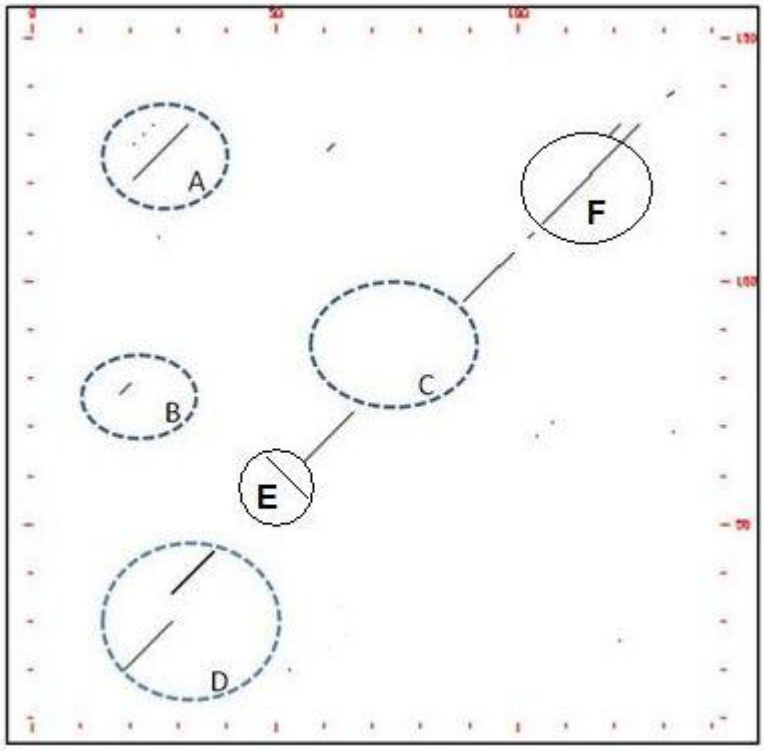
**Biol 47800/59500 Homework 2**

URLs used in this homework. See also "Resources and links" on blackboard.

- NCBI - http://www.ncbi.nlm.nih.gov/
  - nucleotide sequences: http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuccore
  - protein sequences: http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein
- Dotplot programs
  - dotmatcher from the emboss suite at http://www.genomics.purdue.edu/emboss
- LALIGN (http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=lalign&pgm=lal) at the University of Virginia. The program finds not just one, but also *n* non-overlapping alignment of two sequences according to the SIM algorithm discussed in the text. In these alignments, the same two residues will never be found together more than once.

**Questions**

**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.




**2 a-e)** NCBI provides a comprehensive repository of nucleotide and protein sequences. In the next question you will retrieve several resistin sequences from NCBI and compare them using dotplots. Use dotmatcher for analysis. Note: some features in the dotplot may not be obvious unless you increase the window size.

In dotplots, we use sliding windows to improve visualization of similar regions among sequences, instead of writing down a dot for every character that is identical in both sequences, we compare a number of consecutive positions (window size), and we make a dot whenever there is minimum number (stringency) of identical characters. For DNA sequences, sliding window=15, stringency=10, is

common; for protein sequence, sliding window=2 or 3, stringency=2 is a good place to start. Try different window sizes and stringency levels, and determine the best window/stringency combinations and what features of the plots are enhanced at different settings. Note that the EMBOSS *dotmatcher* program , by default, uses a scoring table (EDNAFULL) that scores +5 for a match and -4 for a mismatch. If you want to see the details of this scoring table, you can use the EMBOSS program *embossdata* to get a copy. In general, you will need to use higher values for stringency (number of matches in window) with such a table. Window lengths in the 5 – 50 range and stringencies in the 5 – 50 range are a good place to start (not a prescription for the only values you might want to consider).

a. Retrieve the nucleotide sequences for the human (AF323081.1), mouse (*mus musculus*, AF323080.1) and rat (*rattus norvegicus*, AF378366.1) RETN nucleotide sequences. Choose "Nucleotide" database and use a query such as "RETN mRNA" or "resistin mRNA complete cds" for sequence retrieval. Be sure to choose the mRNA sequences and not genomic scaffold or gene sequences. Note that some of these sequences have more than one splice variant (labeled as "transcript variant"). The primary sequence will usually be labeled simply as mRNA, or complete coding sequence (CDS). If you don't see the primary sequence, it may help to click on "Related Sequences". The sequences you are looking for should be roughly 475-610 bases in length. You will want to retrieve these sequences in "FASTA" format. **Report the IDs, titles, and lengths, of the matching sequences you find, and the sequences you select.**

b. Using a dotplot analysis, compare the human sequence versus itself. **Based on the dotplot, describe the sequence features (e.g., possible insertions, deletions, duplications, etc.) and the positions in the sequence where they occur. Report the stringency and threshold values that you used to see the features you report. Include the most informative plot in your report.**

d. Using a dotplot analysis, compare the mouse RETN sequence versus the human RETN. **Describe and explain the patterns you observe in the dot plot as in 2b) above.**

e. Using a dot plot, compare it the rat RETN sequence with the mouse RETN. **Is rat RETN more or less similar to mouse RETN than human RETN? Explain your reasoning, and provide one or two plots to support your conclusion?**

**3) For this question, use the program LALIGN, above, and the default conditions provided by the site.**

Align the **protein** sequences for the bacteriophage λ and the *salmonella* epsilon34 phage cI repressors. Make sure you get the correct sequences – bacteriophage have multiple repressors, the ones you are looking for are 210-240 residues long. Look up the sequences at NCBI and **report the IDs and lengths of the sequences you select**. Cut and paste FASTA formatted sequences into the LALIGN site. **Record the resulting percent identity and similarity and briefly comment on what these alignments suggests about the evolutionary relationship between the sequences**. If you see very low or no similarity, you have probably chosen the wrong sequences.

**4 a-d) For this question, use the program LALIGN, above. Use the default Blosum 50 scoring matrix.**

a. Obtain the following two sequences from NCBI in FASTA format: *Escherichia coli* recA **protein** (ADY03050.1), and *Saccharomyces cerevisiae* rad51 protein (CAA45563.1). These proteins have the same function, *i.e.*, promoting the pairing of homologous single-stranded DNAs. They almost certainly

have similar three-dimensional structures, but have diverged enough that they are somewhat difficult to align.

b. Use LALIGN to align the above two sequences with gap penalties of –12 and –2 (the default). **Report the length of the alignment, the percent identity, and the score of the alignment.** There is an alternative alignment for the first approximately 45 residues (after this region the two alignments would have the same c-terminal portion). **Which seems "correct"? Why?**

c. Repeat the alignment with gap penalties of –2 and 0 and **report the score and the features you observe in the alignment.**

d. **Describe what happened when the gap penalties were reduced. Which of these alignments looks like a local alignment and which looks like a global alignment? Which looks more "correct" and why?** A dotplot may help you understand the relationship between these sequences.