R. Durbin
S. Eddy
A. Krogh
G. Mitchison

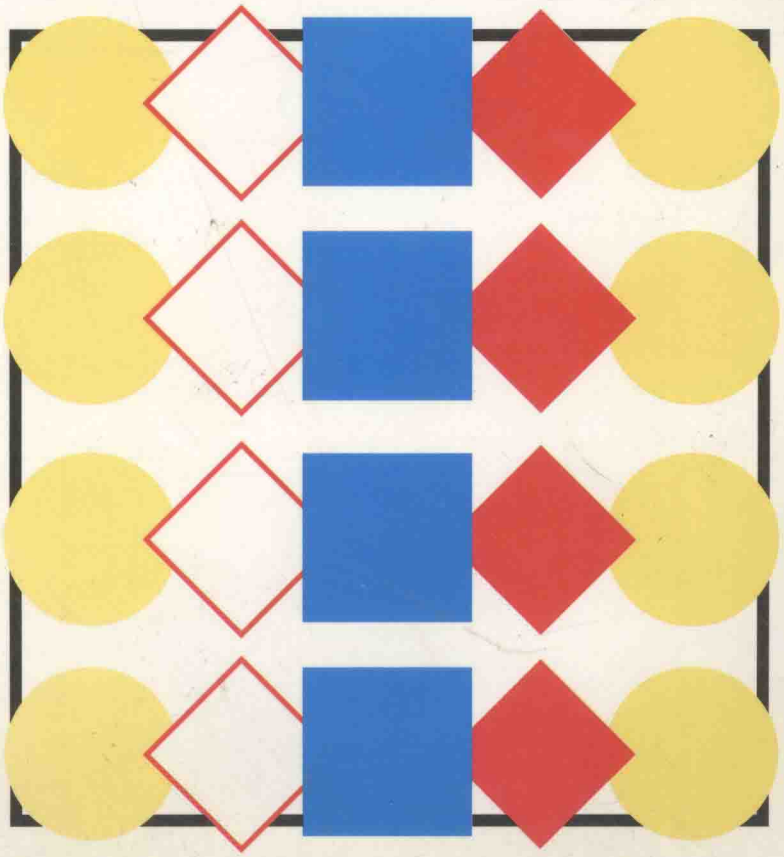
生物信息学 引进版 丛书

生物序列分析

Biological sequence analysis

蛋白质和核酸的概率论模型

Probabilistic models of proteins and nucleic acids



剑桥大学出版社
清华大学出版社

（京）新登字 158 号

Biological sequence analysis

Probabilistic models of proteins and nucleic acids

R. Durbin, S. Eddy, A. Krogh, G. Mitchison

北京市版权局著作权合同登记号： 图字 01-2001-4300

书　名：生物序列分析，蛋白质和核酸的概率论模型
作　者：R. Durbin, S. Eddy, A. Krogh, G. Mitchison
出版者：清华大学出版社（北京清华大学学研大厦，邮编 100084）
http:// www.tup.tsinghua.edu.cn
印刷者：北京市清华园胶印厂
发行者：新华书店总店北京发行所
开　本：787×1092 1/16　印张：23
版　次：2002 年 1 月第 1 版　2002 年 1 月第 1 次印刷
书　号：ISBN 7-302-05097-X/Q・21
印　数：0001～4000
定　价：38.00 元

Biological sequence analysis
Probabilistic models of proteins and nucleic acids

The face of biology has been changed by the emergence of modern molecular genetics. Among the most exciting advances are large-scale DNA sequencing efforts such as the Human Genome Project which are producing an immense amount of data. The need to understand the data is becoming ever more pressing. Demands for sophisticated analyses of biological sequences are driving forward the newly-created and explosively expanding research area of computational molecular biology, or bioinformatics.

Many of the most powerful sequence analysis methods are now based on principles of probabilistic modelling. Examples of such methods include the use of probabilistically derived score matrices to determine the significance of sequence alignments, the use of hidden Markov models as the basis for profile searches to identify distant members of sequence families, and the inference of phylogenetic trees using maximum likelihood approaches.

This book provides the first unified, up-to-date, and tutorial-level overview of sequence analysis methods, with particular emphasis on probabilistic modelling. Pairwise alignment, hidden Markov models, multiple alignment, profile searches, RNA secondary structure analysis, and phylogenetic inference are treated at length.

