NCBI Bookshelf. A service of the National Library of Medicine, National Institutes of Health.

Bergman NH, editor. Comparative Genomics: Volumes 1 and 2. Totowa (NJ): Humana Press; 2007.

Chapter 9 BLAST QuickStart

Example-Driven Web-Based BLAST Tutorial

David Wheeler and Medha Bhagwat.

Summary

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between protein or nucleotide sequences. The program compares nucleotide or protein sequences to sequence in a database and calculates the statistical significance of the matches. This chapter first provides an introduction to BLAST and then describes the practical application of different BLAST programs based on the BLAST Quick Start mini-course (www.ncbi.nlm.nih.gov/Class/minicourses). In each example, emphasis is placed on practical step-by-step procedures, although relevant theory is also given where it affects the choice of BLAST program, parameters, and database.

1. Introduction

BLAST is an acronym for Basic Local Alignment Search Tool and refers to a suite of programs used to generate alignments between a nucleotide or protein sequence, referred to as a "query" and nucleotide or protein sequences within a database, referred to as "subject" sequences. The original BLAST program used a protein "query" sequence to scan a protein sequence database. A version operating on nucleotide query" sequences and a nucleotide sequence database soon followed. The introduction of an intermediate layer in which nucleotide sequences are translated into their corresponding protein sequences according to a specified genetic code allows cross-comparisons between nucleotide and protein sequences. Specialized variants of BLAST allow fast searches of nucleotide databases with very large query sequences, or the generation of alignments between a single pair of sequences. Both the standalone and web version of BLAST are available from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov). The web version provides searches of the complete genomes of *Homo sapiens* as well as those of many model organisms, including mouse, rat, fruit fly, and *Arabidopsis thaliana*, allowing BLAST alignments to be seen in a full genomic context (1).

1.1. Query and Database Sequence Formats

BLAST "query" sequences are given as character strings of single letter nucleotide or amino acid codes, preceded by a definition line, beginning with a ">" symbol and containing identifiers and descriptive information. This format is known as FASTA. BLAST databases are constructed from concatenated FASTA formatted sequences using a program called "formatdb" that produces a mixture of binary- and ascii-encoded files containing the sequences and indexing information used during the BLAST search.

1.2. Scoring of Alignments and Substitution Matrices

A BLAST alignment consists of a pair of sequences, in which every letter in one sequence is paired with, or "aligned to," exactly one letter or a gap in the other. The alignment score is computed by assigning a value to each aligned pair of letters and then summing these values over the length of the alignment. For protein sequence alignments, scores for every possible amino

acid letter pair are given in a "substitution matrix" where likely substitutions have positive values and unlikely substitutions have negative values. By default, BLAST uses the "blosum62" matrix, a member of the most commonly used series of substitution matrix (2), however, several members of the PAM (3) series are also available. For nucleotide alignments, BLAST uses a reward of +2 for aligned pairs of identical letters and a penalty of −3 for each nonidentical aligned pair. The creation of a gap in an alignment results in a negative "gap-creation" penalty, with each extension of a preexisting gap incurring a lesser penalty. For a detailed treatment of the theory of alignment scoring (*see* ref. 4).

1.3. Overview of the Algorithm

BLAST begins a search by indexing all character strings of a certain length within the "query" by their starting position in the query. The length of the string to index, called the "wordsize" is configurable by the user. The allowable range for the "wordsize" varies according to the BLAST program used; typical values are 3 for protein-to-protein sequence searches and 11 for nucleotide to nucleotide searches. BLAST then scans the database looking for matches between the "words" indexed in the "query" and strings found within the database sequences. For nucleotide-to-nucleotide searches, these matches must be exact; for protein-to-protein searches, the score of the match as determined using a substitution matrix, must exceed a specified threshold. When a word match is found, two nearby words in the case of protein searches, BLAST attempts to extend both forward and backward from the match to produce an alignment. BLAST will continue this extension as long as the alignment score continues to increase or until it drops by a critical amount owing to the negative scores given by mismatches. This critical amount is known as the "dropoff." The methods BLAST uses to initiate refine alignments are given more fully in refs. 5 and 6.

