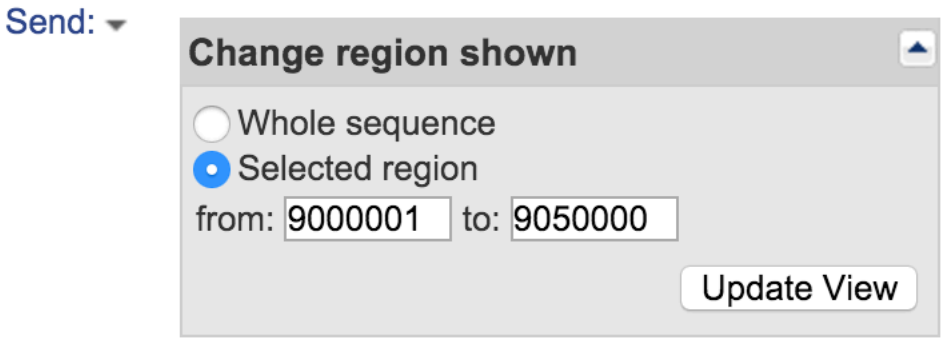
Background and purpose of project

Multiple sequence alignment (MSA) is generally the alignment of three or more biological sequences of similar length. From this output, the homology and evolutionary relationships between sequences can be inferred and studied. However, the computational problem of multiple sequence alignment is difficult. Therefore, heuristic algorithms have been developed in order to combat this problem that involve building phylogenetic trees and doing MSA. Because they are both techniques of joint sequence comparison, they can synergistically help each other. An improved tree can serve as a guide for an improved MSA, and that MSA can be used to build an even better tree.

C. elegans nucleotide sequence acquisition

- The last digit of my Columbia UNI is 8, so n = 8.
- I needed to consider the 50,000 nucleotides starting right after the nucleotide numbered (n+1) × 1,000,000 = 9,000,000.
- I selected the nucleotides 9000001 to 9050000




```
>> data=genbankread('sequence.gb');
>> s=data.Sequence;
>> lc2cap;
```

- The commands above were used in order to read the data from the `sequences.gb` file, save the sequence, and then convert the lowercase bases in the sequence to capital letters.

Identification of the largest open reading frame (ORF) and acquisition of protein sequence

```
>> x = union (findstr(s, 'TAA'), findstr(s, 'TAG'));
>> x = union(x, findstr(s, 'TGA'));
>> x1 = x(rem(x, 3) == 1); x2 = x(rem(x, 3) == 2); x3 = x(rem(x,
3) == 0);
>> [a(1) b(1)] = max(diff(x1));
>> [a(2) b(2)] = max(diff(x2));
>> [a(3) b(3)] = max(diff(x3));
>> %all of this was encapsulated into orf.m, which was used
instead for both the forward and reverse strands
>> s = rcompstrand(s); %The reverse strand
```

- I found that the ORF for s was m (the length of the maximum ORF) was 942 and I = 2 (meaning that it was in the second reading frame)
- I found that the ORF for rcompstrand(s) was m = 903 and I = 1
- Therefore, the largest ORF was in s where m = 942 and I = 2

```
>> orfmax =

ATCAATGAATATTCTTCTAAATTTATATTATTATTCAGATCTGCTGCTCTTGATGAAGGAAATG
AGTTTGTCAACCAACAAAACGCTGATGGAACATTCCTTCGTAACAATACAGGACATAAGAACAC
TGATGAGCATCTCAGTCACAACGTGCTTGATGAGAATGCTCAAATGTCGATTGGAGCAGATGGA
ACTTCCCACAATATTACCAACCGAAAGGGATCAGTTGGGGACTCACATAACGCTGCCTCGGATG
CTCATTCCAACTTTGAAAGCCTTGATGCTCAAGGAAACAAGAAATCTCAAAACTACAGTAAGAA
GGCAGCCTCTGCTTCCGGCTCCAATGCTGACTTCGAGTCTAACTTGGAGTCTCTCAAGAATGCC
GATGGAACATCAATGTCTAACTCAACTGGAAACTTCAACAACACTAGCTATGACAAGGCAACGG
CCGAGGAAGTTATGTCAAAGAAGAATGTCAATGCTGATGGAACATCTTCAATGGAAGCTAGTCA
TGCTGGAAGTAACAGCAGCAAGATCAACTCTGCATCTGGACAATCTTCGGATCTTAGCATGGTT
GGACCAAATGGAATTAAGAGTCACAGTACAAGTAATAAAACAGACAACTATGCTTTGGATGAGG
CTAACCAAAGTGCTGGAAGTATCAGTGAGCAAATTGGAAAGAATGGACAAAGATCTCTTAATGA
ATCAAGCATCGAGAGTGGAAGAAAGGTTAGTTTAACTGAAAAGTTAACAGAGATTTCTAAACTA
ACAATTTCCAGGCAGAAAGCAGAAACAACACCGCTGCCGATACTCTTGACTCAGTCGATGCCAA
CGGAACTGTTAGCTCATCTCATAGTAAATCTGCAAGTGGAACATCTTTGGATGAGAATCATAAT
AAGACTCATGCTCTTCAAGCTTCAGTCGATGAGCACGGAAACA
```

- The max ORF that I found was an exon in one of the genes mentioned:

```
join(17607..17745,17809..18065,18242..18378,18421..19111,

19162..19494,19538..19663,19725..20267,20323..20853)
                     /gene="F59B2.12"
```

- a = {807 942 777}
- b = {50 346 337}
- I looked at the x2 array in my MATLAB workspace. The ORF was between positions 18380 and 19322 (346 and 347 from b)
-

```
>>data=genpeptread('protein.gp');
>> p1=data.Sequence;
```

- I created p1 for the full amino acid sequence that describes the protein encoded by the gene that contains this exon I found.
- **QUESTION:** Are there any amino acids that resulted from <u>codons that were split between two consecutive exons</u>? (in other words one of the nucleotides of the codon is in one exon and the other two in the other). If yes, identify them; if not, explain how you reached that conclusion.
    - There were amino acids that resulted from codons that were split between two consecutive exons. You can see that when you look at where exons are joined, they are not exact multiples of 3. I did not pursue this sequence because according to the note on NCBI this sequence was incomplete at both ends, so I would not have been able to determine the exact locations of the amino acids that were split between two exons effectively.

```
join(17607..17745,17809..18065,18242..18378,18421..19111,

19162..19494,19538..19663,19725..20267,20323..20853)

  (A)  17607..17745 = 138/3 = 46
  (B)  17809..18065 = 256/3 = 85.3333 (likely 255/3 --> 85)
  (C)  18242..18378 = 136 /3 = 45.33 (likely 138 /3 --> 46)
  (D)  18421..19111 = 690 / 3 = 230
  (E)  19162..19494 = 332/3 = 110.66666
```

```
(F)  19538..19663 = 125 / 3 = 41.666666
(G)  19725..20267 = 542 / 3 = 180.66666
(H)  20323..20853 = 530 / 3 = 176.666
```

Likewise, this uncharacterized sequence, when run against the BLAST database, had exactly 5 hits, almost all from the *Caenorhabditis* genera, so this entire portion was repeated again with the next largest ORF.

REPEAT: ORF and protein acquisition (WITH SECOND LARGEST ORF)

- I found that the ORF for rcompstrand(s) was m = 903 and I = 1 (this is the second largest one)
- Note that this protein found (IKKϵ), when activated, promotes the activation of NFKβ, which is involved in inflammatory cytokine expression.

```
TCAAAAACGTCAGCTAAGATGATAAAACACATTCAATTTCAGCACCCCTTCCTGGCACACGAGA
TGGTCGATCCACTGATGGCTCAAAACAGACATAATTGGAAAACAAAATCAGCTTATACTTCGGA
ACAATGTGATCTATGGGCTCTGGGATGTACACTATATTTTTGTGCCACTGGAAAGTTCCCATTC
GAGCACGAAAGGAATAACAAGAGCCTGTATCATAAAGCTGTAGTGGCTCTTACTCAGAATCCGG
ATGCAATCGCAATGGTTTTAGTGCAAAAAGGACGAGATCCAGGAAGAAGAACTGATATATTTGA
ATTCCAACCAGTTACAGAACTTCCTGCTAAATTCACAAGATATCCGAAATGGCTCGTAAGCACA
ATGACATGCCTTCTACGCAGCTTTTTCCATGAACCTTCAATCGAGTATTATGCAAAAGTGGCTG
ATGCTATGAGAAATTCTAAACGAAGAACATTTTCATCAGTGGATCAAATGTCGATTGTTGAGCA
CACGGATATGTCTAATGTTCCACATCTTGGATTCAGTATTCCAAGTATTTCAAAATGTCTTGGA
TATCCAGAAGGAACCGATATACTTCTTCTCTCAAATACCTCCACCCACTACCTTGATTCCAAAC
AAAAATCCGTTGACGGACTTCCCGATGATTTGTATTTGGTCGTTCCCCAGACAAGTCACGTTGA
TATGAGAAAGATTTTGGCGAGAAATATCGAATTCCACGAATTCGATGATATGACAGACAGAAAA
CTTTCTGAAATTCGAATCAAAAAATGTTATGAAGGATTGAGCATGCTAACTGAAATTGACGAGT
ATTTGGCTCTTTTTGATCGAGTCTCCACAATTTTGTCAACCCAATTTTCACTGGTGAGTTTTGC
ATTA
```

- a = 903 897 684
- b = 2 21 727
- x1 = 10 913 (pos 1 and 2)

```
complement(join(46355..46411,46498..46585,47676..47731,
                47843..48646,49104..>49946))
                /gene="ikke-1"
                /locus_tag="CELE_R107.4"
```

/standard_name="R107.4c"
/note="Confirmed by transcript evidence"
/codon_start=1
/product="Inhibitor of nuclear factor kappa-B kinase
epsilon subunit homolog 1"

(A) 46355..46411 = 56/3 = 18.66666 (57/3 → 19)
(B) 46498..46585 = 87 / 3 = 29 (84 / 3 → 28)
(C) 47676..47731 = 55 / 3 = 18.33333 (57 /3 → 19)
(D) 47843..48646 = 803 / 3 = 267.6666 (804 / 3 → 268)
(E) 49104..>49946 = 842 / 3 = 280.66666 (approximately 841, but something might be cut off at the end of this sequence, hence the ..>)

THE ONES HIGHLIGHTED IN RED ARE SPLIT BETWEEN TWO EXONS

MAVSPHKTYPIVITHGEK**Y** % EXON (A)

TLFNDESIGKGAYSEVYRGRTESGRLVA % EXON (B)

VKTACKKLEVAAIGIEIEI % EXON (C)

LKKLKGASNIVQYFGSNHTKMAPGSVTSETISFAMEYASSSLEAEMRRPKNHRGLSSNALIDLV
VDCSMALSALREHNIAHRDIKHMNILLFPGTPTRGRRSTHLFKLCDMGCSKSLSENSSHEMRTL
VGTPNLLHPFLAHEMVDPLMAQNRHNWKTKSAYTSEQCDLWALGCTLYFCATGKFPFEHERNNK
SLYHKAVVALTQNPDAIAMVLVQKGRDPGRRTDIFEFQPVTELPAKFTRYPKWLVSTMTCLLRS
FFHEPSIEYYA**K** % EXON (D)

To answer the question: yes there are codons split between two exons. They are bolded and underlined above.

