BIOSEQ | Bioinformatics Activity

Using BLAST

BLAST (Basic Local Alignment Search Tool) is an online search tool provided by NCBI (National Center for Biotechnology Information). It allows you to “find regions of similarity between biological sequences” (nucleotide or protein). The NCBI maintains a huge database of biological sequences, which it compares the query sequences to in order to find the most similar ones. Using BLAST, you can input a gene sequence of interest and search entire genomic libraries for identical or similar sequences in a matter of seconds.

The amount of information on the BLAST website is a bit overwhelming — even for the scientists who use it on a frequent basis! You are not expected to know every detail of the BLAST program.

BLAST results have the following fields:

E value: The E value (expected value) is a number that describes how many times you would expect a match by chance in a database of that size. **The lower the E value is, the more significant the match.**

Percent Identity: The percent identity is a number that describes how similar the query sequence is to the target sequence (how many characters in each sequence are identical). The higher the percent identity is, the more significant the match.

Query Cover: The query cover is a number that describes how much of the query sequence is covered by the target sequence. If the target sequence in the database spans the whole query sequence, then the query cover is 100%. This tells us how long the sequences are, relative to each other.

FASTA format

FASTA format is used to represent either nucleotide or peptide sequences. The first line is a comment line, beginning with “>” and describing the sequence. All the following lines are the sequence, in plain text.

Example DNA sequence in FASTA format:

```
>gi|23423|ref|NM_23542.0| Homo sapiens protein
ATGAATCGATACGATAGCTAGCTATCGATGCA
GATCAGAGAGGGGCTTTAGCTAGCTAAGCTAG
```

Example protein sequence in FASTA format:

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

Part 1: Identify sequences with Blast

Sequences at http://ase.tufts.edu/chemistry/walt/sepa/activities/exampleSequences.txt

Identify unknown sequences

1. Navigate to the main BLAST page (https://blast.ncbi.nlm.nih.gov/Blast.cgi)
2. Select the appropriate type of BLAST for your sequence
3. Paste the first unknown sequence into the box (for this activity, you can ignore the search options)
4. Click on the "BLAST" button and wait for the results. BLAST is usually fairly quick for short sequences, but should still take a few seconds.
5. Once the results are displayed, notice there are three main headings: **Graphic Summary**, **Descriptions**, and **Alignments** (these may be expanded so you'll have to scroll down).
6. Use these results to answer the questions below.
7. Repeat steps 1-6 for the other unknown sequence. How is this sequence related to the first sequence?

Questions

1. In the Descriptions section, look at the top result, which should be the result with the highest score. Write down information about the best match:

   Description
   (no need to write the whole thing)

   E value

   Identity

   Query cover

2. Now scroll down to the Alignments heading. Look at the top result, which should be the same one. Look at the alignment between your query and the reference. Do you see any mismatches?

3. How can you judge whether this is a good match?

4. What is this gene? Google the name of the gene and write down something significant you learned about it.

Part 2: Investigating sets of sequences

Each of the following sets of sequences were obtained from a sequencing experiment. These sequences can be found in the exampleSequences.txt file.

For each experiment, answer these questions:

- What do these sequences have in common?
- What is your best guess about the original purpose of this experiment?

Set 1

```
>Sequence1a
GTAATGTACATAACATTAATGTAATAAAGA
>Sequence1b
ATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACC
>Sequence1c
ATGGACTAATGGCTAATCAGCCCATGCTCACACATA
```

Set 2

```
>Sequence2a
TTTGGTTGTTCGACGACGGATGCAGAGCTCAGGGAAGTGGGGACGTGTTTTGGCTATCCT
>Sequence2b
GCGATGCATCAGGATGCATCCTCTGATCTTAGGGTGGTACGAGAAAAATTGAAGAATGTA
>Sequence2c
GCGGTTCCACAAGACCCTGAGGCGCCTGGTGCCTGACTCGGACGTCCGGTTCCTCCTCTC
```

Set 3

```
>Sequence3a
TAACCTACGGGTGGCCGCAGTGGGGAATATTGCACAATGGACACAAGTCTGATGCAGCGACGCCG
CGTGGGGGATGAAGGCTTTCGGGTTGTAAACTCCTTTCAGTACAGAAGAAGCATTTTTGTGACGG
TATGTGCAGAAGAAGCGCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGCGCGAGCG
TTGTCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTTGCGCCTGCTGTG
>Sequence3b
TGTCCTACGGGGGGCTGCAGTGAGGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTCG
CGTGAAGGATGACTGTCTTATGGATTGTAAACTTCTTTTATACGGGAATAACAAGAGTCACGTGT
GGCTCCCTGCATGTACCGTATGAATAAGCATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACG
GAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGC
>Sequence3c
GGCCTACGGGGGGCTGCAGTGGGTACGGGCAGACTAGAGTGTGGTAGGGGTAATTGGAATTCCTG
GTGTAGCGGTGGAATGCGCAGATATCAGGAGGAACACCGATGGCGAAGGCAGGTTACTGGCCAT
TACTGACGCTGAGGAGCGAAAGCGTGGGTAGCGAACAGGATTAGATACCCTAGTAGTCT
```