Joshua Allen
CS466
Final mini project

Introduction

I have written and implemented multiple sequence matching program that takes several protein sequences and attempts to match with dynamic programming. The first step in implementing this was to cluster the sequences together by how closely related they were using an algorithm called neighbor-joining made popular by the Clustal program. I found this to be an instructive and interesting challenge, even if the results weren't as impressive as I'd hoped. It performed very well on my small test samples, but had multiple problems when trying to align long, real-world sequences.

Programming Approach

My approach to writing this program was to investigate the clustal approach with neighbor joining. I used pandas dataframes because of the convenience of indexing with strings, instead of trying to keep track of everything with index numbers. Pandas has a lot of powerful features, even though it is sometimes a pain to implement. I started by implementing a UPGMA algorithm based on this youtube video (https://www.youtube.com/watch?v=c2y9s_E2184&t=340s), since neighbor-joining is a modification of this apporach. Once that part was working, I modified the equations to introduce the more complex neighbor-joining approach, based on this video https://www.youtube.com/watch?v=AGSuDxQ7gP8. The main problem I ran into was that my alignment approach, based on our inclass discussions of dynamic programming, assigned higher scores to better matches, whereas the approaches outlined in the videos and elsewhere online (mainly: http://www.srmuniv.ac.in/sites/default/files/files/1(7).pdf) used a lower score = better match algorithm. Therefore, they sought to minimize the equations and I needed to maximize them to get the correct results.

The main idea of the clustering algorithm is to compare all the sequences in turn to each other, creating a symmetrical scoring matrix. Once that matrix is in place, you treat every string as a cluster and then base matches on the best scores, then recalulate the new table given the clusters. The UPGMA approach to coming up with the new matches is simply to average the scores in the new clusters to the other clusters and take the maximum. In the neighbor-joining algorithm, you first compute the new scoring table by averaging the scores between the new cluster and the others, same as for UPGMA, but instead of simple finding the minimum distance (or maximum match, as I did), you find the best score for Q(i,j), where

$$Q(i,j) = (r-2)(d\left(C_i, C_j\right) - u(C_i) - u(C_j)$$

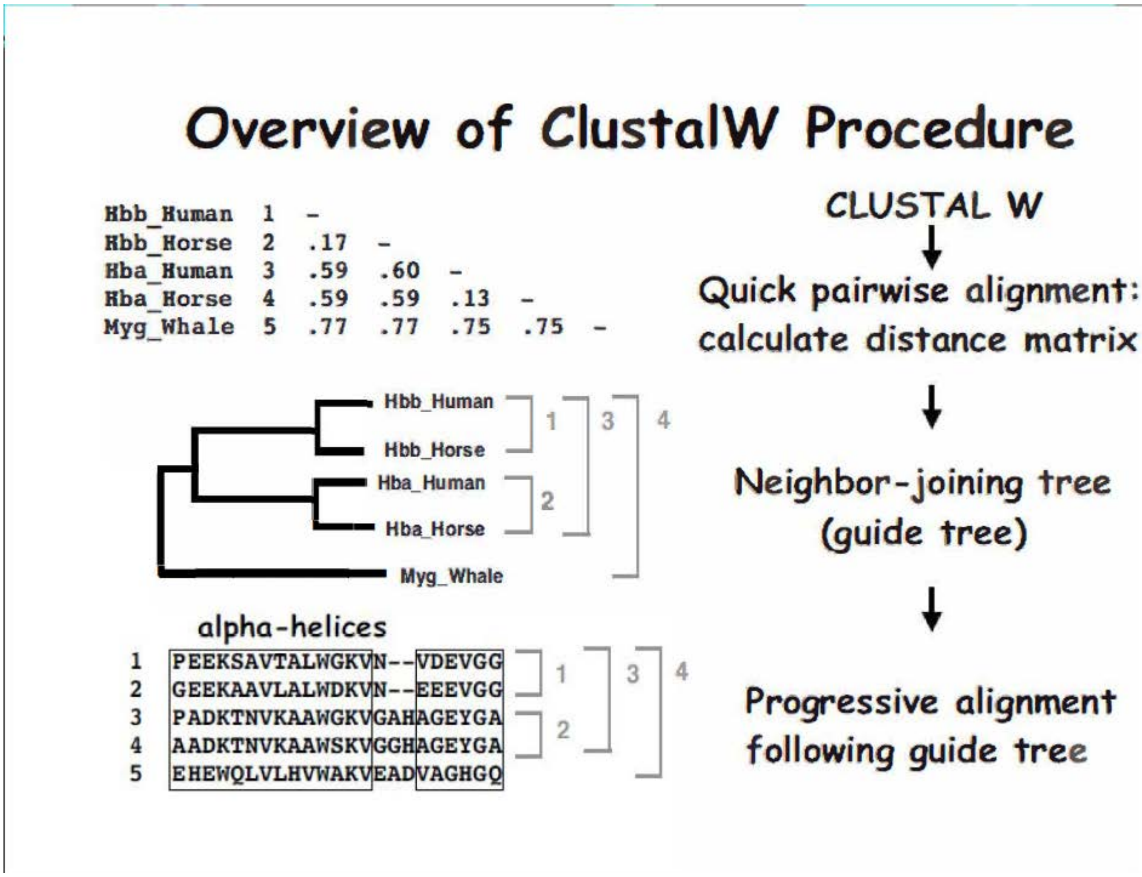
where r is the total number of current clusters. Programmatically, this translates to is setting up a new matrix with the new clusters indexing both rows and columns. It's understood that Q(i,i) = 0 for all i's, so those locations in the matrix are set to zero. The rest are done by computing the distance between the clusters (d(C_i, C_j)), which simply means looking them up in the averaged scoring table, and subtracting the u values, which is the vector formed by summing across the rows of the scoring table. Because the algorithm always groups them two at a time, this approach only relies on the previous version of the scoring table, without pulling in any outside information. The main drawback of this approach, as you'll see below, is that if it is off when you initially score all the matches, those errors will propogate through to the end. Everything depends on the accuracy of your scoring algorithm. I used dynamic programming, which is supposed to be the slow and steady way to find a definite score, but as you'll see, it doesn't always work the way on real data that it does on test data.

