1

Introduction

The reader will quickly discover that the organization of this book was chosen to be parallel to the organization of *Biological Sequence Analysis* by Durbin *et al.* (1998). The first chapter of *BSA* contains an introduction to the fundamental notions of biological sequence analysis: sequence similarity, homology, sequence alignment, and the basic concepts of probabilistic modeling.

Finding these distinct concepts described back-to-back is surprising at first glance. However, let us recall several important bioinformatics questions. How could we construct a pairwise sequence alignment? How could we build an alignment of multiple sequences? How could we create a phylogenetic tree for several biological sequences? How could we predict an RNA secondary structure? None of these questions can be consistently addressed without use of probabilistic methods. The mathematical complexity of these methods ranges from basic theorems and formulas to sophisticated architectures of hidden Markov models and stochastic grammars able to grasp fine compositional characteristics of empirical biological sequences.

The explosive growth of biological sequence data created an excellent opportunity for the meaningful application of discrete probabilistic models. Perhaps, without much exaggeration, the implications of this new development could be compared with implications of the revolutionary use of calculus and differential equations for solving problems of classic mechanics in the eighteenth century.

The problems considered in this introductory chapter are concerned with the fundamental concepts that play an important role in biological sequence analysis: the maximum likelihood and the maximum *a posteriori* (Bayesian) estimation of the model parameters. These concepts are crucial for understanding statistical inference from experimental data and are impossible to introduce without notions of conditional, joint, and marginal probabilities.

The frequently arising problem of model parameterization is inherently difficult if only a small training set is available. One may still attempt to use methods suitable for large training sets. But this move may result in overfitting and the generation of biased parameter estimates. Fortunately, this bias can be eliminated to some degree; the model can be generalized as the training set is augmented by artificially introduced observations, pseudocounts.

Problems included in this chapter are intended to provide practice with utilizing the notions of marginal and conditional probabilities, Bayes' theorem, maximum likelihood, and Bayesian parameter estimation. Necessary definitions of these notions and concepts frequently used in *BSA* can be found in undergraduate textbooks on probability and statistics (for example, Meyer (1970), Larson (1982), Hogg and Craig (1994), Casella and Berger (2001), and Hogg and Tanis (2005)).

1.1 Original problems

Problem 1.1 Consider an occasionally dishonest casino that uses two kinds of dice. Of the dice 99% are fair but 1% are loaded so that a six comes up 50% of the time. We pick up a die from a table at random. What are $P(\text{six}|D_{\text{loaded}})$ and $P(\text{six}|D_{\text{fair}})$? What are $P(\text{six}, D_{\text{loaded}})$ and $P(\text{six}, D_{\text{fair}})$? What is the probability of rolling a six from the die we picked up?

Solution All possible outcomes of a fair die roll are equally likely, i.e. $P(\text{six}|D_{\text{fair}}) = 1/6$. On the other hand, the probability of rolling a six from the loaded die, $P(\text{six}|D_{\text{loaded}})$, is equal to $1/2$. To compute the probability of the combined event $(\text{six}, D_{\text{loaded}})$, rolling a six and picking up a loaded die, we use the definition of conditional probability:

$$P(\text{six}, D_{\text{loaded}}) = P(D_{\text{loaded}})P(\text{six}|D_{\text{loaded}}). \tag{1.1}$$

As the probability of picking up a loaded die is $1/100$, Equality (1.1) yields

$$P(\text{six}, D_{\text{loaded}}) = \frac{1}{100} \times \frac{1}{2} = \frac{1}{200}.$$

By a similar argument,

$$P(\text{six}, D_{\text{fair}}) = P(\text{six}|D_{\text{fair}})P(D_{\text{fair}}) = \frac{1}{6} \times \frac{99}{100} = \frac{33}{200}.$$

The probability of rolling a six from the die picked up at random is computed as the total probability of event "six" occurring in combination either with event D_{loaded} or with event D_{fair}:

$$P(\text{six}) = P(\text{six}, D_{\text{loaded}}) + P(\text{six}, D_{\text{fair}}) = \frac{34}{200} = \frac{17}{100}. \qquad \square$$

Problem 1.2 How many sixes in a row would we need to see in Problem 1.1 before it is more likely that we had picked a loaded die?

