COMS4761: FALL 2007

ASSIGNMENT 4.4 SOLUTIONS

Problem 1: Using Profile HMM to analyze ezRNA structure

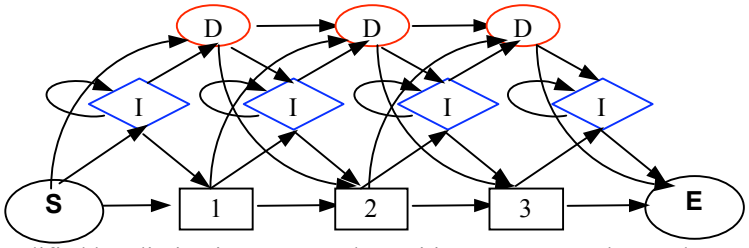
Figure (1) below shows a Multiple Sequence Alignment (MSA) of ezRNA gene components recovered from 10 ezLife species.

```
1110101-00001-111111
1110000-00001-101111
1110101000001-110111
1110-1--00000-110111
1110101-00001-101111
1110-0100000011-111
1110110-0000--111111
1110001-0000101111111
1110110--0000--11-111
1110101-00001-111111
```

Figure 1: An MSA of ezLife Genes

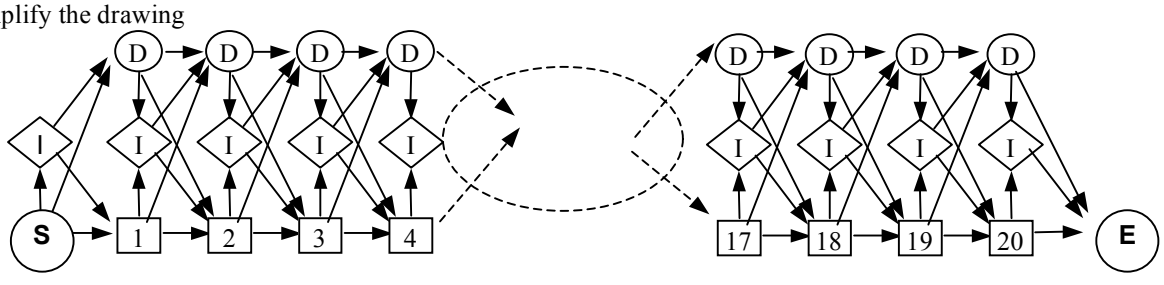
A. Create an HMM profile model for these gene components (hint: prune the HMM graph to aggregate single-symbol columns; explain your model.

First, some brief notes on HMM profile modeling. There are three types of hidden states. A Match state (represented by a square) corresponding to columns with under 50% indels; Delete states (represented by a circle) corresponding to deletions of individual characters within a match column; and Insert states (represented by a diamond) corresponding to columns with over 50% indels. This is depicted below.

