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 *[5 pts]*

| Pos | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| L | 1.867267 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | 2.108429 | 1.415282 |
| P | -0.53063 | 2.177422 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |
| I | 0.162519 | -0.53063 | -0.53063 | -0.53063 | 0.162519 | -0.53063 | -0.53063 |
| M | 0.567984 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |
| Y | 0.162519 | -0.53063 | -0.53063 | -0.53063 | 1.867267 | -0.53063 | -0.53063 |
| E | -0.53063 | -0.53063 | 2.108429 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |
| Q | -0.53063 | -0.53063 | 0.162519 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |
| A | -0.53063 | -0.53063 | -0.53063 | 0.567984 | -0.53063 | -0.53063 | -0.53063 |
| F | -0.53063 | -0.53063 | -0.53063 | 1.954278 | 0.855666 | -0.53063 | 0.855666 |
| N | -0.53063 | -0.53063 | -0.53063 | 0.162519 | -0.53063 | -0.53063 | 0.162519 |
| K | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | 0.162519 |
| T | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | 0.162519 | -0.53063 |
| R | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | 0.567984 |
| S | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | 0.162519 |
| C | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |
| D | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |
| G | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |
| H | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |
| V | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |
| W | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 | -0.53063 |

#### 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?

| L | 1.867267 |
| --- | --- |
| S | -0.53063 |
| L | -0.53063 |
| I | -0.53063 |
| Y | 1.867267 |
| H | -0.53063 |
| V | -0.53063 |
| Total Score | 1.081393 |

1.081393, Our PSSM correctly identifies as belonging 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 = -0.5306 | T = -0.5306 | D = -0.5306 | E = -0.5306 | L = -0.5306 | M = -0.5306 | E = -0.5306 |
| --- | --- | --- | --- | --- | --- | --- |

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

-3.7144

The PSSM correctly identifies the iron transporter as NOT part of the S-locus motif.

### 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. *[3 pts]*

Align the sequences with Clustal-Omega and upload them to the cluster. To build an HMM with HMMer:

module load hmmer

hmmbuild OUTPUTNAME.hmm YOURALIGNMENT\_FILE

Look at the summary HMMer outputs to the screen (or use hmmstat on your output file). How many match states did HMMer use?

Hmmer used 229 match states.

Is the length of the HMM the same as the length of the alignment?

The length of the HMM is not the same as the alignment.

#### Question 5: Search the pineapple protein sequences using the profile to find S-locus family members *[5 pts].*

To get the HMMer profile ready to be searched, first run:

hmmpress YOUR-HMMER-FILE

Then, upload the pineapple sequences to the cluster, and run:

hmmscan --tblout OUTPUTNAME YOUR-HMMER-FILE pineapple.faa

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?

There are 34 statistically significant proteins found in the pineapple data set, protein ACO009110.1 being the most significant.