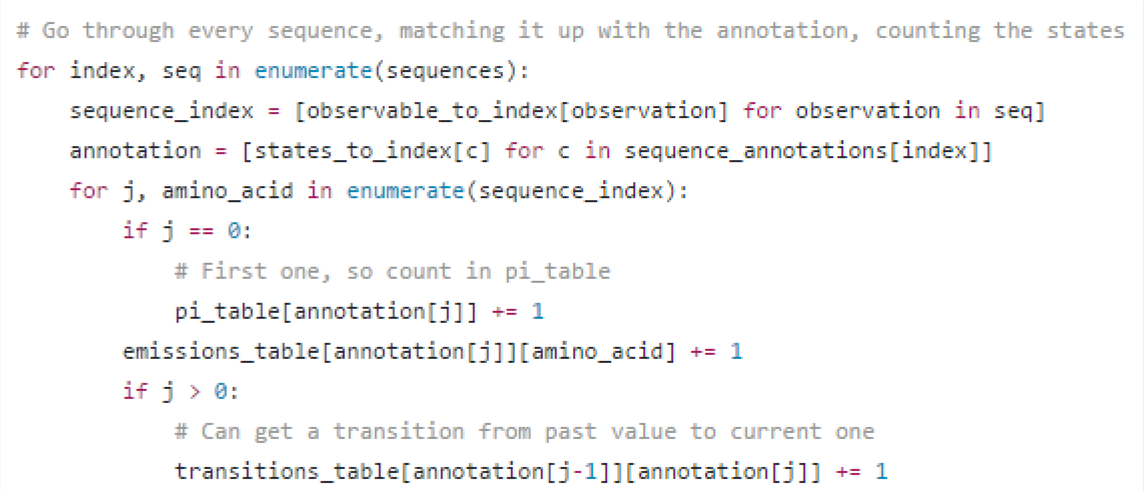
Using a hidden Markov model for prediction of transmembrane helices

Jacob Bieker & Kim Pham Nguyen

Introduction

- Given: Dataset consisting of 160 membrane proteins annotated with their transmembrane helices
- Assignment: Train a 3 state and 4 state model by training parts 0-8 dataset using ”training-by-counting”
- Assignment: Use 3 and 4 state model to make a 10-fold-experiment, training-by-counting and viterbi decoding for prediction.

Train the 3-state model



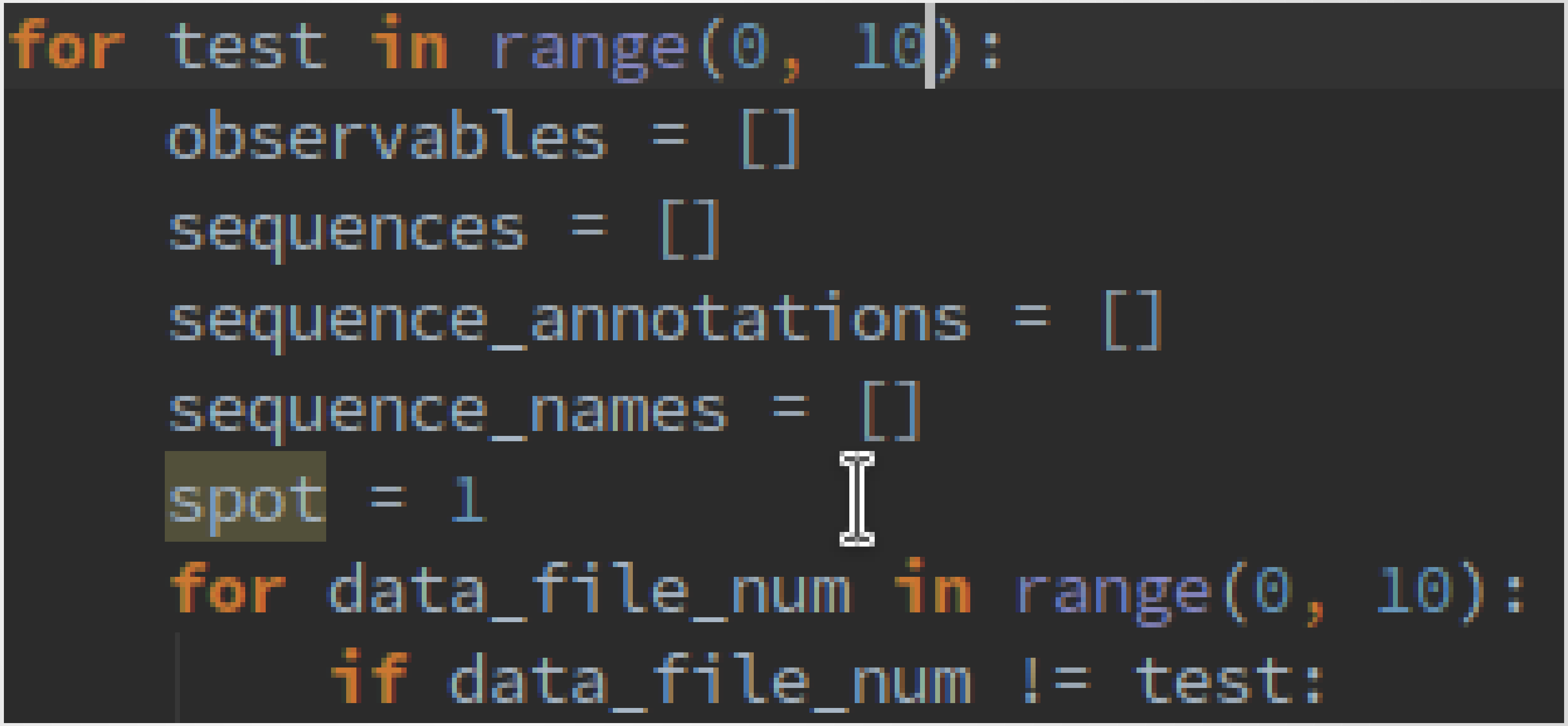
```python
# Go through every sequence, matching it up with the annotation, counting the states
for index, seq in enumerate(sequences):
    sequence_index = [observable_to_index[observation] for observation in seq]
    annotation = [states_to_index[c] for c in sequence_annotations[index]]
    for j, amino_acid in enumerate(sequence_index):
        if j == 0:
            # First one, so count in pi_table
            pi_table[annotation[j]] += 1
        emissions_table[annotation[j]][amino_acid] += 1
        if j > 0:
            # Can get a transition from past value to current one
            transitions_table[annotation[j-1]][annotation[j]] += 1
```

