Genomics Assignment 1

Shivangi Das (1919523)

March 17, 2020

Part 1: Theory

1. Central dogma: $DNA \rightarrow RNA \rightarrow protein$

   DNA makes RNA makes protein.

2. a Escherichia Coli
   Length: 4.6 Mb, Chromosomes: 1

   b Yeast (Saccharomyces cerevisiae)
   Length: 12.5 Mb, Chromosomes: 16

3. Since the probabilities of A, C, T, G are independent and identically distributed, we can apply multinomial model.

   $$P(s) = \prod_{i=1}^{n} p(s(i))$$

   Given $p(A) = \frac{1}{2}$, $p(C) = \frac{1}{6}$, $p(G) = \frac{1}{6}$ and $p(T) = \frac{1}{6}$
   Probability of sequence 'ACGTACGTACGT'

   $$\begin{aligned} &= p(A) * p(C) * p(G) * p(T) * p(A) * p(C) * p(G) * p(T) * p(A) * p(C) * p(G) * p(T) \\ &= 1/2 * 1/6 * 1/6 * 1/6 * 1/2 * 1/6 * 1/6 * 1/6 * 1/2 * 1/6 * 1/6 * 1/6 \\ &= 1/80621568 \\ &= 1.24 * 10^{-8} \end{aligned}$$

4. Probability of an ORF of length 150 codons in a random sequence, requires that we find a start codon (ATG) followed by none of the 3 stop codons for next 149 codons. We'll assume **the distribution of codons to be uniform** in which case it is 3/64 probability to pick a stop codon and 61/64 to pick a non-stop codon (total codons = 64).

   This can be calculated as P(run of 148 codons) = $(\frac{61}{64})^{148} = 0.00082$

   There is a 0.00082 probability of a codon of length 150 which means we can conclude with 99.92% confidence that it is highly significant.

5. Homology refers to similarity between DNA sequences which is a result of common ancestors and not by chance. To measure similarity:

   - first perform sequence alignment using global alignment or local alignment

   - use the Needleman-Wunch algorithm (for global alignment) or Smith-Waterman algorithm (local alignment) to find the best alignment score (uses scoring function for getting alignment score).

   - compare with score obtained from alignment of random sequences of similar type to determine statistical significance. Probability of the two sequences alignment score is calculated (scores equal and above / number of permutations) which is the p-value. If the p-value is less than 0.01 (or the confidence we decide), the two sequences are considered homologous with 99% confidence.

Part 2: Practical

Note: all code in ass1_1919523.m

6. Mitochondrial genome of Pan troglodytes (chimpanzee)

   Scientific name: Pan troglodytes

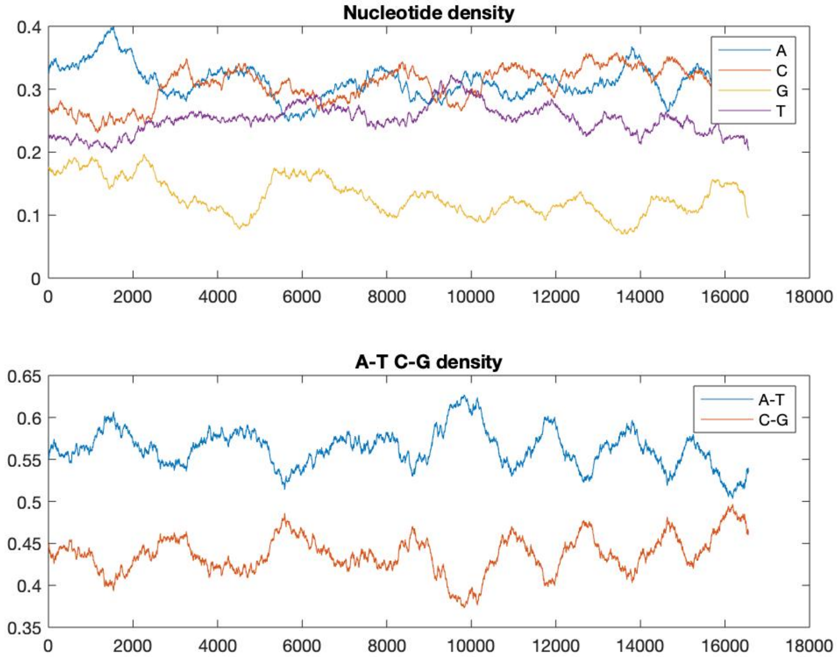
   Accession number: NC_001643

   Sequence length: 16554

   Base count: A: 5154 C: 5099 G: 2133 T: 4168

   GC % : 43.7, Size: 0.02 Mb

   Nucleotide density shown below:



```
Code:
mito_gbk = getgenbank('NC_001643'); %chimp
chimp = mito_gbk.Sequence;
mitochondria_length = length(chimp)
figure
ntdensity(chimp)
```

