Notation

x represents the data.

With discrete-state discrete-time Markov chains, the data are typically the state of the system at times $1, 2, \ldots, N$.

With DNA and protein sequence analysis, the data are often the amino acid or nucleotide residues that occupy positions $1, 2, \ldots, N$ of a sequence of length N.

Position i of the sequence will be denoted x_i.

For example, if

$$x = AGTCT,$$

then N=5. Also, $x_1 = A$, $x_2 = G$, $x_3 = T$, $x_4 = C$, and $x_5 = T$.

To represent subsequences, I will use x^i to indicate positions $1, 2, \ldots, i$ in the sequence.

In the example above,

$$\begin{aligned} x^1 = x_1 &= A \\ x^2 = x_1x_2 &= AG \\ x^3 = x_1x_2x_3 &= AGT \\ x^4 = x_1x_2x_3x_4 &= AGTC \\ x^5 = x_1x_2x_3x_4x_5 &= AGTCT \end{aligned}$$

Notice that $x = x^5$ in this example.

Conditional Probability

For events y and z,

$$\begin{aligned}\Pr(y, z) &= \Pr(y \mid z)\Pr(z) \\ &= \Pr(z \mid y)\Pr(y)\end{aligned}$$

Events y and z are independent if and only if

$$\Pr(y \mid z) = \Pr(y).$$

Equivalently, events y and z are independent if and only if

$$\Pr(z \mid y) = \Pr(z).$$

Equivalently (again), events y and z are independent if and only if

$$\Pr(y, z) = \Pr(y)\Pr(z).$$

Of course, $\Pr(y, z) = \Pr(z, y)$

Getting back to sequences . . .

If x_1 and x_2 are independent of each other, then

$$\Pr(x_1, x_2) = \Pr(x_1)\Pr(x_2).$$

If x_1, x_2, and x_3 are all independent of one another, then

$$\Pr(x_1, x_2, x_3) = \Pr(x_1)\Pr(x_2)\Pr(x_3).$$

If x_i and x_j are independent of one another for all i and all j, then

$$\Pr(x) = \Pr\left(x^N\right) = \Pr(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} \Pr(x_i)$$

and

$$\Pr(x_{i+1} \mid x_1, x_2, \ldots, x_i) = \Pr(x_{i+1}).$$

Markov chains

When $\Pr(x_i \mid x_1, x_2, \ldots, x_{i-1}) = \Pr(x_i)$ for all $i > 1$, we have a **Markov chain of order 0**.

When $\Pr(x_i \mid x_1, x_2, \ldots, x_{i-1}) = \Pr(x_i \mid x_{i-1})$ for all $i > 1$, we have a **Markov chain of order 1**.

Note: When people refer to Markov chains but do not refer to the order of the Markov chain, they are usually thinking about Markov chains of order 1

When $\Pr(x_i \mid x_1, x_2, \ldots, x_{i-1}) = \Pr(x_i \mid x_{i-2}, x_{i-1})$ for all $i > 2$, we have a **Markov chain of order 2**.

When $\Pr(x_i \mid x_1, x_2, \ldots, x_{i-1}) = \Pr(x_i \mid x_{i-k}, x_{i-k+1}, \ldots, x_{i-1})$ for all $i > k$, we have a **Markov chain of order k**.

Assume that sequence x has length N and that it can be described by a Markov chain of order 1. This means that

$$\begin{aligned}
\Pr(x) &= \Pr(x_1, x_2, \ldots, x_N) \\
&= \Pr(x_1)\Pr(x_2 \mid x_1)\Pr(x_3 \mid x_1, x_2) \\
&\quad \times\Pr(x_4 \mid x_1, x_2, x_3)\ldots \\
&\quad \times\Pr(x_N \mid x_1, x_2, \ldots, x_{N-1}) \\
&= \Pr(x_1)\Pr(x_2 \mid x_1)\Pr(x_3 \mid x_2) \\
&\quad \times\Pr(x_4 \mid x_3)\ldots\Pr(x_N \mid x_{N-1})
\end{aligned}$$

The same point in different notation …

$$\begin{aligned}
\Pr\left(x\right) &= \Pr\left(x^N\right) \\
&= \Pr\left(x^1\right)\Pr\left(x^2 \mid x^1\right)\Pr\left(x^3 \mid x^2\right) \\
&\quad \times\Pr\left(x^4 \mid x^3\right)\ldots\Pr\left(x^N \mid x^{N-1}\right) \\
&= \Pr\left(x_1\right)\Pr\left(x_2 \mid x^1\right)\Pr\left(x_3 \mid x^2\right) \\
&\quad \times\Pr\left(x_4 \mid x^3\right)\ldots\Pr\left(x_N \mid x^{N-1}\right) \\
&= \Pr\left(x_1\right)\Pr\left(x_2 \mid x_1\right)\Pr\left(x_3 \mid x_2\right) \\
&\quad \times\Pr\left(x_4 \mid x_3\right)\ldots\Pr\left(x_N \mid x_{N-1}\right)
\end{aligned}$$

Parameter Estimation

Assume that we have a DNA sequence x but we only pay attention to whether each position x_i is a purine or a pyrimidine.

