

PGSS Biology Core 2022 Homework #1

Due Thursday, June 30

The sequence of DNA in Figure 1A is from an exon near the center of a specific wild type (normally functioning) human gene. This gene is transcribed, or “expressed”, by many human tissues including the intestinal epithelium.

1. Take the DNA sequence and transcribe it into mRNA.

2. Take the mRNA sequence and translate it into protein (use the single letter abbreviations for the amino acids) using the genetic code.

3. Used by biologists around the world, BLAST is a freely available computational tool provided by the National Center for Biotechnology Information (NCBI) for analyzing DNA and protein sequences. It does this by comparing your sequence (the query), to the vast database of DNA and protein sequences maintained by the NCBI.

(<https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>)

Take the amino acid sequence and identify the protein using BLAST. See Figure 2 for guidance. After the search is complete you will need to scroll down to find the part of the results that shows you alignments between your sequence and sequences from the database.

4. Take the **mutant** DNA sequence in Figure 1B and transcribe it into mRNA.

5. Take the mRNA sequence and translate it into protein using the genetic code.

6. Predict what will happen to the protein. Then, go to protein schematic in Figure 3 to see where this mutation is located in the gene and the corresponding amino acid location in the protein. Based on that information about this specific protein, predict the consequence to the mutant protein (e.g. what functions will it retain, which will it lose?)

7. PubMed is a database of biomedical literature maintained by the NIH. Go to PubMed, search using the name of the protein as you identified it using BLAST including the terms “mutations” and “signaling” in the search and find out what effect mutations in this gene has in real life. In addition, discover something about the normal function of the protein. Write a brief synopsis (a few sentences) of your findings.

<https://www.ncbi.nlm.nih.gov/pubmed>

8. a. Examine the plot that shows mutation frequency in Figure 3 and you will notice that mutations in our gene of interest that are associated with human disease can be either somatic or germline. Diseased human cells with mutations in our gene of interest (you should know what kind of cells these are from your answer to question 7) always seem to be homozygous mutant; in the majority of patients, both disease alleles will be the result of somatic mutation, while in other patients, one allele will result from somatic mutation, and the other from a germline mutation. In light of that information, describe where you think the mutations came from in those two classes of patients.

b. The frequency of mutation varies across the gene. Speculate about the basis for the observed pattern and what that tells you about the function of the different parts of the protein.

Figure 1A

DNA sequence (wild type)

The top strand is the coding strand and the bottom strand is the template strand.

```
5' CAAGAGGCTGATAGCGCCAATACACTTCAAATCGCTGAGATCAAAGAAAAAATCGGGACACGAAGTGCTGAGGATCCCGTC 3'
3' GTTCTCCGACTATCGCGGTTATGTGAAGTTTAGCGACTCTAGTTTCTTTTTTAGCCCTGTGCTTCACGACTCCTAGGGCAG 5'
```

Figure 1B

DNA sequence (mutant found in a patient population)

The top strand is the coding strand and the bottom strand is the template strand.

```
5' CAAGAGGCTGATAGCGCCAATACACTTCAAATCGCTGAGATCAAATAAAAAAATCGGGACACGAAGTGCTGAGGATCCCGTC 3'
3' GTTCTCCGACTATCGCGGTTATGTGAAGTTTAGCGACTCTAGTTTATTTTTTAGCCCTGTGCTTCACGACTCCTAGGGCAG 5'
```

Figure 2

1. Enter the single letter amino acid sequence here

2. Do not adjust the parameters

3. BLAST!

The screenshot shows the NCBI BLAST Standard Protein BLAST interface. The top navigation bar includes the NIH logo, "U.S. National Library of Medicine", the NCBI logo, "National Center for Biotechnology Information", and a "Sign in to NCBI" link. The main header shows "BLAST® >> blastp suite" with links for "Home", "Recent Results", "Saved Strategies", and "Help". Below the header, the "blastp" tab is selected among other options like "blastn", "blastx", "tblastn", and "tblastx". The "Enter Query Sequence" section contains a large text input field for the "Enter accession number(s), gi(s), or FASTA sequence(s)", a "Clear" button, and a "Query subrange" section with "From" and "To" input fields. Below this is an "Or, upload file" section with a "Browse..." button and "No file selected." text, followed by a "Job Title" input field and a prompt to "Enter a descriptive title for your BLAST search". A checkbox for "Align two or more sequences" is also present. The "Choose Search Set" section includes a "Database" dropdown set to "Non-redundant protein sequences (nr)", an "Organism" section with a text input and an "Exclude" checkbox, and an "Exclude" section with checkboxes for "Models (XM/XP)" and "Uncultured/environmental sample sequences". There is also an "Entrez Query" section with a text input and a "Create custom database" link. The "Program Selection" section shows the "Algorithm" dropdown set to "blastp (protein-protein BLAST)", which is marked as "New". Other algorithms listed include "Quick BLASTP (Accelerated protein-protein BLAST)", "PSI-BLAST (Position-Specific Iterated BLAST)", "PHI-BLAST (Pattern Hit Initiated BLAST)", and "DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)". At the bottom, a "BLAST" button is next to a summary of the search: "Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)". A checkbox for "Show results in a new window" is also visible. A link for "+ Algorithm parameters" is at the very bottom.

