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 *sativa* Indica group, Oryza *sativa* Japonica group, and Zea *mays* 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 *mays* 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. Rice varieties come in many shapes, colors and sizes, and may be different in genetics, grain length, thickness, stickiness, aroma and other characteristics. The list of rice varieties is also known as rice cultivars. Asian rice (Oryza *sativa*) is most widely known and most widely grown, having two major subspecies and over 40,000 varieties.[iii] Those two subspecies, indica and japonica, can be distinguished by length and stickiness. Indica rice is long-grained and not sticky, while japonica is short-grained and glutinous. [iv]

The first protein analyzed is granule-bound starch synthase, which density is positively correlated with starch concentration in seed and amylose concentration in starch.[v] Starch in grass seeds consists of amylopectin and amylose, and these two component determines whether the seed is glutinous after it is cooked. High amylopectin amount makes the seed sticky.[vi]

The second protein is the GS3 protein, also known as seed length and weight protein. Grain yield in many cereal crops is largely determined by grain size, and GS3 functions as a negative regulator of grain size and organ size.[vii]

The third protein is betaine aldehyde dehydrogenase, also known as badh2. An allele located on the gene is a major factor associated with aroma.[viii]

For each species or cultivars, three protein sequences are found on the NCBI database. As these sequences contain spaces and numbers, I defined a clean function to remove numbers, split spaces and concatenate the strings. Then these sequences are going to be aligned for each protein, pairwise, using the modified global alignment function. However, before applying this clean function, I must first manually delete all line breaks.

I obtained the PAM250 matrix from Anna and put it in a file. Then, as the protein sequences are all lower-cased, I create a new file and put into it a lower-cased pam250 matrix. Afterwards, in my main function, I open this file and read its lines without the spaces. From this pam list of 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.

```
print(globalAlign(A1,A2,-5))#2630
print(globalAlign(A1,A3,-5))#2626
print(globalAlign(A2,A3,-5))#2994

print(globalAlign(B1,B2,-5))#631
print(globalAlign(B1,B3,-5))#626
print(globalAlign(B2,B3,-5))#1401

print(globalAlign(C1,C2,-5))#2371
print(globalAlign(C1,C3,-5))#2349
print(globalAlign(C2,C3,-5))#2513
```

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 difference in length. Granule-bound starch synthase is almost identical for the two cultivars of rice.

Yifei Zhang
Anna Ritz
May 7, 2018

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 strictly identical. Badh2 protein is the same for the two cultivars of rice, which explains the similar fragrance.

The unexpectedly tricky part of this project is finding same protein sequence for each species. What I did was searching the NCBI database and try to find identical proteins, but usually they are named differently, so they are not always placed under the same list. Later I learnt some searching skills from Anna and try to use BLAST, which almost always gives me the sequences I want in about 2 minutes.

I learnt (unexpectedly) that white rice and brown rice are actually milled and unmilled rice instead of being two different species; also, there is a species of rice called Oryza rufipogon which is red in color.

Given this method that I use, what I will do next is to analyze more species from the grass family, for example to construct a simple phylogenetic tree using alignment results; I can also dig into different proteins and find out more about the similarities across species.

If the method is generalized, I can possibly obtain a simple version of BLAST for protein alignment, applying it to multiple pairs of sequences at the same time.

Appendix page: Raw data

A1. Granule-bound starch synthase 1 [Zea *mays*]
NCBI Reference Sequence: NP_001105001.3
609
maalatsqlvatraglgvpdastfrrgaaqglrggrtasaadtlsmrtsaraaprlqhqq
qqqarrgarfpslvvcasagmnvvfvgaemapwsktgglgdvlgglppamaanghrvmvv
sprydqykdawdtsvvseikmgdryetvrffhcykrgvdrvfvdhplflervwgkteeki
ygpdagtdyrdnqlrfsllcqaaleaprilslnnnpyfsgpygedvvfvcndwhtgplsc
ylksnyqshgiyrdaktafcihnisyqgrfafsdypelnlperfkssfdfidgyekpveg
rkinwmkagileadrvltvspyyaeelisgiargceldnimrltgitgivngmdvsewdp
srdkyiavkydvstaveakalnkealqaevglpvdrniplvafigrleeqkgpdvmaaai
pqlmemvedvqivllgtgkkkfermlmsaeekfpgkvravvkfnaalahhimagadvlav
tsrfepcgliqlqgmrygtpcacastgglvdtiiegktgfhmgrlsvdcnvvepadvkkv
attlqraikvvgtpayeemvrncmiqdlswkgpaknwenvllslgvaggepgvegeeiap
lakenvaap

