Homework 3

Dieu My Nguyen | MCDB 5520 | Mar 09, 2018

1. (a) 1pt What is the advantage of a seeded method like BLAST compared to a Needleman-Wunsch alignment?

The Needleman-Wunsch algorithm performs global sequence alignment, most useful when sequences are similar and of roughly equal size. This dynamic programming method is less efficient than word methods such as BLAST. BLAST is useful in large-scale database searches for optical local alignments to a query. It breaks the query and database sequences into fragments and seeks matches between them, whereas dynamic programming methods like the Needleman-Wunsch performed full alignment and are therefore slower than BLAST. BLAST also offers minimal sacrifice of sensitivity to distant relationships between sequences.

(b) 3pt Which flavor of BLAST is most sensitive for comparing your sequence to a species that are NOT closely related? Explain in one sentence.

Position-Specific Iterative BLAST (PSI-BLAST) (blastpgp) is usually more sensitive in finding distant relatives of a protein. PSI-BLAST uses a score threshold to derive a position-specific scoring matrix, which is used to further search the database for new matches and is updated for subsequent iterations with newly detected sequences. (Source: https://www.ncbi.nlm.nih.gov/books/NBK2590/ (https://www.ncbi.nlm.nih.gov/books/NBK2590/))

(c) 3pt In Karlin-Altshul statistics, what are four main assumptions?

1. A positive score must be possible
2. The expected score is negative
3. The letters of the sequences are independent and identically distributed
4. The sequences are infinitely long

(Also 5. The alignments do not contain gap)
(Source: http://www.utdallas.edu/~prr105020/biol6385/2018/lecture/Stat_sig.pdf (http://www.utdallas.edu/~prr105020/biol6385/2018/lecture/Stat_sig.pdf))

(d) 3pt Describe a scenario where BLAST will be incapable of finding a high quality match, even when one exists in the database. To get full credit, you must specify the word hit length utilized by BLAST in your scenario AND state the actual percent identity of your query to the best sequence in the database (which BLAST is unable to find).

BLAST, under standard settings, will be incapable of finding the high quality of match in case of searching short sequences (<20 bases), such as primers and small RNAs. Standard nucleotide BLAST uses word size of 11-- too high for short sequences as BLAST requires an uninterrupted stretch of 11 identical nucleotides to align 2 sequences. To work with short sequences, we would have to decrease the word size (some recommend 7) to increase sensitivity. The percent identity of the query to the best sequence in the database would be low (<50%).

2. You will need the Hmwk3.fa file on Canvas for this question. This fasta file contains a segment of an archaeal genome that is not quite finished. The goal is to analyze this segment.

Using NCBI Blast: http://blast.ncbi.nlm.nih.gov/Blast.cgi (http://blast.ncbi.nlm.nih.gov/Blast.cgi) follow the directions below and answer the following questions. The documentation for NCBI Blast is available at: ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf

You decide to use BLASTX (select Blastx from the front page). Search only the first 6000 nucleotides (query subrange 1-6000) against the nonredundant (nr) database. You are only interested in the top hits, so set the "Max target sequences" parameter to 50 and word size to 6. (Note that these are "Algorithm Parameters" and can be adjusted by expanding the form at the bottom.)

(a) 2 pt What species do you think this sequence originated from? Why?

Ignisphaera aggregans. Its type III restriction protein res subunit is the best alignment for our unknown segment, with E value of 0 and query cover 27%.

(b) 2 pt What fraction of the query is included in the best alignment? Is this hit a complete gene?

27% query cover. This hit is a subunit of the type III restriction protein, not a complete gene.

(c) 2 pt What coordinates of the query match and on which strand in the best alignment? (Note that you must view the actual alignment to answer this question.)

