Name (please write your name on eevery page):

Bioinformatics Midterm **All questions worth 1pt**

If not instructed otherwise, mark the single most correct alternative.

1. Which of the following is not included in bioinformatics?
    A. Using the primary sequence to detect homology
    B. Extracting useful information from biological data
    C. Modeling evolution of sequences in the computer
    D. Using twitter feeds to find trending biological studies
    E. Gene Linkage analysis where alleles are linked to disease

2. Which of the following are RNA nucleotides or modified RNA nucleotides?
    A. NAD and NADP
    B. GTP and ATP
    C. FMN and FAD
    D. All of the above

3. True/False - RNA cofactors provide support for Metabolism First hypotheses for the origin of life.

4. Which of the following is NOT true regarding proteins that evolved from the same ancestral protein?
    A. They may have significant similarity in their primary sequence.
    B. They can have different functions
    C. They could have diverged so much that they have only limited homology.
    D. They do not necessarily retain detectable similarity is primary sequence.
    E. They will ALWAYS be homologous, even if it is no longer detectable.

5. True/False - Sequence space is so big that stumbling onto a significantly similar sequence by chance is very unlikely, explaining why convergent evolution of primary sequence similarity has not yet been observed.

6. Which of the following IS considered strong evidence for homology?
    A. Several secondary structure elements are alignable with the Swiss Protein DataBank Viewer
    B. Significant primary sequence similarity
    C. Significant primary sequence identity

D. Two sequences both showing homology over 100% of their length to a third sequence.
E. All of the above

7. Which of the following is the BEST working definition of life?
A. Anything that can reproduce itself perfectly with no errors.
B. Cells. Life could only exist in a form compartmentalized by a lipid bilayer.
C. Self-sustained metabolic system that does not require input from any other living system.
D. A metabolizing system capable of evolution with heredity.
E. All of the above are equally good answers.

8. What is homology?
A. Percent of nucleotides or amino acids that are identical between two sequences.
B. Similarity due to shared ancestry, i.e. both got it from a common ancestor.
C. A difference found because of diverging evolutionary paths since the last common ancestor.
D. When two proteins share a function, such as nucleotide binding, they also share sequence similarities, because of the limited size of protein space.
E. Shared sequence similarity based on convergent evolution, i.e. the ancestor did not have it.

9. Can a protein be 45% homologous to another protein?
A. Yes.
B. Yes, if they share 45% sequence identity.
C. Mostly no, with the exception of cases of domain shuffling.
D. No, without exception.

10. The Swiss Protein data bank file viewer, aka Deep View CAN be used to do which of the following things?
A. Determine the ancestral sequence of two primary sequences.
B. Compare the structures of two sequences.
C. Color a structure by how closely it matches the ancestral sequence.
D. Detect homology with 100% certainty.
E. Determine the function of a hypothetical protein with no matches in the non-redundant database and no known homologs.

11. Which elements make up the secondary structure of proteins?
A. The four nucleotides, A, T, C, and G.

B. The 20 amino acids.
C. A cofactor and its binding site.
D. Alpha helices, beta sheets, and loops.
E. Multiple protein chains interacting to form one macromolecule.

12. Which structural elements are often represented as yellow arrows?
A. Beta sheets
B. Loops
C. Alpha helices
D. Beta arrows
E. Beta Barrels

13. True/False - Proteolipids are one of the subunits that form the hexamer (the head) of nucleotide binding subunits in the F1 ATPase?

14. True/False - In the catalytic cycle of the ATPase, the catalytic subunits work by moving through the different phases of the catalytic cycle in an asynchronous way, s that each of the subunits is a different phase of the catalytic cycle.

15. True/False - The 3 alpha and 3 beta subunits of the Bacterial type ATPsynthase and the 3 A and 3 B subunits in the Archaeal/Eukaryal ATPase are an example of convergent evolution, because the 3 and 3 subunit arrangement was settled on independently by these 2 enzymes.

16. True/False - The ancestral ATPase was most likely a heterodimer composed of one proteolipid subunit and one catalytic head subunit.

17. True/ False - Viruses, cellular parasites, and individual genes, depending on the chosen definition of life, could be considered alive?

18. True/False - Computer program exist that mimic evolution by means of artificial selection, and they are valuable tools because they are capable of finding new solutions that a human has never thought of before.

19. True / False – The Gaia hypothesis believes that Earth maintains itself in homeostatic balance by a series of runaway feed-forward loops, like glaciers leading to increased glacier formation.

