Bioinformatics Take Home Test #2

(This is an open book exam based on the honors system -- you can use notes, lecture notes, online manuals, and text books. 

Teamwork is not allowed on the exams, write down your own answers, do not cut and paste from webpages. If your answer uses a citation, give the source of the quoted text.)

1. Given that two homologous sequences start off with 100% similarity and then diverge over time, what percent similarity will they share when saturation has been reached, (assume equal frequency for the different letters) 1pt  
   a. For nucleotide sequences?  
   b. For protein sequences?

2. Questions of time- 2 pts  
   a. How old is the universe (approximately)?  
   b. How old is planet Earth (approximately)?  
   c. How long has life inhabited the planet Earth (approximately)?  
   d. When did the Bacteria diverge from the Archaea and Eukaryotes, i.e. how old is LUCA (approximately)?

3. What is the late heavy bombardment? 50 words or less- 1pt

4. Which type of sequences can be used to look further back in time, nucleotide or protein? Give a short justification of your reasoning- 1pt

5. What is Among Site Rate Variation (ASRV) and how does it affect saturation in protein and nucleotide sequences? 2 pts

6. True or false- Entrez is so effective because it only uses a non redundant database. 1 pt

7. True/False- Entrez is so effective because it uses pre-computed links to other databank entries and links to the output of previously performed databank searches. 1 pt

8. True/False- The participation of nucleotide derived co-factors in many protein catalyzed reactions supports the RNA world hypothesis. 1 pt

9. True/False- The finding that the ribosomal RNA alone cannot perform translation is an argument against the RNA world hypothesis 1 pt
10. Give short definitions of- 2 pts
   a. mRNA:
   b. tRNA:
   c. rRNA:
   d. transcription:
   e. replication:
   f. translation:

11. What are inteins, introns, exons, exteins? 2 pts

12. True/False- When doing a search on the NCBI database, it is not possible to search for articles in PubMed written by J. P. Gogarten on the ATPsynthase and pull up relevant nucleotide sequences, protein sequences, and crystal structures at the same time. i.e. All of these databases must be search independently because they are not linked. 1pt

13. True/False- Inteins are molecular parasites that splice themselves out at the protein level. 1pt

14. Inteins are composed of two domains. What are they and what is their function? 2pts

15. True/False- When inteins first begin to decay they lose the protein-binding domain first, while the DNA binding domain must stay functional or it will destroy the function of the host proteins. 1pt

16. Which of the following are databases available through NCBI aka Entrez? Circle all that apply- 1pt
   a. BioProject (formerly Genome Project)
   b. Bookshelf
   c. Database of Genome Survey Sequences (dbGSS)
   d. GenBank
   e. Genome Reference Consortium (GRC)
   f. NCBI C++ Toolkit Manual
   g. NCBI Help Manual
   h. Nucleotide Database
   i. Protein Database
   j. PubMed Central (PMC)
   k. Taxonomy
   l. All of the above and many many more.

17. What Boolean operations can be used in NCBI/Entrez searches? 1pt

18. If the following searches were conducted in PubMed for articles, what would the searches return? Please draw ven diagrams to illustrate your answers (i.e. depict each of the individual searches as a circle). 2pts.
   a. Gogarten J NOT Gogarten JP

   b. Gogarten JP AND Doolittle WF

   c. Gogarten J OR ATPsynthase
d. (Gogarten JP OR Swithers K) AND Inteins

19. What does the abbreviation NCBI stand for and why is this site important in the field of bioinformatics? Limit your answer to 30 words or less. 1pt

20. There are two types of databanks: those with a gatekeeper and those without. What are the advantages and disadvantages to each? Limit your answer to 50 words or less. 1pt

21. What is BLINK (hint, it is from NCBI) and how is it useful? Limit your answer to 40 words or less. 1pt

22. What is BLAST (hint, it is from NCBI) and how is it useful? Limit your answer to 40 words or less. 1pt

23. What can be done with BLAST? If you find a significant hit with a BLAST what does that mean? 2 pts

24. Sequences that do not show significant similarity- 1pt
   A) are not homologous
   B) are homologs
   C) might never-the-less be homologs

Graduate questions- Short essays please.

25. How do intein population dynamics allow them to be retained in a population over millions of years? Why do they not simply decay and become extinct from the population once every member of the population is infected with the intein? 3pts

26. If protein space is so big, how is it that complex functional molecules were assembled? 3 pts

Extra credit question for all -- A little exercise in combinatorics:

1. In question 1 you assumed equal frequency of the different letters (the nucleotides AGCT or the 20 different amino acids). 3 pts
   a. How would the result for a nucleotide sequence change, if the frequencies for the two nucleotides are not equal. Use composition with 40%G 40%C and 10%A, 10%T as an example.
   b. How similar would two random sequences be with this composition, on average and before alignment?
   c. What is the general formula?