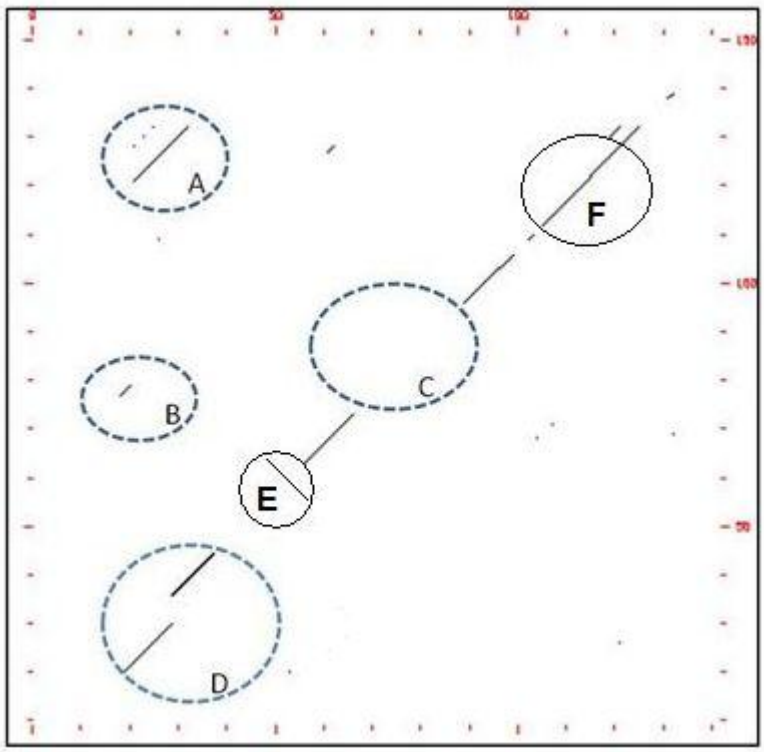
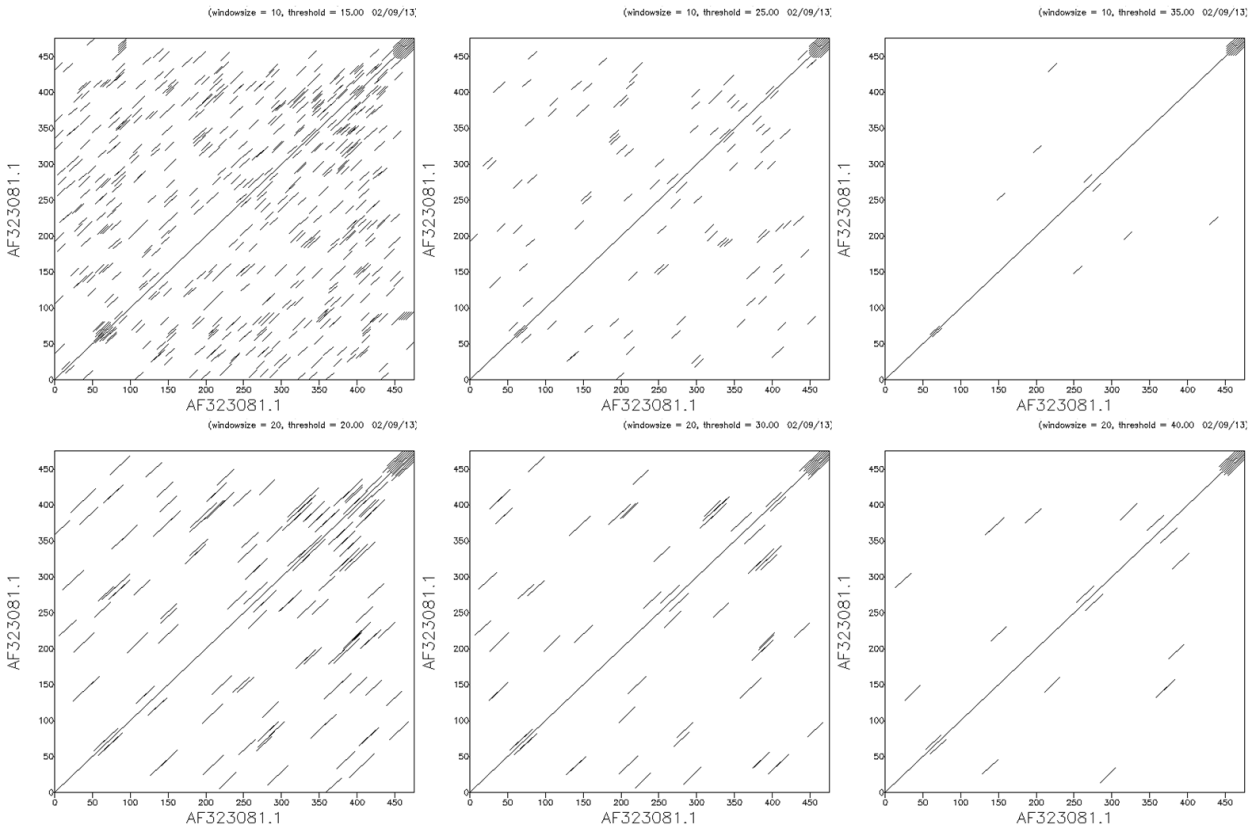
1) In the dotplot below, assuming that the diagonal lines indicate a true relationship, explain the meaning of each of the labeled regions A-F.



