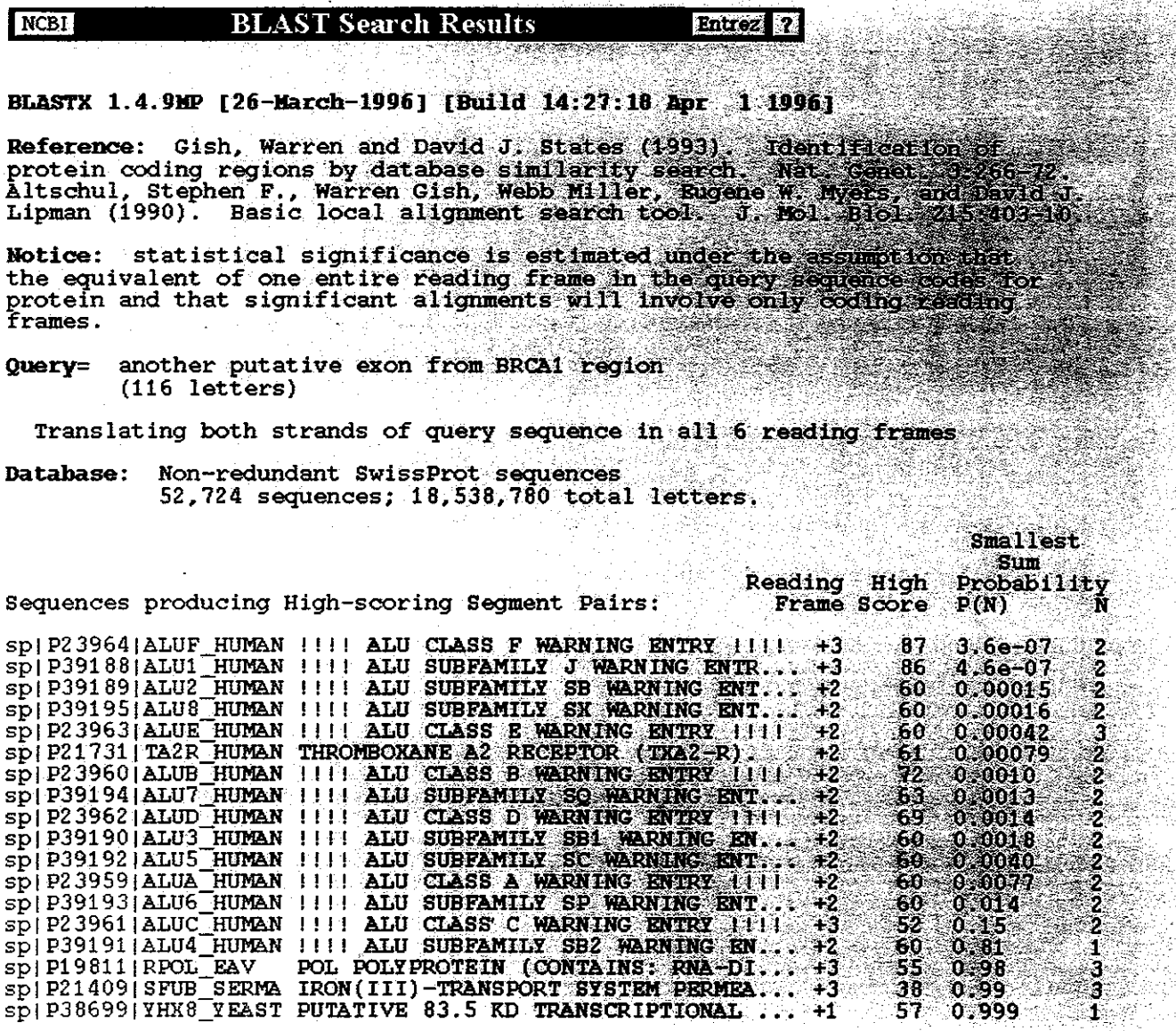
Query= another putative exon from BRCA1 region
(116 letters)

Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences
244,573 sequences; 341,283,545 total letters.

