# Contents

# Chapter 2

# Discrete Time Markov Chain

So far we have considered very simple sequence models where the nucleotides at different sites are independent. It does not take much biological study to realize that DNA/RNA sequences carry out functions largely through the interaction of nucleotides at different positions in the sequence. For example, coding sequence is translated to amino acid sequences via codons, triplets of contiguous nucleotides encoding the genetic code. RNA sequences often carry out function via secondary structure that is formed by base pairing of nearby nucleotides or tertiary structure that can be formed by long distance pairings. Nucleotides that encode amino acids that interact via protein structure can also not be considered independent. In this chapter, we use the Markov chain to model some of this dependence, particularly local dependence between very nearby nucleotides.

## 2.1 Stochastic process

The Discrete Time Markov Chain (DTMC) is a special kind of stochastic process that is used frequently to model sequence data in Bioinformatics. Almost every course/book on stochastic processes has a chapter or two on the DTMC.

- Taylor & Karlin, *An Introduction to Stochastic Modeling*, 3rd edition. Chapters 3-4.
- Ross, *Introduction to Probability Models*, 8th edition, Chapter 4.

We start with some basic definitions.

**Definition 2.1** (stochastic process)**.** *A stochastic process is a collection of random variables* $\{X_t : t \in T, X_t \in \Omega_X\}$. *We may collectively refer to the random variables as* $\boldsymbol{X}$.

**Definition 2.2** (state space)**.** *The state space* $\Omega_X$ *is the collection of all possible values the random variables can take on, i.e. it is the sample space of the random variables. For example, if* $X_t \in [0, \infty)$ *represent random lengths for all* $t \in T$*, then the state space of the stochastic process is* $[0, \infty)$.

Often, the *index set* $T$ is associated with time, even when it does not actually represent time. In this representation, the stochastic process has a *state* that evolves in time. For example, the process may start in state $X_1 = 3$, then evolve to state $X_2 = 4$, and much later enters state $X_{100} = 340$. Another common index set is space, for example $T = \Re^2$ for the real plane.

| | | State Space | |
|---|---|---|---|
| | | discrete | continuous |
| **Index** | discrete | DTMC | not covered |
| **Set** | continuous | CTMC | diffusion processes |

Table 2.1: Classification of stochastic processes.

**Example 2.1** (Stochastic Processes)**.**

- ***HW1a.*** *We built codons from possibly non-independent nucleotides. The state space is* $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$*, the index set is the codon positions* $\{1, 2, 3\}$.
- ***HW1b.*** *We built ORFS, possibly CDS, from (assumed) independent codons. The state space is the set of codons, the index set is the ORF positions* $\{1, 2, 3, \ldots\}$.
- ***HW2.*** $\boldsymbol{d}_i = (b_{i1}, b_{i2}, \ldots, b_{il_i})$ *are the nucleotides observed in the* $l_i$ *(assumed) independent reads observed from individual* $i$*. The state space is* $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$*, the index set counts the reads.*
- ***Homology.*** *Given an alignment, we observed pairs of (assumed) independent nucleotides and recorded whether they were matched or not. The state space is* $\{0, 1\}$*, the index set is the position in the alignment.*

**Definition 2.3** (sample path)**.** *A sample path or realization of a stochastic process is the collection of values assumed by the random variables in one realization of the random process, e.g. the sample path* $x_1, x_2, x_3, \ldots$*, when* $X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots$.

We may speak of the probability or density of a realization, and we mean the joint distribution, either $P(X_1 = x_1, X_2 = x_2, \ldots \mid \boldsymbol{\theta})$ for discrete random variables or $f_{\boldsymbol{X}}(x_1, x_2, \ldots \mid \boldsymbol{\theta})$ for continuous random variables. The stochastic process is defined in terms of parameters $\boldsymbol{\theta}$. (These same functions are the *likelihoods* if we focus on $\boldsymbol{\theta}$.)

Stochastic processes can be classified by whether the *index set* and *state space* are discrete or continuous (Table 2.1). The DTMC is discrete in time and space, and appears frequently in bioinformatics to model sequences, with "time" genome position and "state" the sequence alphabet. The Continuous Time Markov Chain (CTMC) is continuous in time and discrete in space. It is used to model evolution of sequences (discrete) in time (continous). For discrete time processes, $T = \{0, 1, 2, \ldots\}$ are the natural numbers. For continuous time processes, the index set is often $T = [0, \infty)$, and we prefer to write the random variables as $X(t)$ to make it clearer that they are functions of a continuously varying parameter $t$.

A short history of stochastic processes illustrates the close connection with physical processes.

- 1852: DTMC invented to model rainfall patterns in Brussels
- 1845: branching process (type of DTMC) invented to predict the chance that a family name goes extinct.
- 1905: Einstein describes Brownian motion mathematically

- 1910: Poisson process describes radioactive decay
- 1914: birth/death process (type of CTMC) used to model epidemics

Relationship to other mathematics:

- mean behavior of the CTMC is described by ordinary differential equations (ODEs)
- diffusion processes satisfy stochastic differential equations (SDEs), from stochastic calculus

## 2.2 The DTMC

**Definition 2.4** (Markov property)**.** *A discrete time stochastic process* $\{X_n \in \Omega_X : n = 0, 1, 2, \ldots\}$ *with discrete state space is a Markov chain if it satisfies the Markov property.*

$$\begin{aligned} P(X_n = i_n \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) \\ = P(X_n = i_n \mid X_{n-1} = i_{n-1}), \end{aligned}$$

*where* $i_k \in \Omega_X$ *for all* $k = 0, 1, \ldots, n$ *are realized states of the stochastic process.*

**Brief History**

- Markov chain named after Andrei Markov, Russian mathematician who published first mathematical results in 1906.
- Andrey Kolmogorov, another Russian mathematician, generalized to countably infinite state spaces.
- Markov Chain Monte Carlo (MCMC) technique is invented by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller in 1953 in statistical physics.

## 2.3 Parameterization

### 2.3.1 Transition probabilities

**Definition 2.5** (one-step transition probabilities)**.** *The one-step transition probability is the probability that the process, when in state* $i$ *at time* $n$*, will next transition to state* $j$ *at time* $n + 1$*. We write*

$$p_{ij}^{(n,n+1)} = P(X_{n+1} = j \mid X_n = i).$$

- $0 \leq p_{ij}^{(n,n+1)} \leq 1$ since the transition probabilities are (conditional) probabilities.
- $\sum_{j=0}^{\infty} p_{ij}^{(n,n+1)} = 1$ since the chain must transition somewhere and summing over all $j$ is an application of the addition law for a set of disjoint and exhaustive events.

**Definition 2.6** (time homogeneity)**.** *When the one-step transition probabilities do not depend on time, so that*

$$p_{ij}^{(n,n+1)} = p_{ij}$$

*for all* $n$*, then the Markov chain is said to be time homogeneous.*

**Definition 2.7** (one-step transition matrix)**.** *When the state space is discrete, the one-step transition matrix, $P$, is formed by arranging the one-step transition probabilities into a matrix:*

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

- $P$ is a square matrix, possibly of infinite dimension if the state space is countable.
- The rows sum to 1, by properties of one-step transition probabilities given above.

**Example 2.2** (DNA sequence)**.** *Let $X_i$ be indicate the nucleotide at position $i$ of a genome. The index set $T = \{0, 1, 2, \ldots\}$ is discrete and represents position. The discrete and finite state space is $\Omega_X = \{0, 1, 2, 3\}$, where A$=$ 0, C$=$ 1, G$=$ 2, and T$=$ 3. Often, we abuse notation and let the random variable equal a nucleotide base, as in $X_i = A$.*

*Suppose the next nucleotide depends only on the nucleotide that came immediately before it, and no previous nucleotides (Markov property).*

*Let $p_{ij}$ be the probability of nucleotide $j \in \Omega_X$ given nucleotide $i \in \Omega_X$ proceeds it. The Markov matrix is*

$$P = \begin{pmatrix} - & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & - & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & - & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & - \end{pmatrix}.$$

**Definition 2.8** ($n$-step transition probabilities)**.**

$$p_{ij}^{(n)} = P(X_{n+k} = j | X_k = i),$$

*for $n \geq 0$ and states $i, j \in \Omega_X$.*

*By analogy to the 1-step case, we can define $n$-step transition probability matrices $P^{(n)} = (p_{ij}^{(n)})$.*

**Theorem 2.1** (Chapman-Kolmogorov equations)**.**

$$p_{ij}^{(n+m)} = \sum_{k=0}^{\infty} p_{ik}^{(n)} p_{kj}^{(m)}$$

*for all $n, m \geq 0$ and all states $i, j$.*

*Proof.*

$$\begin{aligned} p_{ij}^{(n+m)} &= P(X_{n+m} = j | X_0 = i) \\ &= \sum_{k=0}^{\infty} P(X_{n+m} = j | X_n = k, X_0 = i) P(X_n = k | X_0 = i) \\ &= \sum_{k=0}^{\infty} p_{kj}^{(m)} p_{ik}^{(n)}. \end{aligned}$$

□

- Another compact way to write Chapman-Kolmogorov equations:

$$P^{(n+m)} = P^{(n)} P^{(m)}$$

- By induction,
$$P^{(n)} = P^n.$$

**Example 2.3** (First order R/Y sequence)**.** *Suppose a two state (purine/pyrimidine) Markov chain has one-step transition probability matrix*

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

*What is the probability there will be another purine $4$ positions ahead, given there is a purine in the current position? We need $P^4$.*

$$P^2 = P \cdot P = \begin{pmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{pmatrix}$$

*and*

$$P^4 = P^2 \cdot P^2 = \begin{pmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{pmatrix}.$$

*The entry we seek is $p_{11}^4 = 0.5749$, so there is approximately a $57\%$ chance there will be another purine $4$ nucleotides ahead.*

### 2.3.2 Initial state distribution

In order to compute unconditional probabilities, like "What is the probability of R at site $j$?", we'll need to define the initial state distribution.