Written by an interdisciplinary team of authors, the book is accessible to molecular biologists, computer scientists and mathematicians with no formal knowledge of each others' fields. It presents the state-of-the-art in this important, new and rapidly developing discipline.

Richard Durbin is Head of the Informatics Division at the Sanger Centre in Cambridge, England.

Sean Eddy is Assistant Professor at Washington University's School of Medicine and also one of the Principle Investigators at the Washington University Genome Sequencing Center.

Anders Krogh is a Research Associate Professor in the Center for Biological Sequence Analysis at the Technical University of Denmark.

Graeme Mitchison is at the Medical Research Council's Laboratory for Molecular Biology in Cambridge, England.

出 版 前 言

生物信息学是当今生命科学和自然科学的重大前沿领域之一，同时也将是21世纪自然科学的核心领域和最具活力的领域之一。20世纪信息技术蓬勃发展的态势为其与许多学科交叉提供了基础。生物信息学（Bioinformatics）是在生命科学的研究中，以计算机为工具对生物信息进行储存、检索和分析的科学。其研究重点主要体现在基因组学（Genomics）和蛋白组学（Proteomics）两方面，具体的说就是从核酸和蛋白质序列出发，分析序列中表达的结构功能的生物信息。为了适应21世纪生命科学的蓬勃发展，我们从国外众多优秀出版社引进了这套《引进版　生物信息学丛书》，分为影印版和中文版，以飨读者。

希望这套丛书能对国内的生物信息学研究有所帮助，同时对我国的生命科学赶超世界先进水平起到一定的推动作用。

欢迎广大读者将使用本系列后的意见反馈给我们，更欢迎国内外专家、教授积极向我社推荐国外的优秀生命科学图书和优秀的作者，为生物学教学和科研起到积极的促进作用。

清华大学出版社

2001年11月

Preface

At a Snowbird conference on neural nets in 1992, David Haussler and his colleagues at UC Santa Cruz (including one of us, AK) described preliminary results on modelling protein sequence multiple alignments with probabilistic models called 'hidden Markov models' (HMMs). Copies of their technical report were widely circulated. Some of them found their way to the MRC Laboratory of Molecular Biology in Cambridge, where RD and GJM were just switching research interests from neural modelling to computational genome sequence analysis, and where SRE had arrived as a new postdoctoral student with a background in experimental molecular genetics and an interest in computational analysis. AK later also came to Cambridge for a year.

All of us quickly adopted the ideas of probabilistic modelling. We were persuaded that hidden Markov models and their stochastic grammar analogues are beautiful mathematical objects, well fitted to capturing the information buried in biological sequences. The Santa Cruz group and the Cambridge group independently developed two freely available HMM software packages for sequence analysis, and independently extended HMM methods to stochastic context-free grammar analysis of RNA secondary structures. Another group led by Pierre Baldi at JPL/Caltech was also inspired by the work presented at the Snowbird conference to work on HMM-based approaches at about the same time.

By late 1995, we thought that we had acquired a reasonable amount of experience in probabilistic modelling techniques. On the other hand, we also felt that relatively little of the work had been communicated effectively to the community. HMMs had stirred widespread interest, but they were still viewed by many as mathematical black boxes instead of natural models of sequence alignment problems. Many of the best papers that described HMM ideas and methods in detail were in the speech recognition literature, effectively inaccessible to many computational biologists. Furthermore, it had become clear to us and several other groups that the same ideas could be applied to a much broader class of problems, including protein structure modelling, genefinding, and phylogenetic analysis. Over the Christmas break in 1995–96, perhaps somewhat deluded by ambition, naiveté, and holiday relaxation, we decided to write a book on biological sequence analysis emphasizing probabilistic modelling. In the past two years, our original grand plans have been distilled into what we hope is a practical book.

This is a subjective book written by opinionated authors. It is not a tutorial on practical sequence analysis. Our main goal is to give an accessible introduction to the foundations of sequence analysis, and to show why we think the probabilistic modelling approach is useful. We try to avoid discussing specific computer programs, and instead focus on the algorithms and principles behind them.

