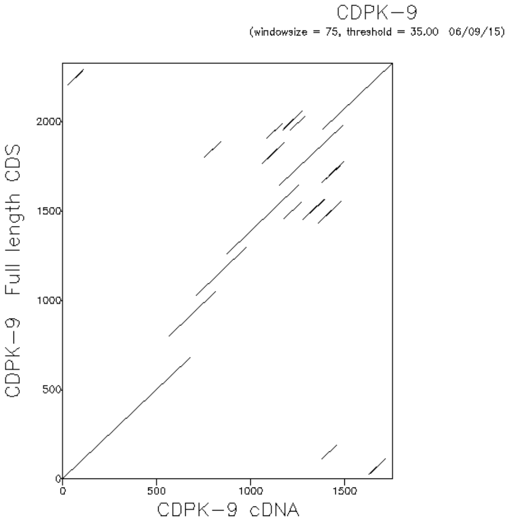
Lab 2 Biol 4150/6150 Fall 2015 Pairwise Alignments

Due 1:00 pm Wednesday, September 9. Paste your responses and results below and upload to T-square Assignments folder.

1. Create and show a dot plot of an *Arabidopsis* calmodulin-like domain protein kinase (CDPK) mRNA (cDNA) sequence versus the corresponding genomic sequence. Adjust parameters (window size, threshold value) until the background is reduced to a suitable level. Paste a screenshot of the final dotplot.



What does the resulting plot tell you?
There are 6 exons, with the last exon being the largest

2. Perform a global pairwise alignment between the yeast DAP1 protein (GI 6325087) and a cytochrome b5 protein from any bacterial species, using the **needle** program in EMBOSS or a Needleman-Wunch implementation on the web.
a) Use the BLOSUM62 matrix.

```
# Identity:      41/242 (16.9%)
# Similarity:    71/242 (29.3%)
# Gaps:         111/242 (45.9%)
# Score: 75.0

NP_015155.1        1 -------------------------MSFIKNLLFGGVKTSEDPTGLTGNG     25
                                              ...:...::...||.|:|.|
EHJ01201.1         1 MESNCINKQGSYSHYYINKMFIGRNSQEVSISIYFDFKTIEEPIG-----     45

NP_015155.1       26 ASNTNDSNKGSEPVVAGNFFPRTLSKFNGHDDEKIFIAIRGKVYDCTR--     73
                          |::...:.    .|....||:::|.:.:..::|:.|.|||.:.
EHJ01201.1        46 -----DNHLRQQK----QFILEELSQYDGSNGKSAYVAVDGIVYDLSNVE     86

NP_015155.1       74 ----GRQFYGPSGP--YTNFAGHDASR---------GLALNS--FDLDVI    106
                         |:.|...:|.  .:.|..|...:         |:.:.|  .::|.|
EHJ01201.1        87 AWAGGKHFGLTAGKDLTSEFNSHHGIKKVLNDKPKVGILIESKQKNMDSI    136

NP_015155.1      107 -------------KDWDQPIDPLDDLTKEQIDALDEWQ-----EHFENKY    138
                                  .||.:.|.||.|      :||:|..     ||...||
EHJ01201.1       137 ARLTLADTYDFSPDDWIEYIMPLVD------NALEEATGGVSLEHLFQKY    180

NP_015155.1      139 PCIGTLIPEPGVNV----------------------------    152
                     ..||.|:.: |:..
EHJ01201.1       181 IMIGILVGQ-GMTFKEATGEIEDWEKTGISKLLDQSKGKQVY    221
```