For testing this algorithm I used several sequences I found from the above-mentioned PDF which were already scored and clustered. I also tested it using short peptide chains randomly modified with a slightly modified version of the program we wrote for our last homework. Since my test peptide chains started as ten characters and could have at most 2 changes (indels and mutations combined), I'd expect close alignments and indeed this seemed to work just fine. I also tested the alignmenst against previous class assignments (though with a different scoring matrix than BLOSUM62, since those examples were nucleotides), so I knew those were working as expected, and found they performed perfectly. So I was confident that the algorithm essentially works, but was surprised by what I found when I ran it against real protein chains.

Data Processing

I examined the monarch butterfly (*Danaes plexippus*) protein related to circadian clocks as found on MonarchBase. Scientists who study monarch butterflies are very interested in their migration patterns. They seek to answer questions like how do the butterflies know when to migrate, and how do they know where to go, even across multiple generations, as a single migration might span three or four generations. A key component to this migration pattern is the monarchs' circadian rythms. I focused specifically on gene DPOG206046 and the protein it expresses, which acts as the 'clock' for the circadian rhythm. This is the core protein others will modify the expression of to create the circadian rhythm, and ultimately the migration behavior. The MonarchBase lists several analogous proteins in other species, including HMEL008784 in the genus *Heliconius*, a genus of butterfly in the same family as *D. plexippus*; BGIBMGA000498-TA in the genus *bombyx*, a genus of silkmoths in the same order (*Lepidoptera*) as *D. plexippus*; Clk-PD in the genus *Drosophila* (the fruit fly), in the same class as *D. pleixppus* (*insecta*); UniRef50_GOYQM3 a protein present in certain pest moths, which are also lepidopterans that are more distantly related, but better studied due to their economic importance. All of these proteins are relatively close to the query protein (55.83%, 50.43%, 55.53%, and 78.75%, respectively, according to MonarchBase). My hope was that my program, though it would likely be slower, would produce some usable results from comparing these five proteins.

When I did my test runs with small proteins, the results of the program were pretty solid. The program runs quickly and produces satisfactory results. For example, on this test data from http://www.srmuniv.ac.in/sites/default/files/files/1(7).pdf,



Overview of ClustalW Procedure

```
Hbb_Human  1  -
Hbb_Horse  2  .17  -
Hba_Human  3  .59  .60  -
Hba_Horse  4  .59  .59  .13  -
Myg_Whale  5  .77  .77  .75  .75  -
```

CLUSTAL W

Quick pairwise alignment: calculate distance matrix

Neighbor-joining tree (guide tree)

Progressive alignment following guide tree

```
1  PEEKSAVTALWGKVN--VDEVGG
2  GEEKAAVLALWDKVN--EEEVGG
3  PADKTNVKAAWGKVGAHAGEYGA
4  AADKTNVKAAWSKVGGHAGEYGA
5  EHEWQLVLHVWAKVEADVAGHGQ
```

Matching these five small subsections produced an adequate output of

```
seq1  PEEKSAVTALWGKV--NVDEVGG
seq2  GEEKAAVLALWDKV--NEEEVGG
seq3  PADKTNVKAAWGKVGAHAGEYGA
seq4  AADKTNVKAANSKVGGHAGEYGA
seq5  EHEWQLVLHVWAKVEADVAGHGQ
```

Though the gaps were a little off of the example, it preserved the major regions highlighted, and produced a clustering that exactly matched the phylogenetic tree in the example.

The longer matches, however did less well. While the speeds were actually comparable, matching for the entire proteins listed above produced a result with far too many gaps between peptides. These problems led to poor scoring even in the first iteration. My grouping algorithm produced a final cluster of:

(Clk-PD, (BGIBMGA000498-TA, (G0YQM3_SPOEX, (DPOGS206046-PA, HMEL008784-PA))))

Clustal Omega gives a phylogenetic tree of (((BGIBMGA000498-TA, Clk-PD), (HMEL008784-PA, G0YQM3_SPOEX)), DPOGS206046-PA). This probably reflects this specific protein more than the evolution of the total organism. This volatility of individual proteins is why phylogenetic trees are done with mitochondrial DNA, which is more conserved between species.

Joshua Allen
CS466
Final mini project

An interesting note is when you examine the strict phylogeny of the species as a whole, my grouping makes some sense. It groups the two from the same family, then two from the same order. I'm not sure which of the two moths would be closer related, but the fruit fly would definitely be the least related to the others. They are all in the same class, but totally different orders. It's possible that while my program did poorly on finding related regions in the protein strand, it was able to basically distinguish a rough distance between the species. This could also just be coincidental noise.