7. Potential protein coding genes in a Chimpanzee mitochondrial genome can be found using:

```
orf = seqshoworfs(chimp,‘MINIMUMLENGTH’,3,‘geneticcode’,2,‘frames’,‘all’);
```

   Genetic code for Vertebrate Mitochondrial is 2. All frames are considered.

   Threshold for the ORFs: random sequence is generated from the chimp genome by 16554 random permutations and the length of the longest ORF is selected as threshold. The possibility of seeing an ORF of length longer than the random ones makes it a good candidate to be a potential protein encoding gene, since it will be unlikely to be by chance.

```
orf1 = seqshoworfs(chimp(randperm(length(chimp))),‘MINIMUMLENGTH’,3,‘geneticcode’,2,
‘frames’,‘all’);

ORFLength1=[];
for i=1:6
```

```
    for j=1:length(orf1(i).Stop)
     ORFLength1=[ORFLength1; orf1(i).Stop(j)+2 - orf1(i).Start(j)];
    end
end

max_threshold=max(ORFLength1)
n_max=length(find(ORFLength>=max_threshold))
```

Since a codon is made up of 3 nucleotide, by specifying MINIMUMLENGTH as $max\_threshold/3$, we get 12 potential genes through this process. In real there are 13 genes. We could reduce the threshold in order to detect potential genes of shorter length. (Note: running the code multiple times may result in answers between 10-13. This is because we are using random permutations which change every time, changing the threshold.)

8. The first potential protein encoding gene can be extracted using

   potential_protein = chimp(orf(1).Start(1):orf(1).Stop(1)+2)

   The p-value is calculated using the method in answer 4 assuming equal probability of each codon (normal distribution).

   Test statistic: ORF length calculated as $length(potential\_protein)/3 = 317$ codons.

   Null hypothesis: the chosen ORF is generated by a random process.

   Significance level = 0.01

   Using uniform distribution assumption, the probability of seeing a codon of length 317 is $\frac{61}{64}^{315} = 0.00000027$ (p-value).

   The p-value is much smaller than the significance level. Therefore, there's a 99.999973% confidence of this being a significant gene.

9. To convert sequence to aminoacid we use

   aminoacid = nt2aa(sequence, 'geneticcode',2)

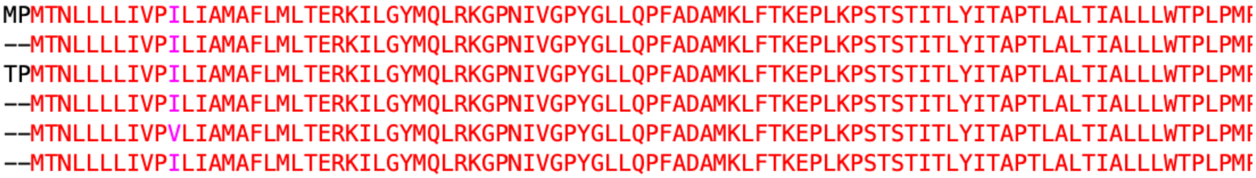
   Amino acid (first 50) = 'MTNLLLLIVPILIAIAFLMLTERKILGYIQLRKGPNIVGPYGLLQPFADA'

   Common chimpanzee and 5 closest organisms

Name	Taxonomic Name	Protein	Identity score
Common chimpanzee	Pan troglodytes	NADH-ubiquinone oxidoreductase chain 1	618
Nigeria-Cameroon chimpanzee	Pan troglodytes ellioti	NADH dehydrogenase subunit 1	616
Western chimpanzee	Pan troglodytes verus	NADH dehydrogenase subunit 1	615
Eastern chimpanzee	Pan troglodytes schweinfurthii	NADH dehydrogenase subunit 1	615
Bonobo	Pan paniscus	NADH dehydrogenase subunit 1	613
Central chimpanzee	Pan troglodytes troglodytes	NADH dehydrogenase subunit 1	613

