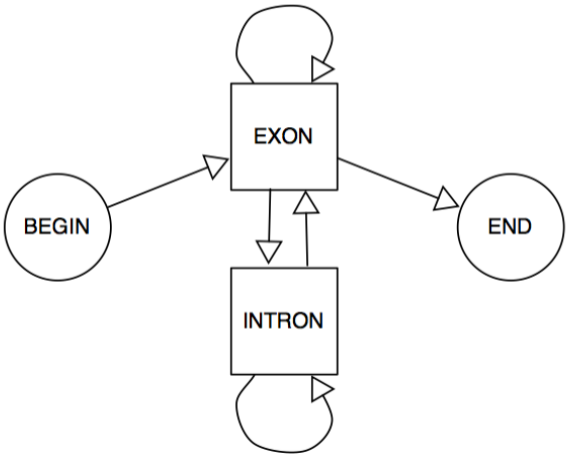
Mounir Abderrahmani mounir1badi@gmail.com

ECE 551 Homework 2 (Spring 2016)
March 25th, 2016

Mounir Abderrahmani mounir1badi@gmail.com

Q1: State Transition



For the small training set
ATGATACTTgtccgagTATATAG

ATGTTTTgtggcagAAAGA

ATGAATgtcgcgagTTTTATAG

	Exon	Intron
Exon	0,59	0,049
Intron	0.049	0.312

Emission Probabilities

	A	T	C	G
EXON	0.4	0.42	0.041	0.14
INTRON	0.123	0.123	0.213	0.541

Testing the set	gives PATH :	OUTPUT file
atgttaagtggccagttaatga atgaaaagtggggccagtaatga atggtcagtag	EEIIIIIIIIIIIIIEEEEEEE EEEEEEEIIIIIIIIIIEEEEEE EEEIIIIIEEE	ATgttaagtggccagTTAATGA ATGAAAAgtggggccagTAATGA ATGgtcagTAG

E denotes Exon state, and I to Intron state. The codes are here

Q2:

1)ATG = 0.7*0.1*0.7*0.2*0.7*0.4*0.9

Begin, M_1, M2, M3, end=4.3218E-3

2)ATG__=0.1*0.25*1*0.5*0.1*0.25*1*0.1*0.2*1

Begin, I_1, M_1, I_2, M_2, D3, end=

3) _ATG_=0.2*1*0.4*0.1*.25*1*0.4*0.9

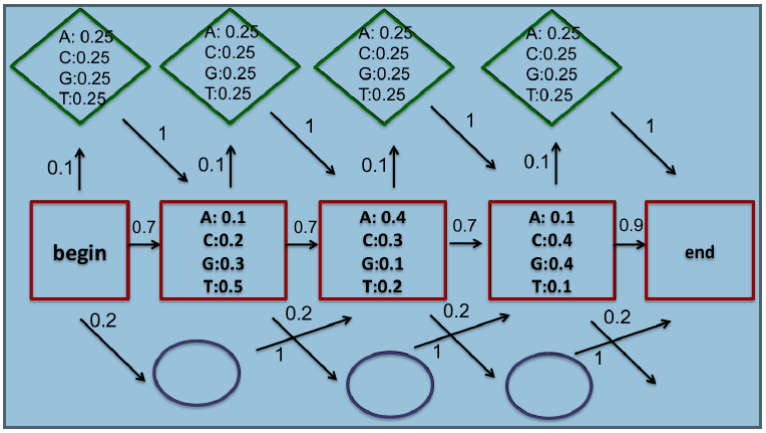
Begin, D_1, M_2,, I_3, M_3, end,

4) AT__G=0.1*0.25*1*0.5*0.2*1*0.4*0.9

Begin, I_1, M_1, D_2, M_3, end

5)A__TG = 0.7*0.1*0.2*1*0.1*0.1*0.25*1

Begin, M_1, D_2, M_3, I_4 , end.



Mounir Abderrahmani mounir1badi@gmail.com

Q3:

Motifs: GGAGG, GGAG or GAGG with probability 72%, 16% and 12%, comes after 4^{th} or 6^{th} position of a codon !

