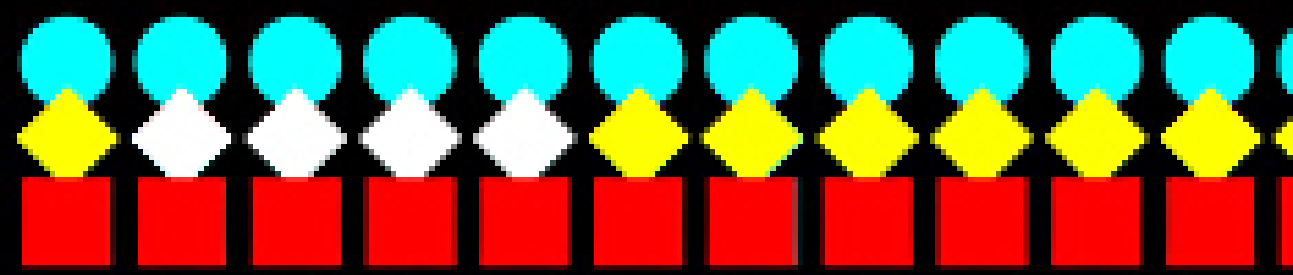
Problems and Solutions in Biological Sequence Analysis



Mark Borodovsky
and Svetlana Ekisheva

CAMBRIDGE www.cambridge.org/9780521847544

This page intentionally left blank

PROBLEMS AND SOLUTIONS IN BIOLOGICAL SEQUENCE ANALYSIS

This book is the first of its kind to provide a large collection of bioinformatics problems with accompanying solutions. Notably, the problem set includes all of the problems offered in *Biological Sequence Analysis* (*BSA*), by Durbin *et al.*, widely adopted as a required text for bioinformatics courses at leading universities worldwide. Although many of the problems included in *BSA* as exercises for its readers have been repeatedly used for homework and tests, no detailed solutions for the problems were available. Bioinformatics instructors had therefore frequently expressed a need for fully worked solutions and a larger set of problems for use in courses.

This book provides just that: following the same structure as *BSA*, and significantly extending the set of workable problems, it will facilitate a better understanding of the contents of the chapters in *BSA* and will help its readers develop problem solving skills that are vitally important for conducting successful research in the growing field of bioinformatics. All of the material has been class-tested by the authors at Georgia Tech, where the first ever M.Sc. degree program in Bioinformatics was held.

MARK BORODOVSKY is the Regents' Professor of Biology and Biomedical Engineering and Director of the Center for Bioinformatics and Computational Biology at Georgia Institute of Technology in Atlanta. He is the founder of the Georgia Tech M.Sc. and Ph.D. degree programs in Bioinformatics. His research interests are in bioinformatics and systems biology. He has taught Bioinformatics courses since 1994.

SVETLANA EKISHEVA is a research scientist at the School of Biology, Georgia Institute of Technology, Atlanta. Her research interests are in bioinformatics, applied statistics, and stochastic processes. Her expertise includes teaching probability theory and statistics at universities in Russia and in the USA.

PROBLEMS AND SOLUTIONS IN BIOLOGICAL SEQUENCE ANALYSIS

MARK BORODOVSKY AND SVETLANA EKISHEVA

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York
www.cambridge.org
Information on this title: www.cambridge.org/9780521847544

First published in print format 2006

ISBN-13 978-0-511-33190-9 eBook (Adobe Reader)
ISBN-10 0-511-33190-8 eBook (Adobe Reader)

ISBN-13 978-0-521-84754-4 hardback
ISBN-10 0-521-84754-0 hardback

ISBN-13 978-0-521-61230-2 paperback
ISBN-10 0-521-61230-6 paperback

M. B.:
To Richard and Judy Lincoff

S. E.:
To Sergey and Natasha

Contents

Preface

Bioinformatics, an integral part of post-genomic biology, creates principles and ideas for computational analysis of biological sequences. These ideas facilitate the conversion of the flood of sequence data unleashed by the recent information explosion in biology into a continuous stream of discoveries. Not surprisingly, the new biology of the twenty-first century has attracted the interest of many talented university graduates with various backgrounds. Teaching bioinformatics to such a diverse audience presents a well-known challenge. The approach requiring students to advance their knowledge of computer programming and statistics prior to taking a comprehensive core course in bioinformatics has been accepted by many universities, including the Georgia Institute of Technology, Atlanta, USA.

In 1998, at the start of our graduate program, we selected the then recently published book *Biological Sequence Analysis* (*BSA*) by Richard Durbin, Anders Krogh, Sean R. Eddy, and Graeme Mitchison as a text for the core course in bioinformatics. Through the years, *BSA*, which describes the ideas of the major bioinformatic algorithms in a remarkably concise and consistent manner, has been widely adopted as a required text for bioinformatics courses at leading universities around the globe. Many problems included in *BSA* as exercises for its readers have been repeatedly used for homeworks and tests. However, the detailed solutions to these problems have not been available. The absence of such a resource was noticed by students and teachers alike.

