Probable Cause: Modeling with Markov Chains

By Angela B. Shiflet and George W. Shiflet
Wofford College, Spartanburg, South Carolina

NOTE: An effective introduction to Markov chains and some of their computational science applications can be accomplished by covering the first two parts, "Scientific Questions" and "Computational Models," and omitting the third on "Bioinformatics and Markov Chains." Exercise 1 and Projects 1-5 are accessible with only this background. The third part provides the necessary background for Exercises 2 and 3 and Projects 6-10. The sections entitled "The Area" and "High Performance Computing and Bioinformatics" of that part can be covered on their own as an overview of the need for high performance computing in this important new area of biology and computational science.

1. Scientific Questions

Introduction

To the U.S. Navy and shipping companies around the world, barnacles can be a real drag, and they are out to get rid of them. How can such a small animal be so despised by so many? Though seemingly insignificant, they are one of the main causes of fouling of ship hulls. Growing on the submerged hull surfaces, they interfere with the smooth movement of ships through the water. More fuel must be used to drive the ship, and this adds up to tremendous costs. Millions of dollars have been expended to find ways to eliminate or at least greatly inhibit attachment. Various types of paints have been tried, but many of them leach toxic compounds into the water. Recently, researchers have developed some non-toxic coatings, which help to change the mechanical properties of the hull surface, so that barnacle larvae and other fouling organisms are less likely to attach.

Incidentally, barnacles also help to foul intake pipes for coastal power stations. So, finding an effective, nontoxic method to prevent such fouling would be a significant benefit to human populations.

As adults, barnacles are mostly small, sessile animals—they remain attached to firm surfaces. They adapted to various naturally occurring surfaces before human beings began exploring and harvesting the seas. The 900 or so species can be found on whale skin, crab and mollusk shells and on rocky shores. You may have seen them as you explored a rocky beach or examined a seashell that had washed up on a sandy beach. They are prominent members of a community of organisms that call the intertidal zone home.

Intertidal regions, which lie between high and low tide lines, represent a transition between the marine and terrestrial ecosystems. Although these regions include sand beaches, estuaries, and bays, barnacles particularly like rocky shores. In fact, rocky intertidal areas include very dense and diverse communities, highly adapted to the periodic exposure to drying, wave action, and extremes of temperature. The organisms of this habitat are often found in distinct, vertical zones, arranged according to degree of

exposure—low, middle, high intertidal, and splash zones. The width of each zone is determined somewhat by the degree of protection from wave action—narrower in more protected areas.

So, barnacles are important members of these communities, which are rich in numbers of taxa. Like their neighbors in this zone, barnacles must live under some fairly extreme physical conditions (e.g., heavy wave action, desiccation, high temperature), while trying to supply themselves with sufficient food and available oxygen, overcoming competition and predation, and producing gametes for reproduction. They are unable to control the physical environment, and none of the biological challenges is easily met. Any additional physical or biological stress would put even organisms as hardy as barnacles in jeopardy.

What if environmental conditions changed so that a barnacle species disappears? Many scientists suggest that the world oceans are warming. What effects might ocean temperature change have on intertidal communities? Well, increasing temperature would add to the often-extreme temperatures that these organisms already have to endure, and they might not be able to withstand them. Temperature cues are also important for development and reproduction of many animals. From 1993-1996, researchers at Hopkins Marine Station in California surveyed transects in a rocky intertidal community that was first surveyed in the 1930's. They found a dramatic shift in species, where southern species (warm-adapted) increased significantly over northern species (cold-adapted), during a time period where ocean and summer air temperatures had both increased over the 60-year span of time. What this study suggests is that such a change can eliminate some species from the community—perhaps a barnacle species. So what? (California 1987; Foster 2009. Intertidal 2007)

Barnacles are filter feeders that occur in large numbers in their communities. They form hiding places for small animals, and they serve as food for others. Their role or niche is interwoven into the community structure and function, and their loss might have serious ramifications. Each species is integrated so that it has multiple interactions with other community members and the environment. The extinction of a barnacle species would certainly affect other constituents of the ecosystem (Barnacles 2012; Barnacles 2003; Secret Life 2012; Stout 2009).

Understanding the effects of losses in diversity is and will continue to be critical to the implementation of judicious conservation policies, but that understanding is problematic in the multifarious, natural ecosystems. Mathematical models may offer us an effective approach to estimating the impact of species losses to a community. For this type of study, we can employ **Markov chain models** (**MCM**), which are based on the probability of passing from one state to another. Normally, the parameters of these models depend on the observed and experimental data available, but MCMs allow us to utilize parameters without extensive experimentation.

Problems from Psychology to Genetics

Besides predicting effects of species loss to a community, Markov chain models are useful in quite a variety of problems from predicting the behavior of animals to locating genes in the DNA. In this module, we start with a problem from psychology in which we have observed the various activities of an animal and the likelihood of moving from one pursuit to another. Using this information, with MCMs we can estimate the average

amount of time a typical animal spends performing each endeavor and, given the activities of a group of animals, to predict their behavior in the near future.

Employing the same modeling technique we can pursue vastly different problems in genetics. One such problem involves locating genes in DNA. Such determination can lead to targeted drug therapies and greater understanding of genetic diseases. Some of the gene-finding programs the scientists employ have Markov chain models at their core.

2. Computational Models

Probability

Markov chain models involve matrices in which all the elements are probabilities, so we start with a brief introduction to probability theory. The **probability** of an event, or the occurrence of something, is a number between 0 and 1, inclusively, indicating the chance of the event happening. A probability of 0 means that the event can never occur, while 1 says that that the situation is always true. As an example, suppose a certain kind of seed has a 50-50 chance of germinating. Thus, the probability or chance of germinating is $P(\text{germinating}) = ½ = 0.5 = 50\%$. For each seed, one of two events can occur, germination or no germination; and the results are equally likely to occur. We expect that if we observe many seeds, about half the seeds will germinate.

Definition The **probability** of an event, E, written $\boldsymbol{P(E)}$, is the chance of its occurrence and is a number between 0 and 1, inclusively.

Quick Review Question 1 Suppose at a site on a strand of DNA, an equal likelihood exists for any of the four bases (A, C, T, G). Give the probability of the base T occurring at a particular site.

The sum of all the possible events for a situation, such as germinating and not germinating, sums to 1. If a seed has only a 30% chance of germinating, $P(\text{germinating}) = 0.3$, then it has a 70% chance of not germinating. $P(\text{not germinating}) = 1 - P(\text{germinating}) = 1 - 0.3 = 0.7$.

Rule The probability of an event not occurring is 1 minus the probability of the event,

$$\boldsymbol{P(\text{not } E) = 1 - P(E)}$$

Quick Review Question 2 Suppose at a site on a strand of DNA, an equal likelihood exists for any of the four bases (A, C, T, G). Give the probability of T *not* being at a particular site.

Suppose an ant is equally likely to go in any one of eight directions, N, NE, E, SE, S, SW, W, NW. For example, $P(\text{N}) = 1/8$ and $P(\text{S}) = 1/8$. The probability that the ant will move in the north or south direction is $P(\text{N or S}) = P(\text{N}) + P(\text{S}) = 1/8 + 1/8 = 1/4$.

The ant cannot move in two directions at the same time, so moving to the north and moving to the south are **mutually exclusive**; the events cannot occur at the same time. If events E_1 and E_2 are mutually exclusive, then the probability of E_1 *or* E_2 is the sum of the probabilities of the individual events, $P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)$.

Rule If events E_1 and E_2 are **mutually exclusive**, and, thus, cannot occur at the same time, then the probability of E_1 or E_2 is the sum of the probabilities of the individual events,

$$\boldsymbol{P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)}$$

Rule If $E_1, E_2, \ldots, E_n$ are **all possible mutually exclusive events** for a situation so that no two of the events cannot occur at the same time, then

$$\boldsymbol{P(E_1) + P(E_2) + \cdots + P(E_n) = 1}$$

Quick Review Question 3 Suppose at a site on a strand of DNA, an equal likelihood exists for any base. Give the probability of a site containing A or T.

To calculate the probability that the ant will go in a northerly (N, NE, NW) or westerly (W, NW, SW) direction, we must subtract the probability of where the events overlap, going NW, as follows:

$$P(\text{northerly or westerly}) = P(\{\text{N, NE, NW}\}) + P(\{\text{W, NW, SW}\}) - P(\text{NW})$$
$$= 3/8 + 3/8 - 1/8 = 5/8$$

We must subtract P(NW) to avoid counting that direction twice. The two events, heading in a northerly direction and heading in a westerly direction, are not mutually exclusive. If events are not mutually exclusive, then for the probability of one or the other we must subtract the probability of overlap from the sum of the probabilities.

Rule $\boldsymbol{P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)}$

Quick Review Question 4 Suppose a certain medicine causes nausea in one out of every ten patients. On the average, 4% of those taking the drug experience diarrhea. The probability of a patient who is using the drug experiencing nausea and diarrhea is 0.01. Give the probability that a patient taking the drug has nausea or diarrhea.

Considering again the seeds that have a 30% chance of germinating, suppose we have two seeds, S_1 and S_2. Each has 0.3 probability of germinating, and the state of one seed has no bearing on the state of the other. We say these events are **independent**. Certainly, the probability of both seeds germinating is even less likely than any one germinating. In fact, the probability of S_1 germinating *and* S_2 germinating is the product of their individual probabilities:

$$P(S_1 \text{ germinating and } S_2 \text{ germinating}) = P(S_1 \text{ germinating}) \cdot P(S_2 \text{ germinating})$$
$$= (0.3)(0.3) = 0.09$$

Only a 9% chance exists of both seeds germinating.

Definition Events are **independent** if the occurrence of one event has no impact on the occurrence of the other.

Rule For **independent events** E_1 and E_2, the probability of both events occurring is the product of their individual probabilities:

$$\boldsymbol{P(E_1 \text{ and } E_2) = P(E_1) \cdot P(E_2)}$$

Quick Review Question 5 Suppose at a site on a strand of DNA, an equal likelihood exists for any of the four bases. Give the probability of one site containing A and another unrelated site containing T.

Frequently, we wish to know the probability of one event, E_2, given the occurrence of another event, E_1. The notation for such a **conditional probability** is $\boldsymbol{P(E_2 \mid E_1)}$. For example, suppose a public health agency wages an aggressive campaign to stop the spread of a particular disease by trying to quarantine any individual who has come in contact with someone who has the disease. The probability that an exposed individual is quarantined can be written as a conditional probability, $P(\text{quarantined} \mid \text{exposed})$, the probability of quarantine given exposure. This quantity is equal to probability of the individual being quarantined and exposed divided by the probability of being exposed:

$$P(\text{quarantined} \mid \text{exposed}) = P(\text{quarantined and exposed}) / P(\text{exposed})$$

For example, suppose in a group of 100 people, 10 have been exposed and 2 have been exposed and quarantined. Thus, picking an individual at random from the group of 100, we have a 10/100 = 10% = 0.10 chance of selecting an exposed person and a 2/100 = 2% = 0.02 chance of the person being quarantined and exposed. However, if our selection is only from the subset of 10 exposed people, then the probability of picking one of the two individuals who is also quarantined is 2/10 = 0.20 = 20%; the probability that an exposed individual is quarantined is 0.02 / 0.10 = 0.2 = 20%.

