**Questions:**

1. Read in the supplied small file test.fasta (example from “Biological Sequence Analysis”). What is the number of match states needed?

There are 8 match states

2. Train the HMM, i.e. the emission and transition probabilities. For emissions in the insert states, use the values supplied in pa. Print the estimated probabilities. Explain whether they make sense.

**mm** (1.0, 0.8571428571428571, 0.8571428571428571, 0.8571428571428571, 0.8571428571428571, 0.8571428571428571, 0.8571428571428571, 1.0, 1.0)

**mi** (0.0, 0.0, 0.0, 0.14285714285714285, 0.0, 0.0, 0.0, 0.0, 0.0)

**md** (0.0, 0.14285714285714285, 0.14285714285714285, 0.0, 0.14285714285714285, 0.14285714285714285, 0.14285714285714285, 0.0, 0.0)

**im** (0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.0)

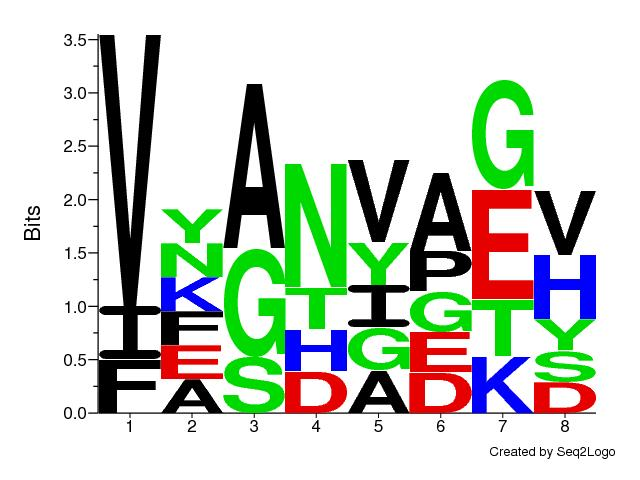
**ii** (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)

**dm** (1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0)

**dd** (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)

3. Print the emission probability matrix in a format demanded by the SeqLogo server (http://www.cbs.dtu.dk/biotools/Seq2Logo-2.0/), e.g. the PSSM

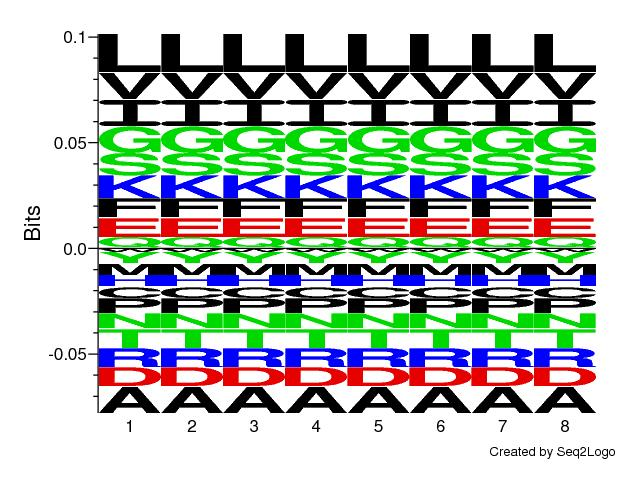
format (http://www.cbs.dtu.dk/biotools/Seq2Logo-2.0/bin/weight.txt), and use SeqLogo to generate a logo. Put both the matrix and logo in your answers.



4. List at least 10 random sequences generated using your HMM.

|  |  |
| --- | --- |
| 1 | VAANIDED |
| 2 | VEDAPH |
| 3 | VAAHAGH |
| 4 | VFSHVAEV |
| 5 | FYAGGAKD |
| 6 | INADGAGY |
| 7 | VGNETV |
| 8 | FEDVDV |
| 9 | FKGHVAS |
| 1- | VYNAAS |

**5. Change from using the estimated emission probabilities (EQ6.25) in your code to the background frequencies (EQ6.26) and repeat steps 2-4. How do the results change?**



The logo is the same in every position obviously. The sampled sequences are also different due to different background distributions. Perhaps I am misunderstanding this question but it seems trivial that for different frequencies the results will be different.

