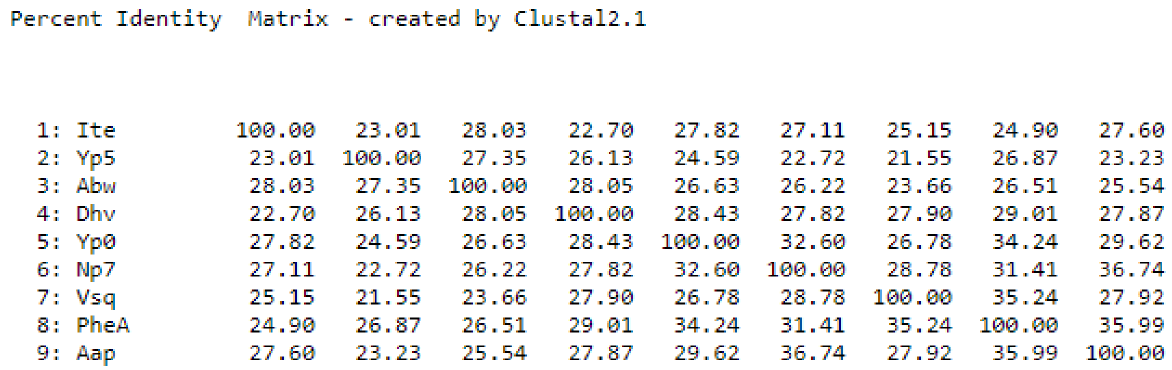
Sequence Alignment Sofrware Challenge

02-604

Owais Sarwar

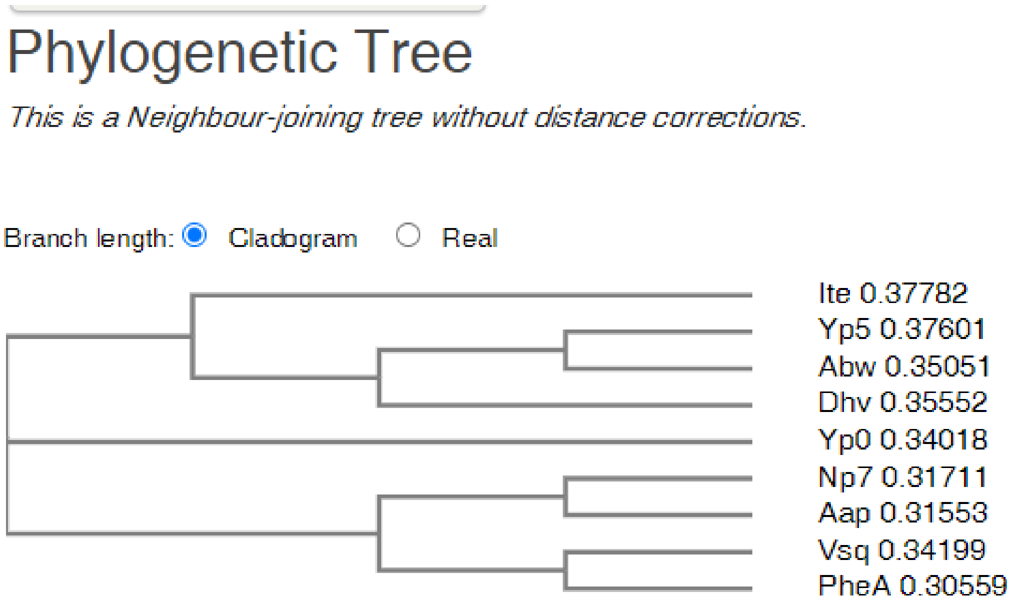
1. Best match: *uncharacterized protein LOC116175982 [Photinus pyralis]*
   a. Score: 214
   b. E-value: 3e-57
   c. Percent Identity: 28.55%
2. Yes it was. BLAST itself reports 100 "Significant alignments" with E-values below their significance threshold. Looking at the alignments, we see that they are extremely small—mostly very very near to zero—and thus the probability of these hits are tiny had they been generated against a random database of the same length.
3. Start: 42; End: 612
4. A local alignment is more appropriate than a global one, the latter which fails to identify important areas of local similarity. As EMBOSS Needle shows, the percent identity of the global alignment of the strings is only 10.3%, with 22.3% similarity, 60.2% gaps, and a total score of 152. The local alignment from EMBOSS Water improves upon all of these metrics—a higher percent identity (21.3%), similarity (44.9%), and score (171), with less gaps (19.4%).
5. Global: Start - End = 6 - 535 ; Local: Start - End = 67 - 525 ; These values are somewhat similar as in BLAST—that is, in the same neighborhood considering the sequence is about 1100 residues long. The local sequence overlaps more than from the global, with a closer start point (if slightly farther end point than the global). But both of these ranges are 10-15% shorter or so than the BLAST sequence.
6. The (20/0.1) alignment is much shorter than the original (about 100 residues vs 500), has higher similarity and identity (about 46% and 29%, respectively) but more percentage gaps (also lower score, which is to be expected since it is much shorter). The third (5/1) alignment has higher similarity and identity (about 46% and 26%, respectively) than the original too and a much higher score while being only slightly longer than the first alignment and having slightly more percentage gaps. While the (20/0.1) alignment has very slightly higher similarity/identity metrics and lower gap than the (5/1) alignment, I would say that the (5/1) is more biologically relevant because it only has marginally worse metrics but is much longer and closer to the length of other A-domains (which are in the neighborhood of 500 residues).



```
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
```

7.

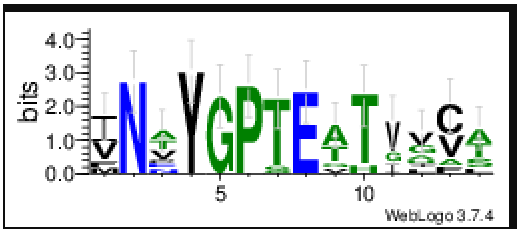


Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

8. They do not appear to code for the same amino acid since the positions corresponding to the PheA sequence do not align with identical amino acids in the other A-domains. If they did, we would expect to see that the relative positions of the other A-domains also contain those amino acids.



9. Among the non-conserved columns (i.e. those with low entropy) I would likely first consider 322, 328, and 330. Clearly, the conserved core cannot differentiate the amino acids produced so the remaining variable columns are the best bed. I picked positions in close proximity to conserved regions because if positions close to a highly conserved region differ, it makes sense to expect that those positions influence the amino acid being coded for.

10. These positions form the conserved core of the sequences. It is possible that these positions help to define the A-domains as A-domains, and are related to their functionality as such.