Rule Conditional probability of event E_2 given event E_1 is

$$\boldsymbol{P(E_2 \mid E_1) = P(E_2 \text{ and } E_1) / P(E_1)}$$

Thus,

$$\boldsymbol{P(E_2 \text{ and } E_1) = P(E_2 \mid E_1)\, P(E_1)}$$

Quick Review Question 6 Suppose the DNA for a certain animal contains the sequence, s_1, of 20 bases (A, C, T, G) that evolves to another sequence, s_2, as follows:

s_1	C	A	C	T	T	G	T	G	A	G	C	C	C	A	C	T	T	C	G	T
s_2	C	A	T	T	T	G	T	G	A	C	C	C	T	A	C	T	T	A	G	T

Determine the following probabilities:

a. That C occurs in s_1, written $P(E_1 = \text{C})$

b. That C occurs in s_2

c. That C occurs in s_1 and T occurs in the corresponding site in s_2, written $P(E_2 = \text{T and } E_1 = \text{C})$

d. T occurs in the corresponding site in s_2, given that C occurs in s_1, written $P(E_2 = \text{T} \mid E_1 = \text{C})$

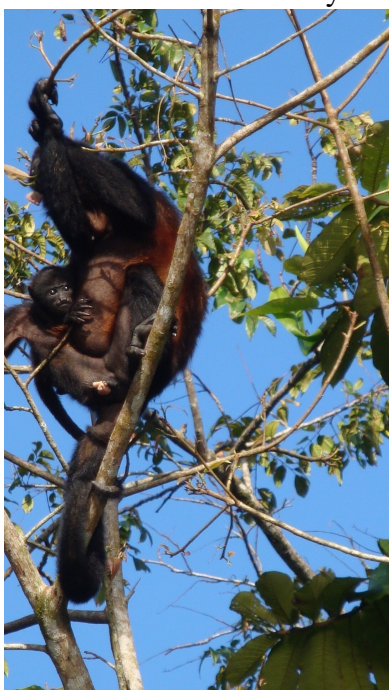
e. Calculate $P(E_2 = \text{T}$ and $E_1 = \text{C}) / P(E_1 = \text{C})$, which is your answer from Part a divided into your answer for Part c.

f. How do your answers from Parts d and e compare?

Transition Matrix

We can employ a matrix of conditional probabilities to estimate the long-term behavior of an animal. For example, the Red Howler Monkey's primary food is leaves. Because leaves are hard to digest, the monkey spends about half of its waking hours resting. Resting requires less energy than other activities and gives time for digestion. Suppose we consider a simplified system where the monkey is only in two states, eating (E) and resting/sleeping (R); and $S = \{\text{E}, \text{R}\}$ is the **state space**, or set of possible states.

Figure 1 Red Howler Monkey mother and infant in Costa Rica



Let us consider some hypothetical data. If one state (X_n) of the monkey is eating, then the probability that the state of the monkey one hour later (X_{n+1}) is eating is 0.6. We express this information as a conditional probability, $P(X_{n+1} = \text{E} \mid X_n = \text{E}) = 0.6$. Because we assume the monkey is either eating or resting at any time, the probability that the monkey is resting one hour after eating is $P(X_{n+1} = \text{R} \mid X_n = \text{E}) = 1 - 0.6 = 0.4$.

Quick Review Question 7 With a state of resting at time n, let us suppose that one hour later the monkey is eating with a probability of 0.2.

a. Express this information in conditional probability notation.

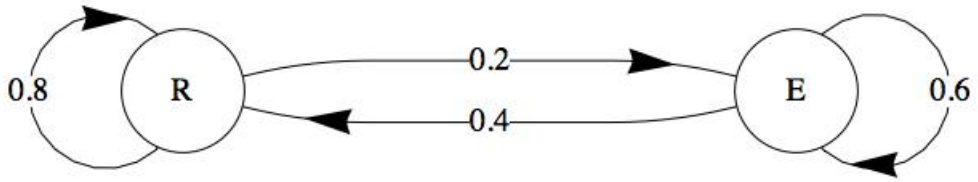
b. Give the conditional probability notation and value for the monkey resting one hour later.

We can express the data in the above paragraph and quick review question with the following matrix, T:

$$T = \begin{array}{c} X_{n+1} \backslash X_n \\ \text{E} \\ \text{R} \end{array} \begin{array}{c} \begin{array}{cc} \text{E} & \text{R} \end{array} \\ \begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix} \end{array}$$

The first column indicates the probabilities of the indicated values (E or R) of state X_{n+1} given that the monkey is initially eating, X_n = E. Note that the sum of this column's values is 1, because we are considering only one of two possible states for the monkey at any time. Similarly, the second column sums to 1 and presents the probabilities of the monkey eating or resting/sleeping given that the animal was resting the previous hour. Figure 2 presents a **state diagram** of the system with the nodes representing the states and probabilities of going from one state to another labeling the directed edges.

Figure 2 State diagram of the system



We call T a **transition matrix** (**Markov matrix**, **probability matrix**, or **stochastic matrix**). A **Markov chain** consists of a sequence of variables $X_1, X_2, X_3, \ldots$ in which the value of any variable, X_{n+1}, only depends on the value of its immediate predecessor, X_n. That is, $P(X_{n+1} = x \mid X_n = x_n, \ldots, X_2 = x_2, X_1 = x_1) = P(X_{n+1} = x \mid X_n = x_n)$.

Definition A **transition matrix** (**Markov matrix**, **probability matrix**, or **stochastic matrix**) is a matrix in which all the entries are nonnegative and the sum of the elements in each column is 1. A **Markov chain** consists in a sequence of variables $X_1, X_2, X_3, \ldots$ in which the value of any variable, X_{n+1}, only depends on the value of its immediate predecessor, X_n.

Suppose initially 90% of a group of Howler monkeys are eating and 10% resting, represented by the **probability vector** $\boldsymbol{v}_0 = \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix}$, where the components are nonnegative and sum to 1. We can predict the percentage of monkeys eating and resting an hour later by evaluating $T\boldsymbol{v}_0$, as follows:

$$\boldsymbol{v}_1 = T\boldsymbol{v}_0 = \begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix} \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0.56 \\ 0.44 \end{bmatrix}$$

The calculations predict that at the next hour 56% will be eating, while 44% will be resting.

Definition A **probability vector** is a vector whose components are nonnegative and sum to 1.

Using T and $\boldsymbol{v}_1$, we can predict the situation at hour 2, as follows:

$$\boldsymbol{v}_2 = T\boldsymbol{v}_1 = \begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix}\begin{bmatrix} 0.56 \\ 0.44 \end{bmatrix} = \begin{bmatrix} 0.424 \\ 0.576 \end{bmatrix}$$

Thus, we predict, 42.4% of the monkeys will be eating, and 57.6% resting.

Note that by substitution of $\boldsymbol{v}_1 = T\boldsymbol{v}_0$ in $\boldsymbol{v}_2 = T\boldsymbol{v}_1$, we see that $\boldsymbol{v}_2 = T(\boldsymbol{v}_1) = T(T\boldsymbol{v}_0) = TT\boldsymbol{v}_0 = T^2v_0$. Similarly, at the next hour, the vector is $\boldsymbol{v}_3 = T\boldsymbol{v}_2 = T(T^2\boldsymbol{v}_0) = T^3\boldsymbol{v}_0 = \begin{bmatrix} 0.3696 \\ 0.6304 \end{bmatrix}$. In general, $\boldsymbol{v}_n = T^n\boldsymbol{v}_0$. Table 1 presents several calculations for T^n and $\boldsymbol{v}_n$. Notice that as n gets larger and larger, written $n \to \infty$, T^n approaches, or **converges to**, $\begin{bmatrix} 1/3 & 1/3 \\ 2/3 & 2/3 \end{bmatrix}$ and $\boldsymbol{v}_n$ converges to $\boldsymbol{v} = \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix}$, an **equilibrium** or **steady-state vector** associated with T. Thus, $\boldsymbol{v}$ is a probability vector with $T\boldsymbol{v} = \boldsymbol{v}$ where each coordinate of $\boldsymbol{v}$ is the long-term probability that the system will be in the corresponding state. As time progresses, at any one time approximately one-third of the monkeys will be eating and two-thirds resting. Moreover, regardless of the starting vector giving the percentages in each category, with time the percentages will approach $33\frac{1}{3}\%$ and $66\frac{2}{3}\%$ for eating and resting, respectively. Even if all monkeys are eating initially, eventually about one-third will be eating at any one time. When all the entries of a transition matrix are positive, it can be shown that T^n will converge to a matrix M and $\boldsymbol{v}_n = T^n\boldsymbol{v}_0$ will converge to a steady-state vector. (We will cover a technique for calculating these limiting steady-state values shortly.)

Definition An **equilibrium** or **steady-state vector**, $\boldsymbol{v}$, of the Markov chain associated with the transition matrix T is a probability vector, where $T\boldsymbol{v} = \boldsymbol{v}$.

Table 1 Markov matrix, T, and probability vector, v, to several powers

n	$\boldsymbol{v}_n = T^n\boldsymbol{v}_{n-1}$	T^n
0	$\begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix}$	
1	$\begin{bmatrix} 0.56 \\ 0.44 \end{bmatrix}$	$\begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix}$

2	$\begin{bmatrix}0.424\\0.576\end{bmatrix}$	$\begin{bmatrix}0.44 & 0.28\\0.56 & 0.72\end{bmatrix}$
3	$\begin{bmatrix}0.3696\\0.6304\end{bmatrix}$	$\begin{bmatrix}0.376 & 0.324\\0.624 & 0.688\end{bmatrix}$
4	$\begin{bmatrix}0.34784\\0.65216\end{bmatrix}$	$\begin{bmatrix}0.3504 & 0.3248\\0.6496 & 0.6752\end{bmatrix}$
10	$\begin{bmatrix}0.333393\\0.666607\end{bmatrix}$	$\begin{bmatrix}0.333403 & 0.333298\\0.666597 & 0.666702\end{bmatrix}$
100	$\begin{bmatrix}0.333333\\0.666667\end{bmatrix}$	$\begin{bmatrix}0.333333 & 0.333333\\0.666667 & 0.666667\end{bmatrix}$

Theorem 1 If all the entries of a Markov matrix are positive, then as n gets larger and larger, T^n converges to a matrix, M, and $\boldsymbol{v}_n = T^n\boldsymbol{v}_0$ converges to a vector, $\boldsymbol{v} = M\boldsymbol{v}_0$.

Quick Review Question 8 Suppose baboons are observed to be eating (E), grooming (G), or resting (R). A biologist records their activities every 15 minutes and estimates that if a baboon is eating at one period, at the next 15 minute period the animal will be eating or resting with the probabilities 0.3 and 0.6, respectively. If grooming at one observation, in 15 minutes they are likely to be grooming with a 0.3 probability or eating with a 0.4 probability. If resting at one time period, at the next observation the probabilities a baboon will still be resting or will instead be eating are 0.8 and 0.2, respectively.

- **a.** Using the order E, G, and R for rows and columns, develop a transition matrix, T, for this problem.
- **b.** Suppose when the study began, 30% of the baboons were eating, 10% were grooming, and 60% were resting. Using the model from Part a, give estimates for the percentages of baboons in each state 15 minutes later.
- **c.** Using a computational tool, estimate the matrix to which T^n converges as n gets larger and larger.
- **d.** Using a computational tool, estimate the vector to which a probability vector for the system converges n goes to infinity.

Using a computational tool, we can calculate that the **dominant eigenvalue** of the Markov matrix, T, for the Howler monkey example is $\lambda = 1$, and a corresponding **eigenvector** is $\boldsymbol{x} = (-0.447214, -0.894427)$, so that,

$$\begin{bmatrix}0.6 & 0.2\\0.4 & 0.8\end{bmatrix}\begin{bmatrix}-0.447214\\-0.894427\end{bmatrix} = 1\cdot\begin{bmatrix}-0.447214\\-0.894427\end{bmatrix}$$

$$T\boldsymbol{x} = \lambda\boldsymbol{x}$$

The ratio of the first coordinate of $\boldsymbol{x}$ to the second coordinate is one-third to two-thirds. That is, if we add the coordinates of $\boldsymbol{x}$, s = -0.447214 + -0.894427 = -1.34164, and divide the sum s into each coordinate of $\boldsymbol{x}$, we obtain -0.447214/-1.34164 = 1/3 = $0.33\overline{3}$ = $33\frac{1}{3}\%$ and -0.894427/-1.34164 = 2/3 = $0.66\overline{6}$ = $66\frac{2}{3}\%$. These values are the exact proportions to which the components of $\boldsymbol{v}_n = T^n\boldsymbol{v}_{n-1}$ tend as n goes to infinity (see Table 1). The vector $\boldsymbol{x}$ = (1/3, 2/3) = ($0.33\overline{3}$, $0.66\overline{6}$) = ($33\frac{1}{3}\%$, $66\frac{2}{3}\%$) is the equilibrium vector associated with the transition matrix T.

