

## Modeling Molecular Evolution

Natural selection is the fundamental mechanism through which evolution occurs, but for selection to be possible there must be some underlying variability in genetic makeup within a species. Since selection usually acts to reduce variability, there must also be a source of new genetic variation. This is introduced at the molecular level, in the DNA of individuals, through what are viewed as random changes as the molecules are copied into new generations.

Depending on the nature of these changes in the DNA, offspring may be more, less, or equally viable than the parents. Many of the molecular changes are believed to be selectively neutral, and so are passed on to further descendants and preserved. The DNA within a particular gene may continue to mutate from generation to generation, gradually accumulating more differences from its ancestral form. Thus several species arising from a common ancestor will have similar, but often not identical, DNA forming a particular gene. The similarities hint at the common ancestor, while the differences point to the evolutionary divergence of the descendants.

Since we can now ‘read’ the structure of DNA with relative ease, a natural and compelling question arises: Can we reconstruct evolutionary relationships between several modern species by comparing the DNA sequences of their versions of a certain gene?

We of course expect that species that have more similar genetic sequences are probably more closely related. However, this observation really isn’t enough to make clear how to deduce an evolutionary tree relating a large number of different species, all with varying degrees of similarity in the chosen gene. In fact, we need to first decide what we might mean by a phrase like ‘degree of similarity’.

In this chapter we’ll develop mathematical models of DNA mutation processes, that is, of *molecular evolution*. Since the language of probability is needed to describe random mutations, we’ll present the basics of that subject along the way. We’ll then see that probability naturally leads us to linear models to describe molecular evolution. The concept of a *phylogenetic distance* as a measure of sequence similarity will emerge from these models. Then, in the next chapter, the material developed here will help address the issue of deducing evolutionary relationships.

### 1. Background on DNA

Genetic information is encoded by DNA molecules, which are passed from parent to offspring. For this transfer, the DNA must be copied. Despite rather elaborate mechanisms to ensure the correctness of the copying process, sections of the molecule may be altered in various ways. Before modeling the most important of these mutations, we need to briefly review the basic structure of DNA.

The DNA molecule forms a double helix, a twisted ladder-like structure. At each of the points where the ladder's upright poles are joined by a rung, one of four possible molecular subunits appears. These subunits, called *nucleotides* or *bases*, are adenine, guanine, cytosine, and thymine, and are denoted by the letters *A*, *G*, *C*, and *T*. Because of chemical similarity, adenine and guanine are called *purines*, while cytosine and thymine are called *pyrimidines*.

Each base has a complementary base with which it can form the rung of the ladder through a hydrogen bond. We always find either *A* paired with *T*, or *G* paired with *C*. Thus knowing one side of the ladder structure is enough to deduce the other. For example if along one pole of the ladder we have a sequence of bases

$$AGCGCGTATTAG,$$

then the other would have the complementary sequence

$$TCGCGCATAATC.$$

Finally, the DNA molecule has a directional sense so that we can make a distinction between a sequence like *ATCGAT* and the inverted sequence *TAGCTA*. The upshot of all this structure is that we will be able to think of DNA sequences mathematically as simply sequences composed of the four letters *A*, *T*, *C*, and *G*.

Some sections of DNA form *genes* that encode instructions for the manufacturing of proteins (though the production of the protein is accomplished through the intermediate production of messenger RNA). In these genes, triplets of consecutive bases form *codons*, with each codon specifying a particular amino acid to be placed in the protein chain according to the *genetic code*. For example the codon *TGC* always means that the amino acid cysteine will occur at that location in the protein. Certain codons also signal the end of the protein sequence. Since there are  $4^3 = 64$  different codons, and only 20 amino acids and one 'stop' command, there is some redundancy in the genetic code. For instance, in many codons the third base has no effect on the particular amino acid the codon specifies.

While originally it was thought that genes always encoded for proteins, we now know that some genes encode the production of other types of RNA which are the 'final products' of the gene, with no protein being produced. Finally, not all DNA is organized into the coding sections referred to as genes. About 97% of human DNA, for example, is believed to be non-coding. Some of this is likely to be meaningless raw material (sometimes called *junk DNA*) which may, of course, become meaningful in future generations through evolution. Other parts of the DNA molecules may serve regulatory purposes. The picture is quite complicated and still not fully understood.

When DNA is copied, the hydrogen bonds forming the rungs of the ladder are broken, leaving two single strands. Then new double strands are formed on these, by assembling the appropriate complementary strands. The biochemical processes are elaborate, with various safeguards to ensure that few mistakes are made. Nonetheless, changes of an apparently random nature sometimes occur.

The most common mutation that is introduced in the copying of sequences of DNA is a *base substitution*. This is simply the replacement of one base for another at a certain site in the sequence. For instance, if the sequence *AATCGC* in an ancestor becomes *AATGGC* in a descendent, then a base substitution  $C \rightarrow G$  has occurred at the fourth site. A base substitution that replaces a purine with a

purine, or a pyrimidine for a pyrimidine, is called a *transition*, while an interchange of these classes is called a *transversion*. Transitions are often observed to occur more frequently than transversions, perhaps because the chemical structure of the molecule changes less under a transition than a transversion.

Other DNA mutations sometimes observed include the deletion of a base or consecutive bases, the insertion of a base or consecutive bases, and the inversion (reversal) of a section of the sequence. All these mutations tend to be seen more rarely in natural populations. Since these types of mutations usually have a dramatic effect on the protein for which a gene encodes, this is not too surprising. We'll ignore such possibilities, in order to make our modeling task both clearer and mathematically tractable.

Focusing solely on base substitutions, a basic problem to be addressed is how to deduce the amount of mutation that must have occurred during the evolutionary descent of DNA sequences. For instance, suppose we know that a descendent species  $S_2$  descended from an intermediate species  $S_1$ , which in turn descended from an ancestral species  $S_0$ . Imagine that for each of these a certain gene included the sequences:

$$S_0 : \mathbf{ACCTGCGCTA} \dots$$

$$S_1 : \mathbf{ACGTGCACTA} \dots$$

$$S_2 : \mathbf{ACGTGCGCTA} \dots$$

Here boldface marks the two sites among the first ten where changes have occurred. (We'll always assume the sequences have been *aligned* so that we can match ancestral and descendent sites. The mathematical methods by which this can be done could be the subject of another chapter or book.)

Now if we only saw the sequences for  $S_0$  and  $S_2$ , we would notice only one base substitution among the first 10 sites, the one appearing in the third site. It might seem reasonable that the ratio  $\frac{1}{10}$  of mutations per site would be a good measure of how much mutation has occurred from  $S_0$  to  $S_2$ .

However, since we have the sequence for  $S_1$  as well, we know things are more complicated. At the seventh site we notice that we've had the substitutions  $G \rightarrow A \rightarrow G$ . The original mutation has been *hidden* since a *back mutation* has occurred, leaving the final base the same as it initially was. Comparing  $S_0$  to  $S_1$  and then  $S_1$  to  $S_2$  has shown 3 mutations among the first ten sites, leading to the much larger measure of  $\frac{3}{10}$  mutations per site.

It could also happen that at another site substitutions such as  $A \rightarrow T \rightarrow G$  occur. Here, even though there were two consecutive substitutions, we would notice only one if we only saw the initial and final sequences. Once again, a mutation has been hidden by a subsequent one.

Thus a simple ratio of mutations per site obtained from comparing the first and last sequence may well give too low an estimate of the amount of mutation that actually occurred. Unless we believe that mutations have been quite rare, so that no hidden mutations occurred, we will need a mathematical model to be able to reconstruct the number of mutations that are likely to have occurred from those we see in comparing only the initial and final DNA sequences.

## 2. An introduction to probability

Describing the random mutation of DNA mathematically requires a facility with basic probability. While we'll keep our discussion as informal as possible, we will need to be careful on a few points and that requires some terminology. Looking at some familiar non-biological examples, such as coin flips and die tosses, will help make the ideas clearer.

Suppose we flip a coin or toss a die. When we refer to the *probability* of a certain outcome, such as getting a heads in the coin flip, or a 4 in the die toss, we mean a number  $\mathcal{P} = \mathcal{P}(\text{outcome})$ , with  $0 \leq \mathcal{P} \leq 1$ , that indicates the likelihood of that outcome occurring. For instance, if we flip a fair coin, we would say the probability of the outcome 'heads' is

$$\mathcal{P} = \frac{1}{2}, \text{ or } \mathcal{P}(\text{heads}) = \frac{1}{2},$$

since we expect to get see a heads in roughly 1 of every 2 tosses. This doesn't mean that if we flip the coin twice we will get one head and one tail, but rather that if we flipped it a very large number of times, we should find that in about  $\frac{1}{2}$  of the tosses each outcome occurred. For the die toss, to express the chance of a 4 turning up we'd say that  $\mathcal{P}(4) = \frac{1}{6}$ , since we expect roughly 1 of every 6 in a large number of tosses to produce a 4.

We might say that a probability measures the chance of a 'random' outcome occurring. Alternately, we may believe the outcome of a die toss is not random (it is, after all, governed by the deterministic laws of physics), but predicting it is too complicated to be practical. With this viewpoint, we are willing to give up trying to say exactly what will happen with any particular toss, and instead accept a description of how often outcomes are likely to occur in the long run. More precisely, the probability  $\mathcal{P}$  of an outcome gives our expectation of the percentage of trials in which that outcome will occur, assuming a very large number of trials are performed. The smaller  $\mathcal{P}$  is, the less likely we believe an outcome is to occur in any given trial.

Usually, a probability will not indicate exactly what will happen in any trial. However, there are two exceptions. A probability of  $\mathcal{P} = 1$  means an outcome is sure to happen — it will occur 100% of the time. Likewise, a probability of  $\mathcal{P} = 0$  means the event is sure not to happen.

Don't assume that the probability of a heads in a coin flip is  $\frac{1}{2}$  just because there are only two possible outcomes, heads and tails. For a weighted coin, there are still only two possible outcomes, but it might be that with such a coin we expect to get heads in 80% of the flips and so we have  $\mathcal{P}(\text{heads}) = .8$ . Such a coin is not 'fair', but it is still capable of being described through probability. Similarly, for a fair die, the probability of any particular outcome is  $\frac{1}{6}$ , but for a weighted die, the probabilities of some of the outcomes might be more than  $\frac{1}{6}$ , while for others they are less than  $\frac{1}{6}$ .

Given a weighted coin, how can we determine the probability of it producing an outcome of heads? We simply perform many trials by flipping it repeatedly. After recording how often heads comes up in these trials, we can compute the estimate

$$\mathcal{P}(\text{heads}) \approx \frac{\text{number of heads produced}}{\text{number of trials}}.$$

For instance if out of 10 trials, we got 4 heads, we'd estimate  $\mathcal{P}(\text{heads}) \approx \frac{4}{10} = .4$ . Performing 100 trials might turn up 56 heads, leading us to the improved estimate  $\mathcal{P}(\text{heads}) \approx \frac{56}{100} = .56$ . The more trials we perform, the more confidence we have in our estimate of the probability. While we can't prove a typical coin gives us heads and tails with probability  $\frac{1}{2}$ , we can gather evidence to back up that belief.

EXAMPLE. To apply this language to a DNA sequence, suppose a 40 base sequence reads as follows:

*AGCTTCCGATCCGCTATAATCGTTAGTTGTTACACCTCTG*

What is the probability that the next base, in site 41, should be an *A*?

If we really know nothing about the function of this DNA, then we might proceed by imagining that the bases here have been chosen at random. If each site is treated as a trial of some random selection process, we have the outcomes of 40 trials before us. A quick tally shows that there are 8 *As*, 7 *Gs*, 11 *Cs*, and 14 *Ts*. Thus we estimate

$$\mathcal{P}(A) \approx \frac{8}{40} = .200, \quad \mathcal{P}(G) \approx \frac{7}{40} = .175, \quad \mathcal{P}(C) \approx \frac{11}{40} = .275, \quad \mathcal{P}(T) \approx \frac{14}{40} = .350.$$

We've used the frequency of the occurrence of the various bases to estimate the probabilities. Just as for the flip of a weighted coin, with a longer sequence of trials, we'd have more confidence in our estimates. Nonetheless, with the limited number of trials at our disposal, we've done the best we can. We thus estimate the probability of an *A* in site 41 as .2.

Often we'll need to group several outcomes into a set, which we call an *event*. For instance for the coin flip there are four possible events, corresponding to the four ways we can make sets of the outcomes:

$$\begin{aligned} E_{heads} &= \{\text{heads}\} & E_{either} &= \{\text{heads, tails}\} \\ E_{tails} &= \{\text{tails}\} & E_{neither} &= \{ \} \end{aligned}$$

We say an event occurs if any of the outcomes in the event is observed.

EXAMPLE. In our DNA example, viewing each site as a trial, the possible basic outcomes are the appearance of the 4 bases. Events which might be of interest are 'the base is a purine' and 'the base is a pyrimidine', or even 'the base is not *A*'. In more formal notation

$$E_{purine} = \{A, G\}, \quad E_{pyrimidine} = \{C, T\}, \quad E_{not\ A} = \{G, C, T\}.$$

When we know the probability of the basic outcomes, we can then assign probabilities to all events. For an event containing only a single outcome, the probability is simply the probability of that outcome. Thus for the fair coin

$$\mathcal{P}(E_{heads}) = \mathcal{P}(\text{heads}) = \frac{1}{2} \quad \text{and} \quad \mathcal{P}(E_{tails}) = \mathcal{P}(\text{tails}) = \frac{1}{2}.$$

Now the event  $E_{either}$  means 'either heads or tails' happens. Since this is a sure thing, its probability is 1 and so  $\mathcal{P}(E_{either}) = 1$ . Similarly, the event  $E_{neither}$  means we get neither a head nor a tail, and this is sure *not* to occur, so its probability is 0.

EXAMPLE. For the DNA sequence example, what should  $\mathcal{P}(E_{\text{purine}})$  be?

One way to estimate it is to go back to our data and simply tally the frequency with which purines occur. For instance, since in our 40 base sequence there were 8 As and 7 Gs, there were a total of 15 purines out of the 40 bases so we estimate  $\mathcal{P}(E_{\text{purine}}) \approx \frac{15}{40} = .375$ .

► Explain why  $E_{\text{pyrimidine}} = .625$  and  $E_{\text{not A}} = .800$ .

There is another way we could estimate  $\mathcal{P}(E_{\text{purine}})$ . Notice that

$$\begin{aligned}\mathcal{P}(E_{\text{purine}}) &= \mathcal{P}(A) + \mathcal{P}(G) \\ \frac{8+7}{40} &= \frac{8}{40} + \frac{7}{40}.\end{aligned}$$

The way fractions are added ensures that the probability of a purine appearing is the same as the sum of the probabilities of the bases  $A$  and  $G$  in the class of purines. In fact, we can generalize this example to the rule:

**Addition Rule (Special case):** *The probability of any event is the sum of the probabilities of the individual outcomes making up that event.*

Consider the toss of a fair die to make this clearer. Our basic one-outcome events are  $E_1, E_2, \dots, E_6$ , where  $E_i = \{\text{'the die shows an } i\} = \{i\}$ . The probabilities of getting any of the outcomes 1, 2, 3, 4, 5, or 6 are all  $\frac{1}{6}$ , since experience shows us that each outcome is equally likely and occurs in roughly 1 out of 6 trials. Since the event  $E = \{1, 2, 3, 4, 5, 6\}$  is a sure thing, its probability is 1. But now events such as ‘the die shows an odd number’ can be given probabilities by

$$E_{\text{odd}} = \{1, 3, 5\}$$

so

$$\mathcal{P}(E_{\text{odd}}) = \mathcal{P}(1) + \mathcal{P}(3) + \mathcal{P}(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

► Explain why for a toss of a fair die the probability of the event ‘the die shows an even number’ is  $\frac{1}{2}$ . What outcomes make up this event?

