100 QUESTIONS FOR BIPG PHD PRELIMINARY EXAM
AND BIPG MSBS QUALIFYING EXAM

Every PhD student must pass a preliminary exam at the end of their first year (distinct from the qualifying exam). The preliminary exam will consist of a set of **20 questions** taken from the list below, representing **all seven categories** (A-G). Each question will be scored on a "pass/fail" basis, with the following results:

- Pass all 20 – you pass, and will be taken to lunch by two or more members of the BIPG faculty.
- Pass 16-19 (80% or more) – you have passed the exam.
- Pass 14-15 (70-75%) – you will be allowed to retake the exam with different questions.
- Pass 0-13 (<70%) – you will have to leave the program. This is because the questions are being provided well in advance, so strong performance is an expectation. [PhD students might still qualify for a masters in BIPG, and masters students for a certificate, if your course and laboratory performance have been sufficient.]

NOTES: a) Because you are given the QE questions one year in advance, MINIMAL answers will NOT be accepted. You must give complete answers. b) There is one "extra credit question". This is more complex, but there is no penalty associated with an incorrect answer (*i.e.*, it adds to the numerator of correct answers, but not to the denominator of total questions).

A. Research methods

1. What is the difference between a "selection" and a "screen"? To illustrate your answer, compare selection for growth on lactose to blue/white screening on XGal indicator agar.
2. What does it mean to say two genes are "orthologous", and what is the difference between "percent identity" and "percent similarity" (for amino acid sequences) in this context?
3. How are tissue- or cell type-specific transgenic and knock-out mice generated?
4. What is the polymerase chain reaction (PCR), what are its major research uses, and what can "real-time" PCR measure that standard PCR cannot?
5. Explain the terms "biological replication" and "technical replication".
6. What is meant by "genotyping"?
7. For transcriptomic studies, it is crucial to rapidly stabilize the RNA preparations. Why is this a problem for RNA but much less so for DNA? Why is it a worse problem in bacteria?
8. What is a restriction-modification system, and what is an example of how are they used in biomedical research?
9. What is a CRISPR-Cas system, and what is an example of how are they used in biomedical research?
10. What is the basic structure of an IgG-type antibody molecule? Label the chains, and indicate which part(s) determine its antigen specificity.
11. What are key differences between primary cell cultures and cell lines?
12. How does Sanger sequencing work, and what is a dideoxy nucleotide?
13. How does NextGen sequencing (such as Illumina) work?
14. In protein purification, what is an affinity tag? Give two examples, and explain how each of them is used.

B. Basic cell and molecular biology

1. Draw the structure of any one of the 20 standard amino acids, and then of a dipeptide (to show a peptide bond).
2. Draw the structure of any one of the four standard DNA mononucleotides, indicating the position that distinguishes DNA nucleotides from RNA nucleotides, and then draw a dinucleotide (to show a phosphodiester bond).
3. Why are DNA sequences usually shown in the 5'→3' orientation?

4. Introns were originally considered to be "junk DNA", but they can play several important roles. What roles are played by introns?
5. What are the major types and consequences of alternative splicing?
6. What is an allele?
7. What is epigenetics?
8. What is meant by the terms: dominant, recessive, epistasis, and mixed dominance?
9. What are the basic steps in getting from a gene to a protein in mammalian cells? How is this process different in bacterial cells?
10. What is "codon bias"? Does it have biological functions?
11. How are genes controlled (turned "on" and "off") in mammalian cells?
12. What are the major steps in cell division for mammalian cells?
13. Name 8-10 different types of noncoding RNA.
14. Briefly describe several evolutionary scenarios for the origin of novel genes.

C. Proteomics

1. Describe primary, secondary, tertiary, and quaternary protein structures and give examples.
2. What is "post-translational modification"? Give at least four examples.
3. How does mass spectroscopy work?
4. What are the benefits and disadvantages of MALDI and ESI relative to one another?
5. How do you extract peptide induced signals (*i.e.*, peaks) from raw MS spectra?
6. What is "collision-induced dissociation", and how does LC/MS/MS data give amino acid sequences of polypeptides?
7. What are the major expected difficulties in mass spectrometry data analysis?
8. What is the basis for separating polypeptides in standard two-dimensional gel electrophoresis? How are the protein identities of the 2D gel spots commonly determined?
9. Describe and compare the following methods of protein structure prediction: homology modeling, threading (fold comparison), and *ab initio* prediction.
10. What are the preprocessing issues in proteomic data analysis?
11. What are the differences between shotgun and multiple reaction monitoring (MRM) proteomic approaches?
12. Describe SILAC and iTRAQ labelling approaches and the major advantages and disadvantages compared to label free quantification.
13. In general terms, how do proteins destined to be embedded in a membrane differ from those destined to be free in the cytoplasm?
14. What is the single-letter code for each of the 20 standard amino acids?

D. Genomics

1. What is the difference between a gene, a genome, and a chromosome?
2. What are the approximate sizes of the human, fruit fly, Arabidopsis, and *E. coli* genomes? What are the sizes of the shortest and the longest known genomes among all living creatures?
3. What is the average size of a human exon and intron? What is the maximal number of introns in a human gene? What is the size (in kbp) of the longest human gene?
4. What is a "SNP"? Approximately how many SNPs are currently known for humans and mice?
5. Describe the composition of the human genome. What are the approximate percentages of coding sequences, introns, interspersed repeats, simple repeats, and satellites?
6. How are microarrays created? What comprises the probes on a microarray? What is a "tiling array"?
7. You are planning a microarray experiment comparing cancer cell lines and normal cell lines. What is the minimum number of samples of each line that should be evaluated in order to yield statistically significant results?