1.4. Statistical Significance

The alignments found by BLAST during a search are scored, as previously described, and assigned a statistical value, called the "Expect Value." The "Expect Value" is the number of times that an alignment as good or better than that found by BLAST would be expected to occur by chance, given the size of the database searched. An "Expect Value" threshold, set by the user, determines which alignments will be reported. A higher "Expect Value" threshold is less stringent and the BLAST default of "10" is designed to ensure that no biologically significant alignment is missed. However, "Expect Values" in the range of 0.001 to 0.0000001 are commonly used to restrict the alignments shown to those of high quality.

2. Course and Website

This BLAST Quickstart chapter illustrates the use of the principal BLAST programs to solve problems that arise in the analysis of protein and nucleotide sequences. Each section provides a succinct description of a protocol with two problems that serve as practical examples. Relevant theory is given when it affects the selection of a search strategy or search parameter, however, the emphasis is on the procedure itself. The sections follow closely the structure of the BLAST QuickStart Mini-Course found at www.ncbi.nlm.nih.gov/Class/minicourses. The BLAST QuickStart is one of 10 2-h format Mini-Courses offered by NCBI on campus at the National Institutes of Health and at locations around the country to over 4000 students a year. The courses use a paired problems approach in which the first of two similar problems or problem sets is solved by the instructor during the first hour on a computer linked to a projection system, while the students watch; in the second hour, the students tackle the second problem, or set of problems

at their own computers. These courses have been effective as practical introductions to bioinformatics procedures. To get the most from the sections next, it will be necessary to navigate to the URL previously listed and click on the “BLAST Quickstart” link to reach the online exercises, although the liberal collection of screen shots will allow the reader follow along for the most part without web access.

3. Nucleotide BLAST

Nucleotide BLAST refers to the use of a member of the BLAST suite of programs, such as “blastn” to search with a nucleotide “query” against a database of nucleotide “subject” sequences.

3.1. Available Nucleotide-Level Searches

There are two members of the BLAST suite of programs that are designed to make nucleotide-to-nucleotide alignments. The first is the original BLAST nucleotide search program known as “blastn.” The “blastn” program is a general purpose nucleotide search and alignment program that is sensitive and can be used to align tRNA or rRNA sequences as well as mRNA or genomic DNA sequences containing a mix of coding and noncoding regions. A more recently developed nucleotide-level BLAST program called MegaBLAST (7) is about 10 times faster than “blastn” but is designed to align sequences that are nearly identical, differing by only a few percent from one another. MegaBLAST allows the rapid mapping of a transcript onto a typical 3 billion base mammalian genome in seconds, and is useful for processing large batches of sequences. A refinement of MegaBLAST, known as discontiguous MegaBLAST, uses a discontiguous template to define an initial “word” in which characters in some positions, such as those in the wobble base position of codons, need not match. Discontiguous MegaBLAST allows rapid cross-species mappings involving coding regions in cases where species differences in codon usage would prevent alignments using the original MegaBLAST program.

3.2. Examples of Nucleotide BLAST Searches

3.2.1. Problem 1

Click on the link indicated by “P” next to the “Nucleotide-nucleotide BLAST (blastn)” to access the problem. This problem demonstrates how to use BLAST to find human sequences in GenBank that can be amplified with a particular primer pair. Access the nucleotide–nucleotide BLAST page (by clicking on the Nucleotide–nucleotide BLAST link). Paste both the forward and reverse primers into the BLAST input box. Insert a string of about 30 N’s after the first primer sequence to separate the two sequences to be found in separate, not overlapping alignments. Limit your search to human sequences by selecting “*Homo sapiens*” from the “All organisms” pull down menu under the Options for advanced blasting and click the BLAST! link. Retrieve results by clicking on the “Format” button. Look for two hits to the same database sequence.

