This assignment will describe how to do analyses using the NCBI WWW BLAST Server (http://www.ncbi.nlm.nih.gov/BLAST/).

Data for the Assignment

The files dna1, dna2, pep1 are given to you as text files on the course web page. The files are plain text files with just a single nucleotide or peptide sequence in FASTA format.\(^1\) In most browsers the sequences are opened up in the window of the browser and the sequences can then be selected and copied and pasted in BLAST.

Summary of Results (Problem Zero)

0. Several searches will be performed as part of this assignment. As you do the various parts, copy the appropriate results into the following table so that it is easier to keep an overview. *Be sure to do this, it will be graded.*

<table>
<thead>
<tr>
<th>Program</th>
<th>Scoring matrix</th>
<th>Best E-value Sequence 1</th>
<th>Number of matches with $E \leq 0.001$ Sequence 1</th>
<th>Best E-value Sequence 2</th>
<th>Number of matches with $E \leq 0.001$ Sequence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTN</td>
<td>default (^a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLASTN</td>
<td>-r 1 -q -1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLASTP</td>
<td>default (^b)</td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>TBLASTN</td>
<td>default (^b)</td>
<td></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

\(^a\) +1 for identity, -3 for difference (-r 1 -q -3)  
\(^b\) BLOSUM62

Using BLASTN to Find DNA Sequences Similar to dna1 and dna2.

The sequences called dna1 and dna2 are genomic DNA sequences. In this assignment you will compare results of doing the analysis a few different ways.

---

\(^1\) FASTA format if defined by having a line that starts with a greater than sign (>) and the description of the sequence. All remaining lines to the end of the file, or the next line that starts with a greater than sign (>, is interpreted as sequence. Most programs that accept a FASTA format file will allow blank spaces (and sometimes even numbers) in the lines with sequence data.
1. Use the "blastn" program through the NCBI BLAST server to find the sequences in the non-redundant (nr) nucleotide database that are most similar to dna1 and dna2. Do a search for gapped alignments. Paste the title and sequence of dna1 or dna2 into the query sequence box.² Use the default search parameters. Click the "BLAST!" button.

On the next page, click the "Format!" button, and wait for the results. [By default a new window is opened, which will allow you to use the Web browse "Back" button to go back to the BLAST page, change some things, and do another search. This can be done several times, as long as you are willing to keep track of the windows.]

Start by looking at the descriptive information at the top of the results page.

1a. What version of the BLAST program was used in the search?

1b. How many nucleotides (letters) were in the query sequences for dna1 and dna2?

1c. What is the date that you are doing this analysis (the databases change size every day)?

1d. How many sequences were in the non-redundant nucleotide database?

1e. How many nucleotides (letters) were in the non-redundant nucleotide database?

1f. What was the average length of an entry in the non-redundant nucleotide database?

The significance of a BLAST match to a database entry is summarized by the $E$-value. In words, the $E$-value is the expected number of times that a random query sequence would align with database sequences with scores greater than or equal to $S$.

1g. For the best matching database sequences for dna1 and dna2, what is the value of $S$ (in bits)? Two answers are needed here!

1h. For the best matching database sequences for dna1 and dna2, what is the $E$-value for observing one or more database alignments with scores greater than or equal to this value of $S$? Again for both sequences.

1i. About how many database sequences for dna1 and dna2 have regions of similarity to the query sequence with an $E$-value less than or equal to 0.001? (You can do the search with the $E$-value set to 0.001 and let BLAST do the counting for you!) Two answers.

Changing the BLASTN Scoring Matrix

2. BLASTN may not always be much help in finding sequences similar to the query. Part of this is the default BLASTN scoring matrix (+1 for identical residues and −3 for differing residues). In this part of the problem, you will use BLAST options to change the scoring matrix.

As noted above, you should still have a WWW browser window with the title "BLAST". If you are still on the page that starts with "formatting BLAST" or "Your request has been successfully

---

² Be sure to start the title with the > sign so that it is not interpreted as part of the sequence.
submitted and put into the Blast Queue", use the browser "Back" button to return to the main BLAST search page. Your program and database selections, and query sequence should still all be there.3

Scroll down the region of the page labeled "Options for advanced blasting" and find the box labeled "Other advanced:". In that box enter the text

\[-r 1 -q -1\]

in order to change the nucleotide comparison scoring matrix.

Repeat the search of the nr nucleotide database for sequences similar to dna1 and dna2 with the new comparison matrix (i.e., click "BLAST!" then "Format!").

2a. For the best matching database sequences for dna1 and dna2, what is the value of $S$ (in bits)?

2b. For the best matching database sequences for dna1 and dna2, what is the $E$-value for observing one or more database alignments with scores greater than or equal to this value of $S$?

2c. For the best matching database sequences for dna1 and dna2, what is the percentage DNA sequence identity in the highest scoring sequence alignment?

2d. How many database sequences for dna1 and dna2 have regions of similarity to the query sequence with an $E$-value less than or equal to 0.001?

