Name:

.

Bioinformatics Take Home Test #3

Due Date Monday 10/10/2014 before class

All questions worth 1pt

1. What does PAM stand for, and what does it mean?

2. Selection for function can preserve sequence similarity in the pairwise comparison of homologous proteins, across domains separated by how many years of independent evolution?
A. Thousands B. Millions C. Billions
D. Limited only by how long life has existed
E. All of the above

3. Command line versions of the BLAST programs are available for which platform?
A. Macs B. PCs C. Unix/Linux
D. All of the Above
E. Only Macs and PCs

4. When searching a database with a query sequence, which of the following is true regarding the E-value?
A. It is NOT proportional to the size of the databank and can be larger than 1.
B. It is NOT proportional to the size of the databank and canNOT be larger than 1.
C. It is proportional to the size of the databank and can be larger than 1.
D. It is proportional to the size of the databank and canNOT be larger than 1.

5. What is a Z-value?
A. Number of matches one can expect due to chance.
B. Probability of obtaining a match of that quality due to chance.
C. Number of standard deviations a match is above mean, generated by randomizing sequences.
D. The measure derived from primary sequence similarity divided by the length of the match.
E. A measure of how similar two secondary structures are.

6. True/False RNA alone CAN have catalytic activity, it does NOT need to collaborate with proteins to do so, and it is capable of doing more than providing specificity due to base pairing.

7. In a BLAST search, what does the filter for low-complexity do?
A. It allows retrieving of "Warning Sequences" that are part of the databank and alerts to the fact that a query is of low complexity.

B. It replaces regions of low complexity *in the databank* with the symbol for any residue.
C. It replaces regions of low complexity in the query sequence with the symbol for any residue.
D. None of the above.

8. Usually E values smaller than a certain threshold are considered to demonstrate homology. This threshold is usually about
A) about 10^{3}, B) about 1, C) about 10^{-4}, D) about 10^{-20}

9. If you want to do a BLAST search of the non-redundant database using a new catalytic RNA sequence as query, which is the BEST search program to use?
A) blastn, B) blastp, C) blastx, D) tblastx, E) PRSS, F) blastrna

10. If you load a multiple sequence FASTA formatted file into an alignment program and the program only recognizes a single sequence, what could have gone wrong?
A. the ">" signs at the beginning of the annotation line are not part of the ASCI code.
B. the program expects the sequences to be in the Genbank flat file format.
C. the text file used different end of line conventions than the alignment program.
D. the program expects the sequences to be in Clustal alignment format with the word "CLUSTAL" written in the header.
E. The program expects the sequences to be in ASN format.

11. Usually a Z values of which magnitude is considered to demonstrate homology?
A. Smaller than 10^{-4}
B. larger than 3
C. smaller than 3
D. This can only be determined by the distribution of alignment scores when shuffling the data

12. What is a GI number?
A. A unique number given to every submitted sequence. If the sequence is changed, it retains this number. This makes it easy to track changes that occurred to a sequence.
B. The Genomic Isoform number given to every type of enzyme, providing easy access to enzymes from different organisms with the same or similar function.
C. A unique number given to every submitted sequence. If the sequence is changed, it receives a new GI number.
D. A unique number given to every submitted sequence. If the sequence is changed, a suffix is added to the number. This makes it easy to track changes that occurred to a sequence.

13. When aligning two sequences that are about 75% identical, which of the following scoring matrices would be **most** appropriate:
(A) PAM 0.75
(B) PAM 1
(C) PAM 7.5
(D) PAM 25
(E) PAM 250

14. If you want to align two sequences that are about 35% identical, which of the following scoring matrices would be most appropriate:
(A) Blosum 65
(B) Blosum 35
(C) Blosum 50
(D) Blosum 80
(E) Blosum 90

15. A databank search is performed with each of a collection of 5000 genes, with the aim for an overall probability to identify a false positive of 5%. Using the Bonferroni correction, which E-value should be applied to each of the 5000 individual databank searches?

16. Using a random shuffling approach (PRSS) you find that two sequences have an E value (assuming 10000 comparisons) of 950. This
A) proves homology
B) disproves homology
C) proves sequence similarity, but not homology
D) does not exclude the possibility that the two sequences might be homologous

17. One databank search is done using FASTA with an amino acid sequence as query and the only reported match has an E-value of 0.000005. What does this mean for the homology of the two sequences?
A) This proves (beyond reasonable doubt) that the two sequences are homologs.
B) the target sequence is a candidate for a homologous sequence, but an E-value of this magnitude does not prove homology
C) this proves (beyond reasonable doubt) that the target sequence is not homologous to the query
D) None of the above

18. If BLAST returns a match with an E-value of 5.4 e-11, what is the probability that this match represent a false positive?
A) 0
B) $5.4\ 10^{-11}$
C) $5.4\ e^{-11}$
D) The rate of false positive cannot easily be estimated.

19. In the above example, what is the frequency of false negatives in the databank?
A) 0
B) $5.4\ e^{-11}$
C) $5.4\ 10^{-11}$
D) The rate of false negatives cannot easily be estimated.