In this result, shown in Fig. 1, there are 13 GenBank entries that align to both the forward and reverse primers at different locations (indicated by thick bars) with a gap in between (indicated by a thin gray bar). There are two GenBank entries that align only to the reverse primer. One alignment of the primer pair to the GenBank entry L78833.1 is shown in Fig. 2. The forward primer aligns to the sequence L78833.1 on the forward strand (as indicated by Strand Plus/Plus) at nucleotides 3252..3270. The reverse primer aligns to the reverse strand (as indicated by Strand Plus/Minus) at nucleotides 3475..3457. Thus, the two primers will amplify the sequence from nucleotides 3252..3475 of the entry L78833.1. Retrieve the entry L78833.1 in Entrez, by clicking on it. The annotation shows that the amplified region covers the Exon 1a and the upstream

sequence of the *BRCA1* gene. Refer to the Note 1 for the multiple hits. You may perform similar search against the human genome BLAST database (*see* Note 2).

Fig. 1

Graphical overview of primer hits from the nucleotide–nucleotide search problem 1. The result shows that 13 GenBank entries align to the forward and reverse primers at different locations (indicated by thick bars) with a gap in between (indicated (more...)

Fig. 2

Alignment view of one of the primer hits to the GenBank entry L78833.1. The forward primer (query nucleotides 1..19) aligns to the sequence L78833.1 on the forward strand (indicated by Strand Plus/Plus) at nucleotides 3252..3270. The reverse primer (query (more...)

3.2.2. Problem 2

Click on the link indicated by "H" next to the Nucleotide–nucleotide BLAST (blastn) to access the problem. This problem describes how to obtain single-nucleotide polymorphism (SNP) information in similar sequences in the database. Hermankova et al. (8) studied the HIV-1 drug resistance profiles in children and adults receiving combination drug therapy. To identify the SNPs in the HIV-1 isolates from these patients, or other similar sequences in the database, use the sequence from one of the patients given next and run a nucleotide–nucleotide BLAST search as described in the problem previously listed. Format the results using the "Flat Query with Identities" option from the "Alignment View" pull down menu under the "Format" options (*see* Note 3). Identify the SNP observed at alignment position 6 (query nucleotide number 10) in Fig. 3. There is an A/G SNP in many of thedatabase sequences.

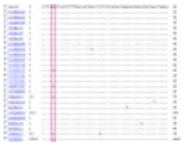

Fig. 3

Query-anchored alignment view for finding single-nucleotide polymorphism (SNP) from the nucleotide–nucleotide search problem 2. Nucleotides in the database sequences identical to the query HIV-1 sequence are indicated by dots and SNPs are indicated (more...)

4. Protein BLAST

Protein-to-protein sequence searches are performed using the original member of the BLAST suite of programs, known as "blastp."

4.1. Available Protein-Level Searches

The default wordsize for a blastp search is three; the default substitution matrix is the blosum62 matrix. Changing the wordsize from three to two increases the sensitivity of the search. Using a different substitution matrix can also have an effect on search sensitivity. During a "blastp" search, low-complexity regions of the query sequence are filtered to reduce the construction of spurious alignments and enhance search speed (*see* Note 4).

4.2. Examples of Protein BLAST Searches

4.2.1. Problem 1

Click on the link indicated by “P” next to “Protein–protein BLAST (blastp)” to access the problem. It describes how to use blastp to determine the type of protein. For this purpose, we will choose the database containing the curated and annotated protein sequences, such as RefSeq or Swissprot. Use the query sequence provided in the problem. This sequence was generated by translating a *5 exon* gene from *Drosophila*. To determine the nature of this protein, run a blastp search. Access the “Protein–protein BLAST (blastp)” page by clicking on the link, paste in the query sequence, select the Swissprot database from the “Choose database” pull down menu and click on the BLAST! link. For each protein–protein search, the query is also searched against the Conserved Domain Database (*see* Note 5). Retrieve results by clicking on the “Format” button. The protein is similar to a number of aspartate amino transferases.