Train 4-state model

- Preprocessing annotation
    - from ”o” to ”M” = ”m”
    - from ”i” to ”M” = ”M”

- Go through every sequence, matching it up with the annotation and counting the states (just as 3 state model)

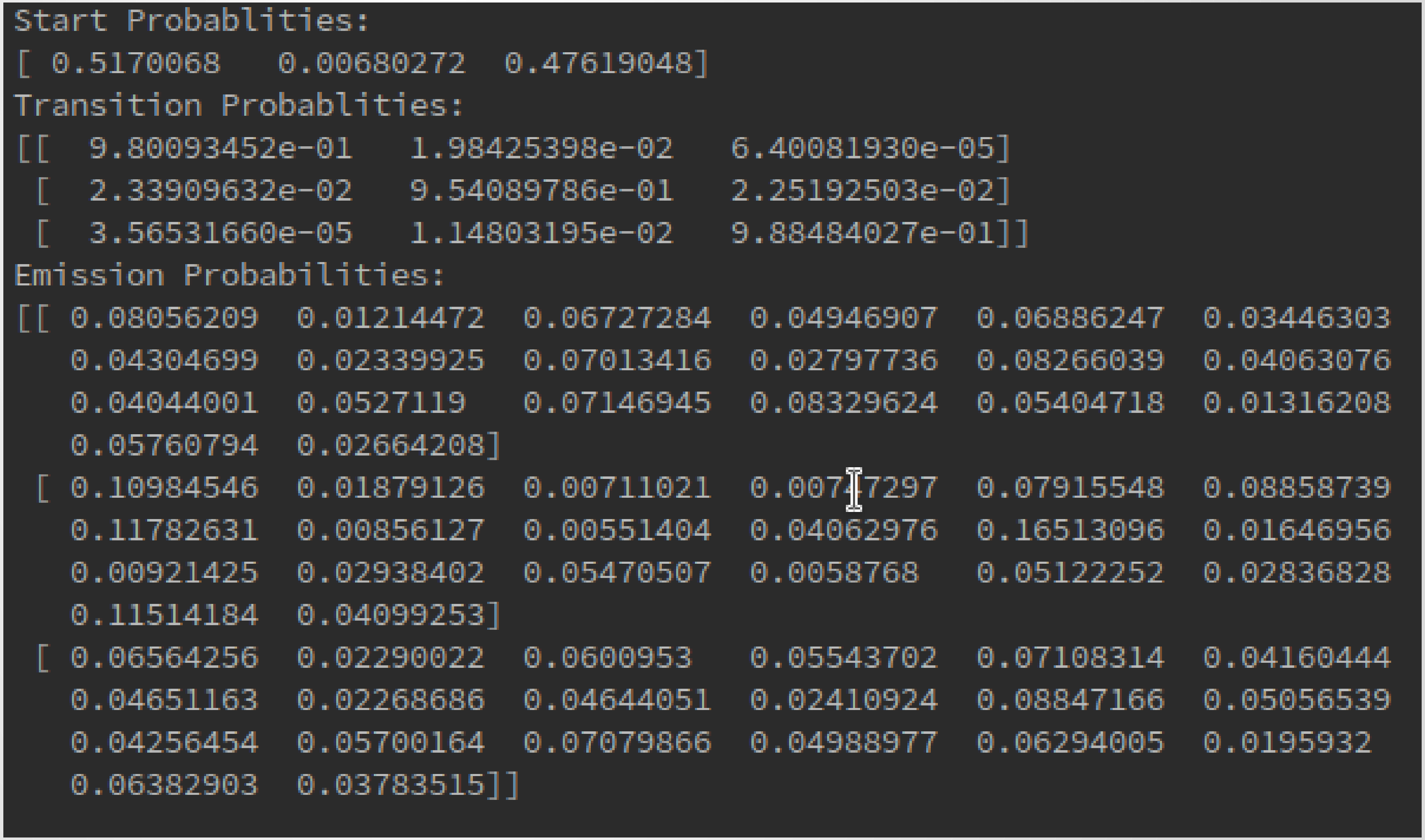
10-fold experiment



```
for test in range(0, 10):
    observables = []
    sequences = []
    sequence_annotations = []
    sequence_names = []
    spot = 1
    for data_file_num in range(0, 10):
        if data_file_num != test:
```