We have carefully cited the work of the many authors whose work has influenced our thinking. However, we are sure we have failed to cite others whom we *should* have read, and for this we apologise. Also, in a book that necessarily touches on fields ranging from evolutionary biology through probability theory to biophysics, we have been forced by limitations of time, energy, and our own imperfect understanding to deal with a number of issues in a superficial manner.

Computational biology is an interdisciplinary field. Its practitioners, including us, come from diverse backgrounds, including molecular biology, mathematics, computer science, and physics. Our intended audience is any graduate or advanced undergraduate student with a background in one of these fields. We aim for a concise and intuitive presentation that is neither forbiddingly mathematical nor too technically biological.

We assume that readers are already familiar with the basic principles of molecular genetics, such as the Central Dogma that DNA makes RNA makes protein, and that nucleic acids are sequences composed of four nucleotide subunits and proteins are sequences composed of twenty amino acid subunits. More detailed molecular genetics is introduced where necessary. We also assume a basic proficiency in mathematics. However, there are sections that are more mathematically detailed. We have tried to place these towards the end of each chapter, and in general towards the end of the book. In particular, the final chapter, Chapter 11, covers some topics in probability theory that are relevant to much of the earlier material.

We are grateful to several people who kindly checked parts of the manuscript for us at rather short notice. We thank Ewan Birney, Bill Bruno, David MacKay, Cathy Eddy, Jotun Hein, and Søren Riis especially. Bret Larget and Robert Mau gave us very helpful information about the sampling methods they have been using for phylogeny. David Haussler bravely used an embarrassingly early draft of the manuscript in a course at UC Santa Cruz in the autumn of 1996, and we thank David and his entire class for the very useful feedback we received. We are also grateful to David for inspiring us to work in this field in the first place. It has been a pleasure to work with David Tranah and Maria Murphy of Cambridge University Press and Sue Glover of SG Publishing in producing the book; they demonstrated remarkable expertise in the editing and LaTeX typesetting of a book laden with equations, algorithms, and pseudocode, and also remarkable tolerance of our wildly optimistic and inaccurate target dates. We are sure that some of our errors remain, but their number would be far greater without the help of all these people.

We also wish to thank those who supported our research and our work on this book: the Wellcome Trust, the NIH National Human Genome Research Institute, NATO, Eli Lilly & Co., the Human Frontiers Science Program Organisation, and the Danish National Research Foundation. We also thank our home institutions: the Sanger Centre (RD), Washington University School of Medicine (SRE), the Center for Biological Sequence Analysis (AK), and the MRC Laboratory of Molecular Biology (GJM). Jim and Anne Durbin graciously lent us the use of their house in London in February 1997, where an almost final draft of the book coalesced in a burst of writing and criticism. We thank our friends, families, and research groups for tolerating the writing process and SRE's and AK's long trips to England. We promise to take on no new grand projects, at least not immediately.

Contents

1

Introduction

Astronomy began when the Babylonians mapped the heavens. Our descendants will certainly not say that biology began with today's genome projects, but they may well recognise that a great acceleration in the accumulation of biological knowledge began in our era. To make sense of this knowledge is a challenge, and will require increased understanding of the biology of cells and organisms. But part of the challenge is simply to organise, classify and parse the immense richness of sequence data. This is more than an abstract task of string parsing, for behind the string of bases or amino acids is the whole complexity of molecular biology. This book is about methods which are in principle capable of capturing some of this complexity, by integrating diverse sources of biological information into clean, general, and tractable probabilistic models for sequence analysis.

Though this book is about computational biology, let us be clear about one thing from the start: the most reliable way to determine a biological molecule's structure or function is by direct experimentation. However, it is far easier to obtain the DNA sequence of the gene corresponding to an RNA or protein than it is to experimentally determine its function or its structure. This provides strong motivation for developing computational methods that can infer biological information from sequence alone. Computational methods have become especially important since the advent of genome projects. The Human Genome Project alone will give us the raw sequences of an estimated 70 000 to 100 000 human genes, only a small fraction of which have been studied experimentally.

Most of the problems in computational sequence analysis are essentially statistical. Stochastic evolutionary forces act on genomes. Discerning significant similarities between anciently diverged sequences amidst a chaos of random mutation, natural selection, and genetic drift presents serious signal to noise problems. Many of the most powerful analysis methods available make use of probability theory. In this book we emphasise the use of probabilistic models, particularly *hidden Markov models* (HMMs), to provide a general structure for statistical analysis of a wide variety of sequence analysis problems.