Definition For square matrix M, the constant λ is an **eigenvalue** and $\boldsymbol{v}$ is an **eigenvector** if multiplication of the constant by the vector accomplishes the same results as multiplying the matrix by the vector, that is, the following equality holds:

$$M\boldsymbol{v} = \lambda\boldsymbol{v}$$

The **dominant eigenvalue** for a matrix is the largest eigenvalue for that matrix.

In general, $\lambda = 1$ is always an eigenvalue for the transition matrix of a Markov chain. Moreover, if each of the components of the corresponding eigenvector $\boldsymbol{x}$ is nonnegative and s is the sum of these components, then $(1/s)\boldsymbol{x}$ is the equilibrium vector for T, and this vector is a probability vector. If we start with a probability vector, $\boldsymbol{v}_0$, where each component gives the fraction in each corresponding state, such as eating (E) and resting/sleeping (R), then $T^n\boldsymbol{v}_0$ converges to $\boldsymbol{v} = (1/s)\boldsymbol{x}$ as n becomes larger and larger. Moreover, each coordinate of this equilibrium vector, $\boldsymbol{v}$, is the ultimate proportion of the corresponding state.

Theorem 2 Suppose T is a Markov chain transition matrix. Then, T has an eigenvalue $\lambda = 1$. Moreover, if each of the components of the corresponding eigenvector, $\boldsymbol{x}$, is nonnegative and s is the sum of these components, then $(1/s)\boldsymbol{x}$ is a steady-state vector for T.

Theorem 3 Suppose T is a Markov chain transition matrix. If T^n has all positive entries for some positive integer n, then T has a unique equilibrium vector $\boldsymbol{v}$. Moreover, if $\boldsymbol{y}$ is a probability vector, then $T^n\boldsymbol{y}$ converges to $\boldsymbol{v}$ as n becomes larger and larger. (Agnew and Knapp, 2002)

Quick Review Question 9 For the baboon example in Quick Review Question 8, using a computational tool, determine

- **a.** the dominant eigenvalue.
- **b.** the principal eigenvector.
- **c.** the steady-state vector associated with T.
- **d.** the ultimate percentages, expressed in whole numbers, in each state.

3. Bioinformatics and Markov Chains

The Area

A newly developing area of computational science, called **bioinformatics**, deals with the organization of biological data, such as in databases, and the analysis of such data, which often makes extensive use of probabilities. Recently, enormous strides have been made in genetics, due in part to the power of bioinformatics and high performance computing. In the next few sections, we give some of the biological background that will enable us to discuss a number of bioinformatics examples, such as those involving Markov chains.

Proteins

Proteins are basic building blocks of life, performing many critical functions. Some proteins are the fundamental, structural components of cells and tissue, while others (**enzymes)** are catalysts for chemical reactions. A simple protein is a linear polymer or chain of **amino acids**. Table 2 lists the twenty amino acids common to proteins along with their one-letter and three-letter codes. Each amino acid contains an **amino group** ($-NH_3^+$) at one end and a **carboxyl group** ($-COO^-$) at the other, connected by a carbon (**α-carbon**). A variable side-chain (**R-group**) and a hydrogen are attached to the α-carbon (see Figure 3). The R-group is responsible for the chemical nature (acidic, nonpolar, etc.) of each amino acid. Chains of amino acids are linked by **peptide bonds,** which form through the interaction of an amino group of one amino acid with the carboxyl group of another (see Figure 4). This interaction results in condensation, or release of water. Because one end of a protein has a free amino group (**N-terminal**) and the other has a free carboxyl group (**C-terminal**), we can assign an orientation to the chain and list the amino acids from the "beginning" (N-terminal) of the chain to the "end" (C-terminal).

Table 2 The twenty commonly occurring amino acids along with their one-letter and three-letter codes. (Note: B is used when one cannot distinguish between D and N because of amino acid analytical processing. Similarly, Z is used when it is ambiguous whether the amino acid is E or Q. X represents an unknown or nonstandard amino acid.)

One-Letter Code	Three-Letter Code	Name
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic Acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic Acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine

M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine

Figure 3 Structure of an amino acid (Rupp 2000)

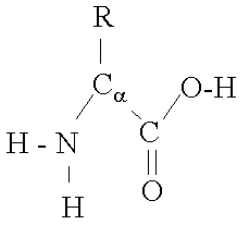
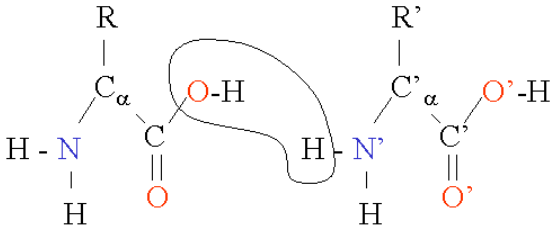


Figure 4 Chain of two amino acids (Rupp 2000)



Quick Review Question 10

a. Give the name of the area of computational science that deals with the organization and the analysis of biological data.

b. Give the number of amino acids common to proteins.

Match each phrase in the following parts with the best term:

α-carbon	amino acids	amino group	C-terminal
carboxyl group	enzymes	N-terminal	peptide bonds
proteins	R-group		

c. Basic building blocks of life

d. Proteins that are catalysts for chemical reactions.

e. A simple protein is a linear chain of these

f. Free amino group that is the beginning of the chain of amino acids

g. Free carboxyl group that is the end of the chain of amino acids

Nucleic Acids

In the cell, the nucleic acid **DNA** (**deoxyribonucleic acid**) contains the encoded information for the manufacture of all the proteins a cell needs. However, DNA does not

oversee protein synthesis directly but acts through an intermediary nucleic acid, **RNA** (**ribonucleic acid**). The RNA sequences subsequently specify the amino acid sequences of proteins. Both DNA and RNA are polymers, or long chains, of molecules called **nucleotides**. A nucleotide is a compound molecule made up of a sugar (either **deoxyribose** or **ribose**), a phosphate, and a nitrogen base (**adenine** (**A**), **guanine** (**G**), **cytosine** (**C**), and **thymine** (**T**) in DNA or **uracil (U)** in RNA). A and G are **purines**, while C, T, and U are **pyrimidines**. DNA is a double strand of nucleotides, whereas RNA is a single strand. Thus, we can say a particular DNA molecule has 300 bases or 300 nucleotides. As with proteins, because the backbone of a strand always has specific chemical structures at opposite ends, we can canonically give direction to the sequence of nucleotides (or bases) in a strand.

Bases in one strand may bond with bases in another. Because of their structure, A and T always bond together, and C and G always bond together. Each pair is said to be made up of **complementary bases** and is referred to as a **base pair (bp).** The number of such **base pairs** is use to describe the **length** of a DNA molecule. Because of pairing consistency, by knowing the sequence of bases in one strand, we can deduce the sequence of bases in the other strand through **reverse complementation**. For example, suppose one sequence is *s* = ATGAC. Because of the required pairing, A - T and C - G, we know the base pairs must appear as follows:

s:	A	T	G	A	C
	|	|	|	|	|
	T	**A**	**C**	**T**	**G**

Quick Review Question 11 Match each phrase in the parts with the best term(s):

A	C	DNA	G	protein
purine	pyrimidine	RNA	T	U

a. Contains the encoded information that is stored to direct the manufacture of all the proteins a cell needs
b. An intermediary nucleic acid in protein synthesis
c. Compound molecule made of a sugar, a phosphate, and a nitrogen base
d. Type of molecule in DNA and RNA sequences
e. Bases in DNA
f. Bases in RNA
g. Purines
h. Pyrimidines
i. Always bonds with base A in DNA
j. Always bonds with base A in RNA
k. Always bonds with base C in DNA or RNA
l. Always bonds with base T in DNA
m. Always bonds with base U in RNA
n. Always bonds with base G in DNA or RNA

In contrast to DNA, RNA is a single strand of nucleotides made up of ribose sugars and bases A, C, G, and U instead of the nitrogen base thymine (T) (see Table 3). Several types of RNA with different functions exist in the cell.

Table 3 Bases in DNA and RNA

Base	Abbreviation	Complement	In DNA	In RNA	Group
adenine	A	T in DNA, U in RNA	yes	yes	purine
cytosine	C	G	yes	yes	pyrimidine
guanine	G	C	yes	yes	purine
thymine	T	A	yes	no	pyrimidine
uracil	U	A	no	yes	pyrimidine

A mutation in a DNA sequence can occur with the **insertion** or **deletion** of a base or the **substitution** of one base for another. One type of substitution, called a **transition**, occurs between purines, from A to G or from G to A, or between pyrimidines, from T to C or vice versa. A **transversion** substitution occurs between a purine and a pyrimidine or vice versa. In a substitution, a transition is much more likely to occur than a transversion.

Quick Review Question 12 Match each phrase in the parts with the best term(s):

deletion DNA insertion nucleotide protein
purine pyrimidine RNA transition transversion

a. Single strand of nucleotides
b. Double strand of nucleotides
c. DNA mutations
d. Substitution between A and G or between T and C
e. Substitution between purine and pyrimidine
f. More likely substitution

From Genes to Proteins

For the genetic application of locating genes with Markov chain models and for some projects, we need some additional background concerning genes. Each cell contains **chromosomes**, which are very long DNA molecules. A **gene** is a contiguous section of a chromosome that encodes information to build a protein or an RNA molecule. In humans, a gene is composed of about 10,000 base pairs (bp). A chromosome contains genes and contiguous sections that are not part of any gene. Some scientists believe that genes (coding sequences) compose only about 10% of a human chromosome. The function of these non-gene bits of DNA is still debated. Some are known to be important for regulation of gene expression and other are important for matching homologues and structure. A complete set of chromosomes in a cell contains the organism's hereditary information and is called the **genome**. For example, a human genome has 46 chromosomes in 23 pairs.

For simplicity, we assume that a particular protein in an organism corresponds to exactly one gene. In a gene, a sequence of three nucleotides (**triplet**) specifies an amino acid. For example, the sequence ACG or ACA encodes the information for the amino acid Threonine (Thr) (See Table 2). The **genetic code** represents such a correspondence between these triplets and the amino acids they specify. With four base choices, a pair of bases could only encode information for (4)(4) = 16 amino acids. With three bases,

(4)(4)(4) = 64 possible triplets exist. Several, such as ACG and ACA, encode the same amino acid; and three sequences do not encode for any amino acid.

Protein synthesis is the process of using genetic code to direct the building of proteins. Synthesis begins in the nucleus, where enzymes catalyze the production of a molecule of RNA, termed **messenger RNA** or **mRNA**. Each DNA triplet specifies a complementary sequence of three nucleotides, which we call a **codon**, in the RNA. The synthesis of RNA is called **transcription**. During transcription, base pairing ensures formation of a strand of RNA that is complementary to the gene sequence with U replacing T.

Quick Review Question 13 Match each phrase in the parts with the best term(s):

chromosome	codon	DNA	gene	genome
mRNA	protein synthesis	transcription	triplet	tRNA

a. Very long DNA molecule in a cell makes up a _____.
b. A contiguous section of a chromosome that encodes information to build a protein or an RNA molecule is called a _____.
c. A complete set of chromosomes contains an organism's hereditary information and is called its _____.
d. A sequence of three nucleotides in a gene is called a _____.
e. A molecule of RNA produced in the nucleus that contains information to synthesize a protein is _____.
f. Sequence of three nucleotides in RNA that is complementary to a DNA triplet is called a(n) _____.
g. The synthesis of RNA is called _____.

