Software Challenge: Genome Rearrangements and Sequence Alignment
02-604: Fundamentals of Bioinformatics

Pretend that you are working with Mohamed Marahiel in 1996 to discover the non-ribosomal code by determining the structural components of gramicidin synthetase, a protein that makes the non-ribosomal peptide gramicidin. You know that it has an adenylation domain (A-domain), but you do not know where it is located in the sequence of amino acids making up the protein.

Fortunately, Peter Brick just published the 3-D structure of firefly luciferase, which has similarities to A-domains. You think that this information might prove useful in finding the A-domain of gramicidin synthetase, and so you obtain the amino acid sequence of gramicidin synthetase, which you decide to compare against the much shorter sequence for firefly luciferase to locate the A-domain in the gramicidin synthetase sequence. Throughout this challenge, you will need two datasets (in FASTA format):

- the amino acid sequence of gramicidin synthetase (grs.fa)
- the amino acid sequence of firefly luciferase (firefly_luc.fa)

First, let's check whether gramicidin synthetase is similar to firefly luciferase. To this end, we could run a local alignment algorithm that we encountered in the main text. But what we would like to do is align gramicidin synthetase against **all** firefly proteins to see if firefly luciferase really is the most similar to gramicidin synthetase. Unfortunately, such a task is computationally very intensive -- especially in 1996!

Instead, we will use a heuristic called **BLAST** (the **B**asic **L**ocal **A**lignment **S**earch **T**ool) that does not guarantee an optimal alignment, but which quickly returns a measure of similarity hits of a sequence against a database. BLAST was published in 1990 in one of the most cited scientific papers of all time.

In general, if we are searching a protein against a database and find a hit with score S, then the **E-value** of S is the expected number of hits in searches of this protein against a *random* database of the same size. Thus, the *smaller* the E-value, the *less* likely that the hit resulted from random noise, and the *more* statistically significant the result.

For a given match of two sequences, the **percent identity** corresponds to the percentage of residues that are identical in the two sequences at the same positions in the alignment.

Run only the gramicidin synthetase sequence (grs.fa) on BLASTp, the version of BLAST used for aligning an amino acid sequence against a database of proteins: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Use the non-redundant protein database, and specify the organism to be the "North American firefly (taxid: 7054)"; otherwise, use default parameters.

Three points each.

1. **Consult the "descriptions" section and report the E-value and the percent identity of the best match.**

2. **Was firefly luciferase identified as a statistically significant match? Explain your answer.**

3. **Consult the "alignments" section. According to the first alignment, report the start and end index of the putative A-domain of the gramicidin synthetase sequence.**

Now that we have examined the statistical significance of the protein match, we will align gramicidin synthetase (grs.fa) with firefly luciferase (firefly_luc.fa).
In particular, we will use EMBOSS to perform a global and local alignment of the two sequences.

- EMBOSS Needle (Global alignment)
- EMBOSS Water (Local alignment)

In both cases, click on "More Options" and select the "PAM150 scoring matrix"; otherwise, use default parameters.

4. **Is global alignment or local alignment more appropriate in this case? Give a short explanation that includes a comparison of the percent identity and total score of each alignment.**

5. **What do each of these alignments suggest is the A-domain of gramicidin? Report the start and end index of the putative A-domain for each alignment. How do these values compare with what BLAST reported?**

Rerun EMBOSS Water (but not EMBOSS Needle) with the following parameter values for an alignment with affine gap penalties (continue using PAM150):

- GAP OPEN = 20, GAP EXTEND = 0.2
- GAP OPEN = 5, GAP EXTEND = 1.0

6. **How do these alignments compare with the local alignment that you generated using the default parameters? Which of the three alignments is likely to be the most biologically relevant in this case? Explain your answer.**

Now that we have verified the similarity of gramicidin synthetase to firefly luciferase, we would like to construct a multiple sequence alignment between the gramicidin synthetase sequence and other known A-domains.

For this task, we will use an extremely popular program called **Clustal Omega** (like BLAST, the original Clustal paper is one of the most cited scientific papers of all time). We will examine PheA, which corresponds to a segment of gramicidin synthetase that codes for phenylalanine.

Run Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) on a_domains.fa, which includes PheA as well as the A-domains of eight other non-ribosomal peptide synthetases from various bacteria. Use default parameters with the Output Format "**Clustal w/ Numbers**." Examine the resulting output (use the Show Colors button and the Result Summary tab).

7. **Upload a snapshot of the phylogenetic tree and the percent identity matrix generated for this alignment as an image file.**

Marahiel determined a handful amino acid positions that are responsible for determining the amino acid that binds to the A-domain. Five of those amino acids correspond to positions 236, 239, 278, 299, and 301 of the PheA sequence.

8. **Consult the multiple sequence alignment produced by Clustal Omega. Based on the positions reported by Marahiel, do the A-domain sequences appear to code for the same amino acid? Explain your answer.**

The remaining residues appear in a window from positions 320 – 332 of the PheA sequence which are reproduced below (with gaps represented by the symbol X):

Ite	323	VNVYGPTEVTIGCS	336
Yp5	724	FNTYGPTEATVVAT	737
Abw	755	INAYGPSEAHXLVS	767
Dhv	291	MNTYGPTEATVAVT	304
Yp0	793	INEYGPTETTVGCT	806
Np7	755	VNVYGPTEATGHCL	758
Vsq	750	INCYGPTEGTXVFA	762
PheA	320	INAYGPTETTXICA	332
Aap	659	VNNYGPTETTXVVA	671

Can we locate the positions in this window that are responsible for determining the amino acid that binds to the A-domain?

9. **Report your top three candidates as positions in the PheA sequence. (Hint: construct a sequence logo from these sequences as a starting point via WebLogo – http://weblogo.threeplusone.com/.) Why did you choose these candidates?**

10. **Some positions in the multiple alignment show very high conservation between all sequences. What is a possible biological interpretation for this conservation?**