   Multialignment can be generated using

```
seq1 = getgenpept('Q9T9W3','SequenceOnly',true);
seq2 = getgenpept('AIV00479','SequenceOnly',true);
seq3 = getgenpept('AEQ36262','SequenceOnly',true);
seq4 = getgenpept('ANQ92411','SequenceOnly',true);
seq5 = getgenpept('AMB65312','SequenceOnly',true);
seq6 = getgenpept('AJO25286','SequenceOnly',true);
seqs = {seq1, seq2, seq3, seq4, seq5, seq6};
showalignment(multialign(seqs))
```



```
MPMTNLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNIVGPYGLLQPFADAMKLFTKEPLKPSTSTITLYITAPTLALTIALLLWTPLPM
--MTNLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNIVGPYGLLQPFADAMKLFTKEPLKPSTSTITLYITAPTLALTIALLLWTPLPM
TPMTNLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNIVGPYGLLQPFADAMKLFTKEPLKPSTSTITLYITAPTLALTIALLLWTPLPM
--MTNLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNIVGPYGLLQPFADAMKLFTKEPLKPSTSTITLYITAPTLALTIALLLWTPLPM
--MTNLLLLIVPVLIAMAFLMLTERKILGYMQLRKGPNIVGPYGLLQPFADAMKLFTKEPLKPSTSTITLYITAPTLALTIALLLWTPLPM
--MTNLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNIVGPYGLLQPFADAMKLFTKEPLKPSTSTITLYITAPTLALTIALLLWTPLPM
```

We can see that the sequence discovered using p-value analysis is an actual protein found in many different types of chimpanzees (even human). This shows that we can detect genes accurately with probability and statistical analysis.

10. Human mitochondrion genome

    Accession no: NC_012920

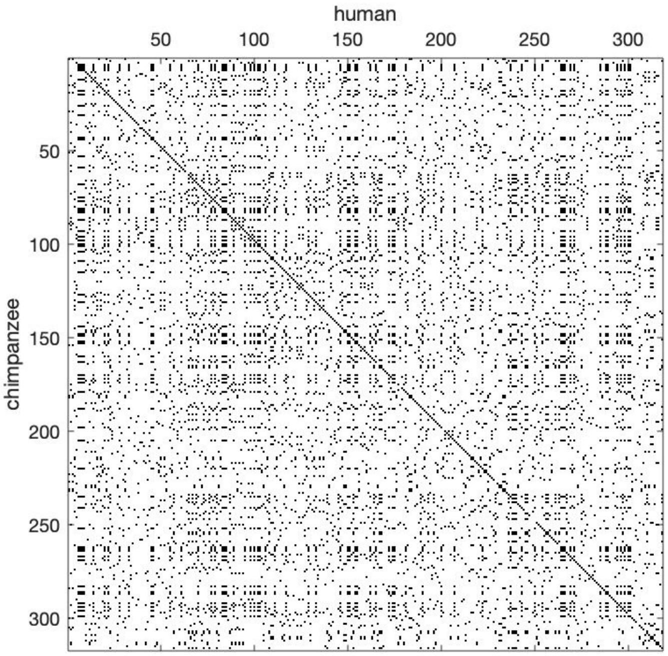
    Name : Homo sapiens mitochondrion

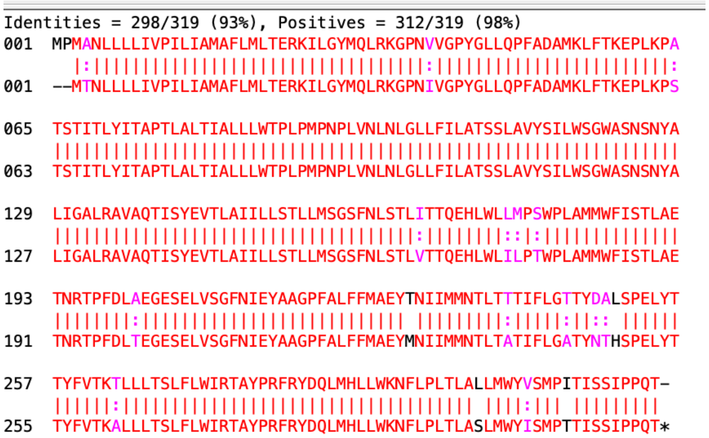
    GC%: 44.4

    Can be seen from BLASTing the discovered protein that it is also found in human mitochondrion. The protein is NADH dehydrogenase subunit 1 (mitochondrion) AFF91323. We can get the sequence using