Locating Genes with Markov Models

The most dependable method of discovering a gene in a new genome is observing a close homolog, or a gene from the same ancestral origin, in another organism. However, when homologs to known genes do not exist, we must employ computational methods to help identify genes (Salzberg *et al*. 1998).

In mammals, the sequence of bases **CG** frequently transforms to (methyl-C)G and then mutates to TG. Thus, the pair CG appears less that we would expect from random occurrences of C and G independently. However, this process of transformation from CG to TG is suppressed in small regions, called **CpG islands**, upstream of, or before, many genes; so CpG islands can be employed to locate genes. The "p" in "CpG" indicates a phosphate that links the two bases C and G in DNA. The classical definition of a CpG island is a DNA segment of length 200 that has CG occurring 50% of the time and a ratio of observed-to-expected number of CpG's above 0.6 (Gardiner-Garden & Frommer 1987).

We can use Markov chains to determine whether a short segment of genomic data is from a CpG island or not. First, we use **training sequences** that we know contain CpG islands, called **positive ("+") samples**, to derive for each base four probabilities—the probabilities that A, C, G, and T follow the base. For example, consider the sequence ACGTCTATTC, which is exceptionally small for the sake of illustration. To calculate the probability that T is followed by A, written as $P(x_i = \text{A} \mid x_{i-1} = \text{T})$ or $P(\text{A} \mid \text{T})$, we

divide the number of occurrences of TA in the sequence, here 1, by the number of pairs that begin with T, here 4 (TC, TA, TT, and TC). Thus, $P(\text{A} \mid \text{T}) = ¼ = 0.25$ for this sequence. 25% of the time the next base after T is A. Moreover, the sum of the probabilities $P(\text{A} \mid \text{T}) + P(\text{C} \mid \text{T}) + P(\text{G} \mid \text{T}) + P(\text{T} \mid \text{T}) = 0.25 + 0.50 + 0.00 + 0.25 = 1.00$.

Figure 5a presents a transition matrix for such positive samples determined from 60,000 nucleotides from a database of human DNA sequences with 48 CpG islands. As in the example in the last paragraph, the sum of the elements on each row is 1.00, while the column sum is not necessarily 1.00. In that matrix, the probability of the pair CG (or the probability that G occurs, given that C has just appeared) is 0.274, written as $P_+(x_i = \text{G} \mid x_{i-1} = \text{C}) = P_+(\text{G} \mid \text{C}) = 0.274$. We also employ training sequences for known **negative** ("**-**") **samples** to derive another transition matrix, such as in Figure 5b. Thus, for these training sequences, the probability that the sequence CG occurs in the positive samples with CpG islands is 0.274, while we find that such a sequence is much less likely (probability of $P_-(\text{G} \mid \text{C}) = 0.078$) to occur in the negative samples that do not contain CpG islands.

Figure 5 Possible transition matrix for (a) positive and (b) negative samples (Durbin *et al*. 1998)

a

x_{i-1} \ x_i (+)	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

b

x_{i-1} \ x_i (-)	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

Quick Review Question 14 Compute the transition matrix using the training sequence ACGTCTATTC.

We can now use Markov chains to determine if a short sequence, $\boldsymbol{x} = (x_1x_2x_3\ldots x_n)$ is more likely to come from a positive or a negative sample by considering the ratio of the probability that the sequence is from a positive sample over the probability that the sequence is from a negative sample:

$$\frac{P(x \mid \text{positive model})}{P(x \mid \text{negative model})}$$

If this ratio is greater than 1, the sequence is more likely to be from a CpG island.

To derive the formulas for the numerator and denominator, let us consider a very short sequence of four bases $\boldsymbol{x} = (x_1x_2x_3x_4)$. Regardless of the positive or negative model, the probability that $\boldsymbol{x}$ occurs, $P(x_1x_2x_3x_4)$ is $P(x_4 \text{ and } x_1x_2x_3)$, the probability of x_4 and $x_1x_2x_3$. As we saw earlier $P(x_4 \text{ and } x_1x_2x_3)$ is $P(x_4|x_1x_2x_3)\cdot P(x_1x_2x_3)$, the probability that x_4

occurs given that the sequence $x_1x_2x_3$ occurs immediately before it times the probability that $x_1x_2x_3$ appears. Thus, we have the following:

$$P(x_1x_2x_3x_4) = P(x_4|x_1x_2x_3)\cdot P(x_1x_2x_3) \quad (1)$$

Now, with Markov chains, x_4 only depends on the value of its immediate predecessor, x_3, so that $P(x_4|x_1x_2x_3) = P(x_4|x_3)$, and we can simplify Equation (1) as follows:

$$P(x_1x_2x_3x_4) = P(x_4|x_3)\cdot P(x_1x_2x_3) \quad (2)$$

We then repeat the process to compute $P(x_1x_2x_3)$:

$$P(x_1x_2x_3) = P(x_3 \text{ and } x_1x_2) = P(x_3|x_1x_2)\cdot P(x_1x_2) = P(x_3|x_2)\cdot P(x_1x_2) \quad (3)$$

Substituting (3) into (1), we have the following:

$$P(x_1x_2x_3x_4) = P(x_4|x_3)\cdot P(x_3|x_2)\cdot P(x_1x_2) \quad (4)$$

Using the same reasoning, we have

$$P(x_1x_2) = P(x_2|x_1)\cdot P(x_1) \quad (5)$$

and finally

$$P(x_1x_2x_3x_4) = P(x_4|x_3)\cdot P(x_3|x_2)\cdot P(x_2|x_1)\cdot P(x_1) \quad (6)$$

The probability of the sequence $x_1x_2x_3x_4$ is "unzipped" from right to left as the product of probability of obtaining x_4 given that x_3 is immediately preceding, the probability of x_3 given x_2 is immediately preceding, the probability of x_2 given x_1 immediately preceding, and the probability of x_1. Generalizing, we have the following formula:

$$P(x_1x_2x_3\ldots x_n) = P(x_n|x_{n-1})\cdot P(x_{n-1}|x_{n-2})\cdots P(x_3|x_2)\cdot P(x_2|x_1)\cdot P(x_1) \quad (7)$$

The probability of x_1, $P(x_1)$, is the proportion of the time x_1 occurs in a sequence or the total number of occurrences of x_1 over the total number of bases in the sequence. For example, in Quick Review Question 6a, we determined that base C appears 7 times in the sequence s_1 of 20 bases, so that $P(\text{C}) = 7/20$. We use the training sequences to determine such probabilities. Moreover, the Markov matrices as in Figure 5 contain the other probabilities. Again, for the sake of example, suppose we have the probabilities of bases in training sequences that contain CpG islands as in Figure 6a. Then, we can calculate the probability that the sequence ACGTC is from a CpG island as follows:

$$P_+(\text{ACGTC}) = P_+(\text{C|T})\ P_+(\text{T|G})\ P_+(\text{G|C})\ P_+(\text{C|A})\ P_+(\text{A})$$

We calculate the first four probabilities using the transition matrix for the positive model in Figure 5a and the probability of A using Figure 6a, as follows:

$$\begin{aligned} P_+(\text{ACGTC}) &= P_+(\text{C|T})\ P_+(\text{T|G})\ P_+(\text{G|C})\ P_+(\text{C|A})\ P_+(\text{A}) \\ &= 0.355 \cdot 0.125 \cdot 0.274 \cdot 0.274 \cdot 0.258 \\ &= 0.00085953 \end{aligned}$$

Figure 6 Probability of bases for (**a**) positive (frequencies from gene-rich human chromosome 19) and (**b**) negative samples (frequencies from reference human genome sequence) (Guide 2010)

a

$P_+(\text{A}) = 0.258$
$P_+(\text{C}) = 0.242$
$P_+(\text{G}) = 0.242$
$P_+(\text{T}) = 0.259$

b

$P_-(\text{A}) = 0.295$
$P_-(\text{C}) = 0.205$
$P_-(\text{G}) = 0.205$
$P_-(\text{T}) = 0.296$

Similarly, we calculate the probability that ACGTC does not come from a CpG island using probabilities Figures 5b and 6b, as follows:

$$P_-(\text{ACGTC}) = P_-(\text{C|T})\ P_-(\text{T|G})\ P_-(\text{G|C})\ P_-(\text{C|A})\ P_-(\text{A})$$

$$= 0.239 \cdot 0.208 \cdot 0.078 \cdot 0.205 \cdot 0.295$$
$$= 0.00023449$$

The calculations indicate a greater probability that ACGTC contains a CpG island than not. Moreover, the quotient of the probabilities being larger than 1 also indicates a CpG island:

$$\frac{P(\text{ACGTC} \mid \text{positive model})}{P(\text{ACGTC} \mid \text{negative model})} = \frac{P_+(\text{ACGTC})}{P_-(\text{ACGTC})} = \frac{0.00085953}{0.00023449} = 3.6655$$

Quick Review Question 15 Using the transition matrices from Figure 5 and probabilities from Figure 6, calculate the following:

a. $P_+(\text{CCGTCGA})$
b. $P_-(\text{CCGTCGA})$
c. The quotient of Parts a and b
d. Is CCGTCGA more likely to be from a CpG island or not?

However, the sequence ACGTC is much shorter than the usual sequence of 200 to 250 bases. If we were to multiply together 200 probabilities, each less than 1, the result would be on the order of 10^{-200}. To avoid such a small magnitude number, the use of division, and a large number of multiplications, we employ logarithms. With the logarithm of a quotient being the difference of the logarithms, we can replace a division with a subtraction:

$$\ln\left(\frac{P_+(\text{ACGTC})}{P_-(\text{ACGTC})}\right) = \ln(P_+(\text{ACGTC})) - \ln(P_-(\text{ACGTC}))$$

Moreover, the log of a product is the sum of the logs:

$$\ln(P_+(\text{ACGTC})) = \ln(0.355 \cdot 0.125 \cdot 0.274 \cdot 0.274 \cdot 0.258)$$
$$= \ln(0.355) + \ln(0.125) + \ln(0.274) + \ln(0.274) + \ln(0.258)$$
$$= -7.0591$$

$$\ln(P_-(\text{ACGTC})) = \ln(0.239 \cdot 0.208 \cdot 0.078 \cdot 0.205 \cdot 0.295)$$
$$= \ln(0.239) + \ln(0.208) + \ln(0.078) + \ln(0.205) + \ln(0.295)$$
$$= -8.3581$$

Thus, we have

$$\ln(P_+(\text{ACGTC})) - \ln(P_-(\text{ACGTC})) = -7.0591 - -8.3581 = 1.2990$$

We then normalize this score by dividing by the length of the sequence to obtain 1.2990/5 = 0.2598. The larger this **length-normalized log-odds score** is the more likely that the sequence is from a CpG island (Tang; Gropl and Huson 2005).

Definition The **length-normalized log-odds score** for a sequence x is

$$\ln\left(\frac{P_+(x)}{P_-(x)}\right) \Big/ |x|$$

Quick Review Question 16 Calculate the length-normalized log-odds score for the sequence CCGTCGA of Quick Review Question 15.

A High Performance Computing Approach to Locating Genes

With a long DNA sequence, computing the length-normalized log-odds score for each segment of 200 bases involves an enormous amount of computation. For example, the genome for *Escherichia coli* (*E. coli*) contains over 5.5 million base pairs, and, thus, over 5.5 million segments of length 200. Computation with the human genome is even more daunting with about 3 billion nucleotides (International 2004).