The goal of this book, *Problems and Solutions in Biological Sequence Analysis* is to close this gap, extend the set of workable problems, and help its readers develop problem-solving skills that are vitally important for conducting successful research in the growing field of bioinformatics. We hope that this book will facilitate understanding of the content of the *BSA* chapters and also will provide an additional perspective for in-depth *BSA* reading by those who might not be able to take a formal bioinformatics course. We have augmented the set of original *BSA* problems with many new problems, primarily those that were offered to the Georgia Tech graduate students.

Probabilistic modeling and statistical analysis are frequently used in bioinformatics research. The mainstream bioinformatics algorithms, those for pairwise and multiple sequence alignment, gene finding, detecting orthologs, and building phylogenetic trees, would not work without rational model selection, parameter estimation, properly justified scoring systems, and assessment of statistical significance. These and many other elements of efficient bioinformatic tools require one to take into account the random nature of DNA and protein sequences.

As it has been illustrated by the *BSA* authors, probabilistic modeling laid the foundation for the development of powerful methods and algorithms for biological sequence interpretation and the revelation of its functional meaning and evolutionary connections. Notably, probabilistic modeling is a generalization of strictly deterministic modeling, which has a remarkable tradition in natural science. This tradition could be traced back to the explanation of astronomic observations on the motion of solar system planets by Isaac Newton, who suggested a concise model combining the newly discovered law of gravity and the laws of dynamics.

The maximum likelihood principle of statistics, notwithstanding the fashion of its traditional application, also has its roots in "deterministic" science that suggests that the chosen structure and parameters of a theoretical model should provide the best match of predictions to experimental observations. For instance, one could recognize the maximum likelihood approach in Francis Crick and James Watson's inference of the DNA double helix model, chosen from the combinatorial number of biochemically viable alternatives as the best fit to the X-ray data on DNA three-dimensional structure and other experimental data available.

In studying the processes of inheritance and molecular evolution, where random factors play important roles, fully fledged probabilistic models enter the picture. A classic cycle of experiments, data analysis, and modeling with search for a best fit of the models to data was designed and implemented by Gregor Mendel. His remarkable long term research endeavor provided proof of the existence of discrete units of inheritance, the genes.

When we deal with data coming from a less controllable environment, such as data on natural biological evolution spanning time periods on a scale of millions of years, the problem is even more challenging. Still, the situation is hopeful. The models of molecular evolution proposed by Dayhoff and co-authors, Jukes and Cantor, and Kimura, are classical examples of fundamental advances in modeling of the complex processes of DNA and protein evolution. Notably these models focus on only a single site of a molecular sequence and require the further simplifying assumption that evolution of sequence sites occurs independently from each other. Nevertheless, such models are useful starting points for understanding the

function and evolution of biological sequences as well as for designing algorithms elucidating these functional and evolutionary connections.

For instance, amino acid substitution scores are critically important parameters of the optimal global (Needleman and Wunsch) and local (Smith and Waterman) sequence alignment algorithms. Biologically sensible derivation of the substitution scores is impossible without models of protein evolution.

In the mid 1990s the notion of the hidden Markov model (HMM), having been of great practical use in speech recognition, was introduced to bioinformatics and quickly entered the mainstream of the modeling techniques in biological sequence analysis.

Theoretical advances that have occurred since the mid 1990s have shown that the sequence alignment problem has a natural probabilistic interpretation in terms of hidden Markov models. In particular, the dynamic programming (DP) algorithm for pairwise and multiple sequence alignment has the HMM-based algorithmic equivalent, the Viterbi algorithm. If the type of probabilistic model for a biological sequence has been chosen, parameters of the model could be inferred by statistical (machine learning) methods. Two competitive models could be compared to identify the one with the best fit.

The events and selective forces of the past, moving the evolution of biological species, have to be reconstructed from the current biological sequence data containing significant noise caused by all the changes that have occurred in the lifetime of disappeared generations. This difficulty can be overcome to some extent by the use of the general concept of self-consistent models with parameters adjusted iteratively to fit the growing collection of sequence data. Subsequently, implementation of this concept requires the expectation–maximization type algorithms able to estimate the model parameters simultaneously with rearranging data to produce the data structure (such as a multiple alignment) that fits the model better. *BSA* describes several algorithms of expectation–maximization type, including the self-training algorithm for a profile HMM and the self-training algorithm for a phylogenetic HMM. Given that the practice with many algorithms described in *BSA* requires significant computer programming, one may expect that describing the solutions would lead us into heavy computer codes, thus moving far away from the initial concepts and ideas. However, the majority of the *BSA* exercises have analytical solutions. On several occasions we have illustrated the implementations of the algorithms by "toy" examples. The computer codes written in C++ and Perl languages for such examples are available at opal.biology.gatech.edu/PSBSA. Note, that in the "Further reading" sections we include mostly papers that were published later than 1998, the year of *BSA* publication. Finally, we should mention that the references in the text to the pages in the BSA book cite the 2006 edition.