4.2.2. Problem 2

Click on the link indicated by “H” next to the “Protein–protein BLAST (blastp)” to access a similar problem to determine the type of protein. Use the query sequence provided in the problem. This sequence was generated by translating a *4 exon* gene from Drosophila. To determine the nature of this protein, run a blastp search against the Swissprot database as described in Subheading 2. The protein is similar to a number of phosphoglucomutases.

5. Translated BLAST

Translated BLAST searches use a genetic code to translate either the “query,” database “subjects,” or both, into protein sequences, which are then aligned as in “blastp.” The translations are performed in the three forward as well as the three reverse reading frames so that no possible translation is missed.

5.1. Available Translated Searches

There are three varieties of translated BLAST search; “tblastn,” “blastx,” and “tblastx.” In the first variant, “tblastn,” a protein sequence query is compared to the six-frame translations of the sequences in a nucleotide database. In the second variant, “blastx,” a nucleotide sequence query is translated in six reading frames, and the resulting six-protein sequences are compared, in turn, to those in a protein sequence database. In the third variant, “tblastx,” both the “query” and database “subject” nucleotide sequences are translated in six reading frames, after which 36 (6 × 6) protein “blastp” comparisons are made. Protein sequences are better conserved than their corresponding nucleotide sequences. Because the translated searches make their comparisons at the level of protein sequences, they are more sensitive than direct nucleotide sequence searches. A common use of the “tblastn” and “blastx” programs is to help annotate coding regions on a nucleotide sequence; they are also useful in detecting frame-shifts in these coding regions. The “tblastx” program provides a sensitive way to compare transcripts to genomic sequences without the knowledge of any protein translation, however, it is very computationally intensive. MegaBLAST can often achieve sufficient sensitivity at a much greater speed in searches between the sequences of closely related species and is preferred for batch analysis of short transcript sequences such as expressed sequence tags.

5.2. Examples of Translated BLAST Searches

5.2.1. Problem 1

Click on the link indicated by “P” next to the “Translated query vs protein database (blastx)” to access the problem. This problem describes how to identify a frame shift in a nucleotide sequence by comparing its translated amino acid sequence to a similar protein in the database. Access the Blastx page by clicking on the link “Translated query vs protein database (blastx),” paste the nucleotide sequence provided in the problem in the query box and run the Blast search. The translation of the query sequence is similar to the sequences of envelope glycoproteins in the database. Compared to the similar proteins in the results, there appears to be a frame shift around nucleotide 268 as seen in Fig. 4. The query whentranslated in reading frame 2 (as indicated by a rectangle) up to nucleotide 268 is similar to only the first 89 amino acids of the database protein AAL71647.1. The translation of the query needs to be shifted to reading frame 1 (as indicated by an oval) to find similarity to the rest of the protein sequence. To discover the nucleotide difference around 268, refer to Note 6

Fig. 4

Detecting frame shifts using a translated search: Blastx problem 1. The query, when translated in reading frame 2 (highlighted by a rectangle) up to nucleotide 268, is similar to only the first 89 amino acids of the database protein AAL71647.1. The translation (more...)

5.2.2. Problem 2

Click on the link indicated by “P” next to “Translated query vs protein database (blastx)” to access the problem. Paste in the sequence provided in the problem and run the blastx search to obtain a result similar to that shown in Fig. 5. The translation of the querysequence 1..564 in reading frame 1 (as indicated by an oval) to find similarity to the rest of the protein sequence. There is a frame shift in the query nucleotide around 564. To find out the nucleotide difference around 564, refer to Note 7.

Fig. 5

Detecting frame shifts using a translated search: blastx problem 2. The translation of the query sequence 1..564 in reading frame 1 (highlighted by a rectangle) is similar to the first 184 amino acids of the database protein AAL91985.1. The translation (more...)

6. Genome BLAST

Genome BLAST refers to the application of any of the BLAST search programs to the complete genomic sequence of an organism or the transcript and protein sequences derived from its annotation.

