BIOINFORMATICS

Vol. 15 no. 1 1999
Pages 38–52

Bayesian inference on biopolymer models

Jun S. Liu[1] and Charles E. Lawrence[2]

[1]Department of Statistics, Stanford University, Stanford, CA and [2]Wadsworth Center for Laboratories and Research, Albany, NY, USA

Received on March 30, 1998; revised on October 14, 1998; accepted on October 27, 1998

Abstract

Motivation: *Most existing bioinformatics methods are limited to making point estimates of one variable, e.g. the optimal alignment, with fixed input values for all other variables, e.g. gap penalties and scoring matrices. While the requirement to specify parameters remains one of the more vexing issues in bioinformatics, it is a reflection of a larger issue: the need to broaden the view on statistical inference in bioinformatics.*

Results: *The assignment of probabilities for all possible values of all unknown variables in a problem in the form of a posterior distribution is the goal of Bayesian inference. Here we show how this goal can be achieved for most bioinformatics methods that use dynamic programming. Specifically, a tutorial style description of a Bayesian inference procedure for segmentation of a sequence based on the heterogeneity in its composition is given. In addition, full Bayesian inference algorithms for sequence alignment are described.*

Availability: *Software and a set of transparencies for a tutorial describing these ideas are available at http://www.wadsworth.org/res&res/bioinfo/*

Contact: *lawrence@wadsworth.org; jliu@stat.stabford.edu*

Introduction

Computational approaches to molecular and structural biology are becoming increasingly important and have spawned the new field of bioinformatics. In the past decade, we have witnessed the developments of the likelihood and minimum message length approaches to pairwise alignments (Bishop and Thompson, 1986; Thorne *et al*., 1991, 1992; Allison *et al*., 1992), the probabilistic models for RNA secondary structure (Zuker, 1989; McCaskill, 1990); the expectation-maximization (EM) algorithm for finding regulatory regions (Cardon and Stormo, 1992; Lawrence and Reilly, 1990), the hidden Markov models for DNA composition analysis and multiple alignments (Churchill, 1989; Baldi *et al*., 1994; Krogh *et al*., 1994), the Gibbs sampling strategies for subtle motif detections and subtle multiple alignments (Lawrence *et al*., 1993; Liu, 1994; Neuwald *et al*., 1997), etc., all of which show that algorithms resulting from statistical thinking are invaluable tools in this field.

A main advantage of statistical approaches is that explicit probabilistic models are employed to describe relationships between various quantities with consideration of the underlying uncertainty. Then available statistical theory can automatically lead to an efficient use of available information in making predictions regarding biopolymer sequences. To date, however, statistical approaches have been primarily used for deriving efficient computational strategies. The utility of these methods to make statistical inferences about unobserved variables has received far less attention. With one exception (Zhu *et al*., 1997, 1998), methods which give complete statistical inferences on all unknowns for biopolymer sequences, either classical or Bayesian, are unavailable.

In this article, we show that the Bayesian methodology provides a useful way to formulate mathematically a bioinformatics problem which yields an assessment of the uncertainty in all unknowns. We also show that many existing recursive dynamic programming (DP) algorithms can be modified to solve the difficult computational problems required by Bayesian analysis. Following the Introduction, we provide a brief overview of Bayesian statistics. In subsequent sections, we apply the basic Bayes procedures to a simple coin example; describe applications in bioinformatics using two specific examples: sequence segmentation and global sequence alignment; and discuss the relationship of the Bayesian approach to other existing methods.

Basic Bayesian statistics

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 inference by explicitly quantifying uncertainties involved in the premise and thus the conclusions.

Classical statistics arrives at its inferential statements by using point estimates of unknown variables, with the maximum likelihood estimates being most popular. Uncertainty in estimation is addressed by studying the frequency behavior (or more properly, the pre-data behavior) of these esti-

mates and then putting confidence limits on the unknown parameters accordingly.

Bayesian statistics seeks a more ambitious goal by modeling all sources of uncertainty (physical randomness, subjective opinions, prior ignorance, etc.) with probability distributions and then trying to find the a posteriori distribution of all unknown variables of interest after considering the data. It uses the calculus of probability as the guiding principle in manipulating data and information, and derives its inferential statement purely based on the post-data probability distributions.

The value of using probability distributions to describe unknown quantities is indicated 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 always among the most coherent, consistent and efficient statistical methods.

The joint and posterior distributions

Bayesian statistics treats all quantities under consideration, be they observed data, unknown parameters, or missing data, as random variables. The full process of a typical Bayesian analysis can be described as consisting of three main steps (Gelman *et al*., 1995): (i) setting up a full probability model, the joint distribution, that captures the relationship among all the variables in consideration; (ii) 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; (iii) evaluating the appropriateness of the model and suggesting improvements (model criticizing and selection).

A standard procedure for carrying out step (i) is first to write down the likelihood function, i.e. the probability of the observed data given the unknowns, and multiply it by the a priori distribution of all the unobserved variables (typically unknown parameters). Let y_{obs} denote the observed data and θ the unobserved parameter. The joint probability can be represented as *Joint = likelihood* × *prior*, i.e.

$$p(y_{obs}, \theta) = p(y_{obs} \mid \theta)\pi(\theta) \qquad (1)$$

where $p(y_{obs} \mid \theta)$ is often denoted as $L(\theta; y_{obs})$ and referred to as the likelihood in classical statistics.

The Bayesian inference is drawn by examining the probability of all possible values of the variables of interest after considering the data. Accordingly, step (ii) is completed by obtaining the posterior distribution through the application of the Bayes theorem:

$$p(\theta \mid y_{obs}) = \frac{p(y_{obs} \mid \theta)\pi(\theta)}{\int p(y_{obs} \mid \theta)\pi(\theta)\mathrm{d}\theta} = \frac{p(\theta, y_{obs})}{p(y_{obs})} \propto p(y_{obs} \mid \theta)\pi(\theta) \qquad (2)$$

When θ is discrete, the integral is replaced by summation. The denominator $p(y_{obs})$ is the marginal distribution of the data, the so-called marginal likelihood of the model. It is sometimes convenient to realize that $p(y_{obs})$ is a normalizing constant, i.e. the constant that is required so that the whole function integrates to one. This constant is obtained by integrating out or summing over all variables, except for the observed data, from the joint distribution.

When there is more than one unknown, e.g. $\theta = (\theta_1, \theta_2)$, and interest focuses only on one component, say θ_1, those unknown quantities that are not of immediate interest, but are needed by the model, nuisance parameters, are removed by integration:

$$p(\theta_1 \mid y_{obs}) = \frac{\int p(y_{obs} \mid \theta_1, \theta_2)\pi(\theta_1, \theta_2)\mathrm{d}\theta_2}{\int\int p(y_{obs} \mid \theta_1, \theta_2)\pi(\theta_1, \theta_2)\mathrm{d}\theta_1\mathrm{d}\theta_2} = \frac{p(y_{obs}, \theta_1)}{p(y_{obs})} \qquad (3)$$

Note that computations required for completing a Bayesian inference are the integrations (sums for discrete variable) 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 equation (3), the challenging aspects of Bayesian statistics are 2-fold: (i) the development of a model, $p(y_{obs} \mid \theta)\pi(\theta)$, which must effectively capture the key features of the underlying scientific problem; and (ii) the necessary computation for deriving the posterior distributions.

Conjugate priors

An early effort at making the integration required by equation (2) accessible was the development of the so-called conjugate priors (Gelman *et al*., 1995). Abstractly, a conjugate prior is a family of distributions for $\pi(\theta)$ that has the same functional form as the likelihood function. As a consequence, when a conjugate prior is used, the functional form of the posterior distribution is the same as that of the prior. Although we can choose any functional form for $\pi(\theta)$, the conjugate prior enjoys the greatest mathematical and computational advantage.

In computational biology, because the data to be analyzed are usually categorical (e.g. DNA sequences with a four-letter alphabet or protein sequences with a 20-letter alphabet), the binomial and multinomial distributions are most commonly used. The unknown parameters often correspond to the frequencies of each letter in the alphabet. The conjugate priors for the multinomial families are the Dirichlet distributions, among which the Beta distribution is a special

case for the binomial family. In analyzing DNA sequences, we often let $\theta = (\theta_a, \theta_t, \theta_g, \theta_c)$ represent the unknown probabilities of the four nucleotides (e.g. P $\theta_i = 1$). With the simple model that each residue in the observed sequence is independent and identically distributed (iid) with frequency θ, the likelihood of an observed DNA sequence can be written as $p(n_a, n_t, n_g, n_c \mid \theta) \propto \theta_\alpha^{n_a} \theta_t^{n_t} \theta_g^{n_g} \theta_c^{n_c}$, where ($n_a$, n_t, n_g, n_c) is the count of the four types of nucleotides in the sequence. Thus, the conjugate prior for θ is of form $\pi(\theta) \propto \theta_a^{\alpha_{a-1}} \theta_t^{\alpha_{t-1}} \theta_g^{\alpha_{g-1}} \theta_c^{\alpha_{c-1}}$, which is a Dirichlet distribution with parameter $\alpha = (\alpha_a, \alpha_t, \alpha_g, \alpha_c)$, where α is often called the 'pseudo-counts'.

The missing data framework

In many problems, it is often fruitful to distinguish two kinds of unknowns: (population) parameters and missing data. Although there is no absolute distinction between the two types, missing data are usually directly related to the individual data and their dimensionality tends to increase as more and more data are observed. On the other hand, the parameters usually characterize the entire population of observations and are fixed in number. For example, in a multiple alignment problem, alignment variables that must be specified for each sequence (observation) are missing data. Residue frequencies or scoring matrices, which apply to all the sequences regardless of their number, are population parameters. Whereas this distinction is essential to the maximum likelihood method, it is employed primarily for conceptual clarity in Bayesian statistics.

When missing data y_{mis} are present in a statistical problem, the inference can be achieved by using the 'observed-data likelihood', defined as $L_{obs}(\theta; y_{obs}) = p(y_{obs} \mid \theta)$, which can obtained by integrating out the missing data from the 'complete-data likelihood', i.e.:

$$L_{obs}(\theta; y_{obs}) = \int p(y_{obs}, y_{mis} \mid \theta)\, \mathrm{d}y_{mis}$$

Since it is often difficult to complete this integral, the maximum likelihood methods often employ advanced computational tools such as the EM algorithm (Dempster *et al.*, 1977).

Bayesian analysis for missing data problems can be achieved coherently through integration:

$$p(\theta_1 | y_{obs}) \propto \int\int p(y_{obs}, y_{mis} \mid \theta_1, \theta_2) p(\theta_1, \theta_2) \mathrm{d}y_{mis} \mathrm{d}\theta_2$$

Since everything is treated as random variables in Bayesian statistics, the integration for eliminating the missing data is no different than that for eliminating nuisance parameters.

Model selection and Bayes evidence

At times, biology indicates that more than one model may be appropriate, and interest often focuses on assessing model fitness and conducting model selections [step (iii) described in the previous section 'The joint and posterior distributions'). The classical hypothesis testing can be seen as a model selection method in which one selects either the null hypothesis or 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 Bayesian model selection procedure, 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:

$$p(y_{obs}, \theta, M) = p(y_{obs} \mid \theta, M) p(\theta, M)$$

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

$$p(y_{obs}, \theta, M) = p(y_{obs} \mid \theta_m) p(\theta_m \mid M = m) p(M = m)$$

where $p(\theta_m \mid M = m)$ is the prior for the parameters in model m, and $p(M = m)$ is the prior probability of model m. The posterior probability for model m is obtained as:

$$\begin{aligned} p(M = m \mid y_{obs}) &\propto \int p(y_{obs} \mid \theta_m) p(\theta_m \mid M = m) p(M = m) \mathrm{d}\theta_m \\ &= p(y_{obs} \mid M = m) p(M = m) \end{aligned}$$

The choice of $p(M = m)$ is dependent on the problem, and we often set $p(M = 0) = p(M = 1) = 0.5$ a priori if we expect that both models are equally likely. Parameter θ_m can change meaning and dimensionality as the model type m changes. 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. We might set $p(M = 1)$ inversely proportional to the number of sequences in the database. Often in hypothesis testing, we wish to compare the null to a family of alternative models whose priors are not well specified, i.e. to a diffused alternative. Then the Bayesian evidence can be summarized as:

$$\sup_{p(\theta \mid M = 1)} p(M = 1 \mid y_{obs})$$

where the supremum is taken with respect to all allowable priors for θ in model 1.

Computational issues

In many practical problems, the required computation is the main obstacle for applying the Bayesian method. In fact, until recently, this computation has often been so difficult that Bayesian statistics was largely a field restricted to specialists. The introduction of iterative simulation methods, such as the data augmentation and the more general Markov chain Monte Carlo (MCMC) (Tanner and Wong, 1987; Gelfand and Smith, 1990), which provide Monte Carlo approximations to the required integrations and summations, has brought the Bayesian method into the mainstream of statistical analysis. The MCMC strategy has also led to some useful sequence analysis algorithms (Lawrence *et al*., 1993; Neuwald *et al*., 1995, 1997). As we illustrate below, by appealing to the rich history of computation in bioinformatics, the required summations can often be performed exactly, which gives rise to either an exact Bayesian inference or an improved MCMC method.

A coin example

To illustrate the basic ideas just described, in this section we consider a simple coin game in which one cannot expect the coins to be fair. In the game, *n* coins are tossed and laid out in a row. You are asked to make an inference about the probability of heads for this sequence of coins.

Single coin type

Suppose the *n* coins are identical with probability of heads θ_1. Let h_n be the number of observed heads and t_n the number of tails. The likelihood function for the observed sequence can be written as the product of *n* Bernoulli trials:

$$L(\theta_1; y_{obs}) = P(y_{obs}|\theta_1) = \theta_1^{h_n}(1 - \theta_1)^{t_n} \tag{4}$$

We model the prior of θ_1 by a Beta distribution defined as:

$$\pi(\theta_1) = B(\theta_1; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_1^{\alpha-1}(1-\theta_1)^{\beta-1} \propto \theta_1^{\alpha-1}(1-\theta_1)^{\beta-1} \tag{5}$$

where $\Gamma(\cdot)$ is the complete gamma function and $\alpha, \beta > 0$ are the parameters set by the user. A useful fact to show that equation (5) does integrate to one is:

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{6}$$

Figure 1 shows this distribution for $\alpha = 2$ and $\beta = 4$, as well as for the special case $\alpha = 1$, and $\beta = 1$ which corresponds to a uniform distribution. The joint distribution of the data and θ_1 is:

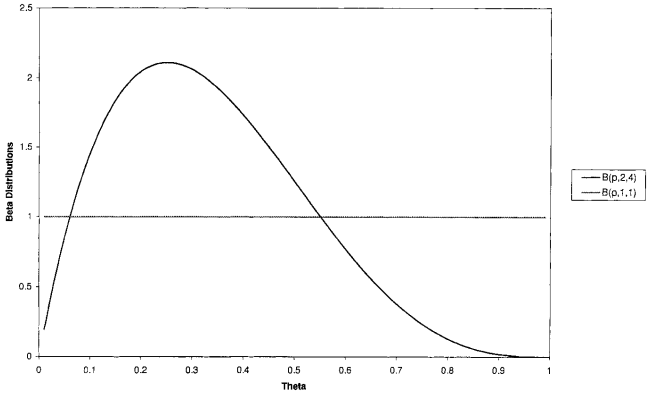

Fig. 1. Two Beta distributions with parameters $\alpha = 2$ and $\beta = 4$, and parameters $\alpha = 1$ and $\beta = 1$.

$$P(y_{obs}, \theta_1) = L(\theta_1; y_{obs})\pi(\theta_1) = \frac{\Gamma(\alpha = \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_1^{h_n+\alpha-1}(1-\theta_1)^{t_n+\beta-1}$$

from which we derive the marginal likelihood by using formula (6):

$$P(y_{obs)} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\int(\theta_1)^{h_n+\alpha-1}(1 - \theta_1)^{h_n+\beta-1}d\theta_1 =$$

$$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad \frac{\Gamma(h_n + \alpha)\Gamma(t_n + \beta)}{\Gamma(n + \alpha + \beta)} \tag{7}$$

and the posterior distribution:

$$P(\theta_1 \mid y_{obs}) = \frac{P(y_{obs}, \theta_1)}{P(y_{obs})}$$

$$= \frac{\Gamma(n + \alpha + \beta)}{\Gamma(h_n + \alpha)\Gamma(n - h_n + \beta)}\theta_1^{h_n+\alpha-1}(1 - \theta_1)^{t_n+\beta-1} \tag{8}$$

As expected, the posterior is a Beta distribution with updated parameters, i.e. $B(\theta_1; h_n + \alpha, t_n + \beta)$. Notice that in this posterior distribution the prior parameter (α, β) and the number of heads or tails (h_n, t_n) are exchangeable. Accordingly, these prior parameters are often referred to as pseudo counts.

The posterior distribution of θ_1, with y_{obs} = {01010000000001100110100101101 1} (1 = heads, 0 = tails; so $h_n = 12$ and $t_n = 18$) and $\alpha = \beta = 1$, is shown in Figure 2, which graphically summarizes the Bayesian inference for this problem: after considering the data, the inference on θ_1 is made by specifying a probability density on all its possible values. The inferred posterior distribution of θ_1 is a probability density with main probability mass surrounding the empirical frequency 0.4 with an appropriate spread.

If four-sided or 20-sided coins are employed, the foregoing game serves as a model for residue composition of DNA or protein sequences, respectively. Generally, if the outcome of

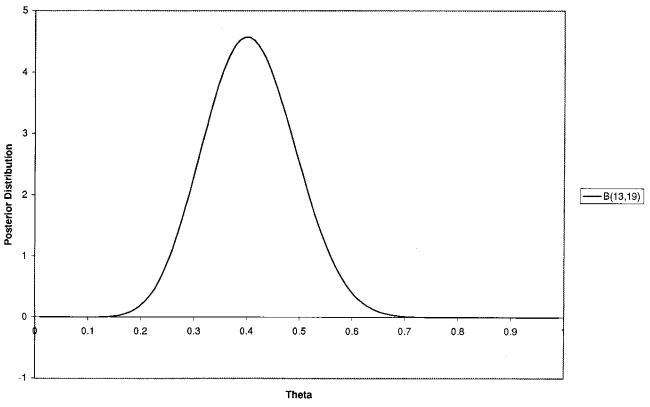

Fig. 2. Posterior Beta distribution for coin-tossing game assuming the use of only one coin: $B(\theta_1; 13,19)$.

each trial (or observed residue) takes value in an alphabet {1, …, *D*} with probability θ_d for $d = 1, \ldots, D$, then the binomial distribution is generalized to the multinomial distribution and the conjugate Beta prior distribution generalizes to the conjugate Dirichlet distribution. With this generalization, the marginal likelihood (7) becomes:

$$P(y_{obs}) = \frac{\Gamma(\Sigma_d \alpha_d)}{\Pi_d \Gamma(\alpha_d)} \frac{\Pi_d \Gamma(n_d + \alpha_d)}{\Gamma(n + \Sigma_d \alpha_d)} \tag{9}$$

where α_d are the parameters for the prior Dirichlet distribution, and n_d are the counts of residue type *d* observed after tossing the *D*-sided object $n = \Sigma_d n_d$ times. Furthermore, the posterior distribution (8) generalizes to:

$$P(\theta|y_{obs}) = \frac{\Gamma(\Sigma_d(n_d + \alpha_d))}{\Pi_d \Gamma(n_d + \alpha_d)} \prod_d \theta_d^{n_d + \alpha_d - 1} \tag{10}$$

Two types of coins: Bayesian segmentation

Suppose you are told that, in the game described above, two types of coins rather than one have been used: the first *A* coins which make up the first segment have probability θ_1 of heads, and the remaining $n - A$ coins have probability θ_2 of heads, with *A* unknown. Treating *A* (called the change point) as missing data, we can write the complete-data likelihood of the sequence as:

$$L(\theta_1, \theta_2; y_{obs}, A = a) = \theta_1^{h_1}(1 - \theta_1)^{t_1} \theta_2^{h_2}(1 - \theta_2)^{t_2} g(a),$$

where h_i is the number of heads in the *i*th segment, and t_i the corresponding number of tails, and an arbitrary prior distribution $g(a)$ for *A*. Note that h_i and t_i are both functions of *a*. We sometimes also write $h_i(a)$ and $t_i(a)$ for clarity's sake.

We choose conjugate priors for the θ, e.g. the Beta distributions, i.e. $\pi(\theta_1, \theta_2) = B(\theta_1; \alpha_1, \beta_1) \times B(\theta_2; \alpha_2, \beta_2)$,. The joint distribution of all the variables including the missing data becomes:

$$P(\theta_1, \theta_2, y_{obs}, A = a) = L(\theta_1, \theta_2; y_{obs}, a)\pi(\theta_1, \theta_2)g(a)$$

$$= g(a) \prod_{i=1}^{2} \frac{\Gamma(\alpha_i + \beta_a)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \theta_i^{h_i + \alpha_i}(1-\theta_i)^{t_i + \beta_i - 1}$$

The exact posterior distribution for *A* is obtained as follows:

$$P(A = a, y_{obs}) = \int\int p(\theta_1, \theta_2, a, y_{obs}) d\theta_1 d\theta_2$$

$$= g(a) \prod_{i=1}^{2} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \int_o^1 \theta_i^{h_i(a) + \alpha_i - 1}(1-\theta_i)^{t_i(a) + \beta_i - 1} d\theta_i \tag{11}$$

By using formula (6), we further derive that:

$$P(A = a, y_{obs}) = g(a) \prod_{i=1}^{2} \left[\frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \frac{\Gamma(h_i + \alpha_i)\Gamma(t_i + \beta_i)}{\Gamma(h_i + t_i + \alpha_i + \beta_i)} \right] \tag{12}$$

For an alphabet of size *D*, the expression in square brackets is replaced by an expression analogous to (9). The marginal likelihood is obtained by summing over *a*: $P(y_{obs}) = \Sigma_a P(A = a, y_{obs})$ and the posterior distribution for *a* is:

$$P(A = a \mid y_{obs}) = P(A = a, y_{obs})/P(y_{obs}) \tag{13}$$

For given $A = a$, the two parameters, θ_1 and θ_2, are mutually independent and have Beta distributions. Thus, the marginal posterior distribution of, say, θ_1, can be expressed as a mixture of Beta distributions:

$$p(\theta_1|y_{obs}) = \sum_{a=0}^{n} p(\theta_1 \mid y_{obs}, A = a) p(A = a|y_{obs})$$

$$= \sum_{a=0}^{n} B(\theta_1; h_1(a) + \alpha_1, t_1(a) + \beta_1) P(A = a \mid y_{obs})$$

As an alternative to this messy expression of mixtures, we can perform a Monte Carlo approximation by drawing random samples $a_1, a_2, \ldots, a_m$ from $P(A = a \mid y_{obs})$ and then averaging the Beta distributions determined by the sampled values:

$$\tilde{p}(\theta_1 \mid y_{obs}) = \frac{1}{m} \sum_{j=1}^{m} B(\theta_1; h_1(a_j) + \alpha_1, t_1(a_j) + \beta_1)$$

The above approach can be illustrated by a data set of size $n = 30$ generated from $\theta_1 = 0.2$ and $\theta_2 = 0.6$. We observe y_{obs} ={00101000000001011110111100010}. The true *a* equals 13 and is generated from Binomial(30,1/2). We conducted the Bayesian two-coin analysis. Setting the prior parameters as $\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1$, and assuming all change points are equally likely, i.e. $g(a) = \frac{1}{n+1}$. The marginal likelihood for the model with one coin is $P(y_{obs}|\textit{one coin}) = 2.69 \times 10^{-10}$ and for the two-coin model $P(y_{obs}|\textit{two coins}) = 5.9 \times 10^{-10}$.

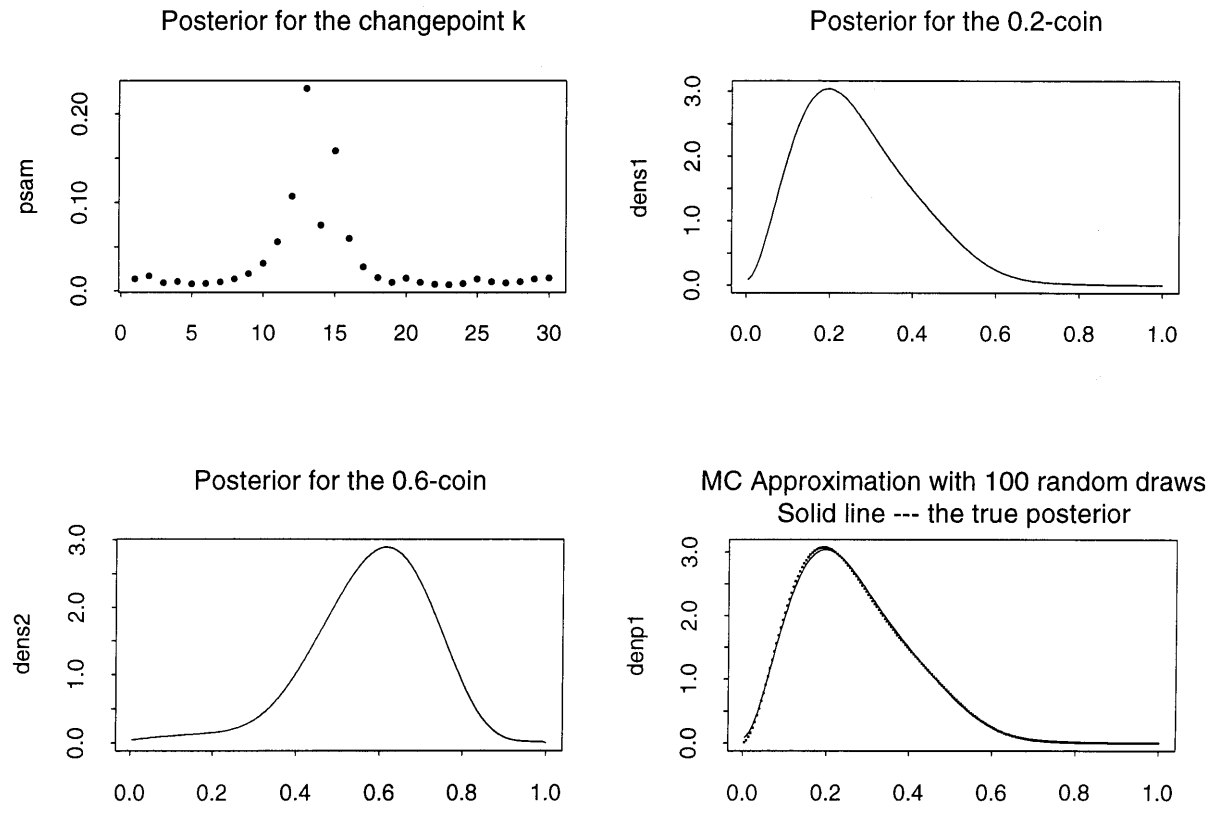

Fig. 3. Posterior distributions for the coin-tossing game assume two coins are used. (**a**) Posterior distribution for the change point from coin #1 to coin #2. (**b**) Posterior distribution for probability of heads on coin #1. (**c**) Posterior distribution for the probability of heads on coin # 2. (**d**) Sampling approximations for the posterior distribution of the probability of heads on coin #1 and the exact posterior distribut ion.

Assuming that the two models are equally likely a priori, the posterior probability that these data were generated by two coins, $P(\textit{two coins}|y_{obs}) = 0.69$. The posterior distributions of a, θ_1 and θ_2 are in Figure 3. A Monte Carlo approximation to the posterior of θ_1 is also shown, just to demonstrate that this approximation can be quite accurate.

Gibbs sampling: a method of Monte Carlo approximation

As indicated above, it is usually the case in applied statistics that the computation involved in eliminating the missing data or a nuisance parameter is so difficult that one needs to use numerical approximations, Monte Carlo methods, other heuristics or a combination of these to complete the required sums and integrals. The Gibbs sampling approach is a special MCMC method that allows one to draw samples of high-dimensional random variables in an iterative fashion. While such approximations are not necessary for the example in the previous paragraph, we consider the application of Gibbs sampling to this problem for illustration purposes. The Gibbs sampler proceeds by drawing samples from each component at a time from its conditional distribution with the rest of the components fixed. In particular, the following procedure seizes the essence of the Gibbs sampler (Gelfand and Smith, 1990):

- Fix $A = a$ and θ_2, draw a new θ_1 from its conditional posterior distribution

$$p(\theta_1 \mid A = a, \theta_2, y_{obs}) = P(\theta_1 \mid A = a, y_{obs}) = \text{Beta}(\theta_1; h_1(a) + \alpha_1, t_1(a) + \beta_1)$$

  to substitute for the old θ_1.
- Now we move to θ_2: fix $A = a$ and the θ_1 just drawn, we sample a new θ_2 from

$$p(\theta_2 \mid A = a, \theta_1, y_{obs}) = p(\theta_2 \mid A = a, y_{obs}) = \text{Beta}(\theta_2; h_2(a) + \alpha, t_2(a) + \beta_2).$$

- Given θ_1 and θ_2, draw A from its conditional distribution

$$P(A = a|\theta_1, \theta_2, y_{obs}) \propto \theta_1^{h_1(a)}(1-\theta_1)^{t_1(a)}\theta_2^{h_2(a)}(1-\theta_2)^{t_2(a)}g(a) \qquad (14)$$

This step can be done by computing the right-hand side of equation (14) for $a = 0, 1, \ldots, n$ and then summing them to renormalize. Asympotically, this algorithm will converge

and yield samples from the posterior distribution of (θ_1, θ_2, A). After convergence, samples from this distribution can be used to approximate posterior distributions of interest in a manner similar to that described at the end of the last section. The major weakness of MCMC sampling algorithms is that in general there is no way to guarantee that convergence has been achieved. Accordingly, such MCMC samples are approximate. On the other hand, when samples can be shown to be drawn from the posterior distribution, as is the case in the previous section, the samples are said to be exact.

A Bayesian bioinformatics paradigm

A probabilistic model is often used as a mechanism through which one connects observed data with a scientific premise or hypothesis about the real-world phenomena. Such models are at the core of all statistical analysis. Since bioinformatics explicitly or implicitly concerns the analysis of data, such models are also at the core of bioinformatics. Because 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 few parameters are often used to achieve the purpose.

When the model is given, some efficient methods 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. Nearly all bioinformatics methods employ score functions, which often are functions of likelihoods or likelihood ratios, at least implicitly. The specification of priors, required for Bayesian statistics, is less well understood in bioinformatics, although not completely foreign. Specifically, the setting of parameters for an 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. At the other extreme is the specification of the so-called uninformed priors, which assigns equal probability to all possible values of the unknowns. The introduction of non-degenerate priors can usually give more flexibility in modeling reality without the use of a more complicated likelihood.

To obtain desired posterior distributions, we must complete the summations and integrations in equations such as (2) and (3). Recursions have been employed with great advantage in bioinformatics as the basis of numerous DP algorithms. In the following, we show by two examples how the basic principles in Bayesian statistics can be applied and existing DP algorithms can be adapted to solve bioinformatics problems.

Bayesian sequence segmentation algorithm

Sequence segmentation models have been developed for many purposes in bioinformatics. These include models of protein sequence hydrophobicity (Kyte and Doolittle, 1982; Auger and Lawrence, 1989), models of protein secondary structure (Schmidler *et al*., 1998), models of sequence complexity (Wootton, 1994), models of sequence composition (Churchill, 1989) and models for gene identification (Krogh *et al*., 1994a; Snyder and Stormo, 1995). What is common to all these methods is that a single sequence is characterized by a series of models which only involve local properties. To facilitate the presentation of these concepts, we begin with a simple case in which each segment is described by an independent model. This approach is applicable to studies of sequence composition and sequence complexity. In the next subsection, we outline the approach to a more complicated case which requires that each segment may be described by one of several models, e.g. protein secondary structure models.

The basic segmentation model. This segmentation model is a generalization of the two-coin example in the previous section. Suppose you are told that a dealer has $k_{\max}$ different coins available to toss instead of just two. The probabilities of heads are different from one another, i.e. $\theta_1 \neq \theta_2 \neq \ldots \neq \theta_{k\max}$, and unknown to you. The dealer flips the first coin C_1 times and records the results, the second coin C_2 times, and so on until she or he has used $\kappa \leq k_{\max}$, coins with a total of $J = \sum_{k=1}^{\kappa} C_\kappa$ flips. You are given only the complete sequence of heads and tails, $R = (r_1, \ldots, r_j)$, where $r_j = \{H,T\}$, with C_1, C_2, …, C_κ and κ unknown. The change points in this sequence occur each time a new coin is used, i.e. at $A_k = \sum_{v=0}^{k} C_v + 1$ with $C_0 = 0$ and $k = 1, \ldots, \kappa$.

Of interest are the values of all the unknowns. Since the number of parameters changes with κ, the choice of the number of change points is a model selection problem. Compared with the two-coin example, the new game is more complicated: there are more change points and the number of these changes is unknown. This example will illustrate a general characteristic of Bayesian bioinformatics: the use of recursions for completing large summations.

The use of a four-sided or 20-sided coin in the foregoing game serves as a model for heterogeneity in residue composition of DNA or protein sequences. Generally, we assume that each outcome (or residue) can take values in an alphabet $\{1, \ldots, D\}$. If residue j is in the kth segment, then $P(r_j = d) = \theta_{kd}$ for $d = 1, \ldots, D$. We let $\Theta_k = (\theta_{k1}, \ldots, \theta_{kD})$. To facilitate computation, we introduce the following notations for a segment of the sequence:

$R_{[i:j]} = (r_i, \ldots, r_j)$; $R_{(i:j]} = (r_i + 1, \ldots, r_j)$; $R_{[i:j)} = (r_i, \ldots, r_{j-1})$

Assuming that the segments are independent of each other given the change points, the complete data likelihood is again the product of the likelihoods of the κ segments weighted by the prior distribution of the change points:

$$P(R, A|\Theta, \kappa) = \prod_{k=1}^{\kappa} P(R_{[A_{k-1}:A_k)} \mid \Theta_k) P(A \mid \kappa)$$

where $P(A \mid \kappa)$ plays the role of $g(a)$ in the previous section 'Model selection and Bayes evidence', and κ are the parameters of the prior segmentation model. With this likelihood, the joint distribution for all the variables can be written as follows:

$$P(R, A, \Theta, \kappa) = L(R, A; \Theta; \kappa)\pi(\Theta, \kappa)$$

$$= P(R|A, \Theta)P(A|\kappa)P(\kappa)P(\Theta) \qquad (15)$$

where Θ and κ are assumed independent a priori.

Computations. The unknowns are the number of segments κ, segmentation A of the sequence, and residue composition Θ in each segment. The posterior distributions of these quantities can all be derived from equation (15), if the necessary summations and integrations can be completed.

We assume again, as with the two-coin example, that a priori all the segmentations with κ change points are equally likely, and thus have prior probability inversely proportional to the number of ways to segment the sequence into κ parts, i.e. $P(A \mid \kappa) = \binom{N}{\kappa}^{-1}$. Furthermore, we assign a prior probability 0.5 to the null model and assume that all of the $k_{\max}$ models which have $\kappa > 0$ change points are equally likely, i.e. $P(\kappa) = \frac{0.5}{\kappa + 1}$.

From equation (15), we obtain the marginal likelihood:

$$P(R) = \sum_{k=1}^{k_{\max}} P(\kappa = k)P(R|\kappa = k)$$

$$= \sum_{k=1}^{k_{\max}} P(\kappa = k) \sum_{A:\|A\|=k} \int P(R, A \mid, \Theta)P(\Theta)\mathrm{d}\Theta \qquad (16)$$

where $\|A\|$ is the number of segmentations implied by A. As in the coin example, we model the residues in each segment by a product multinomial model and a prior product Dirichlet model. With the segment independence and model independence assumptions, we have:

$$P(R|\kappa = k) = \sum_{\|A\|=k} \prod_k \frac{\Gamma(\Sigma_d \alpha_d)}{\Pi_d \Gamma(\alpha_d)} \frac{\Pi_d \Gamma(n_{k,d} + \alpha_d)}{\Pi(n_k + \Sigma_d \alpha_d)} \qquad (17)$$

where $n_{k,d}$ is the count of residue type d in the kth segment $R[A_{k-1}:A_k)$. Apparently, a brute-force computation of equation (17) is prohibitive. Fortunately, the dynamic programming approach of Auger and Lawrence (1989) can be adapted to complete the summation.

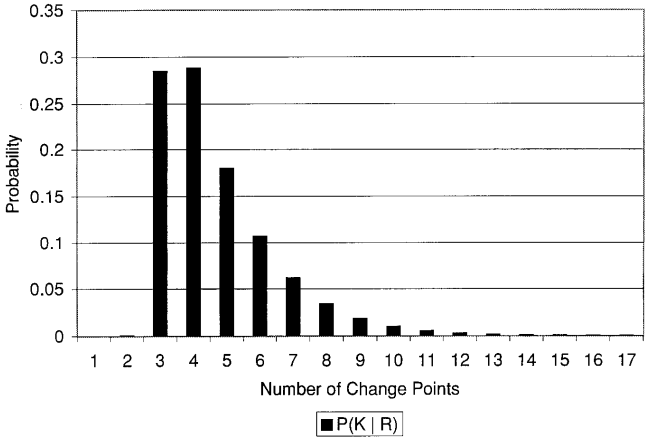

Fig. 4. Posterior distribution of the number of change points in the nucleotide composition of the 500 bp upstream of the translational start site of histone H1 from *Saccharomyces cerevisiae* (*H1: 500bp*).

Let $P(R_{[i:j]} \mid k)$ denote the probability of observing the subsequence $R_{[i:j]}$ given that it consists of k segments. These quantities can be computed using equation (17) with R substituted by $R_{[i:j]}$, $j = 1, \ldots, N$, $i = j, \ldots, N$, and stored in advance. The DP recursion of Auger and Lawrence can then be adapted as:

$$P(R_{[1:j]}|k) = \sum_{v<j} P(R_{[1:v]}|k-1)P(R_{[v:j]}|1) \qquad (18)$$

With $P(R \mid \kappa)$ computed, we use the Bayes rule to obtain $P(\kappa \mid R)$. Figure 4 gives the distribution of the number of change points in composition for a fragment of the genome sequence of *Saccharomyces cerevisiae*, the 500 base pairs upstream of the translational start site of the histone H1 gene (*H1: 500bp*).

The marginal probability that a change point will occur at position v can be obtained:

$$P(A_k = v \text{ for some } k \mid R)$$

$$= \frac{1}{P(R)} \sum_{\kappa} \sum_{k} P(R_{[1:v]} \mid k)P(R_{(v:j]} \mid \kappa - k) \qquad (19)$$

This distribution is illustrated for *H1: 500bp* in Figure 5.

Backward sampling. Since the locations of the change points are mutually dependent, an analytic expression for the distribution of A is not available. However, we can draw exact and independent samples from this distribution by using a recursive backward sampling algorithm.

The first step of the backward sampling algorithm is to draw $\kappa = k$ from its marginal posterior distribution obtained by inverting equation (17) with the Bayes theorem, and to set

$A_k = J$. Then the change points $(A_1, \ldots, A_{k-1})$ are obtained by recursively sampling backward from the following distribution:

$$P(A_{q-1} = j, \mid R, A_q = m) =$$

$$\frac{P(R_{[1:j]} \mid q-1)PR_{(j:m]}|1)}{P(R_{[1:m]}|q)}$$

This forward/backward process mirrors the usual dynamic programming in which the forward step finds the optimal value of the objective function and the backward step traces the solution corresponding to that optimum. Here, the forward step sums over all segmentation variables to yield necessary marginal and conditional distributions, and the backward step samples a solution in proportion to its posterior probability. Averaging over these draws yields a histogram which approaches the distribution of equation (19) to any desired degree. The posterior distribution of the residue frequency f_j at each position j, after considering all possible segmentations, may be examined as well: suppose position j is covered by the kth segment $[u, v]$ of a given segmentation, then:

$$P(f_j \mid R) = P(\Theta_k \mid R) = P(R_{[u:v]} \mid \Theta_k, 1)P(\Theta_k)$$

Averaging these over the sampled segmentations yields the desired distribution. Figure 6 shows this distribution for *H1: 500bp*.

Further analysis and extensions. In the previous subsection, our development employed the following assumptions: residues are independent of one another and the same model with independent parameters is applicable to all segments. Generalization to more complicated segmentation models, e.g. Markov Chain models of sequence composition or application-specific models, e.g. intron/exon models can be obtained through the specification of individual segment models (Liu and Lawrence, 1996; Schmidler *et al*., 1998).

The most well known of these is protein secondary structure prediction in which any subsequences may be classified into any of the three models: α helix, β strand or random coil [simple and useful probabilistic models for the helices and strands have been proposed by Schmidler *et al*. (1998)]. In this case, not only are the locations of the change points unknown, but also the identity of the model appropriate to each segment is unknown. The class of a segment can also depend on the classes of the adjacent segments. Traditionally, these methods employ fixed parameters which have been estimated using a training set. In the Bayesian context, this corresponds to the specification of priors for each of the classes. They thus may be described by hidden Semi-Markov models for which appropriate recursions can be employed (Schmidler *et al*., 1998).

The recursive Bayesian approach is also useful when training data are available. When training data yield exact determination of change points A and model types M, the dis-

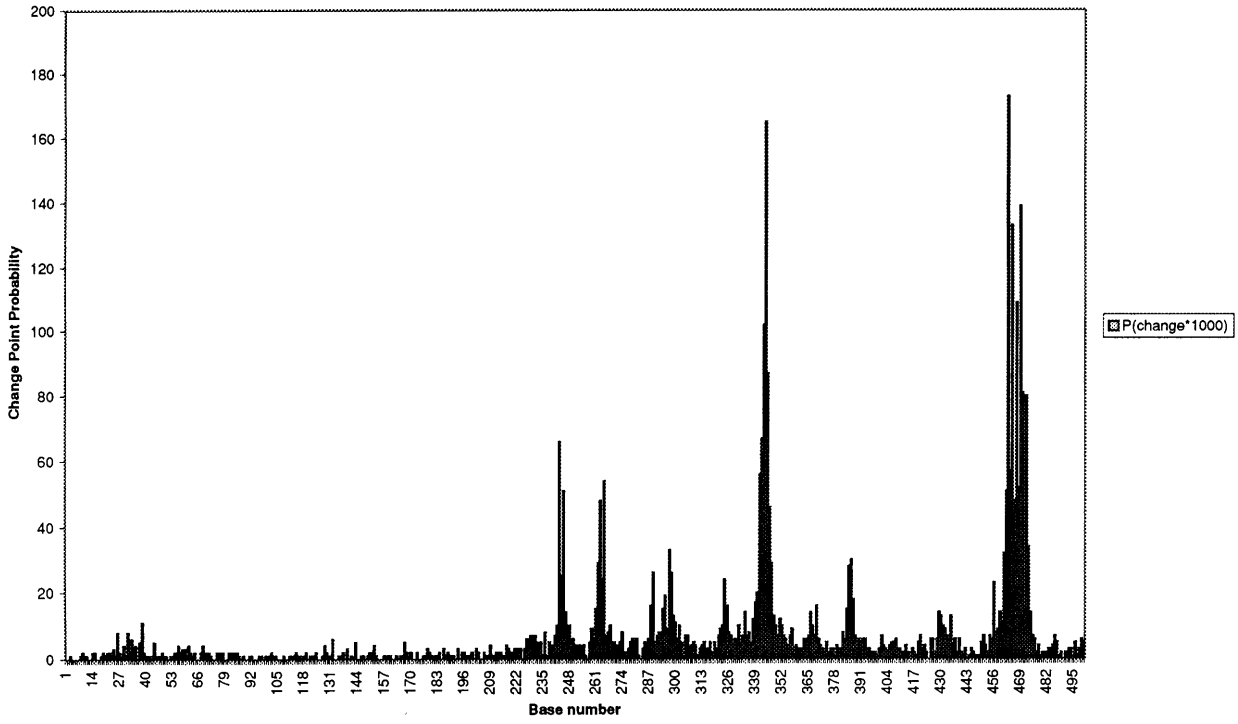

Fig. 5. Posterior marginal distribution of the change point positions for *H1: 500bp*.

