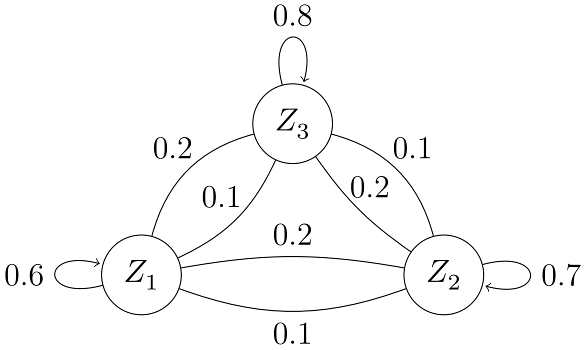
Genome Informatics mid-term exam, fall 2023.

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: $\alpha_{ij} \overset{\text{def}}{=} \mathrm{P}[X_{1..i}, S_i = Z_j|\lambda]$ $\beta_{ij} \overset{\text{def}}{=} \mathrm{P}[X_{i+1..n}|S_i = Z_j, \lambda]$

with λ meaning the HMM model and its parameter values described above.

Problem 1a. 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.

Solution: Step 1. Note that all quantities are conditioned on λ, so we can omit it. We can also simplify notation by writing just S_i with some value implicit, instead of $S_i = Z_j$. Restating the problem with λ omitted and expanding α_{ij}, β_{ij} according to their definitions we obtain:

Proof the following equality:

$$\mathrm{P}[X_{1..i}, S_i]\,\mathrm{P}[X_{i+1..n}|S_i] \;=\; \mathrm{P}[X_{1..n}, S_i]$$

$$\begin{aligned}
& \mathrm{P}[X_{1..i}, S_i]\ \mathrm{P}[X_{i+1..n} | S_i] & \\
= \ & \mathrm{P}[X_{1..i}, S_i]\ \mathrm{P}[X_{i+1..n} | S_i, X_{1..i}] & \text{Markov Assumption} \\
= \ & \mathrm{P}[X_{1..i}, S_i, X_{i+1..n}] & \mathrm{P}[A,B]\ \mathrm{P}[C|A,B] = \mathrm{P}[A,B,C] \\
= \ & \mathrm{P}[X_{1..n}, S_i] \checkmark &
\end{aligned}$$

In words, the Markov assumption is that given an HMM and trying to predict the future past time i, knowing the state of the HMM at time i (i.e. S_i) is the best you can do. You get no extra information by also considering $X_{1..i}$. (Or course if you do not know S_i, $X_{1..i}$ can help you predict the future because it gives you information about S_i).

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:

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

Solution: The sum be computed using

$$\mathrm{P}[S_i = Z_2 | X, \lambda] \quad = \quad \frac{\mathrm{P}[S_i = Z_2, X, \lambda]}{\mathrm{P}[X, \lambda]}$$

$$\mathrm{P}[X, \lambda] \quad = \quad \sum_j \alpha_{nj} \quad = \quad \alpha_{61} + \alpha_{62} + \alpha_{63} \approx 0.0015 + 0.0061 + 0.0057 = 0.0133$$

$$\forall_i \ \mathrm{P}[S_i = Z_2, X, \lambda] \quad = \quad \alpha_{i2}\beta_{i2}$$

$$\sum_i \mathrm{P}[S_i = Z_2 | X, \lambda] \quad = \quad (0.4000)(0.0176) + (0.0620)(0.0387) + (0.0134)(0.1072) +$$

$$(0.0045)(0.4506) + (0.0079)(0.6900) + (0.0061)(1.0000) \quad = \quad 0.02445458$$

$$\mathrm{P}[S_i = Z_2 | X, \lambda] \quad = \quad \frac{0.02445458}{0.0133} \approx 1.83869$$

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.

Solution: For a base case, the simplest approach is $i \geqq 0, C(i, 0) = C(0, i) = 1$.

If counting alignments of empty sequences seems strange, one can use the recurrence:

$$\forall_{i>1} C(1, i) \;\; = \;\; 1 + 1 + C(1, i-1)$$

which follows by considering three cases of the first column of the alignments

```
case 1  0 -------   one possible alignment
        - 0 ... i

case 2  0 ------    one possible alignment
        0  ... i

case 3  - 0-----    C(1, i-1)
        0  ... i    possible alignments
```

Counting Number of Alignments

The recurrence is

$$C(i, j) \quad = \quad C(i-1, j) + C(i, j-1) + C(i-1, j-1)$$

For positive integers i and j.

Using this method one may compute $C(3, 7) = 575$.

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.

Solution: We split into cases. Let $M(i, j)$ be the number of alignments of sequences of lengths i and j which end in a (mis)match, and $G(i, j)$ be the number of alignments which end with a gap in the second sequence.

The recurrence is:

$$\begin{aligned} M(i,j) &= M(i-1,j-1) + G(i-1,j-1) + G(j-1,i-1) \\ G(i,j) &= G(i-1,j) + M(i-1,j) \end{aligned}$$

For positive integers i and j.

For a base cases: $G(0, 0) = 0$, $i > 0$; $G(i, 0) = 1$, $j \geqq 0$ $G(1, j) = 1$.
$j > 0$; $M(1, j) = M(j, 1) = 1$.

