**Locating the Adenylation Domain of the Non-ribosomal Peptide Gramicidin via Comparison to Firefly Luciferase**

In this Application Challenge, we are going to hop in the Bioinformatics Time Machine and head back to 1996.

Pretend that you are working with Mohamed Marahiel 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 be 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](https://en.wikipedia.org/wiki/FASTA_format)):

* the amino acid sequence of gramicidin synthetase ([grs.fa](http://bioinformaticsalgorithms.com/data/challengedatasets/grs.txt))
* the amino acid sequence of firefly luciferase ([firefly\_luc.fa](http://bioinformaticsalgorithms.com/data/challengedatasets/firefly_luc.txt))

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](http://www.ncbi.nlm.nih.gov/pubmed/2231712) 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](http://bioinformaticsalgorithms.com/data/challengedatasets/grs.txt)) 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.

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

uncharacterized protein LOC116175982 [Photinus pyralis]

E-value = 3e-57 Percent identity = 28.55%

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

Yes, there are several versions of firefly luciferase in the results, each with E-values on the order of 1e-12 to 1e-14. It is highly unlikely that the extent of alignment occurred by random chance.

**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.**

Start: 41 End: 613

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): [http://www.ebi.ac.uk/Tools/psa/emboss\_needle/](%20http://www.ebi.ac.uk/Tools/psa/emboss_needle/%20)
* EMBOSS Water (Local alignment):<http://www.ebi.ac.uk/Tools/psa/emboss_water/>

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

**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.**

Even theoretically, local alignment is better for answering our question. Since we are looking for a specific domain in the sequence, it is better to find long regions of similarity rather than an alignment that introduces excessive indels to generate the highest score it can for the whole sequence.

Looking at the results, the percent identity for local alignment (21.3%) is higher than that of global alignment (10.3%) and the local alignment score (171.0) is higher than the global alignment score (152.0).

**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?**

For the global alignment, the start index is 1 and the end index is 535.

For the local alignment, the start index is 67 and the end index is 525.

These values are somewhat close to what BLAST reported (start: 41 end: 613), suggesting that the adenylation domain is the first domain of the sequence. However, the global alignment starts at the very first amino acid, while the local alignment and BLAST start later.

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

**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.**

GapOpen20\_GapExtend0.2 has 29.1% identify and a score of 91.6. The aligned region is only about 100 amino acids long (368-461) due to the strict penalty of beginning an insertion or deletion. There are some gaps, and they are longer because of the lower extension penalty.

GapOpen5\_GapExtend1.0 has 25.7% identity and a score of 309.0. The aligned region is larger (35-548), similar to the alignment with default parameters. There are more gaps, but they are shorter due to a more flexible penalty for beginning indels but higher penalty for extending.

Since GapOpen5\_GapExtend1.0 has the highest score, it is the most likely to be biologically relevant. We do note that having too low a penalty may create an artificially high score, so the original (with middle-ground penalty) may be the safer choice.

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**. 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](http://bioinformaticsalgorithms.com/data/challengedatasets/a_domains.txt), 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).

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

**A screenshot of a graph

AI-generated content may be incorrect.**

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.

**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.**

There is not much amino acid conservation at these sites. The positions only have up to two sequences with a common amino acid. So, the A-domain sequences do not appear to code for the same amino acid. Perhaps these amino acids are responsible for the functional differences between the nine sequences, not their similarities.

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 |

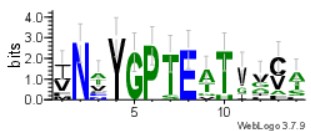
Can you locate the positions in this window that are responsible for determining the amino acid that binds to the A-domain?

**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/**](http://weblogo.threeplusone.com/)**.) Why did you choose these candidates?**

Since all of these sequences have a similar function, we can infer that highly conserved underlying structures (in this case amino acids) are responsible.

The three most conserved positions between these sequences (using indexing from PheA) are Y323, N321, and P325. However, these are not the positions that determine which amino acid binds to the A domain. If the positions differentiating amino acid binding were the same, then the same amino acid would bind all of them. Therefore, we are looking for differences between these sequences.

Three of the least conserved positions are 322, 330, and 332 (positions 320, 328, 330, and 331 are also not well conserved). These are the positions that determine which amino acid binds to the A domain.

****

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

Since all of these sequences have a similar function, we can infer that highly conserved underlying structures (in this case amino acids) are responsible. These regions could be important binding motifs for common regulatory proteins or essential amino acids for the 3D structure of the proteins.