CS1820 HW1

Due: Thursday, Feb. 6 11:59pm

Problem 0

This problem does not require a written solution. Throughout the class, you will be required to use NCBI's Basic Local Alignment Search Tool (BLAST). You can find BLAST online at: `http://blast.ncbi.nlm.nih.gov/Blast.cgi`

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 "database," which allows you to enter the name of an organism, or select one of the organisms from a drop-down menu. Another option of significance is the "Program Selection": megablast (in blastn) will only return strong alignments, discontiguous megablast allows for a less strong alignment, and so on.

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, as you will learn. It is best to be comfortable with this tool early. The three tools you will use the most in the class are the alignment tools `blastn` (Nucleotide BLAST) and `blastp` (Protein BLAST), and then the protein structure database `PDB` (`http://pdb.rcsb.org/pdb/home/home.do`). We will study BLAST in detail in future lectures but the BLAST help page remains a useful reference for how to run and interpret BLAST (should not be required to finish this homework): `http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs`.

Problem 1

Please read the following and be prepared to discuss in class:

- *How to Lie with Statistics* by Huff and Geis [2010], Chapters 1,2 (available online)
- *A structure for Deoxyribose Nucleic Acid* by WATSON and CRICK [1953], which can be found on the course web site: `http://www.cs.brown.edu/courses/csci1820/handouts/watsoncrick.pdf`. Curiously, this DNA structure may be of "considerable biological interest". "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." Hmm...

(Optional) For deeper computer science literature on the method of dynamic programming, visit `http://www.cs.berkeley.edu/~vazirani/algorithms/chap6.pdf` (Dasgupta et al. [2008]).

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 `blastn` score for DNA **and** `blastp` protein similarity between genes and the BRCA1 gene in the following organisms. You need only report the top two BLAST hits each time.

- 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. Go to the NCBI homepage (`http://www.ncbi.nlm.nih.gov/`). At the top, where it says "All Databases" select Nucleotide for database and enter BRCA1 in the blank. Unfortunately, there are much too many results to browse: many organisms, each with possibly hundreds of variants (homo sapiens alone has 1337). So, we'll give you a shortcut: `http://www.ncbi.nlm.nih.gov/nuccore/555931`. 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 "Enter accession

number(s), gi(s), or FASTA sequence(s)" box.
NB. When using BLAST for nucleotide searches, you have to change the database to "Nucleotide collection (nr/nt)".

Protein alignment

From the NCBI homepage, you can type `BRCA1` in the search box at the top, set the database of the search to "Protein", and click Search this will search the *Entrez Protein* database. The tenth 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

A **substring** is a contiguous segment of a string, such as "mad" in "madman".

A **subsequence** has a subset of the characters from the parental string, and preserves order, but may omit characters in between. Example: "mama" in "madman"

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. For example, for the sequences:

X = A C G C T G T
Y = A T C T G C T C G A

the longest common subsequence 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 BLAST score (max score entry in the BLAST results page), and (c) find which has a match with a longer string length. For (b,c) use the Protein search for *short, nearly exact matches* over the **nr** database. **nr** refers to the non-redundant database – it contains all known proteins in all species *exactly once*.

- "computerscience" versus "biology"

- "protein" versus "aminoacid"
- "dynamicprogramming" versus "divideandconquer"

Problem 5

Which of the following problems do you think could be solved in a reasonable amount of time using an alignment-based approach? Explain why, but in no more than two sentences each.

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 OPTIONAL Extra-credit 3pts

Use NCBI tools to perform a multiple alignment of the proteins, and document your method: TPNVSVVDLTVRLGKG, LEKPAKYDDIK, LDDDVTESDVNAA, KGASYEDVKAA, DVSVVDLTV, LTCRLEKPAKY, NKETTYDEIKKV. For this problem, it will be helpful to know what a FASTA file is. FASTA is a text-based format for defining genomic sequences where the first line starts with a ">" character and represents a comment line and subsequent lines are the genomic sequences (coded in the usual way).

References

Sanjoy Dasgupta, Christos H. Papadimitriou, and Umesh V. Vazirani. *Algorithms*. McGraw-Hill, 2008.

D. Huff and I. Geis. *How to Lie with Statistics*. W. W. Norton, 2010.

J. D. WATSON and F. H. CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.