Solution Set 0

Due 10am, Tuesday, September 7th

This homework is intended to be a self-administered placement quiz, to help you (and me) determine if you have the background for the course. You **may** use textbooks, your own class notes from previous courses, and material you find on the internet. Please include a list of the reference materials you used when you hand in your assignment. You **may not** refer to previous homework assignments and/or solution sets from prior years or discuss this assignment with anyone other than Dr. Durand and Dr. Stolzer.

Canvas will be used to distribute assignments, for handing in completed assignments, and for returning graded assignments with comments. To facilitate this process, assignments will be made available in both pdf and latex formats. You may submit your assignments in any of the following ways:

- Download the assignment in pdf format and print it out. Write your solutions in the space provided. Scan your handwritten solution and upload it to Canvas.
- Download the assignment in pdf format. Use the commenting features in adobe or a similar tool to enter your solution directly on the pdf. Upload the completed assignment, annotated with your solutions.
- Download the assignment in latex format. Enter your solutions in latex format, compile the assignment, and turn in the resulting pdf.

For options 1 and 2, make sure that your answers are clear to read.

Please provide answers to all questions. On each question, give your reasoning and show all intermediate steps leading to your solution. Attach additional pages, if needed. Return this assignment by **10am on Tuesday, September 7th.**

1. To solve this problem you will need a table of the genetic code.

   Below you see a hypothetical stretch of genomic DNA that contains a protein coding sequence interrupted by an intron, shown in lower case. *All* letters shown in lower case are part of the intron, not just the x's.

```
CTAAG ATGGT GCTAA Ggu xxxxxxxxxxxxxxxxxxxx ag G CCTAG TTAAC TCAAT
1     6     11    16                          17      23    28
```

   (a) Assume that the first start codon you see as you scan from left to right is the beginning of the open reading frame. Draw a box around the start and stop codons of this open reading frame.

   (b) What is the peptide encoded by this open reading frame?

(c) Consider the third codon. Write down all codons that can result from a single base-pair replacement in the third codon. For each codon, give the amino acid that it encodes.

(d) What is the probability that a single base change in this codon results in a substitution of the associated amino acid for a different amino acid?

(e) Suppose that an "A" is inserted between positions 14 and 15. How will the protein sequence be changed?

2. Cystic fibrosis, a genetic lung disease, is caused by a rare, recessive autosomal allele (*Cftr*). A phenotypically normal man whose father had cystic fibrosis marries a phenotypically normal woman from outside the family, and the couple consider having a child.

   (a) What are the possible genotypes of the husband? Of the husband's father and mother? Of the wife? (Use the symbol "a" for the mutant and "A" for the normal or *wild type* allele.)

   (b) Given this information, what is the chance that the couple's first child will have cystic fibrosis, if the frequency in the population of heterozygotes is 1 in 50? Assume that no new mutations occur.

   (c) If the first child does have cystic fibrosis, what is the probability that the second child will be normal?

(d) If the first child has cystic fibrosis, and the couple has 6 more children, what is the probability that 4 of those 6 will be normal?

(e) If the first child has cystic fibrosis, and the couple has 6 more children, what is the probability that at most 1 of the 6 will have cystic fibrosis?

3. Maple Syrup Urine Disease (MSUD) is an autosomal recessive disease characterized by the inability to metabolize leucine, isoleucine and valine[1]. When these amino acids build up to toxic levels, brain injury occurs resulting in mental retardation and seizures. If infants with MSUD are identified early via genetic screening, the impact of the disease can be lessened through proper diet.

   The incidence of MSUD is roughly 1 in 100,000 individuals worldwide. Genetic tests have a false positive rate of roughly 0.15. Assuming that the false negative rate is zero, what is the probability that a randomly selected person who tests positive, actually has the disease? (Hint: use Bayes' theorem)

[1]The disorder is so called because a derivative of isoleucine has a characteristic sweet smell.

4. A clique is an undirected graph $G = (V, E)$ in which every pair of vertices is connected by an edge. If $|V| = n$, we say that G is a clique of size n.

   Let G be a clique with n nodes and let T be a spanning tree for G that was generated by a depth-first traversal of G.

   Suppose the number of edges in G exceeds the number of edges in T by 15. What is n?

5. Let G be a clique with n nodes and let T_d and T_b be spanning trees resulting from, respectively, a depth-first traversal and a breadth-first traversal of G.

   Suppose the height of T_d and the height of T_b differ by 22. What is n?

6. In each of the following questions, let T be a rooted tree, in which every internal node has exactly k children. Note that if T has L leaves, then the number of nodes in T is

$$\frac{kL-1}{k-1}.$$

   (a) Suppose that T has 176 edges when $k = 4$. How many leaves does T have?

   (b) Can T have 144 nodes when $k = 5$? Why or why not?

   (c) Three k-ary trees have 10, 25, and 43 leaves, respectively. The value of k is greater than 2 and is the same for all trees. What is k?

7. Algorithms X and Y have worst case running times no greater than $150N \cdot \sqrt{N}$ and N^2, respectively, where N is the number of inputs.

   (a) Which algorithm has better asymptotic running time, X or Y?

   (b) For which values of N would you choose algorithm X over algorithm Y? (You may give an algebraic or graphical answer.)

8. Suppose you have n bacteria and a single virus in a Petri dish. In the first generation, the virus kills one bacterium and doubles in number. The remaining bacteria also double in number, resulting in $2n - 2$ bacteria and two viruses. In the next generation, each virus kills one bacterium and then reproduces itself, resulting in a total of four viruses. The remaining bacteria divide, yielding $4n - 8$ bacteria. This pattern is repeated in subsequent generations.

   (a) Let $B(t)$ and $V(t)$ be the number of bacteria and of viruses in the Petri dish at generation t. Write down recurrence relations giving $B(t)$ and $V(t)$ in terms of $B(t-1)$ and $V(t-1)$.

   (b) Solve your recurrence relation. Give expressions for $B(t)$ and $V(t)$ in terms of n and t, assuming the initial state occurs at $t = 0$.

   (c) How many generations will it take for the viruses to eradicate the bacteria? Give your answer in terms of n.