2e. Compare the results of 1i and 2d. Which DNA scoring matrix gives a more sensitive detection of significant similarities (for now, $E \leq 0.001$) of sequences as distantly related as these are?

2f. Look at the descriptions of the DNA sequences with significant similarity. How many of them have a single defined gene or function associated with them (as opposed to being a genomic region that might have multiple genes)?

2g. From these descriptions, what, if any, function would you assign to dna1 and dna2, and why?

---

Using BLASTP to Find Proteins Similar to pep1, the Presumptive Product of dna1

Although BLASTX will translate a sequence for you, it is sometimes preferable to translate the sequence into an amino acid query before doing the analysis.4 The sequence pep1 is the desired translation of dna1. Go back to the main BLAST page (http://www.ncbi.nlm.nih.gov/BLAST/).

---

3 If at any point in the assignment you lose them, you must reenter the query.
Follow the link to "Standard protein-protein BLAST [blastp]". Paste the pep1 title and sequence into the "Search" box on the protein-protein BLAST page. Click the "BLAST!" button and the "Format!" button.

3a. What version of the BLAST program was used in the search?

3b. How many amino acids (letters) were in the query sequence?

3c. For the best matching database sequence, what is the value of $S$ (in bits)?

3d. For the best matching database sequence, what is the $E$-value for observing one or more database alignments with scores greater than or equal to this value of $S$?

3e. How many database sequences have regions of similarity to the query sequence with an $E$-value less than or equal to 0.001?

3f. Why is there such a difference between the amount of database sequences for dna1 and pep1?

3g. From the different database sequences obtained from pep1 and dna2, what can you conclude about the kind of proteins we are dealing with here (in terms of evolution, etc.).

Using TBLASTN to Find DNA Sequences with Translations Similar to pep1

4. In comparing BLASTN results with searches of the protein databases, we failed to consider the possibility that rather than a difference in sensitivity, there might more proteins in the protein database than there are corresponding genes in the DNA databases. We will now address this by using TBLASTN to search the nucleotide databank for DNA sequences that could be translated to yield proteins similar to pep1.5

Go back to the main BLAST page (http://www.ncbi.nlm.nih.gov/BLAST/). Follow the link to "Protein query - Translated db [tblastn]". Paste pep1 into the "Search" box. Make sure that the pop-up menu for search programs says "PROTEIN Query - TRANSLATED database [tblastn]". Click "BLAST!" and then "Format!".6

4a. For the best matching database sequence, what is the value of $S$ (in bits)?

---

4 Just because BLASTX finds a similarity in one translation frame does not mean that this is the correct translation. There are a number of incorrectly translated proteins in the databases, and they create misleading similarities.

5 Because TBLASTN translates in all six frames, it does not depend on sequence annotations defining presumptive gene locations and reading frames.

6 Once again, the change between protein and nucleotide database is automatic with the change of program.
4b. For the best matching database sequence, what is the $E$-value for observing one or more database alignments with scores greater than or equal to this value of $S$?

4c. How many database sequences have regions of similarity to the query sequence with an $E$-value less than or equal to 0.001?

4d. Can you explain the difference in number of sequences found in 1 and 3 by a lack of genes encoding proteins related to pep1? Explain your reasoning.
APPENDIX
Definitions for Homework 3