Sequences producing High-scoring Segment Pairs:

		High Score	Smallest Sum Probability P(N)	N
gb|U15595|HSU15595	Human breast and ovarian cancer sus...	580	1.2e-42	1
emb|X76757|HSSTGEX8	H.sapiens serotonin transporter gen...	260	2.8e-23	2
emb|X76809|HSDNAA3	H.sapiens genomic DNA clone a3	224	9.8e-22	2
emb|Z15025|HSBAT2	H.sapiens Bat2 gene	255	1.5e-21	2
gb|L39874|HUMDODDA	Homo sapiens deoxycytidylate deamin...	255	4.7e-21	2
gb|U07562|HSABLGR2	Human ABL gene, intron 1b, partial ...	251	4.8e-21	2
gb|L11910|HUMRETBLAS	Human retinoblastoma susceptibility...	251	9.7e-21	2
emb|X76808|HSD22	H.sapiens genomic DNA clone d2	224	1.0e-20	2
emb|Z11711|HSA2MGLB1	H.sapiens gene for alpha-2 macroglo...	228	1.3e-20	2
gb|M63377|HUMTRPM2A2	Human TRPM-2 protein gene, exon 4.	227	2.0e-20	2
gb|U52112|HSU52112	Human Xq28 genomic DNA in the regio...	226	2.4e-20	2
emb|X70413|HSTRANXYD	H.sapiens sequence involved in XY ...	260	2.6e-20	2
emb|X87344|HSEVMHC	H.sapiens DNA, DMB, HLA-Z1, LPP2+, L...	260	2.6e-20	2
gb|M87915|HUMALNE54	Human carcinoma cell-derived Alu RN...	260	4.7e-20	2
gb|L36092|HUMTCRB	Homo sapiens (clones: K41A, K35, K2...	250	4.7e-20	2
emb|Z68321|HSL79F5A	Human DNA sequence from cosmid L79F...	218	5.2e-20	2
gb|U51281|HSU51281	Human chromosome 11 cosmid cSRL-87f11.	237	7.5e-20	2
gb|U07000|HSU07000	Human breakpoint cluster region (BC...	230	7.8e-20	2
emb|X68793|HSAT3	H.sapiens gene for antithrombin III	228	9.0e-20	2
gb|M26434|HUMHPRTB	Human hypoxanthine phosphoribosyltr...	265	1.1e-19	2
emb|Z49862|HSL58B6	Human DNA sequence from cosmid L58b...	246	1.2e-19	2
gb|U04737|HSU04737	Human breakpoint cluster region 11q...	269	1.4e-19	2
emb|X83604|HSALL1GEN	H.sapiens all-1 gene	269	1.6e-19	2
emb|Z68870|HSL30G1	Human DNA sequence from cosmid L30G...	224	1.8e-19	2
emb|Z69890|HSRJ14	Human DNA sequence from cosmid RJ14...	243	2.1e-19	2

Figure 2a Results of a BLASTN query against nr using a second putative exon from the *BRCA1* region as the query sequence.



NCBI **BLAST Search Results** Entrez ?

BLASTX 1.4.9MP [26-March-1996] [Build 14:27:18 Apr 1 1996]

Reference: Gish, Warren and David J. States (1993). Identification of protein coding regions by database similarity search. Nat. Genet., 3:266-72. Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-10.

Notice: statistical significance is estimated under the assumption that the equivalent of one entire reading frame in the query sequence codes for protein and that significant alignments will involve only coding reading frames.

Query= another putative exon from BRCA1 region
(116 letters)

Translating both strands of query sequence in all 6 reading frames

Database: Non-redundant SwissProt sequences
52,724 sequences; 18,538,780 total letters.

