Bioinformatics 1 Assignment 2

a) *Sicle cell anemia* is a genetically inherited disease associated with a mutant in the human *Hemoglobin beta (HBB) gene*.
b) *The human leukocyte neutrophil elastase (ELANE) gene* is associated with inflammatory reactions or
c) *Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E) gene* that is associated with gene regulatory protein

You are required two analyse

Gene (a)

Gene (b) or gene (c)

Using online resources such as the NCBI **Download the FASTA DNA/mRNA file and the FASTA amino acid file for the genes**. (hint *HBB is around mRNA 600 nucleotides in length; ELANE is about 5000 DNA nucleotides in length ;and the eIF4E gene is around is ~ 2881 bp in size*).

Introduction

The assignment's main purpose is to discuss sickle cell anaemia's gene in relation to mutations, exons, and introns. It also discusses the results of when two genes are compared together and the various discrepancies that exist. Overall it is to show the knowledge gained using various methods such as Dot Plot Matrix, Blosum Matrix, and using the online software BLAST.

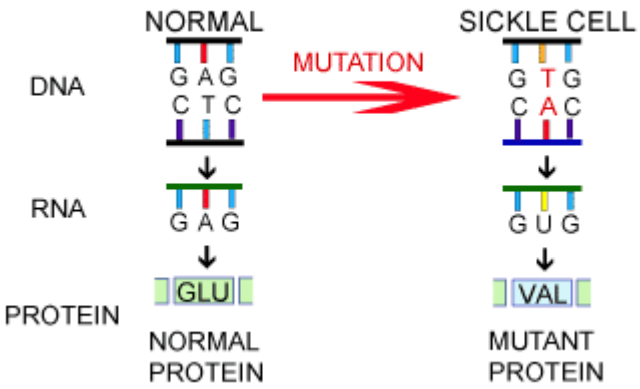
Write a paper of approximately *2000* words on:

1. Briefly describe the DNA and amino acid sequence difference between the normal and mutant form of the gene associated with sickle cell anaemia. **(2 marks).**

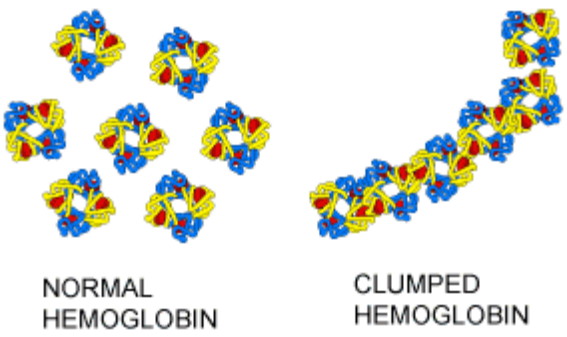
Sickle cell anaemia is a genetic disease with severe symptoms, including pain and anaemia. The disease is caused by a mutated version of the gene that helps make haemoglobin — a protein that carries oxygen in red blood cells. People with two copies of the sickle cell gene have the disease. People who carry only one copy of the sickle cell gene do not have the disease, but may pass the gene on to their children. The disease gets its name from to the shape of the red blood cells under certain conditions. Some red blood cells become sickle-shaped and these elongated cells get stuck in small blood vessels so that parts of the body don't get the oxygen they need.

The mutations that cause sickle cell anaemia have been extensively studied and demonstrate how the effects of mutations can be traced from the DNA level up to the level of the whole organism. Sickle cell anaemia is caused by a single code letter change in the DNA. This in turn alters one of the amino acids in the haemoglobin protein. Valine sits in the position where glutamic acid should be. The valine makes the haemoglobin molecules stick together, forming long fibres that distort the shape of the red blood cells, and this brings on an attack.