**Definition 2.9** (initial state distribution)**.** *The initial state distribution is a probability distribution defined over the first state of the chain $X_0$.*

$$P(X_0 = i) = \alpha_i,$$

*for all $i = 0, 1, \ldots \in \Omega_X$.*

- A Markov chain is fully specified once the transition probability matrix and the initial state distribution have been defined.
- How many parameters are there in an $m$-state Markov chain?

## 2.4 Likelihood

**Unconditional Probabilities**

Now, we can compute unconditional probabilities. Computing probability of state $j$ at particular time $n$:

$$\begin{aligned} P(X_n = j) &= \sum_{i=0}^{\infty} P(X_n = j | X_0 = i) P(X_0 = i) \\ &= \sum_{i=0}^{\infty} p_{ij}^{(n)} \alpha_i. \end{aligned}$$

Computing probability of a chain realization:

$$\begin{aligned} P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) &= \\ P(X_0 = i_0) P(X_1 = i_1 \mid X_0 = i_0) P(X_2 = i_2 \mid X_0 = i_0, X_1 = i_1) & \\ \cdots P(X_n \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}). &= \\ P(X_0 = i_0) P(X_1 = i_1 \mid X_0 = i_0) \cdots P(X_n \mid X_{n-1} = i_{n-1}), & \end{aligned}$$

and substituting in our parameters, we obtain

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \alpha_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}.$$

**Example 2.4** (R/Y sequence)**.** *Using the simple R/Y sequence model, what is the probability of a* Y *four nucleotides ahead, given that there was a* $90\%$ *chance of* Y *at current site?*

$$\begin{aligned} P(X_4 = 1) &= \alpha_0 p_{01}^{(4)} + \alpha_1 p_{11}^{(4)} \\ &= 0.10 \times 0.4251 + 0.90 \times 0.4332 = 0.43239. \end{aligned}$$

**Practice**

$$P = \begin{pmatrix} 0.4 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.4 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.4 \end{pmatrix}$$

$$\alpha_i = 0.25, \text{ for } i \in \{A, C, G, T\}.$$

1. What is $P(AA)$?
2. What is $P(AAAAAA)$?
3. What is $P(ANA)$ where $N$ represents any base?

---

4. What is the likelihood of AACGTAAAATTGCAGGGTTTGCT?
5. What is the likelihood of

   AACGTAAAATTGCAGGGTTTGCT, and
   TACAGCGGGGCTCCTCCTAACCGCAACGAAAATGT

   if these two sequences are independent?

## 2.5 Higher order Markov chains

We can handle some kinds of longer range dependences by increasing the *order* of a chain.

**Definition 2.10** (chain order)**.** *The order* $k$ *of a chain is the length of the context that impacts the next state. The Markov property for an order* $k$ *chain is*

$$P(X_n \mid X_0, X_1, \ldots, X_{n-1}) = P(X_n \mid X_{n-k}, X_{n-k+1}, \ldots, X_{n-1}) \tag{2.1}$$

*for* $n \geq k$*.*

Another way to think about higher order chains is to redefine the state space as $\Omega_X^k$, the $k$-dimensional product space of the original state space. For example, for the nucleotide model with $k = 2$, the state space becomes $\Omega_X^k = \{\text{AA, AC, AG}, \ldots, \text{TG, TT}\}$. Now, the state $X_n$ at time $n$ is a dinucleotide, and a transition probability $P(X_{n+1} = j \mid X_n = i)$ is non-zero only if state $j$ is compatible with $i$. For example AA can transition to (is compatible with) AC, but it cannot transition to CA. Thus, $p_{\text{AA,AC}} \equiv 0$, and the one-step transition probability matrix is increasingly sparse as $k$ increases. Specifically, though the one-step transition probability matrix for the $k$th order DNA model is of dimension $4^k \times 4^k$, there are at most $4$ non-zero entries in each row. For the sake of your computer's RAM, it is much more efficient to store it as a $4^k \times 4$ matrix, even if mathematically we think of a square matrix transition probability matrix.

Which formulation is better, using Eq. (2.1) or the state space $\Omega_X^k$, depends on context. I may use both interchangeably. Notice, however, that we cannot compute a likelihood without defining an initial state distribution on extended state space $\Omega_X^k$.

## 2.6 Uses of Markov chains

### 2.6.1 Run lengths

Given initial state $i$, let $Y$ be the number of consecutive visits to state $i$ before departing for state $j \neq i$. $Y$ is a random variable in sample space $\in \{0, 1, 2, \ldots\}$. The distribution of $Y$ is clearly

$$\begin{aligned} \Pr(Y = 0) &= 1 - p_{ii} \\ \Pr(Y = 1) &= p_{ii}(1 - p_{ii}) \\ \Pr(Y = 2) &= p_{ii}^2(1 - p_{ii}) \\ &\vdots \end{aligned}$$

In general, $\Pr(Y = y) = p_{ii}^y(1 - p_{ii})$ reveals $Y \sim \text{Geometric}(1 - p_{ii})$, the number of failures before the first success when successes occur with probability $1 - p_{ii}$.

As defined above, $Y$ is one less than the run length. In general, the length of a run of $i$'s starting *after* any nucleotide $j$ at position $n$ is

$$\Pr(Y = y \mid X_n = j) = \begin{cases} 1 - p_{ji} & y = 0 \\ p_{ji}p_{ii}^{y-1}(1 - p_{ii}) & y > 0, \end{cases}$$

which is a geometric-like distribution.

Notice that

$$P(Y \geq y) = p_{ii}^y$$

for the general Geometric$(1 - p_{ii})$ and

$$P(Y \geq y) = p_{ji}p_{ii}^{y-1},$$

$y > 0$, for the geometric-like distribution for run lengths in a Markov chain. We can derive this from first principles or from the Geometric cdf $P(Y \leq y - 1) = 1 - p_{ii}^y$.

A test statistic that often comes up with respect to runs is the longest run of state $i$ in a sequence $\boldsymbol{S} = (S_1, S_2, \ldots, S_m)$. Let $Y_l$ be the length of the run of state $i$ starting after the $l$th nucleotide in sequence $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ consisting of all the non-$i$ nucleotides (remove all $i$'s from the original sequence). If $S_j \neq i$ and $S_{j+1} \neq i$, then $Y_l = 0$ if the $j$th nucleotide $S_j$ is the $l$th non-$i$ nucleotide. Note, there is a possible run $Y_0$ before the first nucleotide $X_1$ and another $Y_n$ after the last nucleotide $X_n$.

Now, let $Y_{\text{max}} = \max_{l \in \{0,1,\ldots,n\}} Y_l$. If we condition on the realized states in $\boldsymbol{X}$, then

$$P(Y_l \geq y \mid X_l = j) = p_{ji}p_{ii}^{y-1}$$

for $l > 0$ and $y > 1$. Furthermore, $Y_{l-1}$ and $Y_l$ are independent because given $X_l$, the part of the sequence that determines $Y_{l-1}$ cannot influence $Y_l$ under the Markov property. Therefore, for $y > 0$,

$$\begin{aligned} P(Y_{\text{max}} < y \mid \boldsymbol{X}) &= P(Y_0 < y, Y_1 < y, \ldots, Y_n < y \mid \boldsymbol{X}) && \text{equivalent events} \\ &= \prod_{l=0}^n P(Y_l < y \mid \boldsymbol{X}) && \text{independence} \\ &= \prod_{l=0}^n \left[1 - p_{X_l i}p_{ii}^{y-1}\right], && \text{above derivations} \end{aligned}$$

with the understanding that $p_{X_0 i} = \alpha_i$ is the probability of starting the chain in state $i$. Then $P(Y_{\text{max}} \geq y \mid \boldsymbol{X}) = 1 - P(Y_{\text{max}} < y \mid \boldsymbol{X})$.

**Example 2.5** (2017 HW2 Example)**.** *Given the* 802*nd sequence with nucleotide counts*

```
seq <- paste("ACGACGATCAGCGCAGCAGGCGAGAATCCTCCGCAATGCG",
        "AGCAATCGCGACGGGGGGGGACCCCAAGTGCCACTCTTAA",
        "CGGGGTGGCTTTTCTTAAGTGTAAAAAGCTTTTGGAATAA",
        "GGGCTGGGCAAGACCGGTGCCAGCCGCCGCGGTAACACCG",
        "GCAGCTCTAGTGGTAGCCATTTTTATTGGGCCTTAAAGCGGTT", sep="")
table(strsplit(seq, ""))

##
##  A  C  G  T
## 47 51 64 41
```

*What is the probability the maximum run of* Gs *is* 8 *or longer under the iid model?*

```
p.G <- table(strsplit(seq, ""))["G"]/nchar(seq)
n.G <- sum(strsplit(seq, "")=="G")
1 - (1 - p.G^8)^(nchar(seq) - n.G)

##          G
## 0.01961973
```

