

Team assignment, Due: December 5 midnight

- (30 points) Consider the 13 nucleotide sequences on slide 12 of “motif finding” lecture notes. There are two motif regions and a background region. Build a motif model (M, B) for the second motif region of size 6 using EM. Assess how well it is able to recognize the motif by examining M, B, and the hidden variables upon convergence.

```
TCTCAACGTAACACTTTACAGCGGCG__CGTCATTTGATATGATGC_GCCCCGCTTCCCGATAAGGG
GATCAAAAAAATACTTGTGCAAAAAA__TTGGGATCCCTATAATGCGCCTCCGTTGAGACGACAACG
ATGCATTTTCCGCTTGTCTTCCTGA__GCCGACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
CCTGAAATTCAGGGTTGACTCTGAAA__GAGGAAAGCGTAATATAC_GCCACCTCGCGACAGTGAGC
CTGCAATTTTTCTATTGCGGCCTGCG__GAGAACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
TTTTAAATTTCTCTTGTGTCAGGCCGG__AATAACTCCCTATAATGCGCCACCACCTGACACGGGAACAA
GCAAAAATAAATGCTTGACTCTGTAG__CGGGAAGGCGTATTATGC_ACACCCCGCGCCGCTGAGAA
TAACACCGTGCGTGTGACTATTTTA_CCTCTGGCGGTGATAATGG__TTGCATGTACTAAGGAGGT
TATCTCTGGCGGTGTTGACATAAATA_CCACTGGCGGTGATACTGA__GCACATCAGCAGGACGCAC
GTGAAACAAAACGGTTGACAACATGA_AGTAAACACGGTACGATGT_ACCACATGAAACGACAGTGA
TATCAAAAAGAGTATTGACTTAAAGT_CTAACCTATAGGATACTTA_CAGCCATCGAGAGGGACACG
ACGAAAAACAGGTATTGACAACATGAAGTAACATGCAGTAAGATAC_AAATCGCTAGGTAACACTAG
GATACAAATCTCCGTTGTACTTTGTT__TCGCGCTTGGTATAATCG_CTGGGGGTCAAAGATGAGTG
```

- (20 points) Repeat the above problem using Gibbs sampling. Compare the quality of results with those obtained using EM. Also, compare the running times.
- (20 points) Consider the first and the fifth helix blocks from slide 3 of “multiple-alignment”. Build a profile HMM using the first six sequences. Now, consider substrings from the seventh sequence and estimate the probability that it is a part of the two HMM models.

- First block


```
PEEKSAVTALWGK
GEEKAAVLALWDK
PADKTNVKAAWGK
AADKTNVKAAWSK
AAEKTKIRSAWAP
EGEWQLVLHVWAK
ESQAALVKSSWEE
```
- Fifth block


```
KGTFAT - - LSELHCDKLHVD
KGTFAA - - LSELHCDKLHVD
PNALSA - - LSDLHAHKLRVD
PGALSN - - LSDLHAHKLRVD
__MSSMKDLSGKHAKSFEVD
```

EAELKP - - LAQSHATKHKIP
 __ DATLKNLGSVHVSKGVVA

Note that “-“ represents a deletion (in a sequence with respect to the profile HMM) and “_” represents an insertion (in a sequence with respect to the profile HMM).

4. (30 points) Again consider the helix dataset. For the first and the fifth blocks,
 1. Find the Steiner string and the Center string of the sequences. For Steiner string, you will need to use exhaustive enumeration.
 2. What is the approximation ratio achieved by the Center string using edit distance as the metric?
 3. Compute the optimal SP score using dynamic programming. How does this score compare with that achieved by the Center string?

5. (15 points) Suppose you are given the following two distance matrices: D1 and D2, over 5 taxa.

Matrix D1:

	A	B	C	D	E
A		3	9	10	7
B			8	9	6
C				3	6
D					7
E					

Matrix D2:

	A	B	C	D	E
A		4	8	12	5
B			6	8	4
C				3	6
D					8
E					

For each matrix,

1. Check if it is ultrametric.
 2. Check if it is additive.
 - a. If it is additive, build the phylogenetic tree using Neighborhood Joining.
 - b. If not, run UPGMA and NJ algorithms presented in class and compute the error using the least square error metric.
6. (10 points) Consider the transition matrix Q of an aperiodic and irreducible Markov Chain. Let u be its stationary vector. Consider the “reversed” matrix R as defined by $R(i,j) = Q(j,i)u(i) / u(j)$. Show the following.

- R defines a Markov Chain.
- The stationary vector of R is also u.
- When Q reaches its stationary state, the redistribution of “mass” from state i to state j in a single step is given by $u_i Q(j,i)$. If we go backwards, then the redistribution of “mass” from state j to state i in a single step is given by $u_j Q(i,j)$. Argue that if the condition of detailed balance is satisfied then it is not possible to distinguish between the system evolving forwards or backwards.