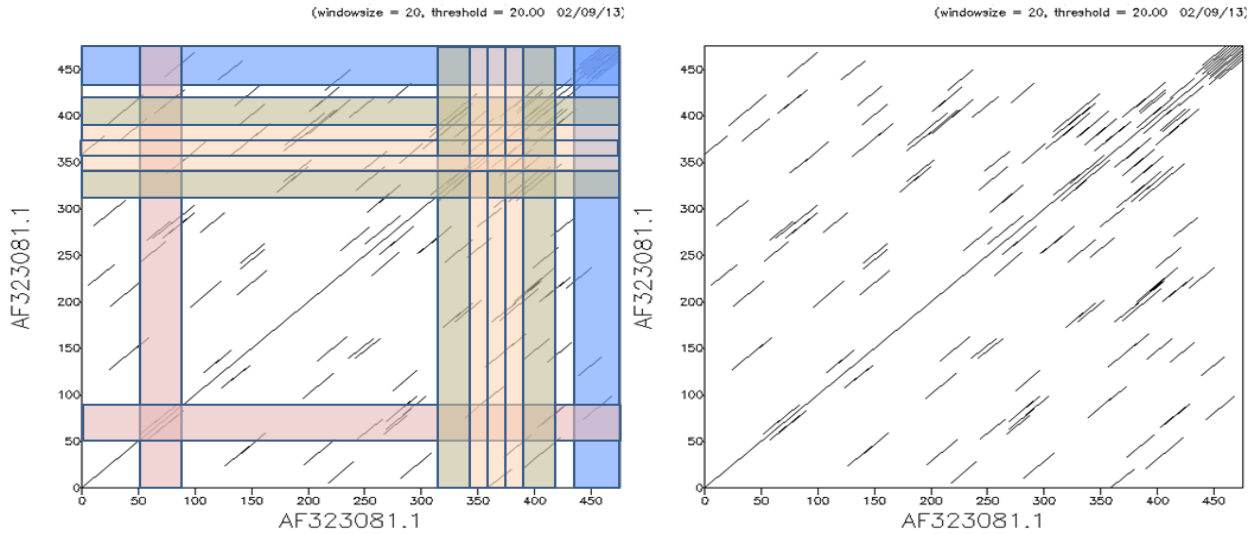
A. similar sequence (possible duplication) of a region at the beginning of the sequence (pos about 20-40) at the end (pos near 120-140), however, since there is no similar match in the southeast corner, this may be noise.
B. Since the match is short, this is most likely noise or possibly a short duplication. Since there are other short similar regions and there is no corresponding line southeast of the main diagonal, it is most likely noise.
C. Region of low similarity
D. Similar region with insertion in the vertical sequence or deletion in the horizontal sequence (must say which sequence)
E. Inverted region, only visible when sequence vs reverse complement is plotted
F. Similar region with no gaps, small duplication near the right hand end.