6.1. Available Genome-Wide Searches

Genome BLAST services are available at NCBI for a variety of organisms including human, mouse, rat, fruit fly, and many others in a growing list. At a minimum, MegaBLAST and “blastn” searches against the complete genome are supported. These are usually offered in conjunction with “tblastn” searches against the genome, “blastp” and “blastx” searches against the proteins

annotated on the genome and MegaBLAST, “blastn” and “tblastn” searches against collections of transcript sequences that have been mapped to the genome. Hits to the genome are displayed graphically within NCBI’s MapViewer to show their genomic context.

6.2. Examples of Genome-Wide BLAST Searches

6.2.1. Problem 1

Click on the link indicated by “P” next to mouse genome BLAST to access the problem. This problem describes how to use mouse genome blast to identify the Hoxb homologues encoded by the mouse genomic assembly sequence. As described in Subheading 5.1., translated searches or protein–protein searches are more sensitive for identifying similarity in the coding regions than the nucleotide–nucleotide searches. Within the translated or protein–protein searches, tblastn will be more appropriate than blastx or blastp for this problem. Both latter programs will use protein databases consisting of already identified protein sequences whereas tblastn will be useful for identifying unannotated coding regions as well.

We will demonstrate the sensitivity of tblastn as compared to the nucleotide–nucleotide search to identify a similarity to a coding region by running two searches: (1) MegaBLAST the query mRNA sequence, NM_008268, against the mouse genomic sequence and (2) tblastn the query protein sequence, NP_032294, against the mouse genomic sequence. Access the mouse genome BLAST page, by clicking on the “mouse” link under the Genomes panel. For the first search, paste the accession number NM_008268 into the query box, accept the default MegaBLAST option, and select the “genome (reference only)” as the database. The results, shown in Figs. 6 and 7, contain only four hits, two to the two *Hoxb5* coding exons and one each to the *Hoxb3* and *Hoxd3* genes. Pay attention to the “Refer to Features in this part of subject sequence.” Three of these hits, two to the *Hoxb5* and one to the *Hoxb3* genes, are on the Contig NT_096135.3 placed on chromosome 11. For the second search, paste the protein accession number NP_032294 into the mouse genome search page, select “genome (reference only)” as the database and tblastn as the program. The result should appear similar to that shown in Fig. 8. This search gives several more hits than the earlier MegaBLAST search. Pay attention to the “Refer to Features in this part of subject sequence.” There is a complete hit to the homeobox B5 protein, shown in Fig. 9, and to the homeodomains of the other members of the homeobox B family, seen in Fig. 10 (corresponding to the amino acids 195..253 in the query), such as B6, B4, B3, B2, B13, and so on, onchromosome 11, homeobox A family members on chromosome 6, and homeobox C family members on chromosome 15 (refer to Note 8 for the locations of conserved domain).

Fig. 6

Genomic MegaBLAST against the mouse genome problem 1: graphical overview. The mRNA query NM_008268.1 gets four hits to the mouse genome as highlighted by an oval. A part of the alignment view of the hit to the homeobox B5 coding region is also displayed (more...)

Fig. 7

Genomic MegaBLAST against the mouse genome problem 1: alignment view. Two of the BLAST hits, for the query NM_008268.1, shown in Fig. 6, are to the homeobox B3 and D3 coding regions (as highlighted by rectangles).

Fig. 8

Genomic tblastn against the mouse genome problem 1: graphical overview. The protein query, NP_032294, on performing tblastn against the mouse genome gets several more hits than the MegaBLAST of the corresponding mRNA against the mouse genome as shown (more...)

Fig. 9

Genomic tblastn against the mouse genome problem 1: alignment view. The protein query, NP_032294, on performing tblastn against the mouse genome, aligns completely to the two coding exons of the homeobox *B5* gene annotated on the mouse genome contig NT_096135.3 (more...)

Fig. 10

Genomic tblastn against mouse genome problem 1: alignment view (continued). Some of the hits in the region of the conserved homeodomain (amino acid residues 195..253) of the query protein NP_032294 to other members of the homeobox protein family are displayed. (more...)