A and G are purines. A generic purine will be denoted by R.

C and T are pyrimidines. A generic pyrimidine will be denoted by Y.

For example, this means that the sequence $x = ACGTC$ would effectively be $x = RYRYY$.

We also assume that we have a first order Markov chain.

In this case,

$$\begin{aligned}\Pr(x) &= \Pr(x_1 = R)\Pr(x_2 = Y \mid x_1 = R) \\ &\quad \times \Pr(x_3 = R \mid x_2 = Y)\Pr(x_4 = Y \mid x_3 = R) \\ &\quad \times \Pr(x_5 = Y \mid x_4 = Y)\end{aligned}$$

We assume E and F denote any two particular states of the Markov chain.

A time-homogeneous Markov chain is one in which $\Pr\left(x_i = E \mid x_{i-1} = F\right)$ is identical for all possible values of i.

In the purine/pyrimidine case, we assume that for all possible values of i

$$\begin{aligned}
\Pr\left(x_i = R \mid x_{i-1} = R\right) &= p_{RR} \\
\Pr\left(x_i = R \mid x_{i-1} = Y\right) &= p_{YR} \\
\Pr\left(x_i = Y \mid x_{i-1} = R\right) &= p_{RY} \\
\Pr\left(x_i = Y \mid x_{i-1} = Y\right) &= p_{YY}
\end{aligned}$$

p_{RR}, p_{RY}, p_{YR}, and p_{YY} are known as the transition probabilities of the Markov chain.

Notice also that

$$\begin{aligned} p_{RR} + p_{RY} &= 1 \\ p_{YR} + p_{YY} &= 1 \end{aligned}$$

The above two constraints mean that we have two degrees of freedom when specifying the values of the 4 parameters p_{RR}, p_{RY}, p_{YR}, and p_{YY}. If we know the values of p_{RR} and p_{YR} then we know the values of p_{RY} and p_{YY}.

Assume that we do know the values of p_{RR} and p_{YR}.

For given values of p_{RR} and p_{YR}, we can write the probability of observing the data $x = RYRYY$ as

$$\Pr(x \mid p_{RR}, p_{YR}) = \Pr(x_1 = R) p_{RY} p_{YR} p_{RY} p_{YY}.$$

If we are interested in estimating p_{RR}, p_{RY}, p_{YR}, and p_{YY} then we will have to decide how to treat $\Pr(x_1 = R)$.

One possibility is that $\Pr(x_1 = R)$ is known *a priori*. For example, maybe $\Pr(x_1 = R) = 1$ because we know the first position of the sequence will be a purine.

Another possibility is that $\Pr(x_1 = R)$ contains information that will help us estimate p_{RR}, p_{RY}, p_{YR}, and p_{YY}.

For example, maybe we are willing to assume that the Markov chain defined by p_{RR}, p_{RY}, p_{YR}, and p_{YY} has achieved *stationarity* prior to x_1.

Intuitively, the probability that a Markov chain occupies a particular state depends on the initial condition of the Markov chain and on the transition probabilities of the Markov chain.

If the Markov chain is sampled long after its beginning, then the state of the chain when it is sampled is almost independent of the initial state of the chain.

At the extreme when the Markov chain is sampled an infinite amount of time after its beginning, the state of the chain will be independent of the initial condition/state. This is *stationarity*.

Instead, the state of the chain will depend only on the transition probabilities p_{ij} and the probability that the Markov chain has state i when it has achieved stationarity will be denoted π_i.

For a stationary Markov chain,

$$\pi_j = \sum_i \pi_i p_{ij}$$

where i and $j \in \{R, Y\}$.

Back to sequence data ...

For simplicity, assume that $\Pr(x_1 = R) = 1$.

Then, we have in the above example ...

$$\Pr(x \mid p_{RR}, p_{YR}) = 1 \times p_{RY} p_{YR} p_{RY} p_{YY}.$$

In words, the above equation is the probability of the data given the values of the model parameters.

For discrete data, the probability of the data given the values of the model parameters is the *likelihood*.

Maximum likelihood estimation of parameters is done by finding the parameter values that maximize the likelihood.

In Bayesian analyses, it is the probability density of the parameter values given the data that guides our inference. Such a probability density is termed a *posterior* distribution.

To perform a Bayesian analysis, we must specify the prior distribution of our parameters.