Coordinates of the hit "type III restriction protein res subunit [Ignisphaera aggregans DSM 17230]" are 5548 and 3881. These query coordinated are inverted, matching the minus strand for this alignment as reported in the Frame number of -3.

type III restriction protein res subunit [Ignisphaera aggregans DSM 17230]
Sequence ID: ADM27872.1 Length: 556 Number of Matches: 1

Range 1: 1 to 556 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
1046 bits(2705)	0.0	Compositional matrix adjust.	556/556(100%)	556/556(100%)	0/556(0%)	-3

```
Query  5548  MSRIIFGTDEWIDEDDFRLFLKFSRYLGRENGFSRFMVDFNKLSESIRSGSLSPNEVLDL  5369
             MSRIIFGTDEWIDEDDFRLFLKFSRYLGRENGFSRFMVDFNKLSESIRSGSLSPNEVLDL
Sbjct  1     MSRIIFGTDEWIDEDDFRLFLKFSRYLGRENGFSRFMVDFNKLSESIRSGSLSPNEVLDL  60

Query  5368  VEGYSVEFEEGSIEDLKKLLEEYMPRVIIRRVSNDIVIIPKVFLGDLVKDFREKGILIYD  5189
             VEGYSVEFEEGSIEDLKKLLEEYMPRVIIRRVSNDIVIIPKVFLGDLVKDFREKGILIYD
Sbjct  61    VEGYSVEFEEGSIEDLKKLLEEYMPRVIIRRVSNDIVIIPKVFLGDLVKDFREKGILIYD  120

Query  5188  KNNKWFKLVKPMYLFDVLDSLRKRNIVIHSEIDIKERIELPIKVMFKGDLRDYQQEALEA  5009
             KNNKWFKLVKPMYLFDVLDSLRKRNIVIHSEIDIKERIELPIKVMFKGDLRDYQQEALEA
Sbjct  121   KNNKWFKLVKPMYLFDVLDSLRKRNIVIHSEIDIKERIELPIKVMFKGDLRDYQQEALEA  180

Query  5008  WRKNSYRGIIALPTGTGKTiiaiaaiAELSEKTLIVTFTKEQMFHWAEKIVSFTDIPRSM  4829
             WRKNSYRGIIALPTGTGKTIIAIAAIAELSEKTLIVTFTKEQMFHWAEKIVSFTDIPRSM
Sbjct  181   WRKNSYRGIIALPTGTGKTIIAIAAIAELSEKTLIVTFTKEQMFHWAEKIVSFTDIPRSM  240

Query  4828  IGYYYGSEKRIAPITITTYQSAFRYVSSLSPYFSFLIIDEVHHLPADKFRYIAENMFARK  4649
             IGYYYGSEKRIAPITITTYQSAFRYVSSLSPYFSFLIIDEVHHLPADKFRYIAENMFARK
Sbjct  241   IGYYYGSEKRIAPITITTYQSAFRYVSSLSPYFSFLIIDEVHHLPADKFRYIAENMFARK  300

Query  4648  RLGLSATVIREDGRHVDLFPLMGGIVYSKSVSELAEKGYIAPFTVKTIKVSltkeekeky  4469
             RLGLSATVIREDGRHVDLFPLMGGIVYSKSVSELAEKGYIAPFTVKTIKVSLTKEEKEKY
Sbjct  301   RLGLSATVIREDGRHVDLFPLMGGIVYSKSVSELAEKGYIAPFTVKTIKVSLTKEEKEKY  360

Query  4468  rkllekykkLAGGREFQTLLEDAKRGDVAALEALKTRAEIRSLVHNAKEKIEALKAIVNR  4289
             RKLLEKYKKLAGGREFQTLLEDAKRGDVAALEALKTRAEIRSLVHNAKEKIEALKAIVNR
Sbjct  361   RKLLEKYKKLAGGREFQTLLEDAKRGDVAALEALKTRAEIRSLVHNAKEKIEALKAIVNR  420

Query  4288  ELENNSKIIVFTQYIEQAEKLAEILNTVYITGELDEDTrrrRLEMFKNNMVKIIVLTTVG  4109
             ELENNSKIIVFTQYIEQAEKLAEILNTVYITGELDEDTRRRRLEMFKNNMVKIIVLTTVG
Sbjct  421   ELENNSKIIVFTQYIEQAEKLAEILNTVYITGELDEDTRRRRLEMFKNNMVKIIVLTTVG  480

Query  4108  DEGIDIPDANVGIIFAGTGSRRQFIQRLGRLLRPMPGKEARLYEIIVKGTFEEAEARKRK  3929
             DEGIDIPDANVGIIFAGTGSRRQFIQRLGRLLRPMPGKEARLYEIIVKGTFEEAEARKRK
Sbjct  481   DEGIDIPDANVGIIFAGTGSRRQFIQRLGRLLRPMPGKEARLYEIIVKGTFEEAEARKRK  540

Query  3928  KALEEVFEGITIMSEE  3881
             KALEEVFEGITIMSEE
Sbjct  541   KALEEVFEGITIMSEE  556
```