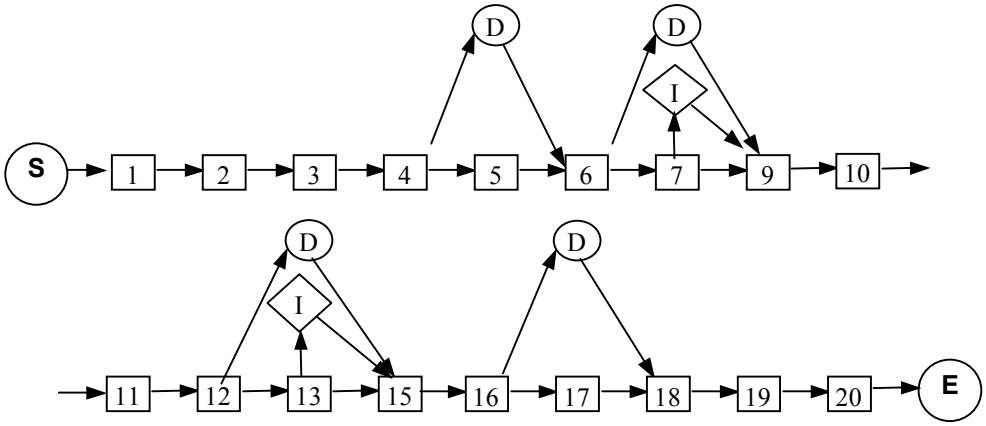


Second, the HMM model may be often simplified by eliminating states and transitions. For example, a column without deletions (e.g., column 6 of 1A) may be represented by a match state, with delete and insert states removed. Similarly, a column with deterministic emissions may be merged with a successor or predecessor match states by redefining the emission symbols; e.g., in figure 1(A) columns 1,2,3,4 may be merged into the start state by associating deterministic emission of 1110 with the start state. We proceed to establish such simplified HMM.

Step1: Construct an HMM graph for the 20 columns MSA. Note that self-loops at Insert state have been eliminated to simplify the drawing

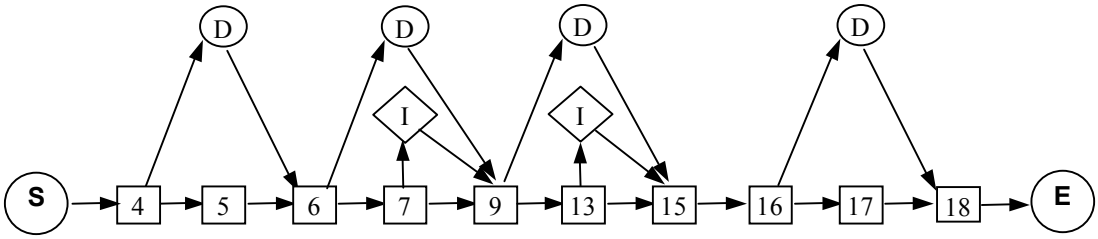


Step 2: Prune the HMM graph: (i) remove "Insert/Delete" states for columns without indels; (ii) remove "Insert" states of columns with less than 50% indels; (iii) remove Match states for columns with over 50% indels; (iv) remove gap arrows between Delete states for columns with a single gap; (v) remove transition arrows from Delete states to Insert states where insertion does not follow deletion.



At this stage one has a profile HMM graph for the MSA and can proceed to compute probabilities. However, the HMM may be further simplified through aggregation of states.

Step 3: Aggregate states of neighboring columns with deterministic output.



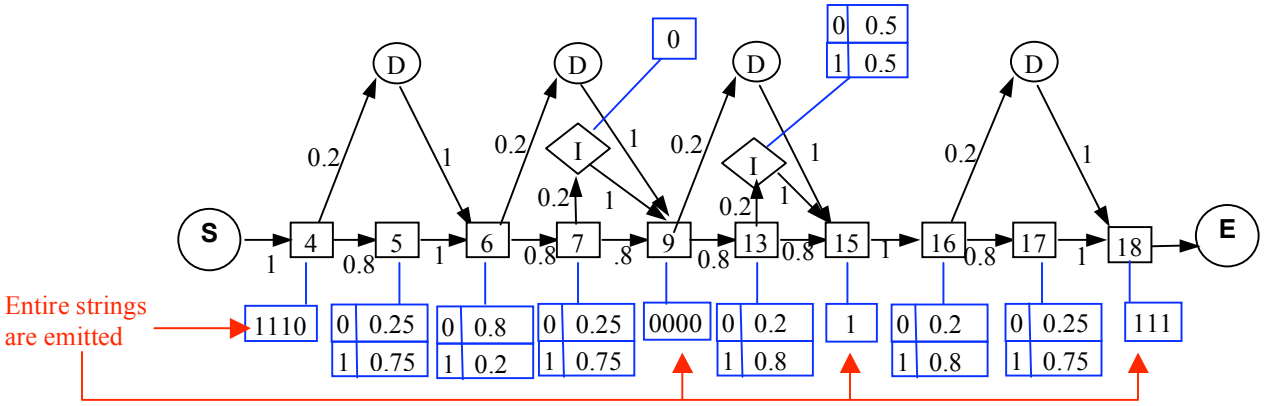
This HMM may be further truncated by aggregating more states as discussed in class. However, to simplify the explanations below I will keep this HMM model.