Solution Bayes' theorem is all we need to determine the conditional probability of picking up a loaded die, $P(D_{\text{loaded}}|n \text{ sixes})$, given that n sixes in a row have been rolled:

$$P(D_{\text{loaded}}|n \text{ sixes}) = \frac{P(n \text{ sixes}|D_{\text{loaded}})P(D_{\text{loaded}})}{P(n \text{ sixes})}$$
$$= \frac{P(n \text{ sixes}|D_{\text{loaded}})P(D_{\text{loaded}})}{P(n \text{ sixes}|D_{\text{loaded}})P(D_{\text{loaded}}) + P(n \text{ sixes}|D_{\text{fair}})P(D_{\text{fair}})}.$$

Rolls of both fair or loaded dice are independent, therefore

$$P(D_{\text{loaded}}|n \text{ sixes}) = \frac{(1/100) \times (1/2)^n}{(99/100) \times (1/6)^n + (1/100) \times (1/2)^n} = \frac{1}{11 \times (1/3)^{n-2} + 1}.$$

This result indicates that $P(D_{\text{loaded}}|n \text{ sixes})$ approaches one as n, the length of the observed run of sixes, increases. The inequality

$$P(D_{\text{loaded}}|n \text{ sixes}) > 1/2$$

tells us that it is more likely that a loaded die was picked up. This inequality holds if

$$\left(\frac{1}{3}\right)^{n-2} < \frac{1}{11}, \quad n \geq 5.$$

Therefore, seeing five or more sixes in a row indicates that it is more likely that the loaded die was picked up. □

Problem 1.3 Use the definition of conditional probability to prove Bayes' theorem,

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}.$$

Solution For any two events X and Y such that $P(Y) > 0$ the conditional probability of X given Y is defined as

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}.$$

Applying this definition once again to substitute $P(X \cap Y)$ by $P(X)P(Y|X)$, we arrive at the equation which is equivalent to Bayes' theorem:

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}.$$

□

Problem 1.4 A rare genetic disease is discovered. Although only one in a million people carry it, you consider getting screened. You are told that the genetic test is extremely good; it is 100% sensitive (it is always correct if you have the disease) and 99.99% specific (it gives a false positive result only 0.01% of the time). Using Bayes' theorem, explain why you might decide not to take the test.

Solution Before taking the test, the probability $P(D)$ that you have the genetic disease is 10^{-6} and the probability $P(H)$ that you do not is $1-10^{-6}$. By how much will the test change this uncertainty? Let us consider two possible outcomes.

If the test is positive, then the Bayesian posterior probabilities of having and not having the disease are as follows:

$$\begin{aligned} P(D|\text{positive}) &= \frac{P(\text{positive}|D)P(D)}{P(\text{positive})} \\ &= \frac{P(\text{positive}|D)P(D)}{P(\text{positive}|D)P(D)+P(\text{positive}|H)P(H)} \\ &= \frac{10^{-6}}{10^{-6}+0.999999\times 10^{-4}} = 0.0099, \\ P(H|\text{positive}) &= \frac{P(\text{positive}|H)P(H)}{P(\text{positive})} = 0.9901. \end{aligned}$$

If the test is negative, the Bayesian posterior probabilities become

$$\begin{aligned} P(D|\text{negative}) &= \frac{P(\text{negative}|D)P(D)}{P(\text{negative})} \\ &= \frac{P(\text{negative}|D)P(D)}{P(\text{negative}|D)P(D)+P(\text{negative}|H)P(H)} \\ &= \frac{0}{0+0.9999\times(1-10^{-6})} = 0, \\ P(H|\text{negative}) &= \frac{P(\text{negative}|H)P(H)}{P(\text{negative})} = 1. \end{aligned}$$

Thus, the changes of prior probabilities $P(D)$, $P(H)$ are very small:

$$|P(D)-P(D|\text{positive})| = 0.0099, \quad |P(D)-P(D|\text{negative})| = 10^{-6},$$
$$|P(H)-P(H|\text{positive})| = 0.0099, \quad |P(H)-P(H|\text{negative})| = 10^{-6}.$$

We see that even if the test is positive the probability of having the disease changes from 10^{-6} to 10^{-2}. Thus, taking the test is not worthwhile for practical reasons. □

Problem 1.5 We have to examine a die which is expected to be loaded in some way. We roll a die ten times and observe outcomes of 1, 3, 4, 2, 4, 6, 2, 1, 2, and 2. What is our maximum likelihood estimate for p_2, the probability of rolling a two? What is the Bayesian estimate if we add one pseudocount per category? What if we add five pseudocounts per category?