Acknowledgements

We thank Sergey Latkin, Svetlana's husband, for the remarkable help with preparation of LaTex figures and tables. We are grateful to Alexandre Lomsadze, Ryan Mills, Yuan Tian, Burcu Bakir, Jittima Piriyapongsa, Vardges Ter-Hovhannisyan, Wenhan Zhu, Jeffrey Yunes, and Matthew Berginski for invaluable technical assistance in preparation of the book materials; to Soojin Yi, and Galina Glazko for useful references on molecular evolution; to Michael Roytberg for helpful discussions on transformational grammars and finite automata. We cordially thank our editor Katrina Halliday for tremendous patience and constant support, without which this book would never have come to fruition. We are especially grateful to Richard Durbin, Anders Krogh, Sean R. Eddy, and Graeme Mitchison, for encouragement, helpful criticism and suggestions. Further, it is our pleasure to acknowledge firm support from the Georgia Tech School of Biology and the Wallace H. Coulter Department of Biomedical Engineering at Georgia Tech and Emory University. Finally, we wish to express our particular gratitude to our families for great patience and constant understanding.

M.B. and S.E.

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}:

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

□

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

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

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*.

Solution We denote the probability that the restriction site starts in a particular sequence position as P and the length of a restriction fragment as L. We associate the number 1 with a sequence position where the restriction site starts and the number 0 otherwise. Then in the generated sequence of ones and zeros the lengths of runs of zeros (equal to the lengths of restriction fragments) can be considered as values of random variable L. If we do not take into account the mutual dependence of occurrences of restriction sites at positions i and j, $|i-j| \leq 6$, the random variable L has the geometric distribution: $P(L = n) = (1 - P)^{n-1}P$. The expected value of L is defined by

$$\mathbf{E}L = \sum_{n=1}^{+\infty} n(1-P)^{n-1}P = P\sum_{n=1}^{+\infty} -\frac{d(1-P)^n}{dP}$$
$$= -P\frac{d\left(\sum_{n=1}^{+\infty}(1-P)^n\right)}{dP} = \frac{P}{P^2} = \frac{1}{P}.$$

For (a) we have

$$P_a = P(CCCGGG) = (0.35)^6 = 1.8 \times 10^{-3},$$

and the average length of restriction fragment is $\mathbf{E}L_a = 1/P_a = 544$ nt.

Similarly, for (b),

$$P_b = P(CCCGGG) = (0.15)^6 = 1.14 \times 10^{-5},$$

and the average length of the restriction fragment is $\mathbf{E}L_b = 87\,788$ nt. The longer average length of restriction fragments in (b) could be expected as the $G + C$-rich restriction site *CCCGGG* would appear less frequently in the $A + T$-rich genomic DNA. □

Problem 1.13 Consider a DNA sequence of length n described by the independence model with equal probabilities of nucleotides. Let X be the number of occurrences of dinucleotide *AA* and Y be the number of occurrences of dinucleotide *AT* in this sequence. What are the expected values and variances of random variables X and Y? For simplicity consider a circular DNA of length n.

Solution Let us define random variables x_i and y_i, $i = 1, \ldots, n$, as follows:

$$x_i = \begin{cases} 1, & \text{if dinucleotide } AA \text{ starts in } i\text{th position in the sequence,} \\ 0, & \text{otherwise;} \end{cases}$$

$$y_i = \begin{cases} 1, & \text{if dinucleotide } AT \text{ starts in } i\text{th position in the sequence,} \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, $X = \sum_{i=1}^{n} x_i$, $Y = \sum_{i=1}^{n} y_i$. The expected values of x_i and y_i, $i = 1, \ldots, n$, under the uniform independence model are given by

$$\mathbf{E}x_i = \mathbf{E}y_i = P(x_i = 1) = P(y_i = 1) = \left(\frac{1}{4}\right)^2 = \frac{1}{16}.$$

Thus, the mean value of X, Y is $\mathbf{E}X = \mathbf{E}Y = n/16$. Similarly, we can state that the expected number of occurrences of any other dinucleotide in the sequence is also $n/16$.

We denote the shortest distance between positions i and j in the circular DNA as $r(i,j)$ and find the second moment of X:

$$\mathbf{E}X^2 = \mathbf{E}\left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n} \mathbf{E}x_i^2 + \sum_{i,j:r(i,j)\geq 2} \mathbf{E}x_i x_j + \sum_{i,j:r(i,j)=1} \mathbf{E}x_i x_j. \tag{1.2}$$

As $x_i^2 = x_i$, the first sum in (1.2) is $\sum_{i=1}^{n} \mathbf{E}x_i^2 = n\mathbf{E}x_i = n/16$.