**Example 2.6** (2017 HW2 Example: Extension)**.** *Now answer the question while modeling the sequence with a Markov chain of order* 1*.*

```
require(stringr, quietly = T)
nucs <- strsplit(seq, "")[[1]]
# count number of transitions out of A, C, G, T
n <- table(nucs)
n[nucs[nchar(seq)]] <- n[nucs[nchar(seq)]] - 1
# MLE of 1-step transition probability matrix
P.hat <- rbind(
   str_count(seq, c("AA","AC","AG","AT"))/n["A"],
   str_count(seq, c("CA","CC","CG","CT"))/n["C"],
   str_count(seq, c("GA","GC","GG","GT"))/n["G"],
   str_count(seq, c("TA","TC","TG","TT"))/n["T"])
rownames(P.hat) <- colnames(P.hat) <- c("A","C","G","T")
```

*We try two different initial state distributions below. It has little effect on the* $p$*-value given it appears just once in the equation.*

```
P.hat

##           A         C         G         T
## A 0.2765957 0.1914894 0.3191489 0.1489362
## C 0.2352941 0.2352941 0.3137255 0.1960784
## G 0.1406250 0.3437500 0.2656250 0.1406250
## T 0.2250000 0.1750000 0.2250000 0.2750000

alpha <- rep(0.25, 4)
1 - prod(1 - c(alpha[3], P.hat[nucs,3]) * P.hat[3,3]^(8-1))

## [1] 0.005348149
```

```
alpha <- str_count(seq, c("A","C","G","T"))/nchar(seq)
1 - prod(1 - c(alpha[3], P.hat[nucs,3]) * P.hat[3,3]^(8-1))

## [1] 0.005354206
```

## 2.7 Irreducibility

There is one more big result in DTMC theory that we need to learn. We need to understand what happens to the chains that "run" for a long time, that is long genomes, long Bayesian samplers, which use Markov chains, *etc*. In short, we want to know $P(X_n = i)$ when $n$ is large for $i \in \Omega_X$.

I prefer, when possible, to provide reasons for results. Although we leave out some proofs, the goal is always to help you understand what you are doing rather than just make you a user of formulae. Finding the balance is not easy, and parts of this section may appear to deviate from a straight line to our goal. The deviations here are the minimum you need to understand the term "ergodic," which I have occassionally seen in the bioinformatics literature.

The states in the state space of a Markov can be classified in multiple ways. Markov chains gain important properties based on the properties of their states, so it is important to know some of these classifications.

Define the 0-step transition probabilities as the Kronecker delta, $p_{ij}^{(0)} = \delta_{ij}$, to communicate the sensible reality that after taking 0 steps, you can only be in the same state you were in when you started.

**Definition 2.11** (accessible)**.** *State $j$ is said to be accessible from state $i$ if $p_{ij}^{(n)} > 0$ for some $n \geq 0$.*

**Definition 2.12** (communicate)**.** *Two states $i$ and $j$ are said to communicate if they are accessible to each other, and we write $i \leftrightarrow j$.*

The relation of communication is an equivalence relation, *i.e.* it has the following three properties:

- **Reflexive:** $i \leftrightarrow i$ because $p_{ii}^{(0)} = 1$.
- **Commutative:** If $i \leftrightarrow j$, then $j \leftrightarrow i$.
- **Transitive:** If $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$.

This equivalence relation divides the state space of a Markov chain into non-overlapping classes.

**Definition 2.13** (class property)**.** *A class property is a property of a state that if true of one member in a class, is true of all members in that class.*

All these definitions lead to the very important irreducibility property of some Markov chains.

**Definition 2.14** (irreducible)**.** *A Markov chain is irreducible if there is only one equivalence class of states, i.e. all states communicate with each other.*

**Examples:**

- Is this Markov chain irreducible?

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

- Or this one?

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix} ??$$

This box defines a technique for checking irreducibility.

**Definition 2.15** (regular)**.** *A transition probability matrix $P$ is regular if there exists $n$, such that $P^n$ has strictly positive entries, i.e. $p_{ij}^{(n)} > 0$ for all $i, j \geq 0$.*

**Lemma 2.1.** *A Markov chain with a regular transition probability matrix is irreducible.*

Note that for the $n$ where $P^n > 0$, $p_{ij}^{(n)}$ for all $i, j \geq 0$, hence all states $i$ in the state space communicate with all other states $j$.

**Method:** One way to check for irreducible Markov chains is to roughly calculate $P^2, P^4, P^8, \dots$ to see if eventually all entries are strictly positive.

**Example 2.7** (checking irreducibility)**.** *Consider, the $3 \times 3$ matrix from the first example above.*

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

*First, encode entries as $+$ or $0$ and call this encoded matrix $Q$.*

$$Q = \begin{pmatrix} + & + & 0 \\ + & + & + \\ 0 & + & + \end{pmatrix}.$$

*Then,*

$$Q^2 = \begin{pmatrix} + & + & + \\ + & + & + \\ + & + & + \end{pmatrix}.$$

*Therefore, the Markov matrix $P$ is irreducible.*

## 2.8 Recurrence and transience

Think of a Markov chain as an infinite sequence of states. We are now interested in whether the Markov chain returns to states it has visited before. There are some fascinating and nonintuitive findings that come out of this line of study that will not be of much interest to us in Bioinformatics. I'll just say, did you know that if you started walking randomly around a 2D grid, you would not be expected to return to the origin in finite time? Good thing the earth is round!

While we won't be studying Markov chains on 2D grids, we do need a little bit of this theory in order to understand how most Markov chains are applied and estimated in Bioinformatics.

Let $f_i$ be the probability that starting in state $i$, the process reenters state $i$ at some later time $n > 0$. Quantity $f_i$ does not depend on $n$; it is the probability that the chain will return to state $i$ in the future. For the drunk wandering aimlessly on the plane, this return may occur at time $\infty$, but that does not mean it will not occur. Those of us who live finite lives might automatically consider an infinite return time as not happening, but not so! Also, a state $i$ may be *accessible* from another state $j$, but accessibility does not mean a guaranteed future visit. In the example below, a chain starting in state 0 can return in two or more steps, but the chain is not guaranteed to return (if it goes to state 2 instead), so $f_0 < 1$.

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

**Definition 2.16** (recurrent)**.** *If $f_i = 1$, then the state $i$ is said to be recurrent.*

**Definition 2.17** (transient)**.** *If $f_i < 1$, then the state $i$ is a transient state.*

**Definition 2.18** (absorbing state)**.** *State $i$ is said to be an absorbing state if $p_{ii} = 1$.*

An absorbing state is a special kinds of recurrent state. Absorption is the process by which Markov chains absorb into an absorbing state.

There are some provable facts about recurrent and transient states.

**Lemma 2.2.** *A recurrent state will be visited infinitely often.*

*Proof.* Suppose the recurrent state $i$ is visited only $T < \infty$ times. Since $T$ is the last visit, there will be no more visits to state $i$ after time $T$. This is a contradiction since the probability that $i$ is visited again after time $T$ is $f_i = 1$. □

In fact, this is an extension of one fact we already new! Our previous result on run lengths is obtained from the next lemma by making all non-run states absorbing states.

**Lemma 2.3.** *The random number of times a transient state will be visited is finite and distributed as a geometric random variable.*

*Proof.* Consider a chain that starts in state $i$. Then, with probability $1 - f_i \geq 0$, the chain will never re-enter state $i$ again. The probability that the chain visits state $i$ $n$ more times is

$$P(n \text{ visits}) = f_i^n(1 - f_i).$$

where we recognize the pmf of a Geometric distribution. □

The expectation of the Geometric distribution is finite.

We end this section with a result we need only to prove later results.

**Theorem 2.2.** *State $i$ is recurrent if $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$ and transient if $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$.*

*Proof.* Let

$$I_n = \begin{cases} 1 & \text{if } X_n = i \\ 0 & \text{if } X_n \neq i \end{cases}$$

indicate whether the chain is in state $i$ at the $n$th timepoint. Then

$$\sum_{n=1}^{\infty} I_n$$

is the total number of visits to state $i$ after chain initiation. Take the expectation,

$$\begin{aligned} E\left[\sum_{n=1}^{\infty} I_n \mid X_0 = i\right] &= \sum_{n=1}^{\infty} E(I_n \mid X_0 = i) = \sum_{n=1}^{\infty} P(X_n = i \mid X_0 = i) \\ &= \sum_{n=1}^{\infty} p_{ii}^{(n)}. \end{aligned}$$

□

## 2.9 Ergodicity

We are getting close to the concept of ergodicity.

**Corollary 2.1.** *Recurrence is a class property: If state $i$ is recurrent and $j$ communicates with $i$, then $j$ is recurrent.*

*Proof.* Because $i$ and $j$ communicate, there exist $m$ and $n$ such that

$$\begin{aligned} p_{ij}^{(m)} &> 0, \text{ and} \\ p_{ji}^{(n)} &> 0. \end{aligned}$$

By Chapman-Kolmogorov,

$$p_{jj}^{(m+k+n)} \geq p_{ji}^{(n)} p_{ii}^{(k)} p_{ij}^{(m)}.$$

Sum over all possible $k$

$$\sum_{k=1}^{\infty} p_{jj}^{(m+k+n)} \geq p_{ji}^{(n)} p_{ij}^{(m)} \sum_{k=1}^{\infty} p_{ii}^{(k)} = \infty.$$

□

In the following, positive recurrence refers to a type of recurrence that I have dropped from the notes as not critical. I quote the adjective parenthetically so you can connect these results to broader Markov chain theory when you learn it.