2a.

>gi|12584201|gb|AF323081.1| Homo sapiens resistin mRNA, complete cds, 476bp
>gi|12584199|gb|AF323080.1| Mus musculus resistin mRNA, complete cds, 591bp
>gi|21309955|gb|AF378366.1| Rattus norvegicus resistin mRNA, complete cds, 608bp



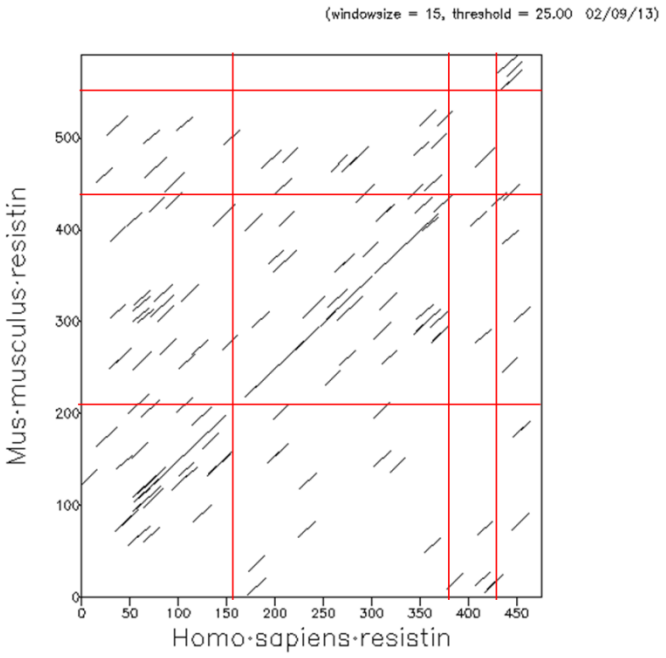
2b. in dotplots of the human sequence vs itself (above) three possibly repetitive regions can be seen at bases 50-90, 310-415, and 440-476. The first and last regions have short repeats (seen by the narrow spacing of the parallel lines. The first region is in a T/C rich region that includes TCTCTGTCTCCTCCTCCTCCCT. The third region is a repeat of GATGATGATGATGATGATGATG. The second region is more complicated. This can be seen in the window=20, threshold=20 plot. The figure below shows one possible way of breaking the repeats down. Note that the brown shaded repeat is approximately twice as long as the yellow shaded repeat, and that there is some similarity between these repeats. This suggests another interpretation of the plot: seven copies of the yellow repeat with some evolutionary divergence in the outer repeats. Because of the restricted sequences possible in coding regions this degree of repetitiveness is common. Note that, except for the third region, the repeats mostly disappear at



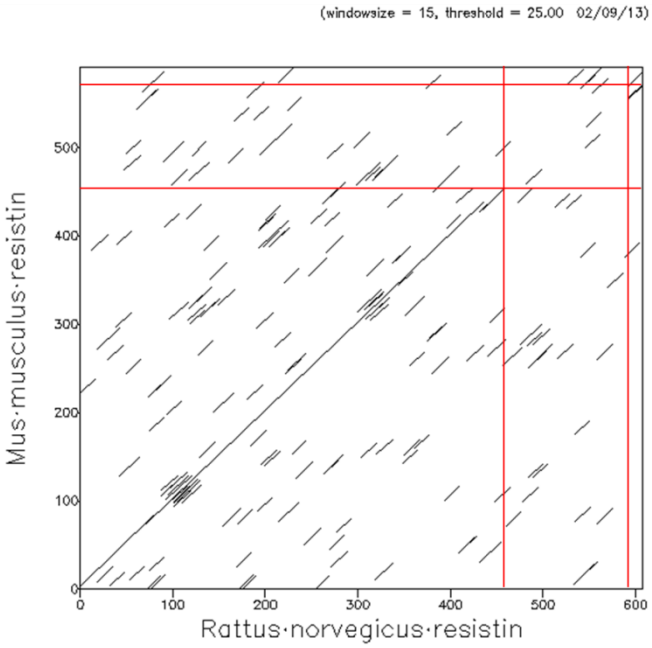
higher stringency, indicating they are weak.