Fortunately, computing the scores for each segment of 200 bases is **embarrassingly parallel** on a high performance system; we can divide computation into many completely independent experiments with virtually no communication except for the initial communication of transition matrices, base probabilities, and sections of the DNA sequence. Thus, we can have multiple nodes on a cluster running the same program with different sequences or sections of a larger sequence and with their own output files. After completion, we can use the scores to predict locations of CpG islands. Because of the embarrassingly parallel nature of this approach with limited communication, the problem scales quite well; so that for a long sequence, the parallel version of the program can run faster with more processes. Projects 6-8 explore the speedup involved with a HPC version of this program.

The Blue Waters website contains serial and parallel versions of a scoring program in MATLAB with the Parallel Toolbox and in C with MPI. Running the MATLAB programs on a MacBook Pro with input data of a segment of 10,000 bases of *E. coli*, the serial version took approximately 17 seconds, while the parallel version using two cores was almost twice as fast, taking about 9 seconds.

Definition An **embarrassingly parallel algorithm** can divide computation into many completely independent parts with virtually no communication.

GeneMark

The technique of locating genes from the previous section "Locating Genes with Markov Models" is a **1st-order Markov model** because the method predicts each base using one preceding base in the DNA sequence. For this method, as in Figure 5a, with positive training sequences that contain CpG islands, $4^2 = 16$ probabilities of base y occurring given base x immediately preceding were calculated. Similarly, as in Figure 5b, 16 probabilities were obtained using negative training sequences that do not contain such islands. Moreover, as in Figure 6, the probabilities of each base occurring in a positive sequence and in a negative sequence were required, resulting in an additional $4 + 4 = 8$ probabilities.

The gene-finding program **GeneMark**, which is a **5th-order Markov model**, employs five previous bases to predict a base. Compared to the 32 probabilities in Figure 5, GeneMark must use $4^6 = 4096$ probabilities for positive and 4096 for negative training sequences. Moreover, comparable to Figure 6, the program must also compute the probability of each sequence of 5 bases occurring in positive and negative training sequence, or $2(4^5) = 2048$ probabilities. Thus, GeneMark calculates $4096 + 4096 + 2048 = 10{,}240$ probabilities from the training sequences alone. Project 6 discusses the GeneMark algorithm in greater detail.

High Performance Computing and Bioinformatics

As with locating genes, biological systems provide us with complexity that challenges our ability to interpret data. To help unravel these complexities, the Human Genome Project set out to map all of the human genome, no simple goal if we consider that our genetic code consists of 20,000-25,000 genes, composed of about 3 billion nucleotides. It is remarkable that the program completed the mapping of the human genome in only 13 years, the last chromosome completed and published in 2006. Now, this tremendous accomplishment seems like the "easy part" in our attempts to unravel the complexities of ourselves. The data generated by this project, which is now combined with data from the genomes of other organisms, is accumulating with ever increasing volume and complexity. To analyze this data and derive any understanding will require the development of genomic-scale technologies. Even with such technologies, biological research in this area is likely to take decades.

A few of the research areas of genetics that will be pursued and expanded include (Human Genome 2012):

- Gene number, exact locations, and functions
- Gene regulation
- DNA sequence organization
- Chromosomal structure and organization
- Noncoding DNA types, amount, distribution, information content, and functions
- Coordination of gene expression, protein synthesis, and post-translational events
- Interaction of proteins in complex molecular machines
- Predicted vs. experimentally determined gene function
- Evolutionary conservation among organisms
- Protein conservation (structure and function)
- Proteomes (total protein content and function) in organisms
- Correlation of SNPs (single-base DNA variations among individuals) with health and disease
- Disease-susceptibility prediction based on gene sequence variation
- Genes involved in complex traits and multigene diseases
- Complex systems biology, including microbial consortia useful for environmental restoration
- Developmental genetics, genomics

If we consider just one of these areas—cataloguing the complete human proteome—we might consider the sequencing of the human genome as a relatively straightforward task. The **proteome** is the set of proteins that are produced and expressed in the cells of our bodies and is of vastly greater size and complexity than the sequences of the human genome. Furthermore, unlike the genome, which is somewhat fixed with time, the proteome is not static and varies considerably with aging in response to various cell signals and other externally derived stimuli. Such studies will require the use sophisticated mathematical and computer techniques and enormous amounts of computational power.

High-performance computing (HPC) has generally been lightly applied to biological problems, but with the size and complexity of biological systems, those days are quickly ending. For instance, during the middle of the 1990's, French researchers had

spent two years searching for the gene associated with the rare genetic disorder (X-linked) adrenoleukodystrophy. The defective gene leads to demyelination of the neurons of the brain and progressive decrease in function of the adrenal gland. The childhood version of this disease leads to coma and death within 10 years of the appearance of the first symptoms (Adrenoleukodystrophy 2011). The scientists had performed the available techniques (chromosome fragmentation, high-throughput sequencing) and aligned the bases for the chromosome, but still could not localize the functional gene. They contacted the computational team at Oak Ridge National Laboratory, who entered the sequence information into GRAIL™ (a suite of tools for sequence recognition). By applying the statistical and pattern recognition tools of this suite, the computer found the gene within two minutes (Mysteries 1999).

To handle the enormous amounts of data, it will certainly be necessary to apply high-performance/distributed computing power. Oehmen and Cannon (2008) suggested three directed ways HPC is applied to biological systems:

1. High throughput data analysis and data mining (pattern recognition) – needed for the tremendous amounts of data from genome sequencing, cell imaging and proteomics
2. HPC Grid and Cluster Computing – large-scale simulations, integrating models that span various time and spatial scales
3. Network Inference / Graphical Analysis—mapping behavior of biological systems onto a network/graph, which will allow us to infer from the mathematical representation various features and relationships of the system

Exercises

***NOTE: Answers to exercises with boxed numbers appear after the* Exercises *section in the section* Answers to Selected Exercises.**

1. In this problem we consider the animal community on a vertical rock wall of a middle intertidal zone. Suppose we have data for large (> 2 cm) and small (≤ 2 cm) mussels *Mytilus californianus (*B and SMC, respectively), goose barnacles *Pollicipes polymerus* (PP), and other crustaceans (Other). Suppose at a fixed point the transition probabilities from ecological state B to ecological states B, SMC, and PP are 0.84, 0.04, and 0.03, respectively; from SMC to B, SMC, and PP are 0.55, 0.26, and 0.03, respectively; from PP to B, SMC, and PP are 0.40, 0.06, and 0.35, respectively; and from other to B, SMC, and PP are 0.15, 0.07, and 0.02, respectively.

- **a.** Develop the 4-by-4 transition matrix for this model, where the sum of the elements in each row is 1.0.
- **b.** Determine the equilibrium vector.
- **c.** Interpret the results.

2. The **Jukes-Cantor Model** for DNA sequence evolution uses a constant α for the probability of substitution of one base for a different base, such as G for T.

- **a.** Under this model, give the formula for the probability that a base at a particular position does not mutate from one evolutionary time step to the next.

b. Give the general transition matrix for this model.
c. Give the transition matrix in the situation where $\alpha = 0.25$, and determine the ultimate distribution of bases.
d. Give the transition matrix in the situation where $\alpha = 0.3$, and determine the ultimate distribution of bases.
e. Give the transition matrix in the situation where $\alpha = 0.1$, and determine the ultimate distribution of bases.
f. Determine the ultimate distribution of bases for the general matrix of Part b.
g. What conclusions do you draw from your calculations?

3. The Kimura model for DNA sequence evolution gives a higher probability for a transition (probability α) than a transversion (probability β) with $\alpha > \beta$. (Sinha 2007)

a. Under this model, give the formula for the probability that a base at a particular position does not mutate from one evolutionary time step to the next.
b. Give the general transition matrix for this model.
c. Give the transition matrix in the situation where $\alpha = 0.25$ and $\beta = 0.10$, and determine the ultimate distribution of bases.
d. Determine the ultimate distribution of bases for the general matrix of Part b.
e. What conclusions do you draw from your calculations?

Answers to Selected Exercises

1. a.
$$\begin{bmatrix} 0.84 & 0.04 & 0.03 & 0.09 \\ 0.55 & 0.26 & 0.03 & 0.16 \\ 0.40 & 0.06 & 0.35 & 0.19 \\ 0.15 & 0.07 & 0.02 & 0.76 \end{bmatrix}$$

b. $(0.25, 0.25, 0.25, 0.25)$

2. a. $1 - 3\alpha$

b.
$$\begin{bmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{bmatrix}$$

d.
$$\begin{bmatrix} 0.1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{bmatrix}$$

with (0.25, 0.25, 0.25, 0.25) ultimate distribution of bases

3. a. $1 - \alpha - 2\beta$

Projects

Develop a sequential or a high-performance computing version of each of the projects below.

1. **Epithelial tissue**, composed of layers of cells, is a covering or lining. For example, the outer portion of the skin, linings of the gastro-intestinal system and the lungs, and the outer surface of the cornea are all epithelial tissue. Usually when a cell divides, the daughter cells have one less side than the parent cell but neighboring cells gain sides. It has been observed that virtually no cells are triangular.

Markov chains can be used to model cell shape, specifically the number of sides of their 2D polygonal structure, in dividing sheets of epithelial cells. A Markov chain model for the number of sides in dividing sheets of epithelial cells hypothesizes that the distribution of sides from a dividing cell to two daughter cells follows a binomial distribution with its coefficients from Pascal's triangle, as indicated in Table 4. The table gives a model of the relative odds of a cell of one shape becoming a cell of another shape after division of that cell and its neighbors. For example, the value in row 7, column 8 is 6; and the sum of the values in column 8 is 16. Thus, 6/16 is the probability that a cell with 8 sides will become a cell with 7 sides after its and its neighbors' divisions. The table incorporates the distribution of sides of a dividing cell to its daughter cells and the observed average gain of one side from the division of neighbors. (Gibson, Patel, Nagpal, Perrim, 2006), (Gibson, Patel, Nagpal, Perrim, Supplementary 2006)

Table 4 A model of the relative odds of a cell of one shape becoming a cell of another shape after division of that cell and its neighbors

		Before Division						
		4	**5**	**6**	**7**	**8**	**9**	**10**
After Division	**4**							
	5	1	1	1	1	1	1	
	6		1	2	3	4	5	
	7			1	3	6	10	
	8				1	4	10	
	9					1	5	
	10						1	

a. Develop a Markov chain model for the number of sides in dividing sheets of epithelial cells where the state of a cell is its number of sides, $s > 3$. In developing the model, draw a state diagram, form a transition matrix, determine the stable equilibrium percentages for categories of the number of cell sides, and the average number of sides.

b. Validate the model by comparing these percentages and this average with observations from time-lapse microscopy of three very different animals:

Drosophila wing disk epithelium, the outer epidermis of the fresh water cnidarian *Hydra*, and the tadpole tail epidermis of the frog *Xenopus* (see Table 5). Employ a histogram for your comparisons.

c. Based on your work, are the scientists who developed this model justified in concluding that "the distribution of polygonal cell types in epithelia is not a result of cell packing, but rather a direct mathematical consequence of cell proliferation"?

Table 5 Observed number of cell sides in *Drosophila* wing disk epithelium, the outer epidermis of the fresh water cnidarian *Hydra*, and the tadpole tail epidermis of the frog *Xenopus* (Gibson, Patel, Nagpal, Perrim, 2006)

	Number of Cell Sides							
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Drosophila	0	64	606	993	437	69	3	0
Hydra	0	16	159	278	125	23	1	0
Xenopus	2	40	305	451	191	52	8	2

2. H. S. Horn used Markov chains to model succession in a forest, perhaps from a virgin forest or from a forest after a catastrophic event, such as fire. Using a tree-by-tree replacement process with synchronous replacement of all trees by a new generation, he assumed that "the probability that a given species will be replaced by another given species is proportional to the number of saplings of the latter in the understory of the former." Besides synchrony, he makes additional simplifying assumptions, such as sapling abundance predicts survival to reach the canopy and transition probabilities are constant. A study of Institute Woods in Princeton, New Jersey, yielded the data in Table 6.