► What outcomes make up the event ‘the die shows a number  $\leq 2$ ’? What is the probability of this event for a fair die?

**Mutually exclusive events and sums of probabilities.** The rule we just used for assigning probabilities to events is actually an important special case of a more general rule that lets us use known probabilities of events to calculate probabilities of more complicated events.

Suppose we have two events,  $E$  and  $F$ , whose probabilities we know, and we are interested in knowing the probability that either  $E$  or  $F$  occurs. This new event, which is denoted by  $E \cup F$ , is the set of outcomes that appears in either  $E$  or  $F$ , or both. This new set is called the *union* of  $E$  and  $F$ . For example, the events ‘the die shows a number  $\leq 4$ ’ and ‘the die shows an even number’ have as their union the event ‘the die does not show a 5’ as we see by

$$E_{\leq 4} \cup E_{\text{even}} = \{1, 2, 3, 4\} \cup \{2, 4, 6\} = \{1, 2, 3, 4, 6\} = E_{\text{not } 5}.$$

We’d like to understand how we can combine probabilities of several events to get the probability of the union.

This is most easily done when the events to be combined are *mutually exclusive*. Informally two events are mutually exclusive if it is impossible for them to occur simultaneously; if one occurs, the other does not. If we've listed the outcomes in the events in sets, then we see they are mutually exclusive when the sets have no outcomes in common. That is, events are mutually exclusive when the sets are *disjoint*.

For instance, for a die toss consider the three events: 'the die shows an odd number', 'the die shows a number  $\leq 3$ ', and 'the die shows a number  $> 4$ '. Writing out the outcomes in each of these events as

$$E_{\text{odd}} = \{1, 3, 5\}, \quad E_{\leq 3} = \{1, 2, 3\}, \quad \text{and} \quad E_{> 4} = \{5, 6\},$$

we see the first two are not disjoint (both events will occur if the die shows a 1 or a 3) while the last two are disjoint (they cannot both occur at once).

For a coin toss, the events  $E_{\text{heads}}$  and  $E_{\text{tails}}$  are mutually exclusive since one precludes the other. However, the composite event  $E_{\text{either}}$  and the event  $E_{\text{heads}}$  are not mutually exclusive: knowing 'heads or tails' was produced does not tell us that 'heads' did not occur.

► Explain why in our DNA example the events  $E_{\text{purine}}$  and  $E_{\text{pyrimidine}}$  are mutually exclusive, while  $E_{\text{pyrimidine}}$  and  $E_{\text{not } A}$  are not.

Now suppose we consider any two events  $E$  and  $F$  which are mutually exclusive. Then their probabilities can be combined according to

**Addition Rule:** *If events  $E$  and  $F$  are mutually exclusive, then the probability of the event ' $E$  or  $F$ ', will be the sum of the probabilities of the two events:*

$$\mathcal{P}(E \cup F) = \mathcal{P}(E) + \mathcal{P}(F), \quad \text{if } E \text{ and } F \text{ are disjoint.}$$

EXAMPLE. Consider a die toss, and the events  $E_{\leq 2}$  = 'the die shows a number  $\leq 2$ ' and  $E_{\text{mult } 3}$  = 'the die shows a multiple of 3'.

- Explain why  $\mathcal{P}(E_{\leq 2}) = \frac{1}{3}$  by listing the outcomes that make up this event.
- Explain why  $\mathcal{P}(E_{\text{mult } 3}) = \frac{1}{3}$  by listing the outcomes that make up this event.
- Are these two events mutually exclusive?

Now the probability of the event  $E_{\leq 2} \cup E_{\text{mult } 3}$  = 'the die shows either a number  $\leq 2$  or a multiple of 3' can be calculated with ease. Since  $E_{\leq 2}$  and  $E_{\text{mult } 3}$  are disjoint,

$$\mathcal{P}(E_{\leq 2} \cup E_{\text{mult } 3}) = \mathcal{P}(E_{\leq 2}) + \mathcal{P}(E_{\text{mult } 3}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

Of course we could also have found this by listing all the outcomes in this event

$$E_{\leq 2} \cup E_{\text{mult } 3} = \{1, 2, 3, 6\},$$

and so

$$\mathcal{P}(E_{\leq 2} \cup E_{\text{mult } 3}) = \mathcal{P}(E_1) + \mathcal{P}(E_2) + \mathcal{P}(E_3) + \mathcal{P}(E_6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{2}{3}.$$

EXAMPLE. Note that the events  $E_{\text{mult } 3}$  and  $E_{< 4}$  are *not* mutually exclusive; it's possible for both to occur simultaneously if the outcome of the toss is a 3. Thus we expect

$$\mathcal{P}(E_{\text{mult } 3} \cup E_{< 4}) \neq \mathcal{P}(E_{\text{mult } 3}) + \mathcal{P}(E_{< 4}).$$

In fact, since

$$E_{mult\ 3} \cup E_{<4} = \{1, 2, 3, 6\} = E_1 \cup E_2 \cup E_3 \cup E_6,$$

we find

$$\mathcal{P}(E_{mult\ 3} \cup E_{<4}) = \frac{2}{3} \neq \frac{1}{3} + \frac{1}{2} = \mathcal{P}(E_{mult\ 3}) + \mathcal{P}(E_{<4}).$$

There is a more general version of the addition rule which can be used on events such as these that are not mutually exclusive. You'll find it in the exercises.

As a final consequence of the addition rule of probabilities of disjoint events, we can understand the probability of an event *not* happening. If  $E$  is any event, let  $E'$  be the *complementary* event composed of all those outcomes not in  $E$ . For example with a die toss

$$(E_{<4})' = E_{>4}.$$

For any event  $E$ , note that  $E$  and  $E'$  are certainly exclusive (they can't both happen at once). Then by the addition rule

$$\mathcal{P}(E \cup E') = \mathcal{P}(E) + \mathcal{P}(E').$$

However, the event  $E \cup E'$  is the event that anything at all happens, and since this is a sure thing,  $\mathcal{P}(E \cup E') = 1$ . Thus  $\mathcal{P}(E) + \mathcal{P}(E') = 1$ , or

$$\mathcal{P}(E') = 1 - \mathcal{P}(E).$$

We now have a rule for calculating probabilities of complementary events.

EXAMPLE. As an application to DNA, the event  $E_{pyrimidine}$  is the same as  $E'_{purine}$ . Thus  $\mathcal{P}(E_{pyrimidine}) = 1 - \mathcal{P}(E_{purine})$ . Of course this is consistent with the example above where  $\mathcal{P}(E_{purine}) = .375$  and  $\mathcal{P}(E_{pyrimidine}) = .625$ .

**Independent events and products of probabilities.** There is another important way we can combine events to get more complicated ones. If  $E$  and  $F$  are events, then  $E \cap F$  denotes the event that both  $E$  and  $F$  occur. The set of outcomes  $E \cap F$  is simply all outcomes appearing in both  $E$  and  $F$ . This is called the *intersection* of the sets. For instance,

$$E_{\leq 4} \cap E_{mult\ 2} = \{1, 2, 3, 4\} \cap \{2, 4, 6\} = \{2, 4\}.$$

Imagine flipping a coin and tossing a die together. Then there are 12 possible outcomes: (heads, 1), (tails, 1), (heads, 2), (tails, 2), ..., (tails, 6). Assuming both the coin and die are fair, each of these outcomes should be equally likely. Since their probabilities must add to 1 (because they are disjoint, and it is certain that one of them occurs), each must have probability  $\frac{1}{12}$ .

► Explain why there are twelve possible outcomes.

Consider the event 'the die shows a 5' and the event 'the coin shows heads':

$$E_5 = \{(\text{heads}, 5), (\text{tails}, 5)\},$$

$$E_{heads} = \{(\text{heads}, 1), (\text{heads}, 2), (\text{heads}, 3), (\text{heads}, 4), (\text{heads}, 5), (\text{heads}, 6)\}.$$

The intersection of these two events is 'the die shows a 5 and the coin shows heads',

$$E_5 \cap E_{heads} = \{(\text{heads}, 5)\} = E_{heads,5}.$$

How are the probabilities of these three events related?

► Explain why  $\mathcal{P}(E_{heads}) = \frac{1}{2}$  and  $\mathcal{P}(E_5) = \frac{1}{6}$  by thinking of each of them as a union of disjoint events and using the addition rule.

Since  $\mathcal{P}(E_{heads,5}) = \frac{1}{12}$ , notice that

$$\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$

so

$$\mathcal{P}(E_{heads}) \cdot \mathcal{P}(E_5) = \mathcal{P}(E_{heads} \cap E_5).$$

At least in this example, the probability of an intersection of two events was simply the product of the probabilities of the two events. The reason that these probabilities behaved this way actually depended on a special feature of the events: the events  $E_5$  and  $E_{heads}$  are independent.

Informally, we say two events are *independent* if knowledge that one of the events has occurred tells us absolutely nothing about whether the other has occurred. In other words, if we were told whether or not the first event occurred, that would have no effect on our belief about the likelihood of the second having occurred.

In this example, knowing whether the die shows a 5 or not tells us nothing about the chance of seeing either of the coin outcomes, a head or a tail.

**Multiplication Rule:** *If events  $E$  and  $F$  are independent, then the probability of the event ‘ $E$  and  $F$ ’ will be the product of the probabilities of the two events:*

$$\mathcal{P}(E \cap F) = \mathcal{P}(E) \cdot \mathcal{P}(F), \text{ if } E \text{ and } F \text{ are independent.}$$

EXAMPLE. Suppose we toss two fair dice in order. There are 36 equally likely outcomes such as  $(1, 1)$ ,  $(1, 2)$ , etc., each with a probability of  $\frac{1}{36}$ . (Because we toss the dice and record what they show in order, the outcome  $(1, 2)$  is not the same as the outcome  $(2, 1)$ .)

Consider the events

$$E_{d2=3} = \text{‘the second die shows a 3’},$$

$$E_{d1=even} = \text{‘the first die is even’}.$$

► Explain why  $\mathcal{P}(E_{d2=3}) = \frac{6}{36} = \frac{1}{6}$  by listing the 6 outcomes that make up the event.

► Explain why  $\mathcal{P}(E_{d1=even}) = \frac{18}{36} = \frac{1}{2}$  by listing the 18 outcomes that make up the event.

Now intuitively, the events  $E_{d1=even}$  and  $E_{d2=3}$  are independent, since one tells us something about die 1 and the other about die 2. Knowledge about one die should communicate nothing about the other. Thus the multiplication rule tells us

$$\mathcal{P}(E_{d1=even} \cap E_{d2=3}) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}.$$

We can confirm this by reasoning a different way. The event  $E_{d1=even} \cap E_{d2=3}$  is the event that the first die is even and the second shows a 3. This means it is composed of the outcomes  $(2, 3)$ ,  $(4, 3)$ , and  $(6, 3)$ . Since each of these outcomes has probability  $\frac{1}{36}$ , we have

$$\mathcal{P}(E_{d1=even} \cap E_{d2=3}) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{12}.$$

EXAMPLE. Continuing with the toss of two dice in order, consider another event

$$E_{sum=9} = \text{'the sum of the results is 9'}$$

► Explain why  $\mathcal{P}(E_{sum=9}) = \frac{4}{36} = \frac{1}{9}$  by listing the 4 outcomes that make up the event.

Now the events  $E_{sum=9}$  and  $E_{d2=3}$  are *not* independent. If we know the sum is a 9, then we know the outcome must have been one of (6, 3), (5, 4), (4, 5), or (3, 6). Since these are all equally likely, we see that knowledge that  $E_{sum=9}$  occurred lets us say there is a 1 in 4 chance that  $E_{d2=3}$  occurred. This is different than the 1 in 6 chance we would have without the knowledge that  $E_{sum=9}$  occurred. Thus knowledge of one event gave us some information about the other, so they are dependent.

To verify that the multiplication rule doesn't hold for this example, we check

$$\mathcal{P}(E_{sum=9} \cap E_{d2=3}) = \mathcal{P}((6, 3)) = \frac{1}{36}$$

while

$$\mathcal{P}(E_{sum=9}) \cdot \mathcal{P}(E_{d2=3}) = \frac{1}{9} \cdot \frac{1}{6} = \frac{1}{54}.$$

While the definition of independent events given here has been an informal one, in the next section will be a bit more precise. Still, this informal way of thinking is often necessary, especially when probability is being used to model complicated processes.

The multiplication and addition rules are very useful in determining the probabilities of events. They allow us to calculate probabilities of complicated events by seeing how they are built from events we already understand by using the words 'or', 'and', and 'not'. An 'or' means we add the probabilities, provided the events being combined are disjoint. An 'and' means we multiply the probabilities, provided the events being combined are independent. A 'not' means we compute the probability of the complementary event, and subtract it from 1.

The key properties of probabilities we've discussed so far can be summarized as:

- The probability of any event  $E$  is a number  $\mathcal{P} = \mathcal{P}(E)$  with  $0 \leq \mathcal{P} \leq 1$ .
- If several events  $E_1, E_2, \dots, E_n$  are mutually exclusive then the probability that any of them occur, *i.e.*, the probability of  $E = E_1 \cup E_2 \cup \dots \cup E_n$ , is  $\mathcal{P}(E) = \mathcal{P}(E_1) + \mathcal{P}(E_2) + \dots + \mathcal{P}(E_n)$ , the sum of the individual probabilities.
- If several events are  $E_1, E_2, \dots, E_n$  independent, then the probability that they all occur, *i.e.*, the probability of  $E = E_1 \cap E_2 \cap \dots \cap E_n$ , is  $\mathcal{P}(E) = \mathcal{P}(E_1) \cdot \mathcal{P}(E_2) \cdot \dots \cdot \mathcal{P}(E_n)$ , the product of the individual probabilities.
- If the probability of an event  $E$  occurring is  $\mathcal{P}$ , then the probability that  $E$  does not occur, *i.e.* the probability of the complementary event  $E'$ , is  $1 - \mathcal{P}$ .

Now let's apply these rules to a very simple model of DNA mutation. Suppose we focus on a particular site in a gene sequence, and on whether at that site a purine or a pyrimidine appears. We only care about these classes, not on the precise bases.

Suppose we also know that with each generation there is a 1.5% chance the base at this site undergoes a transversion, which we'll call simply a 'change'. Thus

there is a 98.5% chance that there is no change (or a transition, which is treated as no change in this model). Then for one generation

$$\mathcal{P}(E_{change}) = .015, \mathcal{P}(E_{no\ change}) = .985.$$

While this probability of a change is much higher than is typically observed, we are not yet concerned with realism.

Now imagine what happens over two generations. There are four possibilities of interest:

$$\left\{ \begin{array}{l} \text{change} \\ \text{no change} \end{array} \right\}, \text{ followed by } \left\{ \begin{array}{l} \text{change} \\ \text{no change} \end{array} \right\}.$$

What are their probabilities?