2c. there was no question 2c.

2d. The most obvious feature in the comparison of human and mouse RETN is the large insertion at bases 440 – 540 in the mouse sequence with respect to the human sequence. This covers part of the second internally repetitive region of the human sequence. You may also note that the 5' end of the sequence appears to be fairly weakly conserved, a a particularly weakly conserved region at bases 160 (human) and 210 (mouse). The weak repetitiveness of the base 50-90 region in the human sequence is also present in mouse.



2e. the mouse and rat sequences are clearly more similar than mouse and human. In a dotplot made at the same window and stringency as the plot above, the main diagonal is continuous and extends all the way to the 5' end. The first and second weakly repetitive regions are clearly seen. The weak conservation of the 3' region (bases 460-590) is seen once again. The consistent weakness of the conservation suggests that there may be few or contradictory constraints on this sequence.



3. lambda repressor

locus:RPC1_LAMBD/accession P03034.2 (NP_040628 Is identical)

length: 237 aa

epsilon34 repressor

locus: ACF16671, or YP_002533506

58.3% identity (82.9% similar) in 216 aa overlap (22-237:1-212)

This level of identity and similarity, which is higher in the c-terminal 2/3 of the sequence, clearly indicates that these sequences are ancestrally related (homologous). Note that the E-value is 2.7 x 10^{-64}.

4a.

```
>gi|323710232|gb|ADY03050.1| RecA [Escherichia coli]
GPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPDTGEQALEICDALARS
GAVDVIVVDSVAALTPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNLKQSNTLLIFINQIRMKIGVMFGN
PETTTGGNALKFYASVRLDIRRIGAVKEGENVVGSETRVKVVKNKIAAPFKQAEFQILYGEGINFYGELV
DLGVKEKLIEKAGAWYSY

>gi|4275|emb|CAA45563.1| RAD51 [Saccharomyces cerevisiae]
MSQVQEQHISESQLQYGNGSLMSTVPADLSQSVVDGNGNGSSEDIEATNGSGDGGGLQEQAEAQGEMEDE
AYDEAALGSFVPIEKLQVNGITMADVKKLRESGLHTAEAVAYAPRKDLLEIKGISEAKADKLLNEAARLV
PMGFVTAADFHMRRSELICLTTGSKNLDTLLGGGVETGSITELFGEFRTGKSQLCHTLAVTCQIPLDIGG
GEGKCLYIDTEGTFRPVRLVSIAQRFGLDPDDALNNVAYARAYNADHQLRLLDAAAQMMSESRFSLIVVD
SVMALYRTDFSGRGELSARQMHLAKFMRALQRLADQFGVAVVVTNQVVAQVDGGMAFNPDPKKPIGGNIM
AHSSTTRLGFKKGKGCQRLCKVVDSPCLPEAECVFAIYEDGVGDPREEDE
```

4b. Alignment 1:

Waterman-Eggert score: 147; 44.1 bits; E(1) < 4.8e-09

28.2% identity (58.8% similar) in 170 aa overlap (21-175:212-375)

Alignment 2:

Waterman-Eggert score: 45; 16.8 bits; E(1) < 0.55

31.4% identity (62.9% similar) in 35 aa overlap (12-45:227-256)