    ```
    humanProtein = getgenpept('AFF91323','SequenceOnly',true);
    ```

    A dotplot of sequence matches shows an almost perfect match. The alignment (BLOSUM50) is shown below



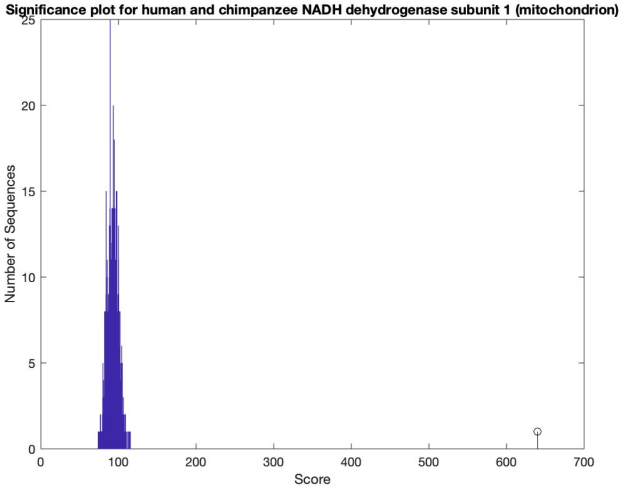


```
Identities = 298/319 (93%), Positives = 312/319 (98%)
001  MPMANLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEPLKPA
       |:||||||||||||||||||||||||||||||||||:||||||||||||||||||||||||:
001  --MTNLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNIVGPYGLLQPFADAMKLFTKEPLKPS

065  TSTITLYITAPTLALTIALLLWTPLPMPNPLVNLNLGLLFILATSSLAVYSILWSGWASNSNYA
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
063  TSTITLYITAPTLALTIALLLWTPLPMPNPLVNLNLGLLFILATSSLAVYSILWSGWASNSNYA

129  LIGALRAVAQTISYEVTLAIILLSTLLMSGSFNLSTLITTQEHLWLLMPSWPLAMMWFISTLAE
     |||||||||||||||||||||||||||||||||||||:||||||||::|:||||||||||||||
127  LIGALRAVAQTISYEVTLAIILLSTLLMSGSFNLSTLVTTQEHLWLILPTWPLAMMWFISTLAE

193  TNRTPFDLAEGESELVSGFNIEYAAGPFALFFMAEYTNIIMMNTLTTTIFLGTTYDALSPELYT
     ||||||||:||||||||||||||||||||||||||| |||||||||:|||||:||:: ||||||
191  TNRTPFDLTEGESELVSGFNIEYAAGPFALFFMAEYMNIIMMNTLTATIFLGATYNTHSPELYT

257  TYFVTKTLLLTSLFLWIRTAYPRFRYDQLMHLLWKNFLPLTLALLMWYVSMPITISSIPPQT-
     ||||||:|||||||||||||||||||||||||||||||||||| ||||:||| |||||||||
255  TYFVTKALLLTSLFLWIRTAYPRFRYDQLMHLLWKNFLPLTLASLMWYISMPTTISSIPPQT*
```

Global alignment gives a score of 640 for scoring matrix BLOSUM50 and 393.6 for BLOSUM30. A significance test gives a p-value of less than 1 in a 1000.

Code:

```
seqdotplot(humanProtein,chimpProtein)
[sc50,globAlig50] = nwalign(humanProtein,chimpProtein);
[sc30,globAlig30] = nwalign(humanProtein,chimpProtein,'scoringmatrix','blosum30');
```



This is expected since humans and chimpanzees are descended from same ancestors, primates and therefore their genes are homologous.

11. First the human mitochondrion genome is processed to find significant protein coding regions following the same process as the chimpanzee genome in answer 7. This results in 11 sequences, and we had 12 sequences from chimpanzee. They are all subjected to global alignment that yields 132 scores. The alignments with p-value more than 0.01 are further considered and significant plot is shown.

    Test statistic: global alignment score compared with scores of 1000 random permutations of similar type.

    Null hypothesis: the alignment is generated by a random process.