1.1 Sequence similarity, homology, and alignment

Nature is a tinkerer and not an inventor [Jacob 1977]. New sequences are adapted from pre-existing sequences rather than invented *de novo*. This is very fortunate for computational sequence analysis. We can often recognise a significant similarity between a new sequence and a sequence about which something is already known; when we do this we can transfer information about structure and/or function to the new sequence. We say that the two related sequences are *homologous* and that we are transfering information *by homology*.

At first glance, deciding that two biological sequences are similar is no different from deciding that two text strings are similar. One set of methods for biological sequence analysis is therefore rooted in computer science, where there is an extensive literature on string comparison methods. The concept of an *alignment* is crucial. Evolving sequences accumulate insertions and deletions as well as substitutions, so before the similarity of two sequences can be evaluated, one typically begins by finding a plausible alignment between them.

Almost all alignment methods find the best alignment between two strings under some scoring scheme. These scoring schemes can be as simple as '+1 for a match, −1 for a mismatch'. Indeed, many early sequence alignment algorithms were described in these terms. However, since we want a scoring scheme to give the biologically most likely alignment the highest score, we want to take into account the fact that biological molecules have evolutionary histories, three-dimensional folded structures, and other features which constrain their primary sequence evolution. Therefore, in addition to the mechanics of alignment and comparison algorithms, the scoring system itself requires careful thought, and can be very complex.

Developing more sensitive scoring schemes and evaluating the significance of alignment scores is more the realm of statistics than computer science. An early step forward was the introduction of probabilistic matrices for scoring pairwise amino acid alignments [Dayhoff, Eck & Park 1972; Dayhoff, Schwartz & Orcutt 1978]; these serve to quantify evolutionary preferences for certain substitutions over others. More sophisticated probabilistic modelling approaches have been brought gradually into computational biology by many routes. Probabilistic modelling methods greatly extend the range of applications that can be underpinned by useful and consistent theory, by providing a natural framework in which to address complex inference problems in computational sequence analysis.

1.2 Overview of the book

The book is loosely structured into four parts covering problems in pairwise alignment, multiple alignment, phylogenetic trees, and RNA structure. Figure 1.1

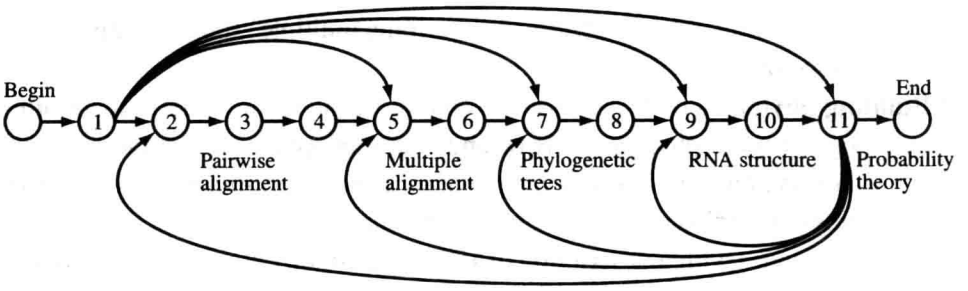


Figure 1.1 *Overview of the book, and suggested paths through it.*

shows suggested paths through the chapters in the form of a *state machine*, one sort of model we will use throughout the book.

The individual chapters cover topics as follows:

2 Pairwise alignment. We start with the problem of deciding if a pair of sequences are evolutionarily related or not. We examine traditional pairwise sequence alignment and comparison algorithms which use dynamic programming to find optimal gapped alignments. We give some probabilistic analysis of scoring parameters, and some discussion of the statistical significance of matches.

3 Markov chains and hidden Markov models. We introduce hidden Markov models (HMMs) and show how they are used to model a sequence or a family of sequences. The chapter gives all the basic HMM algorithms and theory, using simple examples.

4 Pairwise alignment using HMMs. Newly equipped with HMM theory, we revisit pairwise alignment. We develop a special sort of HMM that models aligned pairs of sequences. We show how the HMM-based approach provides some nice ways of estimating accuracy of an alignment, and scoring similarity without committing to any particular alignment.