6.2.2. Problem 2

Click on the link indicated by “H” next to mouse genome BLAST to access the problem. This problem describes how to use mouse genome blast to identify the protocadherin β homologues encoded by the mouse genomic sequence. As described in Subheading 2., tblastn will be useful for identifying unannotated homologues also. Access the mouse genome BLAST page, by clicking on the “mouse” link under Genomic BLAST. Paste the accession number for protocadehrin β 1 protein, AAK26059, in the query box, select “genome(reference only)” as the database and tblastn as the program. The result contains a complete hit to the protocadehrin β 1 protein and to other members of the protocadehrin β and γ subfamily A on chromosome 18 (*see* Note 8 for the locations of conserved domains).

7. BLAST2Sequences

BLAST2Sequences is used to compare two sequences, protein or nucleotide, using any one of the principal BLAST variants, “blastp,” “blastn,” “tblastn,” “blastx,” “tblastx,” or MegaBLAST.

7.1. Comparisons Between Two Sequences

The output of BLAST2Sequences consists of a set of the traditional pairwise alignments generated by the principal BLAST programs it uses, supplemented with a dot plot representation of these alignments. The dot plot is useful for highlighting deletions and duplications of segments between two sequences. The translated variants of BLAST2Sequences are useful for the detection of exons.

7.2. Examples of Two Sequence Comparisons

7.2.1. Problem 1

Click on the link indicated by “P” next to “Align two sequences (bl2seq).” This problem describes the comparison of two nucleotide sequences. The problem provides a genomic sequence and an mRNA (cDNA) sequence. The genomic sequence is a piece from a GenBank HTG record that contains part of the Werner’s syndrome gene *WRN*. This Gene contains 35 exons. The figure in the problem on the BLAST QuickStart website shows the mapping of exons to the cDNA coordinates. We will use BLAST2Sequences to determine which exon, if any, is contained in the supplied HTG sequence by comparing it against the *WRN* gene cDNA sequence. Access the BLAST2Sequences page by clicking on the link “Align two sequences (bl2seq).” Paste the HTG sequence in the top box under “Sequence1” and the cDNA sequence in the bottom box under “Sequence2.” Click on the “Align” button to obtain the outputof Fig. 11. The query (genomic sequence) nucleotides 116..233 are similar to “Sbjct” (cDNA sequence) nucleotides 954..1071. Compare the cDNA coordinates to the exon coordinates provided in the problem. The cDNA coordinates 954..1071 correspond to the exon number 8. Thus, the provided genomic DNA contains exon 8 of the *WRN* gene.

Fig. 11

BLAST2Sequences problem 1 output: detecting an exon on unannotated genomic sequence. The query (genomic sequence) nucleotides 116..233 are similar to “Sbjct” (cDNA sequence) nucleotides 954..1071. Compare the cDNA coordinates to the exon (more...)

7.2.2. Problem 2

Click on the link indicated by “H” next to “Align two sequences (bl2seq).” This problem describes the importance of one of the BLAST parameters. The problem gives one DNA sequence. Paste the sequence in both the Sequence 1 and Sequence 2 windows in the BLAST2Sequences page, and click on Align to reach a display similar to that of Fig. 12. Why is the alignment broken into two parts? The sequence between the nucleotides 655..752 is missing in the alignment of the sequence to itself. This is because of the default Low complexity filter option. Unclick the “Filter” option and perform the search again. Now the query sequence aligns completely to itself (as it should). BLAST masked the nucleotides 655..752, missing in the Alignment View of Fig. 13, as it is a low complexity region (*see* Note 4).

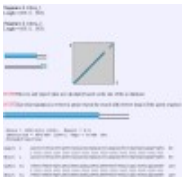

Fig. 12

BLAST2Sequences problem 2 output: the mystery of the missing piece. The alignment of the query sequence to itself is broken into two parts.

Fig. 13

BLAST2Sequences problem 2: alignment view at the junction. The sequence between the nucleotides 655..752 is missing in the alignment of the sequence to itself. The nucleotides at the junction of the two alignments are highlighted by rectangles.

