Name: Taha Ahmad
Student ID: 2546125

BIN 506 : Protein and DNA Sequence Analysis
Assignment : 01
By:
Taha Ahmad
Student ID: 2546125
Instructor: Yesim Aydin Son

Question 1:

a) head() of both input sequence files 'AF230076.1.fna' & 'NM_000558.3.fna'

```
>AF230076.1 Homo sapiens alpha-2-globin (HBA2) gene, complete cds

CGCCCGGCCGGGCGTGCCCCCGCGCCCCAAGCATAAACCCTGGCGCGCTCGCGGCCCGGCACTCTTCTGG

TCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCT

GGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAAGTGAGGCTCCCTCCCCTG

CTCCGACCCGGGCTCCTCGCCCGCCCGGACCCACAGGCCACCCTCAACCGTCCTGGCCCCGGACCCAAAC

>NM_000558.3 Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA

ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTC

AAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCC

TGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGG

CCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTG
```

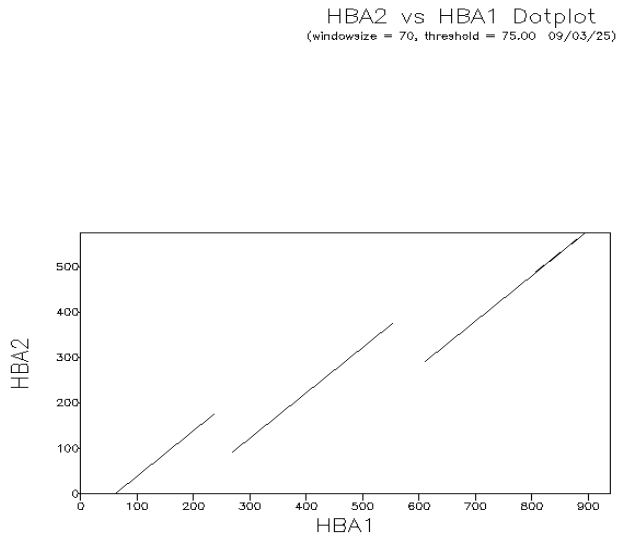
b) <u>explanation of parameters and their uses:</u>

- Window size : determines the number of nucleotides compared at a given time, a smaller window size will increase the sensitivity to short matches but will increase the noise of the data. A larger window size will reduce the noise of the data however it may miss short conserved regions which can be important during analysis.
- Threshold: it is based on the scoring matrix. The scoring matrix chosen for our sequencing was “EDNAFULL”, the matrix defines the mismatch score for nucleotides,

Name: Taha Ahmad
Student ID: 2546125

EDNAFULL score matches as +5 and mismatches as -4. The threshold value will set the minimum score required to plot the dot plot. A lower threshold will include more dots, this means that weaker matches are included whereas higher threshold value rescues noise buy including only strong matched

c) Optimised dot plot:



Explanation:

Windows size 70 : balances sensitivity and reduction of noise at the same time, since hemoglobin alpha genes have long, highly conserved exons (90%) therefore our windows size will smoothen out the minor mismatches in coding regions while ensuring that there is no over something therefore ensuring that shorter conserved motifs are still visible.

Threshold 75: our threshold value ensures the detection of regions with greater than 15 matches in the window and excludes random matches, this reduces the noise and gives a higher resolution clean image

What happens if we increase or decrease the values:

- Although increasing the window size and threshold value is giving us a clear plot much similar to the one above, however it may be missing shorter conserved motifs, it also may lead to over smoothing of the graph.
- Decreasing the values of both threshold and window size (e.g W:30 and T:50) gives us a graph with a lot of noise, which is resembled by the many scattered dots present; this noise in data makes it hard to distinguish true matches.

Name: Taha Ahmad
Student ID: 2546125

Question 2:

Imagine you only have 4 different letters and you need to form words that represent 20 different meanings. How many letters should each word have so that you can create enough unique words for all 20 meanings?

Solution:

Mathematically we know that to find the number of possible combination we had the equation:

a^n , where a is the number of different letters given (4) and n is the length of the word.

- Based on the question statement we have to find n such that **4^n > 20**

The question statement asks us to find the smallest word length that will give us at least 20 unique words, 4^3 is > 20 therefore a **word length of minimum 3 is enough for us**.