Consider someone carrying only one copy of the gene. She does not have the disease, but the gene that she carries still affects her, her cells, and her proteins:



There are also effects at the protein level. When red blood cells carrying mutant haemoglobin are deprived of oxygen, they become "sickle-shaped" instead of the usual round shape. This shape can sometimes interrupt blood flow:



The above picture shows normal haemoglobin (left) and haemoglobin in sickled red blood cells (right) look different; the mutation in the DNA slightly changes the shape of the haemoglobin molecule, allowing it to clump together.

There are negative effects at the whole organism level. Under conditions such as high elevation and intense exercise, a carrier of the sickle cell allele may occasionally show symptoms such as pain and fatigue. Yet there are positive effects at the whole organism level as well. Carriers of the sickle cell allele are resistant to malaria, because the parasites that cause this disease are killed inside sickle-shaped blood cells.

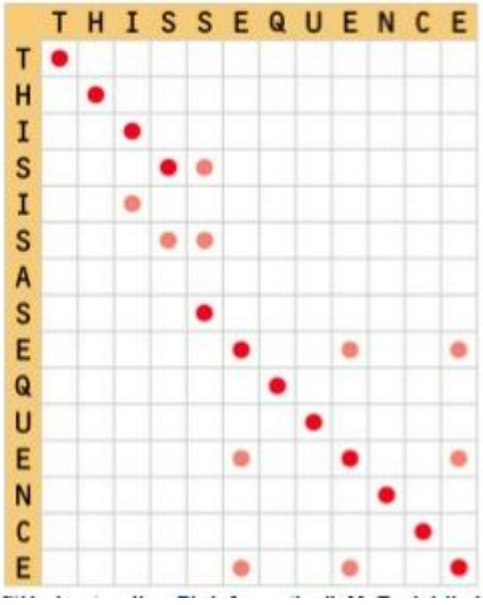
2. If you have a "manually translated" eukaryotic DNA sequence of a gene and the experimentally determined amino acid sequence for the gene, Describe, using appropriate examples:
   - The effect of introns/exons on the ORF (hint: frame shifts and incorrect residues)
   - How you would attempt to find the exact position (bp number) of all the *exons* and *introns* within the DNA sequences using sequence alignment/matching software.

(8 marks)

Genes can be mutated by the deletion or insertion of a number of nucleotide bases. There are different types of deletions or insertions. It is vital for the introns to be removed precisely, as any left-over intron nucleotides, or deletion of exon nucleotides, may result in a faulty protein being produced. This is because the amino acids that make up proteins are joined together based on codons, which consist of three nucleotides. An imprecise intron removal thus may result in a frameshift, which means that the genetic code would be read incorrectly.

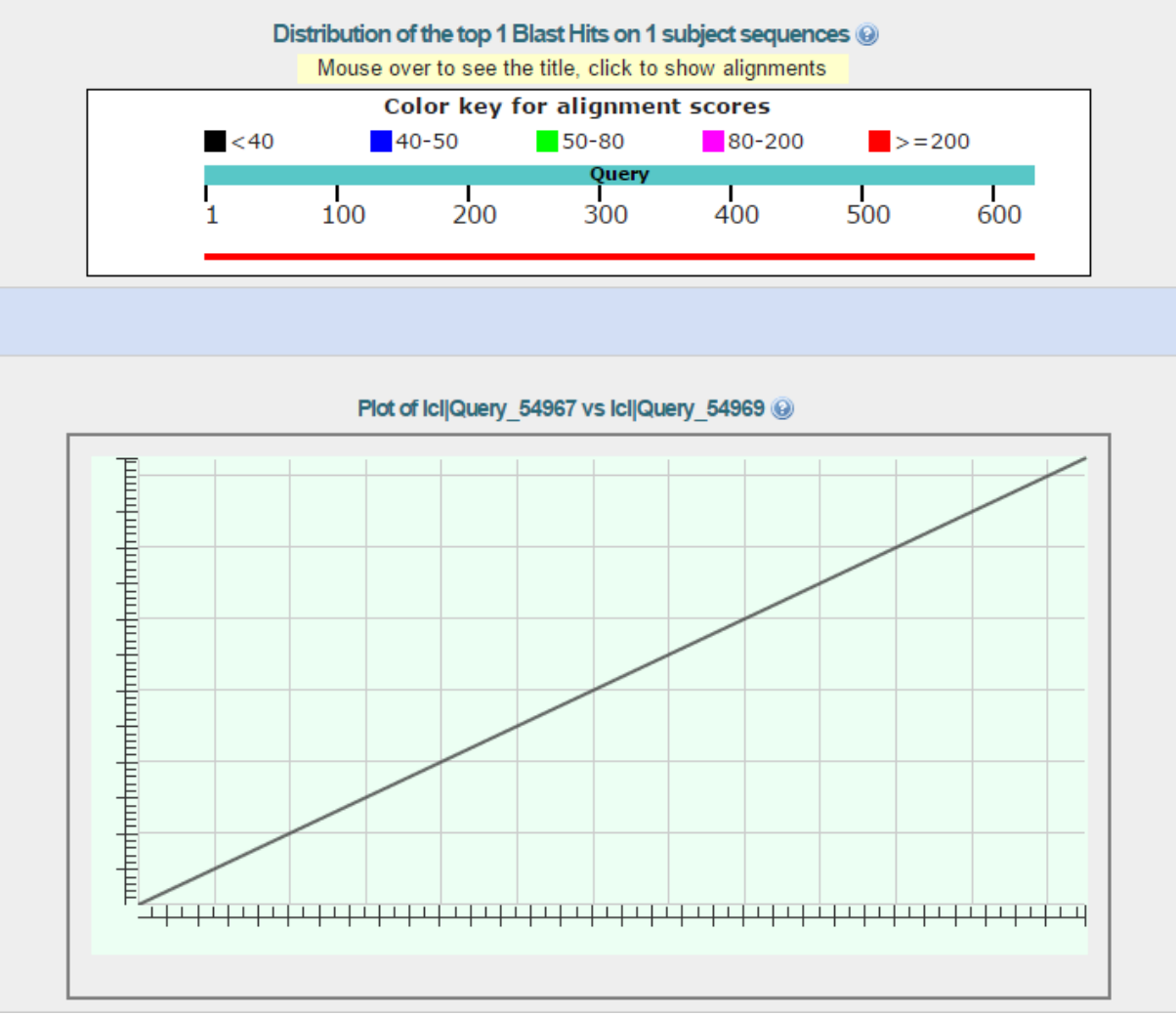
Frameshift mutations occur when the number of deleted or inserted base pairs is NOT a multiple of three. Since codons consist of three base pairs, if, for example, only one or two base pairs are deleted, then the way the DNA is read is shifted at the place of the deletion or insertion. After the place of the mutation, ALL of the amino acids that follow will be different. In this case, either an abnormal protein is made or no protein is made at all. For example: This is an exon "GTA CTG GGA TCT CAA". If the intron before this exon was imprecisely removed, so that the "G" was no longer present, then the sequence would become unreadable: "TAC TGG GAT CTC AA…" and so on. In-frame mutations occur when the number of deleted or inserted base pairs **IS** a multiple of three. This results in a change in only a few amino acids; it may still be possible for the protein to function, even though its sequence may be slightly different.

In order to find the exact position of all the exons and introns, we can use sequence alignment. Sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. We can do this by using the Dot Plot Matrix. This is where each sequence of DNA is represented as a table or matrix, where one sequence represents the rows and the other represents the columns. All elements (row/column) are checked for a match and if there is one, the cell is marked. This will show areas where both sequences and matches occurs. Hence will know and find what part is an intron and part is an exon, by comparison alone.



The above image shows the Dot Plot matrix. 'THISSEQUENCE' can be considered the row sequence which is being compared with the column sequence 'THISISASEQUENCE'. The red dots show where the sequence matches. This can help us find exons and introns of both sequences. If they match more, that means the protein produced should be similar for both. If they match less, then sequences have differences between them.

