BME 110 Midterm Examination

May 10, 2011

Name:

(please print)

Directions: Please circle one answer for each question, unless the question specifies "circle all correct answers". You can use any resource that has been discussed in class to solve these problems (see Resource Links from the main class page). All questions are 2 points each unless otherwise noted. Do the easiest questions first, then go back to do the harder ones that might take more time or thought.

There is to be no communication during the exam (no email, no messaging, no talking, etc). At the end of the exam, please sign your name below the honor code statement.

1. The major source of new sequences in the protein databases each year is:
   a. Real proteins from next-generation peptide sequencing
   b. X-ray crystallography
   c. Predicted proteins from next-generation DNA sequencing
   d. Nanopore chromosome sequencing
   e. Mammalian protein predictions because they have alternative gene splicing

2. High-throughput sequencing is achieved by today’s technologies by
   a. Super-fast ORF/protein prediction
   b. Massively parallel sequencing of millions of very short sequences at once
   c. Being able to sequence an entire long chromosome in one piece
   d. Next-generation sequencing of huge clone libraries
   e. DNA computing chips

3. (True/False) We like to use DNA sequences in FASTA format because they have extensive annotation information and protein prediction information within them.

4. Using any database or resource discussed in class or in the textbook, tell me how many bacterial genomes have been sequenced as of today. You must give your information source (web site or database name) to get credit for your answer.

5. Using PubMed at NCBI, tell me how many publications Eric S. Lander was a co-author on in the journal “Nature” with the word “genome” in the title? (give your method/search query to get credit)
6. According to the UCSC Human Genome Browser (Feb 2009 version), how many different splice variants of the “Parkinson disease protein 7” (chromosome 1), are there in the UCSC Genes track?

7. You are interested in the splice form of this gene (Parkinson disease protein 7 in the UCSC Genes track) that has the longest 3’ UTR because you want to look for microRNA regulation. In order to prepare to do PCR on this splice form, what is the exact genomic size of this splice form (with the longest UTR?)

8. Looking more closely at the Multiz alignment with other species for this gene (human PARK7), which of these species have differences relative to the human genome in the four nucleotides 5’ of the start codon? (Circle all that apply)
   a. Rhesus
   b. Mouse
   c. Dog
   d. Elephant
   e. Opossum
   f. Rabbit

9. You’d like to study microRNAs that might regulate the PARK7 gene, but you need to be able to knock out genes and observe the changes in gene expression. You decide to use mouse as your experimental model. Using the Table Browser for mouse (July 2007), tell me how many total microRNA genes there are in the “miRNA” track.

10. You’d like to study the human RNase P protein 30 in the mouse as well, but you don’t know where it is located in the genome. Go to NCBI to retrieve the human protein sequence “NP_001098016.1”, and using any method you like, give me the chromosome, strand, and start coordinate of the mouse ortholog of this human protein (according to the UCSC Genes track). Describe how you got your answer (just tell me the order of tools and/or databases you used)

11. There has been much debate on the archaeal species / lineage that may have given rise to eukaryotes. A core protein for life in humans is the Sm-like protein (Genbank Accession NP_009011; encoded by the LSM6 gene in the human genome browser), which required for mRNA gene splicing. Using any tool or parameters that you think are most appropriate for a sensitive, distant search, tell me what archaeal species (and the gene) that is most similar to the human LSM6 gene. Tell me which tool you used, and any non-default parameters you set to get credit.
12. Given the e-value, query coverage, percent identity, or total score from your top hit in the last problem, what is the most important single piece of information allowing you to determine if this is indeed a homolog or not? Based on this information, do you believe this is an ortholog or an insignificant hit? (Note: Archaea do not do mRNA gene splicing)

13. (True/False) The definition of “ortholog” is two genes that have the same function.

14. Given the choice of searching a new sequence among closely related species with either the nucleotide sequence or predicted protein query, it is always better to:
   A. Search with BlastX to find potential protein hits missed due to incorrect ORF predictions
   B. Search with BlastP to use the single correct protein prediction
   C. Search with PSI-Blast to detect distant homology
   D. Search with MegaBLAST to find all hits in the database quickly
   E. Search with BlastN to avoid spurious hits due to protein repeats

15. (True/False) Codon usage frequencies is used by GeneMark to estimate phylogenetic distances between species.

16. True/False: BlastX derives its greater sensitivity relative to BlastP by using different, more refined BLOSUM scoring matrices on successive rounds of searching.

17. True/False: In a local alignment, protein matches that are 50% identical can safely be considered to be homologs.

18. True/False: If you need to align two proteins, end-to-end, a Smith-Waterman alignment is preferable to Blast because it guarantees the best global alignment for a given scoring scheme, whereas BlastP does not.

19. A Dot-plot is What is the most efficient way to find out if there are duplicated, inverted, or transposed large regions when comparing two microbial genomes?
   a. BlastN
   b. BlastP
   c. PSI-Blast
   d. MegaBlast (it’s fast)
   e. Dot-plot

Use your new skills with the archaeal genome database, the table browser and the Mobyle EMBOSS website to complete the following questions. You may want to say which tool(s) you used in case you get it wrong, we can try to give you partial credit.

Look up the species *Thermococcus kodakaraensis* in the archaeal genome browser, and…
13. Extract the DNA sequence for coordinates chr:10,000-12,000. What is the G+C content for this region of the genome?

14. Get the protein sequence for TK0404. Using BlastP, all default parameters, searching the NR (archaea only), give the gene name, E-value, and percent identity for the top hit to a protein in another species.

15. Based on the annotation for any “significant” hits to TK0404, tell me:
   a) Do you think this is a real protein? Why or why not.

   b) If you could update the annotation for this protein from its current description to add more information, what could you confidently change this to? Give your evidence in one sentence.
16. The gene TK0001 is a whopping 5,016 nt, which is very large for an archaeal protein coding gene. You want to find out if there might be some alternate small ORFs encoded within it. Grab the sequence for TK0001 and find the next longest ORF using any method discussed in class.

   (a) How long is the next longest ORF?

   (b) What reading frame is it in?

17. Use any method (analysis or database) to tell me all the Pfam domains which occur in TK0001.

18. (3pts) Use the table browser to get all the “sequences” (do not add any extra bases upstream or downstream!) in the track “Genbank ncRNAs” for the entire genome. Use a Mobyle EMBOSS tool to find the most common 4-letter word in all these sequences, and tell me how many times it appears.

19. What is the preferred codon for Glycine (G) in TK0001, and how many times does it occur in this protein?

I have neither given nor received help on this exam.

____________________________

Signature

Student ID number: