Chapter 2. Bayesian Modeling and Computation in Bioinformatics Research

Jun S. Liu

Professor of Statistics

Harvard University

Science Center, One Oxford Street

Cambridge, MA 02138, USA

1 Introduction

With the completion of the human genome and genomes of many other species, the task of organizing and understanding the generated sequence and structural data becomes more and more pressing. These datasets also present great research opportunities to all quantitative researchers interested in biological problems. In the past decade, computational approaches to molecular and structural biology have attracted increasing attentions from both laboratory biologists and mathematical scientists, e.g., computer scientists, mathematicians, and statisticians, and have spawned the new field of bioinformatics. Among all available computational methods, those that are developed based on explicit statistical models play an important role in the field and will be the main focus of this chapter.

The use of probability theory and statistical principles in guarding against false optimism has been well understood by most scientists. The concepts of confidence interval, p-value, significance level, and the power of a statistical test routinely appear in scientific publications. To most scientists, these concepts represent, to a large extent, what statistics is about and what a statistician can contribute to a scientific problem. The invention of clever ideas, efficient algorithms, and general methodologies seem to be the privilege of scientific geniuses and are seldom attributed to a statistical methodology. In general, statistics or statistical thinking is not regarded as very helpful in attacking a difficult scientific problem. What we want to show here is that, quite in contrast to this "common wisdom," formal statistical modeling together with advanced statistical algorithms provide us a powerful "workbench" for developing innovative computational strategies and for making proper inference to account for estimation uncertainties.

In the past decade, we have witnessed the developments of the likelihood approach to pairwise alignments (Bishop and Thompson, 1986; Thorne et al., 1991); the probabilistic models for RNA secondary structure (Zuker, 1989; Lowe and Eddy, 1997); the expectation-maximization (EM) algorithm for finding regulatory binding motifs (Reilly and Lawrence, 1992; Cardon and Stormo 1992); the Gibbs sampling strategies for detecting subtle similarities (Lawrence et al., 1993;

Liu 1994; Neuwald et al., 1997); and the hidden Markov models (HMM) for DNA composition analysis and multiple alignments (Churchill, 1989; Baldi et al., 1994; Krogh et al., 1994); the hidden semi-Markov model for gene prediction and protein secondary structure prediction (Burge and Karlin, 1997; Schmidler et al., 2000). All these developments show that algorithms resulting from statistical modeling efforts constitute a major part of today's bioinformatics toolbox.

Our emphases in this chapter is on the applications of the *Bayesian methodology* and its related algorithms in bioinformatics. We prefer a Bayesian approach mainly for the following reasons: (a) its explicit use of probabilistic models to formulate scientific problems (i.e., a quantitative story-telling); (b) its coherent way of incorporating all sources of information and of treating *nuisance* parameters and missing data; and (c) its ability to quantify numerically uncertainties in all unknowns. In Bayesian analysis, a *comprehensive* probabilistic model is employed to describe relationships among various quantities under consideration: those that we observe (data and knowledge), those about which we wish to learn (scientific hypotheses), and those that are needed in order to construct a proper model (a scaffold). With this Bayesian model, the basic probability theory can automatically lead us to an efficient use of the available information when making predictions and to a numerical quantification of uncertainty in these predictions (Gelman et al., 1995). To date, statistical approaches have been primarily used in computational biology for deriving efficient algorithms. The utility of these methods to make statistical inferences about unobserved variables has received less attention.

An important yet subtle issue in applying the Bayes approach is the choice of a *prior* distribution for the unknown parameters. Since it is inevitable to inject certain arbitrariness and subjective judgements into the analysis when prescribing a prior distribution, the Bayes methods have long been regarded as less "objective" than its frequentist counterpart (Section 2), and thus, disfavored. Indeed, it is often nontrivial to choose an appropriate prior distribution when the parameter space is of high-dimensional. All researchers who intend to use Bayesian methods for serious scientific studies need to put some thought into this issue. On the other hand, however, any scientific investigation has to involve a substantial amount of assumptions and personal judgements from the scientist(s) who conduct the investigation. These subjective elements, if made explicit and treated with care, should not undermine the scientific results of the investigation. More importantly, it should be regarded as a good scientific practice if the investigators make their subjective inputs explicit. Similarly, we argue that an appropriate subjective input in the form of a prior distribution should only enhance the relevance and accuracy of the Bayesian inference. Being able to make an explicit use of subjective knowledge is a virtue, instead of blemish, of Bayesian methods.

This chapter is organized as follows. Section 2 discusses the importance of formal statistical modeling and gives an overview of two main approaches to statistical inference: the frequentist and Bayesian. Section 3 outlines the Bayesian procedure for treating a statistical problem, with an emphasis on using the missing data formulation to construct scientifically meaningful models. Section 4 describes several popular algorithms for dealing with statistical computations: the EM algorithm, the Metropolis algorithm, and the Gibbs sampler. Section 5 demonstrates how the Bayesian method can be used to study a sequence composition problem. Section 6 gives a further example of using the Bayesian method to find subtle repetitive motifs in a DNA sequence. Section 7 concludes the chapter with a brief discussion.

2 Statistical Modeling and Inference

2.1 Parametric statistical modeling

Statistical modeling and analysis, including the collection of data, the construction of a probabilistic model, the quantification and incorporation of expert opinions, the interpretation of the model and the results, and the prediction from the data, form an essential part of the scientific method in diverse fields. The key focus of statistics is on making inferences, where the word *inference* follows the dictionary definition as "the process of deriving a conclusion from fact and/or premise." In statistics, the facts are the observed data, the premise is represented by a probabilistic model of the system of interest, and the conclusions concern unobserved quantities. Statistical inference distinguishes itself from other forms of inferences by explicitly quantifying uncertainties involved in the premise and the conclusions.

In *nonparametric* statistical inference, one does not assume any specific distributional form for the probability law of the observed data, but only imposes on the data a dependence (or independence) structure. For example, an often imposed assumption in nonparametric analyses is that the observations are *independent and identically distributed* (iid). When the observed data are continuous quantities, what one has to infer for this nonparametric model is the whole density curve — an infinite dimensional parameter. A main advantage of nonparametric methods is that the resulting inferential statements are relatively more robust than those from parametric methods. A main disadvantage of the nonparametric approach is, however, that it is difficult, and sometimes impossible, to build into the model more sophisticated structures (based on our scientific knowledge), i.e., it does not facilitate "learning."

Indeed, it would be ideal and preferable if we could derive what we want without having to assume anything. On the other hand, however, the process of using simple models (with a small number of adjustable parameters) to describe natural phenomena and then improving upon them (e.g., Newton's law of motion versus Einstein's theory of relativity) is at the heart of all scientific investigations. Parametric modeling, either analytically or qualitatively, either explicitly or implicitly, is intrinsic to human intelligence, i.e., it is essentially the only way we learn about the outside world. Analogously, statistical analysis based on parametric modeling is also essential to our scientific understanding of the data.

At a conceptual level, probabilistic models in statistical analyses serve as a mechanism through which one connects observed data with a scientific premise or hypothesis about real-world phenomena. Since bioinformatics explicitly or implicitly concerns the analysis of biological data that are intrinsically probabilistic, such models should be also at the core of bioinformatics. No model can completely represent every detail of reality. The goal of modeling is to abstract the key features of the underlying scientific problem into a workable mathematical form with which the scientific premise may be examined. Families of probability distributions characterized by a small number of parameters are most useful for this purpose.

Let $\mathbf{y}$ denote the observed data. In *parametric inference*, we assume that the observation follows a probabilistic law that belongs to a given distribution family. That is, $\mathbf{y}$ is a realization of a random process (i.e., a sample from a distribution) whose probability law has a particular form (e.g., Gaussian, multinomial, Dirichlet etc.), $f(\mathbf{y} \mid \boldsymbol{\theta})$, which is completely known other than $\boldsymbol{\theta}$. Here $\boldsymbol{\theta}$ is called a (population) *parameter* and it often corresponds to a scientific premise for our understanding of a natural process. To be concrete, one can imagine that $\mathbf{y}$ is a genomic segment of length n from a certain species, say, human. The simplest probabilistic model for a genomic segment is the "iid model" in which every observed DNA base pair (bp) in the segment is regarded as *independent* of others and produced randomly by nature based on a roll of a 4-sided die (maybe loaded). Although very simple and unrealistic, this model is the so-called "null model" behind almost all theoretical analyses of popular biocomputing methods. That is, if we want to assess whether a pattern we find can be regarded as a "surprise," the most natural analysis is to evaluate how likely this pattern occurs if an iid model is assumed.

Finding a value of $\boldsymbol{\theta}$ that is most compatible with the observation $\mathbf{y}$ is termed as *model fitting* or *estimation.* We make scientific progresses by iterating between fitting the data to the posited model and proposing an improved model to accommodate important features of the data that are not accounted for by the previous model. When the model is given, an efficient method should be used to make inference on the parameters. Both the maximum likelihood estimation method and the Bayes method use the likelihood function to extract information from data and are efficient; these methods will be the main focus of the remaining part of this chapter.

2.2 Frequentist approach to statistical inference

The frequentist approach, or sometimes simply referred to as the classical statistics procedure, arrives at its inferential statements by using a point estimate of the unknown parameter and addressing the estimation uncertainty by the *frequency behavior* of the estimator. Among all estimation methods, the method of *maximum likelihood estimate* (MLE) is most popular.

The MLE of $\boldsymbol{\theta}$ is defined as an argument $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood function, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg \max_{\text{all } \boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \mathbf{y}),$$

where the *likelihood function* $L(\boldsymbol{\theta} \mid \mathbf{y})$ is defined to be *any* function that is proportional to the probability density $f(\mathbf{y} \mid \boldsymbol{\theta})$. Clearly, $\hat{\boldsymbol{\theta}}$ is a function of $\mathbf{y}$ and its form is determined completely by the parametric model $f(\ )$. Hence, we can write $\hat{\boldsymbol{\theta}}$ as $\hat{\boldsymbol{\theta}}(\mathbf{y})$ to explicate this connection. Any deterministic function of the data $\mathbf{y}$, such as $\hat{\boldsymbol{\theta}}(\mathbf{y})$, is called an *estimator.* For example, if $\mathbf{y} = (y_1, \ldots, y_n)$ are iid observations from $N(\theta, 1)$, a Normal distribution with mean θ and variance 1, then the MLE of θ is $\hat{\theta}(\mathbf{y}) = \bar{y}$, the sample mean of the y_i, which is a linear combination of the y. It can be shown that, under regularity conditions, the MLE $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is asymptotically most efficient among all potential estimators. In other words, no other way of using $\mathbf{y}$ can perform better asymptotically, in terms of estimating $\boldsymbol{\theta}$, than the MLE procedure. But some inferior methods, such as the method of moments (MOM), can be used as alternatives when the MLE is difficult to obtain.

Uncertainty in estimation is addressed by the *principle of repeated sampling.* Imagine that the same stochastic process that "generates" our observation $\mathbf{y}$ can be repeated indefinitely under identical conditions. A frequentist studies what the "typical" behavior of an estimator, e.g., $\hat{\theta}(\mathbf{y}_{\text{rep}})$, is. Here $\mathbf{y}_{\text{rep}}$ denotes a hypothetical dataset generated by a replication of the *same* process that generates $\mathbf{y}$ and is, therefore, a random variable that has $\mathbf{y}$'s characteristics. The distribution of $\hat{\theta}(\mathbf{y}_{\text{rep}})$ is called the *frequency behavior* of estimator $\hat{\theta}$. For the Normal example, the frequency distribution of $\bar{y}_{\text{rep}}$ is $N(\theta, 1/n)$. With this distribution available, we can calibrate the observed $\hat{\theta}(\mathbf{y})$ with the "typical" behavior of $\hat{\theta}(\mathbf{y}_{\text{rep}})$, e.g., $N(\theta, 1/n)$, to quantify uncertainty in the estimation. As another example, suppose $\mathbf{y} = (y_1, \ldots, y_n)$ is a genomic segment and let n_a be the number of "A"s in $\mathbf{y}$. Then $\hat{\theta}_a = n_a/n$ is an estimator of θ_a, the "true frequency of A" under the iid die-rolling model. To understand the uncertainty in $\hat{\theta}_a$, we need to go back to the iid model and ask ourselves: how would n_a fluctuate in a segment like $\mathbf{y}$ that is generated by the same die-rolling process? The answer is rather simple: n_a follows distribution $\text{Binom}(n, \theta_a)$ and has mean $n\theta_a$ and variance $n\theta_a(1 - \theta_a)$.

We want to emphasize that the concepts of an "estimator" and its uncertainty only make sense if a generative model is contemplated. For example, the statement that "$\hat{\theta}_a$ estimates the true frequency of A" only makes sense if we imagine that an iid model (or another similar model) was used to generate the data. If this model is not really what we have in mind, then the meaning of $\hat{\theta}_a$ is no longer clear. A imaginary random process for the data generation is crucial for deriving a valid statistical statement.

A $(1-\alpha)100\%$ confidence interval (or region) for $\boldsymbol{\theta}$, for instance, is of the form $(\underline{\boldsymbol{\theta}}(\mathbf{y}_{\text{rep}}), \overline{\boldsymbol{\theta}}(\mathbf{y}_{\text{rep}}))$, meaning that under repeated sampling, the probability that the interval (the interval is random under repeated sampling) covers the true θ is at least $1-\alpha$. In contrast to what most people have hoped for, this interval statement *does not* mean that "$\boldsymbol{\theta}$ is in $(\underline{\boldsymbol{\theta}}(\mathbf{y}), \overline{\boldsymbol{\theta}}(\mathbf{y}))$ with probability $1-\alpha$." With observed $\mathbf{y}$, the true $\boldsymbol{\theta}$ is either in or out of the interval and no meaningful direct probability statement can be given unless $\boldsymbol{\theta}$ can be treated as a random variable.

When finding the analytical form of the frequency distribution of an estimator $\hat{\boldsymbol{\theta}}$ is difficult, some modern techniques such as the *jackknife* or *bootstrap* method can be applied to numerically simulate the "typical" behavior of an estimator (Efron, 1979). Suppose $\mathbf{y} = (y_1, \dots, y_n)$ and each y_i follows an iid model. In the bootstrap method, one treats the empirical distribution of $\mathbf{y}$ (the distribution that gives a probability mass of $1/n$ to each y_i and 0 to all other points in the space) as the "true underlying distribution" and repeatedly generates new datasets, $\mathbf{y}_{\text{rep},1}, \dots, \mathbf{y}_{\text{rep},B}$ from this distribution. Operationally, each $\mathbf{y}_{\text{rep},b}$ consists of n data points, i.e., $\mathbf{y}_{\text{rep},b} = (y_{b,1}, \dots, y_{b,n})$, where each $y_{b,i}$ is a simple random sample (with replacement) from the set of the observed data points $\{y_1, \dots, y_n\}$. With the bootstrap samples, we can calculate $\hat{\theta}(\mathbf{y}_{\text{rep},b})$, for $b = 1, \dots, B$, whose histogram tells us how $\hat{\theta}$ varies from sample to sample assuming that the true distribution of $\mathbf{y}$ is its observed empirical distribution.

In a sense, the classical inferential statements are *pre-data* statements because they are concerned with the repeated sampling properties of a procedure and do not have to refer to the actual observed data (except in the bootstrap method where the observed data is used in the approximation of the "true underlying distribution"). A major difficulty in the frequentist approach, besides its awkwardness in quantifying estimation uncertainty, is its difficulty in dealing with nuisance parameters. Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2)$. In a problem where we are only interested in one component, θ_1 say, the other component θ_2 becomes a *nuisance parameter*. No clear principles exist in classical statistics that enables us to eliminate θ_2 in an optimal way. One of the most popular practices in statistical analysis is the so-called *profile likelihood* method, in which one treats the nuisance parameter θ_2 as known and fixes it at its MLE. This method, however, underestimates the involved uncertainty (because it treats unknown θ_2 as if it were known) and can lead to incorrect inference when the distribution of $\hat{\theta}_1$ depends on θ_2, especially the dimensionality of θ_2 is high. More sophisticated methods based on orthogonality, similarity, and average likelihood have also been proposed, but they all have their own problems and limitations.

2.3 Bayesian methodology

Bayesian statistics seeks a more ambitious goal by modeling all related information and uncertainty, e.g., physical randomness, subjective opinions, prior knowledge from different sources, etc., with a joint probability distribution and treating *all quantities* involved in the model, be they observations, missing data, or unknown parameters, as random variables. It uses the calculus of probability as the guiding principle in manipulating data and derives its inferential statements based purely on an appropriate conditional distribution of unknown variables.

Instead of treating $\boldsymbol{\theta}$ as an unknown constant as in a frequentist approach, Bayesian analysis treats $\boldsymbol{\theta}$ as a realized value of a random variable that follows a *prior distribution* $f_0(\boldsymbol{\theta})$, which is typically regarded as known to the researcher independently of the data under analysis. The Bayesian approach has at least two advantages. Firstly, through the prior distribution we can inject prior knowledge and information about the value of $\boldsymbol{\theta}$. This is especially important in bioinformatics, since biologists often have substantial knowledge about the subject under study. To the extent that this information is correct, it will sharpen the inference about $\boldsymbol{\theta}$. Secondly, treating all the variables in the system as random variables greatly clarifies the methods of analysis. It follows from the basic probability theory that information about the realized value of any random variable, $\boldsymbol{\theta}$, say, based on observation of related random variables, $\mathbf{y}$, say, is summarized in the conditional distribution of $\boldsymbol{\theta}$ given $\mathbf{y}$, the so-called *posterior distribution.* Hence, if we are interested only in a component of $\boldsymbol{\theta} = (\theta_1, \theta_2)$, say θ_1, we have just to integrate out the remaining components of $\boldsymbol{\theta}$, i.e., nuisance parameters, from the posterior distribution. Furthermore, if we are interested in the prediction of a future observation $\mathbf{y}^+$ depending on $\boldsymbol{\theta}$, we can obtain the posterior distribution of $\mathbf{y}^+$ given $\mathbf{y}$ by completely integrating out $\boldsymbol{\theta}$.

The use of probability distributions to describe unknown quantities is also supported by the fact that probability theory is the only known coherent system for quantifying objective and subjective uncertainties. Furthermore, probabilistic models have been accepted as appropriate in almost all information-based technologies including: information theory, control theory, system science, communication and signal processing, and statistics. When the system under study is modeled properly, the Bayesian approach is coherent, consistent, and efficient.

The theorem that combines the prior and the data to form the posterior distribution (Section 3) is a simple mathematical result first given by Thomas Bayes in 1763. The statistical procedure based on the systematic use of this theorem appears much later (some people believe that Laplace was the first *Bayesian*) and is also named after Bayes. The adjective *Bayesian* is often used for approaches in which subjective probabilities are emphasized. In this sense Thomas Bayes was not really a Bayesian.

A main controversial aspect of the Bayesian approach is the use of the prior distribution, to which three interpretations can be given: (a) as frequency distributions; (b) as objective representations of a rational belief of the parameter, usually in a state of ignorance; and (c) as a subjective measure of what a particular individual believes (Cox and Hinkley 1974). Interpretation (a) refers to the case when $\boldsymbol{\theta}$ indeed follows a stochastic process and, therefore, is uncontroversial. But this scenario is of limited applicability. Interpretation (b) is theoretically interesting but is often untenable in real applications. The emotive words "subjective" and "objective" should not be taken too seriously. (Many people regard the frequentist approach as a more "objective" one.) There are considerable subjective elements and personal judgements injected into all phases of scientific investigations. Claiming that someone's procedure is "more objective" based on how the procedure is derived is nearly meaningless. A truly objective evaluation of any procedure is by how well it attains its stated goals. In bioinformatics, we are fortunate to have a lot of known biological facts to serve as objective judges.

In most of our applications, we employ the Bayesian method mainly because of its internal consistency in modeling and analysis and its capability to combine various sources of information.

Thus, we often take a combination of (a) and (c) for deriving a "reasonable" prior for our data analysis. We advocate the use of a suitable sensitivity analysis, i.e., an analysis on how our inferential statements are influenced by a change in the prior, to validate our statistical conclusions.

2.4 Connection with some methods in bioinformatics

Nearly all bioinformatics methods employ score functions, which are often functions of likelihoods or likelihood ratios, at least implicitly. The specification of priors required for Bayesian statistics is less familiar in bioinformatics although not completely foreign. For example, the setting of parameters for an alignment algorithm can be viewed as a special case of prior specification in which the prior distribution is degenerate with probability one for the set value and zero for all other values. The introduction of non-degenerate priors can typically give more flexibility in modeling reality.

The use of formal statistical models in bioinformatics was relatively rare before the 1990s. One reason is perhaps that computer scientists, statisticians, and other data analysts were not comfortable with big models — it is hard to think about many unknowns simultaneously. Additionally, algorithms for dealing with complex statistical models were not sufficiently well known and the computer hardware was not yet as powerful. Recently, an extensive use of probabilistic models (e.g., the hidden Markov model and the missing data formalism) has contributed greatly to the advance of computational biology.

Recursive algorithms for global optimization have been employed with great advantage in bioinformatics as the basis of a number of dynamic programming algorithms. We show that these algorithms have very similar counterparts in Bayesian and likelihood computations.

3 Bayes Procedure

3.1 The joint and posterior distributions

The full process of a typical Bayesian analysis can be described as consisting of three main steps (Gelman et al. 1995): (a) setting up a full probability model, the *joint distribution*, that captures the relationship among *all* the variables (e.g., observed data, missing data, unknown parameters) in consideration; (b) summarizing the findings for particular quantities of interest by appropriate posterior distributions, which is typically a conditional distribution of the quantities of interest given the observed data; and (c) evaluating the appropriateness of the model and suggesting improvements (model criticism and selection).

A standard procedure for carrying out step (a) is to formulate the scientific question of interest though the use of a probabilistic model, from which we can write down the *likelihood function* of $\boldsymbol{\theta}$. Then a prior distribution $f_0(\boldsymbol{\theta})$ is contemplated, which should be both mathematically tractable and scientifically meaningful. The joint probability distribution can then be represented as $Joint = likelihood \times prior$, i.e.,

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}) f_0(\boldsymbol{\theta}), \tag{1}$$

For notational simplicity, we use $p(\mathbf{y} \mid \boldsymbol{\theta})$, hereafter, interchangeably with $f(\mathbf{y} \mid \boldsymbol{\theta})$ to denote the likelihood. From a Bayesian's point of view, this is simply a conditional distribution.

Step (b) is completed by obtaining the *posterior distribution* through the application of Bayes theorem:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}) f_0(\boldsymbol{\theta})}{\int p(\mathbf{y} \mid \boldsymbol{\theta}) f_0(\boldsymbol{\theta}) d\boldsymbol{\theta}} \propto p(\mathbf{y} \mid \boldsymbol{\theta}) f_0(\boldsymbol{\theta}) \tag{2}$$

When $\boldsymbol{\theta}$ is discrete, the integral is replaced by summation. The denominator $p(\mathbf{y})$, which is a normalizing constant for the function, is sometimes called the *marginal likelihood* of the model and can be used to conduct model selection (Kass and Raftery, 1995). Although evaluating $p(\mathbf{y})$ analytically is infeasible in many applications, Markov chain Monte Carlo methods (Section 4) often can be employed for its estimation.

In computational biology, because the data to be analyzed are usually categorical (e.g., DNA sequences with a 4-letter alphabet or protein sequences with a 20-letter alphabet), the *multinomial distribution* is most commonly used. The parameter vector $\boldsymbol{\theta}$ in this model corresponds to the frequencies of each base type in the data. A mathematically convenient prior distribution for the multinomial families is the *Dirichlet distributions*, of which the Beta distribution is a special case for the binomial family. This distribution has the form

$$f_0(\boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \tag{3}$$

where k is the size of the alphabet and $\alpha_j > 0$ for all j. Here $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ is often called the hyper-parameter for the Dirichlet distribution and the sum $\alpha = \alpha_1 + \cdots + \alpha_k$ is often called the "pseudo-counts," which can be understood heuristically as the total "worth" (in comparison with actual observations) of one's prior opinion. When a simple iid model is imposed on an observed sequence of letters, $\mathbf{y} = (y_1, \ldots, y_n)$, its likelihood function is

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \theta_{y_i} = \prod_{j=1}^{k} \theta_j^{n_j},$$

where n_j is the number of counts of residual type j in $\mathbf{y}$. If a Dirichlet ($\boldsymbol{\alpha}$) prior used for its parameter $\boldsymbol{\theta}$, the posterior distribution for $\boldsymbol{\theta}$ is simply another Dirichlet distribution with hyperparameter $(\alpha_1 + n_1, \ldots, \alpha_k + n_k)$. The posterior mean of, say, θ_j, is $(n_j + \alpha_j)/(n + \alpha)$.

Suppose the parameter vector has more than one component, i.e., $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_{[-1]})$, where $\boldsymbol{\theta}_{[-1]}$ denotes all but the first component. One may be interested only in one of components, θ_1, say. The other components that are not of immediate interest but are needed by the model, *nuisance parameters*, can be removed by integration:

$$p(\theta_1 \mid \mathbf{y}) = \frac{p(\mathbf{y}, \theta_1)}{p(\mathbf{y})} = \frac{\int p(\mathbf{y} \mid \theta_1, \boldsymbol{\theta}_{[-1]}) f_0(\theta_1, \boldsymbol{\theta}_{[-1]}) d\boldsymbol{\theta}_{[-1]}}{\iint p(\mathbf{y} \mid \theta_1, \boldsymbol{\theta}_{[-1]}) f_0(\theta_1, \boldsymbol{\theta}_{[-1]}) d\theta_1 d\boldsymbol{\theta}_{[-1]}} \tag{4}$$

Note that computations required for completing a Bayesian inference are the integrations (or summations for discrete parameters) over all unknowns in the joint distribution to obtain the marginal likelihood and over all but those of interest to remove nuisance parameters. Despite the deceptively simple-looking form of (4), the challenging aspects of Bayesian statistics are:

(i) the development of a model, $p(\mathbf{y} \mid \boldsymbol{\theta}) f_0(\boldsymbol{\theta})$, which must effectively capture the key features of the underlying scientific problem; and (ii) the necessary computation for deriving the posterior distribution. For aspect (i), the *missing data* formulation is an important tool to help one formulate a scientific problem; for (ii), the recent advances in Markov chain Monte Carlo techniques are essential.

3.2 The missing data framework

The missing data formulation is an important methodology for modeling complex data structures and for designing computational strategies. This general framework was motivated in early 1970s (and maybe earlier) by the need for a proper statistical analysis of certain survey data where parts of the data were missing. For example, a large survey of families was conducted in 1967 in which many socioeconomic variables were recorded. Then a follow-up study of the same families were done in 1970. Naturally, the 1967 data had a large amount of missing values due to either recording errors or some families' refusal to answer certain questions. The 1970 data had an even more severe kind of missing data caused by the fact that many families studied in 1967 could not be located in 1970.

The first important question for a missing data problem is under what conditions can ignore the "missing mechanism" in the analysis. That is, does the fact that an observation is missing tell us anything about the quantities we are interested in estimate? For example, the fact that many families moved out of a particular region may indicate that the region's economy was having problem. Thus, if our interested estimand is a certain "consumer confidence" measure of the region, the standard estimate resulting only from the observed families might be biased. Rubin (1976)'s pioneering work provides general guidance on how to judge the ignorability. Since everything in a Bayes model is a random variable, it is especially convenient and transparent in dealing with these ignorability problems in a Bayesian framework. The second important question is that how one should conduct computations, such as finding the MLE or the posterior distribution of the estimands. This question has motivated statisticians to develop several important algorithms: the EM algorithm (Dempster, Laird, and Rubin, 1977), data augmentation

(Tanner and Wong, 1987), and the Gibbs sampler (Gelfand and Smith, 1990).

In late 1970s and early 1980s, people started to realize that many other problems can be treated as a missing data problem. One typical example is the so-called *latent-class* model which is most easily explained by the following example (Tanner and Wong, 1987). In the 1972-74 General Social Surveys, a sample of 3,181 participants were asked to answer the following questions. Question A: whether or not you think it should be possible for a pregnant woman to obtain a legal abortion *if she is married and does not want any more children.* Question B: the italicized phrase in A is replaced with "if she is not married and does not want to marry the man." A latent-class model assumes that a person's answers to A and B are conditionally independent given the value of a dichotomous latent variable Z (either 0 or 1). Intuitively, this model asserts that the population consists of two "types" of persons (e.g., conservative and liberal) and Z is the unobserved "label" of each person. If you know the person's label, then his/her answer to question A will not help you to predict his/her answer to question B. Clearly, variable Z can be thought of as a "missing data" although it is not really "missing" in a standard sense. For another example, in a multiple sequence alignment problem, alignment variables that must be specified for each sequence (observation) can be regarded as missing data. Residue frequencies or scoring matrices, which apply to all the sequences are population parameters. This generalized view eventually made the missing data formulation one of the most versatile and constructive workbenchs for sophisticated statistical analysis and advanced statistical computing.

The importance of the missing data formulation stems from the following two main considerations. Conceptually, this framework helps in making model assumptions explicit (e.g., ignorable versus nonignorable missing mechanism), in defining precise estimands of interest, and in providing a logical framework for causal inference (Rubin, 1976). Computationally, the missing data formulation inspired the invention of several important statistical algorithms. Mathematically, however, the missing data formulation is not well defined. In real life what we can observe is always partial (incomplete) information and there is no absolute distinction between parameters and missing data (i.e., some unknown parameters can also be thought of as missing data and vice versa).

To a broader scientific audience, the concept of "missing data" is perhaps a little odd since many scientists may not believe that they have any missing data. In the most general and abstract form, the "missing data" can refer to any unobserved component of the probabilistic system under consideration and the inclusion of this part in the system often results in a simpler structure. This component, however, needs to be marginalized (integrated) out in the final analysis. That is, when missing data $\mathbf{y}_{\text{mis}}$ is present, a proper inference about the parameters of interest can be achieved by using the "observed-data likelihood," $L_{obs}(\theta; \mathbf{y}_{\text{obs}}) = p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta})$, which can be obtained by integration:

$$L_{\text{obs}}(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}}) \propto \int p(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \boldsymbol{\theta}) d\mathbf{y}_{\text{mis}}.$$

Since it is often difficult to compute this integral analytically, one needs advanced computational methods such as the EM algorithm (Dempster et al., 1977) to compute the MLE.

Bayesian analysis for missing data problems can be achieved coherently through integration. Let $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_{[-1]})$ and suppose we are interested only in θ_1. Then

$$p(\theta_1 \mid \mathbf{y}_{\text{obs}}) \propto \iint p(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \theta_1, \boldsymbol{\theta}_{[-1]}) p(\theta_1, \boldsymbol{\theta}_{[-1]}) d\mathbf{y}_{\text{mis}} d\boldsymbol{\theta}_{[-1]}.$$

Since all quantities in a Bayesian model are treated as random variables, the integration for eliminating the missing data is no different than that for eliminating nuisance parameters.

Our main use of the missing data formulation is to construct proper statistical models for bioinformatics problems. As will be shown in the later sections, this framework frees us from being afraid of introducing meaningful but perhaps high-dimensional variables into our model, which is often necessary for a satisfactory description of the underlying scientific knowledge. The extra variables introduced this way, when treated as missing data, can be integrated out in the analysis stage so as to result in a proper inference for the parameter of interest. Although a conceptually simple procedure, the computation involved in integrating out missing data can be very difficult. Section 4 introduces a few algorithms for this purpose.

3.3 Model selection and Bayes evidence

At times biology may indicate that more than one model are plausible. Then we are interested in assessing model fit and conducting model selection (Step (c) described in Section 2.1). Classical hypothesis testing can be seen as a model selection method in which one chooses between the null hypothesis and the alternative in light of data. Model selection can also be achieved by a formal Bayes procedure. Firstly, all the candidate models are embedded into one unified model. Then the "overall" posterior probability of each candidate model is computed and used to discriminate among the models (Kass and Raftery, 1995).

To illustrate the Bayes procedure for model selection, we focus on the comparison of two models: $M = 0$ indicates the "null" model, and $M = 1$ the alternative. The joint distribution for the *augmented model* becomes

$$p(\mathbf{y}, \boldsymbol{\theta}, M) = p(\mathbf{y} \mid \boldsymbol{\theta}, M)p(\boldsymbol{\theta}, M).$$

Under the assumption that the data depend on the models through their respective parameters, the above equation is equal to

$$p(\mathbf{y}, \boldsymbol{\theta}, M) = p(\mathbf{y} \mid \boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m \mid M = m)p(M = m),$$

where $p(\boldsymbol{\theta}_m \mid M = m)$ is the prior for the parameters in model m, and $p(M = m)$ is the prior probability of model m. Note that the dimensionality of $\boldsymbol{\theta}_m$ may be different for different m. The posterior probability for model m is obtained as:

$$\begin{aligned} p(M = m \mid \mathbf{y}) &\propto p(\mathbf{y} \mid M = m)p(M = m) \\ &= \left\{ \int p(\mathbf{y} \mid \boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m \mid M = m)d\boldsymbol{\theta}_m \right\} p(M = m) \end{aligned}$$

The choice of $p(M = m)$, which is our prior on different models, is assigned independently of data in study. A frequent choice is $p(M = 0) = p(M = 1) = 0.5$ if we expect that both models are equally likely *a priori*. But in other cases we might set $p(M = 1)$ very small. For example, in the context of database searching, the prior probability that the query sequence is related to a sequence taken at random from the database is much smaller. In this case we might set $p(M = 1)$ inversely proportional to the number of sequences in the database.

4 Advanced Computation in Statistical Analysis

In many practical problems, the required computation is the main obstacle for applying both the Bayesian and the MLE methods. In fact, until recently these computations have often been so difficult that sophisticated statistical modeling and Bayesian methods were largely for theoreticians and philosophers. The introduction of the bootstrap method (Efron, 1979), the expectation-maximization (EM) algorithm (Dempster et al., 1977), and the Markov chain Monte Carlo (MCMC) methods (Gilks et al., 1998) has brought many powerful statistical models into the mainstream of statistical analysis. As we illustrate in Section 5, by appealing to the rich history of computation in bioinformatics, many required optimizations and integrations can be done exactly, which gives rise to either an exact solution to the MLE and the posterior distributions or an improved MCMC algorithm.

4.1 The EM algorithm

The EM algorithm is perhaps one of the most well-known statistical algorithms for finding the mode of a *marginal* likelihood or posterior distribution function. That is, the EM algorithm enables one to find the mode of

$$F(\boldsymbol{\theta}) = \int f(\mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) d\mathbf{y}_{\text{mis}}, \tag{5}$$

where $f(\mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) \geq 0$ and $F(\boldsymbol{\theta}) < \infty$ for all $\boldsymbol{\theta}$. When $\mathbf{y}_{\text{mis}}$ is discrete, we simply replace the integral in (5) by summation. The EM algorithm starts with an initial guess $\boldsymbol{\theta}^{(0)}$ and iterates the following two steps:

- **E-step.** Compute

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E_t[\log f(\mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) \mid \mathbf{y}_{\text{obs}}] \\ &= \int \log f(\mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(t)}) d\mathbf{y}_{\text{mis}}, \end{aligned}$$

  where $f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) = f(\mathbf{y}_{\text{mis}}, \mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta})/F(\boldsymbol{\theta})$, the conditional distribution of $\mathbf{y}_{\text{mis}}$.

- **M-step.** Find $\boldsymbol{\theta}^{(t+1)}$ to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

The E-step is derived from an "imputation" heuristic. Since we assume that the log-likelihood function is easy to compute once the missing data $\mathbf{y}_{\text{mis}}$ is given, it is appealing to simply "fill-in" a set of missing data and conduct a complete-data analysis. However, the simple "fill-in" idea is incorrect because it underestimates the variability caused by the missing information. The correct approach is to average the log-likelihood over all the missing data. In general, the E-step considers all possible ways of filling in the missing data, computes the corresponding complete-data log-likelihood function, and then obtains $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ by averaging these functions according to the current "predictive density" of the missing data. The M-step then finds the maximum of the Q function.

It is instructive to consider the EM algorithm for the latent-class model of Section 3.2. The observed values are $\mathbf{y}_{\text{obs}} = (y_1, \ldots, y_n)$, where $y_i = (y_{i1}, y_{i2})$ and y_{ij} is the ith person's answer to jth question. The missing data are $\mathbf{y}_{\text{mis}} = (z_1, \ldots, z_n)$, where z_i is the latent-class label of person i. Let $\boldsymbol{\theta} = (\theta_{0,1}, \theta_{1,1}, \theta_{0,2}, \theta_{1,2}, \gamma)$, where γ is the frequency of $z_i = 1$ in the population and $\theta_{k,l}$ is the probability of a type-k person saying "yes" to the lth question. Then the complete-data likelihood is

$$f(\mathbf{y}_{\text{mis}}, \boldsymbol{\theta}) = p(\mathbf{y}_{\text{obs}}|\mathbf{y}_{\text{mis}}, \boldsymbol{\theta})p(\mathbf{y}_{\text{mis}}|\boldsymbol{\theta}) = \prod_{i=1}^{n}\left[\prod_{k=1}^{2}\left\{\theta_{z_i,k}^{y_{ik}}(1-\theta_{z_i,k})^{1-y_{ik}}\right\}\gamma^{z_i}(1-\gamma)^{1-z_i}\right].$$

The E-step requires us to average over all label imputations. Thus, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is equal to

$$E_t\left[\sum_{i=1}^{n}\left\{\sum_{k=1}^{2}\{y_{ik}\log\theta_{z_i,k} + (1-y_{ik})\log(1-\theta_{z_i,k})\} + z_i\log\gamma + (1-z_i)\log(1-\gamma)\right\}\,\middle|\,\mathbf{y}_{\text{obs}}\right],$$

where the expectation sign means that we need to average out each z_i according to its "current" predictive probability distribution

$$\tau_i \equiv p(z_i = 1 \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(t)}) = \frac{\gamma^{(t)}\theta_{1y_i}^{(t)}}{\gamma^{(t)}\theta_{1y_i}^{(t)} + (1-\gamma^{(t)})\theta_{0y_i}^{(t)}}.$$

Hence, in the E-step, we simply "fill-in" a probabilistic label for each person, which gives

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{m=0}^{1}\sum_{k=1}^{2}\left(\sum_{i:\ y_{ik}=1}\tau_i^m(1-\tau_i)^{1-m}\log\theta_{m,k} + \sum_{i:\ y_{ik}=0}\tau_i^m(1-\tau_i)^{1-m}\log(1-\theta_{m,k})\right) \\ &+ \left(\sum_{i=1}^{n}\tau_i\right)\log\gamma + \left(\sum_{i=1}^{n}(1-\tau_i)\right)\log(1-\gamma). \end{aligned}$$

Although the above expression looks overwhelming, it is in fact quite simple and the M-step simply updates the parameters as $\gamma^{(t+1)} = \sum_{i=1}^{n} \tau_i / n$ and

$$\theta_{m,k}^{(t+1)} = \frac{\sum_{i:\ y_{ik}=1} \tau_i^m (1-\tau_i)^{1-m}}{\sum_{i:\ y_{ik}=1} \tau_i^m (1-\tau_i)^{1-m} + \sum_{i:\ y_{ik}=0} \tau_i^m (1-\tau_i)^{1-m}}.$$

There are three main advantages of the EM algorithm: (a) it is numerically stable (no inversion of a Hessian matrix); (b) each iteration of the algorithm strictly increases the value of the objective function unless it has reached a local optima; (c) each step of the algorithm has an appealing statistical interpretation. For example, the E-step can often be seen as "imputing" the missing data and the M-step can be viewed as the estimation of the parameter value in lights of the current imputation. The idea of iterating between "filling-in" missing data and updating estimate of the parameter has been around much longer than the EM algorithm. But Dempster et al. (1977) provided the first general and mathematically correct formulation of this intuitive idea. See Meng and van Dyk (1997) and discussions therein for an overview of recent advances of the EM algorithm.

4.2 Monte Carlo and Bayesian Analysis

As we have mentioned previously, the Bayesian analysis of a statistical problem can be made based on the joint posterior distribution of *all* unknown variables:

$$p(\boldsymbol{\theta}, \mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}) = \frac{p(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \boldsymbol{\theta}) f_0(\boldsymbol{\theta})}{\iint p(\mathbf{y}_{\text{obs}}, \mathbf{y}'_{\text{mis}} \mid \boldsymbol{\theta}') f_0(\boldsymbol{\theta}') d\mathbf{y}'_{\text{mis}} d\boldsymbol{\theta}'}. \tag{6}$$

Note that this joint distribution is almost completely known — except for the denominator, which is often called the *normalizing constant* (or the *partition function* in physics). Suppose, for example, we are interested only in estimating the first component θ_1 of $\boldsymbol{\theta}$, say. We may need to evaluate its posterior mean (and perhaps other characteristics):

$$E(\theta_1 \mid \mathbf{y}) = \iint \theta_1 p(\boldsymbol{\theta}, \mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}) d\mathbf{y}_{\text{mis}} d\boldsymbol{\theta} = \frac{\iint \theta_1 p(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}} \mid \boldsymbol{\theta}) f_0(\boldsymbol{\theta}) d\mathbf{y}_{\text{mis}} d\boldsymbol{\theta}}{\iint p(\mathbf{y}_{\text{obs}}, \mathbf{y}'_{\text{mis}} \mid \boldsymbol{\theta}') f_0(\boldsymbol{\theta}') d\mathbf{y}'_{\text{mis}} d\boldsymbol{\theta}'}. \tag{7}$$

Neither the numerator nor the denominator in (7) is easy to compute in practice.

If, however, we can generate a random sample $(\mathbf{y}_{\text{mis}}^{(1)}, \boldsymbol{\theta}^{(1)}), \ldots, (\mathbf{y}_{\text{mis}}^{(m)}, \boldsymbol{\theta}^{(m)})$, either independently or dependently (as in a Markov chain), from the joint posterior distribution (6), then we can approximate the marginal posterior distribution of θ_1 by the histogram of the first component, $\theta_1^{(j)}$, of each $\boldsymbol{\theta}^{(j)}$, and approximate (7) by the Monte Carlo sample average

$$\tilde{\theta}_1 = \frac{1}{m}\left(\theta_1^{(1)} + \cdots + \theta_1^{(m)}\right). \tag{8}$$

4.3 Simple Monte Carlo techniques

To begin with basic ideas, we describe two simple algorithms for generating random variables from a given distribution. As a starting point, we assume that independent *uniform* (in region [0,1]) random variables can be produced satisfactorily. Algorithms that serve this purpose are called *random number generators.* In fact, this task is not as simple it looks and the interested reader is encouraged to study further on this topic (Marsaglia and Zaman, 1993).

Inversion method. When we have available the cumulative distribution function (cdf) for a one-dimensional target distribution $\pi(\mathbf{x})$, we can implement the following procedure.

- Draw $U \sim$ Unif [0,1]
- Compute $\mathbf{x} = F^{-1}(U)$, where $F^{-1}(u) = \inf\{x;\ F(x) \geq u\}$.

Then $\mathbf{x}$ such produced must follow π. The interested reader can try to prove this fact. However, because many distributions (e.g., Gaussian) do not have a closed-form cdf, it is often difficult to directly apply the inversion method. To overcome this difficulty, von Neumann (1951) invented the ingenious *rejection method* which can be applied very generally.

Rejection method. Suppose $l(\mathbf{x}) = c\pi(\mathbf{x})$ is known (but c may be unknown) and we can find a sampling distribution $g(\mathbf{x})$ together with a constant M such that the envelope property, i.e., $Mg(\mathbf{x}) \geq l(\mathbf{x})$ for all $\mathbf{x}$, is satisfied. Then we can apply the following procedure.

(a) Draw $\mathbf{x} \sim g(\mathbf{x})$ and compute the ratio $r = l(\mathbf{x})/Mg(\mathbf{x})$ (which should always be ≤ 1);

(b) Draw $U \sim$ Unif[0,1]; accept and return $\mathbf{x}$ if $U \leq r$; reject $\mathbf{x}$ and go back to (a) if $U > r$.

To show that the accepted sample follows distribution π, we let I be the indicator function so that $I = 1$ if sample $\mathbf{X}$ drawn from $g(\ )$ is accepted, and $I = 0$, otherwise. Thus,

$$p(I=1) = \int p(I=1 \mid \mathbf{X}=\mathbf{x}) g(\mathbf{x}) d\mathbf{x} = \int \frac{c\pi(\mathbf{x})}{Mg(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = \frac{c}{M},$$

and

$$p(\mathbf{x} \mid I=1) = \frac{c\pi(\mathbf{x})}{Mg(\mathbf{x})} g(\mathbf{x}) / p(I=1) = \pi(\mathbf{x}).$$

It is seen that the "success rate" for obtaining an accepted sample is c/M. Thus, the key to a successful application of the algorithm is to find a good trial distribution $g(\mathbf{x})$ which gives rise to a small M. Since it is usually difficult to find a good g-function in high-dimensional problems, the rejection method alone tends to be not very useful in difficult problems.

4.4 Markov chain Monte Carlo methods

Markov chain Monte Carlo (MCMC) is a class of algorithms for simulating random variables from a target distribution, $\pi(\mathbf{x})$, given up to a normalizing constant. A major advantage of these algorithms is their ability to "divide-and-conquer" a high-dimensional and complex problem. These algorithms serve our purpose well because in Bayesian analysis we want to draw random samples from the joint posterior distribution (6) without having to know its denominator. The basic idea behind MCMC algorithms is to design and simulate a Markov chain whose equilibrium distribution is exactly $\pi(\mathbf{x})$. Here we describe two methods for constructing such chains — the Metropolis algorithm and the Gibbs sampler — both being widely used in diverse fields. More versatile algorithms and their analyses can be found in Liu (2001).

Metropolis-Hastings Algorithm. Let $\pi(\mathbf{x}) = c\exp\{-h(\mathbf{x})\}$ be the target distribution with unknown constant c. Metropolis et al. (1953) introduced the fundamental idea of Markov chain sampling and prescribed the first general construction of such a chain. Hastings (1970) later provided an important generalization. Starting with any configuration $\mathbf{x}^{(0)}$, the M-H algorithm evolves from the current state $\mathbf{x}^{(t)} = \mathbf{x}$ to the next state $\mathbf{x}^{(t+1)}$ as follows:

- Propose a new state $\mathbf{x}'$ which can be viewed as a small and random "perturbation" of the current state. More precisely, $\mathbf{x}'$ is generated from a *proposal* function $T(\mathbf{x}^{(t)} \to \mathbf{x}')$ (i.e., it is required that $T \geq 0$ and $\sum_{\text{all } \mathbf{y}} T(\mathbf{x} \to \mathbf{y}) = 1$ for all $\mathbf{x}$) determined by the user.

- Compute the Metropolis ratio

$$r(\mathbf{x}, \mathbf{x}') = \frac{\pi(\mathbf{x}')T(\mathbf{x}' \to \mathbf{x})}{\pi(\mathbf{x})T(\mathbf{x} \to \mathbf{x}')} \tag{9}$$

- Generate a random number $u \sim$ Unif[0,1];

  - let $\mathbf{x}^{(t+1)} = \mathbf{x}'$ if $u \leq r(\mathbf{x}, \mathbf{x}')$;

  - let $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$ otherwise.

A more well-known form of the Metropolis algorithm is described as iterating the following steps: (a) a small random perturbation of the current configuration is made; (b) the "gain" (or loss) in an objective function (i.e., $-h(\mathbf{x})$) resulting from this perturbation is computed; (c) a random number U is generated independently; and (d) the new configuration is accepted if $\log(U)$ is smaller than or equal to the "gain," and is rejected otherwise. The well-known *simulated annealing* algorithm (Kirkpatrick, Gelatt and Vecchi, 1983) is built upon this basic Metropolis iteration by adding an adjustable exponential scaling parameter to the objective function (i.e., $\pi(\mathbf{x})$ is scaled to $\pi^{\alpha}(\mathbf{x})$ and $alpha \to 0$).

Metropolis et al. (1953) restricted their choices of the "perturbation" function to be the symmetric ones. That is, the chance of proposing $\mathbf{x}'$ from perturbing $\mathbf{x}$ is always equal to that of proposing $\mathbf{x}$ from perturbing $\mathbf{x}'$. Intuitively, this means that there is no "trend bias" at the proposal stage. Mathematically, this symmetry can be expressed as $T(\mathbf{x} \to \mathbf{x}') = T(\mathbf{x}' \to \mathbf{x})$. Hastings (1970) generalized the choice of T to all those that satisfies the property: $T(\mathbf{x} \to \mathbf{x}') > 0$ if and only if $T(\mathbf{x}' \to \mathbf{x}) > 0$. It is easy to see that the "actual" transition probability function resulting from the M-H transition rule is, for $\mathbf{x} \neq \mathbf{y}$,

$$A(\mathbf{x} \to \mathbf{y}) = T(\mathbf{x} \to \mathbf{y}) \min\{1, r(\mathbf{x}, \mathbf{y})\},$$

where $r(\mathbf{x}, \mathbf{y})$ is the Metropolis ratio as in (9). It is easy to see that

$$\pi(\mathbf{x}) A(\mathbf{x} \to \mathbf{y}) = \min\{\pi(\mathbf{x}) T(\mathbf{x} \to \mathbf{y}), \pi(\mathbf{y}) T(\mathbf{y} \to \mathbf{x})\},$$

which is a symmetric function in $\mathbf{x}$ and $\mathbf{y}$. Thus, the *detailed balance* condition

$$\pi(\mathbf{x}) A(\mathbf{x} \to \mathbf{y}) = \pi(\mathbf{y}) A(\mathbf{y} \to \mathbf{x})$$

is satisfied by A. This condition then implies that π is the *invariant* distribution for the Metropolis-Hastings transition. That is,

$$\int \pi(\mathbf{x}) A(\mathbf{x} \to \mathbf{y}) d\mathbf{x} = \pi(\mathbf{y}).$$

Heuristically, π can be seen as a "fixed point" under the M-H operation in the space of all distributions. It follows from the standard Markov chain theory that if the chain is *irreducible* (i.e., it is possible to go from anywhere to anywhere else in a finite number of steps), *aperiodic* (i.e., there is no parity problem), and not drifting away, then in the long run the chain will settle at its invariant distribution (Neal, 1993). The random samples so obtained eventually are like those drawn directly from π.

The Metropolis algorithm has been extensively used in statistical physics over the past 40 years and is the cornerstone of all MCMC techniques recently adopted and generalized in the statistics community. Another type of MCMC algorithm, the Gibbs sampler (Geman and Geman 1984), differs from the Metropolis algorithm in its extensive use of conditional distributions based on $\pi(\mathbf{x})$ for constructing Markov chain moves.

Gibbs sampler. Suppose $\mathbf{x} = (x_1, \ldots, x_d)$. In the Gibbs sampler, one randomly or systematically chooses a coordinate, say x_1, and then update its value with a new sample x_1' drawn from the conditional distribution $\pi(\cdot \mid \mathbf{x}_{[-1]})$, where $\mathbf{x}_{[-A]}$ refers to $\{x_j,\ j \in A^c\}$. Algorithmically, the Gibbs sampler can be implemented as follows:

Random Scan Gibbs sampler. Suppose currently $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots x_d^{(t)})$. Then

- Randomly select i from $\{1, \ldots, d\}$ according to a given probability vector $(\alpha_1, \ldots, \alpha_d)$.

- Let $x_i^{(t+1)}$ be drawn from the conditional distribution $\pi(\cdot \mid \mathbf{x}_{[-i]}^{(t)})$, and let $x_{[-i]}^{(t+1)} = \mathbf{x}_{[-i]}^{(t)}$.

Systematic Scan Gibbs sampler. Let the current state be $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots x_d^{(t)})$.

- For $i = 1, \ldots, d$, we draw $x_i^{(t+1)}$ from the conditional distribution

$$\pi(x_i \mid x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_d^{(t)}).$$

It is easy to check that *every* individual conditional update leaves π invariant. Suppose currently $\mathbf{x}^{(t)} \sim \pi$. Then $\mathbf{x}_{[-i]}^{(t)}$ follows its marginal distribution under π. Thus,

$$\pi(x_i^{(t+1)} \mid \mathbf{x}_{[-i]}^{(t)}) \times \pi(\mathbf{x}_{[-i]}^{(t)}) = \pi(x_i^{(t+1)}, \mathbf{x}_{[-i]}^{(t)}),$$

implying that the joint distribution of $(\mathbf{x}_{[-i]}^{(t)}, x_i^{(t+1)})$ is unchanged at π after one update.

The Gibbs sampler's popularity in statistics community stems from its extensive use of *conditional distributions* in each iteration. Tanner and Wong (1987)'s data augmentation first linked the Gibbs sampling structure with missing data problems and the EM algorithm. Gelfand and Smith (1990) further popularized the method by pointing out that the conditionals needed in Gibbs iterations are commonly available in many Bayesian and likelihood computations.

Under regularity conditions, one can show that the Gibbs sampler chain converges geometrically and its convergence rate is related to how the variables correlate with each other. Therefore, grouping highly correlated variables together in the Gibbs update can greatly speed up the sampler (Liu, 1994).

Other techniques. A main problem with all MCMC algorithms is that they may, for some problems, move very slowly in the configuration space or may be trapped in the region of a local mode. This phenomena is generally called *slow-mixing* of the chain. When chain is slow-mixing, estimation based on the resulting Monte Carlo samples becomes very inaccurate. Some recent techniques suitable for designing more efficient MCMC samplers in bioinformatics applications include simulated tempering (Marinari and Parisi, 1992), parallel tempering (Geyer, 1991), multicanonical sampling (Berg and Neuhaus, 1992), multiple-try method (Liu et al., 2000), and evolutionary Monte Carlo (Liang and Wong, 2000). These and some other techniques are summarized in Liu (2001).

5 Compositional Analysis of a DNA Sequence

Suppose our observation is a DNA sequence, $R = (r_1, r_2, \ldots, r_n)$, and we are interested in understanding various aspects of it, such as its general compositions, dependence between neighboring base pairs, regions with different statistical characteristics (e.g., G-C rich regions), repeated short sequence patterns, and so on. In this and the next sections we show how progressively complex statistical models can be developed to address these scientific questions. Note that the problem setting is very general since a dataset of multiple sequences can always be regarded as a single "super-sequence" by joining all the individual sequences.

5.1 Multinomial modeling

The simplest statistical model for a DNA sequence is, as we discussed in Section 3.1, the iid multinomial model in which each r_i is assumed to be independently generated according to probability vector $\boldsymbol{\theta} = (\theta_a, \ldots, \theta_t)$. The likelihood function of $\boldsymbol{\theta}$ is then $L(\boldsymbol{\theta} \mid R) = \theta_a^{n_a} \cdots \theta_t^{n_t}$, where $\mathbf{n} = (n_a, \ldots, n_t)$ is the vector of counts of the 4 types of nucleotides. Vector $\hat{\boldsymbol{\theta}} = (n_a/n, \ldots, n_t/n)$ maximizes $L(\boldsymbol{\theta} \mid R)$ and is the MLE of $\boldsymbol{\theta}$. The distribution of $n\hat{\boldsymbol{\theta}}$ under hypothetical replications is Multinom$(n; \boldsymbol{\theta})$. Hence, for example, $n\hat{\theta}_a \sim$Binom(n, θ_a). Inverting this relationship gives us an approximate confidence interval for θ_a.

With a Dirichlet$(\boldsymbol{\alpha})$ prior (3), the posterior of $\boldsymbol{\theta}$ is Dirichlet$(\mathbf{n} + \boldsymbol{\alpha})$ and

$$E(\boldsymbol{\theta} \mid R) = \left(\frac{n_a + \alpha_a}{n + \alpha}, \ldots, \frac{n_t + \alpha_t}{n + \alpha} \right),$$

where $\alpha = \alpha_a + \cdots + \alpha_t$. This result is not that much different from the MLE. If one is interested in the posterior distribution of θ_a, say, an easy calculation gives us

$$\theta_a \mid R \sim \text{Beta}(n_a + \alpha_a, n + \alpha - n_a - \alpha_a).$$

5.2 Homogeneous Markov model

A natural next-step model is the *Markov* model, in which one assumes that the observed sequence follows a Markov chain with transition matrix $P(r_i \to r_{i+1}) = \theta_{r_i, r_{t+1}}$, where the parameter vector is a 4×4 matrix

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_{aa} & \dots & \theta_{at} \\ \vdots & \ddots & \vdots \\ \theta_{ta} & \dots & \theta_{tt} \end{pmatrix},$$

where each row sums to one. The MLE of each component, θ_{at}, say, in the parameter matrix is $n_{at}/n_{a\cdot}$, where n_{at} is the total count of neighboring AT pairs in the sequence and $n_{a\cdot} = n_{aa} + \cdots + n_{at}$ is the total count of A's excluding the first bp r_1. When a conjugate prior is used, a similar procedure to that for the multinomial model gives us the posterior distribution of $\boldsymbol{\theta}$, which is a product of four (one for each row of $\boldsymbol{\theta}$) independent Dirichlet distributions.

5.3 A hidden Markov model

Let us now consider a model that can accommodate compositional heterogeneity in DNA sequences. For this we can think of sequence R as consisting of different segments, and the sequence composition are homogeneous within each segment. Based on this heuristics, Liu and Lawrence (1999) proposed and analyzed a Bayesian segmentation model. Another model, as first proposed by Churchill (1989), is based on the HMM structure shown in Figure 1.

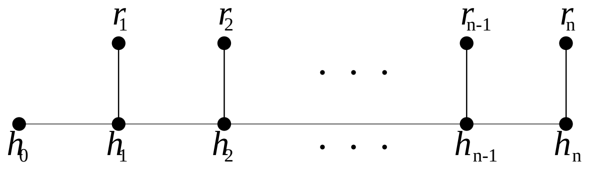

Figure 1: A graphical illustration of the hidden Markov model.

In this HMM model, we assume that the hidden layer $\mathbf{h} = (h_0, h_1, \dots, h_n)$ is a Markov chain Each h_i, for example, may have two possible states where $h_i = 0$ implies that the corresponding

r_i follows one compositional model, Multinom$(\boldsymbol{\theta}_0)$, and $h_i = 1$ indicates that $r_i \sim$ Multinom$(\boldsymbol{\theta}_1)$. Here $\boldsymbol{\theta}_k = (\theta_{ka}, \dots, \theta_{kt})$. A 2×2 transition matrix, $\boldsymbol{\tau} = (\tau_{kl})$, where $\tau_{kl} = P(h_i = k \to h_{i+1} = l)$, dictates the generation of $\mathbf{h}$. A similar model has been developed by Krogh et al. (1994) to predict protein coding regions in E. Coli genome.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\tau})$. The likelihood function of $\boldsymbol{\theta}$ under this HMM is

$$\begin{aligned} L(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\tau} \mid R) &= \sum_{\mathbf{h}} p(R \mid \mathbf{h}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) p(\mathbf{h} \mid \boldsymbol{\tau}), \\ &= \sum_{\mathbf{h}} p_0(h_0) \prod_{i=1}^{n} p(r_i \mid h_i, \boldsymbol{\theta}) p(h_i \mid h_{i-1}, \boldsymbol{\tau}) \\ &= \sum_{\mathbf{h}} p_0(h_0) \prod_{i=1}^{n} \theta_{h_i r_i} \tau_{h_{i-1} h_i} \end{aligned}$$

where h_0 is assumed to follow a known distribution p_0. For a given set of parameter values we can find the exact value of this likelihood function via a recursive summation method as described in (10) below.

However, finding the MLE of $\boldsymbol{\theta}$ is still nontrivial. One possible approach is to maximize L by a Newton-Raphson's method in which the first and the second derivatives of L can all be computed recursively. But this method may be unstable because $\boldsymbol{\theta}$'s dimensionality is a bit too high (the Hessian is a 9×9 matrix). A more stable alternative is the EM algorithm which involves iterations of the following two steps.

- *E-step.* Compute the Q-function:

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) &= E\left[\sum_{i=1}^{n} \log\{\theta_{h_i r_i} \tau_{h_{i-1} h_i}\} \,\middle|\, R, \boldsymbol{\theta}^{(t)} \right] \\ &= \sum_{i=1}^{n} \left[\sum_{h_i} \sum_{h_{i-1}} \left\{ \log \theta_{h_i r_i} + \log \tau_{h_{i-1} h_i} \right\} P_i^{(t)}(h_{i-1}, h_i) \right] \\ &= \sum_{k=0}^{1} \sum_{j=a}^{t} n_{kj}^{(t)} \log \theta_{kj} + \sum_{k=0}^{1} \sum_{l=0}^{1} m_{kl}^{(t)} \log \tau_{kl}. \end{aligned}$$

Here $P_i^{(t)}(h_{i-1}, h_i) = p(h_{i-1}, h_i \mid R, \boldsymbol{\theta}^{(t)})$, the marginal posterior distribution of (h_{i-1}, h_i) when the parameter takes value $\boldsymbol{\theta}^{(t)}$. This quantity can be obtained by using the B-

function defined in (16) and a procedure similar to the computation of (17). The $P_i^{(t)}$ can be derived by a recursive procedure similar to (10). The $n_{kj}^{(t)}$ and the $m_{kl}^{(t)}$ are the sums of the corresponding $P_i^{(t)}(k,l)$.

- *M-step.* Maximize the Q-function. It is obvious that the maximizer of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is
$$\theta_{kj}^{(t+1)} = n_{kj}^{(t)}/n_{k\cdot}^{(t)} \quad \text{and} \quad \tau_{kl}^{(t+1)} = m_{kl}^{(t)}/m_{k\cdot}^{(t)}$$
in which $n_{k\cdot}^{(t)} = n_{ka}^{(t)} + \cdots + n_{kt}^{(t)}$ and $m_{k\cdot}^{(t)} = m_{k0}^{(t)} + m_{k1}^{(t)}$.

To avoid being trapped at a singular point corresponding to zero count of certain base type, we may want to give a nonzero *pseudo* count to each type.

A Bayesian analysis of this problem is also feasible. With a prior distribution $f_0(\boldsymbol{\theta})$, which may be a product of three independent Dirichlet distributions, we have the joint posterior of all unknowns:

$$p(\boldsymbol{\theta}, \mathbf{h} \mid R) \propto p(R \mid \mathbf{h}, \boldsymbol{\theta}) p(\mathbf{h} \mid \boldsymbol{\theta}) f_0(\boldsymbol{\theta}).$$

In order to get the marginal posterior of $\boldsymbol{\theta}$, we may implement a special Gibbs sampler, data augmentation, which iterates the following steps:

- *Imputation:* draw $\mathbf{h}^{(t+1)} \sim p(\mathbf{h} \mid R, \boldsymbol{\theta}^{(t)})$;
- *Posterior Sampling:* draw $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta} \mid R, \mathbf{h}^{(t+1)})$.

The imputation step needs to draw a path, $\mathbf{h}$, from its *posterior* distribution with given parameter value. Its implementation requires a recursive method for summing up all the contributions from h_0 to h_n and then sampling backward. Thus, this method is very similar to dynamic programming and is sometimes called the *forward-backward* method. More precisely, this distribution can be written as

$$\begin{aligned} p(\mathbf{h} \mid R, \boldsymbol{\theta}) &= c\, p(\mathbf{h}, R \mid \boldsymbol{\theta}) = c\, p(R \mid \mathbf{h}, \boldsymbol{\theta}) p(\mathbf{h} \mid \boldsymbol{\theta}) \\ &= c\, p_0(h_0) \prod_{i=1}^{n} \{p(r_i|h_i) p(h_i|h_{i-1})\} = c\, p_0(h_0) \prod_{i=1}^{n} \left(\theta_{h_i r_i} \tau_{h_{i-1} h_i}\right), \end{aligned}$$

where c is the normalizing constant, i.e.,

$$c^{-1} = \sum_{\mathbf{h}} \left\{ p_0(h_0) \prod_{i=1}^{n} \left(\theta_{h_i r_i} \tau_{h_{i-1} h_i}\right) \right\}.$$

The key observation is that c, and also other required marginal distributions, can be computed exactly by a recursive method. Define $F_0(h) = p_0(h)$, and compute recursively

$$F_i(h) = \sum_{h_{i-1}=1}^{2} \{F_{i-1}(h_{i-1}) \tau_{h_{i-1} h} \theta_{h r_i}\}, \text{ for } h = 0, 1. \tag{10}$$

At the end of the recursion we obtain $c^{-1} = F_n(0) + F_n(1)$ and

$$p(h_n \mid R, \boldsymbol{\theta}) = \frac{F_n(h_n)}{F_n(0) + F_n(1)}. \tag{11}$$

In order to sample $\mathbf{h}$ properly, we draw h_n from distribution (11) and then draw h_i recursively backward from distribution

$$p(h_i \mid h_{i+1}, R, \boldsymbol{\theta}) = \frac{F_i(h_i) \tau_{h_i h_{i+1}}}{F_i(0) \tau_{0 h_{i+1}} + F_i(1) \tau_{1 h_{i+1}}}. \tag{12}$$

The posterior sampling step in the Gibbs sampler needs us to draw from the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{h}$ and R. This is a very simple task and only involves finding appropriate counts and sampling from the corresponding Dirichlet distributions. For example, $\boldsymbol{\theta}_0$ should be drawn from Dirichlet$(n_{0a} + \alpha_a, \ldots, n_{0t} + \alpha_t)$, where, n_{0a}, say, is the counts of the r_i whose type is A and whose hidden state h_i is zero.

5.4 HMM with more than two hidden states

It is straightforward to extend the previous 2-state HMM to a k-state HMM so as to analyze a sequence with regions of k different compositional types. In a k-state HMM, we will need a $k \times k$ transition matrix ($k(k-1)$ free parameters) to describe the transitions between the hidden Markov chain, and a probability vector $\boldsymbol{\theta}_j$ for each compositional type ($3k$ free parameters). The total number of free parameters is thus $k(k+2)$.

It is a non-trivial problem, however, to determine what value of k is proper for a given sequence R. A Bayesian model selection procedure as described in Section 3.3 can be applied to resolve this issue. More precisely, we introduce a *model variable* K. For given $K = k$, we can fit a k-state HMM to the sequence and obtain the *model likelihood*

$$p(R \mid K = k) = \iint p_k(R \mid \mathbf{h}, \boldsymbol{\theta}) p_k(\mathbf{h} \mid \boldsymbol{\theta}) f_k(\boldsymbol{\theta}) d\mathbf{h} d\boldsymbol{\theta},$$

where subscript k indicates that the employed distributions correspond to a k-state model. With a prior distribution $p_0(k)$ on K, we can derive the posterior distribution of K given the sequence. Although conceptually simple, this model selection procedure involves a difficult integral which is difficult to solve analytically. One often has to resort to some special MCMC methods designed for estimating the ratio of normalizing constants (Liu, 2001).

As an alternative to the HMM, Liu and Lawrence (1999) and Schmidler et al. (2000) describe a *segmentation* model based on the so-called *hidden semi-Markov model* (HSMM). Sequence segmentation models have been developed for many purposes in bioinformatics, including models for protein sequence hydrophobicity (Kyte and Dolittle, 1982; Auger and Lawrence, 1989), models for protein secondary structure (Schmidler et al., 2000), models for sequence complexity (Wootton, 1994), and models for gene identification (Snyder and Stormo, 1995; Burge and Karlin, 1997). What is common to all these methods is that a single sequence is characterized by a series of models which only involve local properties. That is, we assume in this model that the sequence can be segmented into m parts, where m is unknown, and each segment is described by a "local" model. An advantage of this model is that a Bayesian method for determining the number of segments m is relatively easy (Liu and Lawrence, 1999).

6 Find Repetitive Patterns in DNA Sequence

Similar to the objective of the previous section, our primary interest here is in the analysis of a single "super-sequence." Our focus, however, is one step further than the compositional analysis: we want to find repetitive motif elements in the sequence. The main motivation for this task is that repetitive patterns in biopolymer sequences often correspond to functionally or structurally important part of these molecules. For example, repetitive patterns in noncoding regions of DNA sequences may correspond to a "regulatory motif" to which certain regulatory protein binds so as to control gene expressions. The multiple occurrences of a regulatory motif in R is thus analogous to the multiple occurrences of a *word* in a long sentence. It is of interest to find out what this motif is and where it has occurred. What makes things worse, however, is that although the motif occurs in the sequence multiple times, no two occurrences are exactly identical. In other words, there are often some "typos" in each occurrence of the word. It is therefore rather natural for us to employ probabilistic models to handle this problem.

6.1 Block-motif model with iid background

A simple model that conveys the basic idea of a motif that repeats itself with random variations is the block-motif model as shown in Figure 2. It was first developed in Liu, Neuwald, and Lawrence (1995) and has been employed to find subtle repetitive patterns, such as helix-turn-helix structural motifs (Neuwald et al., 1995) or gene regulation motifs (Roth et al., 1998), in both protein and DNA sequences.

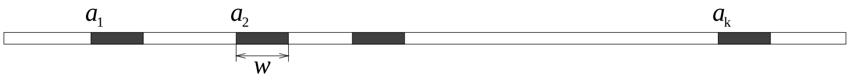

Figure 2: A graphical illustration of the repetitive motif model.

This model says that at unknown locations $A = (a_1, \dots, a_K)$ there are repeated occurrences of a motif. So the sequence segments at these locations should look similar to each other. In other part of the sequence, called the *background*, the residues follow an independent multinomial model. Suppose the motif's width is w, we need $w + 1$ probability vectors to describe the motif and the background: $\boldsymbol{\theta}_0 = (\theta_{0a}, \dots, \theta_{0t})$ describe the base frequencies in the background; and each $\boldsymbol{\theta}_k$ describes the base frequency at position k of the motif. The matrix $\Theta = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_w]$ is called the *profile* matrix for the motif. We again use the generic notation $\boldsymbol{\theta}$ to denote the collection of all parameters, $(\boldsymbol{\theta}_0, \Theta)$.

With a Dirichlet prior Dirichlet($\boldsymbol{\alpha}$), for all the $\boldsymbol{\theta}_i$, we can obtain the Bayes estimates of the $\boldsymbol{\theta}_i$ very easily *if* we know the positions of the motif. To facilitate analysis, we introduce an indicator vector $\boldsymbol{I} = (I_1, \ldots, I_n)$ and treat it as *missing data*. An $I_i = 1$ means that position i is the start of a motif pattern, and $I_i = 0$ means otherwise. We assume *a priori* each I_i has a small probability p_0 to be equal to 1. With this setup, we can write down the joint posterior distribution:

$$p(\boldsymbol{\theta}, \boldsymbol{I} \mid R) \propto p(R \mid \boldsymbol{I}, \boldsymbol{\theta}) p(\boldsymbol{I} \mid \boldsymbol{\theta}) f_0(\boldsymbol{\theta}) \tag{13}$$

where

$$p(\boldsymbol{I} \mid \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p_0^{I_i} (1 - p_0)^{1 - I_i}.$$

If we do not allow overlapping motifs, we need to restrict that in $\boldsymbol{I}$ there are no pair $I_i = 1$ and $I_j = 1$ with $i - j < w$. Since the motif region is a very small fraction of the whole sequence, we may estimate $\boldsymbol{\theta}_0$ based on the whole sequence and treat it as known.

A simple Gibbs sampler algorithm can be designed to draw from this joint posterior distribution (Liu et al. 1995). More specifically, we can iterate the following steps:

- For a current realization of $\boldsymbol{\theta}$, we update each I_i, $i = 1, \ldots, n$, by a random draw from its conditional distribution, $p(I_i \mid \boldsymbol{I}_{[-1]}, R, \boldsymbol{\theta})$, where

  $$\frac{p(I_i = 1 \mid \boldsymbol{I}_{[-i]}, R, \boldsymbol{\theta})}{p(I_i = 0 \mid \boldsymbol{I}_{[-i]}, R, \boldsymbol{\theta})} = \frac{p_0}{1 - p_0} \prod_{k=1}^{w} \left(\frac{\theta_{k, r_{i+k-1}}}{\hat{\theta}_{0 r_{i+k-1}}} \right). \tag{14}$$

  Intuitively, this odds ratio is simply the "signal-to-noise" ratio.

- Based on the current value of $\boldsymbol{I}$, we update the profile matrix Θ column-by-column. That is, each $\boldsymbol{\theta}_j$, $j = 1, \ldots, w$, is drawn from an appropriate posterior Dirichlet distribution determined by $\boldsymbol{I}$ and R.

After a burn-in period (until the Gibbs sampler stabilizes), we continue to run the sampler for m iterations and use (8) to estimate the profile matrix Θ. The estimated Θ can then be used to scan the sequence to find the locations of the motif.

6.2 Block-motif model with a Markovian background

Here the extra complication is that the motif can have a Markovian background. Thus, we need a 4×4 transition matrix, $B_0 = (\beta_{jj'})$, to describe the background. We also assume that the transition from the end of a motif to the next nonsite position follows the same Markov law. Since the total number of bp's that belong to a motif is a very small fraction of the total number of base pairs in R, we may estimate B_0 from the raw data directly, pretending that the whole sequence of R is homogeneous and governed by the transition matrix B_0. In this way, the transition probabilities can be estimated as $\hat{\beta}_{j_1 j_2} = n_{j_1 j_2}/n_{j_1}$, similar to that in Section 5.2. We may then treat B_0 as a known parameter. The joint posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{I})$ in this case differs from (13) only in the description of the residues in the background.

A Gibbs sampler very similar to the one described in Section 6.1 can be implemented. The only difference is in the distribution $p(I_i \mid \boldsymbol{I}_{[-i]}, R, \boldsymbol{\theta})$ which is needed in the conditional update of $\boldsymbol{I}$. That is, conditional on $\boldsymbol{\theta}, R$, we slide through the whole sequence position-by-position to update I_i according to a random draw from $p(I_i \mid \boldsymbol{I}_{[-i]}, R, \boldsymbol{\theta})$, which satisfies

$$\frac{p(I_i = 1 \mid \boldsymbol{I}_{[-i]}, R, \boldsymbol{\theta})}{p(I_i = 0 \mid \boldsymbol{I}_{[-i]}, R, \boldsymbol{\theta})} = \frac{p_0}{1 - p_0} \prod_{k=1}^{w} \left(\frac{\theta_{k, r_{i+k-1}}}{\hat{\beta}_{r_{i+k-2} r_{i+k-1}}} \right).$$

For given $\boldsymbol{I}$, we update the profile matrix Θ in the same way as in Section 6.1.

6.3 Block-motif model with inhomogeneous background

It has long been noticed that DNA sequences contain regions of distinctive compositions. As discussed in Sections 5.3 and 5.4, a HMM can be employed to delineate a sequence with k types of regions. Suppose we decide to use a HMM to model sequence inhomogeneity. As we mentioned before, since the total motif residues is a very small fraction of the whole sequence, we may estimate the background model parameters directly by the methods in Section 5.3, pretending that R does not contain any motifs. Then we treat these parameters as known at the estimated values. After these, there are two strategies to modify the odds ratio formula (14).

In the first strategy, we treat each position in the sequence as a "probabilistic bp" (i.e., having probabilities to be one of the 4 letters) and derive the frequency model it. That is, we need to find $\theta^*_{ij} = p(r^*_i = j \mid R)$ for a future r^*_i and then treat residue r_i in the background as an independent observation from Multinom($\boldsymbol{\theta}^*_i$), with $\boldsymbol{\theta}^*_i = (\theta^*_{ia}, \ldots, \theta^*_{it})$. But this computation is nontrivial because

$$\theta^*_{ij} = p(r^*_i = j|R) = \theta_{0j} p(h_i = 0|R) + \theta_{1j} p(h_i = 1|R), \tag{15}$$

where $p(h_i)$ can be computed via a recursive procedure similar to (10). More precisely, in addition to the series of forward functions F_i, we can define the backward functions B_i. Let $B_n(h) = \sum_{h_n} \tau_{hh_n} \theta_{h_n r_n}$, and let

$$B_k(h) = \sum_{h_k} \{\tau_{hh_k} \theta_{h_k r_k} B_{k+1}(h_k)\}, \quad \text{for} \quad k = n-1, \ldots, 1. \tag{16}$$

Then we have

$$p(h_i = 1 \mid R) = \frac{F_i(1) B_{i+1}(1)}{F_i(1) B_{i+1}(1) + F_i(0) B_{i+1}(0)}. \tag{17}$$

This is the *marginal* posterior distribution of h_i and can be used to predict whether position i is in state 1 or 0. Thus, in the Gibbs sampling algorithm we only need to modify the denominator of the right hand side of (14) to $\prod_{k=i}^{i+w-1} \theta^*_{kr_k}$.

In the second strategy, we seek to obtain the probability of the whole segment,

$$R_{[i:i+w-1]} \equiv (r_i, \ldots, r_{i+w-1}),$$

conditional on the remaining part of the sequence, under the background HMM. Then we modify (14) accordingly. Clearly, compared with the first strategy, the second one is more faithful to the HMM assumption. The required probability evaluation can be achieved by a method similar to that in the first strategy. More precisely,

$$\begin{aligned} p(R_{[i:i+w-1]} \mid R_{[1:i-1]}, R_{[i+w:n]}) &= \frac{p(R)}{p(R_{[1:i-1]}, R_{[i+w:n]})} = \frac{p(R)}{\sum_{\mathbf{h}} p(R_{[1:i-1]}, R_{[i+w:n]}, \mathbf{h})} \\ &= \frac{F_n(0) + F_n(1)}{\sum_{h_1, \ldots, h_w} F_i(h_1) \tau_{h_1 h_2} \cdots \tau_{h_{w-1} h_w} B_{i+w}(h_w)}, \end{aligned} \tag{18}$$

where the denominator can also be obtained via recursions.

6.4 Extension to multiple motifs

Previously, we have assumed that there is only one kind of motif in the sequence and the prior probability for each $I_i = 1$ is known as p_0. Both of these assumptions can be relaxed. Suppose we want to detect and align m different types of motifs of lengths $w_1, \ldots, w_m$, respectively, and each occurring unknown number of times in R. We can similarly introduce the indicator vector $\boldsymbol{I}$, where $I_i = j$ indicates that an element from motif j starts at position i, and $I_i = 0$ means that no elements start from position i. For simplicity, we only consider the independent background model.

Let $p(I_i = j) = \epsilon_j$, where $\epsilon_0 + \cdots + \epsilon_m = 1$, be an unknown probability vector. Given what is known about the biology of the sequences being analyzed, a crude guess k_j for the number of elements for motif j is usually possible. Let $k_0 = n - k_1 - \cdots - k_m$. We can represent this prior opinion about the number of occurrences of each type of elements by a Dirichlet distribution on $\boldsymbol{\epsilon} = (\epsilon_0, \ldots, \epsilon_m)$, which has the form Dirichlet$(b_0, \ldots, b_m)$ with $b_j = J_0 \frac{k_j}{n}$, where J_0 represents the "weight" (or "pseudo-counts") to be put on this prior belief. Then the same predictive updating approach as illustrated in Section 6.1 can be applied. Precisely, the update formula (14) for $\boldsymbol{I}$ is changed to

$$\frac{\pi(I_i = j \mid \boldsymbol{I}_{[-i]}, R)}{\pi(I_i = 0 \mid \boldsymbol{I}_{[-i]}, R)} = \frac{\epsilon_j}{\epsilon_0} \prod_{k=1}^{w_j} \left(\frac{\theta^{(j)}_{k r_{i+k-1}}}{\theta_{0 r_{i+k-1}}} \right),$$

where $\Theta^{(j)} = [\boldsymbol{\theta}^{(j)}_1, \ldots, \boldsymbol{\theta}^{(j)}_{w_j}]$ is the profile matrix for the jth motif. Conditional on $\boldsymbol{I}$, we can then update $\boldsymbol{\epsilon}$ by a random sample from Dirichlet$(b_0 + n_0, \ldots, b_m + n_m)$, where n_j $(j > 0)$ is the number of motif type j found in the sequence, i.e., the total number of i such that $I_i = j$, and $n_0 = n - \sum n_j$. More details can be found in Neuwald, Liu and Lawrence (1995).

7 Discussion

As in classical statistics, optimization has been the primary tool in bioinformatics, in which point estimates of very high-dimensional objects obtained by dynamic programming or other clever computational methods are used. Characterizations of uncertainty in these estimates are mostly limited to simple significance test or completely ignored. The removal nuisance parameters is also problematic, most frequently being the *profile likelihood* method in which the nuisance parameters are fixed at their best estimates. In comparison, the Bayesian method has no difficulties in these important aspects: the uncertainty in estimation is addressed by posterior calculations and the nuisance parameters are removed by summation and integration. When achievable, this class of principled approaches is particularly advantageous in treating bioinformatics problems (Liu et al., 1999; Zhu et al., 1998). In exchange for these advantages, however, one needs to set prior distributions and overcome computational hurdles, none of which are trivial in practice.

The most important limitation on the Bayesian method is the need for additional computational resources. Recursion-based Bayesian algorithms generally have time and space requirements of the same order as their dynamic programming counterparts, although the constants are generally much larger. With the availability of fast workstations with large memories, however, this moderate increase in computing need is not a serious difficulty for most applications. For those problems where there is no polynomial time solution, MCMC methods (and other Monte Carlo methods) provides alternative means to implement a full Bayesian analysis. Although the use of MCMC methods and recursive methods can ease some of the computational concerns, difficulties remain for the specification of sensible prior distributions.

Acknowledgment

This work was supported in part by the NSF grants DMS-9803649 and DMS-0074108. Professor Donald B. Rubin and two anonymous referees for valuable suggestions.

References

[1] I. E. Auger and C. E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54, 1989.

[2] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure. Hidden markov-models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, 91(3):1059–1063, 1994.

[3] M. J. Bishop and E. A. Thompson. Maximum-likelihood alignment of dna-sequences. *Journal of Molecular Biology*, 190(2):159–165, 1986.

[4] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic dna. *Journal of Molecular Biology*, 268(1):78–94, 1997.

[5] L. R. Cardon and G. D. Stormo. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned dna fragments. *Journal of Molecular Biology*, 223(1):159–170, 1992.

[6] G. A. Churchill. Stochastic-models for heterogeneous dna-sequences. *Bulletin of Mathematical Biology*, 51(1):79–94, 1989.

[7] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall, New York, 1974.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39(1):1–38, 1977.

[9] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[10] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[11] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman & Hall, London, reprinted 1997. edition, 1995.

[12] C. Geyer. Markov chain monte carlo maximum likelihood. In E. Keramigas, editor, *Computing Science and Statistics: he 23rd symposium on the interface*, pages 156–163, Fairfax, 1991. Interface Foundation.

[13] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice.* Chapman & Hall, Boca Raton, Fla., 1998.

[14] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[15] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[16] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden markov-models in computational biology : Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531, 1994.

[17] A. Krogh, I. S. Mian, and D. Haussler. A hidden markov model that finds genes in escherichia-coli dna. *Nucleic Acids Research*, 22(22):4768–4778, 1994.

[18] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

[19] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, 1993.

[20] C. E. Lawrence and A. A. Reilly. An expectation maximization (em) algorithm for the idenification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7:41–51, 1990.

[21] F. Liang and W. H. Wong. Evolutionary monte carlo: applications to c_p model sampling and change point problem. *Statistica Sinica*, 10(2):317–342, 2000.

[22] J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene-regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[23] J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, New York, 2001.

[24] J. S. Liu and C. E. Lawrence. Bayesian inference on biopolymer models. *Bioinformatics*, 15(1):38–52, 1999.

[25] J. S. Liu, F. Liang, and W. H. Wong. The use of multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 95:121–134, 2000.

[26] J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *Journal of the American Statistical Association*, 90(432):1156–1170, 1995.

[27] J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Markovian structures in biological sequence alignments. *Journal of the American Statistical Association*, 94(445):1–15, 1999.

[28] T. M. Lowe and S. R. Eddy. trnascan-se: A program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Research*, 25(5):955–964, 1997.

[29] G. Marsaglia and A. Zaman. Monkey tests for random number generators. *Computers & Mathematics With Applications*, 26(9):1–10, 1993.

[30] X. L. Meng and D. van Dyk. The em algorithm : an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society Series B-Methodological*, 59(3):511–540, 1997.

[31] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[32] R. M. Neal. Probabilistic inference using markov chain monte carlo methods. *Tech. Rep., Comp. Sci. Dept., U. of Toronto*, CRG-TR-93-1, 1993.

[33] A. F. Neuwald, J. S. Liu, and C. E. Lawrence. Gibbs motif sampling : Detection of bacterial outer-membrane protein repeats. *Protein Science*, 4(8):1618–1632, 1995.

[34] A. F. Neuwald, J. S. Liu, D. J. Lipman, and C. E. Lawrence. Extracting protein alignment models from the sequence database. *Nucleic Acids Research*, 25(9):1665–1677, 1997.

[35] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature Biotechnology*, 16(10):939–945, 1998.

[36] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–590, 1976.

[37] S. C. Schmidler, J. S. Liu, and D. L. Brutlag. Bayesian segmentation of protein secondary structure. *Journal of Computational Biology*, 7(1-2):233–248, 2000.

[38] E. E. Snyder and G. D. Stormo. Identification of protein-coding regions in genomic dna. *Journal of Molecular Biology*, 248(1):1–18, 1995.

[39] M. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.

[40] J. L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum-likelihood alignment of dna-sequences. *Journal of Molecular Evolution*, 33(2):114–124, 1991.

[41] J. von Neumann. Various techniques used in connection with random digits. *National Bureau of Standards Applied Mathematics Series*, 12:36–38, 1951.

[42] J. C. Wootton. Nonglobular domains in protein sequences : Automated segmentation using complexity-measures. *Computers & Chemistry*, 18(3):269–285, 1994.

[43] J. Zhu, J. S. Liu, and C. E. Lawrence. Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, 14(1):25–39, 1998.

[44] M. Zuker. Computer-prediction of rna structure. *Methods in Enzymology*, 180:262–288, 1989.