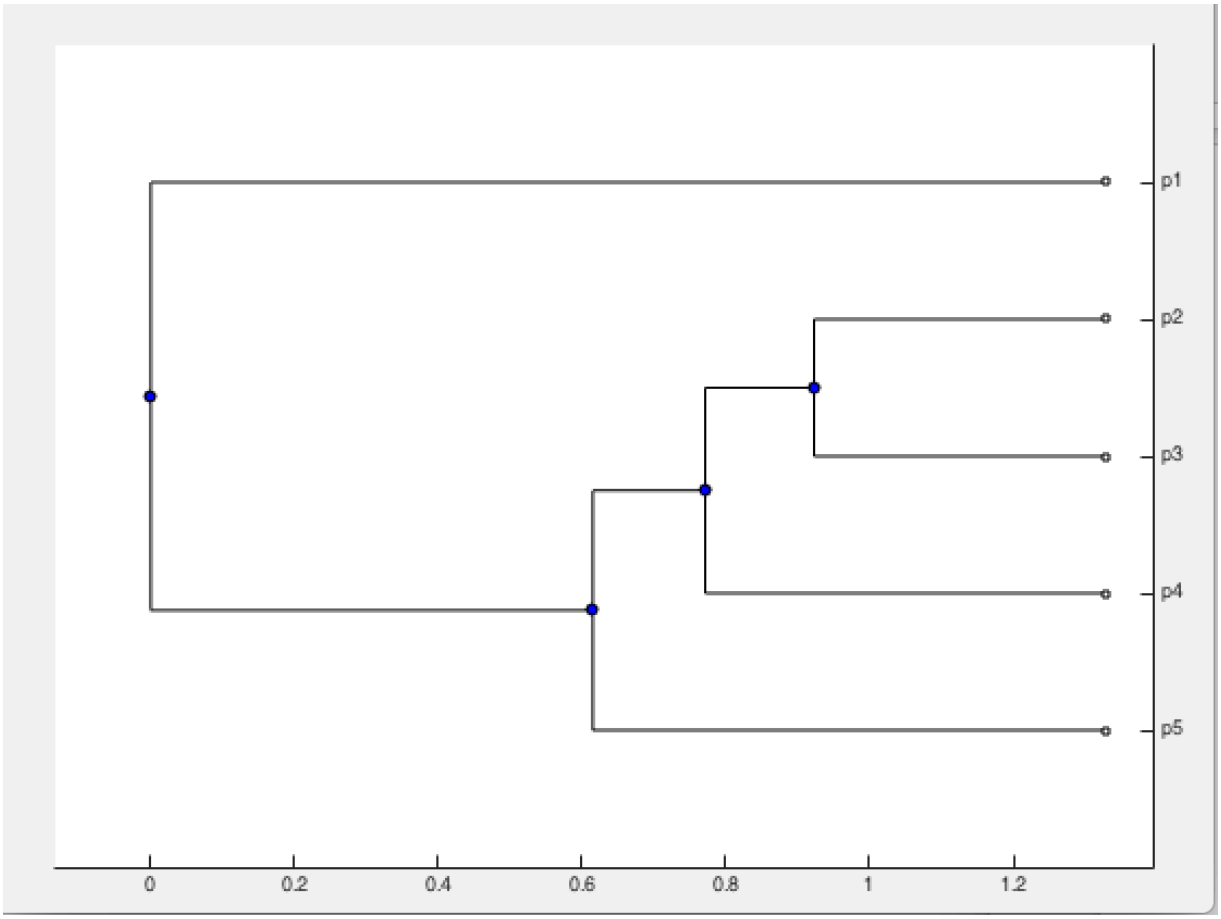
Protein BLAST

5 hits were picked

☑	Inhibitor of nuclear factor kappa-B kinase epsilon subunit homolog 1 [Caenorhabditis elegans]	1704	1704	100%	0.0	100%	NP_871628.1
☐	Inhibitor of nuclear factor kappa-B kinase epsilon subunit homolog 1 [Caenorhabditis elegans]	1699	1699	100%	0.0	99%	NP_871627.1
☐	Inhibitor of NFKappaB Kinase Epsilon subunit homolog [Caenorhabditis elegans]	1606	1606	94%	0.0	99%	NP_001022708.1
☐	Inhibitor of nuclear factor kappa-B kinase epsilon subunit homolog 1 [Caenorhabditis elegans]	1568	1568	91%	0.0	100%	NP_499002.2
☐	Protein CBR-IKKE-1 [Caenorhabditis briggsae]	1203	1203	98%	0.0	71%	CAP27065.2
☐	Uncharacterized protein CELE_Y39G8B.5 [Caenorhabditis elegans]	1187	1187	89%	0.0	78%	NP_496929.2
☐	CRE-IKKE-1 protein [Caenorhabditis remanei]	1177	1177	88%	0.0	74%	XP_003092674.1
☐	hypothetical protein Y39G8B.e - Caenorhabditis elegans	1169	1169	88%	0.0	78%	T26770
☐	hypothetical protein CAEBREN_21440 [Caenorhabditis brenneri]	1111	1111	98%	0.0	66%	EGT56783.1
☐	C. briggsae CBR-IKKE-1 protein [Caenorhabditis briggsae]	947	947	73%	0.0	75%	XP_002642422.1
☑	hypothetical protein Y032_0009g663 [Ancylostoma ceylanicum]	299	299	85%	8e-86	32%	EYC27343.1
☐	hypothetical protein Y032_0009g663 [Ancylostoma ceylanicum]	292	292	85%	3e-83	32%	EYC27344.1
☐	hypothetical protein Y032_0009g663 [Ancylostoma ceylanicum]	289	289	85%	5e-82	31%	EYC27345.1
☑	hypothetical protein OESDEN_13077 [Oesophagostomum dentatum]	275	275	71%	3e-78	33%	KHJ87153.1
☑	Serine threonine protein kinase-related domain containing protein [Haemonchus contortus]	275	275	82%	5e-77	30%	CDJ83327.1
☑	YjeF domain protein [Ancylostoma duodenale]	277	277	83%	7e-76	30%	KIH65189.1

```
>> seqs = fastaread('fiveproteins.txt');
>> seqs(1).Header = 'p1';
>> seqs(2).Header = 'p2';
>> seqs(3).Header = 'p3';
>> seqs(4).Header = 'p4';
>> seqs(5).Header = 'p5';
>> distances = seqpdist(seqs);
>>tree = seqlinkage(distances,'UPGMA',seqs); view(tree)
```

- **QUESTION:** What do the above commands (BOLDED) mean?
  - distance = seqpdist(*Seqs*) returns distance, a vector containing biological distances between each pair of sequences stored in the sequences of *Seqs*.
  - tree = seqlinkage(*Distances*, Method, Name) returns a phylogenetic tree object from the pairwise distances, *Distances*, between the species or products. In this case, UPGMA clustering method was used. The seqs at the end is the name of how to label the leaf nodes in the phylogenetic tree.
  - Finally, view(tree) displays the UPGMA tree



The UPGMA algorithm constructs a rooted tree (dendrogram) that reflects the structure present in a pairwise similarity matrix.

