1. **[/ 5]** Use the DNA sequence below for the following question. Genomic DNA is double-stranded, but by convention only one of the strands is shown below. The presence of a complementary strand is assumed.

```
ATGGGCAGAGACAAAAAAAATACAGCGCTTCTAGATATTGCAAGAGATATAGG
AGGAGATGAAGCTGTAGAAGTTGTAAAAGCCTTAGAAAAGAAAGGAGAAGCAA
CAGATGAGGAATTGGCAGAATTAACTGGAGTAAGAGTTAATACGGTGAGAAAA
ATCTTATACGCCCTGTACGATGCTAAGCTTGCAACCTTTAGAAGAGTTAGAGA
TGACGAGACTGGTTGGTATTATTATTACTGGCGCATTGATACTAAAAGATTAC
CGGAGGTTATTAGAACAAGAAAGTTGCAAGAGCTTGAAAAGTTAAAGCAGATG
CTTCAGGAAGAAACCAGCGAGACCTATTATCACTGTGGAACTCCAGGGCATCC
AAAGCTAACATTTGACGAGGCCTTTGAGTACGGGTTCCAATGTCCAATATGTG
GAGAGATACTTTACGAGTATGACAACTCAAAAATAATTGAAGAACTCAAAAAG
CGAATTGAAGAATTAGAGATTGAACTCGGACTGAGAAGTCCACCAAAAGAAGA
AAAACCAAAAAAAGCAACAAGAAGAAAAAAGTCAAGATCAGGGAAAAAGAAGA
AATAA
```

Your lab has obtained and sequenced genomic DNA from a hyperthermophilic microbe. You want to amplify the full sequence above from the genomic DNA, using a PCR protocol.

(a) **[/ 2]** Use the Primer3Plus software to identify two oligonucleotide primers that will amplify the entire sequence (hint: be sure to use the "Cloning" option in the upper left task selector, and be sure to select the entire sequence). Indicate the actual primer sequences, where they would anneal to the above strand (or its complement), and the 5' and 3' ends of each primer. Indicate the size of the DNA product expected. Are the primers 'acceptable' according to the software? If not, what are the potential problems with the primers?

- Left primer: 5' ATGGGCAGAGACAAAAAAAA 3'
  Anneals with: 3' TACCCGTCTCTGTTTTTTTT 5'
- Right primer: 5' TTATTTCTTCTTTTTCCCTGATCTTG 3'
  Anneals with: 3' AATAAAGAAGAAAAAGGGACTAGAAC 5'
- Product Size: 588 bp
- The left primer is unacceptable due to a long poly-A tail according to the default parameters of the software. This could be a problem, but PCR could still be possible, albeit possibly less effective.

(b) **[/ 3]** You perform the PCR and run an agarose gel to view the results. You expected to see a single band for your PCR product, but you see multiple bands of different sizes, including the size that you expected. You do see bands for a marker DNA size standard, so you know that the gel and staining worked properly. What might have gone wrong with the PCR? Indicate changes you could make to your

PCR protocol that should help to solve the problem.

- The likely culprit in the case about would be the failure of the right primer. Although, thermocycler parameters or other nonspecific binding to other template sequences (hairpins, self binding, wrong 3' sequence for DNA polymerase) may be causing the problem.
- Tags (biotin, fluorescent, etc.) can be added to the primers for easier identification post-PCR.
- Also, non-complementary sequences may be added to the primer at the 5' end since DNA polymerase starts at the 3' end, and it does not affect base paring of primer in use. This allows for cloning, mutagenesis, or specific sequence tags.

2. **[/ 5]** Go to the UCSC genome browser, select the Dec. 2013 (GRCh38/hg38) assembly, and navigate to the human CFTR gene.

(a) **[/ 1]** Which chromosome is the gene on, and which arm of the chromosome?

- Chromosome 7 and the right arm (q arm).

(b) **[/ 2]** Use the Gencode V36 'full' track to answer the following (note: this track is found in the "Genes and Gene Predictions" category and can be activated/altered if it is not initially visible). Use the gene version that is highlighted (blue background) to answer the following:

i. How many exons make up the CFTR **(I chose CFTR, not CFTR-AS1)** gene?

- 21 (transcript and coding)

ii. How much space does the entire gene occupy, including both exons and introns?

