Algorithms in Sequence Analysis
HMM Open Questions

Author: Finn van Vlaandern

Question 1

The posterior distributions acquired by running Baum-Welch with priors $(A_1, E_1), (A_2, E_2), (A_2, E_3), (A_2, E_4)$ are listed under ”Posterior distributions for A and E under different priors”. All simulations were performed with a convergence threshold of 0.01 and iteration count of 1000. Before discussing these results, let's discuss how we chose our prior E_2 using biological information. If hydrophobic residues are more common in domains and less abundant in linker regions, the emission probability of H should be higher given D than given L. Furthermore, if polar and charged amino acids are more common in linker regions than in domains, the emission probabilities of P and C should be higher given L than given D. We therefore chose E_2 as

	C	H	P
D	0.1	0.45	0.45
L	0.5	0.25	0.25

Table 1: *Prior E2*

If linker regions exclusively appear between two domains, we should modify the transition matrix A_1 so that the begin state can only enter into a domain and the end state can only be entered from a domain. Thus, we chose A_2 as

	B	D	L	E
B	0	1	0	0
D	0	0.75	0.25	0
L	0	0.7	0.2	0.1
E	0	0	0	0

Table 2: *Prior A2*

We observe that the priors are crucial in determining both the local optimum converged to and the speed of convergence. Priors $(A_2, E_2), (A_2, E_2), (A_2, E_4)$ converge to similar optima with varying convergence rates. Among these, (A_2, E_2) coheres most with the biological information, and training using this prior converges the fastest. This can be explained by noting that the training sequences are biological sequences, and that using biologically relevant priors gives the model an inductive bias towards learning biological sequences. The prior (A_1, E_1) gives a different optimum from the other priors. This difference can be explained by the differing pattern of zeroes between this prior and the others, making it so that the algorithm is restricted to a subspace of the parameter space different from the subspace explored using the other priors.

Question 2

In question 1, we saw that the prior is important for the achieved optimum and the speed of convergence, and that the incorporation of biological prior information can significantly speed up convergence. Given that we have access to structural information for a small subset of our protein sequences, we would use this structural information to identify which parts of our sequences in the subset are in linker regions and which are in domains, and subsequently estimate priors by maximum likelihood, which comes down to determining empirical frequencies given the structural information.

Posterior distributions for A and E under different priors

Priors A1 and E1

	B	D	L	E
B	0	2.45e-01	7.55e-01	0
D	0	4.76e-01	4.99e-01	2.49e-02
L	0	3.34e-01	6.26e-01	4.02e-02
E	0	0	0	0

Table 3: *Posterior distribution for A (Priors A1, E1)*

	C	H	P
D	4.85e-01	0	5.15e-01
L	4.26e-02	9.57e-01	0

Table 4: *Posterior distribution for E (Priors A1, E1)*

SLL: -2.84×10^3, **convergence:** 79 iterations.

Priors A2 and E2

	B	D	L	E
B	0	1.00e+00	0	0
D	0	8.36e-01	1.18e-01	4.58e-02
L	0	3.42e-01	6.58e-01	0
E	0	0	0	0

Table 5: *Posterior distribution for A (Priors A2, E2)*

	C	H	P
D	1.43e-01	7.14e-01	1.43e-01
L	4.39e-01	1.74e-01	3.87e-01

Table 6: *Posterior distribution for E (Priors A2, E2)*

SLL: -2.83×10^3, **convergence:** 10 iterations.

Priors A2 and E3

	B	D	L	E
B	0	1.00e+00	0	0
D	0	8.35e-01	1.18e-01	4.68e-02
L	0	3.17e-01	6.83e-01	0
E	0	0	0	0

Table 7: *Posterior distribution for A (Priors A2, E3)*

	C	H	P
D	1.42e-01	7.15e-01	1.42e-01
L	4.25e-01	2.01e-01	3.74e-01

Table 8: *Posterior distribution for E (Priors A2, E3)*

SLL: -2.83×10^3, **convergence:** 103 iterations.

Priors A2 and E4

	B	D	L	E
B	0	1.00e+00	0	0
D	0	8.35e-01	1.18e-01	4.68e-02
L	0	3.16e-01	6.84e-01	0
E	0	0	0	0

Table 9: *Posterior distribution for A (Priors A2, E4)*

	C	H	P
D	1.42e-01	7.15e-01	1.42e-01
L	4.24e-01	2.02e-01	3.74e-01

Table 10: *Posterior distribution for E (Priors A2, E4)*

SLL: -2.83×10^3, **convergence:** 82 iterations.