See Appendix 1 for the Clustal Omega alignment and Appendix 2 my output alignment. I decided to run my algorithm on a well-aligned output subsection of the Clustal Omega output to see how the matches compare. I compared sections 6 and 7 of the output in Appendix 1

My output:

```
BGIBMGA000498-TA  MS----I--ELQ-K-PL-T--C--D---LN-E-S-TK-A-SC--N
HMEL008784-PA     YDYYH-FD-DLE-K-VV-T--CH-E-A-LM-Q-K-GELT-SCYYS
Clk-PD            YDYY-HF-DDL-D-S--IV-AC-H-E-EL-R-Q-TGE-GKSCYY-
DPOGS206046-PA    YDYY-HF-DDL-E-K--VVS-C-H-E-AL-M-Q-KGE-LTSCYY-
G0YQM3_SPOEX      YDYY-HF-DDL-E-K--VVT-C-H-E-AL-M-Q-KGE-LTSCYY-


BGIBMGA000498-TA  --ADIA-KNAKQESA-EPD-NVSEA--D--AS-HGT--I--KDA-PS-EE
HMEL008784-PA     -YADID-KSTKQESG-ETD-AVSES--E--QT-RNV--L--KES--S-SE
Clk-PD            RF--L-TKG-QQWI-WLQTD-Y--YVS-YHQ-F-N-SKPDYV-VCTHKVV
DPOGS206046-PA    RF--L-TKG-QQWI-WLQTR-F--YIT-YHQ-W-N-SKPEFV-VCTHRVV
G0YQM3_SPOEX      RF--L-TKG-QQWI-WLQTR-F--YIT-YHQ-W-N-SKPEFV-VCTRRVV


BGIBMGA000498-TA  N---N-------L-----M--V--------M------------S------
HMEL008784-PA     ----D-------A-----M--P--------P------------S------
Clk-PD            SYAEV-LKDSRKEG-Q--K-S-GNSNSITNNGSS-KVIASTGTSSKSASA
DPOGS206046-PA    SYADII-KSTKQERTETEES-VR---DCDHNGSSL-KD----PSTEDAMV
G0YQM3_SPOEX      SYADIA-KSMKQEGAEA-DT-VS---EAEVNRGVM-KE-A--PS-DDAMV


BGIBMGA000498-TA  --------P-S
HMEL008784-PA     ----V---K-S
Clk-PD            T-TTLRDFELS
DPOGS206046-PA    PV----S-P-S
G0YQM3_SPOEX      SM----S-P-S
```

Versus the Clustal Omega output

```
BGIBMGA000498-TA      ----MSIELQKPLTCD-------------------------------------------L
HMEL008784-PA         YDYYHFDDLEKVVTCHEALMQKGELTSCYYSYA---------------------------
Clk-PD                YDYYHFDDLDSIVACHEELRQTGEGKSCYYRFLTKGQQWIWLQTDYYVSYHQFNSKPDYV
DPOGS206046-PA        YDYYHFDDLEKVVSCHEALMQKGELTSCYYRFLTKGQQWIWLQTRFYITYHQWNSKPEFV
G0YQM3_SPOEX          YDYYHFDDLEKVVTCHEALMQKGELTSCYYRFLTKGQQWIWLQTRFYITYHQWNSKPEFV


BGIBMGA000498-TA      NESTKASCNADIAKNAKQESAEPDN-VSEA-------DASHGTIKDAPSEENNLMVMSPS
HMEL008784-PA         ----------DIDKSTKQESGETDA-VSESEQTRNVL--------KESSSEDAMPPSVKS
Clk-PD                VCTHKVVSYAEVLKDSRKEGQKSGNSNSITNNGSSKVIASTGTSSKSASATTTLRDFELS
DPOGS206046-PA        VCTHRVVSYADIIKSTKQERTETEESVRDCDHNGSSL--------KDPSTEDAMVPVSPS
```

```
G0YQM3_SPOEX         VCTRRVVSYADIAKSMKQEGAEAD-TVSEAEVNRGVM--------KEAPSDDAMVSMSPS
```

Again, it suffers from too many gaps. The clusters for this smaller scale match are

((Clk-PD, (DPOGS206046-PA, G0YQM3_SPOEX)), (BGIBMGA000498-TA, HMEL008784-PA))

These results don't improve on the full sequence results significantly at all, and indicate that further work would need to be done to round out this program.

Conclusions

All in all, my program would have a long way to go to compete with the Clustal Omega and other multi-align packages available. Options would be to add more heuristic approaches to sequence matching and introducing an affine gap penalty which would favor grouping the sequences together more. I'd also research more on how proteins specifically are scored to improve the algorithm as it pertains to proteins in particular, since I noticed many of the scoring metrics gave a sense of distance rather than a sense of how good the matches were, in which case you'd seek to minimize distance rather than maximize matchis.

However, I was happy with the clustering part of the program overall, at least as it performed against test strands of few peptides. It took awhile to implement, but for short sequences, where my alignment worked best, the clustering algorithm did too. I feel like this is a promising start, if I wanted to build this up into a larger scale program, because my clustering algorithm was dependent on the scoring algorithm, it's success varied with the alignment as well. So the main focus of further improvement of the program. While it's true many tools already exist that can handle these functions, it was very instructive to try to implement my own version, so I could see what problems the more state-of-the-art versions attempt to solve and what the limitations are of my approach, it gave me the ability to focus what I did compared to what they did. Clustal Omega, for example, is something that has been refined and improved over the years by multiple contributers, so it would have been difficult to compete with that, but by performing my own implementation, I can now examine their code and look at how they improved on the basic ideas.

