

BIOL 495S/ CS 490B/ MATH 490B/ STAT 490B

Spring 2002, Jan. 14 and 16 lectures

Probabilities and probabilistic models

Reading: S. M. Ross, "Introduction to probability models", 7th ed. Chapter 1.

Reference: R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, 1998 "Biological sequence analysis: Probabilistic models of proteins and nucleic acids", Section 1.3 and 3.1.

Probabilities

Let us consider a very simple example. A familiar probabilistic system with a set of discrete outcomes is the roll of a six-sided die. To define probabilities, we must first define the space of possible events. In this example, there are six events for a roll of a die, face 1-6. Following the annotations of the reading (Ross), $S = \{1,2,3,4,5,6\}$. A model of a roll of a die (possibly loaded) would have six parameters $p_1 \dots p_6$; the probability of rolling i is p_i . To be probabilities, the parameters p_i must satisfy the conditions that $p_i \geq 0$ and $\sum_{i=1}^6 p_i = 1$. For example, suppose that all six numbers are equally likely to appear (a fair die), then we will have

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$$

Another example closer to our biological subject matter is to define probabilities of amino acids or nucleotides. For instance, at a position of a protein sequence, we assume an amino acid a occurs at random with probability q_a . In other words, the amino acid at this position is possible to be one of the twenty types, each type a has probability q_a to occur. The probability q_a is a number between 0 and 1, and the sum of all twenty probabilities equals to 1.

Conditional probabilities and independency

Suppose that we toss two dice. There are totally 36 possible outcomes, combining the possible numbers of the first and second toss. Suppose that each of the 36 possible outcomes is equally likely to occur hence has probability $1/36$. If we observe that the first die is a four, then given this information, what is the probability that the sum of the two dice equals six? Given that the initial die is a four, it follows that there can be at most six possible outcomes of our experiment, namely, (4,1), (4,2), (4,3), (4,4), (4,5), and (4,6). Since each of these outcomes originally had the same probability of occurring, they should still have equal probabilities. That is, given that the first die is a four, then the (conditional) probability of each of the outcomes (4,1), (4,2), (4,3), (4,4), (4,5), and (4,6) is $1/6$, while the (conditional) probability of the other 30 points is 0. Hence, the desired probability will be $1/6$.

A conditional probability is the probability that one event will occur given that we already know that some other events have occurred. If we let E and F denote respectively

the event that the sum of the dice is six and the event that the first die is a four, then the probability just obtained is the conditional probability that E occurs given that F has occurred and is denoted by

$$P(E | F)$$

A general formula for $P(E | F)$, which is valid for all events E and F when $P(F) > 0$, is

$$P(E | F) = \frac{P(EF)}{P(F)} \quad (1)$$

where $P(EF)$ is the probability of the intersection of E and F , or the probability of both E and F occur. Back to the previous example, $P(EF)$ is the probability of the first die is a four and the sum of the two dice equals six and $P(F)$ is the probability the first die is a four. Therefore,

$$P(EF) = P(\text{two dice are } (4,2)) = \frac{1}{36}$$

$$P(F) = \frac{1}{6}$$

$$P(E | F) = \frac{1/36}{1/6} = \frac{1}{6}$$

By defining conditional probability, we can also write the probability of both E and F occurrence as

$$P(EF) = P(E | F)P(F)$$

Similarly, we have $P(EF) = P(F | E)P(E)$, if $P(E) > 0$.

Independency

Two events E and F are said to be independent if

$$P(EF) = P(E)P(F)$$

By Equation (1) this implies that E and F are independent if

$$P(E | F) = P(E)$$

That is, E and F are independent if knowledge that F has occurred does not affect the probability that E occurs. That is, the occurrence of E is independent of whether or not F occurs. Two events E and F that are not independent are said to be dependent.

As we will see in the followings, some models for nucleotide sequence (or protein sequence) assume a nucleotide (or amino acid) occurs independent of other residues in the sequence. Therefore, the probability of the sequence will be the product of the probabilities of residues in the sequence.

Probabilistic models of sequences

When we talk about a model normally we mean a system that simulates the object under consideration. A probabilistic model is one that produces different outcomes with different probabilities. A probabilistic model can therefore simulate a whole class of objects, assigning each an associated probability. In our case the objects will normally be sequences, and a model might describe a family of related sequences.