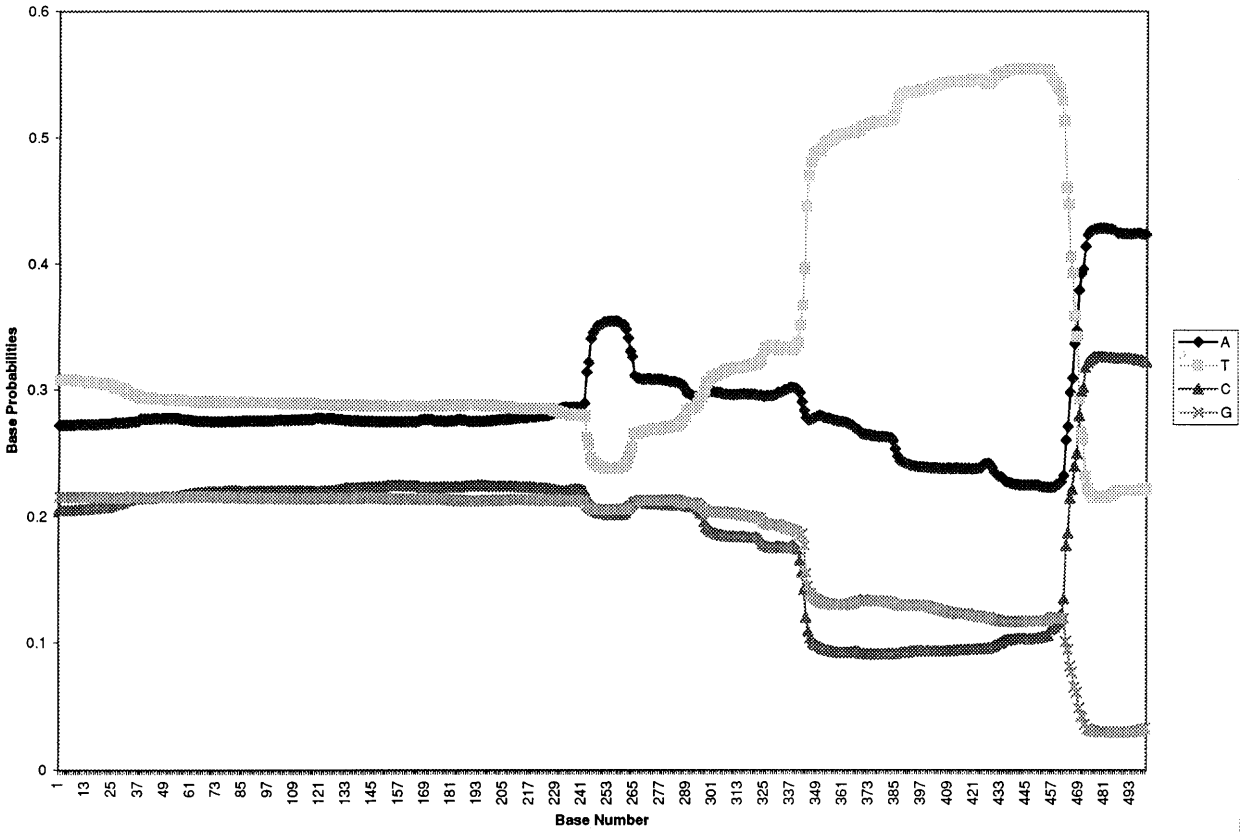

Fig. 6. Posterior distribution of the segmented composition of *H1: 500bp*.

tribution of the observed data parameters Π_k and the missing data parameters, κ, e.g. length distributions of secondary structure types, can be inferred by the Bayesian method without the use of advanced computational methods. There are also situations when the training data are less than perfect. For example, crystal structures provide good data on the secondary structure types for each segment and their locations, but the ends of secondary structure elements are often difficult to pinpoint exactly. In this case, model type variables M are observed exactly, but there is some uncertainty in the change points A, which can be incorporated into training through the assignment of positive probabilities for residues near these ends, and zero probabilities elsewhere. Furthermore, data sets with mixed observations of complete data and incomplete data can be analyzed in one coherent way. The posterior distributions developed in training become the informed priors for the testing phase.

Bayesian pairwise alignment

Sequence alignment has been one of the most important methodologies developed in bioinformatics [see Waterman (1995) for a review]. In this section, we compare two Bayesian alignment algorithms to traditional optimal alignment methods. One is a gap-based alignment procedure based on the recursion of Needleman and Wunsch (1970). The other method is a motif-based alignment algorithm which has been described in detail by Zhu *et al*. (1998). Throughout the section, the observed data consist of two nucleotide or protein sequences, R^1 and R^2, of lengths n_1 and n_2, respectively. The observed data parameter, Θ, is a finite set of matrices which are analogs of scoring matrices, e.g. the PAM (Dayhoff *et al*., 1972) or Blosum (Henikoff and Henikoff, 1992) series. The alignment is characterized by a matrix, A, whose elements $a_{i,j}$ are set to one if residue i of sequence 1 aligns with residue j of sequence 2, and zero otherwise.

Gap-based alignment. Traditionally, the entropic explosion in the number of alignments has been controlled by using penalties, $\log(\lambda_o)$ and $\log(\lambda_e)$, where λ_o, and λ_e are probabilities of gap opening and gap extension, respectively. Here we show how alignment problems of this type may be treated in a Bayesian way by using the statistical models pioneered by Thorne *et al*. (1991, 1992). For the gap-based alignment algorithm, the parameters of the missing data are the gap pen-

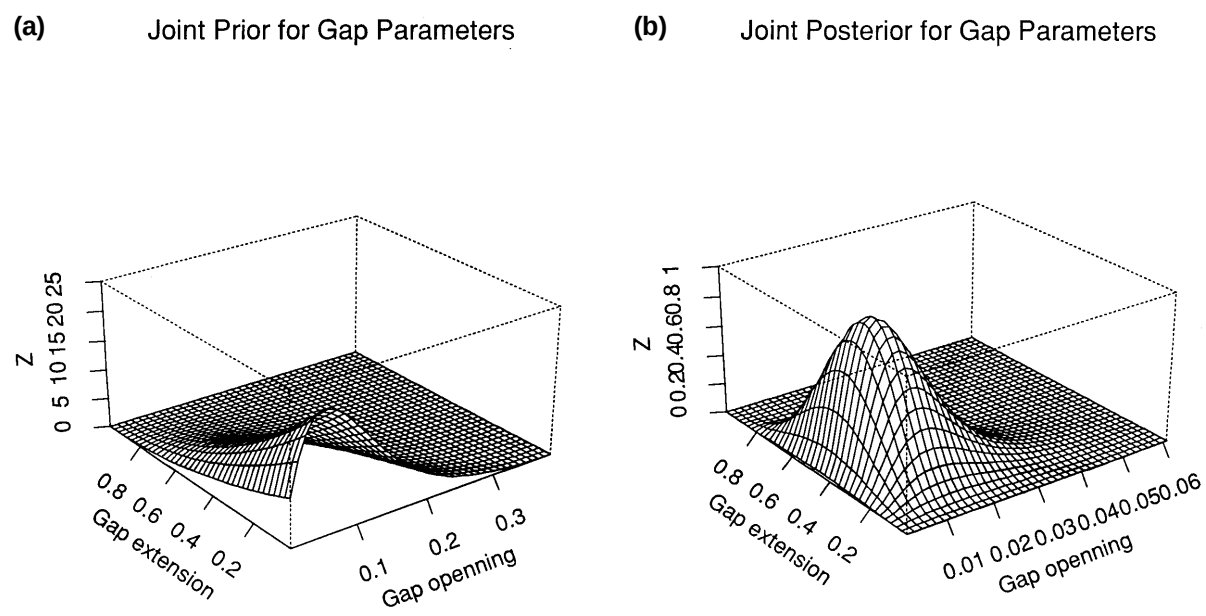

Fig. 7. Prior (**a**) and posterior (**b**) distributions of the gap penalties for the alignment of the hemoglobin α and β chains.

alties, Λ. The prior for alignment in the motif-based model will be described later. The joint distribution is defined as:

$$P(R^1, R^2, A, \Theta, \Lambda) = P(R^1, R^2 \mid A, \Theta)P(\Theta)\, P(A \mid \Lambda)P(\Lambda)$$

Traditional alignment procedures can be seen as optimizing an objective function, usually a similarity score, which is often a log-likelihood (Holmes and Durbin, 1998). More precisely, for a set of fixed values $\Pi = \Pi^0$ and $\Lambda = \Lambda^0$, one finds A^* so that:

$$\log(P(R, A^* \mid \Theta^0, \Lambda^0)) = \max_{all\ A} \{\log(P(R \mid A, \Theta^0) + \log(P(A \mid \Lambda^0))\} \quad (20)$$

The need for setting parameter values Θ^0 and Λ^0 has been the subject of much discussion in bioinformatics. A distinctive advantage of the Bayesian procedure is the added modeling flexibility in the specification of parameters. Here we can regard the selection of Θ^0 and Λ^0 as a special case for the specification of a prior distribution, i.e. the prior is degenerate with probability 1 for the values Θ^0 and Λ^0, and zero for all other values.

A full Bayesian procedure uses a non-degenerate prior distribution for Θ and Λ. Figure 7a shows one such prior distribution $P(\Lambda)$ for the affine gap opening and extension parameters (λ_o, λ_e), which is a product of Beta distributions with the form $\text{Beta}(\lambda_o; 2,18) \times \text{Beta}(\lambda_e; 1,3)$. This choice is just one from a family of distributions described by $\text{Beta}(\lambda_o; \alpha_o, \beta_o) \times \text{Beta}(\lambda_e; \alpha_e, \beta_e)$ which includes degenerate forms and uninformed forms. In the following, we describe a Bayesian Needleman–Wunsch algorithm.