B. What would be the impact on the HMM graph if sequences 4 and 9 had a 0 in column 7; what would be the impact if sequences 2,3,5,6 had indel in column 7?

If sequence 4 had a 0 in column 8 it would have the Delete state of column 7 followed by an Insert state at column 8; this would require adding a transition arrow from the delete state of 7 to the Insert state. An indel in column 6 would add a Delete state for column 6 and transition arrows from the Delete states of 4 to 5 to 6 to handle the gap.

C. Compute the transition and emission probabilities for the HMM based on these sample sequences.

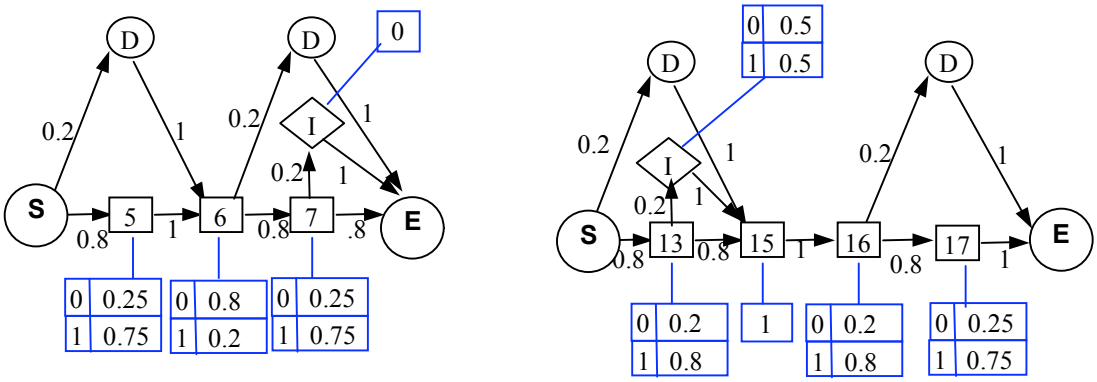
We compute the HMM probabilities by counting transitions and emissions in the MSA of figure 1(A).



The resulting transition probabilities are depicted near respective arrows below; emission probabilities are depicted in blue tables when only one symbol is emitted (probability 1) the table shows only the output symbol.

D. What is the most likely HMM path to generate the sequence: 1110111000001111111.
What is the most likely HMM path to generate the sequence: 1110111000001111111. What is the probability for generating this sequence.

To compute Viterbi decoding one could proceed along the process described in chapter 4.1 slides 18-22. This entails construction of a table with a row for each state and a column for each output symbol; if a full HMM with 20x3+2 states is considered, the dynamic programming table will be 63x18. The reduced HMM still has some 17 states requiring a table of 17x18. Clearly, it is useful to simplify this. Such simplification may be accomplished by using the deterministic columns with perfect matches (columns 1-4,9-12, 15, 18-20) to parse the given sequence as 1110111000001111111, where the underlined parts correspond to columns 1-4, 9-12 and 18-20. The Viterbi path must assign the underlined parts to the respective deterministic columns. This permits us to decompose the decoding in terms of two smaller HMMs, one for columns 5-7, which output 111 and one for columns 13-16 which output 1111. These are depicted in the figure below:



These HMMs are simple enough that Viterbi decoding may be computed by inspection.
The left HMM has two paths that output 111:
(a) {S,5,6,7,E} with probability P=(0.8*0.75)*0.2*(0.8*0.75)*0.8 corresponding to the alignment 111_.
(b) {S,D,6,7,I,E} with probability P=0.2*0.2*(0.8*0.75)*0.2 which is lower than (a) and is thus not part of the Viterbi path. The right HMM too admits two paths generating 1111:
(c) {S, 13,15,16,17,E} P=0.8*0.8*0.8*0.8*0.8*0.75 corresponding to 1_111
(d) {S,13,I,15,16,D,E} P=0.8*0.8*0.2*0.5*0.8*0.2 corresponding to 1111_
The second path (d) has lower probability and is thus not part of the Viterbi path.