Appendix 1

```
BGIBMGA000498-TA      ------------------------------------------------------------    0
HMEL008784-PA         MDEDGDDKDDSK----RRTRNLSEKKRRDQFNMLLNELGSMVSSNNRKMDKSTVLKSTIS    56
Clk-PD                MDDESDDKDDTKSFLCRKSRNLSEKKRRDQFNSLVNDLSALISTSSRKMDKSTVLKSTIA    60
DPOGS206046-PA        MDDDGDDKDDTK----RRTRNLSEKKRRDQFNMLVNELGSMVSTNNRKMDKSTVLKSTIS    56
G0YQM3_SPOEX          MDDDGDDKDDTK----RRTRNLSEKKRRDQFNMLVNELSAMVSTNNRKMDKSTVLKSTIS    56


BGIBMGA000498-TA      ------------------------------------------------------------    0
HMEL008784-PA         FLKNHNEITVRSRAHDVQEDWKPTFLSNEEFTYLVLEALEGFVMVFSATG----------    106
Clk-PD                FLKNHNEATDRSKVFEIQQDWKPAFLSNDEYTHLMLESLDGFMMVFSSMGSIFYASESIT    120
DPOGS206046-PA        FLKNHNEITVRSRAHDVQEDWKPAFLSNEEFTYLVLEALEGFVMVFSASGCIYYVSESVT    116
G0YQM3_SPOEX          FLKNHNEITVRSRAHDVQEDWKPAFLSNEEFTYLVLEALEGFVMVFSATGQIYYVSESIT    116


BGIBMGA000498-TA      ------------------------------------------------------------    0
HMEL008784-PA         -----------------------------------------------QNEVQFQCHLQRG    119
Clk-PD                SQLGYLPQDLYNMTIYDLAYEMDHEALLNIFMNPTPVIEPRQTDISSSNQITFYTHLRRG    180
DPOGS206046-PA        SLLGHTPGDIINKSIFDLAFVDDRPNLYNILQNGGT-LDPTQV-VTTDNPISFRCRLQRG    174
G0YQM3_SPOEX          SLLGHNPADVVNKSIFDLACEDDRPSLYNLLQNPGSATDPMHA-LGKENEIRFQCHLKRG    175


BGIBMGA000498-TA      ------------------------------------------------------------    0
HMEL008784-PA         TLDFRDDITYELVQFNGHFRTNMEQNENNDLSY---------------------------    152
Clk-PD                GMEKVDANAYELVKFVGYFRNDTNTSTGSSSEVSNGSNGQPAVLPRIFQQNPNAEVDKKL    240
DPOGS206046-PA        TLDFRDEVTYELVQFDGHFRKNLESNE-NGH------------------HSYQDEHESRL    215
G0YQM3_SPOEX          SLDFRDETTYELIQFNGHFRTNMEPLDSDDM------------------QSYDPDHESRL    217


BGIBMGA000498-TA      ------------------------------------------------------------    0
HMEL008784-PA         -------------------------QSDQDSSLEWKFLFLDHRAPPIIGYLPFEVLGTSG    187
Clk-PD                VFVGTGRVQNPQLIREMSIIDPTSNEFTSKHSMEWKFLFLDHRAPPIIGYMPFEVLGTSG    300
DPOGS206046-PA        LFVCTGRLYMPQLVRDVSLVDTIRSEFTSRHSLEWKFLFLDHRAPPIIGYLPFEVLGTSG    275
G0YQM3_SPOEX          LFVCTGRLYTPQLIRDVSLVDSSRSEFTSRHSLEWKFLFLDHRAPPIIGYLPFEVLGTSG    277


BGIBMGA000498-TA      ----MSIELQKPLTCD-------------------------------------------L    13
HMEL008784-PA         YDYYHFDDLEKVVTCHEALMQKGELTSCYYSYA---------------------------    220
Clk-PD                YDYYHFDDLDSIVACHEELRQTGEGKSCYYRFLTKGQQWIWLQTDYYVSYHQFNSKPDYV    360
DPOGS206046-PA        YDYYHFDDLEKVVSCHEALMQKGELTSCYYRFLTKGQQWIWLQTRFYITYHQWNSKPEFV    335
G0YQM3_SPOEX          YDYYHFDDLEKVVTCHEALMQKGELTSCYYRFLTKGQQWIWLQTRFYITYHQWNSKPEFV    337
                             :*:. ::*.

BGIBMGA000498-TA      NESTKASCNADIAKNAKQESAEPDN-VSEA-------DASHGTIKDAPSEENNLMVMSPS    65
HMEL008784-PA         ----------DIDKSTKQESGETDA-VSESEQTRNVL--------KESSSEDAMPPSVKS    261
Clk-PD                VCTHKVVSYAEVLKDSRKEGQKSGNSNSITNNGSSKVIASTGTSSKSASATTTLRDFELS    420
DPOGS206046-PA        VCTHRVVSYADIIKSTKQERTETEESVRDCDHNGSSL--------KDPSTEDAMVPVSPS    387
G0YQM3_SPOEX          VCTRRVVSYADIAKSMKQEGAEAD-TVSEAEVNRGVM--------KEAPSDDAMVSMSPS    388
                                :: *. ::*  :                       .       :     *

BGIBMGA000498-TA      YMSDASDAFTNSYHRTSQLA-------------------HGG------------------    88
HMEL008784-PA         ------------------------------------AATSAGST-------GATVASVGT    278
Clk-PD                SQNLDSTLLGNSLASLGTETAATSPAVDSSPMWSASAVQPSGSCQINPLKTSRPASSYGN    480
DPOGS206046-PA        YMSEASDAFATSYN----------S---MSKLASVKSAATSGST-------SATVATLGT    427
G0YQM3_SPOEX          YMSEASDAFGNSSYQ---------P---LSQVS-VS------------------------    411


BGIBMGA000498-TA      -----VANVNEHPCYGS-------------------------------------------    100
HMEL008784-PA         -STTAPASWPRTSHVRY--AGSDTASVS--GES------------------RSSQRNRLS    315
Clk-PD                ISSTGISPKAKRKCYFYNNRGNDSDSTSMSTDSVTSRQSMMTHVSSQSQRQRSHHREHHR    540
DPOGS206046-PA        AITTASATWPPRSSYLLYTTGSDTTSVS--GGS------------------RSSQRN---    464