(d) 1pt Looking at the "Taxonomy Report" (find by first clicking on the "Taxonomy Reports" link at the top of the results), how many hits were observed to Thermoprotei?

67 hits.

(e) 3pt In total, How many genes are in the first 6000 bases of this sequence? For full credit, give the names of each gene.

There are 4 genes total in the first 6000 bases. In the figure of the distrution of the top hits, there are hits along the sequence from 0 to ~1100, from ~1200 to ~2700, one hit from ~3000 to ~3900, and several hits from ~4000 to 5500.

3. You are given the protein sequence for Yfg1 and asked to identify distinct domains within the protein. As a first step you use BLAST to search Yfp1 against the non-redundant database (nr) and the top hit is as follows:

Name: insulin precursor
Score: 320
Query: 100%
E-value: 1e-95
Ident: 100%
Accession: NP_001191615.1

Observing that your best hit was to a RefSeq entry, you re-run your search against the RefSeq database (all parameters exactly the same, only changing the database utilized) and obtain the following top hit:

Name: insulin precursor
Score: 320
Query: 100%
E-value: 9e-111
Ident: 100%
Accession: NP_001191615.1

(a) 3 pt How did you know the hit was to a RefSeq gene?

RefSeq accessions are written in the format of NP_xxxxxx. For this protein sequence for Yfg1, the accession is NP_001191615.1, indicating a RefSeq protein accession.

(b) 7pt Give what you know about how E-values are calculated, why has the E-value changed between the two searches?

The E-value is a parameter represents the expected number of hits we can expect to see by chance when searching a database of a partular size. It decreases exponentially as the score of the match increases. Here, the scores from the 2 searches are the same (320), so this is not why the E-values differ. E-values are also calculated from the length of the query, but if we input the exact same sequence to search, the size of the query is the same and is not the factor to the change in E-value. E-values also decrease as alignments get longer and increase as database gets bigger. Thus, the size of the RefSeq database is probably smaller than the original, making a particular score less easily obtained by chance.

4. 10 pt Explain the difference between an algorithm and a scoring scheme. Use an example from class to clarify your definition.

The scoring scheme is a way for us to evaluate the goodness of an alignment. The scheme consists of character substitution scores for each character replacement and penalties for gaps, which can be one of several types such as linear and affine. The alignment score is the sum of the substitution scores and gap penalties. This score reflects goodness of alignment. One example scoring scheme for proteins is the PAM matrix. The substitution scores of this matrix are dervied from the analysis of known alignments of evolutionarily related proteins.

Sequence alignment algorithms give us ways to arrange sequences to identify regions of similarity that may suggest some relationship. An example is the Needleman-Wunsch algorithm, which consists of 3 steps: 1) Initialization of the score and traceback matrices, 2) Calculation of scores using a selected scoring scheme and filling in the score and traceback matriced, 3) Inferring the alignment from the traceback matrix.

5. Suppose that you are worried that you might have a rare disease. You decide to get tested, and suppose that the testing methods for this disease are correct 99 percent of the time (in other words, if you have the disease, it shows that you do with 99 percent probability, and if you don't have the disease, it shows that you do not with 99 percent probability). Suppose this disease is actually quite rare, occurring randomly in the general population in only one of

every 10,000 people.