Sequences producing High-scoring Segment Pairs:	Reading Frame	High Score	Smallest Sum Probability P(N)	N
sp|P23964|ALUF_HUMAN !!!! ALU CLASS F WARNING ENTRY !!!!	+3	87	3.6e-07	2
sp|P39188|ALU1_HUMAN !!!! ALU SUBFAMILY J WARNING ENTR...	+3	86	4.6e-07	2
sp|P39189|ALU2_HUMAN !!!! ALU SUBFAMILY SB WARNING ENT...	+2	60	0.00015	2
sp|P39195|ALU8_HUMAN !!!! ALU SUBFAMILY SX WARNING ENT...	+2	60	0.00016	2
sp|P23963|ALUE_HUMAN !!!! ALU CLASS E WARNING ENTRY !!!!	+2	60	0.00042	3
sp|P21731|TA2R_HUMAN THROMBOXANE A2 RECEPTOR (TXA2-R).	+2	61	0.00079	2
sp|P23960|ALUB_HUMAN !!!! ALU CLASS B WARNING ENTRY !!!!	+2	72	0.0010	2
sp|P39194|ALU7_HUMAN !!!! ALU SUBFAMILY SQ WARNING ENT...	+2	63	0.0013	2
sp|P23962|ALUD_HUMAN !!!! ALU CLASS D WARNING ENTRY !!!!	+2	69	0.0014	2
sp|P39190|ALU3_HUMAN !!!! ALU SUBFAMILY SB1 WARNING EN...	+2	60	0.0018	2
sp|P39192|ALU5_HUMAN !!!! ALU SUBFAMILY SC WARNING ENT...	+2	60	0.0040	2
sp|P23959|ALUA_HUMAN !!!! ALU CLASS A WARNING ENTRY !!!!	+2	60	0.0077	2
sp|P39193|ALU6_HUMAN !!!! ALU SUBFAMILY SP WARNING ENT...	+2	60	0.014	2
sp|P23961|ALUC_HUMAN !!!! ALU CLASS C WARNING ENTRY !!!!	+3	52	0.15	2
sp|P39191|ALU4_HUMAN !!!! ALU SUBFAMILY SB2 WARNING EN...	+2	60	0.81	1
sp|P19811|RPOL_EAV POL POLYPROTEIN (CONTAINS: RNA-DI...	+3	55	0.98	3
sp|P21409|SFUB_SERMA IRON(III)-TRANSPORT SYSTEM PERMEA...	+3	38	0.99	3
sp|P38699|YHX8_YEAST PUTATIVE 83.5 KD TRANSCRIPTIONAL ...	+1	57	0.999	1

Figure 2b Results of a BLASTX query against SWISS-PROT using a second putative exon from the *BRCA1* region as the query sequence. The PAM40 scoring matrix was used instead of the default BLOSUM62 matrix.