When the different N-terminal of alignment 2 is joined onto alignment 1 (red) you see

```
               20         30         40         50         60         70          80
gb|AD  LQVIAAAQREGKTCAFIDAEHALDPI-YARKLGVDIDNLLCSQPDTGEQALEICDALAR---SGAVDVIVVDSVAAL-TPKA
       ..... ::: :     .: . ::. . :::  ..:               :.. :: :.    .  ..::::::: ::
emb|CA VRLVSIAQRFG-----LDPDDALNNVAYARAYNADHQ-------------LRLLDAAAQMMSESRFSLIVVDSVMALYRTDF
        230             240        250        260                    270        280        290

       90        100        110        120        130          140        150        160
gb|AD  EIEGEIGDSHMGLAARMMSQAMRKLAGNLKQSNTLLIFINQIRMKI--GVMFG-NPETTTGGNALKFYASVRLDIRRIGA
         .::..  .: ::  :  .:...::    : .. ..  ::.  ..  :. :. .:.   ::: .   ...:: ...  .
emb|CA SGRGELSARQMHLAKFM--RALQRLAD---QFGVAVVVTNQVVAQVDGGMAFNPDPKKPIGGNIMAHSSTTRLGFKKGKG
              300         310          320        330        340        350        360

         170
gb|AD  VKEGENVVGS
        ..  .:: :
emb|CA CQRLCKVVDS
         370
```

Both of these alignments have some attractive features. Alignment 1 preserves the positions of charged residues, and residues such as G and P that prefer turns. Alignment 2 aligns a striking and unusual sequence YARKL/YARAY.

Overall, in my opinion, alignment one seems better, especially since it preserves not only the strongest matches, but also many more similar residues (approx 29 vs 20).

```
4c-d. Waterman-Eggert score: 588;  19.1 bits; E(1) <  0.15
22.7% identity (31.7% similar) in 454 aa overlap (11-398:3-220)
Entrez Lookup  Re-search database  General re-search
[Domains]
[alignment]
                 20        30        40        50        60        70        80
emb|C  ESQLQYGNG--SLMSTVPADLSQSVVDGNGNGSSEDIEATNGSGDGGGLQEQAEAQGEMEDEAYDEAALG---SFVPIEK
       ::     .:  .:  :    : : :.            :             : :: :           :   .:.
gb|ADY ES-----SGKTTL--T----L-Q-VI------------A-------------A-AQRE-----------GKTCAFI----
                10                                          20
          90       100       110       120       130       140       150       160
emb|C  LQVNGITMADVKKLRESGLHTA-EAVAYAPRKDLLEIKGISEAKADKLLNEAARLVPMGFVTAADFHMRRSELICLTTGS
              :.    :   : : . . :: :: :    :.     :         .       :      .:.:    :
gb|ADY ---------DA----E---H-ALDPI-YA-RK-L----GV-----D---------I--------D------NLLC----S
               30                  40                                      50
          170       180       190       200       210       220       230       240
emb|C  KNLDTLLGGGVETGSITELFGEFRTGKSQLCHTLAVTCQIPLDIGGGEGKCLYIDTEGTFRPVRLVSIAQRFGLDPDDAL
       .  ::                 ::      :     :      :.:      :                        :::
gb|ADY QP-DT---------------GE------Q-----A------LEI------C-------------------------DAL
                                      60
          250       260       270       280                 290         300         310
emb|C  NNVAYARAYNADHQLRLLDAAAQMMSESRFSLIVVDSVMALYRT----------DFS--GRGELSARQMHL--AKFMRAL
            ::. .:      .:            .:::::: ::  :          : :  :   :.:: : .  :  ::
gb|ADY -----ARS-GA------VD-----------VIVVDSVAAL--TPKAEIEGEIGD-SHMG---LAAR-M-MSQA--MR--
             70                          80          90        100             110
                       320        330         340          350                360
emb|C  QRLA-D--Q------FGVAVVVTNQV-VAQVDGG-MAF-NPD---PKKPIGGNIMA---HSSTTRL-----G-FKKGK--
       .:: .  :      : .     ::. . ..  : : : ::.         :::  :   ..:. ::      :  :.:.
gb|ADY -KLAGNLKQSNTLLIF-I-----NQIRM-KI--GVM-FGNPETTT-----GGN--ALKFYASV-RLDIRRIGAVKEGENV
            120             130           140              150        160       170
          370           380         390
emb|C  -GCQ-RLCKVV-D---SPCLPEAECVFAI-Y-EDGV---G---D---PRE---E
        : . :. ::: .   .: . .::  : : : : :.   :   :    .:   :
gb|ADY VGSETRV-KVVKNKIAAP-FKQAE--FQILYGE-GINFYGELVDLGV-KEKLIE
          180         190         200         210         220
```