If the distance $r(i,j) \geq 2$, the random variables x_i and x_j are independent and

$$\sum_{i,j:r(i,j)\geq 2} \mathbf{E}x_i x_j = \sum_{i,j:r(i,j)\geq 2} \mathbf{E}x_i \mathbf{E}x_j = \frac{n(n-3)}{256}.$$

If $r(i,j) = 1$, then positions i and j are adjacent and, for certainty, we assume that position i precedes j. Then product $x_i x_j$ takes the following values:

$$x_i x_j = \begin{cases} 1, & \text{if triplet } AAA \text{ starts in position } i, \\ 0, & \text{otherwise,} \end{cases}$$

and $\mathbf{E}x_i x_j = P(x_i x_j = 1) = (1/4)^3 = 1/64$. Therefore, the second moment of X becomes

$$\mathbf{E}X^2 = \frac{n}{16} + \frac{n(n-3)}{256} + \frac{2n}{64} = \frac{n(n+21)}{256},$$

and the variance of X is given by

$$\mathbf{Var}X = \mathbf{E}X^2 - (\mathbf{E}X)^2 = \frac{n(n+21)}{256} - \frac{n^2}{256} = \frac{21n}{256}.$$

Similarly, for the second moment of Y we have:

$$\mathbf{E}Y^2 = \mathbf{E}\left(\sum_{i=1}^{n} y_i\right)^2 = \sum_{i=1}^{n} \mathbf{E}y_i^2 + \sum_{i,j:r(i,j)\geq 2} \mathbf{E}y_i y_j + \sum_{i,j:r(i,j)=1} \mathbf{E}y_i y_j. \tag{1.3}$$

The first two sums in (1.3) are the same as in Equation (1.2). However, if $r(i,j) = 1$, the product $y_i y_j$ is always zero, because dinucleotide AT cannot start in two adjacent positions i and j of the sequence. Therefore,

$$\mathbf{E}Y^2 = \frac{n}{16} + \frac{n(n-3)}{256} = \frac{n(n+13)}{256}$$

and

$$\mathbf{Var}Y = \mathbf{E}Y^2 - (\mathbf{E}Y)^2 = \frac{n(n+13)}{256} - \frac{n^2}{256} = \frac{13n}{256}.$$

We see that the variance of the number of occurrences of a dinucleotide depends on its structure: if it consists of different letters (thus, the dinucleotide cannot overlap with the neighbor of the same type), the variance is $13n/256$; if dinucleotide consists of the same letter repeated twice (and can overlap with the neighbor of the same type), the variance increases to $21n/256$.

Remark For an extended discussion of the first and second moments of frequencies of words in biological sequences, see Pevzner, Borodovsky, and Mironov (1989).

□

Problem 1.14 A prokaryotic protein-coding gene normally consists of an uninterrupted sequence of nucleotide triplets, codons. This sequence starts with a specific start codon (*ATG* is most frequent) and ends with one of the three stop codons: *TAA*, *TAG*, *TGA*. A sequence with such a structure is called an "open reading frame' (ORF). However, not every ORF found in prokaryotic genomic DNA is a functional gene. Assuming that *ATG* is the only possible start codon, what is the length distribution of ORFs that occur by chance? Consider an independence model with equal probabilities of four nucleotide types.

Solution There are $4^3 = 64$ triplets (codons) that will appear in the sequence with equal probabilities. Three out of the sixty four are stop codons. Therefore, the probability of encountering a stop codon upon scanning a sequence, triplet by triplet, is $3/64 = 0.047$. For the probability of occurrence of ORF of length L (in codons) we have

$$\begin{aligned} &P(\text{ORF of length } L \text{ starts in a given position}) \\ &\quad = P(ATG) \times P(\text{non-stop codon})^{L-2} \times P(\text{stop codon}) \\ &\quad = \frac{1}{64} \times \left(1 - \frac{3}{64}\right)^{L-2} \times \frac{3}{64} = \frac{3}{4096}\left(\frac{61}{64}\right)^{L-2}. \end{aligned}$$

To derive the ORF length distribution, we use the definition of conditional probability:

$$\begin{aligned} &P(\text{length of ORF is equal to } L) \\ &\quad = \frac{P(\text{ORF of length } L \text{ starts in a given position})}{P(\text{any ORF starts in a given position})} \end{aligned}$$

$$= \frac{P(\text{ORF of length } L \text{ starts in a given position})}{\sum_{L=2}^{+\infty} P(\text{ORF of length } L \text{ starts in a given position})}$$
$$= \frac{(3/4096)(61/64)^{L-2}}{1/64} = \frac{3}{64}\left(\frac{61}{64}\right)^{L-2}.$$

Thus, we have derived the geometric distribution of the lengths of random ORFs along with the parameters of the distribution. □

Problem 1.15 Assuming that non-coding DNA is described by the independence model with probabilities of nucleotides equal to 1/4, show that a gene start (under the assumption that the only start codon is the *ATG* codon) in 75% of cases is expected to coincide with the "longest ORF' start.

