✕ Lessons

Review Classmates: Bioinformatics Application Challenge

Review by September 14, 11:59 PM PDT

Reviews 5 left to complete

Bioinformatics III Final

by Muhammad Nuri
Submitted on September 5, 2016

♡ like ⚐ Flag this submission

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.

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

E-value: 7e-17
Percent Identity: 22%

The best match had an E-value of 7E-17 (7*10^-17) with a 22% percent identity.

○ 0 pts
The learner correctly identifies neither the E-value nor the percent identity.

○ 2 pts
The learner correctly identifies only one of the E-value and the percent identity.

● 4 pts
The learner correctly identifies both the E-value and the percent identity.

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

Yes, because the E-value is near to zero.

GRS amino acid sequence is ~1000 characters. So for it to appear as a random sequence is very low.

Yes, the firefly luciferase (*Photinus pyralis*) was found to be statistically significant because the matches with the five lowest E-values all corresponded to firefly luciferase, and these values were all extremely small, indicating that the matches are unlikely to be due to random chance.

○ 0 pts
The learner fails to identify that firefly luciferase is statistically significant.

○ 2 pts
The learner identifies that firefly luciferase is statistically significant but fails to provide a complete explanation.

○ 4 pts
The learner identifies that firefly luciferase is statistically significant and provides a complete explanation.

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: 65
end: 525

Assuming that the alignment occurs at the A-domain of gramicidin synthetase, the start and end indices of that domain are 65 and 525, respectively. (51-537 are the indices of the luciferase enzyme, an incorrect answer.)

○ 0 pts
The learner fails to identify either 65-525 or 51-537 as the start and end indices of the A-domain.

○ 2 pts
The learner identifies 51 and 537 as the start and end indices of the A-domain.

○ 4 pts
The learner identifies 65 and 525 as the start and end indices of the A-domain.

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):xn-- http://www-.ebi.ac.uk/Tools/psa/emboss_needle/
- EMBOSS Water (Local alignment):xn-- 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.

Help Center

Global Alignment Percent Identity / Total Score: 10.3% / 152.0
Local Alignment Percent Identity / Total Score: 21.3% / 171.0

Local Alignment is appropriate in this case.

Gramicidin Amino Acid Sequence length is 1179 characters and Firefly Luciferase is 521. Firefly luciferase is much smaller. When comparing smaller sequences to longer sequences, Local Alignment should be used.

- (1 point) Percent identities: (local - 21.3% vs. global - 10.3%)
- (1 point) Total score: local - 171 vs. global - 152
- (2 points) Analyzing the more appropriate alignment: the percent identity and total score are significantly higher for the local alignment, suggesting that in this case, local alignment is more appropriate than global alignment.

○ 0 pts
(Please see grading criteria above)

○ 1 pt
(Please see grading criteria above)

○ 2 pts
(Please see grading criteria above)

○ 3 pts
(Please see grading criteria above)

○ 4 pts
(Please see grading criteria above)

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?

Global Alignment Start - End: 1 - 534
Local Alignment Start - End: 67 - 525

Local Alignment (67-525) compares well to BLAST reported result (65-525).

- (1 point) For global alignment, the putative A-domain is 1 to approximately 534 (after this point, there are only gaps in the top line of the alignment).
- (1 point) For local alignment, the putative A-domain extends from position 67 to 525 of gramicidin synthetase (based on the ends of the alignment).
- (2 points) The starting and ending position of the local alignment in particular are very similar to the starting and ending position reported by BLAST.

○ 0 pts
(Please see grading criteria above)

○ 1 pt
(Please see grading criteria above)

○ 2 pts
(Please see grading criteria above)

○ 3 pts
(Please see grading criteria above)

○ 4 pts
(Please see grading criteria above)

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.

Gap Open (20) / Extend (0.2):
Score: 91.6
Percent Identity: 29.1
Start-end: 368-470

Gap Open (5) / Extend (1.0):
Score: 309.0
Percent Identity: 25.7%
Start-end: 35-540

Gap Open=20, Gap Extend=0.2 would be most biologically relevant. The reason is that in biology indels are added in bulk rather than single. So gap extend penalty should be lower.

Gap Open=5, Gap extend = 1.0 lowers the indel penalty and thus directs the algorithm to match as much as possible by insertion/deletions.