**Lemma 2.4.** *Not all states can be transient in a finite-state Markov chain.*

*Proof.* Suppose there are $N$ states in the state space of a finite-state Markov chain. Let $N_i$ is the finite number of visits to state $0 \leq i \leq N-1$. Then after $\sum_{i=0}^{N-1} N_i$ steps in time, the chain will not be able to visit any state $i = 0, \ldots, N-1$, a contradiction. □

**Lemma 2.5.** *All states in a finite-state, irreducible Markov chain are (positive) recurrent.*

*Proof.* Because some states in a finite-state Markov chain must be recurrent, in fact *all* are recurrent since there is only one equivalence class in an irreducible Markov chain and recurrence is a class property. □

**Example 2.8** (transience)**.** *Determine the transient states in the following Markov matrix.*

$$\begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

**Example 2.9** (transience 2)**.** *Determine the transient, (positive) recurrent, and absorbing states in the following Markov matrix.*

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{2} \end{pmatrix}$$

**Example 2.10** (transience 3)**.** *Suppose the transition probability matrix were modified as*

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{2} \end{pmatrix}$$

**Definition 2.19** (periodicity)**.** *The period of state $i$ is the greatest common divisor of all $n$ such that $p_{ii}^{(n)} > 0$. In other words, if we consider all the times at which we could possibly be in state $i$, then the period is the greatest common divisor of all those times.*

- If the state $i$ can be revisited at any time, then the period is 1.
- If the state $i$ can be revisited every two time points, then the period is 2.
- If the state $i$ can never be revisited (i.e. diagonal entry in that $i$th row is 0 for all $P^n$), the the period is *defined* as 0.

**Definition 2.20** (aperiodic)**.** *A Markov chain is aperiodic if every state has period 0 or 1.*

**Lemma 2.6.** *Periodicity is a class property.*

*Proof.* Straightforward. □

**Example 2.11** (period 3)**.** *Confirm the period of the following chain is* 3*.*

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

**Example 2.12** (codons)**.** *Open reading frames encode amino acids via three-letter codes. Many changes at the third site are synonymous, which suggests that transitions out of the second position may be much more uniform than transitions out of the other two sites. To account for these structural differences in transition probabilities, some have used a periodic Markov chain with one-step transition matrices $P_1$ for transitioning from codon position* 1 *to codon position* 2*, $P_2$ for transitioning position* 2 *to* 3*, and $P_3$ from* 3 *back to* 1*. One can formulate this model by letting the state space be $\Omega_X = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} \times \{1, 2, 3\}$, a bivariate state, where the first variable is the nucleotide base and the second variable is the frame. With this state space, the elements of $P_1, P_2$, and $P_3$ can be arranged into a $12 \times 12$ matrix $P$ with a lot of* 0*'s for the position* 2 *that cannot transition to position* 1 *and so on.*

Finally, we are ready for ergodic states.

**Definition 2.21** (ergodic state)**.** *A state is ergodic if it is (positive) recurrent and aperiodic.*

**Lemma 2.7.** *Ergodicity is a class property.*

*Proof.* Recurrence and periodicity are class properties. □

**Definition 2.22** (ergodic chain)**.** *A Markov chain is ergodic if its states are aperiodic and (positive) recurrent.*

**Lemma 2.8.** *A finite state, irreducible, aperiodic Markov chain is ergodic.*

*Proof.* Follows from the last several results. □

Markov chains for genomic data are finite state. They may or may not be irreducible. They may or may not be aperiodic. Therefore, they may or may not be ergodic.

## 2.10 Stationary distribution

And finally we arrive to our goal. The concepts in this section are very important.

### Limiting distribution

Consider the one-step Markov chain transition probability matrix

$$P = \left\| \begin{matrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{matrix} \right\|$$

and examine the powers of the Markov matrix

$$\begin{aligned} P^2 &= \left\| \begin{matrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{matrix} \right\| \\ P^4 &= \left\| \begin{matrix} 0.5749 & 0.4281 \\ 0.5668 & 0.4332 \end{matrix} \right\| \\ P^8 &\approx \left\| \begin{matrix} 0.572 & 0.428 \\ 0.570 & 0.430 \end{matrix} \right\|. \end{aligned}$$

Indeed, it turns out that under certain conditions the $n$-step transition probabilities

$$p_{ij}^{(n)} \to \pi_j$$

approach a number, we'll call $\pi_j$, that is independent of the starting state $i$.

Another way to say this is that for $n$ sufficiently large, the probabilistic behavior of the chain becomes independent of the starting state, i.e.

$$P(X_n = j \mid X_0 = i) \approx P(X_n = j).$$

(caution: true only for $n$ large in certain chains)

**Theorem 2.3.** *For an irreducible, ergodic Markov chain, the limit* $\lim_{n\to\infty} p_{ij}^{(n)}$ *exists and is independent of* $i$*. Let*

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)},$$

*for all* $j \in \Omega_X$*. In addition, the* $\pi_j$ *are the unique, nonnegative solution of*

$$\begin{aligned} \pi_j &= \sum_{i\in\Omega_X} \pi_i p_{ij}, \textit{for all } j \\ \sum_{j\in\Omega_X} \pi_j &= 1 \end{aligned}$$

*Proof.* Please refer to Karlin and Taylor's *A First Course in Stochastic Processes* for a complete proof. □

**Pseudo-Proof**

The following proof (for finite state chains) demonstrates that when the limiting distribution exists, the required equations are satisfied.

Suppose that the limit mentioned in the above theorem exists for all $j$. By the law of total probability, we have

$$\begin{aligned} P(X_{n+1} = j) &= \sum_{i\in\Omega_X} P(X_{n+1} = j \mid X_n = i)P(X_n = i) \\ &= \sum_{i\in\Omega_X} p_{ij} P(X_n = i) \end{aligned}$$

Let $n \to \infty$ on both sides. If one can bring the limit inside the sum, then

$$\pi_j = \sum_{i\in\Omega_X} p_{ij}\pi_i,$$

which is the equation claimed in the theorem.

Finding the limiting distribution amounts to finding an eigenvector.

1. The equation $\pi_j = \sum_{i=0}^{\infty} \pi_i p_{ij}, \forall\, j$, can be written as a matrix equation for finite $\boldsymbol{\pi}^t = (\pi_0, \pi_1, \pi_2, \ldots, \pi_k)$

$$\boldsymbol{\pi}^t = \boldsymbol{\pi}^t \boldsymbol{P}.$$

2. Thus, $\boldsymbol{\pi}$ is the left eigenvector of $\boldsymbol{P}$ with eigenvalue 1.

3. If the Markov chain is irreducible and ergodic, then $\lambda_1 = 1$ is the maximum eigenvalue of $\boldsymbol{P}$ and $\lambda_1$ has multiplicity of 1. Thus $\boldsymbol{\pi}$ is the *first eigenvector* of $\boldsymbol{P}$. (Details not shown. Perron-Frobenius Theorem and lemmas thereof for *stochastic matrices*–those whose rows sum to 1.)

4. The *power iteration* method can be used to numerically compute the first eigenvector $\boldsymbol{\pi}$, even on very large, sparse matrices.

```r
require(gtools, quietly = T)
powerm = function(A, x0, thresh=1e-22)
{# ignore stackoverflow!
    x1 <- x0 %*% A       # MC: no normalization needed
    m1 <- (x1 %*% A %*% t(x1)) / (x1 %*% t(x1))
    if(abs(m1 - 1) < thresh) {   # MC: lambda1 is 1
        return(list(m1, t(x1)))
    } else {
        powerm(A, x1, thresh)
    }
}

A <- rdirichlet(4, alpha=rep(1,4))
pi.0 <- rdirichlet(1, alpha=rep(1,4))
ret = powerm(A, pi.0)
cat("Eigenvector :", ret[[2]], "\n")

## Eigenvector : 0.3157983 0.188925 0.1695012 0.3257755
```

**Definition 2.23** (stationary distribution)**.** *If there exist $\pi_j$ that satisfy $\pi_j = \sum_i p_{ij}\pi_i$ and $\sum_i \pi_i = 1$, then $\pi_j$ is called a stationary distribution. However, be clear that if $\lim_{n\to\infty} p_{ij}^{(n)} \neq \pi_j$, then it is not a limiting distribution.*

1. The *limiting distribution* **does not exist** for periodic chains.

2. The *stationary distribution* (but not the *limiting distribution*) **does exist** for irreducible, (positive) recurrent, periodic chains.

3. A *limiting distribution* is a *stationary distribution*.

An important property of the stationary distribution is given in the following lemma.

**Lemma 2.9.** *If the irreducible, ergodic chain is started with initial state distribution equal to the stationary distribution, then $P(X_n = j) = \pi_j$ for all future times $n$.*

*Proof.* We will do a proof by induction. Show true for $n = 1$.

$$P(X_1 = j) = \pi_j \quad = \quad \sum_i p_{ij}\pi_i \qquad \text{(by limiting distribution eq.).}$$

Assume it is true for $n - 1$, so $P(X_{n-1} = j) = \pi_j$.

Show true for $n$.

$$\begin{aligned} P(X_n = j) &= \sum_i P(X_n = j \mid X_{n-1} = i)P(X_{n-1} = i) \\ &= \sum_i p_{ij}\pi_i \qquad \text{(by induction hypothesis)} \\ &= \pi_j \qquad \text{(by limiting distribution eq.).} \end{aligned}$$

□

**Lemma 2.10.** *$\pi_j$ is the long-run proportion of time the process spends in state $j$.*

*Proof.* We will prove the result for aperiodic chains only. Recall that if a sequence of numbers $a_0, a_1, a_2, \ldots$ converges to $a$, then the sequence of partial averages $s_m = \frac{1}{m}\sum_{j=0}^{m-1} a_j$ also converges to $a$.