Solution Let us assume that a particular *ATG* codon is a *real* start of a gene, not overlapped by an adjacent gene. Then the DNA sequence located upstream to the *ATG* is non-coding DNA described by the independence model. Each possible triplet appears in sequence described by this model with probability 1/64. To find the probability that a given *ATG* situated at the *real* gene start is the 5′-most *ATG* in the ORF, we consider the complementary event that there is yet another *ATG* upstream to *real* start that would make an even longer ORF. By examining non-overlapping triplets upstream to the given *ATG* one at a time, starting with the one immediately adjacent to *ATG*, we observe one of the following possible outcomes. (i) The picked up triplet is one of sixty that are not *ATG*, *TAA*, *TGA*, *TAG*. In this case, we continue the process of triplet examining. (ii) This triplet is one of the three stop codons (*TAA*, *TGA*, *TAG*). We stop and infer that the initially considered *ATG* is the leftmost *ATG* in the ORF. (iii) The triplet under examination is *ATG*. We stop and infer that the initially considered *real* gene start is not the leftmost *ATG* in the ORF. Obviously, the termination of the scanning procedure by reaching one of the stop codons will occur three times more frequently than the termination by reaching the *ATG* codon. Therefore, the *ATG* start of a real gene in 75% of cases coincides with the leftmost *ATG* of the ORF, which defines the longest ORF for the fixed stop codon on the 3′ end. □

Problem 1.16 Suppose we consider two independence models of nucleotide sequence. The first model, M_1, has the same probabilities of nucleotides as defined in Problem 1.6. The second model, M_2, assigns to each nucleotide type the probability 1/4 to appear in any given position. Given the observed sequence $x = ACTGACGACTGAC$, compare the likelihoods of these models.

Solution The likelihood of a model is defined as the conditional probability of data (sequence x) given the model (Durbin *et al.* (1998), p. 6). Thus, we have to compare the probabilities of sequence x under each model. The likelihood of model M_1 is given by

$$P(x|M_1) = \left(\frac{3}{20}\right)^6 \left(\frac{7}{20}\right)^7.$$

Similarly, for the likelihood of model M_2 we have $P(x|M_2) = (1/4)^{13}$. The likelihood ratio is given by

$$\frac{P(x|M_2)}{P(x|M_1)} = \frac{20^{13}}{4^{13} \times 3^6 \times 7^7} = 2.0333 > 1.$$

Therefore, for the observed sequence x model M_2 has a greater likelihood than model M_1. □

Problem 1.17 A circular double-stranded DNA of $L = 3\,400$ nt long was cut by a restriction enzyme. A subsequent gel electrophoresis separation indicated the presence of five DNA pieces. It turned out that the absent-minded researcher could not recall the exact type of restriction enzyme that was used. Still, he knew that the chemical was picked up from a box containing equal number of 4-base cutters and 6-base cutters (restriction enzymes that cut specific 4 nt long sites and specific 6 nt long sites, respectively). What is the posterior probability that the 4-nucleotide cutter was used if the DNA sequence can be represented by the independence model with equal probabilities of nucleotides T, C, A, G.

Solution The probability of appearance of a restriction site in a particular position of DNA sequence is $p_1 = (1/4)^4 = 0.003906$ for the 4-base cutter and $p_2 = (1/4)^6 = 0.000244$ for the 6-base cutter.

We assume that in both cases the number X of restriction sites in the sequence can be approximated by the Poisson distribution with parameter $\lambda_1 = p_1 L = 13.28$ for the 4-base cutter and $\lambda_2 = p_2 L = 0.83$ for the 6-base cutter (see solution to Problem 1.11). Then we obtain

$$P(X = 5|\text{4-cutters}) = e^{-\lambda_1} \frac{(\lambda_1)^5}{5!} = 0.00588,$$

$$P(X = 5|\text{6-cutters}) = e^{-\lambda_2} \frac{(\lambda_2)^5}{5!} = 0.00143.$$

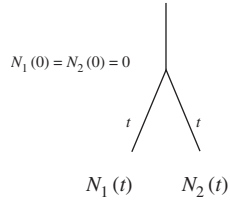


Figure 1.1. The simplest hylogenetic tree T with a pair of the homologous genes x^1 and x^2 being its leaves (see Problem 1.18).