This phylogenetic tree shows the relationships between different protein sequences (p1 – p5). It shows that, for example, the protein sequences p2 and p3 are most closely related to each other. P4 is closely related to (p2, p3). P(5) is closely related to (p4(p3,p2)). And finally, p1 is the least closely related to all other protein sequeuences. UPGMA is used for the creation of phenetic trees (phenograms). In a phylogenetic context, UPGMA assumes a constant rate of evolution (molecular clock hypothesis), and is not a well-regarded method for inferring relationships unless this assumption has been tested and justified for the data set being used. UPGMA was initially designed for use in protein electrophoresis studies, but is currently most often used to produce guide trees for more sophisticated phylogenetic reconstruction algorithms.

Alignment using the Smith-Waterman algorithm

- Locally align two sequences using Smith-Waterman algorithm
- The alignments of the pairs of sequences with the highest and lowest score were emailed to the TA as `swalignHIGHEST.mat` and `swalignLOWEST.mat`.

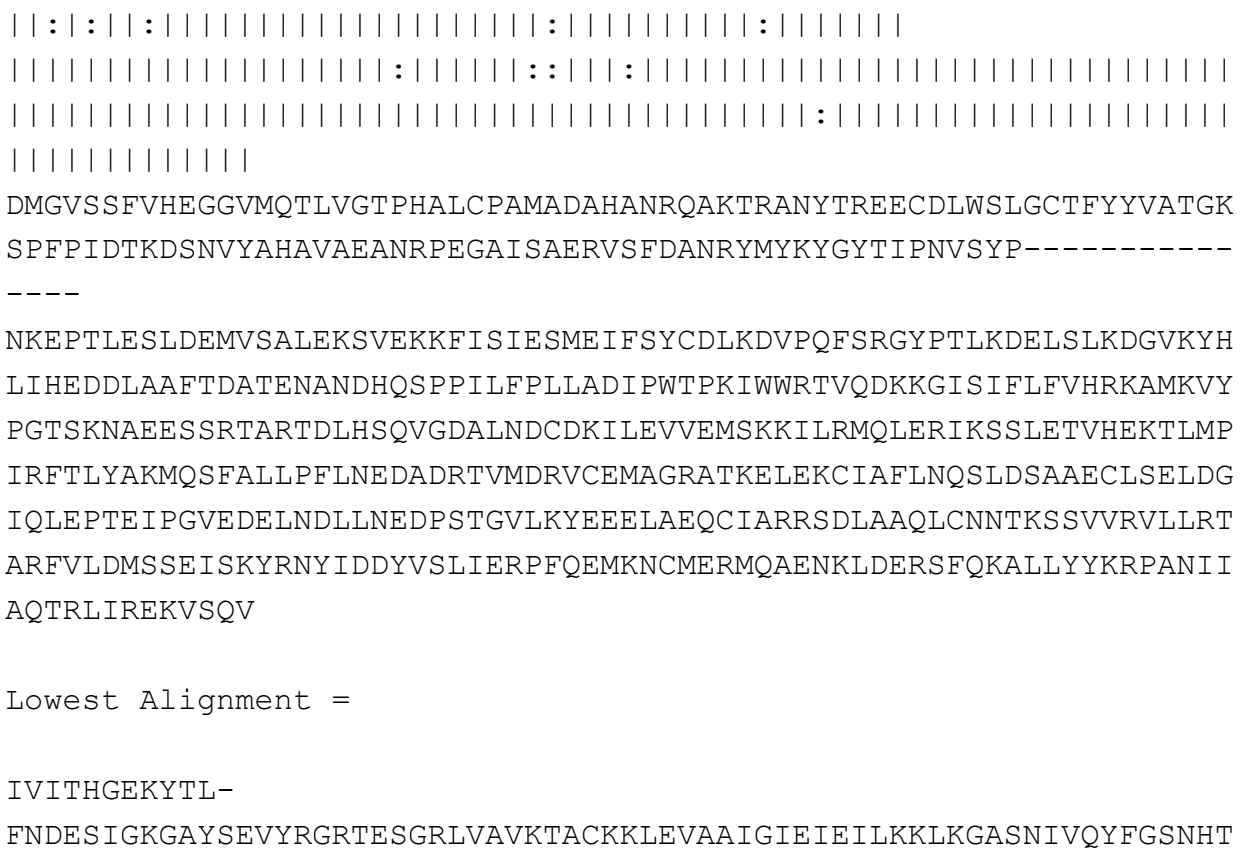
```
>> Score = swalign(seqs(1), seqs(2))

Score =

  375.3333
>> Score = swalign(seqs(1), seqs(3))

Score =

  331.3333

>> Score = swalign(seqs(1), seqs(4))

Score =

  333.6667

>> Score = swalign(seqs(1), seqs(5)) ***LOWEST SCORE

Score =

  316.3333
>> Score = swalign(seqs(2), seqs(3))

Score =

        1042
>> Score = swalign(seqs(2), seqs(4))

Score =

  991.3333
```