Consider the partial sums $\frac{1}{m}\sum_{k=0}^{m-1} p_{ij}^{(k)}$. In the limit, as $m \to \infty$, these partial sums converge to $\pi_j$. But recall

$$\begin{aligned} \sum_{k=0}^{m-1} p_{ij}^{(k)} &= \sum_{k=0}^{m-1} \mathbb{E}\left[1\{X_k = j\} \mid X_0 = i\right] \\ &= \mathbb{E}\left[\sum_{k=0}^{m-1} 1\{X_k = j\} \mid X_0 = i\right] \\ &= \mathbb{E}\left[\#\text{ timesteps spent in state } j\right]. \end{aligned}$$

Here, we have used $1\{X_k = j\}$ is the indicator function that is 1 when $X_k = j$ and 0 otherwise. Therefore, the partial sums created above converge to the proportion of time the chain spends in state $j$. □

**Example 2.13** (estimation under stationarity)**.** *We end by discussing the implications for estimation.*

*Suppose you observe realizations*

$$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} MC(\boldsymbol{P}, \boldsymbol{\alpha}),$$

*where the $i$th realization $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})$ is of length $n_i$. Further, assume $\boldsymbol{\alpha} = \boldsymbol{\pi}$. The log likelihood to maximize is*

$$\ln L\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \mid \boldsymbol{P}, \boldsymbol{\pi}(\boldsymbol{P})\right) \quad = \quad \sum_{i=1}^{n}\left[\ln \pi_{X_{i1}} + \sum_{l=2}^{n_i} \ln p_{X_{i,l-1},X_{i,l}}\right]$$

*Because $\boldsymbol{\pi}$ is an unknown function of $\boldsymbol{P}$, it is not tractable to find MLEs analytically.*

*However, $\sum_{i=1}^{n} \ln \pi_{X_{i1}}$ contributes relatively little to the log likelihood if $n_i$ are large, thus*

$$\begin{aligned}\hat{\boldsymbol{P}} &= \textit{argmax}_{\boldsymbol{P}} \ln L\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \mid \boldsymbol{P}, \boldsymbol{\pi}(\boldsymbol{P})\right) \\ &\approx \textit{argmax}_{\boldsymbol{P}} \sum_{i=1}^{n} \sum_{l=2}^{n_i} \ln p_{X_{i,l-1}, X_{i,l}},\end{aligned}$$

*What are $\hat{p}_{jk}$?*

*If we need to produce estimates of $\boldsymbol{\pi}$, we can use the functional invariance of MLE to find $\hat{\boldsymbol{\pi}}$ as the first eigenvector of $\hat{\boldsymbol{P}}$. (Technically, of course, this is not precisely the MLE of $\boldsymbol{\pi}$ because $\hat{\boldsymbol{P}}$ is not precisely the MLE of $\boldsymbol{P}$, but both should be close when $n_i$ are large for all $i = 1, 2, \ldots, n$.)*

*Or we can use the long-run proportion to produce Method of Moments (MOM) estimators for $\boldsymbol{\pi}$:*

$$\tilde{\pi}_j = \frac{\sum_{i=1}^{n} \sum_{l=1}^{n_i} 1\{X_{il} = j\}}{\sum_{i=1}^{n} n_i}.$$

*You can read about MOM estimators in the Stat 342 notes. In this case, both MLEs and MOMs are unbiased and both should be close to each other and the true $\boldsymbol{\pi}$ as sample size increases.*

## 2.11 More uses of Markov chains

### 2.11.1 Number of occurrences

**Example 2.14** (features in subgenomic samples)**.** *With modern NGS it is possible to collect reads from promoters or other subregions in a genome with particular characteristics in a tissue of interest. For example, you may identify actively transcribed regions, or regions of open chromatin, or regions bound to a known protein, or regions marked with certain histone markers, etc. After isolating these regions, you may annotate them with known transcription factor bindings sites, for example. Questions that arise in this context are*

- *Is there a particular feature over-represented in the sample?*
- *Is there a particular combination of features that is over-represented in the sample?*
- *Are features unusually constrained in their placement relative to each other?*

*The latter question may provide further indication that two features are physically interacting to carry out some function in your tissue of interest.*

We will address the first two of these questions by focusing on the random variable $Y$, the number of features in a sequence of length $N$. If we can derive the distribution of $Y$, then we can assess whether there is unusual prevalence of the feature in a particular sequence, and we can extend the analysis to consider whether there is unusual presence in a collection of sequences. Unfortunately, the distribution of $Y$ is not easy to obtain, so we will satisfy ourselves with obtaining $\mathbb{E}[Y]$ and Var$(Y)$. With these, we can at least perform a Wald test (sometimes).

In the era of big data, when we have many sequences each with occurrence data, there is a major advantage to having these two moments. There are generalizations of the CLT that yield

$$\frac{\sum_{i=1}^{n}(X_i-\mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}} \overset{\cdot}{\sim} \mathcal{N}(0,1),$$

for $\mathbb{E}[X_i]=\mu_i$ and $\text{Var}(X_i)=\sigma_i^2$. Now, suppose you observe a bunch of sequences $\boldsymbol{S}_1,\boldsymbol{S}_2,\ldots,\boldsymbol{S}_n$, and for each one, you can compute the motif occurrence rate $Y_i$ along with its mean $\mu_i$ and variance $\sigma_i^2$. Then almost regardless of the true distribution of $Y_i$ (there are some restrictions), one can obtain a sampling distribution under $H_0$ using the generalized CLT and the test statistic above. Since this test statistic is sensitive to over- or under-abundance of the motif (both tails indicate against $H_0$), it provides the basis of a useful hypothesis test.

Let $\boldsymbol{w}=(w_1,w_2,\ldots,w_k)$ be the feature of interest of length $k$, where $w_i\in\Omega$ are letters in some alphabet. For order zero (iid) and first order Markov chains, the alphabet $\Omega$ is also the state space $\Omega_X$. For order $o>1$ Markov chains, we will use the product space $\Omega_X^o$ formulation, so for $o=3$, pattern $\boldsymbol{w}$ =ATTA has nucleotide length $k=|\boldsymbol{w}|=4$ and state length $l=k-3+1=2$ with $\boldsymbol{w}_1=$ ATT and $\boldsymbol{w}_2=$ TTA. (Notice the bolded $\boldsymbol{w}_i\in\Omega_X^o$ in the $o>1$ order Markov chain is distinct from the unbolded single letters $w_i\in\Omega$.) We may at times write $w_1w_2w_3\cdots$ to represent the string obtained by concatenating some or all of the single letters in $\boldsymbol{w}$. For order $o>1$ Markov chains, we may also write the concatenation as $\boldsymbol{w}_1\boldsymbol{w}_2\cdots$, where we understand the overlapped portion is not duplicated, so for the example above $\boldsymbol{w}_1\boldsymbol{w}_2=$ ATTA *not* ATTTTA!

Define $\epsilon_j=\mathbb{1}\{w_1w_2\cdots w_j=w_{k-j+1}w_{k-j+2}\cdots w_k\}$ to indicate if the feature has a length $j$ overlap with itself. Let $\boldsymbol{X}=(X_1,X_2,\ldots,X_N),X_i\in\Omega_X$ be the random sequence where we find the feature(s) of interest; we could write $X_1X_2X_3\cdots$ for the sequence as a string. In the product state space $\Omega_X^o$ for an order $o>1$ Markov chain, we also have $\boldsymbol{X}=(\boldsymbol{X}_o,\boldsymbol{X}_{o+1},\ldots,\boldsymbol{X}_N),\boldsymbol{X}_i\in\Omega_X^o$, where $\boldsymbol{X}_o=X_1X_2\cdots X_o,\boldsymbol{X}_{o+1}=X_2X_3\cdots X_{o+1},\ldots$.

Define random variable

$$\begin{aligned} I_i &= \mathbb{1}\{X_{i-k+1}=w_1,X_{i-k+2}=w_2,\ldots,X_i=w_k\} \\ &= \mathbb{1}\{\boldsymbol{X}_{i-l+1}=\boldsymbol{w}_1,\boldsymbol{X}_{i-l+2}=\boldsymbol{w}_2,\ldots,\boldsymbol{X}_i=\boldsymbol{w}_l\}, \end{aligned}$$

first represented in the single letter alphabet, second in the state space representation, to indicate if there is a match of the pattern ending at site $i$ in the sequence. It should be clear that

$$Y=I_k+I_{k+1}+\cdots+I_N.$$

Therefore,

$$\mathbb{E}[Y]=\sum_{i=k}^{N}\Pr(\boldsymbol{X}_{i-l+1}=\boldsymbol{w}_1,\boldsymbol{X}_{i-l+2}=\boldsymbol{w}_2,\ldots,\boldsymbol{X}_i=\boldsymbol{w}_l).$$