The default parameters has Gap Open = 10 and Gap Extend = 1.0. It is better than Gap Open = 5, Gap Extend=1.0 because it penalizes indels a bit more.

Gap Open = 20 and Gap Extend = 0.2 looks to be more biologically relevant.

The reason is that biologically insertions/deletions happen in batches rather than singles. Secondly if indel penalty is low (as in Gap Open=5), one increases the alignment at the expense of insertion/deletions mostly.

- (1 point) The first alignment is much shorter and has a larger number of matches but also longer gaps than with the default parameters.
- (1 point) The second alignment has fewer contiguous matches with more short gaps.
- (2 points) The second alignment is most likely to be biologically relevant because it has the highest score (309.0) of any alignment considered, but a reasonable argument can be made for the default parameters being relevant as well.

○ 0 pts
(Please see grading criteria above)

○ 1 pt
(Please see grading criteria above)

○ 3 pts
(Please see grading criteria above)

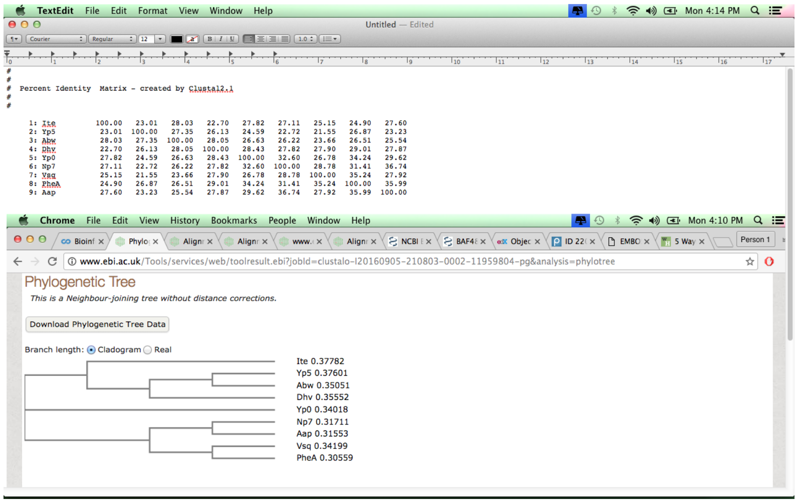
○ 4 pts
(Please see grading criteria above)

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, 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.



```
#
#
# Percent Identity  Matrix - created by Clustal2.1
#
#

     1: Ite          100.00   23.01   28.03   22.70   27.82   27.11   25.15   24.90   27.60
     2: Yp5           23.01  100.00   27.35   26.13   24.59   22.72   21.55   26.87   23.23
     3: Abw           28.03   27.35  100.00   28.05   26.63   26.22   23.66   26.51   25.54
     4: Dhv           22.70   26.13   28.05  100.00   28.43   27.82   27.90   29.01   27.87
     5: Yp0           27.82   24.59   26.63   28.43  100.00   32.60   26.78   34.24   29.62
     6: Np7           27.11   22.72   26.22   27.82   32.60  100.00   28.78   31.41   36.74
     7: Vsq           25.15   21.55   23.66   27.90   26.78   28.78  100.00   35.24   27.92
     8: PheA          24.90   26.87   26.51   29.01   34.24   31.41   35.24  100.00   35.99
     9: Aap           27.60   23.23   25.54   27.87   29.62   36.74   27.92   35.99  100.00
```

www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20160905-210803-0002-11959804-pg&analysis=phylotree

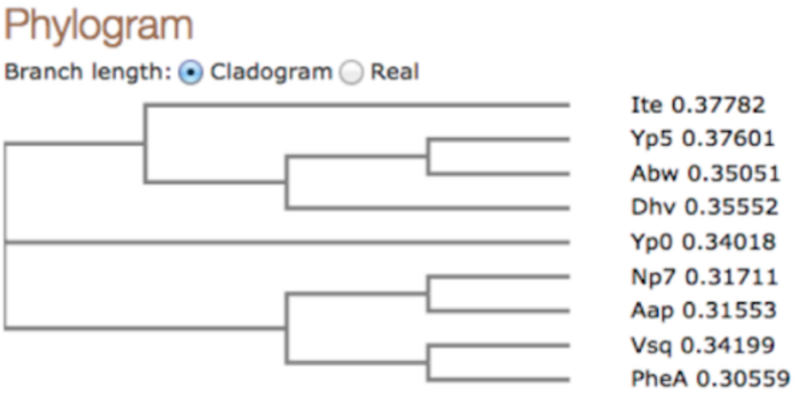
Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