Here I made 10 upstream start of codons respecting the probabilities of the motifs:

CODON1: _ _ _ _ _ _ GGAGG

CODON2: _ _ _ _ GGAGG _ _

CODON3: _ _ _ _ _ GGAGG _

CODON4: _ _ _ _ _ _ GGAGG

CODON5: _ _ _ _ _ GGAGG _

CODON6: _ _ _ _ _ _ GGAGG

CODON7: _ _ _ _ _ GGAGG _

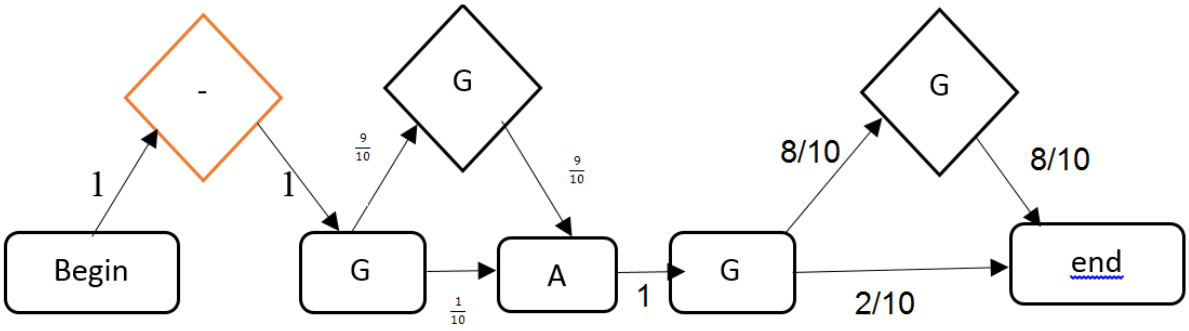
CODON8: _ _ _ _ _ GGAG _ _

CODON9: _ _ _ _ GGAG _ _ _

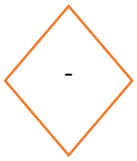
CODON10: _ _ _ _ _ GAGG _

The probabilities won't be accurate unless if we fix the conservative columns, if we consider that alignment we can do this: since they are randomly generated, but the emission observations are almost constant (G,A) starts with G and ends with G and has an A, Gs are duplicate 72% in both sides and 16% in first and 12% in the last side so we can have only 3 columns that is one for matching As and other two for matching Gs, and calculate the looping probabilities for insertions and deletion. We have number of max length of string is 11.

insertion

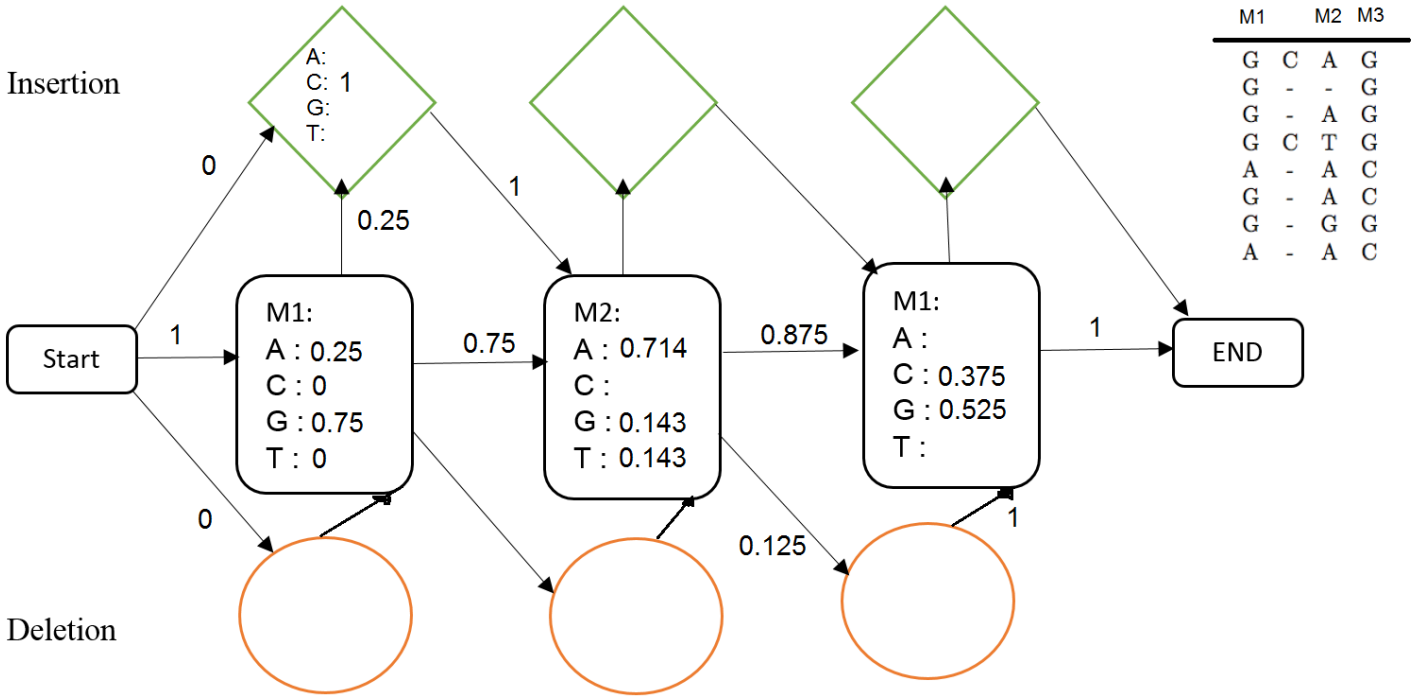


The orange insertion is a different HMM that need to be calculated with different emissions its length between 4 and 6