G0YQM3_SPOEX          ------------------------------------------------------------    411


BGIBMGA000498-TA      ----------------------------MMFPQSYKAGSEPALVP---------------    117
HMEL008784-PA         P------------------------HTGVPQQLMTHRVPEPSLVP---------------    336
Clk-PD                ENHHNQSHHHMQQQQQHQNQQQQHQQHQQLQQQLQHTVGTPKMVPLLPIASTQIMAGNAC    600
```

```
DPOGS206046-PA        -------------------------SSQEL-----QRLPEPALVP---------------    479
G0YQM3_SPOEX          ------------------------------------------------------------    411

BGIBMGA000498-TA      -----------QHGIGAQYLEPDPYVSAVSLPGVL-PL---------------PLPPLPI    150
HMEL008784-PA         -----------QHGIGAQYLEPAPYVGAVGVPGVL-PL---------------SLPPLPV    369
Clk-PD                QFPQPAYPLASPQLVAPTFLEPPQYLTAIPMQPVIAPFPVAPVLSPLPVQSQTDMLPDTV    660
DPOGS206046-PA        -----------QHGIGAQYLEPAPYVGAVGVPAV-LPL---------------SLPPIPV    512
G0YQM3_SPOEX          ------------------------------------------------------------    411

BGIBMGA000498-TA      IVSPDQAQI----QEQLQRKHEELQQMILRQQEELRQVKEQLLLARLGILQPLINVPNPY    206
HMEL008784-PA         IVSPDQAQLQWPWQEQLQRTHRELQQMIVRQQEELRQVKEQLLLARLGILQPIINVQDPY    429
Clk-PD                VMTPTQSQL----QDQLQRKHDELQKLILQQQNELRIVSEQLLLSRYTYLQPMMSMGFAP    716
DPOGS206046-PA        IVAQDQAQL----QEQLQRTHRELQQMIVRQQEELRQVKEQLLFARLGILQPVINVQDPF    568
G0YQM3_SPOEX          ------------------------------------------------------------    411

BGIBMGA000498-TA      ANTEETQQMQRLPTQVSYDRPA-QRSISGYSQTQPPP---------HLSA----------    246
HMEL008784-PA         AHPEEIQQNQRLPAQILYEGGPRPIGYPPHTQAPPGNQNQHHHMPPPLALFHIAAVLCSL    489
Clk-PD                GNMTAA-AVGN-----LGASGQRGLNFTGSNAVQPQF-----------NQYGFALNSEQM    759
DPOGS206046-PA        TNPEQMPNRSS-----IMYDGNRQLSYPQTSHQ---------------------------    596
G0YQM3_SPOEX          ------------------------------------------------------------    411

BGIBMGA000498-TA      ---------LQPHSQLQHQHQHQHPNQHQHPRML--------------------------    271
HMEL008784-PA         WSKL----------NLWLES-FGIVKTALSRPDLI-------------------------    513
Clk-PD                LNQQDQQMMMQQQQNLHTQHQHNLQQQHQSHSQLQQHTQQQHQQQQQQQQQQQQQQQQQQ    819
DPOGS206046-PA        ---QN----------------HNMPPQ---------------------------------    604
G0YQM3_SPOEX          ------------------------------------------------------------    411

BGIBMGA000498-TA      ------------------------------------------------------------    271
HMEL008784-PA         ------------------------------------------------------------    513
Clk-PD                QQQQQQQQQQQQQQQLQLQQQNDILLREDIDDIDAFLNLSPLHSLGSQSTINPFNSSSNN    879
DPOGS206046-PA        ------------------------------------------------------------    604
G0YQM3_SPOEX          ------------------------------------------------------------    411

BGIBMGA000498-TA      ------------------------------------------------------------    271
HMEL008784-PA         ------------------------------------------------------------    513
Clk-PD                NNQSYNGGSNLNNGNQNNNNRSSNPPQNNNEDSLLSCMQMATESSPSINFHMGISDDGSE    939
DPOGS206046-PA        ------------------------------------------------------------    604
G0YQM3_SPOEX          ------------------------------------------------------------    411

BGIBMGA000498-TA      ------------------------------------------------------------    271
HMEL008784-PA         ------------------------------------------------------------    513
Clk-PD                TQSEDNKMMHTSGSNLVQQQQQQQQQQQILQQHQQQSNSFFSSNPFLNSQNQNQNQLPND    999
DPOGS206046-PA        ------------------------------------------------------------    604
G0YQM3_SPOEX          ------------------------------------------------------------    411

BGIBMGA000498-TA      ----------------------------       271
HMEL008784-PA         ----------------------------       513
Clk-PD                LEILPYQMSQEQSQNLFNSPHTAPGSSQ       1027
DPOGS206046-PA        ----------------------------       604
G0YQM3_SPOEX          ----------------------------       411
```