First, we make the important assumption that what happens in passing to the first generation is independent of what happens in passing to the second. This is reasonable if we think mutations are caused by errors and accidents, since the DNA should have no memory of what had happened before. With this assumption, we can use the multiplication rule for combining probabilities of independent events to get

$$\begin{aligned} \mathcal{P}(E_{change,change}) &= (.015)(.015) = .000225 \\ \mathcal{P}(E_{change,no\ change}) &= (.015)(.985) = .014775 \\ \mathcal{P}(E_{no\ change,change}) &= (.985)(.015) = .014775 \\ \mathcal{P}(E_{no\ change,no\ change}) &= (.985)(.985) = .970225 \end{aligned}$$

► What is the sum of these four probabilities? Why did it have to be that?

What is the probability of seeing no change from the original base in generation 0 to the descendent in generation 2? This event is actually composed of two events: either there was no change in each generation, or there was a change in each generation producing no net change (*i.e.*, the changes are hidden). Since these two events are mutually exclusive, we find the desired probability is

$$\mathcal{P}(E_{no\ change,no\ change}) + \mathcal{P}(E_{change,change}) = .970225 + .000225 = .97045.$$

Thus the probability of observing no change when comparing a base across two generations is slightly greater than the chance of no change having actually occurred. Mutations followed by other mutations may result in no net observable change, yet they do affect the likelihood of what we observe.

Notice that to deduce this result we used both the multiplication rule for probabilities of independent events, and the addition rule for probabilities of disjoint events. This sort of analysis will form the basis of all of our modeling of molecular evolution. We'll just need to deal with very large numbers of generations, and with all four of the bases.

### Problems:

1. Use a coin to conduct an experiment to determine the probability of it producing heads or tails when flipped.

- a. Flip the coin ten times, recording your results. Use your data to estimate the probability of heads.
  - b. Flip the coin ten more times, for a total of 20 flips. Use your data to estimate the probability of heads.
  - c. Flip the coin 20 more times, for a total of 40 flips. Use your data to estimate the probability of heads.
  - d. If you believe your coin is fair, then you believe  $\mathcal{P}(\text{heads}) = .5$ . Do your experiments support this? If your experiments did not exactly produce .5 should you be doubtful that the coin is fair? Which experiment produced the result closest to .5? Is that what you would have expected?
2. Suppose a fair coin is flipped 10 times ( $H$  =heads,  $T$  =tails).
    - a.  $HTTHTHHHTH$  is produced in 10 independent trials. What is the probability of this particular sequence of outcomes?
    - b.  $TTTTTTTTTT$  is produced in 10 independent trials. What is the probability of this particular sequence of outcomes?
    - c. Your answers to (a) and (b) should be the same. Why might this be surprising to some people? Are you convinced they are equally likely?
  3. Consider the 20 base sequence

*AGGGATACATGACCCATACA.*

- a. Use the first five bases to estimate the four probabilities  $p_A$ ,  $p_G$ ,  $p_C$ , and  $p_T$ .
  - b. Repeat (a) using the first 10 bases.
  - c. Repeat (a) using all the bases
  - d. Is there a pattern to the way the probabilities you computed in (a-c) changed? If so, what features of the original sequence does this pattern reflect?
4. Consider the 20-base sequence

*CGGTTGCCTGCGTAGTGCG*

- a. Give the best estimates you can for the probability that each base would appear at site 21.
  - b. Give the best estimates you can for the probabilities of a purine and of a pyrimidine at the site 21.
  - c. Which base is most likely to appear at site 21? Is it a purine or a pyrimidine? Does this make sense in light of your answer to (b)? Explain.
5. A simple model for human offspring is that each child is equally likely to be male or female. With this model, a three-child family can be thought of as three random determinations of sex, in order.
    - a. What are the 8 possible outcomes? What is the probability of each?
    - b. What outcomes make up the event ‘the oldest child is a daughter’? What is the event’s probability?
    - c. What outcomes make up the event ‘the family has one daughter and two sons’? What is its probability?
    - d. What is the complement of the event in (c)? List the outcomes in it and describe it in words. What is its probability?
    - e. What outcomes make up the event ‘the family has at least one daughter’? What is its probability?
  6. For a coin toss, there are 2 possible outcomes but 4 events listed in the text. More generally, if a trial has  $n$  possible outcomes, there will be  $2^n$  events.

- a. If one of the bases  $A$ ,  $G$ ,  $C$ , and  $T$  is chosen at random, so there are 4 possible outcomes, then there are  $16 = 2^4$  different events. List them all.
  - b. Explain why if there are  $n$  possible outcomes then there are  $2^n$  possible events.
7. Many genetic traits can be modeled using probability. Imagine picking a person at random from the world population. Then we can consider events such as ‘the person has brown eyes’ or ‘the person is male.’ For each of the following pairs of events, decide whether the two events are mutually exclusive, and if it is reasonable to think of them as independent.
    - a. ‘the person is male’ and ‘the person has brown eyes’
    - b. ‘the person has black hair’ and ‘the person is an albino’
    - c. ‘the person has blue eyes’ and ‘the person has blond hair’
  8. If two events are mutually exclusive, can they also be independent? Explain.
  9. The definition of ‘mutually exclusive’ events given in the text was in words. Explain why it could be expressed more concisely as

$E$  and  $F$  are mutually exclusive means  $E \cap F = \{ \}$ .

10. There is a more general version of the addition rule for probabilities that does not require that events be mutually exclusive: For any events  $E$  and  $F$ ,

$$\mathcal{P}(E \cup F) = \mathcal{P}(E) + \mathcal{P}(F) - \mathcal{P}(E \cap F).$$

- a. Explain why if  $E$  and  $F$  are disjoint then this agrees with the addition rule in the text.
  - b. Show the general version holds in an example for a die toss using the events  $E_{mult\ 3}$  and  $E_{<4}$ .
11. Explain informally why if events  $E$  and  $F$  are independent then the complementary events  $E'$  and  $F'$  must also be independent.
  12. The text presents a model of DNA sequence mutation considering only the classes of purines and pyrimidines, and computes the probability of observing ‘no change’ at a site when comparing an ancestral sequence and a sequence two generations later. Continue that discussion by answering:
    - a. What is the probability of observing a ‘change’ when comparing an ancestral sequence and a sequence two generations later?
    - b. What 4 outcomes (ordered triples of ‘change’/‘no change’) make up the event ‘no change is observed at a site when comparing an ancestral sequence and a sequence three generations later’?
    - c. What is the probability of the event in (b)?

### 3. Conditional probabilities

When base substitutions occur, the probability of a particular base appearing at a site in the descendent sequence might depend on the ancestral base. For example, if the ancestral base is a  $T$ , we'd expect the probability of a  $T$  in the descendent to be high. If the ancestral base is a  $C$ , we'd expect a lower probability of the descendent having a  $T$ , since a transition is less likely than no change. If the ancestral base is an  $A$  or  $G$ , we might expect an even lower probability that the descendent has a  $T$ , since transversions might be rarer than transitions.

To formalize this, we need the concept of *conditional probability*. This is the probability of one event given that we know another event has occurred. Letting  $S_0$  refer to the ancestor and  $S_1$  the descendent, we'll use notation like ' $S_0 = C$ ' to mean that the ancestral site has base  $C$ , and ' $S_1 = T$ ' to mean the descendent site has base  $T$ . Then

$$\mathcal{P}(S_1 = T \mid S_0 = C) = .02$$

will mean that there is a 2% chance that the descendent base is a  $T$  given that the ancestral base is a  $C$ . Notice that the vertical bar ' $\mid$ ' in this conditional probability notation is read as 'given that'. We now have a good way to refer to the fact the probability of a 'final' base appearing depends on the 'initial' base that appeared.

► Taking into account the previous comments on the likelihood of transitions and transversions, which of  $\mathcal{P}(S_1 = A \mid S_0 = C)$ ,  $\mathcal{P}(S_1 = G \mid S_0 = C)$ ,  $\mathcal{P}(S_1 = C \mid S_0 = C)$ , and  $\mathcal{P}(S_1 = T \mid S_0 = C)$  are likely to be smallest? which is likely to be biggest?

The properties of probabilities discussed earlier carry over to the setting of conditional probabilities, as long as we keep in mind we are always assuming something particular happened — the given condition. For instance,

$$\begin{aligned} \mathcal{P}(S_1 = A \mid S_0 = C) + \mathcal{P}(S_1 = G \mid S_0 = C) + \\ \mathcal{P}(S_1 = C \mid S_0 = C) + \mathcal{P}(S_1 = T \mid S_0 = C) = 1. \end{aligned}$$

After all, given that  $S_0 = C$ , the four events  $S_1 = A, G, C$ , and  $T$  are mutually exclusive, yet certainly one of them will occur, and so the probabilities must add to 1.

EXAMPLE. The conditional probability  $\mathcal{P}(S_1 = T \mid S_0 = C)$  is not the same as the probability  $\mathcal{P}(S_1 = T \text{ and } S_0 = C)$ . To see this clearly, suppose we have aligned sequences

$S_0 : AGCTTCCGATCCGCTATAATCGTTAGTTGTTACACCTCTG$   
 $S_1 : AGCTTCTGATACGCTATAATCGTGAGTTGTTACATCTCCG$

Then of the 40 sites shown (which we think of as 40 trials) we find 2 sites with a  $T$  in  $S_1$  and a  $C$  in  $S_0$ . Thus we'd estimate

$$\mathcal{P}(S_1 = T \text{ and } S_0 = C) \approx \frac{2}{40} = .05.$$

However, of the 11 sites that have a  $C$  in  $S_0$ , we find only 2 of these have a  $T$  in  $S_1$ , so we estimate

$$\mathcal{P}(S_1 = T \mid S_0 = C) \approx \frac{2}{11} \approx .182.$$

Pay particular attention to this last calculation. We divided not by the total number of trials, but only by the number of trials that satisfied the given criterion  $S_0 = C$ . The trials in which  $S_0 \neq C$  are irrelevant to the calculation of this conditional probability.

There is another way to find conditional probabilities, which is convenient if we've already computed some other probabilities. From this last example, we know that the probability that both  $S_0 = C$  and  $S_1 = T$  is

$$\mathcal{P}(S_1 = T \text{ and } S_0 = C) \approx \frac{2}{40} = .05.$$

The probability that  $S_0 = C$  can be found to be

$$\mathcal{P}(S_0 = C) \approx \frac{11}{40} = .275.$$

Then

$$\frac{\mathcal{P}(S_1 = T \text{ and } S_0 = C)}{\mathcal{P}(S_0 = C)} \approx \frac{\frac{2}{40}}{\frac{11}{40}} = \frac{2}{11} \approx \mathcal{P}(S_1 = T \mid S_0 = C).$$

The denominators of 40 canceled one another out, leaving us with the ratio we found above.

More formally, we can capture what has happened in this approach by the following general definition.

**Definition of Conditional Probability:** If  $E$  and  $F$  are two events, then the conditional probability of  $F$  given  $E$  is defined by

$$(1) \quad \mathcal{P}(F \mid E) = \frac{\mathcal{P}(F \cap E)}{\mathcal{P}(E)}.$$

The concept of conditional probability also clarifies the notion of independence of events. Earlier, we informally said that events  $E$  and  $F$  were independent if knowledge that one had occurred gave us no information as to whether the other occurred. This could be expressed as

$$(2) \quad \mathcal{P}(F \mid E) = \mathcal{P}(F) \text{ and } \mathcal{P}(E \mid F) = \mathcal{P}(E).$$

Using the definition of conditional probability, the first of these becomes

$$\frac{\mathcal{P}(F \cap E)}{\mathcal{P}(E)} = \mathcal{P}(F),$$

or

$$\mathcal{P}(F \cap E) = \mathcal{P}(E)\mathcal{P}(F).$$

► Explain why the second equation in (2) gives the same result.

This leads us to the formal mathematical definition of independence as

**Definition of Independence:** Events  $E$  and  $F$  are said to be independent if

$$\mathcal{P}(E \cap F) = \mathcal{P}(E)\mathcal{P}(F).$$

Of course, this is essentially the same as the multiplication rule for independent events stated earlier. All the new definition really says is that the word ‘independent’ is simply a concise way of saying the multiplication rule applies. In practice, to recognize whether events are independent or not, it’s usually better to stick with the more informal definition given in the last section, which has been formalized in equations (2).

EXAMPLE. Suppose a 40-base ancestral DNA sequence is

$S_0 : ACTTGTCGGATGATCAGCGGTCCATGCACCTGACAACGGT$

while its descendent aligned sequence is

$S_1 : ACATGTTGCTTGACGACAGGTCCATGCGCCTGAGAACGGC.$

Thinking of each site as a trial of the same probabilistic process, we can estimate 16 conditional probabilities describing the likelihood of observing different types of base substitutions when comparing the sequences of ancestor to descendent:  $\mathcal{P}(S_1 = j \mid S_0 = i)$ , where  $i, j = A, G, C, T$ .

To do this we begin by tallying the number of sites with an occurrence of each pair  $S_0 = i, S_1 = j$  in the aligned sequences, recording the information in a *frequency array* such as Table 1.

$S_1 \setminus S_0$	A	G	C	T
A	7	0	1	1
G	1	9	2	0
C	0	2	7	2
T	1	0	1	6

TABLE 1. Frequencies of  $S_1 = i$  and  $S_0 = j$  in 40-site sequence comparison

► What is the sum of the 16 numbers in the table? Why?

If we add the numbers in a *column* of this table, we obtain the total number of sites with a particular base in  $S_0$ . For instance, the number of sites with  $S_0 = A$  is  $7 + 1 + 0 + 1 = 9$ . In general, the number of sites with  $S_0 = j$  is the sum of the entries in column  $j$ .

► What is the meaning of a row sum in the table?

Now for any bases  $i, j$  we estimate the conditional probabilities  $\mathcal{P}(S_1 = i \mid S_0 = j)$  by dividing the number of sites with  $S_1 = i$  and  $S_0 = j$  by the number of sites with  $S_0 = j$ . That means we must divide the entry in row  $i$ , column  $j$  of the table by the sum of the entries in column  $j$ . We find all the conditional probabilities by dividing all table entries by their corresponding column sums. Rounding the results to 3 digits yields Table 2.

► What is the sum of the entries in any column of this new table? Why?

► If instead of dividing by column sums, you divided by row sums, would you get the same results? What conditional probabilities would you be calculating?

$S_1 \setminus S_0$	$A$	$G$	$C$	$T$
$A$	.778	0	.091	.111
$G$	.111	.818	.182	0
$C$	0	.182	.636	.222
$T$	.111	0	.091	.667

TABLE 2. Estimates of conditional probabilities  $\mathcal{P}(S_1 = i \mid S_0 = j)$ **Problems:**

- Assuming births of each sex are equally likely, a two-child family may have 4 outcomes in the sexes of the children.
  - List the outcomes and give the probability of each.
  - What is the probability that at least one child is a female?
  - What is the probability that the youngest child is a female?
  - What is the conditional probability that the youngest child is a female given that at least one child is a female?
  - What is the conditional probability that at least one child is a female given that the youngest child is a female?
  - Are the events in (b) and (c) independent? Explain.
- Consider the toss of a single die.
  - Show the events  $E_{odd}$  and  $E_{\leq 2}$  are independent by using the formal definition.
  - Show the events  $E_{odd}$  and  $E_{\leq 3}$  are not independent by using the formal definition.
  - Explain as intuitively as possible why the events of (a) were independent, but those of (b) were not.
- Medical tests, such as those for diseases, are sometime characterized by their *sensitivity* and *specificity*. The sensitivity of a test is the probability that a diseased person will show a positive test result (a correct positive). The specificity of a test is the probability that a healthy person will show a negative test result (a correct negative).
  - Both sensitivity and specificity are conditional probabilities. Which of the following are they:
$$\mathcal{P}(- \text{ result} \mid \text{disease}), \quad \mathcal{P}(- \text{ result} \mid \text{no disease}),$$

$$\mathcal{P}(+ \text{ result} \mid \text{disease}), \quad \mathcal{P}(+ \text{ result} \mid \text{no disease})$$
  - The other conditional probabilities listed in (a) can be interpreted as probabilities of false positives and false negatives. Which is which?
  - A study [?] investigated the use of X-ray readings to diagnose tuberculosis. Diagnosis of 1820 individuals produced the data in Table 3. Compute both the sensitivity and specificity for this method of diagnosis.
- Ideally, the specificity and sensitivity of medical tests should be high (close to 1). However, even with a highly specific and sensitive test, screening a large population for a disease which is rare can produce surprising results.