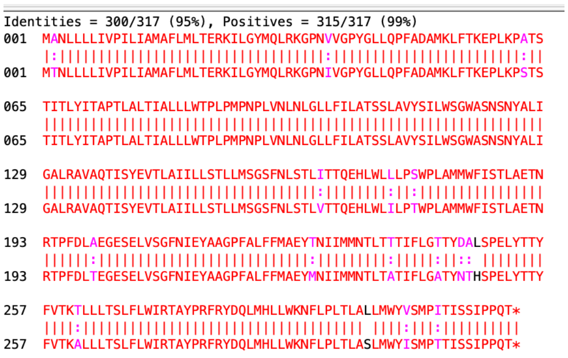
    Significance level = 0.01

    Finding: Doing a statistical test yields 12 of the alignments to be significant since the p-value is less than 1 in a 1000 for most and 0.005 and 0.007 for two. Can also be seen that these two match with other sequences from chimpanzee with a lesser p-value. Since humans and chimpanzees do share ancestors, we end up finding multiple homologous genes.

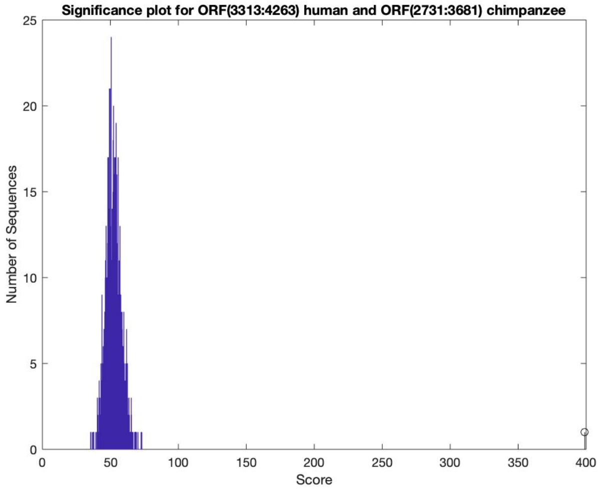
```
[score1, alignment] = nwalign(human_protein, chimp_protein, ...
                    'scoringmatrix','blosum30','gapopen',5,'extendgap',5);
```

    The code in ass1_1919523.m can be used to view all the alignments and significance plots. Below are all the results with alignments and significance plot for two:

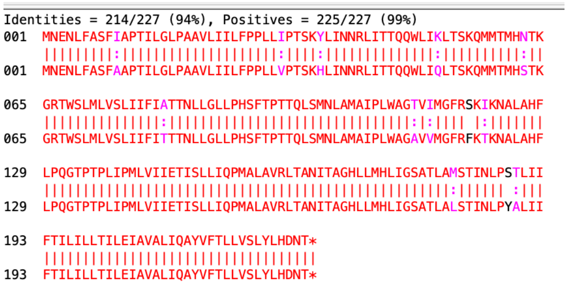
    (a) ORF in position(3313:4263) in human is homologous to ORF in position (2731:3681) in chimpanzee (p-value = 0).



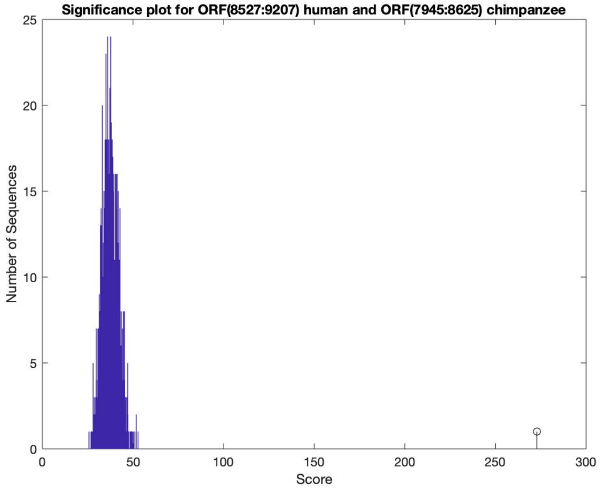
```
Identities = 300/317 (95%), Positives = 315/317 (99%)
001  MANLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEPLKPATS
     |:||||||||||||||||||||||||||||||||||:||||||||||||||||||||||||:||
001  MTNLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNIVGPYGLLQPFADAMKLFTKEPLKPSTS

065  TITLYITAPTLALTIALLLWTPLPMPNPLVNLNLGLLFILATSSLAVYSILWSGWASNSNYALI
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
065  TITLYITAPTLALTIALLLWTPLPMPNPLVNLNLGLLFILATSSLAVYSILWSGWASNSNYALI

129  GALRAVAQTISYEVTLAIILLSTLLMSGSFNLSTLITTQEHLWLLLPSWPLAMMWFISTLAETN
     |||||||||||||||||||||||||||||||||||:||||||||:||:||||||||||||||||
129  GALRAVAQTISYEVTLAIILLSTLLMSGSFNLSTLVTTQEHLWLILPTWPLAMMWFISTLAETN

193  RTPFDLAEGESELVSGFNIEYAAGPFALFFMAEYTNIIMMNTLTTTIFLGTTYDALSPELYTTY
     ||||||:|||||||||||||||||||||||||||:|||||||||:|||||:||:: ||||||||
193  RTPFDLTEGESELVSGFNIEYAAGPFALFFMAEYMNIIMMNTLTATIFLGATYNTHSPELYTTY

257  FVTKTLLLTSLFLWIRTAYPRFRYDQLMHLLWKNFLPLTLALLMWYVSMPITISSIPPQT*
     ||||:|||||||||||||||||||||||||||||||||||| ||||:|||:||||||||||
257  FVTKALLLTSLFLWIRTAYPRFRYDQLMHLLWKNFLPLTLASLMWYISMPTTISSIPPQT*
```