5 Profile HMMs for sequence families. We consider the problem of finding sequences which are homologous to a known evolutionary family or superfamily. One standard approach to this problem has been the use of 'profiles' of position-specific scoring parameters derived from a multiple sequence alignment. We describe a standard form of HMM, called a profile HMM, for modelling protein and DNA sequence families based on multiple alignments. Particular attention is given to parameter estimation for optimal searching for new family members, including a discussion of sequence weighting schemes.

6 Multiple sequence alignment methods. A closely related problem is that of constructing a multiple sequence alignment of a family. We examine existing multiple sequence alignment algorithms from the standpoint of

probabilistic modelling, before describing multiple alignment algorithms based on profile HMMs.

7 Building phylogenetic trees. Some of the most interesting questions in biology concern phylogeny. How and when did genes and species evolve? We give an overview of some popular methods for inferring evolutionary trees, including clustering, distance and parsimony methods. The chapter concludes with a description of Hein's parsimony algorithm for simultaneously aligning and inferring the phylogeny of a sequence family.

8 A probabilistic approach to phylogeny. We describe the application of probabilistic modelling to phylogeny, including maximum likelihood estimation of tree scores and methods for sampling the posterior probability distribution over the space of trees. We also give a probabilistic interpretation of the methods described in the preceding chapter.

9 Transformational grammars. We describe how hidden Markov models are just the lowest level in the Chomsky hierarchy of transformational grammars. We discuss the use of more complex transformational grammars as probabilistic models of biological sequences, and give an introduction to the stochastic context-free grammars, the next level in the Chomsky hierarchy.

10 RNA structure analysis. Using stochastic context-free grammar theory, we tackle questions of RNA secondary structure analysis that cannot be handled with HMMs or other primary sequence-based approaches. These include RNA secondary structure prediction, structure-based alignment of RNAs, and structure-based database search for homologous RNAs.

11 Background on probability. Finally, we give more formal details for the mathematical and statistical toolkit that we use in a fairly informal tutorial-style fashion throughout the rest of the book.

1.3 Probabilities and probabilistic models

Some basic results in using probabilities are necessary for understanding almost any part of this book, so before we get going with sequences, we give a brief primer here on the key ideas and methods. For many readers, this will be familiar territory. However, it may be wise to at least skim though this section to get a grasp of the notation and some of the ideas that we will develop later in the book. Aside from this very basic introduction, we have tried to minimise the discussion of abstract probability theory in the main body of the text, and have instead concentrated the mathematical derivations and methods into Chapter 11, which contains a more thorough presentation of the relevant theory.

What do we mean by a probabilistic model? 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.

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. A model of a roll of a (possibly loaded) die would have six parameters $p_1 \ldots p_6$; 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 $\sum_{i=1}^{6} p_i = 1$. 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, $p_1 p_6 p_3$. We will use dice throughout the early part of the book for giving intuitive simple examples of probabilistic modelling.

Consider a second example closer to our biological subject matter, which is an extremely 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 $x_1 \ldots x_n$, the probability of the whole sequence is then the product $q_{x_1} q_{x_2} \cdots q_{x_n} = \prod_{i=1}^{n} q_{x_i}$.[1] We will use this 'random sequence model' throughout the book as 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 [Bairoch & Apweiler 1997].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 maximises 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 maximises $P(D|\theta)$. This is discussed more formally in Chapter 11.

When estimating parameters for a model from a limited amount of data, there

[1] Strictly speaking this is only a correct model if all sequences have the same length, because then the sum of the probability over all possible sequences is 1; see Chapter 3.

is a danger of *overfitting*, which means that the model becomes very well adapted to the training data, but it will not *generalise* well to new data. Observing for instance the three flips of a coin [*tail, tail, tail*] would lead to the maximum likelihood estimate that the probability of *head* is 0 and that of *tail* is 1. We will return shortly to methods for preventing overfitting.

Conditional, joint, and marginal probabilities

Suppose we have two dice, D_1 and D_2. The probability of rolling an i with die D_1 is called $P(i|D_1)$. This is the *conditional probability* of rolling i given die D_1. If we pick a die at random with probability $P(D_j), j = 1$ or 2, the probability for picking die j and rolling an i is the product of the two probabilities, $P(i, D_j) = P(D_j)P(i|D_j)$. The term $P(i, D_j)$ is called the *joint probability*. The statement

$$P(X,Y) = P(X|Y)P(Y) \tag{1.1}$$

applies universally to any events X and Y.