Table 6 Transition matrix for Institute Woods in Princeton: percent saplings under various species of trees; BTA - Big tooth aspen, GB - Gray birch, SF - Sassafras, BG - Blackgum, SG - Sweetgum, WO - White oak, OK - Red oak, HI - Hickory, TU - Tuliptree, RM - Red maple, BE - Beech (from Table 1, p. 199, Horn, 1975)

		BTA	**GB**	**SF**	**BG**	**SG**	**WO**	**OK**	**HI**	**TU**	**RM**	**BE**
Sapling species (%)	**BTA**	3	-	3	1	-	-	-	-	-	-	-
	GB	5	-	1	1	-	-	-	-	-	-	-
	SF	9	47	10	3	16	6	2	1	2	13	-
	BG	6	12	3	20	0	7	11	3	4	10	2
	SG	6	8	6	9	31	4	7	1	4	9	1
	WO	-	2	3	1	0	10	6	3	-	2	1
	OK	2	8	10	7	7	7	8	13	11	8	1
	HI	4	0	12	6	7	3	8	4	7	19	1
	TU	2	3	-	10	5	14	8	9	9	3	8
	RM	60	17	37	25	27	32	33	49	29	13	6
	BE	3	3	15	17	7	17	17	17	34	23	80

Species Counts		104	837	68	80	662	71	266	223	81	489	405	

a. Using the Table 6 data, develop a Markov chain model of this forest's succession and determine the stable equilibrium percentages.

b. Using the distribution of species in the last row of Table 7 as the initial distribution and the transition matrix from Part a, plot the estimated number of trees of each species for 20 generations.

c. Trees, however, do not have the same life expectancy, as Table 7 indicates. Thus, Horn weighted (i.e., multiplied) the stationary distribution by the longevities in the table, normalized the result (i.e., divide by the sum of the components), and obtained percentages (i.e., multiplied by 100). Perform these calculations on your stable equilibrium distribution to obtain an age-corrected distribution, which is the analog of a climax community.

Table 7 Longevity (years) of trees in Institute Woods (from Table 2, p. 200, Horn, 1975)

BTA	**GB**	**SF**	**BG**	**SG**	**WO**	**OK**	**HI**	**TU**	**RM**	**BE**
80	50	100	150	200	300	200	250	200	150	300

d. Calculate the relative invasiveness of each species as the sum of the percent saplings under other trees divided by the maximum such sum. That is, to calculate this metric, for each row, calculate the row sum minus the diagonal element; find the maximum of these sums; and divide each row sum minus the diagonal element by this maximum. Discuss how the beech's ability to invade under other species is evident in the probabilities of Table 6.

e. Evaluate a metric for each species' resistance to invasion by other species as follows: Calculate the sum of percentages of other saplings under its canopy (column sum excluding diagonal element); determine the minimum such sum; and for each species, compute this minimum divided by the sum of percentages of other saplings under its canopy. Discuss how the beech's resistance to invasion by other species is evident in the probabilities of Table 6.

f. Calculate a metric for each species' self-replacement as the percentage of its own saplings under its canopy (diagonal element) divided by the maximum such percentage (maximum diagonal element). Discuss how the beech's copious self-replacement is evident in the probabilities of Table 6.

g. Compare your results to those of Horn's data for several sub-forests of varying ages in Institute Woods (see Table 8).

h. Discuss the climax abundance of each species in relationship to its possession of the characteristics of Parts d, e, and f.

Table 8 "The empirical approach results from independent measurements of 639 trees in stands that have been fallow for at least the number of years indicated. The

percentages are of total basal area, calculated from diameters measured at breast height." (from Table 2, p. 200, Horn, 1975)

Years fallow	BTA	GB	SF	BG	SG	WO	OK	HI	TU	RM	BE
25	0	49	2	7	18	0	3	0	0	20	1
65	26	6	0	45	0	0	12	1	4	6	0
150	-	-	0	1	5	0	22	0	0	70	2
350	-	-	-	6	-	3	-	0	14	1	76

3. **Fecal shedding** is the elimination of a pathogen through an animal's fecal matter. Because many diseases spread by fecal shedding, an understanding of the dynamics of contagiousness is important in disease prevention and control. (Ivanek *et al*. 2007) used Markov chain models to study in dairy cattle the dynamics of fecal shedding of the pathogen *Listeria monocytogenes* (LM), a bacterium that causes listeriosis, a disease of the central nervous system. Models with two states, shedding (of LM) and non-shedding, were developed for overall (all subtypes*) L. monocytogenes* shedding considering various combinations of time-dependent risk factors, or **covariates** that can change with time. These covariates include silage (feed) contaminated with LM and stress, such as from antiparasitic treatment.

Using data and statistics and considering the situations of presence or absence of contaminated silage and stress, the scientists estimated the probability of fecal shedding or non-shedding one day (time t -1) leading to the presence or absence of LM in a cow's feces the next day (time t). Thus, they determined $2^3 = 8$ probabilities (see Table 9). With 1 indicating presence and 0 absence of each of the three conditions (contaminated silage, stress, and fecal shedding) the day before, Table 9 gives the probabilities of fecal shedding of LM. For example, the first two rows under the headings, consider the situation in which silage contamination and stress did not exist at time t -1. In this case, the probability of changing from a non-shedding state at time t - 1 to a shedding state at time t is $p_{01} = P$(shedding at time t | non-shedding at time t - 1) = 0.038, while the probability of remaining in a shedding state is $p_{11} = P$(shedding at time t | shedding at time t - 1) = 0.116. Using these two probabilities, we can develop a 2-by-2 Markov matrix. In Table 9, each pair of rows below the headings results in a different model.

Table 9 For all subtypes of *Listeria monocytogenes*, presence (1) or absence (0) of overall LM contamination of silage, stress, and LM fecal shedding at time t - 1 with the probability of LM fecal shedding the next day (time t)

	At time t -1			At time t
Subtypes	**Silage Contam.**	**Stress**	**Fecal Shedding**	**Probability of Fecal Shedding**
All	0	0	0	$p_{01} = 0.038$
	0	0	1	$p_{11} = 0.116$
	0	1	0	$p_{01} = 0.174$
	0	1	1	$p_{11} = 0.410$
	1	0	0	$p_{01} = 0.358$

	1	0	1	$p_{11} = 0.648$
	1	1	0	$p_{01} = 0.746$
	1	1	1	$p_{11} = 0.907$

a. Develop four Markov chain models $\begin{bmatrix} p_{00} & p_{10} \\ p_{01} & p_{11} \end{bmatrix}$ for each covariant situation in Table 9. Starting with an initial distribution at time $t = 0$ of 100% non-shedding cows. For each situation, plot the percent of shedding cows from day 0 through day 10. Determine the long-term distributions, which give the equilibrium probabilities of being in non-shedding and shedding states, or the eventual proportion of time in each state. Discuss the results.

b. The time spent in a state of this model has a geometric distribution. If the initial day (day 0) is non-shedding, the probability of the next day (day 1) being non-shedding, or the proportion of time of a non-shedding day 1, is p_{00}; the probability of days 1 and 2 being non-shedding is $p_{00}p_{00} = (p_{00})^2$; the probability of days 1-3 being non-shedding is $(p_{00})^3$; etc. Thus, the mean time spent in the non-shedding state over a period of n - 1 days is $1 + p_{00} + (p_{00})^2 + \cdots + (p_{00})^{n-1}$. This sum is a **finite geometric series**, which equals $\frac{1 - p_{00}^{\ n}}{1 - p_{00}}$. As n goes to infinity, $(p_{00})^n$ goes to 0 because $0 \le p_{00} < 1$. Thus, for a Markov chain model of Part a, we can estimate the mean time for a cow to spend in a non-shedding state as $\frac{1}{1 - p_{00}}$. Similarly, we can estimate the time for a cow to spend in a shedding state as $1/(1 - p_{11})$. Make such estimates for each of the covariant situations in Table 9, and discuss the results.

c. The models of Part a are **homogenous Markov chain models**, which use the same transition matrix throughout. However, we can employ a **non-homogenous Markov chain model**, where we vary the transition matrix depending on the presence or absence of the time-varying covariates (contaminated silage and stress). Thus, for a real or assumed pattern of time-varying covariates, by employing the appropriate transition matrices, we can examine the changing distributions. Develop a program to accept a sequence of time-varying covariates for a period of 20 days and to plot the percentage of shedding cows versus day. Discuss the results for several patterns.

4. For this project, download the PAM1 matrix, *PAM1.dat*, in Table 10 and the frequency data, *freq.dat*, in Table 11. Finding similar sequences in genomic databases can help us determine the biochemistry, physiology, and function of a gene or the protein it produces. In searching such databases, algorithms produce scores that allow us to differentiate sequences that are related to a query sequence from those that are not. One of the main algorithms for database searching is **BLAST** (Basic Local Alignment Search Tool) (BLAST 2012), which uses a **PAM** (**Point Accepted Mutations**) scoring matrix. In the 1970s, a research team lead by Margaret Dayhoff carefully studied the evolution of sequences of amino acids. **PAM** or **PAM 1** is the length of time for 1% of the amino acids to mutate. One

estimate is that a PAM is about a million years. The **PAM1 matrix** is a Markov chain transition matrix with column and row headings of the amino acids where entries represent the amount of evolution over one PAM period of time, or for one mutation per hundred amino acids. Thus, the *ij* element is the probability that the amino acid in the *i*th row will replace the amino acid in the *j*th column after the evolutionary time PAM. A **PAM120 matrix**, which BLAST uses, contains information on the amount of evolution over 120 PAM periods of time. We can obtain this matrix by raising the PAM1 matrix to the 120th power. Use a computational tool as needed to complete the following parts. (Shiflet, 2002)

a. The values are multiplied by 10,000 for clarity. For example, the element in the first row for Ala (A) and third column for Asn (N) is 3. Thus, the probability that the amino acid Asn mutates to the amino acid Ala in about a million years (one PAM epoch) is 3/10,000 = 0.0003 = 0.03%. Draw a partial state diagram using the four amino acids in the top left corner of the matrix.

Table 10 PAM1 matrix with values multiplied by 10,000. The element in row *i*, column *j* is the probability that row *i*'s amino acid will replace column *j*'s amino acid in 1 PAM. "(Adapted from Figure 82. Atlas of Protein Sequence and Structure, Suppl 3, 1978, M.O. Dayhoff, ed. National Biomedical Research Foundation, 1979.")

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

b. Calculate PAM120, *M*. The matrix is usually written with each element multiplied by 100 and rounded to the nearest integer.

c. Each element of the **PAM120 scoring matrix**, *S*, which BLAST uses, is obtained using the following formula:

$$S_{ij} = \text{round}(10 \log_{10}(M_{ij} / f_i)),$$

where f_i is the frequency of the amino acid in row *i* and *M* is the PAM120 matrix from Part b. We will compute *S*, called a **log odds scoring matrix**, using the frequencies in Table 11. Because we do not know what came first, make this matrix symmetric, using the values on and below the diagonal. For example, the score for a mutation over 120 PAM periods from R to N should be the same as the mutation over that period from N to R. (Momand 2006)

Table 11 Normalized frequencies of amino acids (Nakhleh 2010)

Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
8.7%	4.1%	4.0%	4.7%	3.3%	3.8%	5.0%	8.9%	3.4%	3.7%

Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

8.5%	8.1%	1.5%	4.0%	5.1%	7.0%	5.8%	1.0%	3.0%	6.5%

d. Write a function to return the relative position of an amino acid parameter. For example, N is the third amino acid listed in Table 10, so the function returns 3.

e. Write a function to accept two amino acids, such as N and A, as arguments and to return the corresponding PAM120 score using the PAM120 scoring matrix, *S*, from Part c.

f. The BLAST algorithm searches a database for sequences that have a "good," non-gapping local alignment with a segment of the query sequence. The program starts by breaking the query sequence into all possible sequential triplets, or **3-mers**, or **words** of length 3. For example, if the query sequence is *s* = RHQMN, we have three 3-mers, RHQ, HQM, and QMN.