Therefore, the Viterbi path for the observed sequence is {S,1,2,3,4,5,6,7,9,10,11,12,13,15,16,17,E} corresponding to the profile alignment 1110111_00001_111111. The probability of this path is: PS=(0.8*0.75*0.2*0.8*0.75*0.8)*(0.8*0.8*0.8*0.8*0.8*0.75)=(0.8^8)*(0.75^3)*(0.2).

It is often useful to evaluate the log likelihood of this Viterbi alignment to the background probability of {0,1}. The background probability of 0 is P(0)=71/186=0.38 P(1)=105/186=0.62. The probability for generating the observed sequence through background noise is: PB=P(1110111000001111111)= (0.38^5)*(0.62^{13}).

The log-likelihood ratio of the probabilities is:

Log (PS/PB) = [(8*log0.8)+(3*log0.75)+(log0.2)]-[5*log0.38 + 13*log0.62]=-6.14+15.96=9.82.
Therefore, the Viterbi path is 9.82 folds more likely to be generated by the profile HMM than by the background noise.

Problem 2: Computing Conformations of ezProteins

ezProteins are designed for a two-dimensional world; their folds may be entirely described in terms of planar conformations as follows:

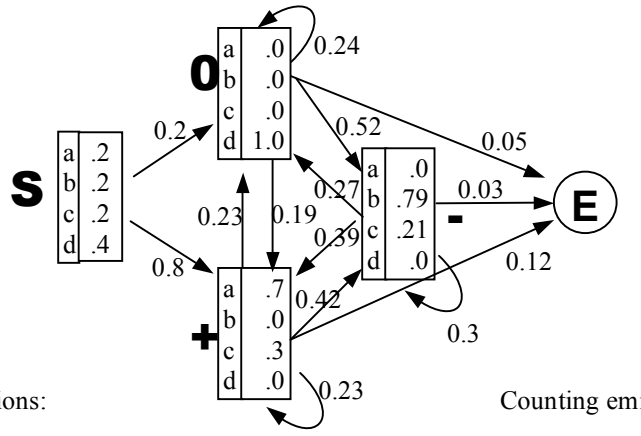
- The ezAA {a,b,c,d} may be described as two dimensional unit vectors (i.e., they have equal length and their width/length ratio is very small).
- The conformation angles of ezPeptide-bonds are $\{0,\pi/2,-\pi/2\}$. That is, the bond angle formed by an ezAA to its predecessor in an ezProtein sequence is either 0 (laying on the same line as the predecessor), or is orthogonal to the predecessor in a counterclockwise direction ($\pi/2$) or in a clockwise direction ($-\pi/2$); we will use `{0,+,-}` to describe these bond angles
- Therefore, ezProteins conformations may be represented as grid paths.

Consider an HMM model to analyze and predict ezProtein conformations. There are 5 hidden states, START (S) and End (E) states and one for each conformation angle `{0,+,-}`; these hidden states can emit the ezAA symbols {a,b,c,d}. Suppose the following database of ezProteins conformation has been obtained through crystallography:

ezProtein	Protein Sequence	Conformation Sequence
pyramidin	aabcbabdcabcba	S+-+-+-0-+-+-+
flagelin	bdddadbdbdcd	S000+0-0-0-0
cactuslin	cadcdbdbbaacbacbbdcdada	S+0+0-0--++--++--0-0+0+
Holin	dcddbcaabbddbbca	S+00--++--00--++
Snooplin	dcadbcabdbadddbbacbcabb	S++0--+-0-+000--+--++--

A. Draw the HMM with the transitions and emissions probabilities.

The HMM graph is simple; each conformation angle has a respective state and the transitions admissible by the sample data. The transition/emission probabilities are computed from the statistics of the samples.