	Persons without TB	Persons with TB
Negative X-ray	1739	8
Positive X-ray	51	22

TABLE 3. Data from TB diagnosis study

- a. Suppose the sensitivity and specificity of a test for disease are both .99. If the test is applied to everyone in a population of 100,000 individuals, only 100 of which have the disease, then compute how many individuals with/without the disease you would expect to test positive/negative. Organize your results in a table like that in the preceding problem.
- b. Use the table you produced in (a) to compute the conditional probability that a person who tests positive actually has the disease.
5. In the text, the data in Table 1 is used to compute the conditional probabilities  $\mathcal{P}(S_1 = i \mid S_0 = j)$ .
  - a. Use the same data to compute  $\mathcal{P}(S_0 = j \mid S_1 = i)$ . Do you get the same results as in Table 2?
  - b. Explain intuitively why you would usually not expect  $\mathcal{P}(S_1 = i \mid S_0 = j)$  and  $\mathcal{P}(S_0 = i \mid S_1 = j)$  to be the same.
6. In tables, such as Table 2, of conditional probabilities describing realistic DNA base substitutions between an ancestor and descendent, there is often a pattern to the sizes of the numbers.
  - a. Which entries refer to no substitution occurring? Why are these likely to be the largest entries?
  - b. Which entries refer to transitions?, to transversions? Does Table 2 support the claim that transitions tend to be more common than transversions?
7. Using the data in Table 1:
  - a. Compute each column sum and divide it by 40. These results can be interpreted as estimates of probabilities. What probabilities are being estimated?
  - b. Compute the row sums, and divide each by 40. What probabilities are being estimated?
8. For the two sequences  $S_0$  and  $S_1$  which are used in producing Table 1:
  - a. Estimate the eight probabilities  $\mathcal{P}(S_0 = i)$  and  $\mathcal{P}(S_1 = j)$  for  $i, j = A, G, C, T$ .
  - b. For each pair  $i, j$ , are the events  $S_0 = i$  and  $S_1 = j$  independent?
  - c. Why does the fact that one sequence is descended from another help explain your answer to (b)?
9. Two DNA sequences of the same length are chosen and labeled  $S_0$  and  $S_1$ , but there is no ancestral relationship between the two.
  - a. Why would you expect that for each pair  $i, j$  the events  $S_0 = i$  and  $S_1 = j$  would be independent?
  - b. If the events  $S_0 = i$  and  $S_1 = j$  are independent, what would the pattern be in the entries in a table like Table 2?
10. Recall from the last section the 2-class model of purine and pyrimidine sequence mutation. Modify the model so that at each generation the probabilities of mutation depend on the current class of the site according to Table 4:
  - a. Explain intuitively why the formula

$S_{t+1} \setminus S_t$	<i>pur</i>	<i>pyr</i>
<i>pur</i>	.98	.01
<i>pyr</i>	.02	.99

TABLE 4. Conditional probabilities  $\mathcal{P}(S_{t+1} = i \mid S_t = j)$ 

$$\begin{aligned} \mathcal{P}(S_2 = \textit{pur} \mid S_0 = \textit{pur}) &= \\ & p(S_2 = \textit{pur} \mid S_1 = \textit{pur}) \cdot p(S_1 = \textit{pur} \mid S_0 = \textit{pur}) + \\ & p(S_2 = \textit{pur} \mid S_1 = \textit{pyr}) \cdot p(S_1 = \textit{pyr} \mid S_0 = \textit{pur}). \end{aligned}$$

is reasonable. Write similar formulas for  $\mathcal{P}(S_2 = \textit{pyr} \mid S_0 = \textit{pur})$ ,  $\mathcal{P}(S_2 = \textit{pur} \mid S_0 = \textit{pyr})$ , and  $\mathcal{P}(S_2 = \textit{pyr} \mid S_0 = \textit{pyr})$ .

- Using these formulas, compute numerical values for  $p(S_2 = j \mid S_0 = i)$  for the 4 possible choices with  $i, j = \textit{pur}, \textit{pyr}$ .
- Using the definition of conditional probability show that the formula in part (a) is valid. You will have to use the assumptions

$$\begin{aligned} \mathcal{P}(S_2 = \textit{pur} \mid S_1 = \textit{pur} \text{ and } S_0 = \textit{pur}) &= \mathcal{P}(S_2 = \textit{pur} \mid S_1 = \textit{pur}), \\ \mathcal{P}(S_2 = \textit{pur} \mid S_1 = \textit{pyr} \text{ and } S_0 = \textit{pur}) &= \mathcal{P}(S_2 = \textit{pur} \mid S_1 = \textit{pyr}). \end{aligned}$$

These assumptions state that probabilities of substitutions between time 1 and time 2 are independent of the base at time 0.

- Suppose  $E_1$  and  $E_2$  are two events, with  $E_2'$  being the event complementary to  $E_2$ . Recall that  $\mathcal{P}(E_2) + \mathcal{P}(E_2') = 1$ .
  - Explain using your intuitive understanding of conditional probabilities why  $\mathcal{P}(E_2 \mid E_1) + \mathcal{P}(E_2' \mid E_1) = 1$  should also hold.
  - Show the formula in part (a) holds more formally by using the definition of conditional probability as a quotient of probabilities. You'll need use that  $(E_2 \cap E_1) \cup (E_2' \cap E_1) = E_1$ .
- MATLAB can be used to compare two sequences and produce a frequency array such as Table 1. While the program `compseq` automates this, the individual steps are useful to know.
  - Try the following command sequence and explain what each line does.

```

S0='AACTGCAGT'
S1='AGCCGCAGA'
S0=='A'
S1=='G'
(S0=='A') & (S1=='G')
sum( (S0=='A') & (S1=='G') )

```
  - What one-line command would find the number of sites with a 'C' in  $S_0$  and a 'G' in  $S_1$ ?
  - What one-line command would count the number of purines in  $S_0$ ?
  - What one-line command would give the number of sites with a purine in  $S_0$  and a pyrimidine in  $S_1$ ?
- Suppose two sequences  $S_0$  and  $S_1$  have been compared, and a frequency table such as that in Table 1 has been produced, and entered into MATLAB as a matrix  $F$ .

- a. Explain why the sequence of commands

$$\text{colsum} = [1, 1, 1, 1] * F, \quad N = \text{colsum} * [1; 1; 1; 1], \quad p0 = \text{colsum}/N$$

will produce the fraction of sites with each base in  $S_0$ .

- b. Give a sequence of commands to produce the fraction of sites with each base in  $S_1$ .

- c. Try the MATLAB command  $D = \text{diag}(\text{colsum})$  to see what it does. Then explain why if  $M$  denotes the matrix of conditional probabilities such as in Table 2, that  $F = M * D$ . Thus  $M$  is easily computed by the command

$$M = F * \text{inv}(\text{diag}(\text{colsum})).$$

#### 4. Matrix models of base substitution

We now can create a basic model of molecular evolution by making use of probability and matrix algebra.

We begin by modeling the ancestral sequence probabilistically. Each site in the sequence is one of the four bases  $A$ ,  $G$ ,  $C$ , or  $T$ , chosen randomly according to some probabilities  $\mathcal{P}_A$ ,  $\mathcal{P}_G$ ,  $\mathcal{P}_C$ , and  $\mathcal{P}_T$ . These four probabilities must satisfy

$$\mathcal{P}_A + \mathcal{P}_G + \mathcal{P}_C + \mathcal{P}_T = 1,$$

since one of the bases is certain to appear. For convenience, we'll always use the order  $A, G, C, T$  for the bases (so the purines come first, and then the pyrimidines) and put these 4 probabilities into a vector as

$$\mathbf{p}_0 = (\mathcal{P}_A, \mathcal{P}_G, \mathcal{P}_C, \mathcal{P}_T).$$

This vector describes the ancestral base distribution, with its entries giving the fraction of sites we would expect to be occupied by each of the four bases.

► To what extent is the assumption that all bases in the sequence are chosen ‘at random’ reasonable? Would it matter whether the DNA sequence was coding or non-coding?

We model the mutation process over one time step assuming that only base substitutions can occur — no deletions, insertions, or inversions are considered. We specify the 16 conditional probabilities of observing a base substitution,  $\mathcal{P}(S_1 = i \mid S_0 = j)$ , for  $i, j = A, G, C$ , and  $T$ . It will be convenient to put these numbers into a  $4 \times 4$  matrix, using the ordering  $A, G, C$ , and  $T$ . In each column of the matrix are entries referring to the same ancestral base, and in each row are entries referring to the same descendent base. Using abbreviated notation such as  $\mathcal{P}_{i|j} = \mathcal{P}(S_1 = i \mid S_0 = j)$ , we let

$$M = \begin{pmatrix} \mathcal{P}_{A|A} & \mathcal{P}_{A|G} & \mathcal{P}_{A|C} & \mathcal{P}_{A|T} \\ \mathcal{P}_{G|A} & \mathcal{P}_{G|G} & \mathcal{P}_{G|C} & \mathcal{P}_{G|T} \\ \mathcal{P}_{C|A} & \mathcal{P}_{C|G} & \mathcal{P}_{C|C} & \mathcal{P}_{C|T} \\ \mathcal{P}_{T|A} & \mathcal{P}_{T|G} & \mathcal{P}_{T|C} & \mathcal{P}_{T|T} \end{pmatrix}.$$

► Why must the sum of the entries in any column of this matrix add to 1?

► How reasonable is it to assume only base substitutions occur? Why would you imagine that these might be the most common mutations, especially in coding regions of DNA?

EXAMPLE. If we have two specific DNA sequences, such as those at the end of the last section, one the ancestor and the other the descendent after one time step, then all these probabilities can be estimated from the data. The data in the frequency array in Table 1 lead to

$$(3) \quad \mathbf{p}_0 \approx (.225, .275, .275, .225) \text{ and } M \approx \begin{pmatrix} .778 & 0 & .091 & .111 \\ .111 & .818 & .182 & 0 \\ 0 & .182 & .636 & .222 \\ .111 & 0 & .091 & .667 \end{pmatrix}.$$

In fact, this estimate of  $M$  is just Table 2 treated as a matrix, while the estimate of  $\mathbf{p}_0$  is just the column sums of Table 1 divided by the number of sites in the sequences.

- Explain why the calculation of  $\mathbf{p}_0$  described here is the correct one to perform.

Expressing our model using a vector and matrix is more than just a concise notation; let's see what happens when we multiply them as

$$(4) \quad M\mathbf{p}_0 = \begin{pmatrix} \mathcal{P}_{A|A} & \mathcal{P}_{A|G} & \mathcal{P}_{A|C} & \mathcal{P}_{A|T} \\ \mathcal{P}_{G|A} & \mathcal{P}_{G|G} & \mathcal{P}_{G|C} & \mathcal{P}_{G|T} \\ \mathcal{P}_{C|A} & \mathcal{P}_{C|G} & \mathcal{P}_{C|C} & \mathcal{P}_{C|T} \\ \mathcal{P}_{T|A} & \mathcal{P}_{T|G} & \mathcal{P}_{T|C} & \mathcal{P}_{T|T} \end{pmatrix} \begin{pmatrix} \mathcal{P}_A \\ \mathcal{P}_G \\ \mathcal{P}_C \\ \mathcal{P}_T \end{pmatrix} \\ = \begin{pmatrix} \mathcal{P}_{A|A}\mathcal{P}_A + \mathcal{P}_{A|G}\mathcal{P}_G + \mathcal{P}_{A|C}\mathcal{P}_C + \mathcal{P}_{A|T}\mathcal{P}_T \\ \mathcal{P}_{C|A}\mathcal{P}_A + \mathcal{P}_{C|G}\mathcal{P}_G + \mathcal{P}_{C|C}\mathcal{P}_C + \mathcal{P}_{C|T}\mathcal{P}_T \\ \mathcal{P}_{G|A}\mathcal{P}_A + \mathcal{P}_{G|G}\mathcal{P}_G + \mathcal{P}_{G|C}\mathcal{P}_C + \mathcal{P}_{G|T}\mathcal{P}_T \\ \mathcal{P}_{T|A}\mathcal{P}_A + \mathcal{P}_{T|G}\mathcal{P}_G + \mathcal{P}_{T|C}\mathcal{P}_C + \mathcal{P}_{T|T}\mathcal{P}_T \end{pmatrix}.$$

To interpret this result, let's focus on the bottom entry

$$\mathcal{P}_{T|A}\mathcal{P}_A + \mathcal{P}_{T|G}\mathcal{P}_G + \mathcal{P}_{T|C}\mathcal{P}_C + \mathcal{P}_{T|T}\mathcal{P}_T.$$

Informally, we expect this to give the probability that a site in  $S_1$  has base  $T$ , since we've multiplied the probability of each initial base occurring by the chance that base mutates to a  $T$ , and summed over all possible initial bases. Checking this more formally, the first product appearing on the left is

$$\mathcal{P}_{T|A}\mathcal{P}_A = \mathcal{P}(S_1 = T \mid S_0 = A)\mathcal{P}(S_0 = A).$$

Using equation 1, this is the same as  $\mathcal{P}(S_1 = T \text{ and } S_0 = A)$ . Applying similar reasoning to the other three products shows

$$\begin{aligned} \mathcal{P}_{T|A}\mathcal{P}_A + \mathcal{P}_{T|G}\mathcal{P}_G + \mathcal{P}_{T|C}\mathcal{P}_C + \mathcal{P}_{T|T}\mathcal{P}_T = \\ \mathcal{P}(S_1 = T \text{ and } S_0 = A) + \mathcal{P}(S_1 = T \text{ and } S_0 = G) \\ + \mathcal{P}(S_1 = T \text{ and } S_0 = C) + \mathcal{P}(S_1 = T \text{ and } S_0 = T). \end{aligned}$$

Notice this is the sum of four probabilities of mutually exclusive events. By the addition rule, it gives the probability of the union of the four events, that is, of the event that  $S_1 = T$ :

$$\mathcal{P}_{T|A}\mathcal{P}_A + \mathcal{P}_{T|G}\mathcal{P}_G + \mathcal{P}_{T|C}\mathcal{P}_C + \mathcal{P}_{T|T}\mathcal{P}_T = \mathcal{P}(S_1 = T).$$

If similar reasoning is applied to the other entries in the right hand side of equation 4, we find  $M\mathbf{p}_0 = \mathbf{p}_1$ , where  $\mathbf{p}_1$  is the vector of probabilities for various bases occurring in the sequence  $S_1$ . We can think of  $M$  as a *transition matrix* that tells us how the probabilities of each base in the ancestral sequence  $S_0$  are transformed into the probabilities of each base in the descendent sequence  $S_1$  one time step later.