8. Notes

Note 1

GenBank and nr. The remaining 12 hits of the primer pair to the database sequences may represent the potential for amplification of different regions of the human genome. Alternatively, the result may stem from the redundant nature of GenBank. The default "nr" database used in this problem includes nucleotide sequences from the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan, the European Molecular Biology Laboratory, and GenBank at NCBI (9, 10). It is redundant in nature as each laboratory can submit the nucleotide sequence that they sequenced even if an identical sequence already exists in the database. The nucleotide "nr" BLAST database is not made nonredundant as opposed to the protein "nr" BLAST database, which is made nonredundant by clustering identical proteins in one group. Thus, further analysis of the annotation on each of the entries is necessary to determine whether the primer pair will amplify a unique region of the human genomic DNA in the *BRCA1* gene. The user may perform the same search against the human genome BLAST database, which is nonredundant (refer to Genomic Blast (http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=9606) and Note 2). Which parameters will you need to change (*see* Note 2)?

Note 2

Primers and human genome BLAST. If a primer pair is used (with some N's in between) as a query on the human genome BLAST page (NCBI home page-BLAST-Genomes-Human), the user will need to use the blastn as the program and increase the*e*-value to, say, 10. Select the "genome (reference only)" as the database. The primer pair finds only one hit on chromosome 17, shown in Fig. 14, aligning to two nearby regions (joined by a thin gray line). The parts of the forward and reverse primers also align to two different regions of the genome (as indicated by two separate hits not joined by a thin gray line) on chromosome X and 2. You may view the hit on the human genome by clicking on the Genome View button at the top and accessing the Map Viewer (Fig. 15).

Fig. 14

Graphical overview of primer hits from the nucleotide–nucleotide search problem 1 on the human genome. The primer pair finds only one hit on chromosome 17 aligning to two separate regions shown by thick bars with a gap in between shown by a thin (more...)

Fig. 15

Graphical overview of primer hits from the nucleotide–nucleotide search problem 1 on the human genome. Press the "Genome View" button highlighted by a rectangle to see hits on the human chromosomes.

Note 3

Alignment views. BLAST offers a variety of alignment views. For example, the pairwise option shows an alignment of the query to one database sequence at a time. There are other options such

as, Query-anchored with identities, Query-anchored without identities, Flat Query-anchored with identities, and Flat Query-anchored without identities. These options show the multiple alignment-like view of the query with the database sequences. They differ in the way identities and gaps are displayed. The option, Flat Query-anchored with identities, is useful to identify the conserved regions (indicated by the dots) in the database sequences with respect to the query and the SNPs (indicated by the one letter code).

Note 4

Low-complexity sequence. The phrase “low-complexity sequence” refers to stretches of nucleotide or protein sequence that are repetitive or simple in composition (11). Extreme examples include runs of As in a nucleotide sequence such as the poly-A tails of eukaryotic mRNAs, or the poly-proline tracts found in some proteins, but the runs need not be limited to repeats of a single base or amino acid. BLAST detects and filters these runs in the “query” by default because they often lead to false starts when BLAST initiates alignments from word hits; beginning an alignment in the poly-a tail of an mRNA is not very likely to lead to a meaningful alignment between related mRNA sequences. This filtering can be turned off on the web using a checkbox, however, the resulting searches will take much longer because BLAST will have to process a great number of false starts. The results returned may also include a larger than usual number of questionable alignments. Nucleotide sequences are filtered using a program called Dust (12); protein sequences are filtered with SEG (13).

Note 5

Automatic CDD search. When a protein–protein BLAST search in ran, the query protein sequence is also searched against the conserved domains database. The presence of a conserved domain in the protein is reported on the page with the request ID before you format the page. For example, for the blastp problem 1, the query protein contains an amino transferase 1_2 conserved domain indicated by the red bar below the query line seen in Fig. 16. Click on the red bar to accessthe conserved domain database and determine the amino acid positions of the domain.

Fig. 16

Automatic conserved domain search: graphical overview. The query protein in the protein–protein BLAST problem 1 contains an amino transferase 1_2 conserved domain indicated by the red bar below the query line.

