Group 6:

Tiffany Cook

Drashti Mehta

Joshua Mikombo

BINF 6201 – Lab 8

Part 1: Build your own Sequence Profile

Question 1: Using the short training data set, construct a Position Specific Scoring Matrix (i.e., a profile). Provide the complete matrix of scores for your completed profile.

AA	1	2	3	4	5	6	7
L	**1.87**	-0.53	-0.53	-0.53	-0.53	2.647	2.87
I	**0.163**	-0.53	-0.53	-0.53	**0.568**	-0.53	-0.53
M	**0.568**	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53
Y	**0.163**	-0.53	-0.53	-0.53	1.95	-0.53	-0.53
P	-0.53	2.177	-0.53	-0.53	-0.53	-0.53	-0.53
E	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53
Q	-0.53	-0.53	0.1625	-0.53	-0.53	-0.53	-0.53
F	-0.53	-0.53	-0.53	1.95	2.177	-0.53	2.3597
A	-0.53	-0.53	-0.53	**0.568**	-0.53	-0.53	-0.53
N	-0.53	-0.53	-0.53	0.1625	-0.53	-0.53	**0.568**
S	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53	0.1625
T	-0.53	-0.53	-0.53	-0.53	-0.53	0.1625	-0.53
R	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53	**0.568**
K	-0.53	-0.53	-0.53	-0.53	-0.53	-0.53	0.1625

Question 2: Test the profile you just constructed with this known S-locus motif from maize *[1 pt]*:

L S L I Y H V

What is your score for this sequence? Does your PSSM correctly identify this as belonging to the S-glycoprotein family?

1.87 - 0.53 - 0.53 - 0.53 + 1.95 + 0 + 0 = **2.23**

And yes, it DOES belong to the S-glycoprotein family

Question 3: Test the profile with this iron transporter motif (that we know is NOT an S-locus protein). *[1 pt]:*

F T D E L M E

What is your score for this sequence? Does your PSSM correctly identify this as not being a member of the family?

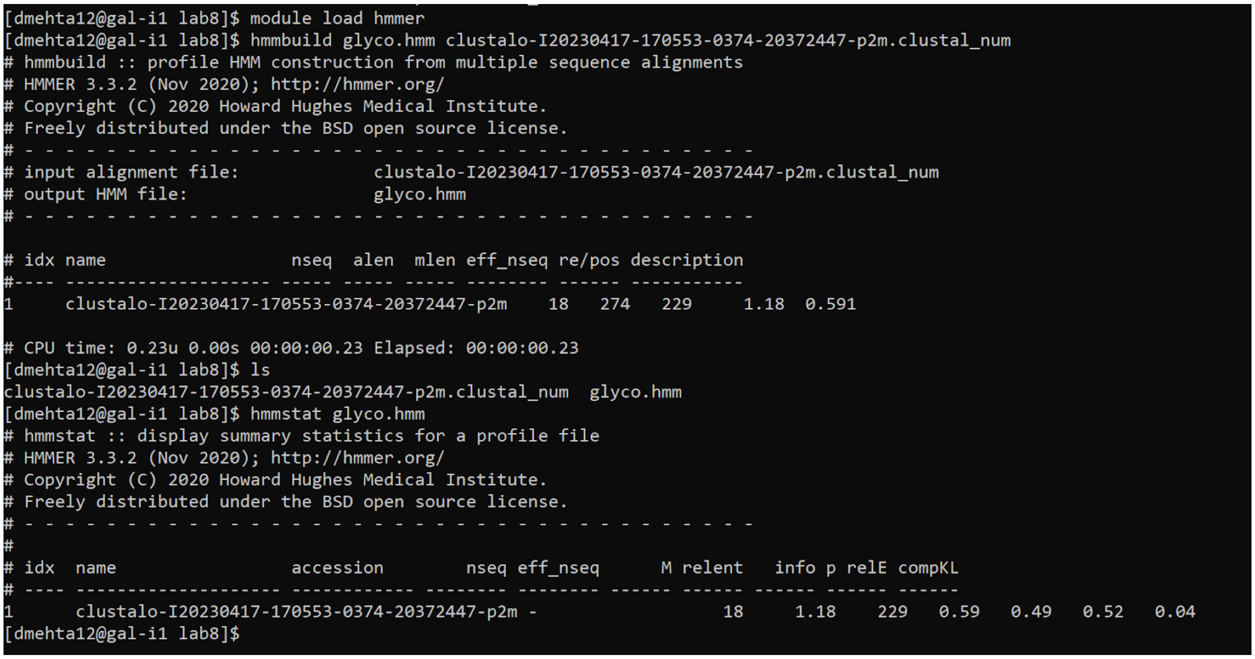
-0.53 - 0.53 + 0 - 0.53 - 0.53 - 0.53 - 0.53 = **-3.15**

Yes, it does correctly identify this as NOT being a member of the family.

Part 2: Use HMMer to build an HMM profile

Question 4: Use the data set of full S-locus glycoprotein sequences to build an HMM.

Look at the summary HMMer outputs to the screen (or use `hmmstat` on your output file). How many match states did HMMer use? Is the length of the HMM the same as the length of the alignment?



```
[dmehta12@gal-i1 lab8]$ module load hmmer
[dmehta12@gal-i1 lab8]$ hmmbuild glyco.hmm clustalo-I20230417-170553-0374-20372447-p2m.clustal_num
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# input alignment file:             clustalo-I20230417-170553-0374-20372447-p2m.clustal_num
# output HMM file:                  glyco.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                  nseq  alen  mlen eff_nseq re/pos description
#---- -------------------- ----- ----- ----- -------- ------ -----------
1     clustalo-I20230417-170553-0374-20372447-p2m    18   274   229     1.18  0.591

# CPU time: 0.23u 0.00s 00:00:00.23 Elapsed: 00:00:00.23
[dmehta12@gal-i1 lab8]$ ls
clustalo-I20230417-170553-0374-20372447-p2m.clustal_num  glyco.hmm
[dmehta12@gal-i1 lab8]$ hmmstat glyco.hmm
# hmmstat :: display summary statistics for a profile file
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
#
# idx  name                  accession        nseq eff_nseq      M relent   info p relE compKL
# ---- -------------------- ------------ -------- -------- ------ ------ ------ ------ ------
1      clustalo-I20230417-170553-0374-20372447-p2m -                 18     1.18    229   0.59   0.49   0.52   0.04
[dmehta12@gal-i1 lab8]$
```

274 and 229 > not same

Question 5: Search the pineapple protein sequences using the profile to find S-locus family members.

HMMer will print a lot of information to the screen, but more usefully it will create a Blast-like output file with the most significant results. When the search finishes, look at this output file to find potential matches.

How many significantly scoring proteins were found in the pineapple data set? Which was the most significant?

It is showing 34 significantly scoring proteins, with the third protein being of the best score as illustrated in the screenshot (circled in red).