Training-by-counting

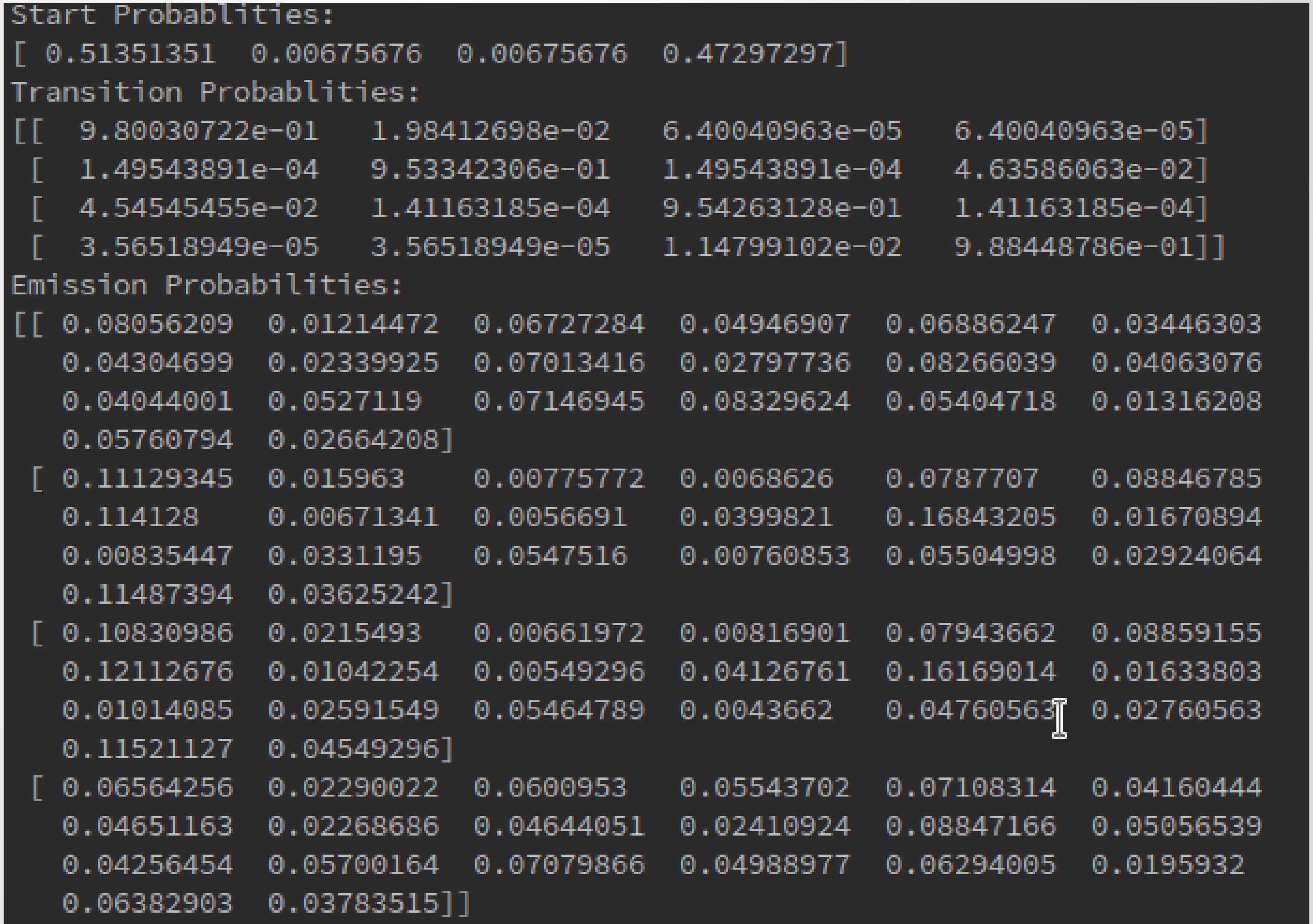
- 3-state model, for parts 0-8:



```
Start Probablities:
[ 0.5170068   0.00680272  0.47619048]
Transition Probablities:
[[  9.80093452e-01   1.98425398e-02   6.40081930e-05]
 [  2.33909632e-02   9.54089786e-01   2.25192503e-02]
 [  3.56531660e-05   1.14803195e-02   9.88484027e-01]]
Emission Probabilities:
[[ 0.08056209  0.01214472  0.06727284  0.04946907  0.06886247  0.03446303
   0.04304699  0.02339925  0.07013416  0.02797736  0.08266039  0.04063076
   0.04044001  0.0527119   0.07146945  0.08329624  0.05404718  0.01316208
   0.05760794  0.02664208]
 [ 0.10984546  0.01879126  0.00711021  0.00747297  0.07915548  0.08858739
   0.11782631  0.00856127  0.00551404  0.04062976  0.16513096  0.01646956
   0.00921425  0.02938402  0.05470507  0.0058768   0.05122252  0.02836828
   0.11514184  0.04099253]
 [ 0.06564256  0.02290022  0.0600953   0.05543702  0.07108314  0.04160444
   0.04651163  0.02268686  0.04644051  0.02410924  0.08847166  0.05056539
   0.04256454  0.05700164  0.07079866  0.04988977  0.06294005  0.0195932
   0.06382903  0.03783515]]
```

