# Protein Sequence

Although we look at the DNA sequence of our samples, often what we are interested in clinically is how the variation on the DNA level affects the protein(s) produced.

Open the protein sequence saved in part two of the DNA sequence work in the document DNASequence.docx. This was saved as ProteinSequence.txt.

## Locate the protein sequence within the genome (hint- use BLAST)

This will tell you where the protein sequence falls within the genome by telling you which gene it comes from

* Find the gene that the protein sequence in ProteinSequence.txt comes from (hint- use blast, restrict the search to reference sequences to make the second part easier to click through entries in NCBI)
* What are the genomic co-ordinates of the gene associated with this protein (hint look up in ncbi)? Remember to record the version of the reference genome used.

## Predict the function of the protein

This may tell you something about the family of a protein and therefore its possible functions if you do not know which protein it is (i.e. you have an amino acid sequence but no idea which protein it could be from).

### On motifs and patterns (hint- use InterPro)

We have already done this as part of the DNA sequence work

* Copy and paste the screenshot from running InterPro below
* What sort of function might this protein have based on the output from InterPro? (Which domains does it find in the proteins and what does it say these might do? Which family or superfamily does the protein belong to and what is the general function of proteins in this family?)
* InterPro also links you to relevant gene ontology terms (see bottom of output- GO terms).
  + What is a gene ontology term and why is it useful (hint- uses of controlled vocabularies)?

### On sequence similarity (hint- use blast)

We have already used blast as part of the DNA sequence work to identify an unknown sequence. We could use the same principle, searching databases to find matches and partial matches to the sequence we are interested in, to compare an unknown protein that has no hit to our species of interest to proteins from other species.

* Run blast on the protein sequence again, restricting the output to exclude human sequences
  + Do we get any hits?
  + Do we get any hits that are not predicted or synthetic constructs from any organisms (e.g. a sequence where the identifier starts with NP)? If so, select one and save a screenshot of the display showing the comparison between our query sequence and the sequence from the organism.
    - How close is the match and what sort of differences do we see?

## Predict an alteration of function in a potentially altered protein

In clinical genomics, we are interested in whether a condition has an identifiable underlying genetic cause. To achieve this, we identify variation in the genome and investigate these variants to determine if they might be clinically relevant to a known phenotype. Part of this process, for variants that are predicted to result in a missense change at the protein level, is in using bioinformatics tools to predict the consequences of the change in amino acid for the function of the protein.

Commonly used tools include Mutation Taster, SIFT, AlignGVGD and PolyPhen, which are commonly accessed in practice through the software Alamut (<https://www.interactive-biosoftware.com/alamut-visual/>)

It is useful to have an idea of how the tools work, as this can help with interpreting the outputs that they generate

* Read the links in the document BioinformaticsPredictionTools.docx for a brief overview of the principles of operation of some of the tools

### Using MSA (Multiple Sequence Alignment)

Conservation of an amino acid across evolution can be useful evidence of its importance to the function of the protein. Where an amino acid is not well conserved and is different in many species it is less likely to be critical for proper functioning of the protein. Why is this?

We have used a multiple sequence alignment (for only two sequences) as part of the DNA sequence work to compare our predicted protein to our known protein. We will use the same software (Clustal Omega) to compare sequences from different species.

* Download protein sequences for the BRCA1 gene for human, rhesus monkey, dog, rat, mouse, and chicken (longest transcripts) as fasta files
  + Note that this is an example. Depending on the gene, the best species to use to form a useful MSA for inferring altered function may vary.
* Make a multiple sequence alignment of these protein sequences and screenshot (hint- use Clustal Omega) (note: this will require you to make a single fasta file containing all of your sequences)
  + Run the tool again with the same data and change the output to fasta. Save the fasta file as BRCA1\_MSA.fasta
* Which tools available in Alamut (Mutation Taster, SIFT, AlignGVGD and PolyPhen) use multiple sequence alignment as part of their evidence?
* What other lines of evidence do the tools available in Alamut (Mutation Taster, SIFT, AlignGVGD and PolyPhen) use (briefly)?

### Other tools

Consider the variant Trp1837Arg (W1837R) in NP\_009225.1 (BRCA1)

* Run AlignGVGD
  + Open your alignment generated above (BRCA1\_MSA.fasta) and cut and paste the human sequence to the top and save
  + Use your alignment (BRCA1\_MSA.fasta) and the library alignment provided by AlignGVGD. Are there any differences in output? Why might this be?
* Run SIFT
* Run PolyPhen
* Run Mutation Taster
  + It requires DNA sequence input
  + The transcript that corresponds to our protein NP\_009225.1 is NM\_007294.3 and the change is c.5509T>C (NM\_007994.3)
* Do the tools agree on whether or not the variant is likely to be pathogenic?
* Why might they not agree?
* Apart from using bioinformatics tools, what other information would you gather to determine if an amino acid variant is pathogenic or not (briefly)?
* Could you use just the tools to decide if a variant is pathogenic or not? Why?

## Resources

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

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

<https://www.interactive-biosoftware.com/alamut-visual/>

<https://www.ebi.ac.uk/Tools/msa/clustalo/>

<http://agvgd.hci.utah.edu/>

<https://sift.bii.a-star.edu.sg/>

<http://genetics.bwh.harvard.edu/pph2/>

<http://www.mutationtaster.org/>

ProteinSequence.txt

BioinformaticsPredictionTools.docx