What would be the meaning of  $M\mathbf{p}_1$ ? For this to make sense biologically, we must assume the probabilistic mutation process over the first time step is identical to that over the next time step. Using the same transition matrix  $M$  of conditional probabilities means each type of base substitution has the same likelihood of occurring as it did before. Furthermore, what happens during the second time step depends only on what the base was at time  $t = 1$  (the information in  $\mathbf{p}_1$ ), and the conditional probabilities (the information in  $M$ ). Whether that site experienced a substitution during the first time step is irrelevant.

To return to our numerical example with  $\mathbf{p}_0$  and  $M$  coming from the data in Table 1, we can compute

$$\mathbf{p}_1 = M\mathbf{p}_0 = \begin{pmatrix} .225 \\ .275 \\ .300 \\ .200 \end{pmatrix}, \quad \mathbf{p}_2 = M\mathbf{p}_1 = \begin{pmatrix} .222 \\ .274 \\ .320 \\ .183 \end{pmatrix}.$$

► What is the sum of the entries in  $\mathbf{p}_1$ ?, in  $\mathbf{p}_2$ ? (You may need to neglect an error due to rounding.) Why must this be the case?

**Markov models.** The model developed above is an example of a *Markov model*. In such a model, we describe a system that must be in one of  $n$  different *states*, but may switch from one state to another with time.

In the DNA substitution model, the system we describe is a site in a DNA sequence. That site is initially in one of 4 states  $A$ ,  $G$ ,  $C$ , or  $T$ , according to the base that occupies it.

We specify initial probabilities that the system is in each of the states by giving a vector of these probabilities,  $\mathbf{p}_0$ . The entries of  $\mathbf{p}_0$  must all be  $\geq 0$  (since they are probabilities) and must add to 1 (since we are certain the system is in one of the states).

We also specify conditional probabilities of the switch from every state to every state over one time step by giving a  $n \times n$  *transition matrix*  $M$ . The entries of  $M$  must all be  $\geq 0$  (since they are probabilities) and each column must add to one (since the conditional probabilities in column  $j$  represent the probabilities of switching from state  $j$  to all states, and we are certain one of these will occur).

An important assumption is made in any Markov model: what happens to the system over a given time step depends only on the state the system is in at the start of that step and the transition probabilities. In particular, there is no ‘memory’ of what state changes might have occurred during earlier time steps that has any effect. We say the conditional probabilities are *independent* of the past history.

► For a DNA substitution model, is it reasonable to assume this independence?

In our DNA model we are also assuming that each site in the sequence behaves identically, and independently of every other site. We used these assumptions in order to find the various probabilities we needed from our sequence data, by thinking of each site as a completely independent trial of the same probabilistic process.

This assumption is probably not very reasonable for DNA in some genes. For instance, since the genetic code allows for many changes in the third site of each codon to have no affect on the product of the gene, one could argue that substitutions in the third sites might be more likely than in the first two sites, violating the assumption that each site behaves identically. Since genes may lead to the production of proteins which are part of life’s processes, the likelihood of change at one site may well be tied to changes at another, violating the assumption of independence.

Nonetheless, we must make simplifying assumptions in order to get anywhere with our model. Further work may find ways around these assumptions, allowing for different conditional probabilities for various sites. Or, we can be careful to take the assumptions into account when using the tools we develop on real data.

For instance, we might only look at the third base of each codon in estimating information from our data, so that it is more reasonable to treat sites as independent and following identical processes.

A matrix whose entries are all  $\geq 0$  and whose columns sum to 1 is called a *Markov matrix*. Actually, you've seen an example of one before in the forest succession model of Chapter ???. That model can be reinterpreted as a Markov model, by imagining it describing one plot in the forest, and tracking the likelihood of the plot being occupied by one type of tree or another.

There are quite a number of theorems concerning certain Markov models that are useful to know about, though we won't go into the proofs. Two that are relevant are:

**THEOREM.** A Markov matrix always has  $\lambda_1 = 1$  as its largest eigenvalue, and has all eigenvalues satisfying  $|\lambda| \leq 1$ . The eigenvector corresponding to  $\lambda_1$  has all non-negative entries.

Unfortunately, this doesn't rule out  $-1$  as an eigenvalue, or having several different eigenvectors with eigenvalue 1. However, there is also:

**THEOREM.** A Markov matrix all of whose entries are positive (*i.e.*, non-zero) always has 1 as a strictly dominant eigenvalue. There will be only one eigenvector (up to scalar multiplication) associated to  $\lambda = 1$ .

Notice that we saw an example of this theorem for the tree model of Chapter ??, where we found the dominant eigenvector was  $(5, 3)$ , with eigenvalue 1. This explains why our numerical experiments with the model led to a stable distribution of  $(A_t, B_t) \approx (625, 375)$ , since  $\frac{625}{375} = \frac{5}{3}$ .

There are a few special Markov models of base substitutions used for DNA sequences which we can analyze very thoroughly.

**The Jukes-Cantor model.** The simplest Markov model of base substitution, the Jukes-Cantor model, adds several additional assumptions to the basic Markov model. First, it assumes all bases occur with equal probability in the ancestral sequence. Thus

$$\mathbf{p}_0 = \left( \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right).$$

Second, in the Jukes-Cantor model the conditional probabilities describing an observable base substitution from any base to any other base are all the same. Thus all possible substitutions are equally likely;  $A \leftrightarrow T$ ,  $A \leftrightarrow C$ ,  $A \leftrightarrow G$ ,  $C \leftrightarrow T$ ,  $C \leftrightarrow G$ , and  $T \leftrightarrow G$  have exactly the same chance of occurring. If we let  $\frac{\alpha}{3}$  denote the conditional probability of a base substitution of any type occurring, so  $\mathcal{P}(S_1 = i \mid S_0 = j) = \frac{\alpha}{3}$  for all  $i \neq j$ , then the 12 off-diagonal entries of the matrix  $M$  will all be  $\frac{\alpha}{3}$ .

► Since the entries in any column of  $M$  add to 1, what should the entries on the main diagonal be?

Therefore, for the Jukes-Cantor model, we use the transition matrix

$$M = \begin{pmatrix} 1 - \alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1 - \alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1 - \alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1 - \alpha \end{pmatrix}.$$

The value of  $\alpha$  will of course depend on the time step we use, and features of the particular DNA sequence we are modeling.

► Why can you think of  $1 - \alpha$  as the probability that no substitution is observed over a time step?

While  $\alpha$  is a probability, we can also interpret it as a rate: it is the rate at which observable base substitutions occur over one time step, and is measured in units of (substitutions per site)/(time step). We emphasize that the *observable* mutations are those that we notice when comparing the ancestral and descendent sequences one time step later; several mutations may actually occur over the time step, but at most one is observable at any site. If back mutations occur during a time step, we may not observe a mutation, even though several occurred.

Mutation rates such as  $\alpha$  for DNA in real organisms are not easily found. Ultimately we'll see how they can be deduced from data. Various researchers have given estimates of  $\alpha$  around  $1.1 \times 10^{-9}$  mutations per site per year for certain sections of chloroplast DNA of maize and barley and around  $10^{-8}$  mutations per site per year for mitochondrial DNA in mammals. The mutation rate for the influenza A virus has been estimated to be as high as .01 mutations per site per year. The rate of mutation is generally found to be a bit lower in coding regions of nuclear DNA than in non-coding DNA. At this point in the development of the model, however, we will treat  $\alpha$  as an unknown constant.

In reality, the mutation rate may not be constant; it may change with time or with location within the DNA. Certainly over the entire evolution of humans from primordial slime it is unreasonable to think that mutation rates have always been the same. However, for shorter periods of time and for DNA serving a fixed purpose, the assumption of a constant mutation rate is sometimes reasonable. When mutation rates are constant, there is said to be a *molecular clock*.

To begin to understand the behavior of the Jukes-Cantor model, let's imagine we have a sequence evolving according to the model and ask ourselves some basic questions about what we will see happening. Remember, our initial sequence has equal proportions of each of the 4 bases, so

$$\mathbf{p}_0 = \left( \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right),$$

and for some small value of  $\alpha$ , the base substitutions occur according to the transition matrix  $M$  given above.

EXAMPLE. For the Jukes-Cantor model, in what proportion of the sites will each base appear after one time step?

To answer this we merely compute

$$\mathbf{p}_1 = M\mathbf{p}_0 = \begin{pmatrix} 1 - \alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1 - \alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1 - \alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1 - \alpha \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}.$$

Thus we find the base composition of the sequence doesn't change under the Jukes-Cantor model. In the language of linear algebra, we'd say that the vector  $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$  is an eigenvector of  $M$  with eigenvalue 1. (In fact, it's the one promised by the two theorems on Markov matrices.) In this context, we might say that  $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$  is an *equilibrium base distribution* for sequences under the Jukes-Cantor model. In earlier chapters, we might have called it a steady state for the model.

EXAMPLE. What proportion of the sites will have a base  $A$  in the ancestral sequence and a  $T$  in the descendent one time step later? In other words, what is  $\mathcal{P}(S_0 = A \text{ and } S_1 = T)$ ?

To answer this, we note

$$\mathcal{P}(S_0 = A \text{ and } S_1 = T) = \mathcal{P}(S_1 = T \mid S_0 = A)\mathcal{P}(S_0 = A).$$

Now the conditional probability  $\mathcal{P}(S_1 = T \mid S_0 = A) = \frac{\alpha}{3}$  can be found as the (4,1)-entry in  $M$  while  $\mathcal{P}(S_0 = A) = \frac{1}{4}$  is an entry in  $\mathbf{p}_0$ . Thus  $\mathcal{P}(S_0 = A \text{ and } S_1 = T) = \frac{\alpha}{12}$ .

EXAMPLE. What is the probability that a base  $A$  in the ancestral sequence will have mutated to become a base  $T$  in the descendent sequence 100 time steps later? In other words, what is the conditional probability  $\mathcal{P}(S_{100} = T \mid S_0 = A)$ ?

To answer this, we first observe that

$$(5) \quad \mathbf{p}_{100} = M^{100}\mathbf{p}_0.$$

Just as the formula  $\mathbf{p}_1 = M\mathbf{p}_0$  holds because the entries of  $M$  are conditional probabilities of various substitutions occurring, the formula in equation (5) must mean that the entries of  $M^{100}$  are conditional probabilities of various net substitutions occurring in the passage from time 0 to time 100. We therefore need to find a certain entry of  $M^{100}$  — the entry in row 4, column 1 — and then we can answer the question.

Of course, finding all entries of  $M^t$  for all  $t$  is of more interest, since that will give us all the conditional probabilities of base substitutions over various numbers of time steps. We base our calculation of  $M^t$  on the insight of Chapter ??: eigenvectors provide the best approach to understanding how powers of matrices behave.

Fortunately the eigenvectors of the Jukes-Cantor matrix  $M$  are easily found. We've already seen one eigenvector (the equilibrium base distribution), but there are 3 more that can be found by trial-and-error or a long computation. The full set is

$$\begin{array}{ll} \mathbf{v}_1 = (1, 1, 1, 1) & \lambda_1 = 1 \\ \mathbf{v}_2 = (1, 1, -1, -1) & \lambda_2 = 1 - \frac{4}{3}\alpha \\ \mathbf{v}_3 = (1, -1, 1, -1) & \lambda_2 = 1 - \frac{4}{3}\alpha \\ \mathbf{v}_4 = (1, -1, -1, 1) & \lambda_2 = 1 - \frac{4}{3}\alpha \end{array}$$

► Check that these are correct by multiplying  $M\mathbf{v}_i$  for each  $i$ .

Notice that the eigenvectors for the Jukes-Cantor model do not depend on the value of the mutation rate  $\alpha$ , though the eigenvalues do.

To find the entries of  $M^t$ , we begin by focusing on the first column of  $M^t$ . The first column can be isolated by taking the product

$$M^t \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \text{first column of } M^t.$$

Now we can express  $(1,0,0,0)$  in terms of the eigenvectors as

$$(1, 0, 0, 0) = \frac{1}{4}\mathbf{v}_1 + \frac{1}{4}\mathbf{v}_2 + \frac{1}{4}\mathbf{v}_3 + \frac{1}{4}\mathbf{v}_4.$$

Thus

$$\begin{aligned} M^t \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} &= \frac{1}{4}M^t\mathbf{v}_1 + \frac{1}{4}M^t\mathbf{v}_2 + \frac{1}{4}M^t\mathbf{v}_3 + \frac{1}{4}M^t\mathbf{v}_4 \\ &= \frac{1}{4}1^t\mathbf{v}_1 + \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \mathbf{v}_2 + \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \mathbf{v}_3 + \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \mathbf{v}_4. \end{aligned}$$

Substituting in the vectors  $\mathbf{v}_i$  we find

$$M^t \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t \\ \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \\ \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \\ \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \end{pmatrix}.$$

The other columns of  $M^t$  are found similarly, giving

$$(6) \quad M^t = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \\ \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \\ \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \\ \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t & \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t \end{pmatrix}.$$

This formula for  $M^t$  is actually quite simple, since it is of the Jukes-Cantor form itself. The value of the Jukes-Cantor parameter for it is just  $\frac{3}{4} - \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t$ .

EXAMPLE. We can now easily answer questions such as: What is the probability that a site that initially has base  $A$  has base  $T$  after 100 time steps. This is the  $(4,1)$ -entry of  $M^{100}$ , which is

$$\frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^{100}.$$

**The Kimura models.** The Jukes-Cantor model is a *one-parameter* model of mutation since it depends on the single parameter  $\alpha$  to specify the mutation rate. Other models use several different parameters to specify mutation rates for several different types of mutations.

A good example of this is the Kimura 2-parameter model, which allows for different rates of transitions and transversions. Imagine that we have mutation rates  $\beta$  for transitions and  $\gamma$  for each of the possible transversions. If we assume these rates are independent of the initial base, then we are saying the off-diagonal entries of the transition matrix are given by:

$$M = \begin{pmatrix} * & \beta & \gamma & \gamma \\ \beta & * & \gamma & \gamma \\ \gamma & \gamma & * & \beta \\ \gamma & \gamma & \beta & * \end{pmatrix}.$$

► Why is it important to use the order  $A, G, C, T$  for the bases to get this matrix?

Since the columns must sum to 1, this means all the diagonal entries must be  $1 - \beta - 2\gamma$ . Notice that if the probabilities of a transition and each transversion are equal so  $\beta = \gamma$ , then this model includes the Jukes-Cantor one as a special case with  $\alpha = 3\beta = 3\gamma$ .

An even more general model is the Kimura 3-parameter model, which assumes a transition matrix of the form

$$M = \begin{pmatrix} * & \beta & \gamma & \delta \\ \beta & * & \delta & \gamma \\ \gamma & \delta & * & \beta \\ \delta & \gamma & \beta & * \end{pmatrix}.$$

By appropriate choice of the parameters, this includes both the Jukes-Cantor and Kimura 2-parameter models as special cases.

Part of the Kimura models is the assumption that the initial base distribution vector is  $\mathbf{p}_0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ . Since this vector is an eigenvector with eigenvalue 1 for both the Kimura 2- and 3-parameter matrices, sequences evolving according to these models have this uniform base distribution at all times. As you'll see in the exercises, all the work done above for the Jukes-Cantor model can be performed for the Kimura 3-parameter model as well.

The general Markov model may well provide the most accurate description of the base substitutions that actually occur in evolution, since it assumes nothing special about the entries in the Markov matrix. It doesn't require any particular relationship between the various conditional probabilities. There are 12 parameters in picking a matrix for this model, since of the 16 entries we may freely pick 3 in each column, but then the fourth is determined by the condition that the columns sum to 1. If we also allow any initial base composition vector  $\mathbf{p}_0$ , then there are 3 additional parameters.