Write a function that has a query sequence parameter and returns a list of all its 3-mers.

g. The BLAST program obtains the evolutionary scores for all possible (20)(20)(20) = 8,000 amino acid triplets in relation to each of the 3-mers in the query sequence and compiles a list of all words that have a score greater than or equal to a certain threshold parameter. For example, using the PAM250 scoring matrix in Table 12, the scoring of QMN relative to pairs QMN, DLL, QSW, and BME is 12, 3, -2, and 8, respectively, as the following computations indicate:

Q		M		N		
Q		M		N		
4	+	6	+	2	=	12

Q		M		N		
D		L		L		
2	+	4	+	(-3)	=	3

Q		M		N		
Q		S		W		
4	+	(-2)	+	(-4)	=	-2

Q		M		N		
B		M		E		
1	+	6	+	1	=	8

If the user picks a threshold value of 5, the program would select QMN and BME and other 3-mers but not DLL and QSW as evolutionary matches.

Develop a function to accept two 3-mers and to return the evolutionary PAM120 score. These scores will differ from those in Table 12.

Table 12 The PAM250 scoring matrix for amino acids. B is used when one cannot distinguish between D and N because of amino acid analytical processing. Similarly, Z is

used when it is ambiguous whether the amino acid is E or Q. X represents an unknown or nonstandard amino acid. Thus, the matrix has 23 rows and 23 columns.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	2																						
R	-2	6																					
N	0	0	2																				
D	0	-1	2	4																			
C	-2	-4	-4	-5	12																		
Q	0	1	1	2	-5	4																	
E	0	-1	1	3	-5	2	4																
G	1	-3	0	1	-3	-1	0	5															
H	-1	2	2	1	-3	3	1	-2	6														
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5													
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6												
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5											
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6										
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9									
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6								
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2							
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3						
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17					
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10				
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4			
B	0	-1	2	3	-4	1	2	0	1	-2	-3	1	-2	-5	-1	0	0	-5	-3	-2	2		
Z	0	0	1	3	-5	3	3	-1	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X

h. Develop a function to have three parameters, a 3-mer (*mer*), a list of 3-mers (*merLst*), and a threshold value (*threshold*), and to return a list of all 3-mers from *merLst* whose evolutionary score relative to *mer* is greater than or equal to *threhold*. For example, as Part g illustrates, if *mer* is QMN, *merLst* is {QMN, DLL, QSW, BME}, and *threshold* is 5, then the function returns {QMN, BME}. Use the PAM120 scoring matrix from Part c.

i. Write a function to return a list of the 8,000 possible amino acid triplets.

j. The second step in the BLAST algorithm is to scan the database for locations of high scoring words from the first step (see Part g). For example, the high scoring word BME occurs at location 6 in the sequence NRSQH**BME**LDLDMFPMST.

Develop a function that has as parameters a list of 3-mers (*merLst*) and a sequence (*sequence*) and that returns a list of integer starting locations for all occurrences 3-mers from *merLst* in *sequence*.

k. The third step of the BLAST algorithm is to extend each of the seeds in both directions until the subsequence score reaches a maximum value according to the matrix scoring. Using a heuristic, the program stops an extension if the score falls below a certain amount less than the highest score so far. For example, suppose the query sequence is in part …SRMCD**RHQ**MNCFPS…, and the program located high scoring word RHQ in the database sequence …NRSQH**RHQ**LDLDMF…. Table 13 shows how we extend from the seed RHQ to find a segment pair (D**RHQ**MN and H**RHQ**LD) with a maximum PAM250 score (1 + 6 + 6 + 4 + 4 + 2 = 23). DRHQMN and HRHQLD are a

locally maximal segment pair, or a segment from the query sequence and a segment from a database sequence with a score that cannot become larger through shrinking or expanding the segments. We repeat this extension process for all seeds looking for all segment pairs with scores above some threshold. The algorithm is fast in part because it does not consider gaps and uses heuristics involving threshold values.

Develop a function that has parameters of two sequences and an integer starting location and returns a list containing the starting location, length, and PAM 120 score (see Part c) of a locally maximal segment pair.

Table 13 Finding the locally maximal segment pair from sequences SRMCD**RHQ**MNCFPS and NRSQH**RHQ**LDLDMF, starting at location 6 and using the PAM250 scoring matrix

In query:	S	D	M	C	**D**	**R**	**H**	**Q**	**M**	**N**	C	F	P	S
In database:	N	R	S	Q	**H**	**R**	**H**	**Q**	**L**	**D**	L	D	M	F
PAM250 Score:	1	-1	-2	-5	**1**	**6**	**6**	**4**	**4**	**2**	-6	-6	-2	-3

l. Write a program to implement the BLAST algorithm as presented in this project. Input should include a query sequence, a list of database sequences, and a threshold value. Obtain sequences from a BLAST database at NCBI (BLAST 2012).

5. (Prerequisite: (Shiflet and Shiflet 2009), sections on "Cellular Automaton Simulation," "Boundary Conditions," and Growth Algorithm) The **Stepping Stone Model** is useful in the study of genetics. For the model, we start with an n-by-n grid (matrix) with each cell (element) having one of k integer values. Repeatedly, we select a cell at random and select one of its eight neighbors at random. We then change the value at the cell to be the value of the selected neighbor. Periodic boundary conditions are employed. A grid represents a state of the system. Thus, with each grid having n^2 cells and each cell having k possible values, the system has k^{n^2} possible states. For example, a small 10-by-10 grid with values only of 0 and 1 has $2^{10^2} = 2^{100} = 1.2677 \cdot 10^{30}$ possible states. A transition matrix with this number of states would have an excessive number of elements: $10^{30} \cdot 10^{30} = 10^{60}$ elements. However, we can employ cellular automaton simulations to simulate the Markov chain (Grinstead and Snell 2003).

a. Develop the Stepping Stone Model using $n = 20$ and $k = 2$ (values *num1* = 1 and *num2* = 2). Employ a random initial configuration, where the probability of p for one of the cell values, *num1*. Using visualizations of the grid with white representing one cell value and black representing the other, develop an animation. Run the animation a number of times with different values of p and observe regions of color and the ultimate "winner." Does the winner seem related to p? Discuss the results.

b. Repeat Part a without the animation but plotting the number of each color at each time step. Discuss the results.

c. Use HPC for this part. Repeat Part a without the animation. Develop a simulation for a large constant number of time steps and calculate the number of cells with value *num1* on completion. For each p value from 0 to 1, varying by 0.1, run the simulation 100 times and compute the average number of cells with final value *num1*. Plot the average versus p. Does the winner seem related to p? Discuss the results.

d-f. Repeat Parts a, b, and c, respectively, using $k > 2$.

6. Download the serial and parallel *cpg* programs and associated data files from the Blue Waters site. Using commands *tic* and *toc*, time the serial program *cpg.c* with data sets *ecoli-sequence.txt* that contains a genomic subsequence of *E. coli* and *ProbabilitiesEcoli.txt* that stores the probability matrices (see comments at the beginning of *cpg.c* or *cpgParallel.c*). Repeat the timings using the parallel program *cpgParallel.c* for an increasing number of processes, n. Plot the speedup factor, $S(n)$, versus the number of processes, n, where the **speedup factor $S(n)$** is as follows:

$$\boldsymbol{S(n)} = \frac{\text{Execution time on sequential computer}}{\text{Execution time on system with } n \text{ processes}}$$

Usually, the maximum speedup possible with n processes is $S(n) = n$, which we call **linear speedup**. This situation is achieved when the time required for execution with n processes is $1/n$ of the time for execution on a sequential computer. Linear speedup is rarely achieved because of overhead factors, such as communications, times when some processes are idle, and additional computations required for the parallel version. Describe the shape of the graph, and discuss how well the problem scales to larger numbers of processes.

7. From the Blue Waters site, download *ProbabilitiesHuman.txt*, which contains the probabilities from Tables 7 and 6, respectively. Also, download all or part of the DNA sequence on chromosome 19 of the human genome at http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?ORG=hum&MAPS=ideogr,est,loc&LINKS=ON&VERBOSE=ON&CHR=19.

a. Employ the techniques of the section on "Locating Genes" to score each subsequence of length 200. Have your program determine the most likely candidates for subsequences being in CpG islands. Do your candidates occur in CpG islands as indicated at http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr19:571325-583493&hgsid=264592883&knownGene=pack&hgFind.matches=uc002loy.3 ? As of this writing (5/23/12), such areas appear in green on the diagram (*Homo Sapiens* 2001; UCSC 2009).

b. Using this data set, repeat Project 6.

8. From the Blue Waters site, download *AE005174v2.txt* (from *AE005174v2.fas* at http://www.genome.wisc.edu/sequencing/o157.htm), which contains the DNA sequence for *Escherichia coli* (*E. coli*), and *Escherichia_coli_O157H7_plasmid_pO157.txt*, which contains training data

generated by generated by GeneMarkSPlusRBS on *E. coli* O157H7 as described in the section on GeneMark (Borodovsky Laboratory 2005). GeneBank at NCBI contains sequence information on "Escherichia coli O157:H7 EDL933, complete genome" (*http://www.ncbi.nlm.nih.gov/nuccore/AE005174*). By clicking on any of the *gene* links, create a data file of a sequence of 200 bases immediately before one of the genes and create another data file of a sequence of 200 bases inside a gene (Escherichia coli 2001; Enterohaemorrhagic 2001).

a. Using the algorithm described in the section on "GeneMark" to score each sequence as containing or not containing a CpG island. Employ the first column of data in *Escherichia_coli_O157H7_plasmid_pO157.txt*.

b. Using this data set, repeat Project 6.

9. Download *Escherichia_coli_O157H7_plasmid_pO157.txt*, which is described in the previous project. For **homogeneous Markov models** involving genomic sequences, probabilities are not dependent upon sequence location, while for **inhomogeneous Markov models** they are. A **reading frame** breaks a sequence of nucleotides into codons. Because we can start the alignment in three possible places an mRNA strand, three possible reading frames exist for such a strand. For example, suppose mRNA contains the sequence of bases AACTGTTAG…. We could have the reading frame begin with AAC, as in AAC-TGT-TAG…; or one base further with ACT-GTT-AG…; or two bases beyond with CTG-TTA-G…. Because DNA has two strands, one complementary to the other, we have six possible reading frames from which transcription can occur. As described in the section on "GeneMark," develop a program that generates transition matrices and probabilities for the six possible reading frames for training sequences and select the model with the highest score. The GeneMark program considers seven possibilities, these six and a model of non-coding DNA.

10. For this project, download *J_SNP.dat*. **SNPs** (pronounced "snips"), single nucleotide polymorphisms, are variations in DNA sequence where one nucleotide in the sequence is changed. In human beings, SNPs occur every 100 to 300 bases (SNP 2012). (Lieberman *et al* 2011) studied an outbreak of the bacterial pathogen Burkholderia dolosa among 14 individuals with cystic fibrosis. Their data contains the sequences for 112 samples taken over 16 years. In the patient id, the letter, such as "J", indicates the patient; the number after the first dash indicates the year after the start of the study; the number after the second dash is the month; and the small letter is a sample. Thus, identifier J_11_8 is a sample for patient J taken 11 years and 4 months after the start of the study. Sometimes difficulties occur in identifying a nucleotide. The following are nucleotide designations besides A, C, T, and G with their meaning in parentheses: R (A or G), Y (C or T), S (G or C), W (A or T), K (G or T), M (A or C), B (C or G or T), D (A or G or T), H (A or C or T), V (A or C or G), N (any base), - (gap). Create transition matrix using J-11-8 to J-11-11, and using this matrix, estimate the ultimate distribution of bases.