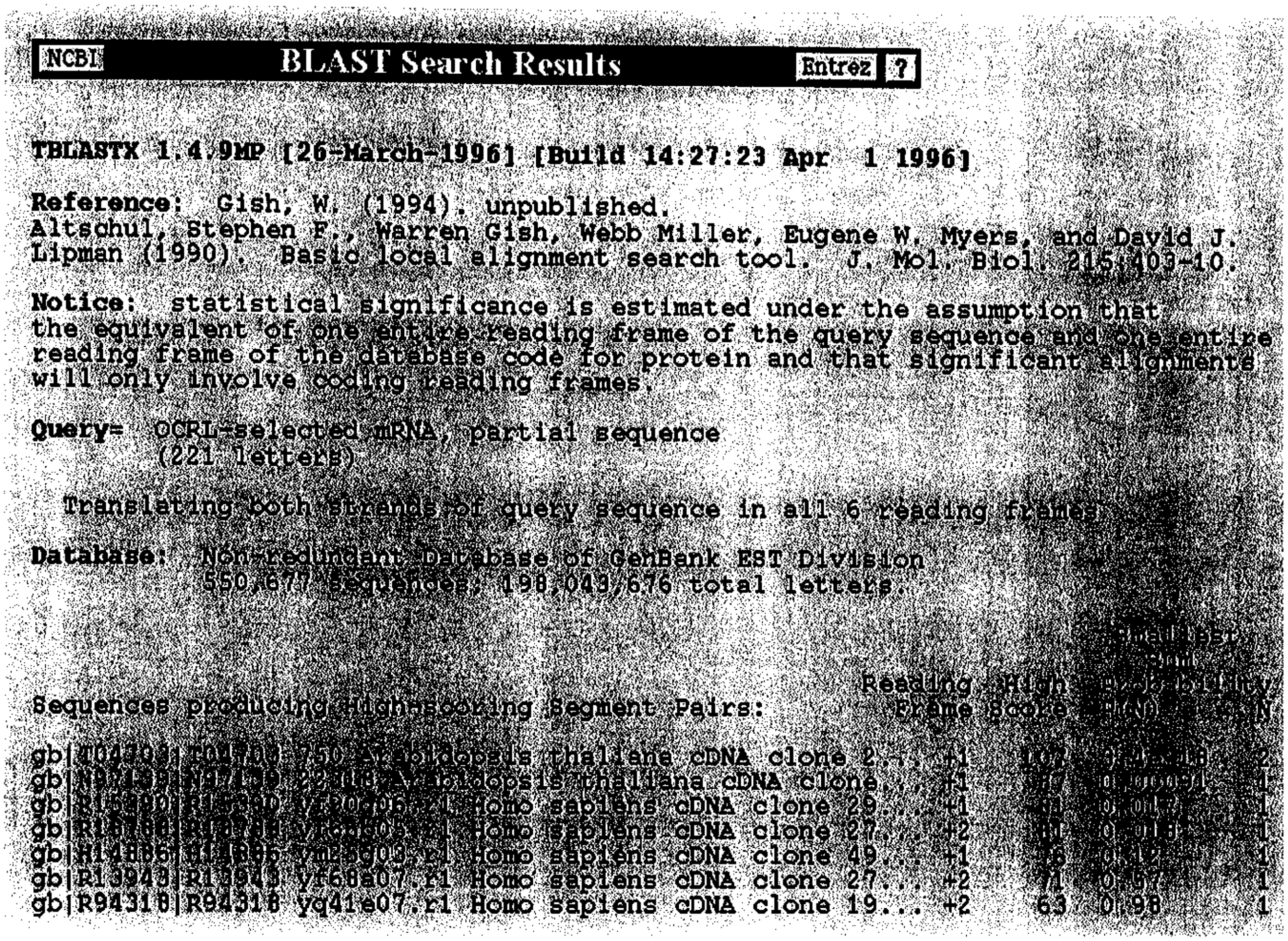

NCBI
BLAST Search Results
Entrez
?
TBLASTX 1.4.9MP [26-March-1996] [Build 14:27:23 Apr 1 1996]
Reference: Gish, W. (1994). unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-10.
Notice: statistical significance is estimated under the assumption that
the equivalent of one entire reading frame of the query sequence and one entire
reading frame of the database code for protein and that significant alignments
will only involve coding reading frames.
Query= OCRL-selected mRNA, partial sequence
(221 letters)
Translating both strands of query sequence in all 6 reading frames
Database: Non-redundant Database of GenBank EST Division
550,677 sequences; 198,043,676 total letters.
Reading
Sequences producing High-scoring Segment Pairs: Frame Score
gb| Arabidopsis thaliana cDNA clone 2... +1 2
gb| Arabidopsis thaliana cDNA clone... +1 1
gb| Homo sapiens cDNA clone 29... +1 1
gb| Homo sapiens cDNA clone 27... +2 1
gb| Homo sapiens cDNA clone 49... +1 1
gb|R13943|R13943 yf68a07.r1 Homo sapiens cDNA clone 27... +2 71 1
gb|R94318|R94318 yq41e07.r1 Homo sapiens cDNA clone 19... +2 63 0.98 1

Figure 3 Results of a TBLASTX query against dbEST using partial cDNA sequence data.

The results are shown in Figure 3. Actually, the mRNA fragment in this example corresponds to nucleotides 1465–1685 of the GenBank entry (M88162) for Lowe's oculocerebrorenal (OCRL-1) syndrome (Attree et al. 1992) and displays a highly significant similarity (p of 3.2×10^{-18}) to an ORF encoded by an *Arabidopsis thaliana* cDNA clone. The homologies between OCRL-1 and plant sequences are not even detectable, let alone significant, at the nucleotide level (data not shown). If the query were human genomic DNA, such a match to plant sequences would be strong evidence of an exon because similarities in noncoding sequences would not be expected to be conserved across such a vast evolutionary distance.