    (b) ORF in position(8527:9207) in human is homologous to ORF in position (7945:8625) in chimpanzee (p-value = 0).



```
Identities = 214/227 (94%), Positives = 225/227 (99%)
001  MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTK
     |||||||||:||||||||||||||||||||:||||:||||||||||||||:||||||||||:||
001  MNENLFASFAAPTILGLPAAVLIILFPPLLVPTSKHLINNRLITTQQWLIQLTSKQMMTMHSTK

065  GRTWSLMLVSLIIFIATTNLLGLLPHSFTPTTQLSMNLAMAIPLWAGTVIMGFRSKIKNALAHF
     |||||||||||||||:|||||||||||||||||||||||||||||||:|:|||| |:|||||||
065  GRTWSLMLVSLIIFITTTNLLGLLPHSFTPTTQLSMNLAMAIPLWAGAVVMGFRFKTKNALAHF

129  LPQGTPTPLIPMLVIIETISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLII
     ||||||||||||||||||||||||||||||||||||||||||||||||||||:|||||| :|||
129  LPQGTPTPLIPMLVIIETISLLIQPMALAVRLTANITAGHLLMHLIGSATLALSTINLPYALII

193  FTILILLTILEIAVALIQAYVFTLLVSLYLHDNT*
     |||||||||||||||||||||||||||||||||||
193  FTILILLTILEIAVALIQAYVFTLLVSLYLHDNT*
```



(c) ORF in position(8527:9207) in human is homologous to ORF in position (14165:15325) in chimpanzee (p-value = 0.0050)

(d) ORF in position(12343:14148) in human is homologous to ORF in position (11761:13566) in chimpanzee (p-value = 0).

(e) ORF in position(7586:8269) in human is homologous to ORF in position (7003:7686) in chimpanzee (p-value = 0).

(f) ORF in position(10760:12184) in human is homologous to ORF in position (10178:11602) in chimpanzee (p-value = 0).

(g) ORF in position(14747:15907) in human is homologous to ORF in position (14165:15325) in chimpanzee (p-value = 0).

(h) ORF in position(4578:5513) in human is homologous to ORF in position (4191:4931) in chimpanzee (p-value = 0).

(i) ORF in position(5904:7445) in human is homologous to ORF in position (2731:3681) in chimpanzee (p-value = 0.0070).

(j) ORF in position(5904:7445) in human is homologous to ORF in position (5321:6862) in chimpanzee (p-value = 0).

(k) ORF in position(9207:10055) in human is homologous to ORF in position (8625:9473) in chimpanzee (p-value = 0).

(l) ORF in position(10323:10766) in human is homologous to ORF in position (9888:10184) in chimpanzee (p-value = 0).

12. We can use BLAST to find alignments easily as done in question 10. A way to do it manually is described in question 11. An alternative can be to check sequence of nucleotide instead of aminoacids.

```
[score1 , alignment] = nwalign(human_NT , chimp_NT , ...
    'scoringmatrix','blosum30','gapopen',5,'extendgap',5,'ALPHABET','NT');
```

Same process is repeated as anwser 11 and if the p-value is less than 0.01, the sequence pair may be homologous.

Few examples:

(a) ORF in position(3313:4263) in human is homologous to ORF in position (2731:3681) in chimpanzee.

(b) ORF in position(8527:9207) in human is homologous to ORF in position (7945:8625) in chimpanzee

Although this process can be very slow (much slower than comparing aminoacid) for large number of matching genes, it still gives accurate results.