Bio 131 Project Final Report
Yifei Zhang
Anna Ritz
May 7, 2018

Aligning Grass Protein Sequences Using PAM-Modified Global Alignment

Project URL: https://repl.it/@yifzhang/Final-Project-Aligning-Grass-Protein-Sequences

The motivation of my project is to utilize the PAM scoring matrix in scoring protein alignments. After reading Margaret Dayhoff s paper[i] concerning the PAM matrices, I want to implement the actual data in the PAM 250 Matrix. The biological question raised here is to compare the same protein across different species to see how they are similar, and how closely are these species related. A total of three different proteins is compared for two species (three cultivars) from the grass family: Oryza Indica group, Oryza Japonica group, and Zea L.

Gramineae, also known as grasses, is a large family of monocotyledonous flowering plants containing aroung 780 genera and 12000 species.[ii] It is also the most economically important plant family, with maize, wheat, rice, barley and millet in its category. Zea L., also called corn, maize, or Indian corn, is the best-known species in genus Zea in the grass family. Rice is the seed of the rice plants.

lists, I build my getPAM function which contains a dictionary of pairwise PAM distances, with any two proteins input it will yield a score output, suggesting the relevant tendency to mutate.

The second step is to modify the Global alignment function in HW6.2. Here, as I have the PAM dictionary, I do not need the match and mismatch weights anymore; I replace match with the match score (a positive number) on the diagonal of the pam matrix, for example int(pam[string1][a-1][string2][b-1]). I also replace the mismatch score with the output from the pam dictionary for any mismatched proteins. Otherwise, the function is similar with that in HW6.2: a blank table is initialized together with a blank backtrack table, and each space in the table is filled in one by one using either directions d, s, e (back track) or scores (table). A retracing of the backtrack table helps building the two actual alignments from the last digit of both sequences using a while loop. Finally, the two alignments are reversed to give the alignments.

By reading and comparing the score output from the modified global alignment function, I do a simple analysis on my data.

Fig 1. Program output (scores)

From the scores, I can see that: scores for the second and third sequences are always higher than either one of the sequences scoring with the first sequence. This means that the second and third species (the two rice cultivars) are more closely related to each other than any of the rice cultivars compared with the maize. This is expected because the two rice cultivars are put in the same species, as they are subgroups, while maize is a different species.

Another thing that I observed is that the scores are all very similar for the first and third protein, but not so similar for the second protein. The second protein is the GS3 protein which regulates grain length and weight. This is also expected because rice and corn are very different in their seed length and weight, with corn being heavier than rice.

Next, I look at the sequences. For the first and second pair of alignments, both sequences are largely similar, with around 5 indels in each alignment sequence. This shows that the structure of granule-bound starch synthase is not so different for maize and rices. The third pair of alignment is almost identical, with only one digit's dif

For the fourth and fifth pair of alignments, both sequences are significantly similar, with around 20 indels in each alignment sequence. This shows that the structure of GS3 protein for maize and rices still has significant similarities. Again, the sixth pair of alignment is almost identical, with only one digit s difference in length. GS3 protein is almost identical for the two cultivars of rice.

For the seventh and eighth pair of alignments, both sequences are largely similar, with around 3 indels in each alignment sequence. This shows that the structure of badh2 is not so different for maize and rices. The ninth pair of alignment is-0.5 (or4.4.4.4.4.40.2 (l) o0.24r4.4.4.4.) 0(l) o0.24r (or4.g10.2

Bio 131 Project Final Report
Yifei Zhang
Anna Ritz
May 7, 2018

Bio 131 Project Final Report
Yifei Zhang
Anna Ritz
May 7, 2018

B2. Seed length and weight protein [Oryza *sa i a* Indica Group]
GenBank: BAH89236.1
233

Bio 131 Project Final Report
Yifei Zhang
Anna Ritz
May 7, 2018

Bibliography