Protein Query Sequences

Given a coding nucleotide sequence and the protein it encodes, it is almost always preferable to use the protein as the query sequence to search a database because of the greatly increased sensitivity to detect more subtle relationships.

This is due to the larger alphabet of proteins (20 amino acids) compared with the alphabet of nucleic acid sequences (4 bases), where it is far easier to obtain a match by chance. In addition, with nucleotide alignments, only a match (positive score) or a mismatch (negative score) is obtained, but with proteins, the presence of conservative amino acid substitutions can be taken into account. Here, a mismatch may yield a positive score if the nonidentical residue has physical/chemical properties similar to the one it replaced. Various scoring matrices are used to supply the substitution scores of all possible amino acid pairs. The best scoring system for general purposes is the BLOSUM62 matrix (Henikoff and Henikoff 1993), and this is currently the default choice for the BLAST programs. It is important to recognize that BLOSUM62 is tailored for alignments of moderately diverged sequences (in practice, the most frequent type of query) and thus may not yield the best results under all conditions. Indeed, the PAM40 matrix used in Example 1b above yielded more convincing results than BLOSUM62 (not shown). Altschul (1993) recommends using a combination of three matrices to cover all contingencies. This may improve sensitivity, but at the expense of slower searches. In practice, a single BLOSUM62 matrix is routinely used but others (PAM40 and PAM250) are tried when an absolutely exhaustive search is necessary. Low PAM matrices are best for detecting very strong but localized sequence similarities, whereas high PAM matrices are best for detecting long but weak alignments between very distantly related sequences.

Confounding Subsequences and Query Masking

One of the most important advances in searching databases for similarities has been the introduction of methods for the automatic masking of low-complexity sequences. Problematic query sequences result in hundreds (or thousands) of spurious matches to nebulous entities with names such as proline-rich protein that may obscure more subtle but biologically interesting matches. The term low-complexity subsequences has a rigorous definition (Wootton and Federhen 1993) but may be thought of as synonymous with regions of locally biased amino acid composition. Biased means that the sequence deviates from the random model which underlies the calculation of the statistical significance (p value) of an alignment. Such alignments among low-complexity sequences are statistically but not biologically significant, i.e., one cannot infer homology (common ancestry) or functional similarity. Thus, it is best simply to note their presence but exclude them from the database search process. Several investigators have worked on the definition, origin, and detection of simple (low-complexity) sequences in protein and DNA (for review, see Wootton 1994b). The SEG (Wootton and Federhen 1996) and XNU (Claverie and States 1993) algorithms have been implemented as options for the BLAST family of programs, and they may be used individually or in combination.

Just how common is the problem of confounding, low-complexity segments among proteins? Wootton (1994b) has shown that they are surprisingly abundant, accounting for 25% or more of the residues in protein sequence databases. In the context of positionally cloned human disease genes, the data are even

more striking: Of the 42 gene products identified at the time of this writing, 83% have an average of five low-complexity segments, each accounting for between 1% and 48% of the residues in an individual protein. Clearly, masking low-complexity sequences is not an option—it should be used in every search, particularly those that involve conceptual translations (BLASTX, TBLASTN, TBLASTX), where, for example, translation of a poly(A) yields a polylysine ORF. Indeed, the BLAST Web servers now perform masking of low-complexity sequences by default, and this function must be turned off if masking is not desired.

NCBI **BLAST Search Results** Entrez ?

BLASTP 1.4.9MP [26-March-1996] [Build 14:27:01 Apr 1 1996]

Reference: Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990), "Basic local alignment search tool," J. Mol. Biol. 215:403-10.

Query= gi|1280437 candidate tub gene
(459 letters)

Database: Non-redundant GenBank CDS
translations+PDB+SwissProt+SPupdate+PIR
205,297 sequences; 58,469,162 total letters.