Counting transitions:

➔	S	0	+	-	E
S		1	4	0	0
0	0	5	4	11	1
+	0	6	6	11	3
-	0	9	13	10	1
E	0	0	0	0	0

Counting emissions:

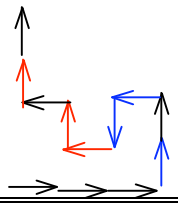
	a	b	c	d
S	1	1	1	2
0	0	0	0	23
+	19	0	8	0
-	0	26	7	0
E	0	0	0	0

B. A newly discovered ezProtein colin, has been sequenced: **dddadacbbacd**
Use Viterbi decoding to compute its likely conformation and draw it.

Consider the path length in terms of log likelihood of the respective transition/emission probabilities. The Viterbi decoding may be computed directly by inspection as seen by computing the first few columns of the table…The entries marked with a * represent very low probability events (large negative log likelihood) and are ignored.

	d	d	d	a	d	a	c	b	b	a	c	d
E		*	*	*	*	*						
-		*	*	*	*	*	-17.15	-19.09	-21.17		-26.55	
+		*	*	-8.61	*	-13.64	-17.50			-23.04	-26.9	
0		-3.64	-5.70	*	-10.73	*						-28.44
S	-1.32	*	*	*	*	*						

The optimal decoding is: S00+0++--+-0. This path has the following conformation:

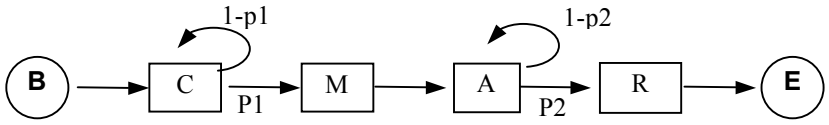


C.

(i) ZPCR proteins consist of 4 domains: (a) a Coupling domain, conformed as a small Zigzag (figure 2.1) of geometrically distributed length, averaging 8 ezAA; (b) a Membrane domain, conformed as a Line (figure 2.5) of 2 ezAAs; (c) a Signal Attractor domain, conformed as a small Train (figure 2.3), of geometrically distributed length, averaging 12 ezAAs; and (d) a Receptor domain, consisting of single small globulin (figure 2.6).
Design an HMM to detect ZPCR proteins; draw the HMM states, transition probabilities and emission probabilities (assume the emission probabilities are identical to those computed at part A). Explain your design choices and computations of probabilities.

I will include here long comments on considerations in HMM modeling. There could be several approaches for building an HMM model for ZPCR. The challenge in selecting a model is to balance the tradeoffs between model complexity, quality of predictions and robustness to noise. We follow first one such approach to illustrate modeling considerations, then consider various alternative modeling choices.

The primary design challenge here is how to model the macro structure of the ZPCR domains. We start with a coarse grain HMM assigning a state for each of the four ZPCR domains, as well as a Begin and End states. This is depicted below.

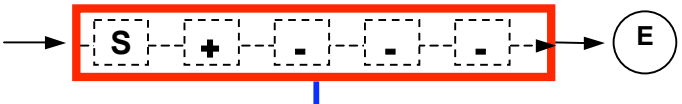


We note first that the M and R domains represent conformation of fixed length while the C and A domains represent variable length conformations. How do we model such variable length paths? The geometric distribution of path length is key to model variable length paths. These paths may be generated by feedback transitions that repeat them by "tossing coins" with respective probabilities to repeat C and A, as depicted in the figure above.

How do we associate emissions by these macro hidden-states? Clearly one has to derive such emissions model from those of micro-states associated with the macro states. For example, the R domain hides a small globulin

conformation. This conformation will emit respective ezAA sequences that may be incorporated with the macro-state R.

The globulin conformation depicted in figure 2.6 as "S+ - - - "; S represents choices of initial orientation; this conformation may be considered as a composite state of R. What sequences may be emitted by such globulin conformation state? The states "S+" can emit sequences of two symbols starting with {a,b,c,d} and terminating in either {a,c}. The states "---" may emit any sequence of length 3 formed from {b,c}. The probabilities of the different ezAA globulin sequences may be computed from the HMM model of part A and described in a tabular form as depited below. It is thus possible to represent the receptor macro state R as a sequence of micro-states and respective emissions of entire globulin sequences, as depicted in the figure below. For example, the Receptor state may emit the sequence "babbb" with probability 0.069. This table may be simplified by ignoring entries with probability under 0.01, as depicted by the table on the right.