Solution The maximum likelihood estimate for p_2 is the (relative) frequency of outcome "two," thus $\hat{p}_2 = 4/10 = 2/5$. If one pseudocount per category is added, the Bayesian estimate is $\hat{p}_2 = 5/16$. If we add five pseudocounts per category, then $\hat{p}_2 = 9/40$. In the last case the Bayesian estimate $\hat{p}_2$ is closer to the probability of the event "two" upon rolling a fair die, $p_2 = 1/6$.

In any case, it is difficult to assess the validity of these alternative approaches without additional information. The best way to improve the estimate is to collect more data. □

1.2 Additional problems

The following problems motivated by questions arising in biological sequence analysis require the ability to apply formulas from combinatorics (Problems 1.6, 1.7, 1.9, and 1.10), elementary calculation of probabilities (Problems 1.8 and 1.16), as well as a knowledge of properties of random variables (Problems 1.13 and 1.18). Our goal here is to help the reader recognize the probabilistic nature of these (and similar) problems about biological sequences.

Basic probability distributions are used in this section to describe the properties of DNA sequences: a geometric distribution to describe the length distribution of restriction fragments (Problem 1.12) and open reading frames (Problem 1.14); a Poisson distribution as a good approximation for the number of occurrences of oligonucleotides in DNA sequences (Problems 1.11, 1.17, 1.19, and 1.22). We will use the notion of an "independence model" for a sequence of independent identically distributed (i.i.d.) random variables with values from a finite alphabet $\mathcal{A}$ (i.e. the alphabet of nucleotides or amino acids) such that the probability of occurrence of symbol a at any sequence site is equal to q_a, $\sum_{a\in\mathcal{A}} q_a = 1$. Thus, a DNA or protein sequence fragment $x_1, \ldots, x_n$ generated by the independence model has probability $\prod_{i=1}^{n} q_{x_i}$. Note that the same model is called the random sequence model in the BSA text (Durbin *et al.,* 1998). The independence model is used to describe DNA sequences in Problems 1.12, 1.14, 1.16, and 1.17.

The introductory level of Chapter 1 still allows us to deal with the notion of hypotheses testing. In Problem 1.20 such a test helps to identify *CpG*-islands in

a DNA sequence, while in Problem 1.21 we consider the test for discrimination between DNA sequence regions with higher and lower $G + C$ content.

Finally, issues of the probabilistic model comparison are considered in Problems 1.16, 1.18, and 1.19.

Problem 1.6 In the herpesvirus genome, nucleotides C, G, A, and T occur with frequencies 35/100, 35/100, 15/100, and 15/100, respectively. Assuming the independence model for the genome, what is the probability that a randomly selected 15 nt long DNA fragment contains eight C's or G's and seven A's or T's?

Solution The probability of there being eight C's or G's and seven A's or T's in a 15 nt fragment, given the frequencies 0.7 and 0.3 for each group C & G and A & T, respectively, is $0.7^8 \times 0.3^7 = 0.0000126$. This number must be multiplied by $\binom{15}{8} = 15!/8!7!$, the number of possible arrangements of representatives of these nucleotide groups among fifteen nucleotide positions. Thus, we get the probability 0.08. □

Problem 1.7 A DNA primer used in the polymerase chain reaction is a one-strand DNA fragment designed to bind (to hybridize) to one of the strands of a target DNA molecule. It was observed that primers can hybridize not only to their perfect complements, but also to DNA fragments of the same length having one or two mismatching nucleotides. If the genomic DNA is "sufficiently long," how many different DNA sequences may bind to an eight nucleotide long primer? The notion of "sufficient length" implies that all possible oligonucleotides of length 8 are present in the target genomic DNA.

Solution We consider a more general situation with the length of primer equal to n. There are three possible cases of hybridization between the primer and the DNA: with no mismatch, with one mismatch, and with two mismatches. The first case obviously identifies only one DNA sequence exactly complementary to the primer. The second case, one mismatch, with the freedom to choose one of three mismatching types of nucleotides in one position of the complementary sequence, gives $3n$ possible sequences. Finally, two positions carrying mismatching nucleotides can occur in $n(n-1)/2$ ways. Each choice of these two positions generates nine possibilities to choose two nucleotides different from the matching types. This gives a total of $9n(n-1)/2$ possible sequences with two mismatches. Hence, for $n = 8$, there are

$$1 + 3 \times 8 + \frac{9 \times 8 \times 7}{2} = 277$$

different sequences able to hybridize to the given primer. □

Problem 1.8 A DNA sequencing reaction is performed with an error rate of 10%, thus a given nucleotide is wrongly identified with probability 0.1. To minimize the error rate, DNA is sequenced by $n = 3$ independent reactions, the newly sequenced fragments are aligned, and the nucleotides are identified by the following majority rule. The type of nucleotide at a particular position is identified as α, $\alpha \in \{T, C, A, G\}$, if more nucleotides of type α are aligned in this position than all other types combined. If at an alignment position no nucleotide type appears more than $n/2$ times, the type of nucleotide is not identified (type N).