>> Score = swalign(seqs(2), seqs(5)) ***HIGHEST SCORE

Score =

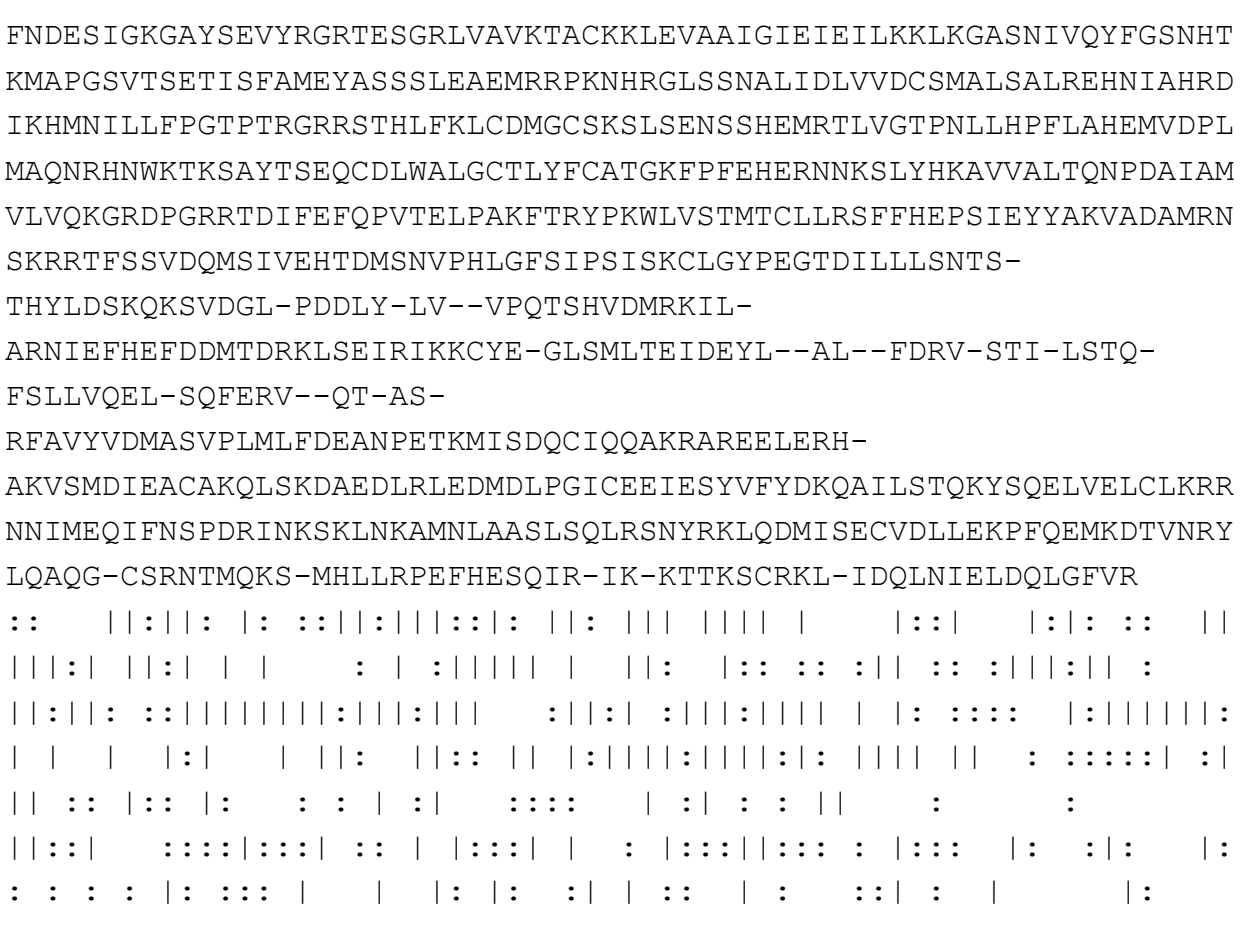
1.2877e+03

```
>> Score = swalign(seqs(3), seqs(4))

Score =

  817.3333

>> Score = swalign(seqs(3), seqs(5))

Score =

  942.6667

>> Score = swalign(seqs(4), seqs(5))

Score =

  882.3333

Highest Alignment =

EIFSYCDLKDVPQFSRGYPTLKDELSLKDGVKYQLIYEDDLVVFADATESGKPHQNSPILFPLL
TDIPWTPKIWWRTIQDKKG---------------P--
SKNGEESSRTARIDMHSQVGDALNDCDKILDVVEMSKKILRVQLERIKSNLESVHEKTLMPIRF
TLYAKMQSFALLPFLSEEADKTVMDRVCEMAGRATKELEKCVAFLNQSLDSATECLSELDKIQL
EPTEIPGVEDELNDLLNDDPSTGVFHYEEQLAEQCIARRSDLAAQLCNNTKSSVVRVLLRTARF
VLDMSSEISKYRNYIDDYVSLIERPFQEMKNCMERMQAEHKLDERSFQKALLYYKRPANIIAQT
RLIREKVSQV

:||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||:
||:||||::|:||||::: ||: |||||||:||||||||||||:|||||
|  |||:||||||||
|:|||||||||||||||:||||||||||:|||||||:||:||||||||||||||||||||||||
```



DMGVSSFVHEGGVMQTLVGTPHALCPAMADAHANRQAKTRANYTREECDLWSLGCTFYYVATGK
SPFPIDTKDSNVYAHAVAEANRPEGAISAERVSFDANRYMYKYGYTIPNVSYP----------
----
NKEPTLESLDEMVSALEKSVEKKFISIESMEIFSYCDLKDVPQFSRGYPTLKDELSLKDGVKYH
LIHEDDLAAFTDATENANDHQSPPILFPLLADIPWTPKIWWRTVQDKKGISIFLFVHRKAMKVY
PGTSKNAEESSRTARTDLHSQVGDALNDCDKILEVVEMSKKILRMQLERIKSSLETVHEKTLMP
IRFTLYAKMQSFALLPFLNEDADRTVMDRVCEMAGRATKELEKCIAFLNQSLDSAAECLSELDG
IQLEPTEIPGVEDELNDLLNEDPSTGVLKYEEELAEQCIARRSDLAAQLCNNTKSSVVRVLLRT
ARFVLDMSSEISKYRNYIDDYVSLIERPFQEMKNCMERMQAENKLDERSFQKALLYYKRPANII
AQTRLIREKVSQV