Answers to Quick Review Questions

1. ¼ = 0.25 = 25%

2. $0.75 = 1 - 0.25 = 1 - P(\mathrm{T})$; alternatively, $0.75 = P(\mathrm{A}) + P(\mathrm{C}) + P(\mathrm{G})$.
3. ½ = 0.5 = 50%
4. 0.13 = 0.10 + 0.04 - 0.01
5. 1/16 = (¼)(¼)
6. **a.** 7/20 = 0.35
b. 5/20 = 0.25
c. 2/20 = 0.10
d. 2/7 = 0.286 because C occurs in s_1 7 times
e. 2/7 = (2/20) / (7/20)
f. They are equal.
7. **a.** $P(X_{n+1} = \mathrm{E} \mid X_n = \mathrm{R}) = 0.2$
b. $P(X_{n+1} = \mathrm{R} \mid X_n = \mathrm{R}) = 0.8 = 1 - 0.2$
8. **a.** $\begin{bmatrix} 0.3 & 0.4 & 0.2 \\ 0.1 & 0.3 & 0.0 \\ 0.6 & 0.3 & 0.8 \end{bmatrix}$
b. E: 25%, G: 6%, R: 69% because the product of T from Part a and (0.3, 0.1, 0.6), expressed as a column vector, is (0.25, 0.6, 0.69), expressed as a column vector.
c. $\begin{bmatrix} 0.229508 & 0.229508 & 0.229508 \\ 0.0327869 & 0.0327869 & 0.0327869 \\ 0.737705 & 0.737705 & 0.737705 \end{bmatrix}$
d. (0.229508, 0.0327869, 0.737705)
9. **a.** 1
b. Any nonzero multiple of (-0.296799, -0.0423999, -0.953998)
c. (0.229508, 0.0327869, 0.737705) obtained by multiplying the vector from Part b by one over the sum of its elements, s = -1.2932
d. E: 23%, G: 3%, R: 74% obtained by expressing as percentages the elements of the vector from Part c
10. **a.** bioinformatics
b. 20
c. proteins
d. enzymes
e. amino acids
f. N-terminal
g. C-terminal
11. **a.** DNA
b. RNA
c. nucleotide
d. nucleotide
e. A, G, C, T
f. A, G, C, U
g. A, G
h. C, T, U
i. T
j. U

k. G
l. A
m. A
n. C

12. a. RNA
b. DNA
c. deletion, insertion, transition, transversion
d. transition
e. transversion
f. transition

13. a. chromosome
b. gene
c. genome
d. triplet
e. mRNA
f. codon
g. transcription

14.

x_{i-1} \ x_i	**A**	**C**	**G**	**T**
A	0.00	0.50	0.00	0.50
C	0.00	0.00	0.50	0.50
G	0.00	0.00	0.00	1.00
T	0.25	0.50	0.00	0.25

15. a. $P_+(\text{CCGTCGA}) = 4.7767 \cdot 10^{-5} = 0.368 \cdot 0.274 \cdot 0.125 \cdot 0.355 \cdot 0.274 \cdot 0.161 \cdot 0.242$
b. $P_-(\text{CCGTCGA}) = 4.5822 \cdot 10^{-6} = 0.298 \cdot 0.078 \cdot 0.208 \cdot 0.239 \cdot 0.078 \cdot 0.248 \cdot 0.205$
c. 10.4245
d. CCGTCGA more likely to be from a CpG island because the quotient is greater than 1.

16. $0.3349 = \ln((0.368 \cdot 0.274 \cdot 0.125 \cdot 0.355 \cdot 0.274 \cdot 0.161 \cdot 0.242)/(0.298 \cdot 0.078 \cdot 0.208 \cdot 0.239 \cdot 0.078 \cdot 0.248 \cdot 0.205))/7 = (\ln(0.368) + \ln(0.274) + \ln(0.125) + \ln(0.355) + \ln(0.274) + \ln(0.161) + \ln(0.242) - \ln(0.298) - \ln(0.078) - \ln(0.208) - \ln(0.239) - \ln(0.078) - \ln(0.248) - \ln(0.205)) / 7$

References

Agnew, Jeanne and Robert C. Knapp 2002. *Linear Algebra with Applications*. Monterey, Calif.: Brooks/Cole Pub. Co.

"Adrenoleukodystrophy," University of Maryland Medical Center. 2011. http://www.umm.edu/ency/article/001182trt.htm. Accessed 02/21/12

"Barnacles." *Science Encyclopedia*. 2012. http://science.jrank.org/pages/752/Barnacles.html. Accessed 02/21/12.

"Barnacles will cling no more with self-cleaning, non-toxic coating for ships developed by Cornell researchers," *Cornell News*. 2003.

http://www.news.cornell.edu/releases/March03/ACS.Ober.deb.html/. Accessed 02/21/12.

BLAST. 2012. Blast Home at National Center for Biotechnology Information. http://blast.ncbi.nlm.nih.gov/Blast.cgi. Accessed 01/19/12.

Borodovsky Laboratory. 2005. Escherichia_coli_O157H7_plasmid_pO157.mat. School of Biology, Georgia Tech.

"California's Rocky Intertidal Zones, " exerpts from the California Coastal Commission's *California Coastal Resource Guide*. 1987. http://ceres.ca.gov/ceres/calweb/coastal/rocky.html. Accessed 02/21/12.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Enterohaemorrhagic *Escherichia coli* (EHEC) O157:H7. *E. coli* Genome Project. 2001. http://www.genome.wisc.edu/sequencing/o157.htm. Accessed 5/23/12.

Escherichia coli O157:H7 EDL933, complete genome. 2001. NCBI. http://www.ncbi.nlm.nih.gov/nuccore/AE005174. Accessed 5/23/12.

Foster, Dr. Rick. "Intertidal Communities Overview." ADF&G. 2000. Alaska Department of Fish and Game. 11 Feb. 2009 .http://www.habitat.adfg.state.ak.us/geninfo/kbrr/coolkbayinfo/kbec_cd/html/ecosys/estuarin/rocky.htm.

Gardiner-Garden, M. and M. Frommer. 1987. "CpG islands in vertebrate genomes." *J. Mol. Biol.* 196, 261-282.

Gibson, Matthew C., Ankit B. Patel, Radhika Nagpal and Norbert Perrim. 2006. "The emergence of geometric order in proliferating metazoan epithel." *Nature* 442, 1038-1041, August 31, 2006.

Gibson, Matthew C., Ankit B. Patel, Radhika Nagpal and Norbert Perrim. 2006. "The emergence of geometric order in proliferating metazoan epithel," Supplementary Material. genetics.med.harvard.edu/~perrimon/papers/GibsonM_Supp_Nature.pdf

Grinstead, Charles and Laurie Snell. 2003. *Introduction to Probability*, Chapter on "Markov Chains." American Mathematical Society.

Gropl, Clemens and Daniel Huson, 2005. "Hidden Markov Models." February17, 2005,19:15 http://www.inf.fu-berlin.de/inst/ag-bio/FILES/ROOT/Teaching/Lectures/WS0405/aldabi/script-hmm.pdf. Accessed 11/18/11.

Guide to the Human Genome, "Chromosomes and DNA," 2010. Cold Spring Harbor Lab Press http://www.cshlp.org/ghg5_all/section/dna.shtml. Accessed 05/05/12.

Homo Sapiens (human), Chromosome 19. NCBI Map Viewer. 2001. http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?ORG=hum&MAPS=ideogr,est,loc&LINKS=ON&VERBOSE=ON&CHR=19. Accessed 5/23/12.

Horn, H. S. 1975. Forest Succession, Scientific American, 232: 90-98.

Horn, H. S. 1975. Markovian properties of forest succession, pp. 196-211 in *Ecology and Evolution of Communities*, M. L. Cody and J. M. Diamond, eds., Harvard University Press, Cambridge, Massachusetts.

"Human Genome Project Information," Oak Ridge National Laboratory. 2008. http://www.ornl.gov/sci/techresources/Human_Genome/faq/seqfacts.shtml. Accessed 02/21/12.

International Human Genome Sequencing Consortium. 2004. "Finishing the euchromatic sequence of the human genome". *Nature* 431 (7011): 931–45.

"Intertidal Stressors," *Ocean News*. 2007. http://oceanlink.island.net/ONews/ONews7/intertidal.html. Accessed 02/21/12.

Ivanek, Renata, Yrjo T. Grohn, Alphina Jui-Jung Ho, Martin Wiedmann. 2007. "Markov chain approach to analyze the dynamics of pathogen fecal shedding—Example of Listeria monocytogenes shedding in a herd of dairy cattle." *J. Theor. Biol*. 245 (2007) 44–58.

Lieberman, T.D. *et al*. 2011. "Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes" *Nature Genetics* 43: 1174-1176.

Momand, Jamil. 2006. "Scoring Matrices," Southern California Bioinformatics Summer Institute. http://instructional1.calstatela.edu/jmomand2/2006/curriculum/ppt/scoring_matrices.ppt. Accessed 01/17/12.

"Mysteries of Life: From Molecules to Mice," Oak Ridge National Laboratory. 1999. http://www.ornl.gov/info/ornlreview/v32_2_99/from.htm. Accessed 02/21/12.

Nakhleh, Luay K. 2010. "Pairwise Sequence Alignment (II)," COMP 571 Bioinformatics: Sequence Analysis. http:// www.cs.rice.edu/~nakhleh/COMP571/Presentation3.ppt. Accessed 01/17/12.

Rupp, Bernhard. 2000. "Protein Structure Basics," UCRL-MI-125269, Lawrence Livermore National Laboratory. Originally accessed in 2000 from http://www.structure.llnl.gov/Xray/tutorial/protein_structure.htm. Currently available at http://www.ruppweb.org/Xray/tutorial/protein_structure.htm. Accessed 10/30/11.

Salzberg, S.L., A.L. Delcher, S. Kasif, and O. White. 1998. "Microbial gene identification using interpolated Markov models." *Nucleic Acids Research*, 26:2, 544-548.

"The Secret Life of "Barnacles," Natural History Museum. 2012. www.nhm.ac.uk/resources-rx/files/10feat_secret_life_of_barnacles-3061.pdf. Accessed 02/21/12.

Shiflet, Angela and George Shiflet. 2009. "Biofilms: United They Stand, Divided They Colonize" UPEP Curriculum Module http://shodor.org/petascale/materials/UPModules/biofilms/. Accessed 01/17/12.

Shiflet, Angela. 2002. "Searching Genomic Sequence Databases." http://wofford-ecs.org/DataAndVisualization/GenomicSearching/index.htm. Accessed 01/17/12.

Sinha, Saurabh. 2007. "Evolutionary Models," lecture notes in CS 498. www.cs.uiuc.edu/class/fa07/cs498ss/lectures/LectureEvoModel.ppt

"SNP Fact Sheet" 2012. Oak Ridge National Laboratory. http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml. Accessed 01/14/12.

Stout, Prentice. "Barnacle," Rhode Island Sea Grant Fact Sheet. 2009. http://seagrant.gso.uri.edu/factsheets/597barnacle.html. Accessed 02/21/12.

"Talking Glossary of Genetic Terms," National Human Genome Research Institute, http://www.nhgri.nih.gov/DIR/VIP/Glossary.

Tang, Haixu. 2007. "Probabilistic sequence modeling II: Markov chains," lecture notes for Bioinformatics in Molecular Biology and Genetics: Practical Applications

http://darwin.informatics.indiana.edu/col/courses/I529/Lecture/lec-4.ppt. Accessed 11/18/11.

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly. 2009. http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr19:571325-583493&hgsid=264592883&knownGene=pack&hgFind.matches=uc002loy.3. Accessed 5/23/12.