Joshua Allen
CS466
Final mini project

Appendix 2

```
DPOGS206046-PA         0    MDDD---GD-DK-DD--T-KRRT---R-N-----LS-E-K-K-R-RDQ-F-N-MLV--NEL-GS---MVST-N-NR-K-M-DK-S-T-V-L-KS-
HMEL008784-PA          0    MDED---GD-DK-DD--S-KRRT---R-N-----LS-E-K-K-R-RDQ-F-N-MLL--NEL-GS---MVSS-N-NR-K-M-DK-S-T-V-L-KS-
BGIBMGA000498-TA       0    M------SI-EL--Q----K--P-----------LT---------CD-------L---NE---S--TKASC---N----A-D------I-A-KN-
Clk-PD                 0    M---DDES-DD-K--DD-TK---SFL-C-RKSRNL-S-E-K-K-RRD-Q-F-NSL-V-ND-L--SA-LIS-T-SS-R-K-MD-K-S-T-V-LK-S
G0YQM3_SPOEX           0    MDDD---GD-DK-DD--T-KRRT---R-N-----LS-E-K-K-R-RDQ-F-N-ML--VNEL-SA---MVST-N-NR-K-M-DK-S-T-V-L-KS-


DPOGS206046-PA         T-I-S-F-L-KN-H-N-EIT-VR--SR-AHD-VQ-ED-WK-PA-F-L-SNE-E-FT-Y-L-VL-E-A-L-EG-F-VM--V-F-S-ASG--CI--YYV--S
HMEL008784-PA          T-I-S-F-L-KN-H-N-EIT-VR--SR-AHD-VQ-ED-WK-PT-F-L-SNE-E-FT-Y-L-VL-E-A-L-EG-F-VM--V-F-S-A--------------
BGIBMGA000498-TA       ----A-----K----Q-E-------S--A------E-----P------DN-----------V----S---E--------------A--------------
Clk-PD                 -T-I-A-F-LK-N-H-NEA-T--DRS-KVF-EI-QQ-DW-KP-A-F-LSN-D-EY-T-H-LM-L-E-S-LD-G-F--MM-V-F-SSM-G--S-IFY-A--
G0YQM3_SPOEX           T-I-S-F-L-KN-H-N-EIT-VR--SR-AHD-VQ-ED-WK-PA-F-L-SNE-E-FT-Y-L-VL-E-A-L-EG-F-VM--V-F-S-AT--G-QI-YY--V-


DPOGS206046-PA         --ES--V-TS---L--L-G--H-T--PG-D--I-IN--KS-I--F-D--L--A--FVD--D----R-PN--L---YN--I-L---Q-N-G-G-T-L---D
HMEL008784-PA          -----------------------T---G---------------------------------------------------------Q-N-----------E
BGIBMGA000498-TA       ----------------------------------------------------------------------------------------------------
Clk-PD                 S-E-S--IT-S-Q-L----G----YLP-Q-D--LY-N-M--TI--Y-D--L--A-Y---E-M-D--H--EA-L-L--N--I-F---M-N-P--T--PVIE
G0YQM3_SPOEX           -SE--SI-T--S---L-L--G-HN--PA---DV-V--NK-S--I-F--D--L--A-C-E---D-DR-PS----L-Y--N----LLQ-N--P-GSAT---D


DPOGS206046-PA         -PT-Q----V--V-T-TD----N--P-I--S-F-R-CR--L-QR--GT--L-D-FR-DE---V--T----YEL-V-QF-D-G--HF-R-KNL--ESN---
HMEL008784-PA          --V-Q---------------------------F-Q-CH--L-QR--GT--L-D-FR-DD---I--T----YEL-V-QF-N-G--HF-R-TNM--EQN---
BGIBMGA000498-TA       D--A---S-------------------------------H------G--TI----K-D-A-P---S-----E----E-------N-----N-L----MVM
Clk-PD                 -P---RQ-T-D--I----SSS--N--Q-I--T-F-YT-H--L--RRG-G--M-E--K------V--DANAYE-L-VK-F---VG-YF-R-N---------
G0YQM3_SPOEX           -PM-H------AL--G-K---E--NE---IR-F-Q-CH--L-KR--GS--L-D-FR-D--E---TT----YEL-I-QF--NG--HF-R-TN--MEPL---


DPOGS206046-PA         E-N-----G-H-H--S--YQD--E--H--E-S--R-L-L--F-V-CTG-RL--YMPQ--L-V-R-D---V---SL---V--DTI--R-------------
HMEL008784-PA          E-N------------N----D-----------------L-------S------Y--Q----------------S-------D----Q-------------
BGIBMGA000498-TA       --S----------------------P--------------------S------Y--M----------------S-------D------------------
Clk-PD                 -D-TNTST-G-S-S-S-EV--SN-G--S--N-G-QP--AVL-----P--R-I-F---QQ--NP-N-A-E-V-DK-K-L-V-F-V-GTGRVQNPQLIREMS
G0YQM3_SPOEX           D-S-------D-D-MQS-Y-D--P--D-HE---SR--LL---F-VCT-GR--LYTPQ---LI-R---D---V-S--L---VDS-S-R-------------


DPOGS206046-PA         -S--E-----F-TS--R-H--SL-E-WK-F-L-F-L-DH-R-APPI-I-G-YL-PFE-V-L-G-TS-G-YD-YYHF-D-D-L-E-K-VV-S-C-H-E-AL
HMEL008784-PA          ----D--------S-------SL-E-WK-F-L-F-L-DH-R-APPI-I-G-YL-PFE-V-L-G-TS-G-YD-YYHF-D-D-L-E-K-VV-T-C-H-E-AL
BGIBMGA000498-TA       -------------A-------S---------------D----A------------F--------TN----S--YH----------R-----T--S-Q--L
Clk-PD                 I-IDPTSNE----FTS-K-H-S-M-EW-K-F-L-F-LD-H-RAPP-I-I-GY-MPF-E-V-L-GT-S-GY-DYYH-F-D-D-L-D-S-IVA--C-H-EEL
G0YQM3_SPOEX           ----S----E-FTS--R---HSL-E-WK-F-L-F-L-DH-R-APPI-I-G-YL-PFE-V-L-G-TS-G-YD-YYHF-D-D-L-E-K-VV-T-C-H-E-AL

DPOGS206046-PA         MQ--KGE-LT-S-C-YYR-F--L--TKG--Q-QWI--WL--QTRF-YI-TY------H--QWN--S--KPE-F-V---V-C-TH--R-V---V-S-YA-D
HMEL008784-PA          MQ--KGE-LT-S-C--------------------------------Y---Y-------------------------------------------S-YA-D
BGIBMGA000498-TA       AH---GG-VA-N-V----N------------E------------------H------------------P----------C---------------Y---
Clk-PD                 --RQTG-E--GKS-CYYR--F--L-TK-G-Q-QW-I-W-L-Q-------T-DYYVSY-H-QF-N--S-KP-D-Y-V-V--C-T-H--K-V-V---SY-A-
G0YQM3_SPOEX           MQ--KGE-LT-S-C-YYR---F--LTK--GQ-QW--IW--LQTR-FY-ITY--------HQW--N--SKPE--F--V--VC-T--RR---V--VS-YA-D
```