- 147,900 bp (transcript, including untranslated regions)
  147,706 bp (coding region)

(c) **[/ 2]** Activate the OMIM genes track to identify the OMIM entry for CFTR (note: this track is found in the "Phenotype and Literature" category and can be activated/altered if it is not initially visible). Visit the OMIM gene page to determine the function of CFTR, and whether it is disease associated.

i. What is the normal biological function of this gene?

- Cystic fibrosis transmembrane conductance regulator; it encodes an ATP-binding cassette (ABC) transporter and regulated tightly by an intrinsically disordered protein segment (termed a regulatory domain).

ii. Which, if any, diseases is it associated with?

- CFTR is associated with cystic fibrosis with pancreatic insufficiency that does not support bicarbonate transport.

3. **[/ 5]** Prediction of protein function and conservation using BLAST.

Example polypeptide sequence:

MGRDKKNTALLDIARDIGGDEAVEVVKALEKKGEATDEELAELTGVRVNTVRKILY
ALYDAKLATFRRVRDDETGWYYYYWRIDTKRLPEVIRTRKLQELEKLKQMLQEET
SETYYHCGTPGHPKLTFDEAFEYGFQCPICGEILYEYDNSKIIEELKKRIEELEIELGL
RSPPKEEKPKKATRRKKSRSGKKKK

(a) **[/ 2]** Use NCBI BLAST to determine the probable function of the protein encoded by the example polypeptide sequence above. Choose the Basic BLAST program "protein BLAST". Paste the query sequence, and search the non-redundant protein sequences within "archaea" (by typing this into the "Organism" search set).

i. Indicate the most similar protein and its function, and identify the organism this gene is from.

- Transcription factor E from *Pyrococcus furiosus*; plays a role the activation of archaeal genes transcribed by RNA polymerase[1].
  - Grünberg, S., Bartlett ☺, M. S., Naji, S., & Thomm, M. (2007). Transcription factor E is a part of transcription elongation complexes. Journal of Biological Chemistry, 282(49), 35482–35490.

ii. To assess the significance of the alignment, give the E-value as well as the identity and positive percentages.

- Total Score: 394
- Query cover: 100%
- Identity: 100%
- E-value: 5.1×10^{-140}

iii. With these values, can you state the function of this protein, and what is your level of confidence in assigning this function?

- Very confident; the lower the E value, the more unlikely the alignment happens by chance. Total score = max score, and identity and query cover is at 100%.

(b) **[/ 1]** Go back to the BLAST home page, and use standard protein BLAST determine if there are proteins with biologically relevant similarity in Homo sapiens (note: it may take a few minutes for results to arrive).

- No significant similarity found.

(c) **[/ 2]** Next, perform a BLAST search again for similarity in Homo sapiens, but this time using the DeltaBLAST option (note: Delta stands for 'Domain Enhanced Lookup Time Accelerated').

i. Does this modify your previous answers in part b, and if so, how? How likely is it that the function of the top hit is the same as the function of the query protein sequence? Which information in the BLAST results allowed you to make this judgement?

- No, no significant degree of confidence in similarity. The E-value is low, but compared to archaea it is much higher. Query cover is less than ideal—with only 76% there could be several missed mutations that significantly change function.
- Interesting that it is describing a general transcription factor IIE, possible subunit of α-subuint.
- However, with a percent identity of only 14.47%, then the number of identical amino acids is quite low and below the "40%" rule of thumb.

ii. Which BLAST approach was most valuable, and why?

- I mean, I suppose it depends on context, but the basic protein-protein blast certainly shows identical amino acid sequence in the specific organism you are interested and thus is pretty useful for identifying a protein. However, it is next to useless if there is no match (well, no match can still be useful data, but I digress) so in this example case then it isn't that helpful.
- If this sequence was generated from a human sample (is this necessary, actually?), then the DELTA blast would be useful to determine whether the sequence you have is just not known or possibly a distant homolog, especially in cross-domain homology; such results must be taken with caution if results are similar to the ones from part (b).
- Running the DELTA blast again, but restricting it to archaea, leads to results similar to that of part (a). Interestingly, the E value is much lower, and the query cover is not at 100% either; the total and max score is much less too. These results indicate that DELTA blast might not be better in general.
- With a sample size of three searches then I suspect I have little data to make an informed conclusion here.