Let A be the alignment matrix which can be seen as a 'path' in a dynamic programming setting. With given $\Lambda = (\lambda_o, \lambda_e)$, the probability of any allowable path, prior to seeing the content of the two sequences to be aligned but conditional on their lengths n_1 and n_2, is:

$$P(A \mid \lambda_o, \lambda_e) = \frac{\lambda_o^{k_g(A)} \lambda_e^{l_g(A) - k_g(A)}}{\sum_{A'} \lambda_o^{k_g(A')} \lambda_e^{l_g(A') - k_g(A')}} \quad (21)$$

where $k_g(A)$ and $l_g(A)$ are the total number and the total length of the gaps in A, respectively. The summation in the denominator is over all possible alignment A of the two sequences. In the following derivation, we assume that the length information, n_1 and n_2, is conditioned on implicitly. Thus:

$$P(\Pi, A, R^1, R^2 \mid \Lambda) = P(R^1, R^2 \mid A, \Theta)P(\Theta)P(A \mid \lambda_o, \lambda_e)$$

where $\Theta = (\Theta(r_1, r_2))$ is the joint distribution of a pair of aligned residues. The marginal distributions are $\Theta(r_1, \cdot)$ and $\Theta(\cdot, r_2)$. In this notation, we can write that:

$$\log P(R^1, R^2 | A, \Theta) = \sum_{j=1}^{n_1} \log \Theta(r_j^1, \cdot) + \sum_{k=1}^{n_2} \log \Theta(\cdot, r_k^2) + a_{j,k} \log \Psi_{r_j^1, r_k^2}$$

where $\log \Psi_{r_i^1, r_j^2} = \log \Theta(r_i^1, r_j^2) - \log \Theta(r_i^1, \cdot) - \log \Theta(\cdot, r_i^2)$ corresponds to a scoring matrix, say a PAM or Blosum matrix.

One can remove two of the three unknowns as follows:

$$P(R^1, R^2 \mid \Lambda) = \frac{\Sigma_\Theta \Sigma_A P(R^1, R^2 \mid A, \Theta) P(\Theta) \lambda_o^{kg(A)} \lambda_e^{lg(A)-kg(A)}}{\Sigma_{A'} \lambda_o^{kg(A')} \lambda_e^{lg(A')-kg(A')}} \quad (22)$$

where in the numerator the Π is marginalized by summing over all the scoring matrices in a given set, each with prior 'weight' $P(\Pi)$. Both the numerator and the denominator of equation (21) can be computed via a recursive algorithm shown as follows, which is similar to the dynamic programming of Needleman and Wunsch (1970).

As with the traditional alignment algorithm, we can describe a path as consecutive moves of three types: $\rightarrow$ (deletion), $\downarrow$ (insertion) and $\searrow$ (match). To ensure uniqueness, one often adds the restriction that an insertion ($\downarrow$) cannot follow a deletion ($\rightarrow$). For example, to obtain the numerator of equation (21), we define $p(k,l)$, $p_m(k,l)$, $p_i(k,l)$ and $p_d(k,l)$, where:

$$\begin{aligned} p_m(k,l) &= p(k-1, l-1)\Theta(r_k^1, r_l^2) \\ p_i(k,l) &= \left[\lambda_e p_i(k-1, l) + \lambda_o p_m(k-1, l)\right]\Theta(r_k^2, \cdot) \\ p_d(k,l) &= \left[\lambda_e p_d(k, l-1) + \lambda_o p_m(k, l-1) + \lambda_o p_i(k, l-1)\right]\Theta(\cdot, r_l^2) \\ p(k,l) &= p_m(k,l) + p_i(k,l) + p_d(k,l) \end{aligned}$$

If the model uses only the interaction term, as is traditional in bioinformatics, instead of the joint distribution, all the marginal terms $\Theta(r_k^2, \cdot)$ and $\Theta(\cdot, r_l^2)$ can be substituted by 1, and $\Theta(r_k^1, r_l^2)$ by $\Psi(r_k^1, r_l^2)$ in the forgoing recursive formulas. The marginal likelihood can be obtained as $P(R^1, R^2) = \int P(R^1, R^2 \mid \Lambda) P(\Lambda) d\Lambda$. We know of no ways to complete this integration analytically. Traditional numerical integration methods work well for this low-dimensional integration. With the marginal likelihood, we can have the desirable posterior distribution such as:

$$P(\Lambda \mid R^1, R^2) = \sum_A \sum_\Theta P(\Theta, A, R^1 R^2 \mid \Lambda) P(\Lambda) / P(R^1, R^2)$$

To illustrate, we examined the alignment of two sequences: hemoglobin α and β chains, and use the PAM-80 matrix. The posterior joint distribution of the gap penalty parameters, i.e.

$$P(\lambda_o, \lambda_e \mid R)$$

is one of the outputs. As shown in Figure 7b, the posterior distribution of these parameters differs substantially from the prior distribution, indicating that the data have a strong influence on the results. The marginal posterior distributions of λ_o and λ_e are shown in Figure 8.

Motif-based alignment. While the gap penalty-based approaches have dominated alignment methods for many years, Bayesian statistics opens up new directions in dealing with insertions and deletions in alignments. In the procedure just described, the gap parameters are primarily used to control the prior distribution for the alignment (e.g. penalizing exponentially growing number of ways of gap opening). In contrast, Zhu *et al*. (1998) attack the problem by directly specifying a prior alignment distribution: all alignments with *k* gaps are equally likely, and the probability on the distribution of *k* is uniform. This prior discounts alignment with many gaps by penalizing it with a factor that is inversely proportional to the number of that type of alignments. Input requirements for the scoring matrices are also more flexible in the Bayesian setting than in traditional methods. For example, Zhu *et al*. (1997, 1998) examine the use of a series of either the PAM or the Blosum matrices as prior input in which all the matrices are assigned equal probability a priori. They report that the posterior distributions of the scoring matrices are often flat and sometimes multimodal, indicating that no one matrix is clearly more preferable to others when aligning the two sequences. One multimodal case is shown in Figure 9, in which there are strong modes at PAM 140 and PAM 80. This result illustrates two further features of Bayesian procedures. To examine these, we consider the expression for the posterior distribution of the scoring matrix.

$$P(\Theta \mid R) = \frac{1}{P(R)} \sum_\Lambda \sum_A P(R^1, R^2 \mid A, \Theta) P(\Theta) P(A \mid \Lambda) P(\Lambda)$$

First, we see that this posterior is obtained by averaging over all alignments. Hence, a 'good' alignment is not required to assess the distances between the sequences. This feature may be of value to distance methods employed in molecular evolution studies, since the requirement that a pair of sequences must be sufficiently close to permit a good alignment is removed. Furthermore, samples from these distance distributions can be employed to incorporate alignment uncertainty into phylogenetic tree construction. Secondly, this posterior distribution of the sequence distance incorporates variations in the alignments, which means that varying levels of sequence conservation in different regions of protein sequences can be detected. Zhu *et al*. (1998) show that bimodality in distance is a reflection of the variable degrees to which motifs in the GTPase sequences they compared are conserved.

Relationship to other methods. In many cases of applying the traditional alignment procedure, one may be tempted to optimize simultaneously over all possible alignments and a set of parameter values of Θ and Λ, but this approach is problematic because it often leads to non-ignorable bias (Little and Rubin, 1987). An approach that does not share this difficulty is based on the observed data likelihood, i.e. the one obtained by summing over (or integrating out) all possible values of the missing data in the complete-data likelihood:

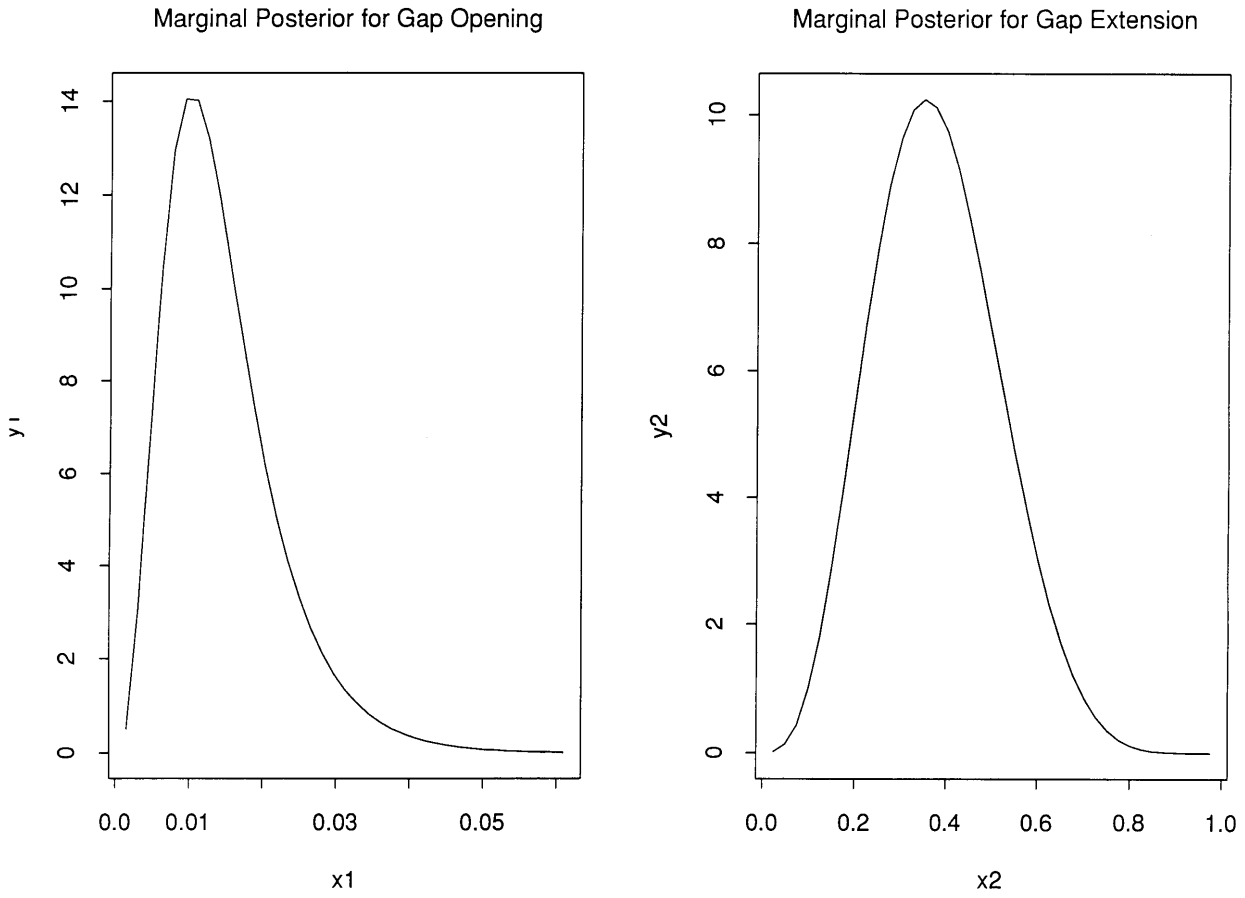

Fig. 8. Marginal posterior distributions of the gap penalties for the alignment of the hemoglobin α and β chains.