We use Bayes' theorem to calculate the posterior probability that the 4-base cutter produced the restriction fragments:

$$\begin{aligned} &P(\text{4-cutters}|X = 5) \\ &\quad = \frac{P(X = 5|\text{4-cutters})P(\text{4-cutters})}{P(X = 5|\text{4-cutters})P(\text{4-cutters}) + P(X = 5|\text{6-cutters})P(\text{6-cutters})} \\ &\quad = \frac{0.00588 \times 0.5}{0.00588 \times 0.5 + 0.00143 \times 0.5} = 0.804. \end{aligned}$$

With 80.4% chance that the 4-base cutter was used, the initial uncertainty seems to be resolved. □

Problem 1.18 One theory states that the latest common ancestor of birds and crocodiles lived 120 million years ago (MYA), while another theory suggests that this time is twice as long. Comparison of homologous genes x^1 and x^2 of two species, the Nile crocodile and the Mediterranean seagull, revealed on average 365 differences in 1000 nt long fragments. It is assumed that mutations at different DNA sites occur independently, and at each site the number of mutations fixed in evolution is approximated by the Poisson process. The rate of mutation fixation, p, per nucleotide site per year, is equal to 10^{-9}. Given the observed number of differences, (a) compare the likelihoods of the two theories, and (b) determine the maximum likelihood estimate of the divergence time. For simplicity, assume that no more than one mutation could occur at any given nucleotide site of the whole lineage.

Solution (a) Assuming that the divergence of the two species occurred t years ago, we consider the simplest phylogenetic tree T with leaves x^1 and x^2. The occurrence of substitutions along branches of the tree can be described by two independent Poisson processes $N_1(\tau)$ and $N_2(\tau)$ both with parameter p. The moment of divergence corresponds to $\tau = 0$ and the present time to $\tau = t$ (see Figure 1.1).

We will compare the likelihoods of tree T for two values of the elapsed time, $t = t_1 = 120$ MYA and $t = t_2 = 240$ MYA, associated with the competing theories. The likelihood of a two-leaves tree with a molecular clock property depends on t only. Then the (conditional) likelihood at site u carrying matching nucleotides in DNA sequences is given by

$$\begin{aligned} L_u(t) &= P(x_u^1 = x_u^2 | t, \text{ no more than one mutation at site } i) \\ &= P(N_1(t) = 0, N_2(t) = 0 | N_1(t) + N_2(t) \le 1) \\ &= \frac{P(N_1(t) = 0, N_2(t) = 0, N_1(t) + N_2(t) \le 1)}{P(N_1(t) + N_2(t) \le 1)} \\ &= \frac{P(N_1(t) = 0, N_2(t) = 0)}{P(N_1(t) + N_2(t) \le 1)}. \end{aligned} \tag{1.4}$$

The numerator of the last expression in Equation (1.4) is equal to

$$P(N_1(t) = 0)P(N_2(t) = 0) = e^{-2pt}$$

due to the independence of processes $N_1(\tau)$ and $N_2(\tau)$, while $N_1(t) + N_2(t)$ is again the Poisson random variable (say $N_3(t)$) with parameter $2p$ due to the known property of the Poisson distribution. Thus, we have

$$\begin{aligned} P(N_1(t) + N_2(t) \le 1) &= P(N_3(t) \le 1) = P(N_3(t) = 0) + P(N_3(t) = 1) \\ &= e^{-2pt} + 2pte^{-2pt}, \end{aligned}$$

and the likelihood $L_u(t)$ from Equation (1.4) becomes

$$L_u(t) = \frac{e^{-2pt}}{e^{-2pt} + 2pte^{-2pt}} = (1 + 2pt)^{-1}. \tag{1.5}$$

Similarly, at site u with mismatching nucleotides the likelihood is given by

$$\begin{aligned} L_u(t) &= P(x_u^1 \neq x_u^2 | t, \text{ no more than one mutation at site } i) \\ &= P(N_1(t) = 0, N_2(t) = 1 | N_1(t) + N_2(t) \le 1) \\ &+ P(N_1(t) = 1, N_2(t) = 0 | N_1(t) + N_2(t) \le 1) \\ &= \frac{2P(N_1(t) = 0)P(N_2(t) = 1)}{P(N_1(t) + N_2(t) \le 1)} = \frac{2e^{-pt}pte^{-pt}}{e^{-2pt} + 2pte^{-2pt}} = 2pt(1 + 2pt)^{-1}. \end{aligned} \tag{1.6}$$

From Equations (1.5) and (1.6) we derive the likelihood of tree T with two leaves which are genomic sequences of length N aligned with M mismatches:

$$L(t) = \prod_{u=1}^{N} L_u(t) = (1 + 2pt)^{-N}(2pt)^M. \tag{1.7}$$

To test the two theories, we calculate the log-odds ratio for $t_1 = 120$ MYA and $t_2 = 240$ MYA:

$$\ln \frac{L(t_1)}{L(t_2)} = N \ln(1 + 2pt_2) - N \ln(1 + 2pt_1) + M \ln(2pt_1) - M \ln(2pt_2)$$
$$= -252.99 < 0.$$

Therefore, the available data support the theory that birds and crocodiles diverged 240 MYA, since this theory has a greater (conditional) likelihood than the competing one.