Sequences producing High-scoring Segment Pairs:		High Score	Smallest Sum Probability P(N)	N
gi|1279766	(U52433) candidate tub gene; similar...	2355	0.0	1
gi|1280437	(U52824) candidate tub gene; similar...	2355	0.0	1
gi|1305497	(U54644) tub homolog [Homo sapiens]	1569	0.0	2
sp|Q09306|YQQ4_CAEEL	HYPOTHETICAL 48.2 KD PROTEIN F10B5.4...	960	3.9e-114	5
sp|P46686|P46_MOUSE	P4-6 PROTEIN. /gi|467578 (X69827) ph...	502	6.0e-112	2
pir||S57757	nucleolar phosphoprotein - African c...	85	0.00028	2
pir||C61615	sericin MG-2 - greater wax moth (fra...	90	0.0014	2
sp|P27476|NSR1_YEAST	NUCLEAR LOCALIZATION SEQUENCE BINDIN...	74	0.0014	2
gi|1197667	(U46857) vitellogenin [Anolis pulche...	90	0.0019	1
gi|295671	(L11275) selected as a weak suppress...	91	0.0020	1
sp|P32583|SR40_YEAST	SUPRESSOR PROTEIN SRP40. /pir||S3817...	91	0.0020	1
pir||S52772	cylicin II - bovine /gi|757754 (Z467...	71	0.015	2
sp|P07856|SERI_BOMMO	SERICIN PRECURSOR.	84	0.027	2
gi|1118177	(U41992) coded for by C. elegans cDN...	55	0.042	4
gi|1206019	(U48298) rtoA gene product [Dictyost...	81	0.046	1
pir||B54415	transcription factor HBP-1a(1) - whe...	54	0.052	4
gi|1118007	(U40858) serine/threonine kinase [Ca...	68	0.052	2
gi|497894	(D12919) transcription factor HBP-1a...	54	0.054	4
sp|P33479|IE18_PRVKA	IMMEDIATE-EARLY PROTEIN IE180. /pir|...	63	0.054	4
gi|688015	(S73518) fibroblast growth factor re...	68	0.060	1
pir||S29268	protein kinase C substrate, 80K - mouse	59	0.062	2
sp|P26645|MACS_MOUSE	MYRISTOYLATED ALANINE-RICH C-KINASE ...	60	0.063	2
pir||S16519	protein kinase C substrate, 80K - mouse	60	0.063	2
sp|P40631|MLH_TETTH	MICRONUCLEAR LINKER HISTONE POLYPROT...	66	0.069	2