A Dot Plot matrix for longer sequences can also be generated using an online software called BLAST. You are basically given with 2 entry boxes where you place the first sequence and then in the other box you placed the second sequence. You then initiate the process for them to be sequentially aligned. The following is what will be generated; a graphic summary of the alignment scores and the Dot Plot matrix.



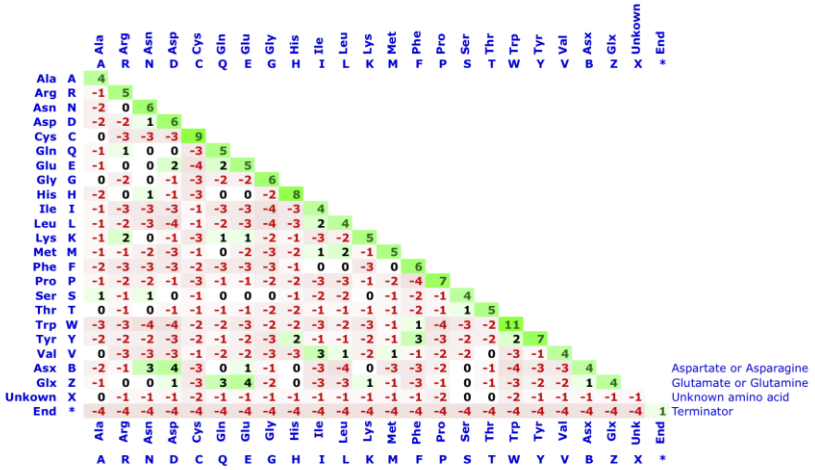
Another panel is also generated which shows the sequences being compared with each other to show with base pairs exist and which don't. Currently the following shows that the sequences match fully.

Range 1: 1 to 626 Graphics ▼ Next Match ▲ Pr

Score	Expect	Identities	Gaps	Strand
1157 bits(626)	0.0	626/626(100%)	0/626(0%)	Plus/Plus

```
Query  1    ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATC  60

Query  61   TGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG  120

Query  121  TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTG  180

Query  181  AGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  AGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTC  240

Query  241  ATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGG  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  ATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGG  300

Query  301  GCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACT  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  GCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACT  360

Query  361  TCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  TCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA  420

Query  421  CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCC  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCC  480

Query  481  ACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCC  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  ACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCC  540

Query  541  CTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCC  600
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  CTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCC  600

Query  601  TAATAAAAAACATTTATTTTCATTGC  626
            |||||||||||||||||||||||||
Sbjct  601  TAATAAAAAACATTTATTTTCATTGC  626
```

The BLOSUM (BLOcks SUbstitution Matrix) matrix is a substitution matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments. The Blosum Matrix for the above 2 sequences aligned can be used as follows, for example, the following shows the Blosum Matrix.



If we compare A and A against each other, we can see that A against A on the Blosum matrix gives us the number 4. If we do the comparison for both sequences similarly, we can then add up all the numbers and find out the sequence of alignment.

3. Using online analytical tools determine the regions of alignment between the manually translated DNA sequences and the experimentally obtained: amino acid sequences:
    - specifying their position (bp numbers);
    - the reading frame on which they reside
    - any other data you consider will be critical in the **annotation** of the gene sequence.

(15 marks)

I compared the original gene A and the translated gene C Amino Acid files against each other. They showed alignment at the base pairs 50 to 60 in gene A against 10 to 20 in gene C. This resides on the second open reading frame. There were 5 positives and 0 gaps when these proteins were compared. For the sequences producing significant alignments, there was a max score of 13.1% only between these 2 genes and they both were only 36% identical. Since they both have different levels of exons and introns present, there is no doubt that these genes will have quite a high difference between each other. Both would be spliced at different regions and both produce complete different proteins altogether. Obviously the sequences and the genes would be different, considered they both have different functions.