> dna1
CCGATGAAATTCACAGGCTTCAATCGGACATCTGACGAGCTTCAATCGGACCTGATAATGTATTACTTTTCTCT
TGACTGATTCGCGCCCCACGGGTGTCGCGCGGCGGAAATTTCTCTCCTGGTACTAGCTACCCGGGTAACCGGTAAAAATAG
GTAAAAATCTGTAGCACAATAATATCTTATGGCCTGTATTDCACGCCAGTTTCCAGCTGTTAAAGATCCCGGTAAACCAAC
AGGACGATCTCCTGCGGCACCTTACCCACGCGGACTTTGCGTGAATTTGACAATCGGCGACTTGGCGCAATACCAGTGATC
ACATCAATAGACCCACAAATACGGACAGCTGCTAATGGAATGCAATCCTGAGCCTTTTCCAGGACGACGAGAATCCCATG
CGCCGTTTGAAACCCCTCATCGCCGCATTTGGCTTGAGTTTTAGCATCGGCAGCTTCGCCAATACCGACATCACCATAGAC
ACCACATGAACCCGCACTTCCGAGTTGGGCTAAGCTATGGAATGCTGTTGGTACTGAGGATTACCACTGCGTGTTGCGTG
TTGACCTGTTGGAGATTCTGGGCGTAAGCTATTGGAATAGCAAGAAATCTATGATCTTGACGTTTGAGATTAACGTGCCCCT
GGTGCGTTGGCCCGGTGGCTGTTTTGCGGATGAATACCACTGGCGTGATGGCATAGGCCGCACCGGACCAGCGCCCGGAAT
GGCTGGGAATATATAGCCGCAAGGCCAATATACCCCTGCGCGGAATTTGCGGCTAAAATTGACCGGGAATACCCCTCAGG
TCCAGTTTTTGCAAAATGGTAAATGAGGCTAGGGATTTGTGGGCCAAAGAGGATCTAGGATTAGTTAACATCAACTGGA
TTTGAATTAGGTCCCCATCAGTCGATATTGGTGAAATTATCCCCAGCTGTAGGGAAGGGGTTATAGACTTTGTTTTTTTGA
TTCTTAAAGAATTGAGAAACTATGGCATGTTCCATCTCAGCAATGGTTTGGGCTTAAAACAATATGTTAATAAAGTTCTGA
AAATAATCCAAATTAAAAGGAGACTAAGCAAATGGTTCATTTCCACCCCTTTGGCAATGTTAACTTTTACGAAATGGATTG
GAGTCTTAAGGGAGATTATGGGCTCATGATCCAGTCATCGCAAAAGAGGGAAGTCGATGGTATGTTTTTCACACCGGGAGC
GGCATTCAAATCAAAACATCTGAAGATGGGGTTCATTGGGAGAATATGGGCCGGGTATTTCCATCGTTGCCGGATTGGTG
AAGCAGTATGTTCCAGAAAAAGACGAAGATCATTTATGGGCGCCCGATATATGTTTTTATAACGGAATTTACTATCTGTAT
TATTCGAGCTCTACATTGGAAAAAACACTCGGGCTTGGAGATTTCTGGGAAATACCTGCTGTTGGATTTGATCACTGAGC
TTATGAAAGTACCAGCTTCGCTTGACCGTTTACCTACCAACAGCTGCTGCTTGATGCTGCAATGCGACAGGCTCTGAGG
CATTGCCTCCTTCATCGTCTCCTGGTACAGCTAGCAGAGCTGGAGCGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
CCGGTTTTTGGTATGCGACGCTACGTTAAATTAAACACATCTGCGCCTGCGGCTGCGGCTGCGGCTGCGGCTGCGGCTGCG
> dna2
AGATCTAGGATGTATGGAACCTATGAAGGCTAGGGATTTGTGGGCCAAAGAGGATCTAGGATTAGTTAACATCAACTGGA
TTTGAATTAGGTCCCCATCAGTCGATATTGGTGAAATTATCCCCAGCTGTAGGGAAGGGGTTATAGACTTTGTTTTTTTGA
TTCTTAAAGAATTGAGAAACTATGGCATGTTCCATCTCAGCAATGGTTTGGGCTTAAAACAATATGTTAATAAAGTTCTGA
AAATAATCCAAATTAAAAGGAGACTAAGCAAATGGTTCATTTCCACCCCTTTGGCAATGTTAACTTTTACGAAATGGATTG
GAGTCTTAAGGGAGATTATGGGCTCATGATCCAGTCATCGCAAAAGAGGGAAGTCGATGGTATGTTTTTCACACCGGGAGC
GGCATTCAAATCAAAACATCTGAAGATGGGGTTCATTGGGAGAATATGGGCCGGGTATTTCCATCGTTGCCGGATTGGTG
AAGCAGTATGTTCCAGAAAAAGACGAAGATCATTTATGGGCGCCCGATATATGTTTTTATAACGGAATTTACTATCTGTAT
TATTCGAGCTCTACATTGGAAAAAACACTCGGGCTTGGAGATTTCTGGGAAATACCTGCTGTTGGATTTGATCACTGAGC
TTATGAAAGTACCAGCTTCGCTTGACCGTTTACCTACCAACAGCTGCTGCTTGATGCTGCAATGCGACAGGCTCTGAGG
CATTGCCTCCTTCATCGTCTCCTGGTACAGCTAGCAGAGCTGGAGCGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
CCGGTTTTTGGTATGCGACGCTACGTTAAATTAAACACATCTGCGCCTGCGGCTGCGGCTGCGGCTGCGGCTGCGGCTGCG
> pep1
MNPASTLSVKTTNKTNTTTLHCATVAATLTVASWAVDKSLNPQNTGETISISKYIYQFAGEILGSGYGIWGGEDSP
NKNGFRNDVILKALQELQVPVIRWPGGCFADEYRWRDGIGPREQRPIRVNTHWGVEPFNTGTFHEFFELVLENTAEVAG
NLGTSPSEQMAEWLYIYSNSNSTVYAERVNNGREEEWEFAVFVYNGRESWGCNGGTLPYEYTLNYRHFSTVFKAAGKRPK
LVASGDDDDTGGTTPSLSKNNIDIGHSYILPTDWSIKGAATGFEDKWELELTERYLDSSFVLAQTTGILKKNPEEG
NIGLYLDEWGTWYDAPETHNPNFLPLQYNTVRDATAVAAVVNLNIPHYNADLRMHANIAQJMNQLAGMILDNEMKLMTTYPHY
FKMYP1FQDATHIPLDIKQQRDSYAHKKTTPVPSASAAAKTPNINIVSLVNLNPNEAEVESIALQGIVKTIGGTTSKQ
MDAHNTDFKPNVINQPRALNSDYSISKNGKTLTVKLPKAVVVLQNK