Lowest Alignment =



IVITHGEKYTL-
FNDESIGKGAYSEVYRGRTESGRLVAVKTACKKLEVAAIGIEIEILKKLKGASNIVQYFGSNHT
KMAPGSVTSETISFAMEYASSSLEAEMRRPKNHRGLSSNALIDLVVDCSMALSALREHNIAHRD
IKHMNILLFPGTPTRGRRSTHLFKLCDMGCSKSLSENSSHEMRTLVGTPNLLHPFLAHEMVDPL
MAQNRHNWKTKSAYTSEQCDLWALGCTLYFCATGKFPFEHERNNKSLYHKAVVALTQNPDAIAM
VLVQKGRDPGRRTDIFEFQPVTELPAKFTRYPKWLVSTMTCLLRSFFHEPSIEYYAKVADAMRN
SKRRTFSSVDQMSIVEHTDMSNVPHLGFSIPSISKCLGYPEGTDILLLSNTS-
THYLDSKQKSVDGL-PDDLY-LV--VPQTSHVDMRKIL-
ARNIEFHEFDDMTDRKLSEIRIKKCYE-GLSMLTEIDEYL--AL--FDRV-STI-LSTQ-
FSLLVQEL-SQFERV--QT-AS-
RFAVYVDMASVPLMLFDEANPETKMISDQCIQQAKRAREELERH-
AKVSMDIEACAKQLSKDAEDLRLEDMDLPGICEEIESYVFYDKQAILSTQKYSQELVELCLKRR
NNIMEQIFNSPDRINKSKLNKAMNLAASLSQLRSNYRKLQDMISECVDLLEKPFQEMKDTVNRY
LQAQG-CSRNTMQKS-MHLLRPEFHESQIR-IK-KTTKSCRKL-IDQLNIELDQLGFVR

```
|:  | : :  |::   :  ||   |::  :: :| : : : :::: |::| |  :|
||::|: | |  |: | : : : : : |:  ::| || :|||:  | :::::::: |: || : :
::||  ::||: :|::: :  |  :  :: || :||:| |: ||:::  |: |:  : : | :
:  |  : :: |:  | :::|:: |:|:|:|||||||: ::| :||::  :: ::||: ::  ||
:| | |: |:::   :: :::::|  |:  ::|
MISQDGEEYTIDFS-
RKIGSGAYGNVHPGRASSGRTVAVKCARDAKEIVAARKEVEFYRRCAGARNIVKYVGSQH-
KNATDH-
SPERFSFAMERALFSLDKLMKQGRHLHGLPKDIIIDLLVDSASALNALKTRGIAHRDIKHLNIL
VFPGL-RKGRKSQYLFKFCDMGVS-SFVHEGG-VMQTLVGTPHALCP--A--MADA-HA-
NRQA-KTRANYTREECDLWSLGCTFYYVATGKSPFPIDTKDSNVYAHA-VAEANRPEG-
AISAERVSFDANRY--MYKY-GYT-IP-NVS-YP----N------K----
EPTLESLDEMVSALEKSVEKKFISIESMEIFSYCDLKDVPQFSRGYPTLKDELSLKDGVKYHLI
HEDDLAAFTDATENANDHQSPPILFPLLADIPWTPKIWWRTVQDKKGISIFLFVHRKAMKVYPG
TSKNAEESSRTARTDLHSQVGDALNDCDKILEVVEMSKKILRMQLERIKSSLETVHEKTLMPIR
FTLYAKMQSFALLPFLNEDAD-RTVMDRVCEMAGRATKELEKCIAFLNQSLDSAAECLS-
ELDGIQLEPTEIPGVEDELNDLLNED-PS-TGVLKYEEELAEQCIARRSDLAAQLCNN-TK-S-
SVVRVLLRTARFVLDMSSEISKYRNYIDDYVSLIERPFQEMKNCMER-
MQAENKLDERSFQKALLYYKRPANIIAQTRLIREKVSQVTAQFPLSNVSIFNDEKDYAR
```