20. Evolution by natural selection requires which four things to occur?

A. Variation among offspring, a niche, heredity, and competition for resources.
B. Heredity, variation among offspring, excess offspring, and a niche.
C. Excess offspring, competition for resources, heredity, and variation among offspring.
D. A niche, heredity, a human to naturally select the best offspring, and oxygen.
E. DNA, RNA, proteins, and lipids.

21. True/False - Having the same or similar function frequently occurs with homologs.

22. Which of the following explains how complex functional molecules were assembled, despite the vastness of protein space?
A. Protein space is made slightly smaller by removing all of the possibilities that cannot be synthesized or they will clog up the ribosome.
B. There are multiple unrelated solutions for the same functionality, exemplified by the fact that there are non-homologous enzymes inhabiting completely different regions of protein space with the same function.
C. An exact function does not need to be hit upon, because natural selection can take a protein with limited function and make it better.
D. Similar structures have similar function, so there are entire regions of protein space occupied by homologs that all function equally well, or nearly so.
E. All of the above.

23. How many peptides (short proteins) of 46 amino acids in length are possible, given that there are 20 possible amino acids? For your answer only consider the principles of combinatorics and ignore possible incompatibilities between amino acids). Give the formula; do not attempt the math.
20^{46}

24. Virus exhibit which of the following characteristics of life?
A. Heredity
B. Metabolism
C. Reproduction
D. Evolution
E. All of the above and more

Name:

25. How quickly does the size of Genbank double?

A. Approximately once every 1.5 years
B. Approximately once every 2-3 years
C. Approximately once every 5 years
D. It varies, depending on the latest sequence technology
E. None of the above

26. True/False - The E-value is proportional to the size of the databank.

27. If a BLAST search returns a match with an E-value of 1 e-20, and 200 searches were done simultaneously, how many false negatives are there?

A. $1 * e^{-20}$
B. $1*10^{-20}$
C. $1*10^{-20}/ 200$
D. There is no way to calculate the number of false negatives
E. None of the above

28. Why might % identity be high, when the E-value is insignificant?

A. The sequence is short
B. The complexity is low
C. The proteins have acquired different functions
D. A and B E. B and C F. A and C

29. What is coalescence?

A. The process through which Gaia regulates the Earth's climate.
B. The process exploited by parasites to ensure their survival over the host's.
C. The process of tracing lineages back in time to their common ancestor.
D. The process by which inteins splice themselves out of the host gene.
E. The process of determining the ancestral sequence of two or more homologs.

30. The first genetic material likely was similar to

A. DNA
B. Lipids
C. Proteins
D. RNA
E. Carbohydrates?

31. Two random nucleotide sequences with equal frequencies of A, G, T, and Cs without alignment have an average percent identity of 25%. How would the average

percent identity change, if the frequencies for the nucleotides are not equal. Use composition with X%G X%C and Y%A, Y%T.
$(2(X/100)^2 + 2(Y/100))^2*100\%$

32. Due to where saturation is reached, ___ can be used to look further back in time.
A. Gene presents/absence
B. Gene order
C. Nucleotides
D. Proteins
E. Functional studies

33. The universe and the earth are approximately how old, respectively?
A. 20 billion years old and 500 million years old
B. 14 billion years old and 4.5 billion years old
C. 16 billion years old and 7 billion years old
D. 2 billion years old and 450 million years old
E. 1 million years old and 4500 thousand years old

34. True / False - There is no way to determine the age of the Earth, because plate tectonics caused all of the rock on the planet to be reworked.

35. What are phylogenetic trees?
A. A species of tree famous for its inteins
B. A pictogram Darwin invented showing changing allele frequencies over time
C. Graphical depictions of how species are related to each other
D. A tool created by virologists to show which viruses can infect which hosts.
E. None of the above

36. True/False - BLINK, from NCBI, is a tool for BLASTing environmental sequences, such as the Sargasso Sea sequenced by Craig Venter.

37. Which of the following are NOT acquired traits capable of being passed on to offspring?
A. Symbiotic gut Bacteria
B. Epigenetic modifications
C. Large antlers on deer
D. Germ-line mutations
E. All of the above.

Name:

38. True/False - Entrez is so effective because it links to pre-computed searches.

39. True/False - Entrez covers many databases simultaneously AND does so quickly.

40. True/False - When inteins first begin to decay they lose their homing endonucleus domain first, while the other domain must stay functional to preserve function of the host proteins.

41. True/False - Among Site Rate Variation is caused by which of the following?

A. Viral infection
B. Certain amino acids or types of amino acids needed in catalytic sites and for structural interactions
C. The rate of evolution changing over time
D. Biases in AT and CG usage
E. Biases in Amino Acid usage

42. True/False - Due to Among Site Rate Variation, many protein sequences take a very long time to become saturated with substitutions.