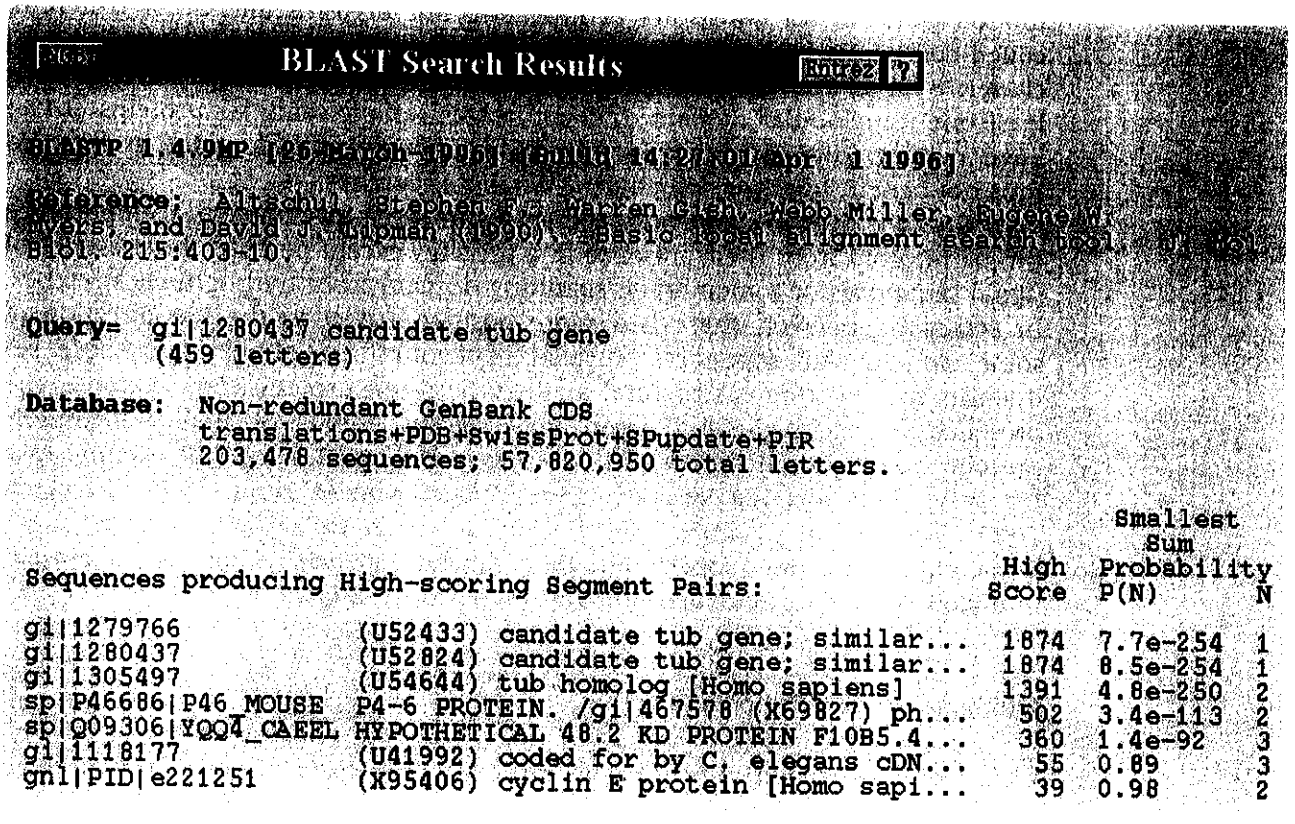
Figure 4a Results of a BLASTP query against nr using the product of the maturity-onset obesity (*tub*) gene.

Example 4

Recently, a gene responsible for maturity-onset obesity, *tub*, was cloned in the mouse (Noben-Trauth et al. 1996). The following is a BLASTP search in which the SEG algorithm is not used for filtering.

```
mail blast@ncbi.nlm.nih.gov
Subject:

PROGRAM blastp
DATALIB nr
BEGIN
>gi|1280437 candidate tub gene
MVQANADGRPRSRRARQSEEQAPLVESYLSSSGSTSYQVQEADSIASVQLGA
TRPPAPASAKKSKGAAASGGQGGAPRKEKKGKHKGTSGPATLAEDKSEAQGP
VQILTVGQSDHDKDAGETAAGGGAQPSGQDLRATMQRKGISSSMSFDEDEDE
DENSSSSSQLNSNTRPSSATSRKSIREAASAPSPAAPEPPVDIEVQDLEEFA
LRPAPQGITIKCRITRDKKGMDRGMYPTYFLHLDREDGKKVFLLAGRKRKKS
KTSNYLISVDPTDLSRGGDSYIGKLRSNLMGTKFTVYDNGVNPQKASSSTLE
SGTLRQELAAVCYETNVLGFKGPRKMSVIVPGMNMVHERVCIRPRNEHETLL
ARWQNKNTESIIELQNKTPVWNDDTQSYVLNFHGRVTQASVKNFQIIHGNDP
DYIVMQFGRVAEDVFTMDYNYPLCALQAFAIALSSFDSKLACE
```



BLAST Search Results

BLASTP 1.4.9MP [26-March-1996] [Build 14:27:01 Apr 1 1996]

Reference: Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). "Basic local alignment search tool." J. Mol. Biol. 215:403-10.

Query= gi|1280437 candidate tub gene
(459 letters)

Database: Non-redundant GenBank CDS
translations+PDB+SwissProt+SPupdate+PIR
203,478 sequences; 57,820,950 total letters.