8. Why is the polymerase chain reaction (PCR) employed during sample preparation in most microarray experiments?
9. Describe a method for incorporation of a fluorescent probe into a sample during PCR reactions.
10. Identify and briefly describe the process by which nucleic acid sample target sequences bind to the probes on a microarray.
11. During a study utilizing microarrays, it was determined that the mRNA for a specific transcription factor was expressed at dramatically higher levels in diseased patients versus normal patients; however, mRNAs from genes regulated by this transcription factor showed no difference in expression. Explain how this could occur.
12. Why might you encounter false-positives in microarray data?
13. For publication of microarray studies in scientific journals, confirmatory experiments must be completed for the significant findings revealed by the microarray data. How, and why?
14. What are the advantages and limitations of using SAM (Significance Analysis of Microarrays) for microarray analysis?
15. What is the "EST database"?

E. Biostatistics

1. What are “Student’s t test” and “ANOVA”, and how are they used?
2. List the computational models for identifying a gene regulator network. Which is currently the best model? Briefly describe this model.
3. Explain the terms “clustering” and “cluster analysis”. What are the major clustering approaches?
4. What are the advantages and limitations of linear discriminant analysis?
5. What is False Discovery Rate (FDR)? Compared to the Bonferroni method, which one is more conservative?
6. What are the sources of experimental variation in data from Affymetrix chips? Do we need to normalize before intensities from different chips are compared?
7. What are volcano plots, and how are they used to find differentially expressed genes?
8. What is the difference between non-hidden Markov models and hidden Markov models? Give an example for each model.
9. What are the advantages of hidden Markov models?
10. Describe the Poisson process for estimating the number of mutations.
11. Describe how principal component analysis is used in analyzing microarray data.
12. What are the bioinformatics challenges for RNA-seq analysis?
13. Highlight the statistical issues in RNA-seq analyses.
14. Do we need to normalize microRNA data? Justify your answer.
15. Identify the major pattern recognition methods. Briefly describe the application of one method in selected area in bioinformatics.

F. Computer science and programming

1. Explain the terms OS, GUI, and programming language.
2. What are java applets?
3. What is a support vector machine (SVM) and how is it used in bioinformatics?
4. Write Linux commands that will show the files in a directory, open one of them to reveal its contents, rename the file, and delete the file (you can make up the file names for this).
5. Write a regular Perl expression that would find a particular pattern in a text file. (For example, all types of Drosophila species in the EST database.)
6. Write a regular Perl expression that would find any string where a word is repeated at least two times. (For example, the correct expression would find “hello world hello”, but not “hello world hi”.)
7. What is dynamic programming?
8. What is a brute force algorithm and its pros and cons?

9. What is pseudocode? How is it different from a programming language?
10. What are "Big-O" and "Big-Omega" Notations?
11. If the following polynomial determines time complexity of an algorithm, what is the Big-O of the algorithm? $f(n) = n^2 + 100n + \log_{10} n + 1000$
12. What are NP-Complete Problems?
13. What is performance guarantee and how to find performance guarantee for a minimization problem?
14. What is clustering? Name at least two clustering methods.
15. What is hashing? What does a hash function do?

G. Bioinformatics

1. Describe the major molecular databases, and tools available to search and retrieve data from them.
2. Describe the methods for searching databases for sequence similarity and how to interpret the significance of the results.
3. Describe the major methods of pair-wise sequence comparisons, their limitations and advantages.
4. Compute the global alignment between the sequence strings s1 = AACGGT and s2 = TGAACT, using these scoring parameters: +1 for match, −1 for mismatch, and −1 for a gap.
   a. What is the maximum alignment score between these two sequences?
   b. Show an alignment with this maximal score.
   c. Is this maximally-scoring alignment unique, or are there multiple alignments that have the maximal similarity score?
5. Describe algorithms for computationally predicting genes in genome sequences.
6. We wish to test the null hypothesis (H_0) that nucleotides occur at random in the human genome, against the alternative that there is a tendency for long repeats of A. How would you conduct such a test?
7. We wish to test whether or not a particular sequence word (GAGA) is randomly distributed in the human genome. How would you conduct such a test?
8. Describe the major methods for constructing phylogenetic trees and their advantages and disadvantages.
9. What are the differences between Jukes-Cantor and Kimura models of DNA evolution?
10. Describe the major methods for multiple sequence alignment of nucleic acids and proteins.
11. Explain the derivation and use of amino acid substitution scoring matrices.
12. Describe the rationale and organization of Gene Ontology classification.
13. What is the Pfam database?
14. What are SAM/BAM files, and what are they used for?

H. Extra credit question: Pseudogene sequences are very similar to those of a given "parent" gene, and their transcription is difficult to identify in microarrays due to potential cross-hybridization between probes and transcripts from the parent genes and the pseudogenes. RNA-seq provides an opportunity to get around this problem, and ascertain the transcription of pseudogenes specifically. A challenge, however, lies in uniquely identifying reads mapped to pseudogene regions, which are typically similar to the parent genes. How can you bioinformatically mitigate this challenge?

Rev. 8/30/2021