A model of a sequence of three consecutive rolls of a die might be that they were all independent, so that the probability of sequence [1,6,3] would be the product of the

individual probabilities, $p_1 p_6 p_3$. For a fair die, each number has probability 1/6 to occur in a roll. The probability of sequence [1,6,3] from three independent consecutive rolls is then 1/216.

Consider a second example of a simple model of any protein or DNA sequence. Biological sequences are strings from a finite alphabet of residues, generally either four nucleotides or twenty amino acids. Assume that a residue a occurs at random with probability q_a , independent of all other residues in the sequence. If the protein or DNA sequence is denoted $x_1 \dots x_n$, the probability of the whole sequence is then the product $q_{x_1} q_{x_2} \dots q_{x_n} = \prod_{i=1}^n q_{x_i}$. This is often used as ‘random sequence model’ of a base-level model, or null hypothesis, to compare other models against.

Maximum likelihood estimation

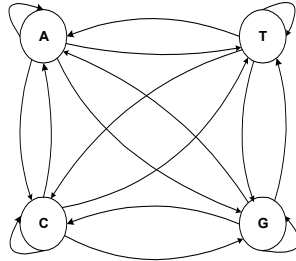
The parameters for a probabilistic model are typically estimated from large sets of trusted examples, often called a training set. For instance, the probability q_a for amino acid a can be estimated as the observed frequency of residues in a database of known protein sequences, such as SWISS-PROT. We obtain the twenty frequencies from counting up some twenty million individual residues in the database, and thus we have so much data that as long as the training sequences are not systematically biased towards a peculiar residue composition, we expect the frequencies to be reasonable estimates of the underlying probabilities of our model. This way of estimating models is called maximum likelihood estimation, because it can be shown that using the frequencies with which the amino acids occur in the database as the probabilities q_a maximizes the total probability of all the sequences given the model (the likelihood). In general, given a model with parameters θ and a set of data D , the maximum likelihood estimate for θ is that value which maximizes $P(D | \theta)$.

Markov chains

Markov chain is an excellent statistical tool for modeling a sequence of any type, because a Markov chain is simply a sequence of events that occur one after another. The main restriction on a Markov chain is that the probability assigned to an event at any location in the chain can depend on only a fixed number of previous events. For example, the identity of a DNA base might depend only on the previous base (a more general model compared to independent bases). Let us use a simple example to introduce Markov models.

In the human genome wherever the dinucleotide CG occurs (frequently written CpG to distinguish it from the C-G base pair across the two stands) the C nucleotide (cytosine) is typically chemically modified. There is a relatively high chance of this modification that mutates C into a T, with the consequence that in general CpG dinucleotides are rarer in the genome than would be expected from the independent probabilities of C and G. For biologically important reasons the mutation modification process is suppressed in short stretches of the genome, such as around the promoters or ‘start’ regions of many genes. In these regions we see many more CpG dinucleotides than elsewhere. Such regions are called CpG islands.

What sort of probabilistic model might we use for CpG island regions? We know that dinucleotides are important. We therefore want a model that generates sequences in which the probability of a symbol depends on the previous symbol. The simplest such model is a classical Markov chain. We can show a Markov chain graphically as a collection of ‘states’, each of which corresponds to a particular residue, with arrows between the states. A Markov chain for DNA can be drawn like this:



where there is a state for each of the four letters A, C, G, and T in the DNA alphabet. A probability parameter is associated with each arrow in the figure, which determines the probability of a certain residue following another residue, or one state following another state. These probability parameters are called the transition probabilities, which we will write p_{st} :

$$p_{st} = P(x_i = t \mid x_{i-1} = s),$$

where s, t indicate one of the four nucleotides.