43. Which of the following is NOT a Boolean operation used in NCBI/Entrez searches?

A. EQUAL
B. NOT
C. AND
D. OR
E. GREATER or LESS THAN

44. Inteins are composed of which of the following domains? **Choose 2.**

A. Self-splicing domain
B. Walker motif
C. Nucleotide binding domain (GRASP)
D. Hydrolase domain
E. Helix-turn-Helix DNA binding domain
F. Homing endonuclease domain

45. True/False - Ribozymes, catalytic RNA molecules, are one of the lines of evidence that supports the RNA World.

46. True/False - Homologous proteins can have different functions.

Name:

47. Why can two proteins, which lack significant sequence similarity, still be homologs?

A. They have the same function
B. There could have been too many mutational events in the protein disguising homology
C. They are found in the same bacteria
D. All of the above
E. None of the above

48. Which of the following databanks can determine if two sequences are 100% homologous?

A. BioProject (formerly Genome Project)
B. Bookshelf
C. Database of Genome Survey Sequences (dbGSS)
D. GenBank
E. Genome Reference Consortium (GRC)
F. NCBI Help Manual
G. Nucleotide Database
H. None of the above

49. What does the abbreviation NCBI stand for?

National Center for Biotechnology Information

50. True/False - Gut Bacteria are passed on to children in utero, and so are only passed from mother to child.

51. What is the Black queen hypothesis?

A. Selection for streamlined genomes will result in all members of a community producing only a subset of the required leaky goods.
B. Parasites are the first thing to evolve, after life springs up *de novo*.
C. Life is like an arms race, where all life forms have to run faster and faster just to stay in place.
D. DNA based organisms took over from the RNA world, after DNA was created by a virus in an act of genome warfare.
E. None of the above.

52. True/False - The Modern Synthesis ignores the significance of mutations themselves.

53. How might mutual aid be selected for?

A. Trick question: it cannot be selected for, because even if a stingy species is going extinct, it cannot decide to stop being stingy.
B. When a parasite enters a system, the host coops it to function in a way that benefits the host, so they both prosper.
C. Rival species kill each other, so that members of their own species have greater access to resources.
D. When cooperation between species results in more offspring for both, the entire community thrives and spreads.
E. B-D.

54. Assuming equal frequency of the different building blocks, two random protein sequences are on average _____ and nucleotide sequences are on average _____?

A. 20% identical and 5% identical.
B. 95% identical and 75% identical.
C. 20% identical and 40% identical.
D. 5% identical and 25% identical.
E. None of the above.

55. Which of the following features of life as we know it is inescapable and will surely be found in all alien life discovered?

A. DNA
B. Parasites
C. The central dogma (DNA →RNA →Proteins)
D. RNA
E. All of the above

56. True/False - Selection for function can preserve sequence similarity in the pairwise comparison of homologous proteins, across domains separated by billions of years.

57. True/False - Selection for function is a reason for Among Site Rate Variation.

58. True/False - The late heavy bombardment might not actually exist and could be nothing more than the tail in on the early heavy bombardment.

59. When searching a database with a query sequence, which of the following is true regarding the E-value?

A. It is proportional to the size of the databank and can be larger than 1.
B. It is proportional to the size of the databank and canNOT be larger than 1.

C. It is NOT proportional to the size of the databank and can be larger than 1.
D. It is NOT proportional to the size of the databank and canNOT be larger than 1.
E. It is the number of standard deviations a match is above mean, generated by randomizing sequences.

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

61. True/False - The late heavy bombardment occurred AFTER the impact that created the moon.

62. True/False - RNA alone canNOT have catalytic activity, it REQUIRES proteins cofactors to function.

63. In a BLAST search, what does the filter for low-complexity do?
A. It replaces regions of low complexity in the databank with the symbol for any residue.
B. It replaces regions of low complexity in the query sequence with the symbol for any residue.
C. It allows retrieving of "Warning Sequences" that are part of the databank and alerts to the fact that a query is of low complexity.
D. None of the above.

64. Usually E values smaller than a certain threshold are considered to demonstrate homology. This threshold is usually about
A) about 5, B) about 1, C) about 10^{-5}, D) about 10^{-25}

65. If you want to do a BLAST search of the non-redundant database using a newly sequenced protein coding gene sequence as query, which is the BEST search program to use?
A) blastn, B) blastp, C) blastx, D) tblastx, E) PRSS, F) blasttranslate

Name:

66. If you load a multiple sequence FASTA formatted file into an alignment program and the program only recognizes a single sequence, what has most likely gone wrong?