What is the expected percentage of (a) correctly and (b) incorrectly identified nucleotides? (c) What is the probability that at a particular site identification is impossible? (d) How does the result of (a) change if $n = 5$; what about for $n = 7$? Assume that there are only substitution type errors (no insertions or deletions) with no bias to a particular nucleotide type.

Solution (a) In a given position, we consider the three sequencing reaction calls as outcomes of the three Bernoulli trials with "success" taking place if the nucleotide is identified correctly (with probability $p = 0.9$) and "failure" otherwise (with probability $q = 0.1$). Then the probabilities of the following events are described by the binomial distribution and can be determined immediately:

$$P_3 = P(\text{"success" is observed three times}) = p^3 = 0.9^3 = 0.729,$$

$$P_2 = P(\text{"success" is observed twice}) = \binom{3}{2} p^2 q$$

$$= 3 \times 0.9^2 \times 0.1 = 0.243.$$

Under the majority rule, the expected percentage $\mathbf{E}$ of correctly identified nucleotides is given by

$$\mathbf{E}^c_{n=3} = P(\text{"success" is observed at least twice}) \times 100\%$$
$$= (P_3 + P_2) \times 100\% = 97.2\%.$$

(b) To determine the probability of identifying a nucleotide at a given site incorrectly, we have to be able to classify the "failure" outcomes; thus, we need to generalize the binomial distribution to a multinomial one. Specifically, in each independent trial (carried out at a given sequence site) we can have "success" (with probability $p = 0.9$) and three other outcomes: "failure 1," "failure 2," and "failure 3" (with equal probabilities $q_1 = q_2 = q_3 = 1/30$). To identify a nucleotide incorrectly would mean to observe at least two "failure i" outcomes, $i = 1, 2, 3$,

among $n = 3$ trials. Therefore,

$$P'_3 = (\text{“failure } i\text{” is observed three times}) = q_i^3 = (1/30)^3 = 0.000037,$$

$$P'_2 = P(\text{“failure } i\text{” is observed twice}) = 2\binom{3}{2} q_i^2 q_j + \binom{3}{2} q_i^2 p$$

$$= 6 \times (1/30)^3 + 3 \times (1/30)^2 \times 0.9 = 0.00356.$$

Finally, for the expected percentage of wrongly identified nucleotides we have

$$\mathbf{E}^w_{n=3} = \left(\sum_{i=1,2,3} (P'_3 + P'_2) \right) \times 100\%$$

$$= 3(P'_3 + P'_2) \times 100\% = 1.1\%.$$

(c) At a particular site, the base calling results in three mutually exclusive events: “correct identification,” “incorrect identification,” or “identification impossible.” Then, the probability of the last outcome is given by

$$P(\text{nucleotide cannot be identified}) = 1 - (P_3 + P_2) - 3(P'_3 + P'_2) = 0.0172.$$

(d) To calculate the expected percentage $\mathbf{E}^c_n$ of correctly identified nucleotides for $n = 5$ and $n = 7$, we apply the same arguments as in section (a), only instead of three Bernoulli trials we consider five and seven, respectively. We find:

$$\mathbf{E}^c_{n=5} = P(\text{at least three “successes” among five trials}) \times 100\%$$

$$= p^5 + 5 \times 0.9^4 \times 0.1 + 10 \times 0.9^3 0.1^2 = 99.14\%.$$

Similarly,

$$\mathbf{E}^c_{n=7} = P(\text{at least four “successes” among seven trials}) \times 100\% = 99.73\%.$$

As expected, the increase in the number of independent reactions improves the quality of sequencing. □

Problem 1.9 Due to redundancy of genetic code, a sequence of amino acids could be encoded by several DNA sequences. For a given ten amino acid long protein fragment, what are the lower and upper bounds for the number of possible DNA sequences that could carry code for this protein fragment?