2/0 alignment

This alignment has many insertions and deletions and looks much like a global alignment since almost all positions of both sequences are used. Only one of the regions from the previous alignment shows in this alignment (red). Since there are so few contiguous matching regions in what are obviously homologous proteins, this alignment looks much poorer than the one with the default gap penalties. Note that the E-value for this alignment is only 0.15 (close to random), although the raw score is higher 588 vs 147.

2/2 alignment (parameters set incorrectly)

Waterman-Eggert score: 333; 29.7 bits; E(1) < 0.0001
31.1% identity (58.1% similar) in 241 aa overlap (1-208:185-394)

```
                10                  20        30            40        50             60
                10                  20        30            40        50             60
gb|AD  GPE-SSGKTTL--TL----QV---IAAAQREGKTCAFIDAEHALDPIY----ARKLGVDIDNLLCS----QP-DTGEQAL
       : :  .::. :  ::    :.   :...  ::: : .::.: .. :.     :...:.: :. : .    .  .. .: :
emb|CA G-EFRTGKSQLCHTLAVTCQIPLDIGGG--EGK-CLYIDTEGTFRPVRLVSIAQRFGLDPDDALNNVAYARAYNADHQ-L
           190       200       210           220       230       240       250
                70         80        90          100       110       120       130
gb|AD  EICDALAR--SGA-VDVIVVDSVAALTPKAEIEGEIGD-S--HMGLAARMMSQAMRKLAGNLKQSNTLLIFINQIRMKI-
       .. :: :.  : .  ..:::::: ::  .... :. :. :  .: :: ..: .:...:: .  : .. ..  ::.  ..
emb|CA RLLDAAAQMMSESRFSLIVVDSVMALY-RTDFSGR-GELSARQMHLA-KFM-RALQRLA-D--QFGVAVVVTNQVVAQVD
     260       270       280       290         300       310          320       330
           140         150       160       170       180       190         200
gb|AD  GVM-FGNPETTT--GGNALKFYASVRLDIRRIGAVKEGENVVGSETRV-KVVKNKIAAP-FKQAE--FQILYGEGINFYG
       : : : ::.     ::: .  ..:.     :.:  :.:.   : . :. ::: .   .: . .::  : : : .:.   :
emb|CA GGMAF-NPDPKKPIGGNIMA-HSSTT----RLG-FKKGK---GCQ-RLCKVV-D---SPCLPEAECVFAI-YEDGV---G
             340       350             360           370           380        390

gb|AD  E
       .
emb|CA D
```

This alignment has is considerably longer than the default alignment, and has many more indels. Large parts of the region from 30 to 150 are quite similar (shown in red). The extension beyond residue 170 (upper scale) shows some enticingly similar regions (although I wouldn't say the alignment is exactly correct). The low penalty alignment looks more like a global alignment in that it uses more of the residues of both sequences, but since it does not include all residues of either sequence, they both look like local alignments. Overall, the -2/-2 alignment has too many gaps with many singletons. This is very unlikely to be correct.