If we assume stationarity, then

$$\pi(\boldsymbol{w}) := \Pr(\boldsymbol{X}_{i-l+1} = \boldsymbol{w}_1, \boldsymbol{X}_{i-l+2} = \boldsymbol{w}_2, \ldots, \boldsymbol{X}_i = \boldsymbol{w}_l)$$

does not depend on $i$. For example, for the first order Markov chain, we have

$$\pi(\boldsymbol{w}) = \pi_{w_1} p_{w_1 w_2} \cdots p_{w_{k-1} w_k},$$

the equilibrium frequency of the pattern $\boldsymbol{w}$. The iid model reaches equilibrium instantaneously, so

$$\pi(\boldsymbol{w}) = p_{w_1} p_{w_2} \cdots p_{w_k}.$$

For the second order Markov chain, we have

$$\pi(\boldsymbol{w}) = \pi_{\boldsymbol{w}_1} p_{\boldsymbol{w}_1 \boldsymbol{w}_2} \cdots p_{\boldsymbol{w}_{l-1} \boldsymbol{w}_l},$$

where, remember, $l = k - o + 1$. Importantly, since $\pi(\boldsymbol{w})$ does not depend on $i$,

$$\mathbb{E}[Y] = (N - k + 1)\pi(\boldsymbol{w}).$$

For the variance, recall

$$\text{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2,$$

so we need

$$\begin{aligned}
\mathbb{E}[Y^2] &= \mathbb{E}[(I_k + I_{k+1} + \cdots + I_N)^2] \\
&= \mathbb{E}[I_k^2 + I_{k+1}^2 + \cdots + I_N^2 + 2\sum_{i<m} I_i I_m] \\
&= \mathbb{E}[I_k^2 + I_{k+1}^2 + \cdots + I_N^2] \\
&\quad + 2(N-k)\,\mathbb{E}[I_i I_{i+1}] \\
&\quad + 2(N-k-1)\,\mathbb{E}[I_i I_{i+2}] \\
&\quad \vdots \\
&\quad + 2(N-2k+2)\,\mathbb{E}[I_i I_{i+k-1}] \\
&\quad + 2\sum_{i<m:m-i\geq k} (N-k-(m-i)+1)\,\mathbb{E}[I_i I_m],
\end{aligned}$$

To convince yourself of this formula, imagine placing pairs of patterns in the sequence with varying amount of overlap. There are, for example, $N - 2k + 2$ places you could place the pattern $w_1 w_2 \cdots w_k w_2 w_3 \cdots w_k$ (notice the 1-overlap). *At stationarity*, each one of those placements has the same expectation $\mathbb{E}[I_i I_{i+k-1}]$ for all the $N - 2k + 2$ possible values of $i$.

For an overlap of $j < k$, we have

$$\begin{aligned}
\mathbb{E}[I_i I_{i+k-j}] &= \epsilon_j \pi(w_1 w_2 \cdots w_k w_{j+1} w_{j+2} \cdots w_k) \\
&= \epsilon_j \begin{cases} \pi(\boldsymbol{w}_1 \boldsymbol{w}_2 \cdots \boldsymbol{w}_l \boldsymbol{w}_{j-o+2} \boldsymbol{w}_{j-o+3} \cdots \boldsymbol{w}_l) & j \geq o \\ \pi(\boldsymbol{w}_1 \boldsymbol{w}_2 \cdots \boldsymbol{w}_l) p^{(o-j)}_{\boldsymbol{w}_l \boldsymbol{w}_1} \pi(\boldsymbol{w}_2 \cdots \boldsymbol{w}_l \mid \boldsymbol{w}_1) & j < o \end{cases}
\end{aligned}$$

where the first version is written in terms of single letter alphabet and the second version is written in terms of states. For $m-i \geq k$, we have

$$\begin{aligned}\mathbb{E}[I_i I_m] &= \pi(\boldsymbol{w})P(X_{m-k+1}=w_1, X_{m-k+2}=w_2, \ldots, X_m = w_k \mid X_i = w_k, \\ &\quad X_{i-1}=w_{k-1}, \ldots, X_{i-o+1}=w_{k-o+1}) \\ &= \pi(\boldsymbol{w})P(\boldsymbol{W}_{m-l+1}=\boldsymbol{w}_1, \boldsymbol{W}_{m-l+2}=\boldsymbol{w}_2, \ldots, \boldsymbol{W}_m=\boldsymbol{w}_l \mid \boldsymbol{W}_i=\boldsymbol{w}_l),\end{aligned}$$

which involves a $m-i-k+\max\{1,o\}$-step transition probability $p_{\boldsymbol{w}_l\boldsymbol{w}_1}^{(m-i-k+\max\{1,o\})}$, so

$$\mathbb{E}[I_i I_m] = \pi(\boldsymbol{w})p_{\boldsymbol{w}_l\boldsymbol{w}_1}^{(m-i-k+\max\{1,o\})}\pi(\boldsymbol{w}_2\boldsymbol{w}_3\cdots\boldsymbol{w}_l \mid \boldsymbol{w}_1),$$

where $\pi(\boldsymbol{w}_2\cdots\boldsymbol{w}_l \mid \boldsymbol{w}_1) = p_{\boldsymbol{w}_1\boldsymbol{w}_2}p_{\boldsymbol{w}_2\boldsymbol{w}_3}\cdots p_{\boldsymbol{w}_{l-1}\boldsymbol{w}_l}$. For the iid model, $\mathbb{E}[I_i I_m] = [\pi(\boldsymbol{w})]^2$ since two patterns are independent.

Finally, given $o < k$, Var$(Y)$ is

$$\begin{aligned}&(N-k+1)\pi(\boldsymbol{w}) - (N-k+1)^2\left[\pi(\boldsymbol{w})\right]^2. \\ &+\mathbb{1}\{o>1\}2\sum_{j=1}^{o-1}(N-2k+j+1)\epsilon_j\pi(\boldsymbol{w}_1\boldsymbol{w}_2\cdots\boldsymbol{w}_l)\left[\prod_{l=1}^{o-j}p_{\boldsymbol{w}_l\oplus_{l-1}\boldsymbol{w}_1,\boldsymbol{w}_l\oplus_l\boldsymbol{w}_1}\right] \\ &\pi(\boldsymbol{w}_2\cdots\boldsymbol{w}_l \mid \boldsymbol{w}_1) \\ &+2\sum_{j=\max\{1,o\}}^{k-1}(N-2k+j+1)\epsilon_j\pi(\boldsymbol{w}_1\boldsymbol{w}_2\cdots\boldsymbol{w}_l\boldsymbol{w}_{j-\max\{1,o\}+2}\boldsymbol{w}_{j-\max\{1,o\}+3}\cdots\boldsymbol{w}_l) \\ &+2\sum_{j=k}^{N-k}(N-k-j+1)\pi(\boldsymbol{w})p_{\boldsymbol{w}_l\boldsymbol{w}_1}^{(j-k+\max\{1,o\})}\pi(\boldsymbol{w}_2\boldsymbol{w}_3\cdots\boldsymbol{w}_l \mid \boldsymbol{w}_1),\end{aligned}$$

where $\boldsymbol{w}_1 \oplus_l \boldsymbol{w}_2$ is the concatenation operator that takes the $(o-l)$-suffix of $\boldsymbol{w}_1$ and prepends it to the $l$-prefix of $\boldsymbol{w}_2$ to create a $o$mer.

Under the iid model, we can drop the conditions and simplify to

$$\begin{aligned}\text{Var}(Y) &= (N-k+1)\pi(\boldsymbol{w}) \\ &\quad+\left[3k^2-4k+1-(2k-1)N\right]\left[\pi(\boldsymbol{w})\right]^2 \\ &\quad+2\sum_{j=1}^{k-1}(N-2k+j+1)\epsilon_j\pi(w_1w_2\cdots w_kw_{j+1}w_{j+2}\cdots w_k).\end{aligned}$$

If there are no overlaps or we do not count overlapped patterns, then the sum is dropped as well. Then, if $\pi(\boldsymbol{w})$ is very small, then the first term is substantially larger than the second term and Var$(Y) \approx \mathbb{E}[Y]$, which indicates $Y$ is nearly Poisson distributed. However, it is not generally true, not for short patterns and not for patterns with overlaps when overlapping matches are counted.

Finally, we can use Wald test to check if the observed number of occurrences $y$ is unusual under the sequence model (iid or Markov chain) using test statistic

$$\frac{(y-\mathbb{E}[Y])^2}{\text{Var}(Y)} \sim \chi^2(1).$$

This statistic has a $\chi^2(1)$ distribution if $y$ is the MLE of $\mathbb{E}[Y]$. Or it is true if $Y$ is approximately normally distributed by some other argument. There is in fact work to show that $Y$ is approximately normal if $\mathbb{E}[Y]$ is large (say, 500 or more). In contrast, when $k \geq 10$ and $\mathbb{E}[Y]$ is not large, the normal approximate is poor, so the Wald test is not reliable. There is additional theoretical work, but we'll drop it here. You resort to Monte Carlo simulation if you need a test.

It is not clear if we have gained much by using theory rather than Monte Carlo simulation to approach this problem for a single sequence. With Monte Carlo simulation, we can estimate the pmf of $Y$. With theory, we stopped short. With theory, we can get the exact expectation and variance, but then we have to rely on the asymptotics of the Wald test. Which cost us more in accuracy? The asymptotics? The Monte Carlo approximation? It probably depends. On $k$, $N$, the pattern $\boldsymbol{w}$, and the model properties. Which costs us more in time? The theory if we only use the results once. The Monte Carlo if we release it as software to run tests and compute $p$-values for the masses.