A2. Granule-bound starch synthase [Oryza *sativa* Indica Group]
GenBank: AAF72562.1
609
msalttsqlatsatgfgiadrsapssllrhgfqglkprspaggdatslsvttsaratpkq
qrsvqrgsrrfpsvvvyatgagmnvvfvgaemapwsktgglgdvlgglppamaanghrvm
visprydqykdawdtsvvaeikvadryervrffhcykrgvdrvfidhpsflekvwgktge
kiygpdtgvdykdnqmrfsllcqaaleaprilnlnnnpyfkgtygedvvfvcndwhtgpl
asylknnyqpngiyrnakvafcihnisyqgrfafedypelnlserfrssfdfidgydtpv
egrkinwmkagileadrvltvspyyaeelisgiargceldnimrltgitgivngmdvsew
dpskdkyitakydattaieakalnkealqaeaglpvdrkipliafigrleeqkgpdvmaa
aipelmqedvqivllgtgkkkfekllksmeekypgkvravvkfnaplahlimagadvlav
psrfepcgliqlqgmrygtpcacastgglvdtviegktgfhmgrlsvdckvvepsdvkkv
aatlkraikvvgtpayeemvrncmnqdlswkgpaknwenvllglgvagsapgiegdeiap
lakenvaap

A3. Granule-bound starch synthase 1, partial [Oryza *sativa* Japonica Group]
GenBank: AGK90263.1
609
msalttsqlatsatgfgiadrsapssllrhgfqglkprspaggdatslsvttsaratpkq
qrsvqrgsrrfpsvvvyatgagmnvvfvgaemapwsktgglgdvlgglppamaanghrvm
visprydqykdawdtsvvaeikvadryervrffhcykhgvdrvfidhpsflekvwgktge
kiygpdtgvdykdnqmrfsllcqaaleaprilnlnnnpyfkgtygedvvfvcndwhtgpl
asylknnyqpngiyrnakvafcihnisyqgrfafedypelnlserfrssfdfidgydtpv
egrkinwmkagileadrvltvspyyaeelisgiargceldnimrltgitgivngmdvsew
dpskdkyitakydattaieakalnkealqaeaglpvdrkipliafigrleeqkgpdvmaa
aipelmqedvqivllgtgkkkfekllksmeekypgkvravvkfnaplahlimagadvlav
psrfepcgliqlqgmrygtpcacastgglvdtviegktgfhmgrlsvdckvvepsdvkkv
aatlkraikvvgtpayeemvrncmnqdlswkgpaknwenvllglgvagsapgiegdeiap
lakenvaap

B1: GS3-like protein [Zea *mays*]
NCBI Reference Sequence: NP_001144472.1216

maaaaaprpksppaspdpcgrhrlqlavdalhreigflegeissiegvhaasrcckevde
fvgrnpdpfltiqqergshdqsqqflkkfrgksclsyylswicgggwwcppplqlkrppa
pscscaprlgklcsstassccscccccrfrvvyaaagcgccapcprcscdctcacprccsc
acpmcxxpxaapraaacaydghekfcvhasssstwr

B2. Seed length and weight protein [Oryza *sativa* Indica Group]
GenBank: BAH89236.1
233
mamaaaprpksppappdpcgrhrlqlavdalhreigflegeinsiegihaasrccrevde
figrtpdpfitissekrshdhshhflkkfrclcrasacclsylswicccssaaggcssss
ssfnlkrpscccncncnccssssssscgaaltkspcrcrrrsccrrcccggvgvracasc
scsppcaccappcagcscrctcpcpcpggcscacpacrcccgvprccppcl

B3. Seed length and weight protein[Oryza *sativa* Japonica Group]
GenBank: BAH89240.1
232
mamaaaprpksppappdpcgrhrlqlavdalhreigflegeinsiegihaasrccrevde
figrtpdpfitissekrshdhshhflkkfrclcrasacclsylswicccssaaggcssss
sssfnlkrpscccncncnccsssssscgaaltkspcrcrrrscccrrcccggvgvracas
cscsppcaccappcagcscrctcpcpcpggcscacpacrcccgvprccppcl

C1. Betaine aldehyde dehydrogenase [Zea *mays*]
NCBI Reference Sequence: NP_001105781.2
506
mmasqamvplrqlfvdgewrppaqgrrlpvvnptteahigeipagtaedvdaavaaaraa
lkrnrgrdwarapgavrakylraiaakvierkqelaklealdcgkpydeaawdmddvagc
feyfadqaealdkrqnspvslpmetfkchlrrepigvvglitpwnypllmatwkvapala
agcaavlkpselasvtcleladickevglppgvlnivtglgpdagaplsahpdvdkvaft
gsfetgkkimaaaapmvkpvtlelggkspivvfddvdidkavewtlfgcfwtngqicsat
srllvhtkiakefnekmvawaknikvsdpleegcrlgpvvsegqyekikkfilnaksega
tiltggvrpahlekgffieptiitdittsmeiwreevfgpvlcvkefstedeaielandt
qyglagavisgdrercqrlseeidagiiwvncsqpcfcqapwggnkrsgfgrelgeggid
nylsvkqvteyisdepwgwyrspskl

