

## CS195 HW1

*Due: Thursday, February 14th 2:30pm*

### Problem 0

Throughout the class, you will be required to use NCBI's Basic Local Alignment Search Tool (BLAST). You can find BLAST online at: <http://www.ncbi.nlm.nih.gov/blast/>

BLAST finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

In this homework, you will encounter problems that ask you perform various searches using BLAST. BLAST has many options, depending on whether you want to search using a protein sequence or a nucleotide sequence for either matching protein or nucleotide sequences. In order to perform a search, you click on the type of search you would like to perform, for example, a protein-protein search (**protein blast**). You will see a large box where you can input the protein or nucleotide sequence at the top - this is the most important step. You will see many other options below, one of which is "organism," which allows you to enter the name of an organism, or select one of the organisms from a drop-down menu.

Once you get to the results page, you will see some reference information at the top, followed by a colorful distribution of hits. Below that you will see a list of sequences producing significant alignments. You can click on the sequence identifier to get more information about the match. If you want to see how the scoring was done, you can also click on the score.

BLAST has many other useful features, so we would recommend playing around a little bit, and trying out some of the other options. There is also a tutorial website you can visit: <http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>

### Problem 1

Please read the following:

- *How to Lie with Statistics* Chapter 1 (handed out in class)
- *A structure for Deoxyribose Nucleic Acid* by Watson and Crick, which can be found on the course web site: <http://www.cs.brown.edu/courses/cs196-1/wc-paper.html>

## Problem 2

BRCA1 is a breast cancer susceptibility gene that was first identified in 1994. People carrying a mutation (abnormality) in this gene are at an increased risk of breast or ovarian cancer. The normal gene plays a role in repairing breaks in DNA. However, when the gene is mutated it is thought that this repair function may become disabled thus leading to more DNA replication errors and cancerous growth.

Report the BLAST score for DNA and protein similarity between genes in the following organisms and the BRCA1 gene.

- *Bos taurus*
- *Rattus norvegicus*
- *Drosophila melanogaster*
- *Mus musculus*
- *Xenopus laevis*

### DNA alignment

NCBI (the National Center for Biotechnology Information), who hosts the BLAST tool we're using, has many other tools and databases as well. If you go to the NCBI homepage (<http://www.ncbi.nlm.nih.gov/>) you can type BRCA1 in the search box at the top, set the type of search to "Nucleotide", and click Go: this will search the *Entrez Nucleotide* database. Unfortunately, there are much too many results to browse: many organisms, each with possibly hundreds of variants (homo sapiens alone has 106). So, we'll give you a shortcut:

<http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=U14680>. The third line says the accession identifier is "U14680": this is a unique id for this entry in the database, which the BLAST tool understands. So, when using BLAST, we can just type this (instead of the whole sequence) in the "Search" box.

**NB. When using BLAST for nucleotide searches, you have to change the database from "Human genomic plus transcript" to "nr".**

### Protein alignment

From the NCBI homepage, you can type BRCA1 in the search box at the top, set the type of search to "Protein", and click Go: this will search the *Entrez Protein* database. The sixth result is labeled "BRCA1 [Homo sapiens]", which looks like what we want. Click its link ([AAC37594](#)), and it'll bring up a page with detailed information. A few lines down says the title of this entry is "Complete genomic sequence and analysis of 117 kb of human DNA containing the gene BRCA1", which tells us we're really in the right place (there are entries for genes with mutations as well as normal genes; we have to read and be careful). The third line says the accession identifier is "AAC37594": as above, when using BLAST, we can just type this instead of the whole sequence.

### Problem 3

Define a similarity matrix that can be used by the Dynamic Programming algorithm discussed in class to determine the longest common subsequence of two words, *with gaps*. For example, for the sequences:

$$\begin{array}{rcccccccc} X & = & A & C & G & C & T & G & T \\ Y & = & A & T & C & T & G & C & T & C & G & A \end{array}$$

the longest common subsequence (with gaps) is ACGCTG.

### Problem 4

For each of the following pairs of strings, (a) omit the letters that do not denote amino acids, (b) determine which has the greatest score BLAST score, and (c) find which has a match with a longer string length. Use the Protein search for *short, nearly exact matches* over the **nr** database.

- “computerscience” versus “biology”
- “protein” versus “aminoacid”
- “dynamicprogramming” versus “divideandconquer”

### Problem 5

Which of the following problems do you think could be solved using an alignment-based approach and why?

1. A lost book from the 16th century has been discovered. We would like to determine if it was written by William Shakespeare.
2. To highlight distinctive phrases within a book, the online retailer Amazon.com, analyzes the text and extracts *statistically improbable phrases*. These phrases commonly occur within a given book, but much less frequently within most other books. For example:
  - *Jurassic Park*: duckbilled hadrosaurs, tyrannosaur roared, soft hooting cry
  - *The Great Gatsby*: old sport
  - *The Hitchhiker’s Guide to the Galaxy*: new hyperspace bypass, large friendly letters, herring sandwiches

You are tasked with developing a similar algorithm.

3. Translate Jupiter’s slang in the Gold-Bug into plain English.

4. Decoding the cipher from the Gold-Bug.
5. Two students may have violated the collaboration policy by turning in term papers which have largely the same content. Analyze the two papers for similarity.

## Problem 6

In making similarity matrices for specific purposes, one may use the three building blocks (match score, mismatch score, and gap score) in different ways to accomplish the same goal. Give two similarity matrices neither of which allow any gaps and both of which prioritize matches over mismatches. Explain the side effects of each.