Matus Halak : 2724858

Open Questions Hidden Markov Models Assignment

Question 1:

In the A2_prior we ensured that Links only occur between domains by only allowing transition from begin to domain state, as well as, only allowing termination from domain to end state. In E2_prior e ensured that hydrophobic residues occur less frequently in linkers by decreasing emission probability of H from L state and keeping it low relative to the other emission probabilities from L. For domains, we increased the probability of H emission and kept it high relative to emission probabilities of other symbols. For E3 & E4 we modified E2_prior according to instructions.

The tables show the posteriors for the transition and emission probability matrices obtained via Baum Welch Training. All training runs were run with the default 0.01 SLL convergence criterion for a maximum of 1000 iterations.

Overall, we notice that all training runs using the A2 prior converge to the same HMM, regardless of the Emission Probability Prior, which only seems to influence the speed of convergence. The biologically, informed E2 prior leads to the fastest convergence (44 iterations), followed by E4 which gives both states the same starting priors (93 iterations) and lastly E3, which contains the flipped values of the biologically informed E2, and therefore represents a 'nonsense' prior that the algorithm needs to somewhat fight against (E3 is the only case where > 100 iterations were needed). Importantly we see that using a different transition probability prior A1 that allows starting / ending a sequence with a linker region converges to a completely different HMM (79 iterations). This is because the A2 prior has the strict requirement that Linkers ONLY occur between domains and can never start or end the sequence, doesn't operate in the same search space as A1, which might identify biologically implausible explanations to explain the data better.

A1_prior:

	B	L	D	E
B	0	0.5	0.5	0
L	0	0.7	0.2	0.1
D	0	0.2	0.7	0.1
E	0	0	0	0

E1_prior:

	H	P	C
L	0.5	0	0.5
D	0	0.5	0.5

```
Converged after 79 iterations.
Final SLL: -2.84e+03
Final parameters:

[A]   B     D     L     E
    B 0.000 0.245 0.755 0.000
    D 0.000 0.476 0.499 0.025
    L 0.000 0.334 0.626 0.040
    E 0.000 0.000 0.000 0.000

[E]   C     H     P
    D 0.485 0.000 0.515
    L 0.043 0.957 0.000
```

A2_prior

	B	L	D	E
B	0	0	1	0
L	0	0.7	0.3	0
D	0	0.3	0.6	0.1
E	0	0	0	0

```
Converged after 44 iterations.
Final SLL: -2.83e+03
Final parameters:

[A]   B     D     L     E
    B 0.000 1.000 0.000 0.000
    D 0.000 0.833 0.121 0.046
    L 0.000 0.337 0.663 0.000
    E 0.000 0.000 0.000 0.000

[E]   C     H     P
    D 0.141 0.717 0.142
    L 0.435 0.183 0.382
```

E2_prior

	H	P	C
L	0.1	0.5	0.4
D	0.5	0.25	0.25

```
Converged after 109 iterations.
Final SLL: -2.83e+03
Final parameters:

[A]   B     D     L     E
    B 0.000 1.000 0.000 0.000
    D 0.000 0.835 0.119 0.047
    L 0.000 0.317 0.683 0.000
    E 0.000 0.000 0.000 0.000

[E]   C     H     P
    D 0.142 0.716 0.142
    L 0.424 0.201 0.374
```

E3_prior

	H	P	C
L	0.5	0.25	0.25
D	0.1	0.5	0.4

```
Converged after 93 iterations.
Final SLL: -2.83e+03
Final parameters:

[A]   B     D     L     E
    B 0.000 1.000 0.000 0.000
    D 0.000 0.835 0.119 0.047
    L 0.000 0.318 0.682 0.000
    E 0.000 0.000 0.000 0.000

[E]   C     H     P
    D 0.142 0.716 0.142
    L 0.425 0.200 0.375
```

E4_prior

	H	P	C
L	0.1	0.45	0.45
D	0.1	0.45	0.45

Question 2:

Imagine you were able to find structural information for a small subset of your protein sequences.
Considering the results of question 1 and the influence that the priors have on the training performance, how would you modify your workflow in order to improve the HMM training?

This structural data would allow us to refine our A & E priors (by checking in our structural data how often linker regions occur within / outside 2 domain regions and what AAs are enriched in linkers vs domains) and make them even more biologically informed than A2, E2, which should lead to the quickest convergence, by starting from an informed point in the search space and help us get to the most biologically accurate HMM.