Written questions – submit to “HW1 Written” in Gradescope

10. Answer the following questions based on your reading of Life and its Molecules.

(a) (1 pt) What are the components (monomers) that make up proteins?

Proteins are made up of a range of amino acids strung together in sequence to define a specific protein.

(b) (2 pts) What is transcription?

Transcription is the process by which the information encoded in DNA is replicated as a strand of RNA.

(c) (2 pts) What is the relationship between transcription and translation?

Transcription takes information on DNA and copies it to RNA and translation takes the information on RNA to produce specific chains of amino acids thus producing a target protein. Effectively transcription moves the information from a format useful to reproduction (DNA) to a format useful for “day-to-day” use (RNA) and translation is the second step in the process making use of that information in production.

(d) (1 pt) How do retroviruses (such as HIV) contradict the “Central Dogma?”

The central dogma states that the flow of information from DNA to RNA to protein is one way but retroviruses convert RNA into DNA thus directly contradicting the above statement.

(e) (1 pt) If we could represent a nucleic acid DNA string of length n using d bits per base, the whole thing would fit in nd bits. If our encoding uses only 0s and 1s, how many bits do we need to represent a single DNA base?

A single DNA base would need to be represented by 2 bits as we have 4 bases to specify and can use 00 01 10 11 to do so.

(f) (1 pt) Explain why 6 bits are enough to specify a codon?

6 Bits are enough to specify a codon because a codon consists of 3 ordered bases (G C U and A) thus we can specify the first base with the most significant 2 bits, the second base with the middle two bits and the third base with the least significant two bits. We can use the same encoding as we do for DNA only swapping T for U.

(g) (1 pt) What year was this article written? (Extra Challenge: can you identify an instance where the weaknesses of ”current” technology have been corrected? Instrumentation and Experimental Technology is one good place to look.)

It was written in 2004 thus is pre-second gen sequencing and is limited to the context of the Sanger/Microarray era.

(h) (1 pt) What is one key difference between eukaryotic and prokaryotic genes?

Prokaryotic genes do not have intron portions of their genes that interrupt coding sections while Eukaryotic genes have both introns and exons

(i) (1 pt) All of the cells in a human contain a genome that is ultimately a copy of the genome in the fertilized egg. Yet, humans have diverse tissues, made up of different kinds of cells. If the genome is the same, what is different about these cells which can explain their different roles and functions?

Cells are differentiated through gene expression effectively limiting what proteins and in what quantities a cell produces by repressing, enhancing, or promoting the transcription of different segments of DNA into RNA.

(j) (1 pt) How many distinct start and stop codons are there and what is their role in protein synthesis?

There are three stop codons and one start codon. Stop codons mark the end of a protein sequence. The start codon marks the beginning of one but can also be used to encode for methionine.

(k) (1 pt) In Hunter's view, what's wrong with saying that scientists have found "the gene for" a complex trait like susceptibility to breast cancer?

Traits like breast cancer are often poly-genetically controlled meaning there are a wide range of genes that will increase or decrease one's propensity towards the trait. Claiming you found "the" gene is a simplification of the fact that one may have found a gene that matters but is far from the only thing that matters.

11. (2 pts) Thinking back to the "Palindromes" programming question: can a Palindromic DNA sequence be odd in length? Support your answer with an example.

No, it cannot be odd in length because the middle base would not have a match. In a language palindrome the middle letter can match with itself, take for example, "RADAR" where we have a middle D matching with itself. For DNA we need matches to also be inverses which a make a self-match impossible. For example, AGT lets us match the first A with the final T but G cannot self-match because its match would have to be the inverse of G (U). Thus, all palindromes of DNA must have even length because each base needs to have a pair of different base than itself as a match.

12. Answer the following questions based on your reading of We Now Can See a Virus Mutate Like Never Before by Sarah Zhang in The Atlantic. Link is: https://proxy1.library.jhu

.edu/login?url=https://catalyst.library.jhu.edu/permalink/01JHU INST/t3c16/alma991060740487807861. You may need to be logged onto the campus Wi-Fi or VPN to access the link.

Note that this article describes an earlier phase of the pandemic, but the discussion of genomic surveillance is quite relevant.

(a) (1 pt) About how many DNA "letters" long is the SARS-CoV-2 genome?

The SARS-CoV-2 genome is about 30,000 letters long.

(b) (1 pt) Before SARS-CoV 2, what are some of the diseases that were studied using DNA sequencing and "genomic surveillance"? Name at least 4.

Zika, dengue, chikungunya, and yellow fever

(c) (2 pts) How did Dr. Oliveira discover that a new variant was responsible for the spike of new cases in South Africa's Eastern Cape? What specific pattern did he notice?

Dr Oliveira found that 50 clinics in the region all yielded remarkably similar DNA sequence results from covid samples. Typically one would expect to see a huge range of diversity in DNA sequences at this scale, but the fact that they are so similar suggests that the recent spike is due to the massive spread of this one specific variant found at all 50 clinics.

(d) (1 pt) From your understanding of the article, list two (2) challenges that genomic surveillance faced during the COVID-19 pandemic, give examples from the article to support your answer.

(Adoption is not universal) - One challenge is an imbalance of resources for genomics surveillance in different regions. For example, the South African surveillance did not receive the same funding as the UK effort making the pace of work limited in certain regions. The US is also not backing genomic surveillance in the same way as countries like Denmark.

(Needed Computational "Catch Up") - Due to the sudden spike in usage of DNA sequencing for genomic surveillance during the pandemic, computational tools are not yet capable of maximally utilizing the massive amounts of data processed.

13. Answer the following questions based on our discussion of DNA replication and secondgeneration sequencing. You might find it helpful to review these brief videos:

- How DNA is Copied - https://youtu.be/vaYxqrKn7Pk
- Sequencing by Synthesis - https://youtu.be/IzXQVwWYFv4
- Base Calling and Sequencing Errors - https://youtu.be/U4QnpciIJhM

(a) (2 pts) Why does it make sense to refer to a single-stranded segment of DNA as a template?

A single stand of DNA has two main functions: producing a second strand of DNA and producing a strand of RNA. In both of these cases individual bases are matched to the DNA strand base by base to create a copy. In this sense, a single stand of DNA is a template for the copying of information onto new strands.

(b) (2 pts) What is the importance of the "terminator" group added to the bases?

The terminator group added to bases in the sequencing process forces only a single base to be added at a time allowing us to stop and take a top-down picture of the DNA strands looking at each of the most recently added bases (really looking at their most recently added fluorophore). We can then remove the terminators allowing the added base to be built upon. The analogy of a smooth "stud-less" lego piece being added on top of a 1x1 to prevent anything from being built on top until it is removed is useful to describe the function of a terminator group.

(c) (2 pts) What is a negative consequence that can occur if a base added to the sequencing reaction is unterminated? How does this affect earlier versus later sequencing cycles?

If a sequence is unterminated then the sequencing cycle will effectively skip a step for that strand by continuing to add bases until eventually terminated. This has no effect on previous steps of the cycles but will mean one of the many clones of a given strand on the plate will be out of sync with the group causing a small amount of disagreement in the light emitted from the cluster. Overtime, if enough of these mistakes happen that it is not just a small percentage of the cluster it will become impossible to come to a trustworthy conclusion about the nature of the sequence, but in the short term, if only 10 of 200 strands emit a different color the noise will not affect the prediction of the sequencer due to the high percentage of agreeing clones.