And for the answer: $D(i, j) = M(i, j) + G(i, j)$

This method can be used to compute $D(3, 7) = 55$. The alignment paths are enumerated on the next page.

Counting Number of Alignments

```
0123456   0123456   0123456-  0123456   012-3456  0123456   0123456
--0-1-2   --0-12-   --0---12  --0--12   --012---  --012--   --01--2

0123456   0123456-  01234-56  0123456   0123456-  0123456-  012345-6
--01-2-   ----0-12  ----012-  ----012   ------01  -----012  -----012

0123456   0123456   0123-456  0123456-  0123456   0123456   0123456
---01-2   ---012-   ---012--  ---0--12  ---0-12   -01-2--   -01---2

0123456   0123456   01-23456  0123456   0123456   0123456-  0123456
-01--2-   -012---   -012----  -0--1-2   -0--12-   -0----12  -0---12

0123456   0123456   0123456   0123456   0123456   0123456   0123456
-0-12--   -0-1--2   -0-1-2-   0-1-2--   0-1---2   0-1--2-   0-12---

0123456   0123456   0123456-  0123456   0123456   0123456   0123456
0---1-2   0---12-   0-----12  0----12   0--12--   0--1--2   0--1-2-

0-123456  0123456   0123456   0123456   0123456   0123456   --012345
012-----  012----   01--2--   01----2   01---2-   01-2---   012-----

-0123456  -0123456  -0123456  -0123456  -0123456  -0123456
012-----  01--2---  01----2-  01-----2  01---2--  01-2----
```

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.

Solution: A straightforward solution is to simply apply the neighbor joining computations with r and D.

First we compute r, according to:

$$r_i \stackrel{\text{def}}{=} \frac{\Sigma_j\, d_{ij}}{\ell - 2} \qquad \ell = |\{\mathtt{A},\mathtt{B},\mathtt{C},\mathtt{D},\mathtt{E}\}| = 5$$

$$\begin{aligned}
r_{\mathtt{A}} = &\quad \tfrac{1}{3}(16+16+14+10) &= 18\tfrac{2}{3}\\
r_{\mathtt{B}} = &\quad \tfrac{1}{3}(16+2+4+8) &= 10\\
r_{\mathtt{C}} = &\quad \tfrac{1}{3}(16+2+4+8) &= 10\\
r_{\mathtt{D}} = &\quad \tfrac{1}{3}(14+4+4+6) &= 9\tfrac{1}{3}\\
r_{\mathtt{E}} = &\quad \tfrac{1}{3}(10+8+8+6) &= 10\tfrac{2}{3}
\end{aligned}$$

Then compute normalized distances D_{ij} accoding to:

$$D_{ij} \;\;=\;\; d_{ij} - r_i - r_j$$

B	**C**	**D**	**E**	
$16 - 18\frac{2}{3} - 10$	$16 - 18\frac{2}{3} - 10$	$14 - 18\frac{2}{3} - 9\frac{1}{3}$	$10 - 18\frac{2}{3} - 10\frac{2}{3}$	**A**
	$2 - 10 - 10$	$4 - 10 - 9\frac{1}{3}$	$8 - 10 - 10\frac{2}{3}$	**B**
		$4 - 10 - 9\frac{1}{3}$	$8 - 10 - 10\frac{2}{3}$	**C**
			$6 - 9\frac{1}{3} - 10\frac{2}{3}$	**D**

B	**C**	**D**	**E**	
$-12\frac{2}{3}$	$-12\frac{2}{3}$	-14	$-19\frac{1}{3}$	**A**
	-18	$-15\frac{1}{3}$	$-12\frac{2}{3}$	**B**
		$-15\frac{1}{3}$	$-12\frac{2}{3}$	**C**
			-14	**D**

Neighbor Joining Phylogenetic Tree Inference

The smallest (most negative) normalized distance is $D(\mathtt{A}, \mathtt{E}) = -19\frac{1}{3}$.

So let internal node `F` be the parent of `A` and `E`; with distance from those child nodes of:

$$\begin{aligned} d(\mathtt{A}, \mathtt{F}) &= \tfrac{1}{2}(d_{\mathtt{AE}} + \mathtt{r}_{\mathtt{A}} - \mathtt{r}_{\mathtt{E}}) = \tfrac{1}{2}(10 + 18\tfrac{2}{3} - 10\tfrac{2}{3}) = 9 \\ d(\mathtt{E}, \mathtt{F}) &= \tfrac{1}{2}(d_{\mathtt{AE}} + \mathtt{r}_{\mathtt{E}} - \mathtt{r}_{\mathtt{A}}) = \tfrac{1}{2}(10 + 10\tfrac{2}{3} - 18\tfrac{2}{3}) = 1 \end{aligned}$$

The (unnormalized) distance from the remaining leafs $i \in l\{\mathtt{B}, \mathtt{C}, \mathtt{D}\}$ to `F`, follow:

$$d(\mathtt{F}, \mathtt{B}) \;= \tfrac{1}{2}(d_{\mathtt{AB}} + d_{\mathtt{EB}} - d_{\mathtt{AE}}) = \tfrac{1}{2}(16 + 8 - 10) = 14$$

This answer unfinished...