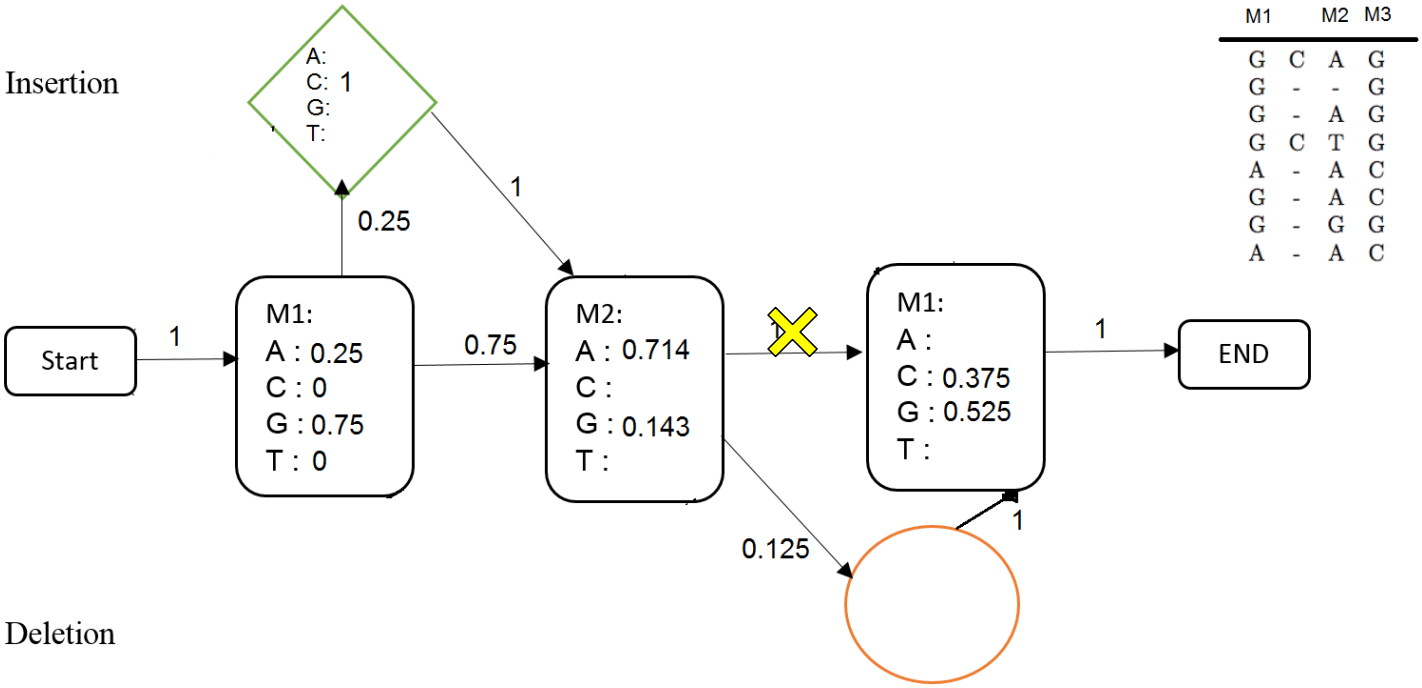


Mounir Abderrahmani mounir1badi@gmail.com

Q4: choosing 3 column out of 4 as most conservative columns, and one as insertion and one deletion occur in the second sequence in the transition from the second to the third column. (empty entries correspond to 0 probability)



M1		M2	M3
G	C	A	G
G	-	-	G
G	-	A	G
G	C	T	G
A	-	A	C
G	-	A	C
G	-	G	G
A	-	A	C



M1		M2	M3
G	C	A	G
G	-	-	G
G	-	A	G
G	C	T	G
A	-	A	C
G	-	A	C
G	-	G	G
A	-	A	C

B : since I have only 3 column and the first insertion doesn't loop that is it produce only once (c) , the sequence GCCAG can't be generated with the above model.