When conditional or joint probabilities are known, we can calculate a *marginal* probability that removes one of the variables by using

$$P(X) = \sum_Y P(X,Y) = \sum_Y P(X|Y)P(Y),$$

where the sums are over all possible events Y.

Exercise

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?

Bayes' theorem and model comparison

In the same occasionally dishonest casino as in Exercise 1.1, we pick a die at random and roll it three times, getting three consecutive sixes. We are suspicious that this is a loaded die. How can we evaluate whether that is the case? What we want to know is $P(D_{\text{loaded}}|3 \text{ sixes})$; i.e. the *posterior probability* of the hypothesis that the die is loaded given the observed data, but what we can directly calculate is the probability of the data given the hypothesis, $P(3 \text{ sixes}|D_{\text{loaded}})$, which is called the *likelihood* of the hypothesis. We can calculate posterior probabilities using Bayes' theorem,

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}. \tag{1.2}$$

The event 'the die is loaded' corresponds to X in (1.2) and '3 sixes' corresponds to Y, so

$$P(D_{\text{loaded}}|3 \text{ sixes}) = \frac{P(3 \text{ sixes}|D_{\text{loaded}})P(D_{\text{loaded}})}{P(3 \text{ sixes})}.$$

We were given (see Exercise 1.1) that the probability $P(D_{\text{loaded}})$ of picking a loaded die is 0.01, and we know that the probability $P(3 \text{ sixes}|D_{\text{loaded}})$ of three sixes given it is loaded is $0.5^3 = 0.125$. The total probability of three sixes, $P(3 \text{ sixes})$, is just $P(3 \text{ sixes}|D_{\text{loaded}})P(D_{\text{loaded}}) + P(3 \text{ sixes}|D_{\text{fair}})P(D_{\text{fair}})$. Now

$$\begin{aligned} P(D_{\text{loaded}}|3 \text{ sixes}) &= \frac{(0.5^3)(0.01)}{(0.5^3)(0.01)+(\frac{1}{6}^3)(0.99)} \\ &= 0.21. \end{aligned}$$

So in fact, it is still more likely that we picked up a fair die, despite seeing three successive sixes.

As a second, more biological example, let us assume we believe that, on average, extracellular proteins have a slightly different amino acid composition than intracellular proteins. For example, we might think that cysteine is more common in extracellular than intracellular proteins. Let us try to use this information to judge whether a new protein sequence $x = x_1 \ldots x_n$ is intracellular or extracellular. To do this, we first split our training examples from SWISS-PROT into intracellular and extracellular proteins (we can leave aside unclassifiable cases).

We can now estimate a set of frequencies q_a^{int} for intracellular proteins, and a corresponding set of extracellular frequencies q_a^{ext}. To provide all the necessary information for Bayes' theorem, we also need to estimate the probability that any new sequence is extracellular, p^{ext}, and the corresponding probability of being intracellular, p^{int}. We will assume for now that every sequence must be either entirely intracellular or entirely extracellular, so $p^{\text{int}} = 1 - p^{\text{ext}}$. The values p^{ext} and p^{int} are called the *prior* probabilities, because they represent the best guess that we can make about a sequence *before* we have seen any information about the sequence itself.

We can now write $P(x|\text{ext}) = \prod_i q_{x_i}^{\text{ext}}$ and $P(x|\text{int}) = \prod_i q_{x_i}^{\text{int}}$. Because we are assuming that every sequence must be extracellular or intracellular, $p(x) = p^{\text{ext}}P(x|\text{ext}) + p^{\text{int}}P(x|\text{int})$. By Bayes' theorem,

$$P(\text{ext}|x) = \frac{p^{\text{ext}} \prod_i q_{x_i}^{\text{ext}}}{p^{\text{ext}} \prod_i q_{x_i}^{\text{ext}} + p^{\text{int}} \prod_i q_{x_i}^{\text{int}}}.$$

$P(\text{ext}|x)$ is the number we want. It is called the *posterior* probability that a sequence is extracellular because it is our best guess *after* we have seen the data.

Of course, this example is confounded by the fact that many transmembrane proteins have intracellular and extracellular components. We really want to be able to switch from one assignment to the other while in the sequence. That