Clustal Omega MSA

```
CLUSTAL O(1.2.1) multiple sequence alignment


p1      MAVSPHKTYPIVITHGEKYTLFNDESIGKGAYSEVYRGRTESGRLVAVKTACKKLEVAAI
p4      ----------MICWDGLEYSIDFSKKIGTGAYGNIHPGRTQSGRRVAVKHARDEKEIRAA
p3      --------------------------------FQVHTGRAASGRSVAVKCARDAKEIVAA
p2      ----------MISQDGEEYTIDFSRKIGSGAYGNVHPGRASSGRTVAVKCARDAKEIVAA
p5      ----------MISQDGEEYTIDFSRKIGSGAYGNVHPGRASSGRTVAVKCARDAKEIVAA
                                      ::: **: *** **** * .  *: *

p1      GIEIEILKKLKGASNIVQYFGSNHTKMAPGSVTSETISFAMEYASSSLEAEMRRPKNHRG
p4      RKEVEFYRRCAGAPNIVKYIGSRHKLATD--HSPERFSFAMERALFSLESLVKQVRNLRG
p3      KKEVEFYKRCAGGRNIVKYIGSQYKEATD--RSPERFSFAMERAMFSLDKLMKQVRHLHG
p2      RKEVEFYRRCAGARNIVKYVGSQHKNATD--HSPERFSFAMERALFSLDKLMKQVRHLRG
p5      RKEVEFYRRCAGARNIVKYVGSQHKNATD--HSPERFSFAMERALFSLDKLMKQGRHLHG
          *:*: ::  *. ***:*.**.:.  :    : * :***** *  **:  ::: :. :*

p1      LSSNALIDLVVDCSMALSALREHNIAHRDIKHMNILLFPGTPTRGRRSTHLFKLCDMGCS
p4      LPREVIIDLLRDSVSALTTLKERGIAHRDIKHLNILVFPGS-TKGRRSQYLFKFCDMGVS
p3      LAKDILIDLLVDCASALNTLKSRGIAHRDIKHLNILVFPGP-RKGRKSQYLFKFCDMGVS
p2      LPKDIIVDLLVDSTSALNALKTRGIAHRDIKHLNILVFPGQ-RRGRKSQYLFKFCDMGVS
p5      LPKDIIIDLLVDSASALNALKTRGIAHRDIKHLNILVFPGL-RKGRKSQYLFKFCDMGVS
        *  : ::**: *.  **.:*: : ********:***:***   :**:* :***:**** *

p1      KSLSENSSHEMRTLVGTPNLLHPFLAHEMVDPLMAQNRHNWKTKSAYTSEQCDLWALGCT
p4      SFVHED--GVMQTLVGTPNALCPELAVAY-------AQRRTRTNDNYTREQCDLWSFGCT
p3      SFVHDG--GVMQTLVGTPHVLCPAMADAH-------AHRQAKTRENYTREQCDLWSLGCT
```

```
p2      SFVHEG--GVMQTLVGTPHALCPAMADAH-------ANRQAKTRANYTREECDLWSLGCT
p5      SFVHEG--GVMQTLVGTPHALCPAMADAH-------ANRQAKTRANYTREECDLWSLGCT
        . : :     *:******. * * :*           .:. :*.  ** *:****::***

p1      LYFCATGKFPFEHERNNKSLYHKAVVALTQNPDAIAMVLVQKGRDPGRRTDIFEFQPVTE
p4      LYYVATGAFPFPIDAKDSSVYAQAVAEGIRPEGAISAERVCCG--INRYV--Y----KYG
p3      FYFVSTGKSPFPIDSKDSNVYANAVAEANRPEGAIAAERVCTD--SGRYM--Y----NYS
p2      FYYVATGKSPFPIDTKDSNVYAHAVAEANRPEGAISAERVSFD--ANRYM--Y----KYG
p5      FYYVATGKSPFPIDTKDSNVYAHAVAEANRPEGAISAERVSFD--ANRYM--Y----KYG
        :*: :**  **  : ::..:* .**.   :   **:   *       *    :

p1      LPAKFTRYPKWLVSTMTCLLRSFF-HEPSIEYYAKVADAMRNSKRRTFSSVDQMSIVEHT
p4      FKIENPVHPKWFCHCLTKMVAMLFSPNRSLEALSEMVNSLVTSTEKRFFSIESMEIYSYC
p3      YDIPNTTYPKWFRHCLAKLIGVLFSKEPTLEALDEMVSALEKSVEKKFISIESMDTFSYC
p2      YTIPNDSYPKWFRHCLAKLIATLFSKEPTLESLDEMVSALEKSVEKKFISIESMEIFSYC
p5      YTIPNVSYPN---------------KEPTLESLDEMVSALEKSVEKKFISIESMEIFSYC
               :*:                : ::*   ::..:: .* .: * *::.*.  .:

p1      DMSNVPHLGFSIPSISKCLGYPEGTDILLLSNTSTHYLDSKQKSVD---GLPDDLYLVVP
p4      NLSNVLSFSAGYPQLKDELGLKPFVEYRLIYENQLAVFTSKCLEANYSHQSPPILFPLLT
p3      DLKEVPQFSRNYPTLEEELGLKEGVEYRLIYENELAMFTSKLENQNRHRQYPPILYPVLA
p2      DLKDVPQFSRGYPTLKDELSLKDGVKYQLIYEDDLVVFADATESGKPH-QNSPILFPLLT
p5      DLKDVPQFSRGYPTLKDELSLKDGVKYHLIHEDDLAAFTDATENANDH-QSPPILFPLLA
        ::.:*  :.   * :.. *.    ..  *: : .   : .   . .       *: ::

p1      QTSHVDMRKILARNIEFHEFDDMTDRKLSEIRIKKCYEG-------------L-------
p4      DIPWS--PKQWGLSIED--VGGT------------------TNQGGNPPNNRMSILAQAG
p3      DIPWT--PKIWWRTIQD--KQGM-----------------SKGTGHSSRSARFDMHSQVG
p2      DIPWT--PKIWWRTIQD--KKG-----------------PSKNGEESSRTARIDMHSQVG
p5      DIPWT--PKIWWRTVQD--KKGISIFLFVHRKAMKVYPGTSKNAEESSRTARTDLHSQVG
        :       *    .::

p1      SMLTEIDEYLALFDRVSTILSTQFSLLVQELS-QFERVQTASRFAVYVDMASVPLMLFDE
p4      EALRDCEKIMEVVEVSRNILRNQLERLKIELESVREKTVMPMRFTIHAKMHSFALMPFVE
p3      EALNDCDKILEIVDVSKRILCMQLEMIKNELESAREKTLMPIRFTIYAKMQSFALLPFTS
p2      DALNDCDKILDVVEMSKKILRVQLERIKSNLESVHEKTLMPIRFTLYAKMQSFALLPFLS
p5      DALNDCDKILEVVEMSKKILRMQLERIKSSLETVHEKTLMPIRFTLYAKMQSFALLPFLN
        . * : :: : :.:    **  *:. :  .*.   *:.    **:::..* *. *: * .

p1      ANPETKMISDQCIQQAKRAREELERHAKVSMDIEACAKQLSKDAEDLRLEDMDLPGICEE
p4      NDTDKAVMDRICELA-DSAAKALETCTNFINQSLTITAEHLSELNAVELQTIEEPGLEDD
p3      ADSDRTVMNRVCEMA-GRATKQLEKCITFLNQSLKSASECVSELEAIELEANEIPGVEEE
p2      EEADKTVMDRVCEMA-GRATKELEKCVAFLNQSLDSATECLSELDKIQLEPTEIPGVEDE
p5      EDADRTVMDRVCEMA-GRATKELEKCIAFLNQSLDSAAECLSELDGIQLEPTEIPGVEDE
         : :  ::.  *      * : **    .  :    : :  .: : :.*:  : **: ::

p1      IESYVFYDKQAILSTQKYSQELVELCLKRRNNIMEQIFNSPDRINKSKLNKAMNLAASLS
p4      LNAMLSDD-PHSSAACEYEIRLANKCVARRAELAAQLCNKSQKSIVR---VLLRVARFTV
p3      LNGLLNDD-PST-GVFPYERDLANQCIARRADLAAQLCNNTKSSVVR---VLLRTAR---
p2      LNDLLNDD-PST-GVFHYEEQLAEQCIARRSDLAAQLCNNTKSSVVR---VLLRTARFVL
p5      LNDLLNED-PST-GVLKYEEELAEQCIARRSDLAAQLCNNTKSSVVR---VLLRTARFVL
        ::  :  *     ..  *.  *.: *: ** ::  *: *. .          :. *

p1      QLRSNYRKLQDMISECVDLLEKPFQEMKDTVNRYLQAQGCSRNTMQKSMHLLRPEFHESQ
p4      DMRVELARYGNIIDDCISDIERPFQELKNCMERLQASSNLDENSFKKALQYYKRASN---
p3      ------------------------------------------------------------
p2      DMSSEISKYRNYIDDYVSLIERPFQEMKNCMERMQAEHKLDERSFQKALLYYKRPAN---
p5      DMSSEISKYRNYIDDYVSLIERPFQEMKNCMERMQAENKLDERSFQKALLYYKRPAN---


p1      IRIKKTTKSCRKLIDQLNIELDQLGFVRLGDILIKAESEQTLTRSEEIQETQ--------
p4      --ITNHTRLIYERAQLLLQM------VQQKDQSRNSPSVNM--------------------
p3      ------------------------------------------------------------
```

```
p2      --IIAQTRLIREKVSQVFDL------VQQVNNSK--------------------------
p5      --IIAQTRLIREKVSQVTAQ------FPLSNVSIFNDEKDYARRMEHVRKLLPVLSNNLR

p1      ----------------------------------VV----SAESSPNKEQF-------PKP
p4      ------------------------------------------------------------
p3      ------------------------------------------------------------
p2      ------------------------------------------------------------
p5      KGSCGKIAVIGGSIEYTGAPFFAAISALRLGADLVHVMCAPEAAPVIKGFSPELIVHPGL

p1      EQDSILESSIDEGSTSFESTPPSSPPDVGSNNY----FQAVFQYFAKPTSSPSSS-AK--
p4      ------------------------------------------------------------
p3      ------------------------------------------------------------
p2      ------------------------------------------------------------
p5      EPETVIPKLERM-------DAIVLGPGLGRNPRLAPLFGNVLEFVRKTD-VPFVMDADGL

p1      ------------------------------------------------------------
p4      ------------------------------------------------------------
p3      ------------------------------------------------------------
p2      ------------------------------------------------------------
p5      WFLCEAIRQGVPPLPSAILTPNIVEFSRLCESALGIPDVLTIKDQDKLEDLASRLSTHLG

p1      ------------------------------------------------------------
p4      ------------------------------------------------------------
p3      ------------------------------------------------------------
p2      ------------------------------------------------------------
p5      TSLFVKGRVDIITNPDGKVTLGDDEGCPRRCGGQGDVSSGTLAMFLLWATRMSSSHDAKN

p1      ----------------------------------------------------
p4      ----------------------------------------------------
p3      ----------------------------------------------------
p2      ----------------------------------------------------
p5      AAGLACSQLVRRCAKLAYARVGRSMITSDLIDEIPGVLRELDSKKLEDCPRA
```

