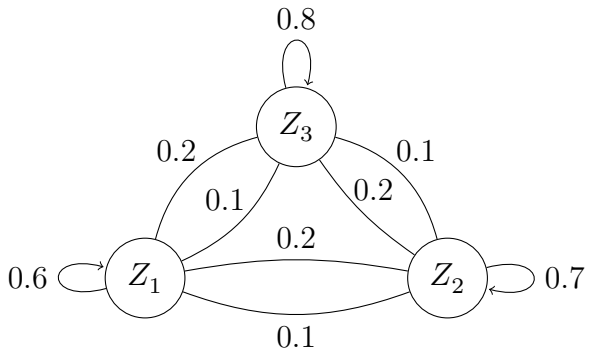
Genome Informatics mid-term exam, fall 2023.

November 23, 2023

3-state coin flipping HMM

Name & student ID: _____

Problem 1.



	Z_1	Z_2	Z_3
H	0.7	0.2	0.5
T	0.3	0.8	0.5

Emission probabilities of {H,T} for the 3 states.

Suppose the model starts in Z_1 or Z_2 each with 50% probability,
in other words: $\mathrm{P}[S_1 = Z_1] = \mathrm{P}[S_2 = Z_2] \;=\; 0.5$.

Define: $\alpha_{ij} \stackrel{\text{def}}{=} \mathrm{P}[X_{1..i}, S_i = Z_j|\lambda]$ $\beta_{ij} \stackrel{\text{def}}{=} \mathrm{P}[X_{i+1..n}|S_i = Z_j, \lambda]$
with λ meaning the HMM model and its parameter values described above.

Problem 1**a.** Formally prove that:

$$\mathrm{P}[X_{1..n}, S_i = Z_j|\lambda] \;=\; \alpha_{ij}\beta_{ij}$$

Explicitly state any assumptions used, including assumptions that are part of the definition of a Hidden Markov Model. I'm looking for a formal proof based mainly on symbolic manipulation.

3-state coin flipping HMM

Name & student ID: _____

Problem 1b.
Fill in the values missing from this table.

i	1	2	3	4	5	6
X	T	H	H	H	T	T
α_{i1}	0.1500	0.0910	0.0464	0.0230	0.0049	0.0015
α_{i2}	0.4000	0.0620	0.0134	0.0045	0.0079	0.0061
α_{i3}	0.0000	0.0550	0.0373	0.0209	0.0111	0.0057
β_{i1}	0.0415	0.0792	0.1463	0.2406	0.4400	1.0000
β_{i2}	0.0176	0.0387	0.1072	0.4506	0.6900	1.0000
β_{i3}	0.0328	0.0663	0.1348	0.2724	0.5100	1.0000

Question. Given $X =$ `THHHTT`, what is the expected number of time steps the model spent in state Z_2? In other words:

$$\sum_{i=1}^{n} \mathrm{P}[S_i = Z_2 | X, \lambda]$$

Counting Number of Alignments

Name & student ID: _____

Problem 2.

In the problem we consider the number of ways to align to sequences X and Y of lengths n and m respectively. Let x and y denote some character in X and Y respectively (x and y could be the same or different).

The basic rule is there are three kinds of columns.

```
(mis)match: x   gap in X:  -    gap in Y:  x
            y              y               y
```

Sometimes alignments with alternating gaps are disallowed. By "alternating gaps" I mean a column with a gap in X *immediately* following a column with a gap in Y. In other words an alignment containing

```
-x   or  x-
y-       -y
```

Has an alternating gap.

Problem 2a.
Let $C(n, m)$ denote the number of ways to align sequences of length n and m, **allowing** alternating gaps in the alignment. You can confirm by hand that $C(1, 1) = 3$.

Stipulate the recursive relations and base cases necessary to use dynamic programming to compute $C(n, m)$ for positive integers n, m.

Use this method to compute $C(3, 7)$. Make a table to show your work.

Counting Number of Alignments

Name & student ID: _____

Problem 2b.

Let $D(n, m)$ denote the count the number of ways to align sequences of length n and m, **disallowing alternating gaps** in the alignment. Consider $D(n, 0) = D(0, m) = 1$. You can confirm by hand that $D(1, 1) = 1$.

Stipulate the recursive relations and base cases necessary to use dynamic programming to compute $D(n, m)$ for positive integers n, m. Hint, you can use the same technique as that used for affine gap cost alignment.

Use this method to compute $D(3, 7)$. Make a table to show your work.

Neighbor Joining Phylogenetic Tree Inference

Name & student ID: _____

B	**C**	**D**	**E**	
16	16	14	10	**A**
	2	4	8	**B**
		4	8	**C**
			6	**D**

Problem 3.
Use the Saitou & Nei Neighbor joining algorithm to infer a plausible tree for species **A**,**B**,**C**,**D**,**E** (topology & edge lengths) from the distance matrix shown. Use symbols **F**,**G**,**H** to denote internal nodes.

DNA Dinucleotide Order Markov Model

Name & student ID: _____

Problem 4.

Background A zero order (plain) Markov model for single stranded DNA generates DNA sequences $X = X_1...X_n$ under the assumption that $\mathrm{P}[X_{i+1}|X_{1..i}] = \mathrm{P}[X_{i+1}]$. Since $\mathrm{P}[\mathtt{a}] + \mathrm{P}[\mathtt{c}] + \mathrm{P}[\mathtt{g}] + \mathrm{P}[t] \equiv 1$, the model has 3 degrees of freedom. Given a suitable prior distribution (don't worry about it for this question) and the length n of a training sequence X; the frequency of three nucleotides (`a`, `c`, and `g` for example) is sufficient (and minimal) information needed to train the model.

Problem 4a.
A 1st order (plain) Markov model for single stranded DNA generates DNA sequences $X = X_1...X_n$ under the assumption that $\mathrm{P}[X_{i+1}|X_i] = \mathrm{P}[X_{i+1}|X_{1..i}]$.

Given a suitable prior distribution and the length n of a training sequence, list a sufficient (**and as minimal as possible**) set of statistics on from X which can be used to train the model.

2-Stranded DNA Dinucleotide Order Markov Model

Name & student ID: _____

Problem 4b.

A 1st order (plain) Markov model for double stranded DNA generates double stranded DNA sequences $D = D_1 ... D_n$. Where each element D_i represents a pair from the set:
$\{\mathtt{a} = \mathtt{t}, \mathtt{c} \equiv \mathtt{g}, \mathtt{g} \equiv \mathtt{c}, \mathtt{t} = \mathtt{a}\}$. The assumption is that $\mathrm{P}[D_{i+1}|D_i] = \mathrm{P}[D_{i+1}|D_{1..i}]$.

So for example D might be:

```
gataca
ctatgt
```

One strand, `gataca`, happens to be written on top, but often this is arbitrary. In that case the opposite strand `tgtatc` should be equivalent for the purpose of training a Markov model.

Let D be double stranded data we want to train on. Assume we train on a sequence X which is one of the two strands of D. We constrain our result to be the same no matter which strand is used. So in the example above $X =$ `gataca` or $X =$ `tgtatc` should both give the same result.

Given a suitable prior distribution and the length n of the training sequence; list a sufficient (and as minimal as possible) set of statistics on from X which can be used to train the model. Explain as necessary to demonstrate your reasoning.