Training-by-counting

- 4-state model, for parts 0-8:



```
Start Probablities:
[ 0.51351351  0.00675676  0.00675676  0.47297297]
Transition Probablities:
[[  9.80030722e-01   1.98412698e-02   6.40040963e-05   6.40040963e-05]
 [  1.49543891e-04   9.53342306e-01   1.49543891e-04   4.63586063e-02]
 [  4.54545455e-02   1.41163185e-04   9.54263128e-01   1.41163185e-04]
 [  3.56518949e-05   3.56518949e-05   1.14799102e-02   9.88448786e-01]]
Emission Probabilities:
[[ 0.08056209  0.01214472  0.06727284  0.04946907  0.06886247  0.03446303
   0.04304699  0.02339925  0.07013416  0.02797736  0.08266039  0.04063076
   0.04044001  0.0527119   0.07146945  0.08329624  0.05404718  0.01316208
   0.05760794  0.02664208]
 [ 0.11129345  0.015963    0.00775772  0.0068626   0.0787707   0.08846785
   0.114128    0.00671341  0.0056691   0.0399821   0.16843205  0.01670894
   0.00835447  0.0331195   0.0547516   0.00760853  0.05504998  0.02924064
   0.11487394  0.03625242]
 [ 0.10830986  0.0215493   0.00661972  0.00816901  0.07943662  0.08859155
   0.12112676  0.01042254  0.00549296  0.04126761  0.16169014  0.01633803
   0.01014085  0.02591549  0.05464789  0.0043662   0.04760563  0.02760563
   0.11521127  0.04549296]
 [ 0.06564256  0.02290022  0.0600953   0.05543702  0.07108314  0.04160444
   0.04651163  0.02268686  0.04644051  0.02410924  0.08847166  0.05056539
   0.04256454  0.05700164  0.07079866  0.04988977  0.06294005  0.0195932
   0.06382903  0.03783515]]
```

3-State Viterbi Results

- Over all 10 folds:

Variance: 0.00442402539894

Mean: 0.676804974853

- Individual ACs:

0.630622152503856

0.758868567899825

0.651262692466251 5

0.570194775552633 8

0.710101154509694 7

0.713966925168569

0.700230173603168 9

0.654556299001142 6

0.590335997464210 8

0.787911010358954 6

4-State Viterbi Results

- Over all 10 folds:

Variance:0.00468726047895

Mean: 0.680697300092

- Individual ACs:

0.5825665066458083
0.7400915818648046
0.643551590962399
0.6066165791228548
0.7359067133402921
0.7177319941203848
0.7404435898416963
0.6368863895360404
0.6119622374686893
0.7912158180171618