- Patterns observed: The beginning of the amino acid sequence is similar among three of the five sequences (p4, p2, and p5).They have the same number of gaps and start at methionine followed by isoleucine, followed by a polar amino acid.
- Interestingly, all five sequences have pretty conserved sequences (either identical or closely related in terms of amino acid chemical properties), except for p1 towards the beginning of the amino acid sequence.
- The third row, looks to be the most highly conserved among all five sequences, particularly that stretch of IAHRDIKHMN, indicating that this section of the protein sequence may serve a vital function in the protein (i.e. could be a binding site or site of enzymatic activity)
- There are a few stretches of just “red” in the colored version of this CLUSTAL MSA in which you see an abundance of nonpolar and aromatic amino acids next to each other, indicating that this might be an area that is probably embedded in the interior of the protein (because it is hydrophobic, may not want to interact with hydrophilic aqueous environment of the cell) or a lipid binding region. This is specific to the seventh row and

ninth row.

- There are long stretches of gaps in p5 and p1, which is why a lot of their amino acid sequence at the beginning and the end do not match up with p2, p3, and p4.

Neighbor-joining tree

- The topology of the neighbor joining tree is different from the topology of the UPGMA tree derived from MATLAB earlier (even though these are both distance methods). It makes a lot of sense that the two trees are not the same because the entire point of this project was to show that the UPGMA tree served as a guide tree for the MSA to construct the neighbor joining tree. If I ran through more iterations, I would eventually converge to a tree. These algorithms start with the known sequences and attempt to reconstruct the history of changes that had to take place from a common ancestor.
- Each branch on a tree of this kind has a length equal to the number of substitutions (or mutations) required to get from one node to the next.
- In this case, you can see that p1 seems to be more closely related to p4, and then p3 is closely related to (p1, p4), and p5 with (p3,(p1,p4)) and finally p2 being the least closely related.
- Some differences to note: UPGMA gives you a rooted tree while NJ gives an unrooted tree. Moreover, NJ does not assume that all lineages evolved at the same rate (hence, it follows the molecular clock hypothesis.
- **NOTE: the major difference between these two trees is that the branch lengths are different. Overall, the NJ method shows shorter branch lengths than UPGMA. The reason for this is probably because the clusters that are merged in NJ are not only close to each other (as in UPGMA) but also far apart from the rest.**
- NJ is generally considered more accurate for identifying trees representing evolution than the UPGMA algorithm. The NJ and UPGMA algorithms are similar in that they both specify tree topology by iteratively determining pairs of neighboring nodes that share the same parent node, using distance-based techniques. However, the UPGMA assignment of the closest pair of leaves to become neighbors does not necessarily apply in the problem that we consider here. For example, UPGMA may fail to identify a pair of neighboring leaves, one of which contains an exceptionally long edge.

```
(
(
(
p1:0.51197,
p4:0.19251)
:0.07435,
p3:0.09362)
:0.06924,
p2:0.03548,
p5:0.03765);
```

Phylogram

Branch length: Cladogram Real