► Why are there only 3 parameters for  $\mathbf{p}_0$ , even though it has 4 entries?

Unless we have specific parameter values in mind for the general Markov model, it's hard to derive detailed results for it of the sort we found for the Jukes-Cantor model. However, as long as all entries of the matrix are positive, the two theorems

stated above do tell us that there must be an equilibrium base distribution. Furthermore, by applying the Strong Ergodic Theorem of Chapter ??, we know that over time, the general Markov model will result in  $\mathbf{p}_t$  approaching this equilibrium distribution, even if the initial base distribution is something else.

### Problems:

1. Review the forest succession model in the text of Chapter ??, in order to interpret it as a Markov model of a single plot in the forest.
  - a. What are the ‘states’ for this model?
  - b. The matrix used in that model was  $\begin{pmatrix} .9925 & .0125 \\ .0075 & .9875 \end{pmatrix}$ . Explain why this is a Markov matrix.
  - c. Explain what conditional probabilities are given by each of the entries in this matrix.
  - d. In the text, we considered a forest that initially had 10 trees of species  $A$  and 990 trees of species  $B$ . What are the initial probabilities of a plot being in each of the states; that is, what is  $\mathbf{p}_0$ ?
2. Recall the Leslie models of Chapter ?. The matrices used in these models are typically not Markov matrices. Why not?
3. Although the Jukes-Cantor model assumes  $\mathbf{p}_0 = (.25, .25, .25, .25)$ , a Jukes-Cantor matrix could describe mutations even with a different  $\mathbf{p}_0$ . Investigate the behavior of a model using a Jukes-Cantor matrix as you vary  $\mathbf{p}_0$  by using a computer. For instance, with  $\alpha = .03$ , and  $\mathbf{p}_0 = (.2, .3, .4, .1)$  you might use the MATLAB commands such as
 

```
a=.03, b=a/3
M=[1-a,b,b,b;b,1-a,b,b;b,b,1-a,b;b,b,b,1-a]
p=[.2; .3; .4; .1]
P=p
for i=1:10
p=M*p
P=[P p]
end
plot(P')
```

  - a. With the value of  $M$  and  $\mathbf{p}_0$  suggested, do you see  $\mathbf{p}_t$  approach its equilibrium value? Approximately how many time steps are necessary for all the  $\mathbf{p}_t$  to be within .05 of the equilibrium? within .01?
  - b. Make several other choices of  $\mathbf{p}_0$  and repeat step (a).
  - c. Using  $\mathbf{p}_0 = (.25, .25, .25, .25)$ , what do you observe? Why?
  - d. Using  $\mathbf{p}_0 = (0, 1, 0, 0)$  what do you observe? What is the biological meaning of this  $\mathbf{p}_0$ ?
4. Investigate the effect of varying  $\alpha$  on the behavior produced by the Jukes-Cantor matrix. Let  $\mathbf{p}_0 = (.2, .3, .4, .1)$  and use MATLAB commands such as those in the previous exercise to:
  - a. Compare the behavior of the model for  $\alpha = .03$  and  $\alpha = .06$ . For which value of  $\alpha$  does the model approach the equilibrium fastest?

- b. Does your observation in part (a) hold for other initial choices of  $\mathbf{p}_0$ ?
- c. Explain in intuitive terms why larger values of  $\alpha$  should result in a quicker approach to the equilibrium.
5. The Markov matrices that describe real DNA mutation tend to have their largest entries along the main diagonal in the (1,1), (2,2), (3,3), and (4,4) positions. Why should this be the case?
6. Make up a  $4 \times 4$  Markov matrix  $M$  with all positive entries, and an initial  $p_0$ . To be biologically realistic, make sure the diagonal entries of  $M$  are the largest.
- a. Use a computer to observe that after many time steps  $\mathbf{p}_t = M^t \mathbf{p}_0$  appears to approach some equilibrium. Estimate the equilibrium vector as accurately as you can.
- b. Is your estimate in part (a) an eigenvector of  $M$  with eigenvalue 1? If not, does it appear to be close to having this property?
- c. Use a computer to compute the eigenvectors and eigenvalues of  $M$ , for instance with the MATLAB command `[S D]=eig(M)`. Is 1 an eigenvalue? Is your estimate of the equilibrium close to its eigenvector?
- d. Are your computations in part (c) consistent with the two theorems about Markov matrices appearing in the text?
7. Express the Kimura 2-parameter model using a  $4 \times 4$  matrix, but with the bases in the order  $A, C, G, T$ . Is this the same as the matrix in the text? Explain.
8. Consider the Markov matrix appearing in equation (3).
- a. Use a computer to find its eigenvectors and eigenvalues. Are they explained by the two theorems of this section?
- b. What is the equilibrium base distribution for this model? Be sure you give a vector whose entries sum to 1.
9. An ancestral DNA sequence of 40 bases was
- CTAGGCTTACGATTACGAGGATCCAAATGGCACCAATGCT*,
- but in a descendent it had mutated to
- CTACGCTTACGACAACGAGGATCCGAATGGCACCAATTGCT*.
- a. Give an initial base distribution vector and a Markov matrix to describe the mutation process.
- b. These sequences were actually produced by a Jukes-Cantor simulation. Is that surprising? Explain. What value would you choose for the Jukes-Cantor parameter  $\alpha$  to approximate your matrix by a Jukes-Cantor one?
10. Data from two comparisons of 400-base ancestral and descendent sequences are shown in Table 5.

$S_1 \setminus S_0$	A	G	C	T	$S'_1 \setminus S'_0$	A	G	C	T
A	92	15	2	2	A	90	3	3	2
G	13	84	4	4	G	3	79	8	2
C	0	1	77	16	C	2	4	96	5
T	4	2	14	70	T	5	1	3	94

TABLE 5. Frequencies from 400 site comparisons for two pairs of sequences

- a. For one of these pairs of sequences a Jukes-Cantor model is appropriate. Which one, and why?

- b. What model would be appropriate for the other pair of sequences? Explain.
11. In MATLAB, type `load seqdata` to read in some simulated sequence data. The three pairs of sequences `s0` and `s1`, `t0` and `t1`, `u0` and `u1`, are simulated ancestor and descendent sequences produced according to three different models. Which one was made according to the Jukes-Cantor model? the Kimura 2-parameter model? a general Markov model? Explain how you can tell. To easily compare sequences by producing a frequency array, use a command like `compseq(s0,s1)`.
12. Suppose we wish to model molecular evolution not at the level of DNA sequences, but rather at the level of the proteins that genes encode.
- a. Create a simple one-parameter mathematical model (similar to the Jukes-Cantor model) describing the process. You will need to use that there are 20 different amino acids from which proteins are constructed in linear chains.
- b. In this situation, how many parameters would the general Markov model have?
13. The MATLAB program `mutate` can be used to simulate the mutation of a DNA sequence according to a Markov model. It will allow you to specify a  $4 \times 4$  Markov matrix  $M$  and initial base distribution vector  $\mathbf{p}_0$ , as well as the number of bases you would like in your sequences.
- a. Use the MATLAB program `mutate` to perform a 10 base simulation for the Jukes-Cantor model with  $\alpha = .1$  and  $\mathbf{p}_0 = (.25, .25, .25, .25)$ . Now imagine that the results of your simulation were two data sequences. Use them to estimate probabilities for an initial base distribution vector and a Markov matrix. (The program `compseq` will be useful for this.) Are your estimates close to what you began with?
- b. Repeat part (a) but using sequences of length 100, and then of length 1000.
- c. The difference between a probabilistic model's description and what actually happens under that model when only a finite number of trials are performed is sometimes called *stochastic error*. What conclusions can you draw from parts (a) and (b) about the stochastic error for short sequences as opposed to long ones?
14. Repeat the last problem, but using your own choice of a  $4 \times 4$  Markov model and initial base distribution. Are the results similar?
15. Suppose you have compared two sequences  $S_\alpha$  and  $S_\beta$  of length 1000 sites and obtained the data in Table 6 for the number of sites with each pair of bases.

$S_\beta \setminus S_\alpha$	A	G	C	T
A	105	25	35	25
G	15	175	35	25
C	15	25	245	25
T	15	25	35	175

TABLE 6. Frequencies of  $S_\beta = i$  and  $S_\alpha = j$  in 1000 site sequence comparison

- a. Assuming  $S_\alpha$  is the ancestral sequence, find an initial base distribution  $\mathbf{p}_0$  and a Markov matrix  $M$  to describe the data. Is your matrix  $M$  Jukes-Cantor? Is  $\mathbf{p}_0$  an equilibrium distribution for  $M$ ?

b. Assuming  $S_\beta$  is the ancestral sequence, find an initial base distribution  $\mathbf{p}'_0$  and a Markov matrix  $M'$  to describe the data. Is your matrix  $M'$  Jukes-Cantor? Is  $\mathbf{p}'_0$  an equilibrium distribution for  $M'$ ?

You should have found that one of your matrices was Jukes-Cantor and the other was not. This can't happen if both  $S_\alpha$  and  $S_\beta$  have base distribution  $(.25, .25, .25, .25)$ .

16. The formula for  $M^t$  for the Jukes-Cantor model can be used to show that powers of  $M$  approach a certain matrix as  $t \rightarrow \infty$ .
- For  $0 < \alpha \leq 1$ , explain why  $-\frac{1}{3} \leq 1 - \frac{4}{3}\alpha < 1$ .
  - Use this to explain how  $(1 - \frac{4}{3}\alpha)^t$  behaves as  $t \rightarrow \infty$ , and thus why

$$M^t \rightarrow \begin{pmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{pmatrix}.$$

Note that each of the columns of this matrix is the equilibrium distribution.

c. Why did we exclude  $\alpha = 0$  from our analysis?

17. Based on the last problem, one might conjecture that powers of a Markov matrix all of whose entries are positive approach a matrix whose columns are the equilibrium distribution. On a computer, investigate this experimentally by creating a Markov matrix, computing very high powers of it to see if the columns become approximately the same, and then checking whether this column is an eigenvector with eigenvalue 1 of the original matrix.
18. Show the product of two Jukes-Cantor matrices is again a Jukes-Cantor matrix as follows: Let  $M(\alpha_1)$  be the Jukes-Cantor matrix with parameter  $\alpha_1$ , and  $M(\alpha_2)$  the Jukes-Cantor matrix with parameter  $\alpha_2$ . Compute  $M(\alpha_1)M(\alpha_2)$  to show it has the form  $M(\alpha_3)$ . Give a formula for  $\alpha_3$  in terms of  $\alpha_1$  and  $\alpha_2$ .
19. Show the product of two Kimura 3-parameter matrices is again a Kimura 3-parameter matrix.
20. Show the Kimura 3-parameter matrix has the same eigenvectors as those given in the text for the Jukes-Cantor matrix. What are the eigenvalues?
21. Use the results of the last problem to give formulas for the entries of the first column of  $M^t$  where  $M = M(\beta, \gamma, \delta)$  is the Kimura 3-parameter matrix. (The other columns could be handled similarly, leading to the result that  $M(\beta, \gamma, \delta)^t = M(\beta', \gamma', \delta')$  where

$$\begin{aligned} \beta' &= \frac{1}{4} + \frac{1}{4}(1 - 2\gamma - 2\delta)^t - \frac{1}{4}(1 - 2\beta - 2\delta)^t - \frac{1}{4}(1 - 2\beta - 2\gamma)^t \\ \gamma' &= \frac{1}{4} - \frac{1}{4}(1 - 2\gamma - 2\delta)^t + \frac{1}{4}(1 - 2\beta - 2\delta)^t - \frac{1}{4}(1 - 2\beta - 2\gamma)^t \\ \delta' &= \frac{1}{4} - \frac{1}{4}(1 - 2\gamma - 2\delta)^t - \frac{1}{4}(1 - 2\beta - 2\delta)^t + \frac{1}{4}(1 - 2\beta - 2\gamma)^t. \end{aligned}$$

22. The Jukes-Cantor model can be presented in a different form as a  $2 \times 2$  Markov model. Let  $q_t$  represent the fraction of sites that agree between the ancestral sequence and the descendent sequence at time  $t$ , and  $p_t$  the fraction that differ, so  $q_0 = 1$  and  $p_0 = 0$ . Assume that over each time step, the probability that a base substitution occurs is  $\alpha$ , and that each of the 3 possible base substitutions

is equally likely. Then

$$\begin{pmatrix} q_{t+1} \\ p_{t+1} \end{pmatrix} = \begin{pmatrix} 1 - \alpha & \frac{\alpha}{3} \\ \alpha & 1 - \frac{\alpha}{3} \end{pmatrix} \begin{pmatrix} q_t \\ p_t \end{pmatrix}, \quad \begin{pmatrix} q_0 \\ p_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

- a. Explain why each entry in the matrix has the value it does. (Observe that  $1 - \frac{\alpha}{3} = (1 - \alpha) + \frac{2\alpha}{3}$ .)
  - b. Compute the steady-state of the model by finding the eigenvector with eigenvalue 1.
  - c. Find the other eigenvalue and eigenvector for the matrix.
  - d. Use (b) and (c), together with the initial conditions  $(q_0, p_0) = (1, 0)$  to give a formula for  $q_t$  and  $p_t$  as functions of time.
23. This exercise will derive one of the entries in equation (6) another way, in the style of Chapter ???. Let  $q_t$  denote the probability that the base at a fixed site at time  $t$  is the same as it was at time 0, and let  $\alpha$  denote the probability of a substitution in a single time step for the Jukes-Cantor model.
- a. Explain why

$$q_{t+1} = (1 - \alpha)q_t + \frac{\alpha}{3}(1 - q_t)$$

(You'll need to think about two ways the base at time  $t + 1$  might agree with that at time 0: either it agreed at time  $t$  and didn't change, or didn't agree at time  $t$  and changed back to the original base.) What value should  $q_0$  have? Investigate the behavior of this model in MATLAB using `onpop`.

The equation in (a) simplifies to

$$q_{t+1} = \frac{\alpha}{3} + \left(1 - \frac{4\alpha}{3}\right) q_t.$$

Notice that this model is a little different from those we dealt with in Chapter ???. If we graphed  $q_{t+1}$  as a function of  $q_t$ , we'd get a straight line, but because the form of the equation is  $q_{t+1} = s + rq_t$  rather than just  $q_{t+1} = rq_t$ , we can't call it linear. (The term 'linear' in this context requires that there be no constant term.) Instead a model of the form  $q_{t+1} = s + rq_t$  is called an *affine* model. Affine models can be converted to linear models and analyzed as outlined in the next few steps:

- b. Find the equilibrium  $q^*$  of the model by solving  $q^* = \frac{\alpha}{3} + \left(1 - \frac{4\alpha}{3}\right) q^*$ .
- c. Let  $q_t = q^* + \epsilon_t$  to focus on the perturbation  $\epsilon_t$  from equilibrium. Substitute this and a similar expression for  $q_{t+1}$  into the model equation, and simplify to get an equation expressing  $\epsilon_{t+1}$  in terms of  $\epsilon_t$ . Your result should be linear.
- d. What is  $q_0$ ? Use this value to give the value of the initial perturbation  $\epsilon_0$ .
- e. Based on your work in parts (c) and (d), give a formula for  $\epsilon_t$  in terms of  $t$ .
- f. From (c) and (e) show that

$$q_t = \frac{1}{4} + \frac{3}{4} \left(1 - \frac{4}{3}\alpha\right)^t.$$

### 5. Phylogenetic distances

With a model of DNA mutation in hand, we can better understand how to relate the amount of mutation that we observe in comparing an ancestral and descendent sequence to the amount of mutation that must have actually occurred. We'll be able to uncover the amount of hidden mutation that was obscured by subsequent mutations at the same site.