Note 6

Single nucleotide differences in blastx Problem 1. To discover the nucleotide difference in the BLASTX Problem 1, we will compare the query nucleotide sequence to the nucleotide sequence on which the protein AAL71647.1 is annotated. Click on the accession number AAL71647.1. The protein is annotated on the nucleotide entry AY077250.1 as shown in “DBSOURCE.” From the BLAST mini-course page, click on the link “Align two sequences (bl2seq)” in the panel labeled “Special.” Paste the query nucleotide sequence from the problem in the box for Sequence 1 and the accession number, AY077250, in the second box. Unclick the filter box (*see* Note 4) and click the “Align” button to create the output shown in Fig. 17. The query nucleotide lacks an “A” corresponding to the nucleotide 266 in AY077250.1 causing a frame shift. There are other differences between the two nucleotide sequences (such as a nucleotide substitution or deletion of three nucleotides), which do not cause a frame shift.

Fig. 17

Identification of a frame shift from BLASTX problem 1. The region of the query nucleotide lacks an "A," corresponding to the nucleotide 266 in AY077250.1, causing a frame shift that is highlighted by a rectangle.

Note 7

Single nucleotide differences in blastx Problem 2. To discover the nucleotide difference in the blastx problem 2, click on the accession number AAL91985.1. The protein is annotated on the nucleotide entry AF482979.1 as shown in "DBSOURCE." From the BLAST mini-course page, click on the link "Align two sequences (bl2seq)" in the panel labeled "Special." Paste the query nucleotide sequence from the problem in the box for Sequence 1 and the accession number, AF482979, in the second box. Unclick the filter box and click on the "Align" button to produce the alignment of Fig. 18. The querynucleotide sequence contains an extra "T" at nucleotide 565.

Fig. 18

Identification of a frame shift from BLASTX problem 2. The region of the query sequence containing an extra "T," compared to AF482979.1, at position 565 is highlighted by a rectangle.

Note 8

Manual CDD search. A protein query can be also manually searched against the conserved domain database. The option is provided under the Protein panel at the "Search the conserved domain database (rpsblast)" link. Perform this search for the protein accession number NP_032294 from the Genomes problem 1 (Fig. 19).

Fig. 19

Conserved domain search results. The query protein in the Genome BLAST Problem 1, NP_032294, contains a homeodomain between amino acid 195..253, highlighted by ovals. Perform this search for the protein accession number NP_032294 from the Genomes Problem (more...)

References

1. Madden TL, McGinnis S. Blast: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32:W20–W25. [PMC free article: PMC441573] [PubMed: 15215342]
2. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992;89:10,915–10,919. [PMC free article: PMC50453] [PubMed: 1438297]
3. Natl Biomed Res Found. Washington, DC; 1978. Atlas of Protein Sequence and Structure, chapter Matrices for detecting distant relationships.
4. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol. 1996;266:460–480. [PubMed: 8743700]
5. Madden TL, Schaffer AA, Zhang J, et al. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–3402. [PMC free article:

PMC146917] [PubMed: 9254694]
6. Madden T. The NCBI Handbook. The blast sequence analysis tool.
7. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning dna sequences. J Comput Biol. 2000;7:203–214. [PubMed: 10890397]
8. Hermankova M, Ray SC, Ruff C, et al. Hiv- 1 drug resistance profiles in children and adults with viral load of <50 copies/ml receiving combination therapy. JAMA. 2001;286:196–207. [PubMed: 11448283]
9. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. Genbank. Nucleic Acids Res. 2006;34:16–20. [PMC free article: PMC1347519] [PubMed: 16381837]
10. Mizrachi I. The NCBI Handbook. Genbank.
11. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 1996. pp. 554–571. [PubMed: 8743706]
12. Tatusov R, Lipman DJ. Dust. Unpublished data.
13. Wootton JC, Federhen S. Computers and Chemistry. Elsevier Science; Amsterdam, The Netherlands: 1993. Statistics of local complexity in amino acid sequences and sequence databases.

Bookshelf ID: NBK1734