(b) We determine the maximum likelihood estimate t^* of the time of divergence of the two species as a maximum point of the logarithm of likelihood $L(t)$, formula (1.7):

$$\frac{d \ln L(t)}{dt} = -\frac{2Np}{1 + 2pt^*} + \frac{2pM}{2pt^*} = 0,$$
$$t^* = \frac{M}{2p(N - M)} = 2.874 \times 10^8.$$

Thus, $t^* = 287.4$ MYA is the maximum likelihood divergence time, while the maximum likelihood value *per se* is given by

$$L_{\max} = L(t^*) = (1 + 2pt^*)^{-N}(2pt^*)^M = 10^{-285}. \qquad \square$$

Problem 1.19 It is known that *CpG*-islands in high eukaryotes are relatively rich with *CpG* dinucleotides, while these dinucleotides are discriminated in the rest of a chromosome. It is assumed that the frequency of occurrences of *CpG* dinucleotides in a *CpG*-island can be approximated by the Poisson distribution with twenty-five *CpG* dinucleotides per 250 nt long fragment on average, while in the rest of the DNA this average is ten *CpG* per 250 nt. Suggest the Bayesian type algorithm for *CpG*-island identification. How will this algorithm characterize a 250 nt long DNA fragment containing nineteen *CpG* dinucleotides?

Solution We assume that that the numbers of occurrences of *CpG* dinucleotides in *CpG*-islands and non-*CpG*-islands are both described by the Poisson distribution with parameter $\lambda_1 = 25$ and $\lambda_2 = 10$, respectively.

If a given 250 nt long DNA fragment contains n dinucleotides *CpG*, how likely is it that the DNA fragment belongs to a *CpG*-island? We have to compare two *a posteriori* probabilities: $P_1 = P$(being a *CpG*-island given n observed *CpG* dinucleotides) and $P_2 = P$(being a non-*CpG*-island given n observed *CpG* dinucleotides). Assuming that both alternatives, being a *CpG*-island and being a non-*CpG*-island,

are *a priori* equally likely, we use Bayes' theorem to calculate P_1 and P_2:

$$\begin{aligned} P_1 &= P(\text{DNA fragment with } n \text{ } CpC \text{ has Poisson distribution with } \lambda_1 = 25) \\ &= \frac{P(n \text{ } CpG|\lambda_1 = 25)\frac{1}{2}}{P(n \text{ } CpG|\lambda_1 = 25)\frac{1}{2} + P(n \text{ } CpG|\lambda_2 = 10)\frac{1}{2}} \\ &= \frac{25^n e^{-25}}{10^n e^{-10} + 25^n e^{-25}}. \end{aligned}$$

In the above we applied the formula for Poisson distribution and canceled the common factor $n!$. Similarly,

$$\begin{aligned} P_2 &= P(\text{DNA fragment with } n \text{ } CpC \text{ has Poisson distribution with } \lambda_2 = 10) \\ &= \frac{P(n \text{ } CpG|\lambda_2 = 10)\frac{1}{2}}{P(n \text{ } CpG|\lambda_1 = 25)\frac{1}{2} + P(n \text{ } CpG|\lambda_2 = 10)\frac{1}{2}} \\ &= \frac{10^n e^{-10}}{10^n e^{-10} + 25^n e^{-25}}, \end{aligned}$$

or $P_2 = 1 - P_1$. The simple identification algorithm for a *CpG*-island works as follows. For a given 250 nt long DNA fragment with n observed *CpG* dinucleotides value P_1 is computed. If $P_1 > 0.5$ ($P_1 > P_2$), the DNA fragment is identified as a part of a *CpG*-island. Otherwise, the fragment is identified as a part of a non-*CpG*-island.

For $n = 19$ we have

$$\begin{aligned} P_1 &= \frac{(25)^{19} e^{-25}}{(25)^{19} e^{-25} + (10)^{19} e^{-10}} = 0.92, \\ P_2 &= 1 - P_1 = 0.08, \end{aligned}$$

and we conclude that the DNA fragment belongs to a *CpG*-island. □

Problem 1.20 Given the conditions stated in Problem 1.19, the following decision-making rule is accepted: if more than eighteen *CpG* dinucleotides are observed in a 250 nt long DNA fragment, it is identified as a *CpG*-island. Determine false positive and false negative rates of this method.