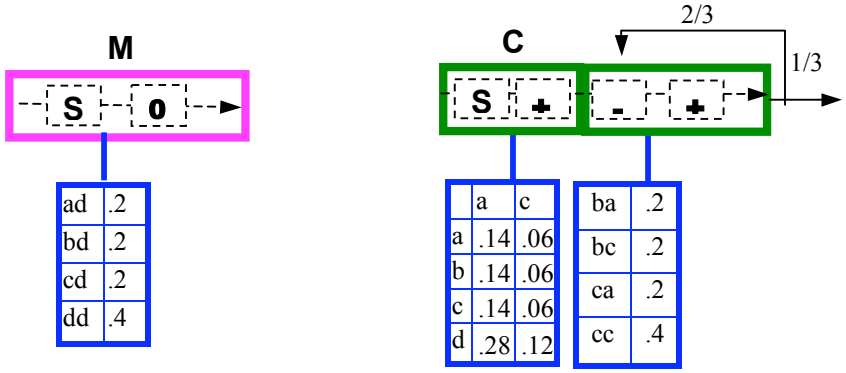


	bbb	bbc	bcb	cbb	bcc	cbc	ccb	ccc
aa	0.069	0.018	0.018	0.018	0.005	0.005	0.005	0.001
ac	0.030	0.008	0.008	0.008	0.002	0.002	0.002	0.001
ba	0.069	0.018	0.018	0.018	0.005	0.005	0.005	0.001
bc	0.030	0.008	0.008	0.008	0.002	0.002	0.002	0.001
ca	0.069	0.018	0.018	0.018	0.005	0.005	0.005	0.001
cc	0.030	0.008	0.008	0.008	0.002	0.002	0.002	0.001
da	0.138	0.037	0.037	0.037	0.010	0.010	0.010	0.003
dc	0.059	0.016	0.016	0.016	0.004	0.004	0.004	0.001

	bbb	bbc	bcb	cbb
aa	0.07	0.02	0.02	0.02
ac	0.03			
ba	0.07	0.02	0.02	0.02
bc	0.03			
ca	0.07	0.02	0.02	0.02
cc	0.03			
da	0.14	0.04	0.04	0.04
dc	0.06	0.02	0.02	0.02

This representation permits us to keep the number of states of the HMM small, with each state describing a respective macro conformation state of a domain and the sequences that it may emit.

The membrane domain M may be similarly modeled as a line segment of length 2 "S0". The emissions associated with this state are {ad,bd,cd,dd} with respective probabilities {0.2,0.2,0.2,0.4}. This is depicted in the figure below.