Solution The lower bound of one would be reached if all ten amino acids are methionine or tryptophan, the amino acids encoded by a single codon. In this case the amino acid sequence uniquely defines the underlying nucleotide sequence. The

Table 1.1. *The maximum number I_α of nucleotides C and G that appear in one of the synonomous codons for given amino acid α*

I_α	Amino acid α
1	Asn, Ile, Lys, **Met**, Phe, Tyr
2	Asp, Cys, Gln, Glu, His, Leu, Ser, Thr, **Trp**, Val
3	Ala, Arg, Gly, Pro

upper bound would be reached if the amino acid sequence consists of leucine, arginine, or serine, the amino acids encoded by six codons each. A ten amino acid long sequence consisting of any arrangement of *Leu*, *Ser*, or *Arg* can be encoded by as many as $6^{10} = 60\,466\,176$ different nucleotide sequences. □

Problem 1.10 Life forms from planet XYZ were discovered to have a DNA and protein basis with proteins consisting of twenty amino acids. By analysis of the protein composition, it was determined that the average frequencies of all amino acids excluding *Met* and *Trp* were equal to 1/19, while the frequencies of *Met* and *Trp* were equal to 1/38. Given the high temperature on the XYZ surface, it was speculated that the DNA has an extremely high $G + C$ content. What could be the highest average $G + C$ content of protein-coding regions (given the average amino acid composition as stated above) if the standard (the same as on planet Earth) genetic code is used to encode *XYZ* proteins?

Solution To make the highest possible $G + C$ content of protein-coding region that would satisfy the restrictions on amino acid composition, synonymous codons with highest $G + C$ content should be used on all occasions. The distribution of the high $G + C$ content codons according to the standard genetic code is as shown in Table 1.1 (where I_α designates the highest number of C and G nucleotides in a codon encoding amino acid α).
Therefore, the average value of the $G + C$ content of a protein-coding region is given by

$$\langle G + C \rangle = \sum_\alpha \frac{I_\alpha}{3} \mathrm{f}_\alpha$$
$$= \frac{1}{3}\left(\frac{1}{19}(5 \times 1 + 9 \times 2 + 4 \times 3) + \frac{1}{38}(1 + 2)\right) = 0.64.$$

Here f_α is the frequency of amino acid α.

Remark Similar considerations can provide estimates of upper and lower bounds of $G + C$ content for prokaryotic genomes (planet Earth), where protein-coding regions typically occupy about 90% of total DNA length. □

Problem 1.11 A restriction enzyme is cutting DNA at a palindromic site 6 nt long. Determine the probability that a circular chromosome, a double-stranded DNA molecule of length $L = 84\,000$ nt, will be cut by the restriction enzyme into exactly twenty fragments. It is assumed that the DNA sequence is described by the independence model with equal probabilities of nucleotides T, C, A, and G. Hint: use the Poisson distribution.

Solution The probability that a restriction site starts in any given position of the DNA sequence is $p = (1/4)^6 = 0.0002441$. If we do not take into account the mutual dependence of occurrences of restriction sites in positions i and j, $|i-j| \leq 6$, the number X of the restriction sites in the DNA sequence can be considered as the number of successes (with probability p) in a sequence of L Bernoulli trials; therefore, X has a binomial distribution with parameters p and L. Since L is large and p is small, we can use the Poisson distribution with parameter $\lambda = pL = 20.5$ as an approximation of the binomial distribution. Then

$$P(X = 20) = e^{-\lambda}\frac{\lambda^{20}}{20!} = 0.088.$$

Notably, the probability of cutting this DNA sequence into any other particular number of fragments will be lower than $P(X = 20)$. Indeed, the ratio R_k of probabilities of two consecutive values of X,

$$R_k = \frac{P(X = k+1)}{P(X = k)} = \frac{\lambda}{k+1},$$

shows that $P(X = k)$ increases as k grows from 0 to λ, and decreases as k grows from λ to L, thus attaining its maximum value at point $k = \lambda$. In other words, if λ is not an integer, the most probable value of the Poisson distributed random variable is equal to $[\lambda]$, where $[\lambda]$ stands for the largest integer not greater than λ. Otherwise, the most probable values are both $\lambda - 1$ and λ. □

Problem 1.12 Determine the average length of the restriction fragments produced by the six-cutter restriction enzyme *SmaI* with the restriction site *CCCGGG*. Consider (a) a genome with a $G + C$ content of 70% and (b) a genome with a $G + C$ content of 30%. It is assumed that the genomic sequence can be represented by the independence model with probabilities of nucleotides such that $q_G = q_C$, $q_A = q_T$. Note that enzyme *SmaI* cuts the double strand of DNA in the middle of site *CCCGGG*.