Take Home Exam 3 Answers.

QUESTION 1

What Boolean operators can be used in NCBI/Entrez searches?

AND OR NOT

QUESTION 2

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 Evalue should be applied to each of the 5000 individual databank searches?

$0.05/5000=1*10^{-5}$

QUESTION 3

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

QUESTION 4

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

QUESTION 5

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).

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.

QUESTION 6

You do a data bank search with a single protein (amino acid sequence) as query. You obtain an E-value of 1.5 e-3.What is the probability that this match is a false positive?

0.0015

QUESTION 7

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.

QUESTION 8

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.

QUESTION 9

If multiple search terms are connected by Boolean operators without parenthesis, the NCB interface will start evaluation

A. from the left
B. from the right
C. Using a priority of AND over OR

QUESTION 10

If you want to align two sequences that are about 30% identical, which of the following scoring matrices would be most appropriate:

A. Blosum 25
B. Blosum 45
C. Blosum 60
D. Blosum 80

QUESTION 11

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

QUESTION 12

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.

QUESTION 13

In the unix operating system, which command would one use to enter a subdirectory?

A. ls
B. chmod
C. pwd
D. cd
E. qlogin

QUESTION 14

In the unix operating system, which command could one use to copy the content of two file into a single file?

A. ls
B. cat
C. pwd
D. cd
E. qlogin

QUESTION 15

In the unix operating system, which command would one use to check if a file is in the current directory?

A. ls
B. cat
C. pwd
D. cd
E. qlogin

QUESTION 16

In the unix operating system, which command would one use to display the content of a file on the screen?

A. ls
B. cat
C. pwd
D. cd
E. qlogin

QUESTION 17

One data bank search is done using FASTA with an amino acid sequence as query and the only reported match has an E-value of 1.8 E-30, what does this mean for the homology between the query and the target sequences?

A. An E-value of this magnitude is suggestive of homology, but further studies need to undertaken to prove homology beyond reasonable doubt.
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.

QUESTION 18

What was the name of the scientist who developped the PAM substitution matrix and who first used the single letter amino acid code?

Margaret Dayhoff

QUESTION 19

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

True **False**

QUESTION 20

In BLAST searches using only a single genome as target, proteins can have more than one match because of paralogy.

True False

QUESTION 21

True/False- The Modern Synthesis does not give any weight to the effects of mutations themselves.

True False

QUESTION 22

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

QUESTION 23

Usually E values smaller than a certain threshold are considered to demonstrate homology. This threshold is usually about

A. about 103,
B. about 1,
C. about 10^{-4}
D. about 10^{-20}

QUESTION 24

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

QUESTION 25

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:

QUESTION 26

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

Point accepted Mutation - Accepted point mutations per 100 residues. The PAM matrix (20x20) gives information on how frequently one amino acid is substituted by another one. The PAM001 matrix was calculated for very closely related sequences, separated by only one substitutions per 100 residues. While the matrix was calculated from closely related sequences, the PAM250 matrix (calculated from applying PAM001 250 times) is appropriate for very divergent sequences.

QUESTION 27

What does the abbreviation NCBI stand for?

National Center for Biotechnology Information

QUESTION 28

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.

QUESTION 29

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.

QUESTION 30

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

QUESTION 31

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

QUESTION 32

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.

QUESTION 33

_____ sequences reach saturation before _____ sequences reach saturation, so _____ sequences can be used to look further back in time.

A. Nucleotide, protein, nucleotide
B. Protein, nucleotide, nucleotide
C. Nucleotide, protein, protein
D. Protein, nucleotide, protein
E. None of the above

QUESTION 34

You do a Blast search and a match to a sequence in the database is listed with an E-value of 3.2 e-23.

A. This means that the number of matches expected due to chance is 3.2 x e^-23 (e= Euler's number and ^ stands for to the power of)
B. This means that the number of matches expected due to chance is 3.2 x 10^-23 (^ stands for "to the power of")
C. This means that the probability for matches of this quality due to chance is 3.2 x e^-23 (e= Euler's number and ^ stands for to the power of)
D. This means that the probability for matches of this quality due to chance is 3.2 x 10^-23 (^ stands for to the power of)
E. Both B and D are correct
F. Both A and C are correct

QUESTION 35

Describe at least two processes that could be considered to go 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). Do not feel restricted by the possibility that this process might not actually occur in nature. (If in doubt list more than 2 two processes).

Possible answers include: Fusion between two independent lineages to form a new organism (e.g., aneupolyploidization; uptake of endosymbionts during eukaryogenesis).

Non-random, targeted mutations.

The holobiont could evolve through the acquisition of symbionts (which might be described as Lamarckian evolution, in case the symbionts are passed on to the next generation, which seems to be the case with the human microbiom)

Others were accepted within reason.