# BIO INF CW [to be copy pasted into LaTeX]

Part 1:

I discovered four transcripts for CDH7, all of different lengths and all were protein encoding, although three of the proteins had equal lengths. Of the three transcripts whose lengths were similar, they also had the same distribution of nucleotides. However, the shortest transcript had a significantly different distribution of nucleotide to the others as can be seen in the normalised graph.

In terms of proteins, three of the transcripts encode the exact same protein despite differing in length. While NM\\_001317214.3 encodes a smaller protein that does not follow the same distribution of amino acids as the other transcript's proteins.

Part 2:

Three of the encoded proteins in Part 1 were identical so for this task I first extracted the proteins from NM\\_004361.5 (encodes the joint longest protein of length 785) \& NM\\_001317214.3 (encodes the shortest protein of length 630). I then used the same alignment from Part 2 Task 1, this time choosing the option to align proteins, I used local alignments due to the difference in size but also global to get a larger picture of how the proteins compared to each other.

**discussion**

Comparing the coding regions of the longest and shortest transcripts, (using a local alignment) we can see that there is a complete match from [0-1864], after which the alignment becomes patchier with a sequence of gaps, matches and replacements. The output statistics were length=1898, The identity and similarity were both 99.0% and there were 12 gaps (0.6%). The alignment score was 9332.5.

Comparing the full-length transcript sequences using a local alignment (as they are significantly different in size). We have a complete match between [1-136] at the beginning of both sequences, followed by a large gap of length 802. Then we have a very large matching section of length 2062. After this point there are no large sequential matching sections rather gaps that range from size 1-50, and small sections of matches and replacements. The statistics for this alignment were length=5358, identity and similarity both 56.7% and gaps making up 40.0%. The score given was 11838.0.

Examining the global alignment of the long and short protein isoforms we can see a complete match between [1-621] of both the isoforms and then we can observe a gap of 7 amino acids. Followed by a replacement (of a strongly related amino acid), and by a match, these appear to be just a product of the alignment algorithm rather than a meaningful pattern. Then because the short isoform is significantly shorter than the long there are no more matches. The alignment had an identity and similarity of 79.5%, and 20.3% gaps (although these statistics are misleading and largely a result of the sequences being different lengths and if we instead only looked at the first 643 amino acids of the long sequence then our gaps would be 7/643 = 1.1%, and Identity = 622/643 = 96.7%). Because of the differences in length, I also decided to use a local alignment, this gave 100% identity and similarity for an alignment length of 621, which we already observed in the global alignment. The score for the local alignment was 3187.0, and the score for the global alignment was 3179.0.

Task 5:

After browsing literature related to CDH7, I think that the missing exon from the short protein isoform results in the short protein being soluble compared to the insoluble long protein. This happens because the exonic sequence codes the TM domain, and removing this generates a soluble version of the protein.

Because this short protein isoform is soluble this means it is able to cross cell membranes. This most likely means that this protein is a single-pass transmembrane protein.

A cadherin is a transmembrane protein, these are proteins that span the entire cell membrane and often function as gateways to control the transport of specific substances across the membrane. They are usually highly hydrophobic, the main reasons for this are so they can act as function as barriers and so they can be easily inserted into a membrane (that contains other hydrophobic phospholipids). When acting as barriers they often form channels or pores in the membrane to allow the passage of ions, molecules, and other substances. But we know that the short protein isoform is soluble, thus it is most likely a single-pass transmembrane protein. This is another form of transmembrane protein, but this contains hydrophilic and hydrophobic regions. The hydrophilic region allows them to connect the hydrophilic interior and exterior environments, while the hydrophobic regions allow them to stay anchored in the membrane. The cell membrane needs a protein with this property so that it can interact with its surrounding aqueous environment, rather than blocking all access to water from the cell.

Part 3:

I wrote a query for the Protein database using the Entrez package to find all the human proteins with a Cadherin domain. It took some time to refine my search filters to ensure that I was getting only proteins with a Cadherin domain without including other proteins. The query term I ended up using was "cadherin[All Fields] AND Homo sapiens[porgn] AND swissprot[filter]" on the Protein database. This returned 238 proteins including those that did not seem to be human and have a cadherin domain, therefore I inspected the search results and found where the search results (sorted by relevance) stopped returning protocadherins (118), I then set the max number of returned elements to 118 and used these for the remaining tasks.