20. What are two of the most commonly used scoring matrices for data bank searches and for aligning protein sequences?
A) GTR and Dayhoff Recoding
B) PAM and Blosum
C) Gonnet and JTT
D) none of the above, explain:

21. True/False A multiple sequence fasta file contains hyperlinks to the actual sequences.

22. One databank search is done using FASTA with an amino acid sequence as query and the only reported match has an E-value of 52, what does this mean for the homology of the two sequences?
A) An E-value of this magnitude does not prove homology, but the sequences may never-the-less be homologous.
B) this proves (beyond reasonable doubt) that the two sequences are NOT homologs.
C) this proves (beyond reasonable doubt) that the two sequences ARE homologs.
D) None of the above.

23. Some students still have difficulties to discriminate between the term homology (=shared ancestry) and significant similarity. Which of the following statements is correct:
A. All complex sequences that show significant similarity in a pairwise sequence comparison are homologous.
B. All homologous sequences show significant similarity in a pairwise sequence comparison.
C. Both of the above statements are correct

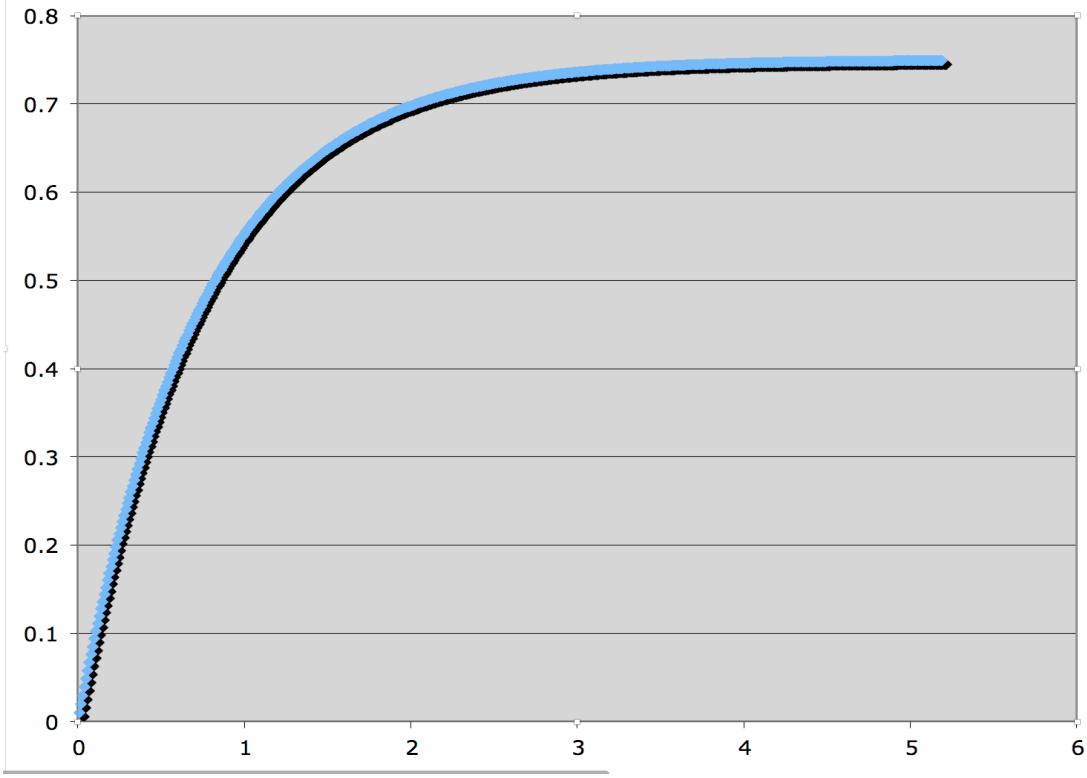
24. Comparing sequence A to sequence B obtains an alignment that matches sequences A and B over their whole length. The P-value for this alignment is $<10^{-13}$. Sequence B also has a significant match to sequence C ($P<10^{-9}$). You consider these P-values as sufficient proof for homology.
A. This shows that sequence A is homologous to sequence C
B. This is suggestive of homology between A and C, but to be sure you need to calculate the P-value for the match between A and C.
C. These findings cannot be used to infer homology between sequences A and C
D. None of the above.

25. When aligning two sequences that are about 20% identical, which of the following scoring matrices would be most appropriate?
(A) PAM 3
(B) PAM 9
(C) PAM 24
(D) PAM 210

26. A pairwise BLAST comparison is performed using two protein sequences that are known to have the same function in two different organisms, but no significant matches are reported.

What can be done to visualize the existing similarity?
A) Increase the "expect value".
B) Use the encoding nucleotide sequence.
C) Turn off the low complexity filter.
D) A + B; E) B + C; F) A + C

27. 2pt The following diagram gives the relation between the number of substitutions that have occurred during evolution (x axis) and the observed fraction of sequence differences. The depicted curve corresponds to the Jukes Cantor correction for a nucleotide sequence. This correction is only correct, if all sites have the same probability to undergo a substitution, and if all nucleotides occur with the same frequency.



a. Provide a rough sketch of how this relationship would change, if the different sites would have different substitution frequencies, and some sites would only very rarely undergo a substitution.

b. Using a different color, indicate how the curve would change, if the sequences have a strong compositional bias (e.g., 35% A, 35% T, 15% G, 15% C), but all have the same probability to undergo a substitution event. (If combinatorics is not your expertise, it might help to think about a sequence that only consists of As and Ts.)

Extra credit:

1. 1pt Describe a process that in your opinion goes beyond the simplest definition of natural selection (offspring similar to parents but random inherited variation, more offspring than necessary for replacement, selection due to limited resources).