**Example 2.15** (Occurrence rates)**.** *Let's do the calculations for* ATA *in a sequence of length* $N = 100$ *under a first order Markov chain.*

```
# Function: return given order Markov chain states from alphabet letters
power.states <- function(states, order=1) {
   state.names <- states
   if (order > 1) {
      for (k in 2:order) {
         state.names <- paste(rep(c(’A’,’C’,’G’,’T’),
            each=length(state.names)),
            rep(state.names, times=4), sep="")
      }
   }
   state.names
}# power.states

# Function: compute E[Y] and Var(Y) for arbitrary order MC
pattern.moments <- function(P, w, N=1000, pi=NULL) {
   order <- log(nrow(P))/log(4)
   w.single <- strsplit(w, ’’)[[1]]
   k <- nchar(w)        # length of pattern
   w <- substring(w, first=0:(nchar(w)-max(order,1))+1,
         last=max(1,order):nchar(w))
   l <- length(w)       # no. states in pattern
   if (l < 2) stop("Does not work for order >= pattern length")
   state.names <- power.states(c(’A’,’C’,’G’,’T’), order=order)
   if (is.null(pi) & order)   # stationary distribution
      pi <- as.vector(powerm(P, rdirichlet(1, rep(1, 4^order)))[[2]])
   else if (is.null(pi)) {
      pi <- P[1,]
      P <- matrix(P, nrow=4, ncol=4, byrow=T)
   }
   rownames(P) <- colnames(P) <- names(pi) <- state.names
   p.AA <- NULL
   P.A <- as.matrix(P[,w[1]])    # A column of P
   for (n in 1:(N-k)) {
      p.AA[n] <- P.A[w[l],]      # p_{AA}^{(n)}
      P.A <- P %*% P.A      # A column of P^(n+1)
   }
   pi.w <- pi[w[1]]     # stationary prob of pattern w
   for (i in 2:l) pi.w <- pi.w * P[w[i-1], w[i]]
   epsilon <- NULL      # overlap indicator
   pi.w.j <- c(rep(1, k-2), P[w[l-1],w[l]])     # pi(w[(j+1):k]|w[1:j])
   if (order > 1)       # prob trans w[l] to w[1] with overlap
      pi.w.1j <- rep(P[w[l], paste(substr(w[l], 2, order),
         substr(w[1], 1, 1), sep="")], order-1)
   for (i in 2:k) {     # overlap + 1
      if (k - i >= order & i < k)
         pi.w.j[k-i] <- pi.w.j[k-i+1]*P[w[l-i],w[l-i+1]]
      if (order > 1 & i < order)
         pi.w.1j[i] <- pi.w.1j[i-1]*P[paste(substr(w[l], i, order),
            substr(w[1], 1, i-1), sep=""), paste(substr(w[l], i+1, order),
             substr(w[1], 1, i), , sep="")]
      epsilon[i-1] <- sum(w.single[(k-i+2):k] == w.single[1:(i-1)]) == (i-1)
   }
   E.Y <- (N-k+1)*pi.w; var.Y <- (N-k+1)*pi.w - E.Y^2
   if (order > 1)
      for (j in 1:(order-1))
         var.Y <- var.Y + 2*(N-2*k+j+1)*epsilon[j]*pi.w*pi.w.1j[order-j]*
            pi.w/pi[w[1]]
   if (order < k)
      for (j in max(order,1):(k-1))
         var.Y <- var.Y + 2*(N-2*k+j+1)*epsilon[j]*pi.w*pi.w.j[j]
   for (j in k:(N-k))
      var.Y <- var.Y + 2*(N-k-j+1)*pi.w*p.AA[j-k+max(1,order)]*pi.w/pi[w[1]]
   return(list("E.Y"=E.Y, "Var.Y"=var.Y, "pi"=pi))
}# pattern.moments

set.seed(876)
require(gtools)
N <- 100                                # sequence of length N
P <- rdirichlet(4, alpha=rep(1,4))      # simulate first order P
ATA.moments <- pattern.moments(P, w = "ATA", N = N)
```

*The selected transition probability matrix and stationary (and limiting) distribution are:*

```
rownames(P) <- colnames(P) <- c('A','C','G','T')
P

##            A         C          G          T
## A 0.28489546 0.5469891 0.01389402 0.15422142
## C 0.49482659 0.1546800 0.30658557 0.04390784
## G 0.14204092 0.5383302 0.29325587 0.02637298
## T 0.02087266 0.3765329 0.05042462 0.55216984

ATA.moments$pi

##         A         C         G         T
## 0.2982236 0.3734116 0.1785385 0.1498263
```

*The expectation is 0.0940785 with standard deviation 0.3069925 (variance 0.0942444). If we observed $y = 1$ occurrence in a sequence of length $N = 100$, then the probability of observed $Y \geq 1$ (the $p$-value from Wald's test) is*

```
y <- 1
pchisq((ATA.moments$E.Y - y)^2/ATA.moments$Var.Y, df=1, lower.tail=F)

##           A
## 0.003167914
```

*This is an approximate test of*

$H_0$ : $\boldsymbol{S} \sim MC(\boldsymbol{P}, \boldsymbol{\pi})$

*“sequence $\boldsymbol{S}$ is generated by the stationary Markov chain with transition probability matrix $\boldsymbol{P}$”*

$H_1$ : *“$\boldsymbol{S}$ generated with an excess of pattern* ATA*”,*

*and an exact test of*

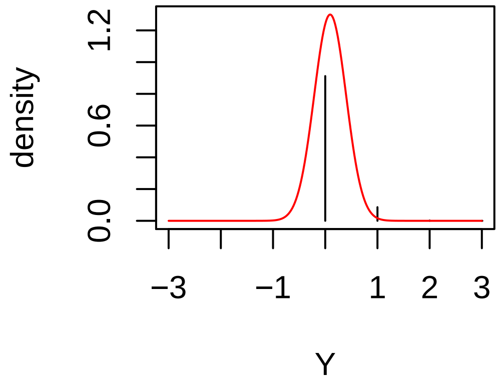
$$\begin{aligned} H_0 &: Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right) \\ H_1 &: \mathbb{E}[Y] > \mu_Y, \end{aligned}$$

*where $\mu_Y$ and $\sigma_Y^2$ are the mean and variance computed by the presented theory. The conditions may not be consistent with normality ($\mathbb{E}[Y]$ is too small), so let’s check the Monte Carlo-estimated* p*-value.*

```
# Markov chain simulator [slow]
sim.mc <- function(P, pi = NULL, len=1000) {
        if (is.null(pi))
                pi <- as.vector(powerm(P, rdirichlet(1, rep(1,4)))[[2]])
        nucs <- c(sample(4, size=1, prob=pi))
        for (i in 2:len) {
                nucs[i] <- sample(4, size=1, prob=P[nucs[i-1],])
        }
        nucs
}# sim.mc
# Count patterns (always allows overlap now) [slow]
cnt.pattern <- function(seq, w = c(1,4,1), allowoverlap = T) {
        cnt <- 0
        k <- length(w)
        for (i in k:length(seq)) {
                if (sum(seq[(i-k+1):i] == w) == k) cnt <- cnt + 1
        }
        cnt
}
```

```
B <- 1000
pmf.ATA <- rep(0, N+1)
for (b in 1:B) {
        seq <- sim.mc(P, pi = ATA.moments$pi, len = N)
        cnt <- cnt.pattern(seq, w = c(1,4,1))
        pmf.ATA[cnt + 1] <- pmf.ATA[cnt + 1] + 1
}
```

```
t <- seq(from=-3, to=3, length.out=1000)
plot(0:N, pmf.ATA/B, xlim=c(-3,3), xlab="Y", ylab="density", type="h",
   ylim=c(0, max(dnorm(t, mean=ATA.moments$E.Y, sd=sqrt(ATA.moments$Var.Y)))))
lines(t, dnorm(t, mean=ATA.moments$E.Y, sd=sqrt(ATA.moments$Var.Y)),
   col="red")
```



*The approximating normal distribution (red curve) certainly does not fit the Monte Carlo-simulated discrete $Y$ distribution well (black lines). Here, the probability of $\geq 1$ occurrences is 0.088, which is no longer significant at the traditional level.*

*The pattern* ACA *with the same transition probability matrix works better because it is much more likely.*

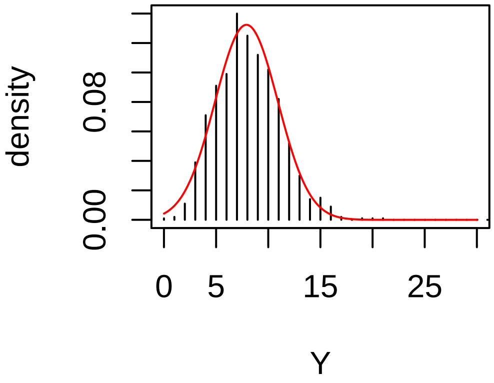
```
ACA.moments <- pattern.moments(P, w = "ACA", N = N)
y <- 22
pchisq((ACA.moments$E.Y - y)^2/ACA.moments$Var.Y, df=1, lower.tail=F)

##             A
## 2.938845e-06
```

*And Monte Carlo:*

```
pmf.ACA <- rep(0, N+1)
for (b in 1:B) {
        seq <- sim.mc(P, pi = ACA.moments$pi, len = N)
        cnt <- cnt.pattern(seq, w = c(1,2,1))
        pmf.ACA[cnt + 1] <- pmf.ACA[cnt + 1] + 1
}
```

```
plot(0:N, pmf.ACA/B, xlim=c(0,30), xlab="Y", ylab="density", type="h")
x <- seq(from=0, to=30, length.out=1000)
lines(x, dnorm(x, mean=ACA.moments$E.Y, sd=sqrt(ACA.moments$Var.Y)),
   col="red")
```



*That looks considerably more normal, and the probability of* $22$ *or more occurrences is* $0$*, similar to the theoretical estimate.*

### 2.11.2 Distance between occurrences

Specialized Markov chains can be constructed to answer many important questions in bioinformatics and beyond. One technique that is useful for solving difficult calculations is first step analysis, an application of the versatile Law of Total Probability, specifically a corollary called the Law of Total Expectation. For two random variables $X$ and $Y$ (there are some conditions), the Law of Total Expectation is

$$\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X \mid Y)]$$

We will show a proof below for the case where we need it. In particular, first step analysis can be used to compute the expectation of the number of nucleotides until the first pattern or between pattern occurrences, one of the motivating questions for this section.