ad	.2
bd	.2
cd	.2
dd	.4

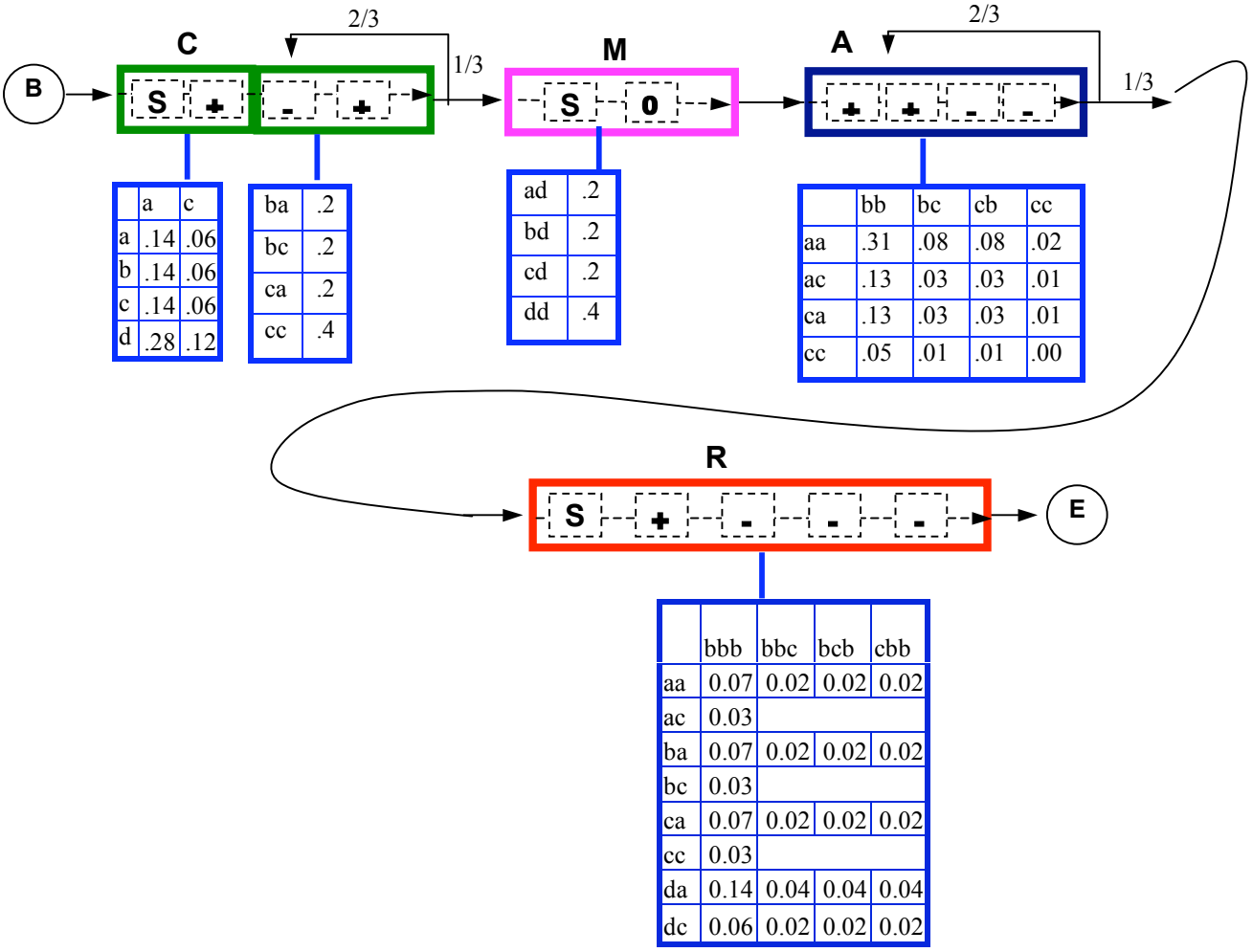
	a	c
a	.14	.06
b	.14	.06
c	.14	.06
d	.28	.12

ba	.2
bc	.2
ca	.2
cc	.4

The coupling domain macro-state C, depicted in the figure above, is conformed as a small zigzag described by the sequence of conformation states " (S+) $(+-)^k$ " where k is a repetition factor. We describe the macro-state in terms of two component states, an initialization state (S+) describing the first stair of a zig-zag and a repetition state (+-). The emissions for (S+) are the product set of pairs {a,b,c,d}x{a,c} with probabilities depicted by the respective table. The (-+) state emits pairs {ba,bc,ca,cc} with probabilities depicted by the table. The average number of repetitions is 8/2-1=4-1=3; therefore, the probability of feedback loop is 2/3. Notice that this model requires at least one stair of the zig-zag to appear in the coupling domain.

Similar considerations may be used to model the attractor domain A. However, instead of pursuing similar considerations as above we consider an alternative model to illustrate HMM modeling variations.
The Attractor domain consists of a train conformation. Such train conformation could be modeled as above with an initialization symbol S followed by square cars. An alternative model is to consider a train conformation as a repetition of the micro conformation states (++--). This would yield the model of A depicted in the figure below. This model, while simpler than one that includes initialization state as C, M, R, comes at a price. The HMM will reject any sequence that does not strictly fit the specific conformation selected here. Therefore, this model may lead to greater sensitivity to noise and true-negative errors.