C3. Betaine aldehyde dehydrogenase [Oryza *sativa* Japonica Group]
GenBank: ABI84118.1
503
mataipqrqlfvagewrapalgrrlpvvnpatespigeipagtaedvdaavaaarealkr
nrgrdwarapgavrakylraiaakiierkselarletldcgkpldeaawdmddvagcfey
fadlaesldkrqnapvslpmenlkcylrkepigvvglitpwnypllmatwkvapalaagc
tavlkpsesasvtcleladvckevglpsgvlnivtglgseagaplsshpgvdkvaftgsy
etgkkimasaapmvkpvslelggkspivvfddvdvekavewtlfgcfwtngqicsatsrl
ilhkkiakefqermvawaknikvsdpleegcmlgpvvsegqyekikqfvstaksqgatil
tggvrpkhlekgfyieptiitdvdtsmqiwreevfgpvlcvkefsteeeaielandthyg
lagavlsgdrercqrlteeidagiiwvncsqpcfcqapwggnkrsgfgrelgeggidnyl
svkqvteyasdepwgwykspskl

C2. Betaine aldehyde dehydrogenase [Oryza *sativa* Indica Group]
GenBank: ACF06149.1
503
mataipqrqlfvagewrapalgrrlpvvnpatespigeipagtaedvdaavaaarealkr
nrgrdwarapgavrakylraiaakiierkselarletldcgkpldeaawdmddvagcfey
fadlaesldkrqnapvslpmenfkcylrkepigvvglitpwnypllmatwkvapalaagc
tavlkpselasvtcleladvckevglpsgvlnivtglgseagaplsshpgvdkvaftgsy
etgkkimasaapmvkpvslelggkspivvfddvdvekavewtlfgcfwtngqicsatsrl
ilhkkiakefqermvawaknikvsdpleegcrlgpvvsegqyekikqfvstaksqgatil
tggvrpkhlekgfyieptiitdvdtsmqiwreevfgpvlcvkefsteeeaielandthyg
lagavlsgdrercqrlteeidagiiwvncsqpcfcqapwggnkrsgfgrelgeggidnyl
svkqvteyasdepwgwykspskl

Bibliography

[i] Dayhoff, M.O., Schwartz, R. and Orcutt, B.C. (1978). "A model of Evolutionary Change in Proteins". Atlas of protein sequence and structure(volume 5, supplement 3 ed.). Nat. Biomed. Res. Found. pp. 345–358. ISBN 0-912466-07-3.

[ii] Christenhusz, M.J.M.; Byng, J.W. (2016). "The number of known plants species in the world and its annual increase". Phytotaxa. Magnolia Press. 261(3): 201–217. doi:10.11646/phytotaxa.261.3.1. Archivedfrom the original on 2016-07-29.

[iii] "Varieties - Rice Association". www.riceassociation.org.uk. Retrieved 2018-05-06.

[iv] [iv] "The two main types of rice: INDICA RICE and JAPONICA RICE". LEGroup Industries - RICE - TUNA - CHEESE - JAMS. 2017-10-09. Retrieved 2018-05-06.

[v] Lindeboom, N. , Chang, P. R., Tyler, R. T. and Chibbar, R. N. (2005), Granule-Bound Starch Synthase I (GBSSI) in Quinoa (Chenopodium quinoa Willd.) and Its Relationship to Amylose Content. Cereal Chemistry, 82: 246-250. doi:10.1094/CC-82-0246

[vi] Xian-Zhong Han, Bruce R. Hamaker, Amylopectin Fine Structure and Rice Starch Paste Breakdown, Journal of Cereal Science, Volume 34, Issue 3,2001, Pages 279-284,

[vii] Hailiang Mao, Shengyuan Sun, Jialing Yao, Chongrong Wang, Sibin Yu, Caiguo Xu, Xianghua Li, Qifa Zhang Linking differential domain functions of the GS3 protein to natural variation of grain size in rice Proceedings of the National Academy of Sciences Nov 2010, 107 (45) 19579-19584; DOI:10.1073/pnas.1014419107

[viii] Singh, Anuradha, Pradeep K. Singh, Rakesh Singh, Awadhesh Pandit, Ajay K. Mahato, Deepak K. Gupta, Kuldeep Tyagi, Ashok K. Singh, Nagendra K. Singh, and Tilak R. Sharma. “SNP Haplotypes of the BADH1 Gene and Their Association with Aroma in Rice (Oryza *Sativa* L.).” Molecular Breeding 26, no. 2 (August 1, 2010): 325–38. https://doi.org/10.1007/s11032-010-9425-1.