Solution The false positive rate (*FPR*) is defined as the probability that the rule would identify a non-*CpG*-island as a *CpG*-island. Since the number X of *CpG* dinucleotides in a non-*CpG*-island is described by the Poisson distribution with

parameter $\lambda = 10$, we have

$$\begin{aligned} FPR &= P(\text{more than eighteen } CpG \text{ out of } 250|\text{non-}CpG\text{-island}) \\ &= P(X > 18|\lambda = 10) = \sum_{n=19}^{+\infty} P(X = n) = 1 - \sum_{n=0}^{18} P(X = n) \\ &= 1 - \sum_{n=0}^{18} e^{-10} \frac{10^n}{n!} \approx 0.007. \end{aligned}$$

The false negative rate (*FNR*) is defined as the probability that a *CpG*-island is identified as a non-*CpG*-island. Since the number Y of *CpG* dinucleotides in a *CpG*-island region has the Poisson distribution with parameter $\lambda = 25$, the false negative rate is given by

$$\begin{aligned} FNR &= P(\text{less or equal to eighteen } CpG \text{ out of } 250|CpG\text{-island}) \\ &= P(Y \leq 18|\lambda = 25) = \sum_{n=0}^{18} P(Y = n) \\ &= \sum_{n=0}^{18} e^{-25} \frac{25^n}{n!} \approx 0.09. \end{aligned}$$

Note that $FPR < FNR$. This means that the classification rule is more likely to decide that *CpG*-island DNA is non-*CpG*-island DNA than vice versa. □

Problem 1.21 An inhomogeneous DNA sequence is known to contain both $C+G$-rich composition regions and regions with unbiased (uniform) nucleotide composition. We assume that the independence model (P model) with parameters $p_T = 1/8$, $p_C = 3/8$, $p_A = 1/8$, $p_G = 3/8$, describes the regions with high $C + G$ content. Regions with uniform nucleotide composition are described by the independence model (Q model) with parameters $q_T = 1/4$, $q_C = 1/4$, $q_A = 1/4$, $q_G = 1/4$. For a given DNA fragment X, the log-odds ratio, $L = \log_2[P(X|P)/P(X|Q)]$ is determined, and, if $L \geq 0$, X is classified as a high $C+G$ composition fragment; if $L < 0$, X is classified as compositionally unbiased. Determine the probabilities of type-one error (false negative rate) and type-two error (false positive rate) of the classification of a DNA fragment of length n. Consider $n = 10, 20$ and 100.

Solution For a DNA sequence X of length n we test the null hypothesis,

$$H_0 = \{X \text{ belongs to a } C + G\text{-rich region}\} = \{X \in P\},$$

versus the alternative hypothesis,

$$H_a = \{X \text{ belongs to a region with uniform composition}\} = \{X \in Q\}.$$

The log-odds ratio is given by

$$L = \log_2[P(X|P)/P(X|Q)] = \log_2\left(\left(\frac{p_A}{q_A}\right)^{n_1}\left(\frac{p_C}{q_C}\right)^{n_2}\left(\frac{p_T}{q_T}\right)^{n_3}\left(\frac{p_G}{q_G}\right)^{n_4}\right)$$
$$= n_1\log_2\frac{1}{2} + n_2\log_2\frac{3}{2} + n_3\log_2\frac{1}{2} + n_4\log_2\frac{3}{2} \approx 0.585(n_2+n_4) - (n_1+n_3),$$

where n_1, n_2, n_3, and n_4 are numbers of nucleotides A, C, T, and G, respectively, observed in fragment X. We accept hypothesis H_0 (and reject H_a) if $L \geq 0$, i.e. if $P(X|P) \geq P(X|Q)$; and we reject H_0 (and accept H_a) otherwise. For the type-one error α (significance level of the test) we have

$$\alpha = P(\text{type-one error}) = P(H_0 \text{ is rejected}|H_0 \text{ is true}) = P(L < 0|X \in P)$$
$$= P(0.585(n_2+n_4) - (n_1+n_3) < 0|X \in P).$$

Next, we define the Bernoulli trial outcomes by interpreting an occurrence of A or T at a given site of sequence X as a "success" and an occurrence of C or G as a "failure." If p is the probability of "success," then the number of "successes" in n Bernoulli trials, $S = n_1 + n_3$, has a binomial distribution with parameters n and p. The type-one error, α, becomes

$$\alpha = P(0.585(n_2+n_4) - (n_1+n_3) < 0|X \in P)$$
$$= P(0.585(n-S) - S < 0|S \in B(n, p = 1/4))$$
$$= P(S > 0.369n|S \in B(n, p = 1/4))$$
$$= \sum_{k:0.369n<k\leq n}\binom{n}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{n-k} = \left(\frac{1}{4}\right)^n\sum_{k:0.369n<k\leq n}\binom{n}{k}3^{n-k}.$$

It follows from the central limit theorem that as $n \to \infty$ the sequence of random variables $(S - \mathbf{E}S)/\sqrt{\mathbf{Var}S}$ weakly converges to the standard normal distribution. Therefore, for large n,

$$\alpha = P(S > 0.369n|S \in B(n, p = 1/4)) = P\left(\frac{S - \mathbf{E}S}{\sqrt{\mathbf{Var}S}} > \frac{0.369n - \mathbf{E}S}{\sqrt{\mathbf{Var}S}}\right)$$
$$= P\left(\frac{S - 0.25n}{0.25\sqrt{3n}} > 0.2748\sqrt{n}\right) \approx 1 - \Phi(0.2748\sqrt{n}).$$

Here

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{t^2}{2}\right)dt$$