A. the text file used different end of line conventions than the alignment program.
B. the program expects the sequences to be in the Genbank flat file format.
C. the ">" signs at the beginning of the annotation line are not part of the ASCI code.
D. The program expects the sequences to be in ASN format.
E. the program expects the sequences to be in Clustal alignment format with the word "CLUSTAL" written in the header.

67. One databank search is done using FASTA with an amino acid sequence as query and the only reported match has an E-value of 10^{-36}. 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

68. What is a GI number?

A. 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.
B. A unique number given to every submitted sequence. If the sequence is changed, it receives a new GI number.
C. The Genomic Isoform number given to every type of enzyme, providing easy access to enzymes from different organisms with the same or similar function.
D. 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.

69. When aligning two sequences that are about 85% identical, which of the following scoring matrices would be **most** appropriate:

(A) PAM 0.85
(B) PAM 1

(C) PAM 8.5
(D) PAM 25
(E) PAM 850

70. If you want to align two sequences that are about 95% identical, which of the following scoring matrices would be most appropriate:
(A) Blosum 65
(B) Blosum 95
(C) Blosum 50
(D) Blosum 80

71. A databank search is performed with each of a collection of 10000 genes, with the aim for an overall probability to identify a false positive of 1%. Using the Bonferroni correction, which E-value should be applied to each of the 10000 individual databank searches?
$0.1/10000=10^{-6}$

72. If BLAST returns a match with an E-value of 2.4 e-11, what is the probability that this match represent a false positive?
A) 0
B) $2.4\ 10^{-11}$
C) $2.4\ e^{-11}$
D) The rate of false positive cannot easily be estimated.

73. Usually a Z values of which magnitude is considered to demonstrate homology?
A. Smaller than 10^{-5}
B. This can only be determined by the distribution of alignment scores when shuffling the data
C. larger than 3
D. smaller than 3

74. Using a random shuffling approach (PRSS) you find that two sequences have an E value (assuming 10000 comparisons) of 1950. 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

Name:

75. In the above example, what is the frequency of false negatives in the databank?

A) 0
B) 2.4 e^{-11}
C) 2.4 10^{-11}
D) The rate of false negatives cannot easily be estimated.

76. What are two commonly used scoring matrices for data bank searches and for aligning protein sequences?

A) GTR and Dayhoff Recoding
B) PAM and Blosum
C) Jukes Cantor 69
D) none of the above

77. One databank search is done using FASTA with an amino acid sequence as query and the only reported match has an E-value of $\mathbf{10^4}$, 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.

78. In a multiple sequence fasta file format, which character begins each new sequence?

A. * B. – C. # D. > E. ^

79. Which of the following statements is correct:

A. All homologous sequences show significant similarity in a pairwise sequence comparison.
B. All complex sequences that show significant similarity in a pairwise sequence comparison are homologous.
C. Both of the above statements are correct

80. 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 $<10^{-15}$. Sequence B also has a significant match to sequence C ($P<10^{-6}$).

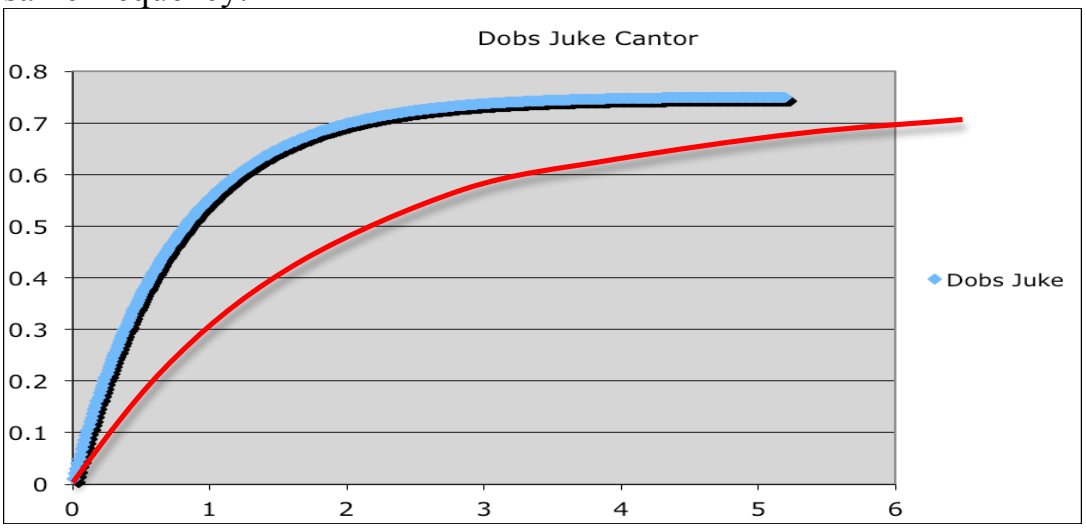
A. These findings cannot be used to infer homology between sequences A and C

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. This shows that sequence A is homologous to sequence C.
D. None of the above.