b) Use the PAM250 matrix. Paste the alignment below and report the score of the alignment.

```
# Length: 243
# Identity:      40/243 (16.5%)
# Similarity:    85/243 (35.0%)
# Gaps:         113/243 (46.5%)
# Score: 122.0
#
#
#=======================================

NP_015155.1        1 -------------MSFIKNLLFG------------GVKTSEDPTGLTGNG     25
                                  ..:|:::::|            :.||.|:|.|
EHJ01201.1         1 MESNCINKQGSYSHYYINKMFIGRNSQEVSISIYFDFKTIEEPIG-----     45

NP_015155.1       26 ASNTNDSNKGSEPVVAGNFFPRTLSKFNGHDDEKIFIAIRGKVYDCTR--     73
                              :::.....:|:...||:::|.::...::|:.|.|||.:.
EHJ01201.1        46 ---------DNHLRQQKQFILEELSQYDGSNGKSAYVAVDGIVYDLSNVE     86

NP_015155.1       74 ----GRQF--YGPSGPYTNFAGHDASRG----------------------     95
                         |::|  .:..:..::|.:|::.:.
EHJ01201.1        87 AWAGGKHFGLTAGKDLTSEFNSHHGIKKVLNDKPKVGILIESKQKNMDSI    136

NP_015155.1       96 --LAL-NSFDLDVIKDWDQPIDPLDDLTKEQIDALDE-----WQEHFENK    137
                       |:| :::|:.. .||.:.|.||.|      :||:|     ..||:.:|
EHJ01201.1       137 ARLTLADTYDFSP-DDWIEYIMPLVD------NALEEATGGVSLEHLFQK    179

NP_015155.1      138 YPCIGTLIPEPGVNV----------------------------    152
                     |..||.|:.: |:..
EHJ01201.1       180 YIMIGILVGQ-GMTFKEATGEIEDWEKTGISKLLDQSKGKQVY    221
```

c) Compare and contrast the two alignments - highlight any differences.

Both alignments show approximately equal identity scores and gap counts but the PAM250 scored alignment has a higher percent similar. The two alignments start at different residues but the terminally aligned residues are the same.

3. Perform a local alignment between the two sequences, using the **water** program in EMBOSS or an implementation on the web.

a) Use the BLOSUM62 matrix. Paste the alignment below, and report the score of the alignment.

```
# Length: 170
# Identity:      40/170 (23.5%)
# Similarity:    65/170 (38.2%)
# Gaps:          57/170 (33.5%)
# Score: 85.0
#
#
#=======================================

NP_015155.1       13 KTSEDPTGLTGNGASNTNDSNKGSEPVVAGNFFPRTLSKFNGHDDEKIFI     62
                     ||.|:|.|          |::...:.    .|....||:::|.:.:..::
EHJ01201.1        38 KTIEEPIG----------DNHLRQQK----QFILEELSQYDGSNGKSAYV     73

NP_015155.1       63 AIRGKVYDCTR------GRQFYGPSGP--YTNFAGHDASR---------G     95
                     |:.|.|||.:.      |:.|...:|.  .:.|..|...:         |
EHJ01201.1        74 AVDGIVYDLSNVEAWAGGKHFGLTAGKDLTSEFNSHHGIKKVLNDKPKVG    123

NP_015155.1       96 LALNS--FDLDVI-------------KDWDQPIDPLDDLTKEQIDALDEW    130
                     :.:.|  .::|.|             .||.:.|.||.|      :||:|.
EHJ01201.1       124 ILIESKQKNMDSIARLTLADTYDFSPDDWIEYIMPLVD------NALEEA    167

NP_015155.1      131 Q-----EHFENKYPCIGTLI    145
                     .     ||...||..||.|:
EHJ01201.1       168 TGGVSLEHLFQKYIMIGILV    187
```

b) Use the PAM250 matrix. Paste the alignment below, and highlight any differences between this alignment and the alignment generated with BLOSUM62.

```
# Length: 195
# Identity:      39/195 (20.0%)
# Similarity:    83/195 (42.6%)
# Gaps:          71/195 (36.4%)
# Score: 131.0
#
#
#=======================================

NP_015155.1        3 FIKNLLFG------------GVKTSEDPTGLTGNGASNTNDSNKGSEPVV     40
                     :|:::::|            :.||.|:|.|              :::...
EHJ01201.1        16 YINKMFIGRNSQEVSISIYFDFKTIEEPIG--------------DNHLRQ     51

NP_015155.1       41 AGNFFPRTLSKFNGHDDEKIFIAIRGKVYDCTR------GRQF--YGPSG     82
                     ..:|:...||:::|.::...::|:.|.|||.:.      |::|  .:..:
EHJ01201.1        52 QKQFILEELSQYDGSNGKSAYVAVDGIVYDLSNVEAWAGGKHFGLTAGKD    101

NP_015155.1       83 PYTNFAGHDASRG------------------------LAL-NSFDLDVIK    107
                     ..::|.:|::.:.                        |:| :::|:.. .
EHJ01201.1       102 LTSEFNSHHGIKKVLNDKPKVGILIESKQKNMDSIARLTLADTYDFSP-D    150

NP_015155.1      108 DWDQPIDPLDDLTKEQIDALDE-----WQEHFENKYPCIGTLIPE    147
                     ||.:.|.||.|      :||:|     ..||:.:||..||.|:.:
EHJ01201.1       151 DWIEYIMPLVD------NALEEATGGVSLEHLFQKYIMIGILVGQ    189
```