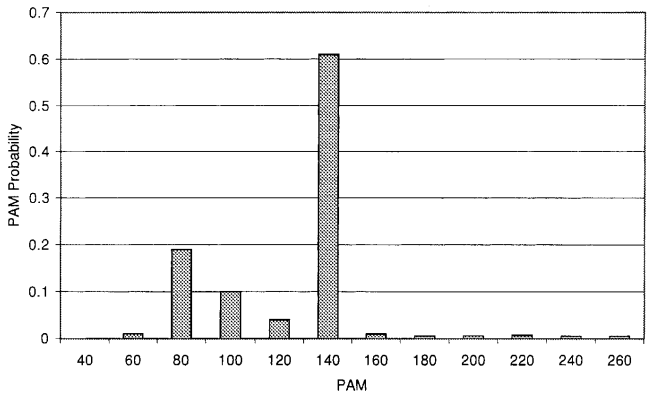

Fig. 9. Posterior distribution of the PAM distance between LETU and 1GIA.

$$P(R \mid \Theta, \Lambda) = \sum_{\text{All } A} P(R, A \mid \Theta, \Lambda) \qquad (23)$$

Just as in the Bayesian algorithms described above, these summations are completed using recursive relations derived from dynamic programming algorithms. Several authors have presented algorithms which find the optimal values of the parameters, Θ*, Λ*, for the observed data likelihood. For example, Churchill (1989) uses the maximum likelihood method to characterize the compositional heterogeneity of nucleotide sequences. Thorne *et al*. (1991) give a maximum likelihood method for the alignment of a pair of nucleotide sequences, and Allison *et al*. (1992) give a related method to choose between alternate alignment models. Their procedures yield point estimates of the Θ and Λ, but provide no information about uncertainties in these estimates or the effect of these uncertainties on the other unknowns. Under certain conditions, confidence limits based on asymptotic normality of these estimates can be obtained. However, no procedures are available to assess either the impact of uncertainties in these parameters, or the effects of using the optimized values of Θ and Λ, on the alignment.

Discussion

Since bioinformatics concerns the analysis of biopolymer sequence data, its main products are inferences about unobserved variables. As in classical statistics, optimization has

been the primary tool for making inference in bioinformatics, in which point estimates of very high-dimensional objects, obtained by using dynamic programming, are frequently used. Characterizations of uncertainties in these estimates have been very difficult and are mostly limited to a simple significance test or completely ignored.

We show in this article that the rich history of computation in bioinformatics can be adapted to meet the requirements of the Bayesian methods. Specifically, dynamic programming recursions can be modified to complete the high-dimensional summations required in Bayesian analyses. Through the use of these recursions, coupled with specific approaches to integrate out or sum over all other variables, the full power of the Bayesian methodology can be brought to bear on a wide range of problems previously addressed by dynamic programming. The fruits of this Bayesian approach include the following: (i) full inferences on all unknowns, with all uncertainties incorporated; (ii) a general and broad relaxation of the traditional fixed parameter settings; (iii) assessments of significance through the use of Bayesian model selection procedures.

The most important limitation on the Bayesian method is the need for additional computational resources. While Bayesian algorithms generally have time and space requirements of the same order as their dynamic programming counterparts, the constants are generally larger by an order of magnitude or more. As a result of the combination of previously developed efficient algorithms and the availability of fast workstations with large memories, this limitation is not a serious one for most applications. As discussed only briefly here, for those problems for which no polynomial time algorithm exists, such as multiple sequence alignment, Markov Chain Monte Carlo (and perhaps other Monte Carlo approaches) provide alternative means to implement a full Bayesian analysis. When this is the case, bioinformatics joins the majority of the field of applied statistics and statistical physics in the need to rely on algorithms whose convergence cannot be guaranteed.

Only recently have explicit statistical approaches, in the form of hidden Markov models and Gibbs sampling algorithms, come to play a significant role in bioinformatics. While these approaches have brought a number of algorithmic advances to bioinformatics, the potential for statistical inference has gone largely underexploited. As illustrated here, because Bayesian statistics is so well suited to bioinformatics, it provides a facile route to unleash the power of statistical inference in bioinformatics.

Acknowledgements

We thank Ivan Auger of the Wadsworth Center for technical help with the segmentation example. We also thank David Landsman and Tyra Wolfsberg for providing sequence data for the histone h1 example. This work was partially supported by NIH grant R011HG01257 and DOE grant DEFG0296ER7266 to C.E.L., and NSF grants NSF-9501570 and NSF-9803649 and by the Terman fellowship from Stanford University to J.S.L.

References

Allison,L., Wallace,C.S. and Yee,C.N. (1992) Minimum message length encoding evolutionary trees and multiple alignment. *Proceedings of 25th Hawaii International Conference on System Science*, **1**, 663–674.

Auger,I.E. and Lawrence,C.E. (1989) Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, **51**, 39–54.

Baldi,P., Chauvin,Y., McClure,M. and Hunkapiller,T. (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059–1063.

Bishop,M.J. and Thompson,E.A. (1986) Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.*, **190,** 159–165.

Cardon,L.R. and Stormo,G.D. (1992) Expectation maximization algorithm for identifying binding sites with variable lengths from unaligned DNA fragments. *J. Mol. Biol.*, **223**, 159–170.

Churchill,G.A. (1989) Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, **51**, 79–94.

Dayhoff,M.E., Eck,R.V. and Park,C.M. (1972) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Vol. 5, 89–99.

Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Ser. B*, **39**, 1–38.

Gelfand,A.E. and Smith,A.F.M. (1990) Sampling-based approach to calculating marginal densities. *J. Am. Stat. Assoc.*, **85**, 398–409.

Gelman,A., Carlin,J.B., Stern,H.S. and Rubin,D.B. (1995) *Bayesian Data Analysis*. Chapman & Hall, New York.

Henikoff,S. and Henikoff,JG. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Holmes,I. and Durbin,R. (1998) Dynamic programming alignment accuracy. *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology*, **2**, 102–108.

Kass,R.E. and Raftery,A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.

Krogh,A., Mian,I.S. and Haussler,D. (1994a) A hidden Markov model that finds genes in *E.coli* DNA. *Nucleic Acids Res.*, **22**, 4768–4778.

Krogh,A., Brown,M., Mian,S., Sjolander,K. and Haussler,D. (1994b) Protein modeling using hidden Markov models. *J. Mol. Biol.*, **235**, 1501–1531.

Kyte,J. and Doolittle,R.P. (1982) A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Little,R.J.A. and Rubin,D.B. (1987) *Statistical Analysis with Missing Data*. Wiley & Sons, New York.

Liu,J.S. (1994) The collapsed Gibbs sampler with applications to a gene regulation problem. *J. Am. Stat. Assoc.*, **89**, 958–966.

Liu,J.S. and Lawrence,C.E. (1996) Unified Gibbs method for biological sequence analysis. *Proc. Am. Stat. Assoc., Biometrics Section*, 194–199.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618–1632.

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

Schmidler,S.C., Liu,J.S. and Brutlag,D.L. (1998) Bayesian segmentation of protein secondary structure. Technical Report, Department of Statistics, Stanford University.

Snyder,E.E. and Stormo,G.D. (1995) Identification of protein coding regions in genomic DNA. *J. Mol. Biol.*, **248**, 1–18.

Tanner,M.A. and Wong,W.H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Stat. Assoc.*, **82**, 528–550.

Thorne,J.L., Kishino,H. and Felsenstein,J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, **33**, 114–124.

Thorne,J.L., Kishino,H. and Felsenstein,J. (1992) Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.*, **34**, 3–16.

Waterman,M.S. (1995) *Introduction to Computational Biology*. Chapman & Hall, New York.

Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.

Zhu,J., Liu,J.S. and Lawrence,C. (1997) Bayesian adaptive alignment and inference. *ISMB*, **5**, 358–368.

Zhu,J., Liu,J.S. and Lawrence,C.E. (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, **14**, 25–31.

Zuker,M. (1989) Computer prediction of RNA structure. *Methods Enzymol.*, **180**, 262–288.