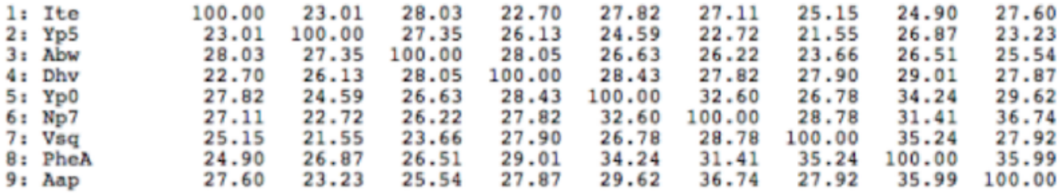
Download Phylogenetic Tree Data

Branch length: ● Cladogram ○ Real

(2 points) Phylogenetic tree:



(2 points) Percent identity matrix:



```
1: Ite          100.00   23.01   28.03   22.70   27.82   27.11   25.15   24.90   27.60
2: Yp5           23.01  100.00   27.35   26.13   24.59   22.72   21.55   26.87   23.23
3: Abw           28.03   27.35  100.00   28.05   26.63   26.22   23.66   26.51   25.54
4: Dhv           22.70   26.13   28.05  100.00   28.43   27.82   27.90   29.01   27.87
5: Yp0           27.82   24.59   26.63   28.43  100.00   32.60   26.78   34.24   29.62
6: Np7           27.11   22.72   26.22   27.82   32.60  100.00   28.78   31.41   36.74
7: Vsq           25.15   21.55   23.66   27.90   26.78   28.78  100.00   35.24   27.92
8: PheA          24.90   26.87   26.51   29.01   34.24   31.41   35.24  100.00   35.99
9: Aap           27.60   23.23   25.54   27.87   29.62   36.74   27.92   35.99  100.00
```

○ 0 pts
(Please see grading criteria above)

○ 2 pts
(Please see grading criteria above)

○ 4 pts
(Please see grading criteria above)

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.

No

The A-domains do not appear to code for the same amino acid because the columns of the A-domain multiple alignment corresponding to positions 236, 239, 278, 299, and 301 of PheA are very poorly conserved.

○ 0 pts
The learner fails to identify that the A-domains do not seem to code for the same amino acid and offers a limited explanation.

○ 2 pts
The learner identifies that the A-domains do not seem to code for the same amino acid but does not offer a reasonable explanation.

○ 2 pts
The learner fails to identify that the A-domains do not seem to code for the same amino acid but attempts to offer a reasonable explanation

○ 4 pts
The learner identifies that the A-domains do not seem to code for the same amino acid and offers a reasonable explanation.

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/.) Why did you choose these candidates?

320, 322 and 328 because they are very different at these places suggesting that amino acids are getting built at these places

The answer depends on the answer to the preceding question. If the previous question was answered correctly, then the correct answer is any 3 out of the 5 indices: 320, 322, 328, 330, 331, and 332 (1 point each).

However, if the previous question was answered incorrectly, then the learner can earn 2 points by selecting the highest conservation indices (3 out of 5 of the following: 321, 323, 324, 325, and 327). Give 1 point if only 1 or 2 indices are found.

(1 point) These indices are chosen because they have the lowest conservation, since the A-domains code for different amino acids. (Note: the true coding indices are 322, 330 and 331.)

○ 0 pts
(Please see grading criteria above)

○ 1 pt
(Please see grading criteria above)

○ 2 pts
(Please see grading criteria above)

○ 3 pts
(Please see grading criteria above)

○ 4 pts
(Please see grading criteria above)

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

321 (N), 323 (Y), 325(P).

These are structural parts of A-domain necessary to build amino acids.

There is high conservation in the non-coding amino acid positions because those positions are important for the function of the peptide-making machinery and must be conserved.

Other correct answers may include the key words “homology” and “function”.

○ 0 pts
The learner provides an unreasonable answer.

○ 2 pts
The learner provides a somewhat reasonable answer that does not follow the guidelines above.

○ 4 pts
The learner provides a reasonable answer, in particular one that follows the guidelines above.

(optional) Please provide any additional feedback that you would like to give here.

Submit Review

Comments

Visible to classmates

share your thoughts...