Joshua Allen
CS466
Final mini project

```
DPOGS206046-PA         -II-K--S-TK-Q-E-RT--ETEE-S-VRDCD--HN-G-S---SL-KDPSTE-DA-M-VPVSPS-YMSE--ASDA-FA-TSYNSMSKLASVKS-AA-TSG
HMEL008784-PA          -ID-K--S-TK-Q-E-SG--ET-D-A-VSESE--QT-R-N---VL-KESSSE-DA-M--P--PS--V-K---S-A--A-T------------S--A---G
BGIBMGA000498-TA       -------------------G-----------------------------S-------M------------------------------------------
Clk-PD                 EV---LK-DS-R-K-E--G-------------QK-S--G-NSNS-I---T--NN-G--------S----SK----V-IA--------------S--T---
G0YQM3_SPOEX           -I-AK--S-MK-Q-E--G---A--E-A----D----T-------V----S-E--A-------------E-------------------------------

DPOGS206046-PA         -STS-ATV--A-T-LGTA-ITT---ASAT-W-PPR-SS---Y--LL-YTT-GS-D-TTS-V---SG-GS-R-SS-QRN---S----------S---QEL-
HMEL008784-PA          -STG-ATV--A-S-VGTS--TT---APAS-W--PR-TS---H--VR-Y-A-GS-D-TAS-V---SG-ES-R-SS-QRNRL-S--------PHTGVPQQL-
BGIBMGA000498-TA       -------M----------------------F--PQ--S-----Y--K---AG-----------------------------S------------------
Clk-PD                 G-T-S---SKSA-S----A---TTT--LRD-F---E--LSS-Q--------N-L-D---S-TLL--G-NS-L-AS-----L-GTETAATS-----P--AV
G0YQM3_SPOEX           -------V-------------N------------R----------------G--------V-------------------------------------M-

DPOGS206046-PA         --Q-R-LPEPA---L-V---PQ--H-G---I--G--A-Q-YLE-PA-PY-VG--A-VG---V-PA-VL-PL-S--------------------L-PPI-P
HMEL008784-PA          MTH-R-VPEPS---L-V---PQ--H-G---I--G--A-Q-YLE-PA-PY-VG--A-VG---V-PG-VL-PL-S--------------------L-PPL-P
BGIBMGA000498-TA       --------EPA---L-V---P--Q-H--G-I----GA--QYLE-PD-PY-V--SA-V--S-L-P-GVL-P-LP--------------------L-PP-LP
Clk-PD                 ---D-S--SPMWS-AS-AVQP-S----G-SCQI-N-PLK--TSRP-AS-SY-G--NI-S-S-T-G-I-SP---KAKRKCYFYNNRGNDSDSTS-MST--D
G0YQM3_SPOEX           ----K---E----A------P---------------S-D---D--A----M-----VS------------M-S--------------------------P

DPOGS206046-PA         V-I-V--A-Q--D-Q-A-Q-L-------Q-E-----Q-L-Q-R-T-H----------------------------------------------RE--L-Q
HMEL008784-PA          V-I-V--S-P--D-Q-A-Q-L---QWPWQ-E-----Q-L-Q-R-T-H----------------------------------------------RE--L-Q
BGIBMGA000498-TA       -II-V--S---P-D-QA--Q---I-----Q-----E-QL--Q-R-KH------------------------------------------------EEL--
Clk-PD                 ---S-VTS--R---Q--S---MM-------THVSS-Q--SQ-R----QRSHHREHHRENHHNQSHHHMQQQQQHQNQQQQHQQHQQLQQQLQH----TV-
G0YQM3_SPOEX           --------S-------------------------------------Y-----------------------------------------------------

DPOGS206046-PA         --Q--MIVR--Q---Q---E--E---LR-----Q-VK----EQ---L-L--F--AR-L-G--I-LQ---PV-I-N--V-----------Q-D--P--F--
HMEL008784-PA          --Q--MIVR--Q---Q---E--E---LR-----Q-VK----EQ---L-L--L--AR-L-G--I-LQ---PI-I-N--V-----------Q-D--P--Y--
BGIBMGA000498-TA       -Q-Q-MIL--R--Q---Q---E--E-L-R-----QV---K-E-Q--L-L---L-A-RL---GI-L--Q-P-LI--N-V------------PN----P--Y
Clk-PD                 G-TPKMVP-L--LPIASTQIM-AG-NAC-QFPQPAYPLA-SPQ-LVAPTFLE-PPQ-YL-T-AI-PM-QP--VIA-PFPVAPVLSPLPVQSQTD-M--L-
G0YQM3_SPOEX           -----M------------------------------S----E------------A-------------------S----------------D--A-----

DPOGS206046-PA         -T----N--------------P-E--Q-----M---P------N-R-----S--S-----I---M--YDGN--R--QL----SY-PQ-T---S-------
HMEL008784-PA          -A----H--------------P-E--E-----I--QQ------NQR--L-PA--Q-----I---L--YEGG-PR--PI----GYPPH-T--QA----PPG
BGIBMGA000498-TA       ----A---N-------------T--E----E---T----Q-Q-M-Q-RL-----P--TQ---V---SYD------R---------P---AQRS--IS---
Clk-PD                 -PDT-V-V-MTPTQSQLQDQLQRKH-DELQ-K-L--ILQ-Q-QN--E--L-RI-VSE--Q-L-LLSRYT--Y--L-QPMMSMGFAP-GN---MTA-A---
G0YQM3_SPOEX           F-----G------------------------------------N---------SS------------Y--------Q--------P--L-------S---

DPOGS206046-PA         ---H--Q-------QN-HNM-----P-P------------------------------------------------------------------------
HMEL008784-PA          ---N--Q----N--QH-HHM----PP-PL-A-L-FH----I--A-AV-L-C-S-L-WS---K---L-N-------------------L--W-----L---
BGIBMGA000498-TA       -G---Y--S-----Q-----T-Q-PP-P---------H----LS-A--L----Q------P--H---S----------------------Q-----L---
Clk-PD                 A-VGN--L-GA-SGQ-R-GL-N-F-TGS-N-A-V--Q-PQ-F--NQYG-F-A---L--NS--E-Q-M-LNQQDQQMMMQQQQNLHTQ-HQHNLQQQHQSH
G0YQM3_SPOEX           --------------Q----V--------------------------------------------------------------------------------
```