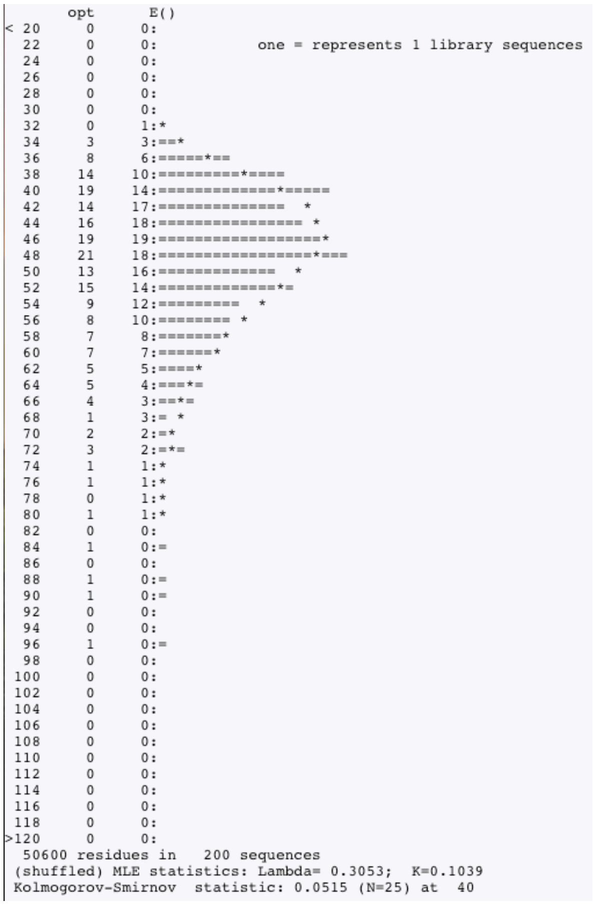
c) Compare and contrast the global and local alignments. Which algorithm gave the better alignment? Write a brief explanation of why the two algorithms gave different alignments in this instance.

Global alignments attempts to align every residue in a sequence against another whereas local attempts to align small, potentially shared/conserved regions between two sequences. Global alignments have better diagnostic power with closely related proteins than local alignments. Local alignments however are useful for sequences of vastly dissimilar length but similar function, that is we can align conserved domains. In this instance the local alignment performed better because it is more tolerant of gaps and is designed to align parts of evolutionarily distant proteins with conserved, similar function.

4. Test the effects of using a different BLOSUM scoring matrices in #3 above - which scoring matrix gives the best alignment? Explain and justify your criterion for the quality of the alignment.

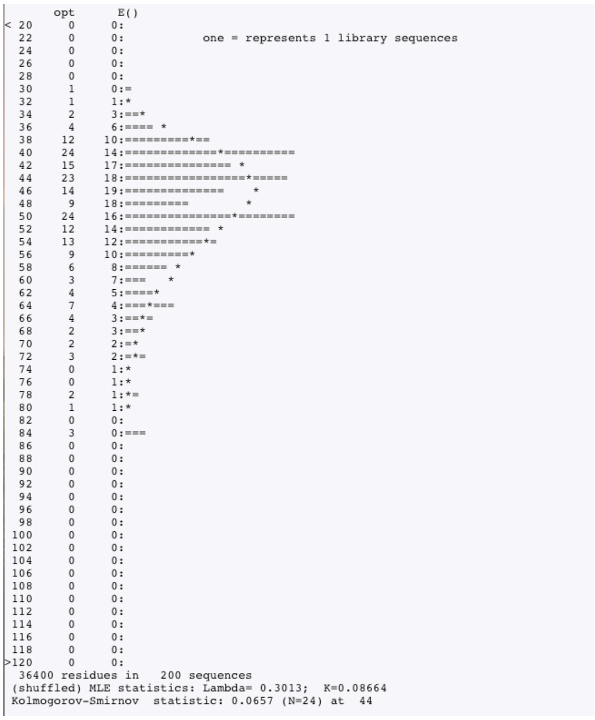
BLOSUM40 matrix gave the best alignment, of length 196, ID 44/196, Gaps 72/196 and a score of 209. As the matrix number decreases the so does the percent similarity between proteins. Because these are two distant protein they likely share some conserved regions but domains with low evolutionary pressure have likely diverged significantly. Alignment with a BLOSUM40 matrix gives the longest alignment, highest scoring alignment with an average numbers of caps.