|  |  |
| --- | --- |
| 1 | VFGNIPEH |
| 2 | VGNAAKS |
| 3 | FYADVATY |
| 4 | VFANVPGH |
| 5 | VNGNPTH |
| 6 | VANVES |
| 7 | VEADYAGH |
| 8 | FFGWGGGY |
| 9 | IYANAGV |
| 10 | IKAHVAGH |

**6. Read in the supplied large file test\_large.fasta and repeat steps 2-5. Report the emission and transition probabilities, the logo (from SeqLogo), and the sampled sequences.**

There are 43 match states

**mm** (1.0, 1.0, 1.0, 0.9067796610169492, 1.0, 0.9915254237288136, 0.9915254237288136, 0.9915254237288136, 0.9915254237288136, 0.9915254237288136, 0.7203389830508474, 0.9152542372881356, 0.940677966101695, 0.940677966101695, 0.8813559322033898, 1.0, 1.0, 1.0, 1.0, 1.0, 0.9915254237288136, 1.0, 1.0, 0.9915254237288136, 1.0, 1.0, 0.9915254237288136, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.9915254237288136, 0.7966101694915254, 0.8813559322033898, 0.9745762711864406, 1.0, 1.0, 1.0, 0.9915254237288136, 1.0, 1.0, 1.0)

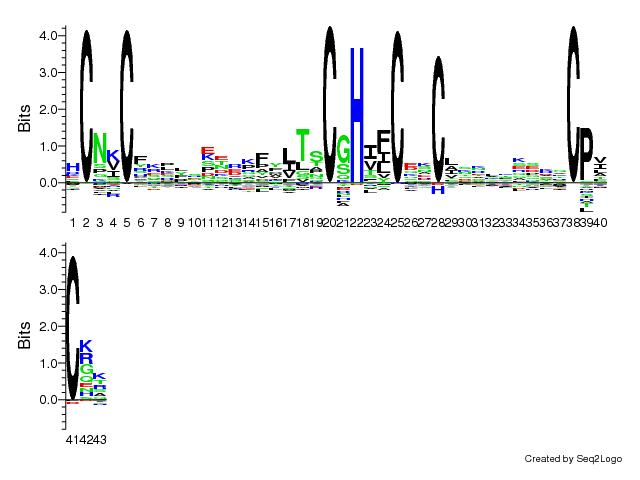
**mi** (0.0, 0.0, 0.0, 0.09322033898305085, 0.0, 0.0, 0.0, 0.0, 0.0, 0.00847457627118644, 0.0, 0.0, 0.0, 0.0, 0.11864406779661017, 0.0, 0.0, 0.0, 0.0, 0.0, 0.00847457627118644, 0.0, 0.0, 0.00847457627118644, 0.0, 0.0, 0.00847457627118644, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.00847457627118644, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.00847457627118644, 0.0, 0.0, 0.0)

**md** (0.0, 0.0, 0.0, 0.0, 0.0, 0.00847457627118644, 0.00847457627118644, 0.00847457627118644, 0.00847457627118644, 0.0, 0.2796610169491525, 0.0847457627118644, 0.059322033898305086, 0.059322033898305086, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.2033898305084746, 0.11864406779661017, 0.025423728813559324, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)

**im** (0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.0)

**ii** (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)

**dm** (1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0)

**dd** (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)

**7. Pick a randomly generated sequence and search for it in the PFAM database (http://pfam.xfam.org/). To what family does the sequence belong?**

Seq used: HCPNCLCMLRASHMPFLTSCFHVTCQECFEKRYSSGCPACGQ

family found: zinc-RING finger domain.

This makes sense as the original sequences also seem to be zinc fingers.

**8. What is the time complexity of estimating the emission and transition probabilities?**

If M is the length of a sequence I expect something like O( M \* 3 ) as for each position there are at least 3 paths possible each time. But I figured the number of sequences should be taken into account as well, only I am not sure how.

**9. Which part of your code is deterministic and which part is randomized?**

Deterministic:

- finding match states

- Calculating probabilities

Randomized:

- Sampling a sequence