The final HMM model resulting from the considerations above is depicted below.



	a	c
a	.14	.06
b	.14	.06
c	.14	.06
d	.28	.12

ba	.2
bc	.2
ca	.2
cc	.4

ad	.2
bd	.2
cd	.2
dd	.4

	bb	bc	cb	cc
aa	.31	.08	.08	.02
ac	.13	.03	.03	.01
ca	.13	.03	.03	.01
cc	.05	.01	.01	.00

	bbb	bbc	bcb	cbb
aa	0.07	0.02	0.02	0.02
ac	0.03			
ba	0.07	0.02	0.02	0.02
bc	0.03			
ca	0.07	0.02	0.02	0.02
cc	0.03			
da	0.14	0.04	0.04	0.04
dc	0.06	0.02	0.02	0.02

NOTE:

1. There could be other valid HMM models. In particular, one could add states to permit greater flexibility at the boundaries between the domains. For example, the initialization states may have some repetitions to allow entire initialization segments. Similarly, one can add termination states to allow flexible termination of a domain conformation.
2. The micro model could represent less constrained transition model than the macro model and thus increase its prediction power. For example, it could admit noisy conformations with imperfect zig-zag, train, lines and globulin. The macro model does not admit such noise, which is a significant deficiency if one wishes to study protein conformations.
3. These considerations illustrate the intrinsic complexities of HMM modeling. In general, in modeling large scale systems (e.g., genome components, protein domains) one wishes to hide details while minimizing the impact of such hiding on prediction power. At one extreme of HMM

one could model all microstates and transitions among them. Such HMM model may require a very large number of states and transitions at the price of both computational complexity as well as overfitting the microstate data and losing prediction of global features. The Viterbi decode of a given protein sequence may fail to predict global features of the protein conformation. At the other extreme a macro model may ignore the microscopic details and reduce the state complexity to focus on global features. This too has its price in noise sensitivity and true-negative errors.

4. The art of HMM modeling is thus to find a balance between the complexity of the HMM and its prediction power.

(ii) (10 points) Is the following ezProtein a ZPCR: **abcbdddabbaabcabcb** ?

We use inspection to try and identify a Viterbi path for the sequence. Start with the tail associated with the R domain: "cabcb" the emission table of R shows that this sequence has probability 0.02 of occurring (this is low, but still better than background noise would explain). However, the sequence "abbaab" preceding the tail cannot be emitted by the attractor state A. But is this a result of the sequence not representing a "train" conformation?

If one uses the HMM of part A to decode "abbaab" the resulting Viterbi path is "+--++-" which looks like a section of a train. Why is it that the Acceptor state of the macro HMM is unable to accept this train conformation? The reason for the failure is that the Acceptor state has a very rigid model of what a "train" conformation is; this conformation must start with ++ and terminates with – which is not the case here. To accommodate such variations in "train" conformation the Attractor model will need to be extended to permit noisy boundaries. Therefore, the more variants of a global feature need to be recognized, the more states and transitions must be incorporated in the model.

The initial segment "abcb" would too be rejected by the coupling state C. This sequence does not fit its zig-zag model. However, the Viterbi path of the micro-HMM decodes this path to a conformation S-+- which has a zig-zag structure. Again we see that the macro-HMM lacks the flexibility to accept various forms of noisy conformations and thus results in true-positive errors.

Finally, the "ddd" segment, likewise, would be rejected by the HMM above. However, by including an initialization state S with the Acceptor state model, the third "d" would become an emission of this initialization state while the initial pair "dd" would be admitted as an emission of the membrane domain M.

One can conclude that the given ezProtein has a ZPCR conformation but cannot be classified by the HMM as such due to its limitations in classifying global conformation features. To overcome these limitations the HMM would have to be greatly expanded to interpret the numerous variations of global features. This will not only increase the complexity of the HMM and computations associated with it, but it will result in false-positive errors as the HMM can accept sequences as ZPCR even when they are not.