5. Use the PRSS Shuffle program (http://www.ch.embnet.org/software/PRSS_form.html) to visualize the distribution of alignment scores if the second sequence is randomly shuffled. Be sure to use the same BLOSUM matrix and gap penalty as in one of your local alignments in 3 or 4 above.
a) Paste a screenshot of the resulting distribution. Based on this distribution, is the score you obtained in 3 or 4 significant? Explain.
Yes, it is significant. The score, 85, falls into the long right tail of the distribution



```
        opt      E()
< 20     0      0:
  22     0      0:           one = represents 1 library sequences
  24     0      0:
  26     0      0:
  28     0      0:
  30     0      0:
  32     0      1:*
  34     3      3:==*
  36     8      6:=====*==
  38    14     10:=========*====
  40    19     14:=============*=====
  42    14     17:==============  *
  44    16     18:================ *
  46    19     19:==================*
  48    21     18:=================*===
  50    13     16:=============  *
  52    15     14:=============*=
  54     9     12:=========  *
  56     8     10:======== *
  58     7      8:=======*
  60     7      7:======*
  62     5      5:====*
  64     5      4:===*=
  66     4      3:==*=
  68     1      3:= *
  70     2      2:=*
  72     3      2:=*=
  74     1      1:*
  76     1      1:*
  78     0      1:*
  80     1      1:*
  82     0      0:
  84     1      0:=
  86     0      0:
  88     1      0:=
  90     1      0:=
  92     0      0:
  94     0      0:
  96     1      0:=
  98     0      0:
 100     0      0:
 102     0      0:
 104     0      0:
 106     0      0:
 108     0      0:
 110     0      0:
 112     0      0:
 114     0      0:
 116     0      0:
 118     0      0:
>120     0      0:
  50600 residues in   200 sequences
 (shuffled) MLE statistics: Lambda= 0.3053;  K=0.1039
 Kolmogorov-Smirnov  statistic: 0.0515 (N=25) at  40
```

b) The score, 85, is likely significant as it still falls in the long right tail of the distribution..



```
        opt      E()
< 20     0      0:
  22     0      0:           one = represents 1 library sequences
  24     0      0:
  26     0      0:
  28     0      0:
  30     1      0:=
  32     1      1:*
  34     2      3:==*
  36     4      6:==== *
  38    12     10:=========*==
  40    24     14:=============*==========
  42    15     17:=============== *
  44    23     18:=================*=====
  46    14     19:==============    *
  48     9     18:=========        *
  50    24     16:===============*========
  52    12     14:============ *
  54    13     12:===========*=
  56     9     10:=========*
  58     6      8:====== *
  60     3      7:===   *
  62     4      5:====*
  64     7      4:===*===
  66     4      3:==*=
  68     2      3:==*
  70     2      2:=*
  72     3      2:=*=
  74     0      1:*
  76     0      1:*
  78     2      1:*=
  80     1      1:*
  82     0      0:
  84     3      0:===
  86     0      0:
  88     0      0:
  90     0      0:
  92     0      0:
  94     0      0:
  96     0      0:
  98     0      0:
 100     0      0:
 102     0      0:
 104     0      0:
 106     0      0:
 108     0      0:
 110     0      0:
 112     0      0:
 114     0      0:
 116     0      0:
 118     0      0:
>120     0      0:
  36400 residues in   200 sequences
 (shuffled) MLE statistics: Lambda= 0.3013;  K=0.08664
 Kolmogorov-Smirnov  statistic: 0.0657 (N=24) at  44
```

.