Joshua Allen
CS466
Final mini project

```
DPOGS206046-PA         --Q-------------------------------------------------------------------------------------------------
HMEL008784-PA          --E------------------------------------------------S---F-----GI--V-----K------T-----A---------------
BGIBMGA000498-TA       ----Q-H-Q--------------------------------------------H-----Q-H---------Q------H----PN---Q-----------
Clk-PD                 SQLQ-Q-HTQQQHQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQLQLQQ-Q-N-DIL-LR-EDIDDIDAFLNLSPLHSLGSQSTINPFNSSSNNNNQS
G0YQM3_SPOEX           --S-----------------------------------------------------------------V-------------------------------

DPOGS206046-PA         ----------------------------------------------------------------------------------------------------
HMEL008784-PA          L----------------------------------------------S-------------R--P---D----L-----I--------------------
BGIBMGA000498-TA       -------------------------------------------------------H-----QH-P--------R---M----------------------
Clk-PD                 YNGGSNLNNGNQNNNNRSSNPPQNNNEDSLLSCMQMATESSPSINFHMGISDDGS-E-TQSEDNKMMHTSGSNLVQQ-Q-QQQQQQQQILQQHQQQSNSF
G0YQM3_SPOEX           ---------------------------------------------------------S------------------------------------------

DPOGS206046-PA         ------------------------------------------------
HMEL008784-PA          ------------------------------------------------
BGIBMGA000498-TA       ------------------------------------------L-----
Clk-PD                 FSSNPFLNSQNQNQNQLPNDLEILPYQMSQEQSQNLFNSPHTAPGSSQ
G0YQM3_SPOEX           ------------------------------------------------
```