To frame the issue we want to address more clearly, let's consider the Jukes-Cantor example of the last section. There, we imagined modeling sequence mutation by the Jukes-Cantor matrix

$$M = M(\alpha) = \begin{pmatrix} 1 - \alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1 - \alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1 - \alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1 - \alpha \end{pmatrix},$$

and computed the entries of  $M^t$  for  $t = 0, 1, 2, 3, \dots$ . The diagonal entries of  $M^t$  turned out to all be

$$\frac{1}{4} + \frac{3}{4} \left(1 - \frac{4}{3}\alpha\right)^t.$$

Now the diagonal entries of  $M^t$  give conditional probabilities that the base at time  $t$  is the same as the base at time 0. In other words, they indicate the probability of observing no change when the site at time 0 is compared to the one at time  $t$ . Since all these diagonal entries are equal that means that at time step  $t$  we would expect to observe that the fraction of sites that agreed with their initial base was given by the formula

$$q(t) = \frac{1}{4} + \frac{3}{4} \left(1 - \frac{4\alpha}{3}\right)^t.$$

The fraction of sites that are different, then, will be

$$p(t) = 1 - q(t) = \frac{3}{4} - \frac{3}{4} \left(1 - \frac{4\alpha}{3}\right)^t$$

► Why could you also get the formula for  $p(t)$  by adding the three off-diagonal entries in any column of  $M^t$ ?

In the graph of  $p(t)$  in Figure 1, we of course see that  $p(0) = 0$  since at time  $t = 0$  no substitutions have yet occurred. More interestingly, we see that the fraction of sites that differ from their original base gradually increases with  $t$ , approaching the value  $\frac{3}{4}$ . This fraction never exceeds  $\frac{3}{4}$  however.

► Even if so much mutation has occurred that the two sequences appear to be completely unrelated, you would expect to find agreement at  $\frac{1}{4}$  of the sites. Why?

The graph also illustrates that for each time  $t$ ,  $p(t)$  has a different value. That means that given any value  $0 \leq p \leq 3/4$  we should be able to find a  $t$  with  $p(t) = p$ . That is, from the proportion of sites that differ between two sequences, we should be able to recover the number of elapsed time steps (assuming we know  $\alpha$ ). For real sequence data  $p$  is easily estimated, though the elapsed time  $t$  and the mutation rate  $\alpha$  usually aren't known. Recovering them from data is now our goal.

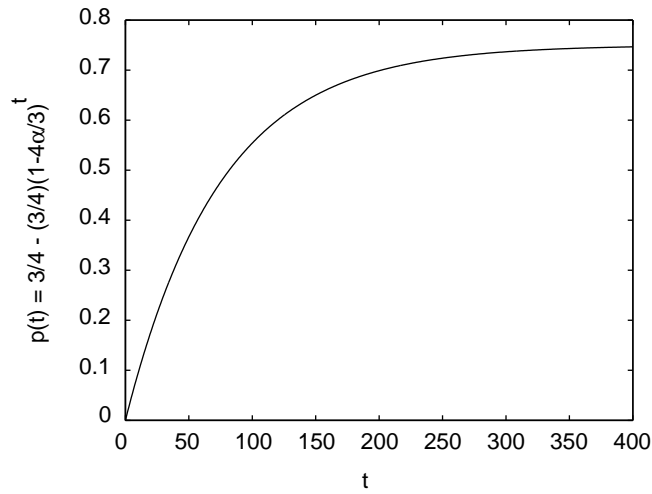


FIGURE 1. The Jukes-Cantor model,  $\alpha = .01$ : Fraction of differing sites at time  $t$

**The Jukes-Cantor distance.** Suppose we have records of an original DNA sequence and also of a mutated version of it from some later time. Suppose we also believe the Jukes-Cantor model describes the mutation process that occurred, but we do not know either the mutation rate  $\alpha$  or the number of elapsed time steps  $t$ .

From the DNA sequence data, we can estimate  $p = p(t)$  by comparing many sites before and after mutation and using the proportion of sites that disagree in the two sequences as an estimate. For instance, if the original sequence were *ATTGAC* and the final one *ATGGCC*, we'd estimate  $p(t) = 2/6 \approx .333$ . Of course with real data it's best to have much longer DNA sequences so that we have more confidence in our estimate.

With  $p = p(t)$  estimated, how do we recover information on the mutation rate  $\alpha$  and the amount of elapsed time  $t$ ? Since

$$p = \frac{3}{4} - \frac{3}{4} \left(1 - \frac{4}{3}\alpha\right)^t,$$

solving for  $t$  yields

$$(7) \quad t = \frac{\ln\left(1 - \frac{4}{3}p\right)}{\ln\left(1 - \frac{4}{3}\alpha\right)}.$$

To go further, we need to realize that our choice of a step size for time in formulating our model affects both the value of the mutation rate  $\alpha$ , and the number of elapsed time steps between ancestor and descendent. We can't really expect to recover both of these. However, the product of the two does have a meaning that is more intrinsic to what we are modeling: let

$$\begin{aligned} d &= t\alpha \\ &= (\text{number of time steps})(\text{mutation rate}) \\ &= (\text{number of time steps})(\text{number of substitutions per site / time step}) \\ &= (\text{expected number of substitutions per site during the elapsed time}). \end{aligned}$$

We emphasize that this expected number of substitutions includes even those we don't observe because they are hidden by subsequent substitutions.

To extract  $d = t\alpha$  from equation (7), we must use an approximation. Now  $\ln(1+x) \approx x$  when  $x$  is near 0 (see the problems). Furthermore we can be sure  $-\frac{4}{3}\alpha$  is near 0 if we assume that we have chosen a time step that is very small, so that the mutation rate per time step,  $\alpha$ , is also very small. Thus

$$\ln\left(1 - \frac{4}{3}\alpha\right) \approx -\frac{4}{3}\alpha.$$

Substituting this into equation (7) gives

$$\begin{aligned} t &\approx \frac{\ln\left(1 - \frac{4}{3}p\right)}{-\frac{4}{3}\alpha} \\ &\approx -\frac{3}{4\alpha} \ln\left(1 - \frac{4}{3}p\right), \end{aligned}$$

or

$$d = t\alpha \approx -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right).$$

If our time steps are made smaller, so the mutation rate  $\alpha$  is also smaller, the approximation used for the logarithm is increasingly accurate. We therefore define the *Jukes-Cantor distance* between DNA sequences  $S_0$  and  $S_1$  as

$$d_{JC}(S_0, S_1) = -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right),$$

where  $p$  is the fraction of sites that disagree in comparing  $S_0$  to  $S_1$ . Provided the Jukes-Cantor model accurately describes the evolution of one sequence into another, it is an estimate of the total number of substitutions per site that occurred during the evolution.

'Distance' here is an abstract notion of how different the sequences are due to mutations. Recall that if the mutation rate is constant over an evolutionary history, we say there is a molecular clock. Provided a molecular clock hypothesis is valid, the distance computed here is proportional to the amount of elapsed time, with the constant of proportionality being the mutation rate. Thus the distance can be thought of as a measure of how much time was required for one sequence to mutate into the other. If the molecular clock hypothesis does not hold, it is still a reconstruction of the average number of substitutions that occurred at any one site. The larger it is, the greater the evolutionary change.

While we were unable to recover either the mutation rate  $\alpha$  or the number of elapsed time periods  $t$  by themselves, we could at least recover the product of the two from comparing sequences. If there is some other data (such as a geological record) suggesting the time involved, then the mutation rate can be found from  $d_{JC}$ . This is one way that real DNA mutation rates are estimated.

EXAMPLE. Consider the two 40 base sequences at the end of Section 3. From Table 1 we find that 11 of the sites have undergone a substitution, so  $p = 11/40 = .2750$ . Thus

$$d_{JC}(S_0, S_1) = -\frac{3}{4} \ln\left(1 - \frac{4 \cdot 11}{3 \cdot 40}\right) \approx .3426$$

Thus while we observed .2750 substitutions per site on average, we estimate that in the course of evolution .3426 substitutions per site occurred. Hidden mutations account for the difference.

**The Kimura distances.** Given any Markov model of base substitutions we could hope to imitate the steps above to derive an appropriate formula reconstructing the amount of mutation that has occurred. For the Kimura models, you'll find an exercise that steps you through the procedure. The final formula for the Kimura 3-parameter model is

$$d_{K3} = -\frac{1}{4}(\ln(1 - 2\beta - 2\gamma) + \ln(1 - 2\beta - 2\delta) + \ln(1 - 2\gamma - 2\delta)),$$

where  $\beta$ ,  $\gamma$ , and  $\delta$  are estimates of parameters for a Kimura 3-parameter matrix describing the mutation of the initial sequence to the final.

Of course, if  $\gamma = \delta$ , this also gives a distance for the Kimura 2-parameter model. In that case,  $\beta$  is the probability of a transition, while  $\gamma + \delta = 2\gamma$  is the probability of a transversion. Thus if from sequence data we estimate the probability of a transition  $p_1$  by counting all transitions and dividing by the length of the sequence, and the probability of a transversion  $p_2$  similarly, we have

$$d_{K2} = -\frac{1}{2}\ln(1 - 2p_1 - p_2) - \frac{1}{4}\ln(1 - 2p_2).$$

If sequence data seems to indicate that transitions and each transversion type did not proceed at equal rates, then the Jukes-Cantor model is a poor one, and so the Kimura distance formulas are better choices for estimating the total amount of mutation.

**Additive and symmetric distances: Log-det.** The distance formulas given so far assume the data are consistent with either the Jukes-Cantor model, or a Kimura 2- or 3-parameter model. Since these models don't necessarily describe all sequence data well, it's natural to ask for a distance formula for the general Markov model.

To motivate such a formula, we won't focus on reconstructing the total number of base substitutions that occurred, but rather on a property shared by both the Jukes-Cantor and Kimura distances.

This property concerns the behavior of the distance formula when we consider two successive mutation processes. Imagine an ancestral sequence  $S_0$  from which has evolved  $S_1$ , from which in turn has evolved  $S_2$ , as shown schematically in Figure 2.

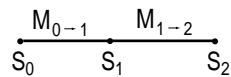


FIGURE 2. Three sequences in evolutionary order

Let  $M_{0 \rightarrow 1} = M(\alpha_1)$  and  $M_{1 \rightarrow 2} = M(\alpha_2)$  be two Jukes-Cantor matrices, describing the two mutation processes as shown. Then we can calculate a mutation

matrix  $M_{0 \rightarrow 2}$  for the full passage from  $S_0$  to  $S_2$  as the product

$$M_{0 \rightarrow 2} = M_{1 \rightarrow 2} M_{0 \rightarrow 1}.$$

► Why are the matrices multiplied in this order?

A short calculation shows that  $M_{0 \rightarrow 2}$  is also a Jukes-Cantor matrix,  $M_{0 \rightarrow 2} = M(\alpha_3)$ , with

$$\alpha_3 = \alpha_1 + \alpha_2 - \frac{4}{3}\alpha_1\alpha_2.$$

As part of the Jukes-Cantor model, suppose the base distribution for each of  $S_0$ ,  $S_1$ , and  $S_2$  is the equilibrium  $(1/4, 1/4, 1/4, 1/4)$ . Then in passing from one sequence to the next by the Jukes-Cantor matrices  $M(\alpha_i)$ , we find the fraction of sites that change is  $p = \alpha_i$ , and so the Jukes-Cantor distances are

$$\begin{aligned} M(\alpha_1) &: -\frac{3}{4} \ln \left( 1 - \frac{4}{3}\alpha_1 \right) \\ M(\alpha_2) &: -\frac{3}{4} \ln \left( 1 - \frac{4}{3}\alpha_2 \right) \\ M(\alpha_1)M(\alpha_2) = M(\alpha_3) &: -\frac{3}{4} \ln \left( 1 - \frac{4}{3}(\alpha_1 + \alpha_2 - \frac{4}{3}\alpha_1\alpha_2) \right) \end{aligned}$$

But a little algebra shows

$$-\frac{3}{4} \ln \left( 1 - \frac{4(\alpha_1 + \alpha_2 - \frac{4}{3}\alpha_1\alpha_2)}{3} \right) = \left( -\frac{3}{4} \ln \left( 1 - \frac{4\alpha_1}{3} \right) \right) + \left( -\frac{3}{4} \ln \left( 1 - \frac{4\alpha_2}{3} \right) \right).$$

This means that *multiplying* two Jukes-Cantor matrices corresponds to *adding* the associated distances.

We can see why this had to be the case if we recall that the Jukes-Cantor distance is recovering the total number of substitutions per site that must have occurred, including hidden ones. If we imagine a sequence mutating first according to one matrix and then the other (*i.e.*, according to the product of the matrices), then the total number of substitutions per site would be the sum of those described by each individual matrix.

Returning to the general Markov model, we'd like a definition of distance between sequences that has the *additive property* that

$$d(S_0, S_2) = d(S_0, S_1) + d(S_1, S_2)$$

in situations described by Figure 2.

To define such a distance, suppose  $F$  is the  $4 \times 4$  frequency array obtained by comparing sites in sequences  $S_0$  and  $S_1$ . Let  $\mathbf{f}_0$  and  $\mathbf{f}_1$  be the frequency vectors for the bases in  $S_0$  and  $S_1$ , respectively. For instance  $F$  might be the entries in Table 1 with  $\mathbf{f}_0$  and  $\mathbf{f}_1$  its column and row sums. Then one version of the *log-det distance* (also called the *paralinear distance* in this form) between  $S_0$  and  $S_1$  is defined by

$$d_{LD}(S_0, S_1) = -\frac{1}{4} \left( \ln(\det(F)) - \frac{1}{2} \ln(g_0 g_1) \right),$$

where  $g_i$  is the product of the 4 entries in  $\mathbf{f}_i$ . Recall from Chapter ?? that 'det' denotes the determinant of a matrix. Since the argument for why this distance is additive depends on some knowledge of linear algebra beyond this text, we leave it to the exercises.

The meaning of the log-det distance is harder to interpret than the other distances we've discussed. Unlike those, it usually isn't just the total number of mutations per site that must have occurred over the evolutionary history. Still, you should think of it as some measure of the amount of mutation that has occurred. In special circumstances, such as when the Jukes-Cantor or Kimura models apply exactly, it gives the same result as they do, as you will also see in the exercises.

The key fact that all the phylogenetic distances we've discussed are additive will be extremely useful in the next chapter, when we turn to constructing phylogenetic trees relating many species.

Another useful property of all of these distances is *symmetry*. While we thought of having an ancestral and descendent sequence in discussing the various distances, in fact none of the final formulas depend on knowing which one of the sequences was the ancestral one. For instance the Jukes-Cantor distance is calculated from first finding the fraction of sites that differ in the 2 sequences. If we had the same sequences, but switched which one we imagined was ancestral, we'd calculate exactly the same distance. This means

$$d(S_0, S_1) = d(S_1, S_0).$$

This property will also be very valuable to us, since usually we don't have an ancestral sequence and a descendent one, but rather two descendents. In the exercises you'll see how symmetry helps us use a distance formula in this circumstance.

### Problems:

1. Calculate  $d_{JC}(S_0, S_1)$  for the two 40 base sequences

$S_0$  : CTAGGCTTACGATTACGAGGATCCAAATGGCACCAATGCT

$S_1$  : CTACGCTTACGACAACGAGGATCCGAATGGCACCAATTGCT.