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 4**a.**

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.

Solution: To train a 1st order model we need to know the frequencies of all dimers and monomers. There are 4x4=16 dimers, with the constraint that their probabilities must sum to one; so the dimer probabilities have at most 16-1=15 degrees of freedom.

One could say there are 12 major degrees of freedom and 3 minor ones. The 12 major ones are the conditional probabilities: $\mathrm{P}[X_{i+1}|X_i]$ there are 4x4= 16 of these with 4x3=12 degrees of freedom due to the sum-to-one constraints:

$$\text{for } b \in \mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t};\;\; \mathrm{P}[b\mathtt{a}|b.] + \mathrm{P}[b\mathtt{c}|b.] + \mathrm{P}[b\mathtt{g}|b.] + \mathrm{P}[b\mathtt{t}|b.] \equiv 1$$

Where $b.$ denotes base b followed by any other base.

The 3 minor degrees of freedom come from the need to stipulate the probabilities of the first base in the sequence

$$\text{for } b \in \mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t};\;\; \mathrm{P}[X_0 = \mathtt{a}] + \mathrm{P}[X_0 = \mathtt{c}] + \mathrm{P}[X_0 = \mathtt{g}] + \mathrm{P}[X_0 = \mathtt{t}] \equiv 1$$

These parameters might be considered minor because they only affect the probability of the base at the sequence start, and unless the start of the sequence is special, one might reasonably use the marginal probability of the dimer frequences to estimate them.

$$\text{for } b \in \mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t};\;\; \mathrm{P}[X_0 = b] \longleftarrow \sum_{b' \in \{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}} \mathrm{P}[b|bb']$$

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.

2-Stranded DNA Dinucleotide Order Markov Model

Solution: Note that for double stranded DNA dimers which form reverse complementary pairs always appear together; but some dimers are "DNA palindromic", in the sense that they are equal to their reverse complement.

aa	ac	ag	at	ca	cc	cg	ct	ga	gc	gg	gt	ta	tc	tg	tt	Forw 5'→3'
tt	tg	tc	ta	gt	gg	gc	ga	ct	cg	cc	ca	at	ag	ac	aa	Back 3'→5'
tt	gt	ct	at	tg	gg	cg	ag	tc	gc	cc	ac	ta	ga	ca	aa	Back 5'→3'

The four dimers `at`, `cg`, `gc`, `ta` are DNA palindromes. The other 12 dimers are not, and can be grouped into 6 pairs of equal probability.

$\mathrm{P}[\texttt{aa}] = \mathrm{P}[\texttt{tt}]$	$\mathrm{P}[\texttt{ac}] = \mathrm{P}[\texttt{gt}]$	$\mathrm{P}[\texttt{ag}] = \mathrm{P}[\texttt{ct}]$
$\mathrm{P}[\texttt{ca}] = \mathrm{P}[\texttt{tg}]$	$\mathrm{P}[\texttt{cc}] = \mathrm{P}[\texttt{gg}]$	$\mathrm{P}[\texttt{ga}] = \mathrm{P}[\texttt{tc}]$

Implying $\mathrm{P}[\texttt{aa}|\texttt{a.}]\,\mathrm{P}[\texttt{a}] = \texttt{P}[\texttt{tt}|\texttt{t.}]\,\texttt{P}[\texttt{t}]$, etc.

In terms of conditional probabilities we can use sum-to-one constraints to write the conditional probabilities of the 4 DNA palindrome dimers in terms of the conditional probabilities of the non-palindromic dimers.

$$\begin{aligned}
\mathrm{P}[\texttt{at}|\texttt{a.}] &= 1 - \mathrm{P}[\texttt{aa}|\texttt{a.}] - \mathrm{P}[\texttt{ac}|\texttt{a.}] - \mathrm{P}[\texttt{ag}|\texttt{a.}] \\
\mathrm{P}[\texttt{cg}|\texttt{c.}] &= 1 - \mathrm{P}[\texttt{ca}|\texttt{c.}] - \mathrm{P}[\texttt{cc}|\texttt{c.}] - \mathrm{P}[\texttt{ct}|\texttt{c.}] \\
\mathrm{P}[\texttt{gc}|\texttt{g.}] &= 1 - \mathrm{P}[\texttt{ga}|\texttt{g.}] - \mathrm{P}[\texttt{gg}|\texttt{g.}] - \mathrm{P}[\texttt{gt}|\texttt{g.}] \\
\mathrm{P}[\texttt{ta}|\texttt{t.}] &= 1 - \mathrm{P}[\texttt{tc}|\texttt{t.}] - \mathrm{P}[\texttt{tg}|\texttt{t.}] - \mathrm{P}[\texttt{tt}|\texttt{t.}]
\end{aligned}$$

Thus the conditional probabilities of the 6 non-palindromic dimer equivalence classes are enough to determine conditional probabilities for all the dimers.

Thus there are 6 degrees of freedome assuming we use the dimer frequencies to estimate the probability of the first base of the sequence.