**Example 2.16** (ATA)**.** *The goal of this example is to determine the expected number of nucleotides before the first occurrence of ATA or between occurrences of ATA in a genome. This calculation is useful to predict the expected fragment length for a restriction enzyme that cuts at pattern ATA. (There will be other uses once we can also compute the variance in the number of nucleotides before the first occurrence.)*

*Let us assume that the genome in question is well-modeled as iid nucleotides with proportions* $\{p_A, p_C, p_G, p_T\}$*. While this model is silly for biological sequences,it is merely a convenient assumption to reduce the number of parameters in our first attempt at first step analysis.*

*To answer this question, we construct a specialized Markov chain on state space* $\Omega_X = \{\overline{\mathrm{A}}, \mathrm{A}, \mathrm{AT}, \mathrm{ATA}\}$*, where* $\overline{\mathrm{A}} = \{\mathrm{C}, \mathrm{G}, \mathrm{T}\}$*. The states AT and ATA are special states that retain enough history to indicate the chain is on its way to possibly matching the target sequence ATA. State AT is entered if the current state is A and the next state is T; state ATA is entered if the current state is AT, and the next state is A, i.e. the target pattern occurs. To specify the Markov chain, we need an initial state distribution on the state space* $\Omega_X$ *and a transition probability matrix*

$$\boldsymbol{P} = \begin{pmatrix} 1-p_A & p_A & 0 & 0 \\ p_C + p_G & p_A & p_T & 0 \\ 1-p_A & 0 & 0 & p_A \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

*We will defer our choice of initial state distribution to later. The choice for the last row of* $P$ *makes* ATA *an absorbing state, which we will define later. Since we end up dropping the last row in our calculations, it actually does not matter how we define the last row.*

*Now, we observe Markov chain $\{X_n : n \geq 0\}$. Let $T_n$ be the time (number of nucleotides) until the end of the first ATA in the sequence at or after position $n$. If $X_n = \text{ATA}$, then $T_n = 0$. The wait time (number of nucleotides) before the next occurrence of* ATA *after $n-1$ is $W_{n-1} = 1 + T_n$. The added 1 is the (time) cost of generating at least one nucleotide at position $n$ to get a match at or after site $n$.*

*Ideally, we want the distribution of $W_{n-1}$ under a reasonable $H_0$, say the iid model or more plausibly MC(1). Then, we could assess if observed wait times were unusual. However, our first goal is more modest. We want to compute $\mathbb{E}[W_{n-1}] = 1 + \mathbb{E}[T_n]$, the expected wait time until the next occurrence after site $n-1$.*

*It is not easy to compute the unconditional expectation $\mathbb{E}[T_n]$, but conditioning can help. The Law of Total Expectation yields*

$$\mathbb{E}[T_n] = \mathbb{E}[\mathbb{E}(T_n \mid X_n)]$$

*and applied again*

$$\begin{aligned} \mathbb{E}[T_n \mid X_n] &= \mathbb{E}[\mathbb{E}(T_n \mid X_n, X_{n+1}) \mid X_n] \\ &= \sum_{x \in \Omega_X} p_{X_n x}\, \mathbb{E}[T_n \mid X_n, X_{n+1} = x]. \end{aligned} \tag{2.2}$$

*The proof of the first result (and by easy implication the second) is*

$$\begin{aligned} \mathbb{E}[\mathbb{E}(T_n \mid X_n)] &= \textstyle\sum_{x\in\Omega_X} \mathbb{E}(T_n = t \mid X_n = x) P(X_n = x) && \textit{defn of expectation} \\ &= \textstyle\sum_{x\in\Omega_X} \sum_{t=0}^{\infty} t P(T_n = t \mid X_n = x) P(X_n = x) && \textit{defn of expectation} \\ &= \textstyle\sum_{x\in\Omega_X} \sum_{t=0}^{\infty} t P(T_n = t, X_n = x) && \textit{conditional probability} \\ &= \textstyle\sum_{t=0}^{\infty} t \sum_{x\in\Omega_X} P(T_n = t, X_n = x) && \textit{rearrange sums} \\ &= \textstyle\sum_{t=0}^{\infty} t P(T_n = t) && \textit{marginal probability} \\ &= \mathbb{E}[T_n]. \end{aligned}$$

*We make use of the Markov property by noting that $T_n = T_{n+1} + 1$ if $X_n \neq$* ATA*. Thus,*

$$\begin{aligned} \mathbb{E}[T_n \mid X_n, X_{n+1} = x] &= \begin{cases} 0 & X_n = \text{ATA} \\ \mathbb{E}[T_{n+1} + 1 \mid X_{n+1} = x] & X_n \neq \text{ATA}. \end{cases} \\ &= \begin{cases} 0 & X_n = \text{ATA} \\ 1 + \mathbb{E}[T_{n+1} \mid X_{n+1} = x] & X_n \neq \text{ATA}. \end{cases} \end{aligned}$$

*In addition, if the Markov chain is time homogeneous, then*

$$\mathbb{E}[T_n \mid X_n = x] = \mathbb{E}[T_{n-1} \mid X_{n-1} = x] = \cdots = \mathbb{E}[T_0 \mid X_0 = x] := \mu_x$$

*for all $n \geq 0$. Using this new notation, Eq.* (2.2) *expands into four equations, one for every possible current state $X_n$ in $\Omega_X$.*

$$\begin{aligned} \mu_{\overline{\text{A}}} &= 1 + (1 - p_A)\mu_{\overline{\text{A}}} + p_A \mu_{\text{A}} \\ \mu_{\text{A}} &= 1 + (p_C + p_G)\mu_{\overline{\text{A}}} + p_A \mu_{\text{A}} + p_T \mu_{\text{AT}} \\ \mu_{\text{AT}} &= 1 + (1 - p_A)\mu_{\overline{\text{A}}} + p_A \mu_{\text{ATA}} \\ \mu_{\text{ATA}} &= 0. \end{aligned} \tag{2.3}$$

*Excluding the last equation, this system of equations can be written in matrix notation as*

$$\boldsymbol{\mu} = \boldsymbol{Q}\boldsymbol{\mu} + \mathbf{1},$$

*where $\boldsymbol{\mu} = (\mu_{\overline{\text{A}}}, \mu_{\text{A}}, \mu_{\text{AT}})'$ and $\boldsymbol{Q} = \{\boldsymbol{P}\}_{i,j=1}^{|\Omega_X|-1}$ excludes the absorbing state(s) from $\boldsymbol{P}$. The solution is*

$$\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{Q})^{-1}\mathbf{1},$$

*which can be easily found in* `R`*.*

```
require(gtools)
p <- rdirichlet(1, alpha=rep(2, 4))      # generate nucleotide proportions
cat("p[A] =", p[1], "p[C] =", p[2], "p[G] =", p[3], "p[T] =", p[4], "\n")

## p[A] = 0.2132227 p[C] = 0.1337596 p[G] = 0.3673962 p[T] = 0.2856215

Q <- matrix(c(1 - p[1], p[1], 0, sum(p[2:3]), p[1], p[4], 1 - p[1],
        0, 0), nrow=3, byrow=T)
colnames(Q) <- rownames(Q) <- c("notA", "A", "AT")
print(Q)

##           notA         A        AT
## notA 0.7867773 0.2132227 0.0000000
## A    0.5011558 0.2132227 0.2856215
## AT   0.7867773 0.0000000 0.0000000

mu <- solve(diag(rep(1,3)) - Q, rep(1, 3))
cat(mu, "\n")

## 81.69907 77.00914 65.27898
```

*Finally, we return to the question of the initial state distribution. If we are starting at the beginning of a sequence, then the first state $X_0$ cannot be* AT *nor* ATA *because there is no* A *nor* AT *to build on. Thus, the initial state distribution is $\alpha_{\mathrm{A}} = p_A, \alpha_{\overline{\mathrm{A}}} = 1 - p_A$ and the expected wait time until the first occurrence is*

$$\mathbb{E}[W_{-1}] = 1 + \mathbb{E}[T_0] = 1 + (1 - p_A)\mu_{\overline{\mathrm{A}}} + p_A\mu_A.$$

```
1 + sum(c(p[1], 1 - p[1]) * mu[1:2])

## [1] 79.00914
```

*The same result would apply if the current state was the last matching ATA and we wanted to know the expected location of the next ATA that did not overlap with the previous one.*

*If instead we are looking for the next occurrence of ATA, including ATA overlapped with itself (e.g. there are three occurrences of ATA in ATATATA), then having just encountered ATA at position $n-1$, the initial state is either AT with probability $p_T$, A with probability $p_A$ or $\overline{\mathrm{A}}$ with probability $p_C + p_G$. Therefore, the expected number of nucleotides between overlapping occurrences of ATA is*

$$\mathbb{E}[W_n \mid X_n = \mathrm{ATA}] = 1 + \mathbb{E}[T_{n+1} \mid X_n = \mathrm{ATA}] = 1 + (p_C + p_G)\mu_{\overline{\mathrm{A}}} + p_A\mu_{\mathrm{A}} + p_T\mu_{\mathrm{AT}}.$$

```
1 + sum(c(sum(p[2:3]),p[1],p[4]) * mu[1:3])

## [1] 77.00914
```

*If there is no overlap in the pattern, e.g.* ATG*, then there is no difference between the overlapped and non-overlapped wait times.*

*Clearly, the easiest way to solve these problems is to skip the formal application of the Law of Total Expectation and simply construct the matrix $\boldsymbol{Q}$.*

- If you used a first order Markov chain instead of the iid model, how would the derivations of first step analysis change?
- When (if ever) does it makes sense to use a Markov chain of order $o \geq k$, the length of the pattern?
- How do we compute the variance of $W_n$?