10 pt You obtain a positive test result. What is the probability that you have the disease? [Hint: Use Bayes Rule]

Assume that the disease is independently and identically distributed throughout the population, we use Bayes Rule:

$$\begin{aligned}
&\text{A} = \text{have disease}\\
&\text{B} = \text{test positive}\\
&P(A) = \frac{1}{10,000}\\
&P(B|A) = 0.99
\end{aligned}$$

To derive P(B), we condition on whether A occurs:

$$\begin{aligned}
P(B) &= P(B|A) * P(A) + P(B|\text{not A}) * P(\text{not A})\\
&= (0.99)\left(\frac{1}{10,000}\right) + (1 - 0.99)\left(1 - \frac{1}{10,000}\right)\\
&= 0.010098
\end{aligned}$$

$$\begin{aligned}
P(A|B) &= \frac{P(A) * P(B|A)}{P(B)}\\
&= \frac{\frac{1}{10,000} * 0.99}{0.010098}\\
&= 0.0098
\end{aligned}$$

6. (Advanced) You are excited about being able to use the human genome browser (at UCSC) to look more closely at the molecular basis of human genetic diseases in the news. To start with, you decide to investigate one of the genes mentioned in the NY Times article "Disease Cause is Pinpointed with Genome" by Nicholas Wade (http://www.nytimes.com/2010/03/11/health/research/11gene.html?_r=0 (http://www.nytimes.com/2010/03/11/health/research/11gene.html?_r=0)).

One of the two papers described in this article has two authors "Lupski JR" and "Gibbs RA", and titled " Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy". [The article is downloadable on Canvas.]

Use the NYT article, the paper and genome.ucsc.edu human genome browser (version NCBI36/hg18) to answer the following questions.

(a) 1pt What is the human gene name, and abbreviation (six letters) that was found to causative of Charcot-Marie-Tooth neuropathy in this subject's genome (you can use this gene name to find the gene quickly, use the "gene" window).

SH3TC2

(b) 2pt How many exons does this gene have (use "RefSeq Genes" track, or dark blue "UCSC Genes" track)? What is the "genomic size" (full length, including exons and introns)?

There are 17 exons.
Genomic size: 81025 base pairs.

(c) 1pt Which DNA sequencing technology was used?

Sequencing by Oligonucleotide Ligation and Detection (SOLiD), a next-generation sequencing technology that involves ligation-based sequencing and 2-base encoding method. Fluoresent dyes are used to tag combinations of dinucleotides.

(d) 1pt The paper describes following two independent mutations through the extended family, and showed only those who inherited both mutations had the disease. What are these mutations? (i.e. Q340R)

Single-base variants (single-nucleotide polymorphisms, SNPs) and copy-number variants.

(e) 1pt For family members with only one bad allele (haploinsufficiency), what were two typical symptoms?

Median-nerve mononeuropathy at the wrist associated with evidence of a more widespread axonal polyneuropathy, and carpal tunnel syndrome.

(f) 1pt Using coordinates and/or protein sequence from Figure 2 from the paper find the position in the UCSC genome browser of the mutation that normally codes for Tyrosine. Figure 2C gives this alignment, but it does NOT mention that there are two species that have the precise mutation variant responsible for this disease. What are these two species? (Hint: in Multiz alignment of 44-vertebrates, click on the settings bar (grey vertical bar on left), and select "+" at the top to select and see all species in the Multiz alignment track. Another hint: note that the gene is on the reverse/minus strand, so to "turn it around" with 5' end on the left, click on the "reverse" button just below the browser window (between the "configure" and "refresh" buttons).

The novel missense mutation Y169H is located at: chr5:149,042,716. Species: Mus musculus and Macaca mulatta.

(g) 1pt You want to develop a genetic test for this mutation, so you need to find the closest "SNP" (single nucleotide polymorphism). You notice there is a SNP in the "Simple Nucleotide Polymorphisms" track right next to your mutation. What is the dbSNP id # (starts with "rs"), and the chromosome coordinate (i.e. chrX:12345443).

dbSNP id # rs17722293; chr5:149042711-149042711

(h) 2pt You notice that this mutation is found in a relatively small exon. If you were to go looking in the largest exon for other genetic mutations, which exon would that be? Give the exon number and first five nucleotides (on the 5' end) of this exon.

Exon 17. Nucleotides: GATGC.