Sequences producing High-scoring Segment Pairs:		High Score	Smallest Sum Probability P(N)	N
gi|1279766	(U52433) candidate tub gene; similar...	1874	7.7e-254	1
gi|1280437	(U52824) candidate tub gene; similar...	1874	8.5e-254	1
gi|1305497	(U54644) tub homolog [Homo sapiens]	1391	4.8e-250	2
sp|P46686|P46_MOUSE	P4-6 PROTEIN. /gi|467578 (X69827) ph...	502	3.4e-113	2
sp|Q09306|YQQ4_CAEEL	HYPOTHETICAL 46.2 KD PROTEIN F10B5.4...	360	1.4e-92	3
gi|1118177	(U41992) coded for by C. elegans cDN...	55	0.89	3
gnl|PID|e221251	(X95406) cyclin E protein [Homo sapi...	39	0.98	2

Figure 4b Results of a BLASTP query against nr using the product of the maturity-onset obesity (*tub*) gene. The results differ from those in Figure 4a since the SEG algorithm is being used for filtering.

The results of this search (Figure 4a) are voluminous and apparently significant. However, look at the alignments for vitellogenin or SRP40: How does one interpret these results?

The SEG algorithm can be applied to a protein sequence to divide it into low-complexity and high-complexity subsequences. This is illustrated for the *tub* gene product in Figure 4b. Note the four internal low-complexity regions that are enriched for serine, glycine, proline, and other amino acid residues. Such subsequences can be masked before a BLASTP search by including the directive FILTER seg between the DATALIB nr and BEGIN statements in the E-mail example presented here. As stated above, filtering is performed by the SEG algorithm by default if the BLAST Web pages are used for the search. Performing a masked search instead of an unmasked search reduces the hit list from nearly 300 matches (almost all false positives) to only seven meaningful matches, with the top five being statistically significant. All spurious matches disappear from the output. (In the filtered output alignments, masked residues appear as Xs, a character that the BLAST programs ignore.)

In addition to low-complexity segments, many proteins also have other regions that may lead to problems in interpreting database search results. Nonglobular domains of proteins, such as the collagen helix and myosin rod, also deviate from the random model of amino acid composition and represent an intermediate level of compositional complexity in comparison with globular structures (Wootton 1994b). Users who have spent time searching databases will have undoubtedly encountered a query sequence that yields a long and monotonous hit list consisting of little else but myosins and other α-helical, coiled-coil proteins. Again, this phenomenon is not irrelevant to the positional cloner: 81% of human gene products isolated in this manner to date have at least one nonglobular domain. Indeed, in this sample, only six proteins (14%) have neither low-complexity segments nor nonglobular domains and would thus not benefit from masking during database searches.

In the case of low-complexity sequences, little, if any, biological importance can be inferred from their presence in a protein, except perhaps the presence of signal sequences or a membrane-spanning domain from hydrophobic stretches. In contrast, however, there is a more positive aspect to the identification of some nonglobular sequences. Instead of evolutionary homology, these sequences may represent structural analogy between proteins. For example, the presence of a coiled-coil domain may signify homo- or hetero-oligomeric protein-protein interactions and thus suggest an experiment. This was the case for polyposis-associated familial colon cancer (Kinzler et al. 1991; Su et al. 1993). The default parameters of the SEG program are set to detect and mask low-complexity subsequences, but they can be adjusted to partition a protein into globular and nonglobular domains (Wootton 1994a).

Additional BLAST Options

BLAST and associated software tools can be further customized by using a large number of options that are more fully described in documents available by sending an E-mail *help* message to:

blast@ncbi.nlm.nih.gov

or on the BLAST World Wide Web page at:

http://www.ncbi.nlm.nih.gov/

Recently, a new network client-server program called PowerBLAST has become available by ftp from NCBI:

ncbi.nlm.nih.gov (directory/pub/sim2/PowerBlast)

This application runs on UNIX, Macintosh, and Microsoft Windows systems. PowerBLAST has a number of new and useful features, including automatic query masking and filtering, search restriction by organism, graphical output of gapped multiple alignments, the ability to analyze 100 kb per hour, and compatibility with NCBI's new database annotation and submission tool called Sequin.