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® >> blastp suite Home Recent Results Saved Strategies Help

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence BLASTP programs search protein databases using a protein query. more... Reset page Bookmark

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange

From To

Or, upload file Browse... No file selected.

Job Title Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism Optional Enter organism name or id—completions will be suggested ☐ Exclude + Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query Optional Enter an Entrez query to limit search YouTube Create custom database

Program Selection

Algorithm

☐ Quick BLASTP (Accelerated protein-protein BLAST) New

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

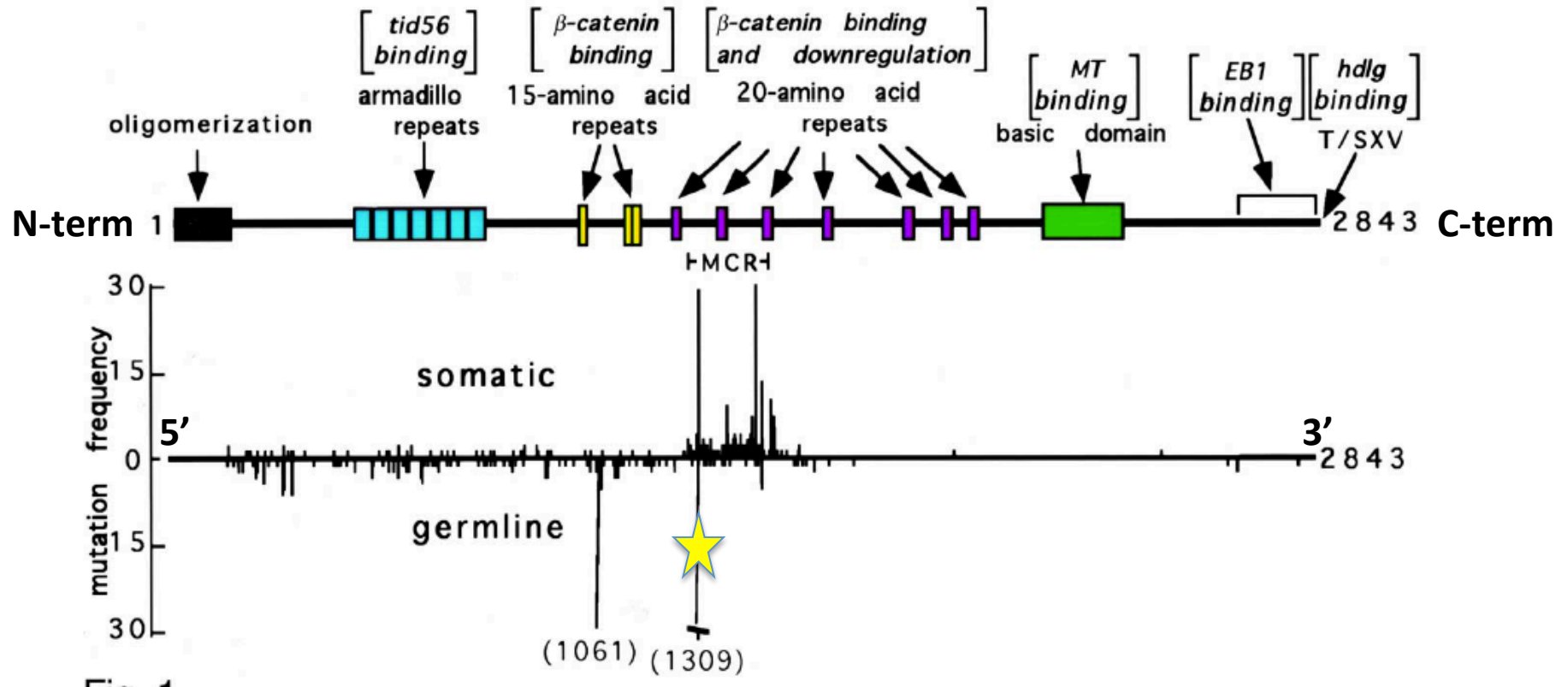
Choose a BLAST algorithm

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

☐ Show results in a new window

+ Algorithm parameters

Figure 3



- Each colored box in the protein cartoon represents a different functional region of the protein
- MCR = mutational cluster region
- the yellow star indicates the position of the mutation described in questions 4 and 5
- mutation frequency refers to how often a mutation at a specific nucleotide has been identified in a cancer patient population. Higher numbers indicate that the mutation is more frequent.