A simple example would have the prior distributions $\Pr(p_{RR})$ and $\Pr(p_{YR})$ be independent of one another. Let's assume

$$\begin{aligned} \Pr(p_{RR}) &= 12{p_{RR}}^2(1-p_{RR}) \quad 0 \le p_{RR} \le 1 \\ \Pr(p_{YR}) &= 4(1-p_{YR})^3 \quad 0 \le p_{YR} \le 1 \end{aligned}$$

For a Bayesian analysis, we want to determine $\Pr(p_{RR}, p_{YR} \mid x)$

$$\Pr(p_{RR}, p_{YR} \mid x) = \frac{\Pr(x \mid p_{RR}, p_{YR}) \Pr(p_{RR}, p_{YR})}{\Pr(x)}$$

$$= \frac{\Pr(x \mid p_{RR}, p_{YR}) \Pr(p_{RR}, p_{YR})}{\int_{p_{RR}=0}^{1} \int_{p_{YR}=0}^{1} \Pr(x \mid p_{RR}, p_{YR}) \Pr(p_{RR}, p_{YR}) dp_{RR} dp_{YR}}$$

In our case, the posterior distribution is

$$\Pr(p_{RR}, p_{YR} \mid x) = 1800 {p_{RR}}^{2} (1 - p_{RR})^{3} p_{YR}^{1} (1 - p_{YR})^{4}$$

The mean of the posterior distributions ($\mathrm{E}(p_{RR} \mid x)$ and $\mathrm{E}(p_{YR} \mid x)$) would be conventional Bayesian estimates of p_{RR} and p_{YR}.

$$\mathrm{E}(p_{RR} \mid x) = \int_{p_{RR}=0}^{1} p_{RR} \Pr(p_{RR} \mid x) dp_{RR}$$

$$\mathrm{E}(p_{YR} \mid x) = \int_{p_{YR}=0}^{1} p_{YR} \Pr(p_{YR} \mid x) dp_{YR}$$

The MAP (*maximum a posteriori*) estimates are the values of the parameters that maximize the posterior density.

MAP estimates are used rather in cases where it is not computationally tractable to compute the posterior means.

(Note that we have been using $\Pr(\cdot)$ for both discrete and continuous probability densities)

Hidden Markov Models (HMMs)

A hidden Markov model is a case where data depend on “hidden” states and these hidden states are organized according to a Markov chain.

We cannot directly observe the hidden states but we can make inferences about the hidden states based on the data.

For example, some genome regions are GC-rich and some genome regions are AT-rich.

In a long DNA sequence, we may be interested in defining which are the GC-rich regions and which are the AT-rich regions.

The tricky part is the A and T can be found in GC-rich regions and G and C can be found in AT-rich regions.

For a sequence $x = x_1 x_2 \ldots x_N$, we will define a corresponding set of states $y = y_1 y_2 \ldots y_N$.

Let y_i be 1 if position x_i is in a GC-rich region and y_i be 0 if position x_i is in an AT-rich region.

Assume that x_i depends directly on y_i. Let,

$$\begin{aligned}
\Pr\left(x_i = A \mid y_i = 0\right) &= p_{A0} \\
\Pr\left(x_i = A \mid y_i = 1\right) &= p_{A1} \\
\Pr\left(x_i = G \mid y_i = 0\right) &= p_{G0} \\
&\text{etc.} \ldots
\end{aligned}$$

Also, assume that the states y_i follow a Markov chain where

$$\begin{aligned}
\Pr\left(y_{i+1} = 0 \mid y_i = 0\right) &= \rho_{00} \\
\Pr\left(y_{i+1} = 0 \mid y_i = 1\right) &= \rho_{10} \\
\Pr\left(y_{i+1} = 1 \mid y_i = 0\right) &= \rho_{01} \\
\Pr\left(y_{i+1} = 1 \mid y_i = 1\right) &= \rho_{11}
\end{aligned}$$

To make things simple, assume that $\Pr\left(y_1 = 1\right) = 1$ and therefore $\Pr\left(y_1 = 0\right) = 0$.

Pretend that we observe a sequence x and the hidden states y.

$$\begin{array}{lcccccccccccc} \mathrm{x} & = & \mathrm{G} & \mathrm{C} & \mathrm{A} & \mathrm{G} & \mathrm{C} & \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{A} & \mathrm{T} & \mathrm{G} \\ \mathrm{y} & = & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{array}$$

We can calculate
$\Pr(x, y \mid p_{A0}, p_{A1}, \ldots, p_{T1}, \rho_{00}, \rho_{01}, \rho_{10}, \rho_{11})$.

To shorten notation, we'll just write $\Pr(x, y)$ to represent this probability.

$$\begin{aligned} \Pr(x, y) &= \Pr(y)\Pr(x \mid y) \\ &= \Pr(y_1 = 1)\rho_{11}^6\rho_{10}\rho_{00}^3\Pr(x \mid y) \\ &= \rho_{11}^6\rho_{10}\rho_{00}^3\Pr(x \mid y) \end{aligned}$$

$$\Pr(x \mid y) = p_{G1}p_{C1}p_{A1}\ldots p_{G0}$$

If we assume that we observe x but not y,

$$\Pr(x) = \sum_{y} \Pr(x \mid y) \Pr(y)$$

In our example, $N = 11$ and so there are $2^N = 2^{11} = 2048$ possible y sequences.

When N is not so small, it is difficult to calculate the above sum without a clever algorithm.

to be continued ...