81. 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?

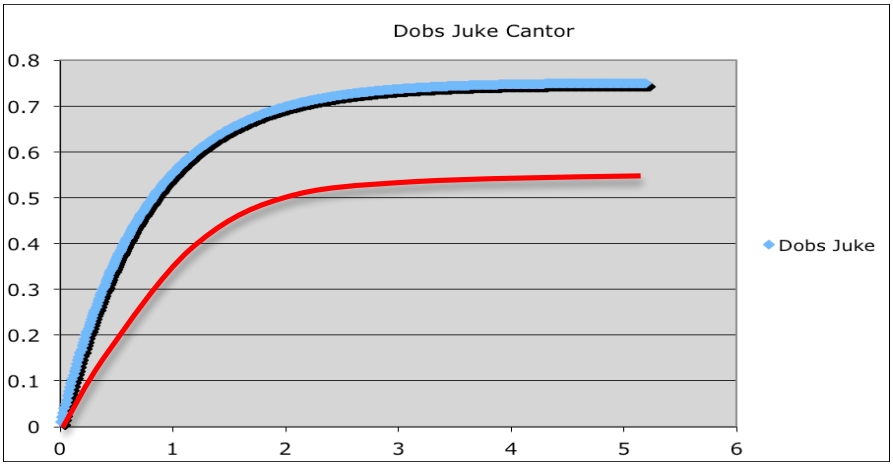
A) Use the encoding nucleotide sequence.
B) Increase the "expect value".
C) Turn off the low complexity filter.
D) A + B
E) B + C
F) A + C

82. 3pt 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.

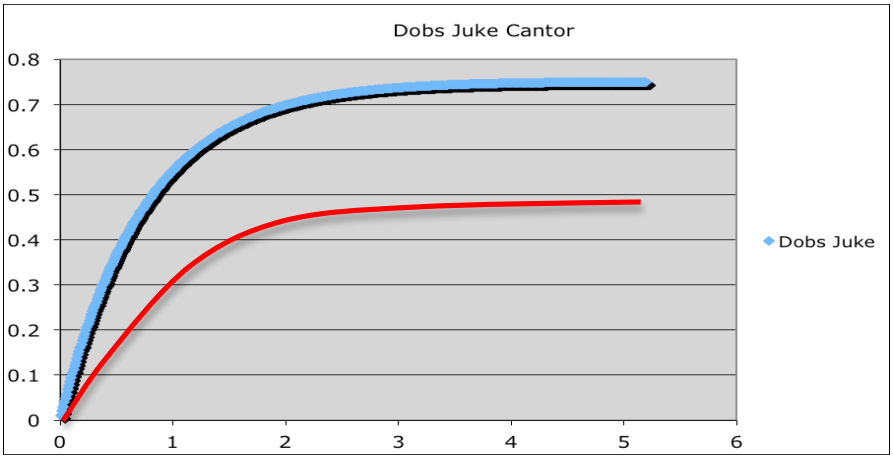


a. Provide a rough sketch of how this relationship would change, if the different sites would have different substitution frequencies, and some sites would only very rarely undergo a substitution.

Name:



b. Provide a rough sketch of how this relationship would change, if some sites never undergo a substitution event (for ease of drawing, assume 20% invariant sites).



c. Using a different color, indicate how the curve would change, if the sequences have a strong compositional bias (e.g., 50% A, 50% T, 0% G, 0% 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.)

Name:

83. If the following searches were conducted in PubMed for articles, what would the searches return? Please draw Venn diagrams to illustrate your answers (i.e. depict each of the individual searches as a circle). 2pts.
A. Inteins NOT ATPsynthase

B. Archaea OR Bacteria

C. Gogarten J AND ATPsynthase

D. (Gogarten JP AND Swithers K) NOT Inteins

84. Considering the combinatorics space given by all possible peptides (small proteins) with a length 10 amino acids (about 10^{13} different theoretical peptides) What is meant by the statement that this space is highly connected?

It only takes 10 substitutions to get from any one of these peptides to any other peptide.

85. Who first used tree-like diagrams to depict the evolution of organisms?
A) Darwin
B) Lamarck

C) Already the bible and the Aztecs used the tree of life imagery to depict the history of life

86. In plotting exponential growth or decay in a semi logarithmic fashion, how would you calculate the decay or growth constant?

A) the Y-axis intercept B) the slope C) both