Markov models for the CpG island example are illustrated below. From a set of human DNA sequences we extracted a total of 48 putative CpG islands and derived two Markov chain models, one for the regions labeled as CpG islands (the ‘+’ model) and the other from the remainder of the sequence (the ‘-’ model). The transition probabilities for each model were set using the equation

$$p_{st}^+ = \frac{c_{st}^+}{\sum_{t'} c_{st'}^+}$$

and its analogue for p_{st}^- , where c_{st}^+ is the number of times letter t followed letter s in the island regions. These are the maximum likelihood estimators for the transition probabilities. The resulting tables are

+	A	C	G	T	-	A	C	G	T
A	0.180	0.274	0.426	0.120	A	0.300	0.205	0.285	0.210
C	0.171	0.368	0.274	0.188	C	0.322	0.298	0.078	0.302
G	0.161	0.339	0.375	0.125	G	0.248	0.246	0.298	0.208
T	0.079	0.355	0.384	0.182	T	0.177	0.239	0.292	0.292

where the first row in each case contains the frequencies with which an A is followed by each of the four bases, and so on for the other rows, so each row sums to one. These numbers are not the same; for example, G following A is much more common than T following A. Notice also that the tables are asymmetric. In both tables the probability for G following C is lower than that for C following G. But the transition probabilities of $C \rightarrow G$ and $G \rightarrow C$ are much higher in the island model than those in the non-island model.

The different transition probabilities make the CpG island to be distinguished from other regions.

For any probability model of sequences we can write the probability of the sequence, denoted by $x = (x_1, x_2, \dots, x_{L-1}, x_L)$, as

$$\begin{aligned} P(x) &= P(x_1, x_2, \dots, x_{L-1}, x_L) \\ &= P(x_L | x_1, x_2, \dots, x_{L-1})P(x_{L-1} | x_1, x_2, \dots, x_{L-2}) \dots P(x_2 | x_1)P(x_1) \end{aligned} \quad (2)$$

by applying $P(EF) = P(F | E)P(E)$ many times. The key property of a Markov chain is that the probability of each symbol x_i depends only on the value of the preceding symbol x_{i-1} , not on the entire previous sequence, i.e.

$$P(x_i | x_1, \dots, x_{i-1}) = P(x_i | x_{i-1}) = p_{x_{i-1}x_i}$$

The equation (2) therefore becomes

$$\begin{aligned} P(x) &= P(x_1, x_2, \dots, x_{L-1}, x_L) \\ &= P(x_L | x_1, x_2, \dots, x_{L-1})P(x_{L-1} | x_1, x_2, \dots, x_{L-2}) \dots P(x_2 | x_1)P(x_1) \quad (3) \\ &= P(x_L | x_{L-1})P(x_{L-1} | x_{L-2}) \dots P(x_2 | x_1)P(x_1) \end{aligned}$$

Although we have derived this equation in the context of CpG islands in DNA sequences, it is in fact the general equation for the probability of a specific sequence from any Markov chain.

Equation (3) is used to calculate the likelihood of a sequence. For instance, we use (3) to obtain the likelihood values of a sequence under ‘+’ model and ‘-’ model respectively. If the likelihood value from ‘+’ model is much higher than that from ‘-’ model, we may conclude the sequence is a CpG island region (see exercise 5).

Exercises Due on Wednesday, Jan. 23rd.

1. In a DNA sequence set, there are totally 2×10^4 nucleotides, among which 6×10^3 are adenine, 4×10^3 are guanine, and 8×10^3 are thymine. Use this data set to estimate the probabilities for the four types of nucleotides.
2. A protein sequence is 500 amino acids length. Assume amino acids occur independently in the sequence. If we are given that the probability for leucine is $q_L = 0.04$, how many leucine residues we are expected to see in this sequence?
3. Prove Bayes’ theorem that

$$\begin{aligned} P(E | F) &= \frac{P(F | E)P(E)}{P(F)} \\ &= \frac{P(F | E)P(E)}{P(F | E)P(E) + P(F | E^c)P(E^c)} \end{aligned}$$

where all the conditional probabilities are assumed well defined.

4. Suppose that the chance of rain tomorrow depends on previous weather conditions only through whether or not it is raining today. Suppose also that if it rains today, then it will rain tomorrow with probability 0.6; and if it does not rain today, then it will rain tomorrow with probability 0.3. Show that the process is a two-state Markov chain. Calculate all the transition probabilities.

5. Consider a nucleotide sequence 'AGCGCG'.
 - (a) Use both the '+' Markov model and the '-' Markov model defined in the text to obtain two probabilities of this sequence. When applying Equation (3), we assume the start residue is equally likely to be each of the four nucleotides. The transition probabilities are from the table in the text.
 - (b) Compare the probabilities calculated in (a). What region we can predict for this sequence, CpG island or normal region?