I also decided to compare the original gene A against the translated gene A, i.e. against itself! This showed that the DNA which I used to translate to obtain the Amino Acid sequence for gene A, it resulted in a 100% match. There were no gaps whatsoever and the base pairs from 1 to 60 base pairs, 61 to 120 base pairs, and 121 to 147 base pairs was where these both resided, on the second open reading frame as well. As these genes carry out the production of the same protein and carry out the same splicing at the same regions, these genes would definitely would be 100% identical to each other, as both have the same function. There was a total max score of 301 in regards to the sequences producing significant alignments and 100% query cover.

Next I decided to compare the original gene C against the translated gene C, again, with itself. In regards to this, the sequences producing significant alignments, was 503 in total yet showed 97% coverage but yet is 100% identical! They are located on the third open reading frames and range from 8 - 67 base pairs on the original gene in comparison to the 19 – 78 base pairs. 68 – 127 base pairs on the original gene were matched with 79 – 138 base pairs on the translated gene. 128 – 187 base pairs were matched to the 139 – 198 base pairs on the translated gene. 188 – 247 base pairs also matched against the translated 199 – 258 base pairs. The last 249 – 248 base pairs were matched against 259 base pairs on the translated gene. There are 0 gaps and 100% identities and 100% positives. This shows that the translate gene and the original gene are 100% identical even though there was only 97% coverage. This also means that not every element in a gene is required as long as the required coding regions i.e. exons are present in order to produce the required protein. Both gene hence produce the same protein, as long as the required building blocks are there.

4. Using the annotated data in the "geneback" record: Discuss the results of your analysis from part B: e.g.
    - the "correctness" of the exon and intron positions;
    - any frame shits and why they occur
    - other information (contained in the genebank record) you consider important to explain differences between the results from part b and the annotated gene sequences (DNA/RNA). **(30 marks)**

When two genes are compared, they show that they are less identical mostly because they share the same Amino Acid sequences needed to produce their own proteins. They also will have different number of exons and different number of introns. As seen when comparing gene A and the translated gene C, they both did not agree with each other considering both genes operated in different ways to produce a different product. They would obviously have different exons and introns. Hence different frame shits would occur also. Frameshift mutations occur when the number of deleted or inserted base pairs is NOT a multiple of three. This causes the DNA sequence to be read differently than it was originally intended and hence abnormal proteins would be produced. This showed there were many discrepancies and also the fact that sequence alignment was quite low, with a total of only 13.1%!

When the original gene and the translated gene of the similar type were compared, it showed they were 100% identical. The exons and introns would be in the correct places, as seen when comparing gene A and the translated gene A. Even though one was translated yet the other was original, they both showed they had similar exons and introns and both were on the same open reading frame. They also showed 100% alignment and 100% coverage. This showed that they produced similar proteins and hence were 100% identical. Both genes had the same operation and hence were, again, 100% identical.

When comparing the translated gene C and the original gene C, there were discrepancies clearly seen, as there was only the coverage of 97%! Yet the comparison showed that they both genes were 100% identical. Both produce the same protein and were similar genes, even if there was some material which did not match or was missing, it seems to not have mattered considering the sequence alignment showed a total of 100%. This should indicate that even though different base pairs matched different base pairs, both were quite identical and hence should produce the same normal protein each. Even if frameshifts occurred, they should have existed slightly considering the numbers showed that these genes are very identical.

Conclusion

This assignment demonstrated what sickle cell anaemia is and how it differs when mutation comes into play. It also showed how base pairs can be found and what are frameshifts in regards to exons and introns. Genes A and C were specifically picked for this assignment and compared the differences in regards to sequence alignments, similarity and the different behaviours each bestows.

References

1. http://evolution.berkeley.edu/evolibrary/article/mutations_06
2. https://www.dnalc.org/resources/3d/17-sickle-cell.html
3. https://www.ncbi.nlm.nih.gov/
4. https://en.wikipedia.org/wiki/BLOSUM