2. Ancestral and descendent sequences of 400 bases were simulated according to the Jukes-Cantor model. A comparison of aligned sites gave the frequency data in Table 7.

$S_1 \setminus S_0$	A	G	C	T
A	90	3	3	2
G	3	79	8	2
C	2	4	96	5
T	5	1	3	94

TABLE 7. Frequencies of  $S_1 = i$  and  $S_0 = j$  in 400 site sequence comparison

- a. Compute the Jukes-Cantor distance to ten decimal digits, showing all steps.
- b. Compute the Kimura 2-parameter distance to ten decimal digits, showing all steps.
- c. Are the answers to (a) and (b) identical? Explain.

3. Ancestral and descendent sequences of 400 bases were simulated according to the Kimura 2-parameter model with  $\gamma = \beta/5$ . A comparison of aligned sites gave the frequency data in Table 8.

$S_1 \setminus S_0$	A	G	C	T
A	92	15	2	2
G	13	84	4	4
C	0	1	77	16
T	4	2	14	70

TABLE 8. Frequencies of  $S_1 = i$  and  $S_0 = j$  in 400 site sequence comparison

- Compute the Jukes-Cantor distance to ten decimal digits, showing all steps.
  - Compute the Kimura 2-parameter distance to ten decimal digits, showing all steps.
  - Which of these is likely to be a better estimate of the number of substitutions per site that actually occurred? Explain.
- Compute the Kimura 3-parameter and log-det (paralinear) distance for the sequences of the last two problems.
  - Graph  $d_{JC}$  as a function of  $p$ .
    - Why does  $d_{JC} = 0$  if two sequences are identical?
    - Why does  $d_{JC}$  not make sense if two sequences differ in  $3/4$  or more of the sites? Should this cause problems when trying to use the formula on real data?
    - Explain in biological terms why if two sequences differ in just under  $3/4$  of the sites, the value of  $d_{JC}$  should be very large.
  - Complete the gaps in the derivation of the formulas for  $d_{JC}$  in the text by doing all the necessary algebra on the equation

$$p = \frac{3}{4} - \frac{3}{4} \left(1 - \frac{4}{3}\alpha\right)^t$$

to find the formula in equation (7) for  $t$  in terms of  $p$  and  $\alpha$ .

- The Jukes-Cantor distance formula is sometimes stated as

$$d_{JC} = -\frac{3}{4} \ln \left( \frac{4q-1}{3} \right),$$

where  $q$  is the proportion of bases that are the same in the ‘before’ and ‘after’ sequences. Derive this formula from the one in the text.

- Give numerical evidence that the approximation  $\ln(1+x) \approx x$  is valid for small  $x$  by making a table of values of  $x$  and  $\ln(1+x)$  for  $x$  close to 0. Give graphical evidence by plotting  $y = \ln(1+x)$  and  $y = x$ .
- (Calculus) Show the approximation  $\ln(1+x) \approx x$  is valid for  $x$  near 0 by using calculus to find the tangent line approximation to  $y = \ln(1+x)$  at the point where  $x_0 = 0$ .
- In practice when applying a distance formula to real DNA sequence data, it’s uncommon to have sequences for both an ancestor and a descendent. Instead we usually have two descendent DNA sequences  $S_1$  and  $S_2$  that mutated from a common, yet unknown, ancestral sequence  $S_0$ , as in Figure 3. From the data,

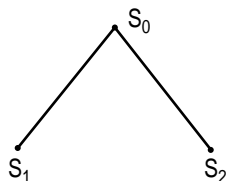


FIGURE 3. An evolutionary tree

we can only compute  $d(S_1, S_2)$ . Show that for an additive, symmetric distance this is the same as  $d(S_0, S_1) + d(S_0, S_2)$ .

11. When transitions are more frequent than transversions, the Kimura 2-parameter distance often gives a larger value than the Jukes-Cantor distance. Explain this informally by explaining why hidden mutations are more likely under this circumstance.
12. The Jukes-Cantor distance is an estimate of the number of mutations that occurred per site over the course of one sequence evolving from another. A simpler estimate for this number is just  $p$ , the proportion of sites that have changed from the initial to final sequence.
  - a. Explain why multiple mutations at the same site would cause  $p$  to be less reliable. Does it give an overestimate or underestimate of the true amount of mutation?
  - b. Give an intuitive explanation of why, if  $p$  is relatively small, so that the sequences have few differences, this simpler estimate might be reasonable anyway.
  - c. Explain why part (b) is consistent with the Jukes-Cantor model. That is, explain why for small  $p$

$$-\frac{3}{4} \ln \left( 1 - \frac{4}{3}p \right) \approx p$$

by using the approximation for  $\ln(1+x)$  valid for small  $x$ .

- d. It has been claimed that if the  $p$  is less than .1, it can be used as a reasonable approximation of the Jukes-Cantor distance. Do you agree? Illustrate by graphing both  $d = -\frac{3}{4} \ln \left( 1 - \frac{4}{3}p \right)$  and  $d = p$  for  $0 \leq p < 3/4$ .
13. Show that the formula for the Jukes-Cantor distance can be recovered from the formula for the Kimura 3-parameter distance by letting  $\beta$ ,  $\gamma$ , and  $\delta$  all be  $\alpha/3$ .
14. Use the MATLAB program `mutate` to simulate a sequence evolving according to the Jukes-Cantor model for  $t = 400$  time steps, using a matrix with parameter  $\alpha = .001$  for each time step. Compute a frequency array of base combinations with `F=compseq(Sinit,Sfinal)` and then compute the Jukes-Cantor distance with `distJC(F)`. Is the computed distance  $\alpha t = .4$ ? If not, explain why not.
15. In MATLAB, type `load seqdata` to read in some simulated sequence data. Type `who` to see the names of the things you just loaded.
  - a. Compute all six Jukes-Cantor distances between the sequences  $a1$ ,  $a2$ ,  $a3$ , and  $a4$ . You can compute a frequency array for base combinations with `F=compseq(a1,a2)` and then compute the distance with `distJC(F)`.
  - b. Suppose these sequences came from currently living species whose evolutionary relationships we would like to deduce. Draw the evolutionary tree that you

believe best describes the relationship. Explain how you've used the distance data in your reasoning.

Note: This problem is the subject of the next chapter.

16. Derive the formula for the Kimura 3-parameter distance as follows. Refer to exercise 21 of the last section which gives formulas for the parameters  $\beta'$ ,  $\gamma'$ , and  $\delta'$  in  $M(\beta', \gamma', \delta') = M(\alpha, \beta, \gamma)^t$ .
- a. Show that

$$\begin{aligned} 1 - 2\beta' - 2\delta' &= (1 - 2\beta - 2\delta)^t, \\ 1 - 2\beta' - 2\gamma' &= (1 - 2\beta - 2\gamma)^t, \\ \text{and } 1 - 2\gamma' - 2\delta' &= (1 - 2\gamma - 2\delta)^t. \end{aligned}$$

- b. Use (a) to show

$$\begin{aligned} \ln(1 - 2\beta' - 2\delta') + \ln(1 - 2\beta' - 2\gamma') + \ln(1 - 2\gamma' - 2\delta') &= \\ t(\ln(1 - 2\beta - 2\delta) + \ln(1 - 2\beta - 2\gamma) + \ln(1 - 2\gamma - 2\delta)). \end{aligned}$$

- c. Assuming  $\beta$ ,  $\gamma$ , and  $\delta$  are all small, use the approximation  $\ln(1 + x) \approx x$  for  $x \approx 0$  in the last equation to get

$$\ln(1 - 2\beta' - 2\delta') + \ln(1 - 2\beta' - 2\gamma') + \ln(1 - 2\gamma' - 2\delta') \approx -4t(\beta + \gamma + \delta).$$

- d. Explain why  $\beta + \gamma + \delta$  should be interpreted as the total rate of base substitution, and thus why it is reasonable to define

$$d_{K3} = -\frac{1}{4}(\ln(1 - 2\beta' - 2\delta') + \ln(1 - 2\beta' - 2\gamma') + \ln(1 - 2\gamma' - 2\delta')).$$

17. (Linear Algebra) The goal of this problem is to show that the Jukes-Cantor distance is a special case of the log-det distance. You will need to know the following two facts about determinants of  $k \times k$  matrices that are proved in a Linear Algebra course:

i)  $\det(cA) = c^k \det(A)$ .

ii)  $\det(A) =$  the product of  $A$ 's  $k$  eigenvalues.

- a. Suppose two sequences  $S_0$  and  $S_1$  of length  $N$  were compared and the frequency table  $F$  was found to be *exactly* described by a Jukes-Cantor matrix  $M(\alpha)$  with base distribution  $(1/4, 1/4, 1/4, 1/4)$  for  $S_0$ . Show that  $F = \frac{N}{4}M(\alpha)$ .

- b. Explain why  $\mathbf{f}_1 = \mathbf{f}_2 = (N/4, N/4, N/4, N/4)$ .

- c. Use the facts above to show that in this case  $d_{LD}(S_0, S_1) = d_{JC}(S_0, S_1)$ .

18. (Linear Algebra) Proceeding as in the last problem, show that the Kimura 3-parameter distance is a special case of the log-det distance.

19. (Linear Algebra) Show the log-det distance formula is additive and symmetric through the following steps. You will need to know the following three facts about determinants of  $k \times k$  matrices that are proved in a Linear Algebra course:

i)  $\det(AB) = \det(A)\det(B)$ .

- ii) If the  $(i, j)$ -entries of  $D$  are all zero for  $i \neq j$ , then

$$\det(D) = D(1, 1) \cdot D(2, 2) \cdots D(k, k).$$

- iii)  $\det(A^T) = \det(A)$ , where  $A^T$ , the *transpose* of  $A$ , is a matrix whose  $(i, j)$ -entry is the  $(j, i)$ -entry of  $A$ .

a. For the situation in Figure 2, with initial base distribution  $\mathbf{p}_0$  in  $S_0$ , explain why  $\mathbf{p}_1 = M_{0 \rightarrow 1}\mathbf{p}_0$  and  $\mathbf{p}_2 = M_{1 \rightarrow 2}\mathbf{p}_1$  are the base distributions in  $S_1$  and  $S_2$  respectively.

b. For the vector  $\mathbf{p}_i = (a, b, c, d)$  let

$$D_i = \begin{pmatrix} \sqrt{a} & 0 & 0 & 0 \\ 0 & \sqrt{b} & 0 & 0 \\ 0 & 0 & \sqrt{c} & 0 \\ 0 & 0 & 0 & \sqrt{d} \end{pmatrix}.$$

Then for each pair  $i, j$  with  $0 \leq i < j \leq 2$ , define the matrix

$$N_{i \rightarrow j} = D_j^{-1} M_{i \rightarrow j} D_i.$$

Show  $N_{1 \rightarrow 2} N_{0 \rightarrow 1} = N_{0 \rightarrow 2}$ , and use fact (i) to conclude

$$\ln(\det(N_{1 \rightarrow 2})) + \ln(\det(N_{0 \rightarrow 1})) = \ln(\det(N_{0 \rightarrow 2})).$$

c. Show the frequency array for comparing  $S_i$  to  $S_j$  is  $F_{i \rightarrow j} = D_j N_{i \rightarrow j} D_i$ , and then use fact (i) to show

$$\ln(\det(F_{i \rightarrow j})) = \ln(\det(N_{i \rightarrow j})) + \ln(\det(D_i)) + \ln(\det(D_j)).$$

d. Combine (b), (c), and fact (ii) to show the log-det distance is additive.

e. Explain why  $F_{j \rightarrow i} = F_{i \rightarrow j}^T$ , and then use fact (iii) to show the log-det distance is symmetric.

### Projects:

1. Investigate how the Jukes-Cantor distance formula performs on simulated sequence data produced according to the Jukes-Cantor model.

The MATLAB program `mutate` can be used to simulate DNA mutations according to any specified Markov model of base substitution. Then the Jukes-Cantor distance can be computed for the sequences so produced. However, seldom does the Jukes-Cantor distance exactly recover the value of  $\alpha t$  used in the simulation.

Explore the performance of the Jukes-Cantor distance formula for recovering  $\alpha t$  on data that is produced by the Jukes-Cantor model under varying circumstances.

Suggestions:

- So that the derivation of the Jukes-Cantor distance formula is valid, pick a small value of the mutation rate, such as  $\alpha = .001$  to use in all your simulations.
- Use `compseq` to compare sequences, and `distJC` to compute distances.
- For some fixed value of  $t$ , say around 300, perform a number of simulations for various values of  $N$ , and compute the Jukes-Cantor distances. Compare these to  $\alpha t$ . Do you see a pattern in how accuracy varies with  $N$ ?
- For some fixed value of  $N$ , say around 400, perform a number of simulations for various values of  $t$ , and compute the Jukes-Cantor distances. Compare these to  $\alpha t$ . Do you see a pattern in how accuracy varies with  $t$ ?

- In comparing the computed values of  $d_{JC}$  to  $\alpha t$ , you could consider either  $d_{JC} - \alpha t$  or  $d_{JC}/(\alpha t)$ . What do each of these mean?
  - For a fixed choice of  $N$  and  $t$ , do many simulations. Then present your results by plotting a histogram of the values you find for  $d_{JC}$ . Do they appear to cluster around  $\alpha t$ ? How spread out is the histogram? Compute means and standard deviations.
  - Repeat the last step, changing  $N$  or  $t$  to a new value.
  - What conclusions can you draw about using the Jukes-Cantor formula for real data? Does it appear to be most accurate when sequences are long or short? when the length of elapsed time is large or small? Can you give an intuitive explanation of why this should be the case?
2. Investigate how the various distance formulas perform on simulated sequence data produced according to models *different* from the one underlying the distance formula.

This is an important issue since real sequence evolution is at best only approximately described by any of these models.

If, for instance, one sequence evolved from another according to the Kimura 2-parameter model with  $\beta \neq \gamma$ , then we shouldn't expect the Jukes-Cantor distance to be a valid reconstruction of the amount of mutation. However, it should be close if  $\beta$  and  $\gamma$  are close to one another.

As in the project above, the MATLAB program `mutate` can simulate DNA mutations according to a specified Markov model of base substitution. Then the various distances can be computed for the sequences so produced, using `compseq`, `distJC`, `distK2` and `distLD`.

Suggestions:

- Explore the performance of the Jukes-Cantor and Kimura 2-parameter distance formulas for recovering the total amount of mutation on data that is produced by various Kimura 2-parameter matrices. You should keep the parameters  $\beta$  and  $\gamma$  very small, and use a large number of time steps for `mutate`. If  $\beta/\gamma = \kappa$ , how different can  $\kappa$  be from 1 for the Jukes-Cantor formula to be close to correct?
- Explore the performance of the Jukes-Cantor distance formula for recovering  $\alpha t$  on data that is produced by a model using Jukes-Cantor matrices but with an initial base distribution other than the Jukes-Cantor one. (Keep  $\alpha$  small, and use a large value of  $t$  for `mutate`.) How different can the initial base distribution be before the distance seems unreliable?
- Repeat the last item for the Kimura 2-parameter model.
- If data is produced according to some general Markov model, how close to a Kimura 2-parameter matrix must it be for the log-det distance and Kimura 2-parameter distance to be close?
- Even if data is produced according to the Jukes-Cantor model, do the Kimura distances and log-det distances give the same results as the Jukes-Cantor distance? Shouldn't they?
- Explain how you would decide which distance formula to use if you were given two sequences.