Bulletin of Mathematical Biology Vol. 51, No. 1, pp. 79–94, 1989.
Printed in Great Britain.

0092–8240/89$3.00+0.00
Pergamon Press plc

STOCHASTIC MODELS FOR HETEROGENEOUS DNA SEQUENCES

■ GARY A. CHURCHILL
Department of Biostatistics,
University of Washington,
Seattle, WA 98195, U.S.A.

The composition of naturally occurring DNA sequences is often strikingly heterogeneous. In this paper, the DNA sequence is viewed as a stochastic process with local compositional properties determined by the states of a hidden Markov chain. The model used is a discrete-state, discrete-outcome version of a general model for non-stationary time series proposed by Kitagawa (1987). A smoothing algorithm is described which can be used to reconstruct the hidden process and produce graphic displays of the compositional structure of a sequence. The problem of parameter estimation is approached using likelihood methods and an EM algorithm for approximating the maximum likelihood estimate is derived. The methods are applied to sequences from yeast mitochondrial DNA, human and mouse mitochondrial DNAs, a human X chromosomal fragment and the complete genome of bacteriophage lambda.

1. DNA Composition. A Deoxyribonucleic acid (DNA) molecule consists of a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, a five carbon sugar and a phosphate group. There are four different bases. The pyrimidines, thymine (T) and cytosine (C) have a six member carbon-nitrogen ring structure. The purines, adenine (A) and guanine (G), have fused five and six member rings. The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with the 5′ carbon of one sugar linked to the 3′ carbon of the next, giving the string an orientation. DNA occurs naturally as a double helix composed of two polynucleotide strands with the bases facing inwards. The two single strands run antiparallel and are connected by hydrogen bonding between complementary bases. Guanine pairs specifically with cytosine, forming three hydrogen bonds, and adenine pairs with thymine forming two hydrogen bonds. Thus, in a double stranded DNA (dsDNA) molecule, the total amount of G equals the amount of C and the total amount of A equals that of T. The amount of G + C is variable and characteristic of individual DNA molecules.

Because the two strands are complementary, it is sufficient to represent a DNA molecule by the sequence of bases on a single strand, written in the standard 5′ to 3′ direction. In addition to this four letter representation, it will be of interest to consider binary representations. The purine–pyrimidine (AG–TC) and the strong–weak hydrogen-bonding (GC–AT) classifications will be considered here. Taken together, these two binary representations uniquely determine the four base representation. Other binary representations, such as the keto-amine classification (GT–AC), might also be of interest.

Early studies of the compositional properties of DNA sequences relied on indirect methods, such as base composition determination or the analysis of nearest neighbor frequencies. The measurement of heterogeneity by density gradient centrifugation was first investigated both experimentally and theoretically by Sueoka (1959). He concluded that most of the observed density heterogeneity is due to base composition variation. Elton (1974) compiled data on compositional heterogeneity, mainly from bacteria. He found that models with homogeneous probabilistic structure could not provide adequate descriptions of the variation and proposed a model in which the DNA consists of a sequence of "segments" with different underlying compositions.

More recently, it has been proposed that the nuclear genomes of warmblooded vertebrates are composed of large segments (>300 kilobases) which have fairly homogeneous G+C content within themselves but fall into a small number of distinct classes with different characteristic proportions of G+C (Bernardi *et al.*, 1985). These large regions, the isochores, are interspersed in some as yet undetermined fashion throughout each chromosome and seem to correspond to the DNA segments seen in Giemsa and reverse chromosomal bands. The distribution of genes is shown to be biased in favor of the GC rich isochores. This fundamental organization of DNA is likely to be related to a number of structural and functional properties of chromosomes most of which can only be conjectured at this time.

2. *Modeling DNA.* With the advent of sequencing technology, it has become possible to study the compositional properties of DNA directly. The DNA sequence data is viewed here as a stochastic process. It is not necessary to suppose that the sequence is generated by an actual random mechanism. Rather, the model is viewed as a useful tool with which the large and complex data contained in a sequence can be summarized and used to answer specific questions. It is a method for extracting information. The forces of evolution which have formed the DNA sequences of living organisms may be in some aspects random but are complex and variable and no simple model could hope to describe the endpoint. A good model aims to reveal important features, not to mimic nature in every detail.

Classically, Markov chains with stationary transition probabilities have been used to model discrete sequences. When the composition of a DNA molecule is fairly homogeneous, such models provide good descriptions of local structure. One problem with the Markov chain approach is that it requires the same properties to hold throughout the length of the sequence. It can often be observed that different regions of the same DNA molecule display quite different patterns of base composition and dependence between neighboring bases. Compositional variation is likely to reflect functional or structural differences between regions. When the homogeneity assumption is found not to hold, it will be of interest to identify distinct regions of the sequence and to characterize their properties locally.

Ad-hoc tests for heterogeneity can be constructed by dividing a sequence into k segments of equal length and using standard chi-square methods to test for differences in the proportions of single bases or dinucleotides. Another method of characterizing heterogeneity is to scan the sequence with a fixed-size window and compute summary statistics of local composition (Staden, 1984). Both of these methods involve an arbitrary choice of either the number of segments or the window size. The first method provides a test statistic but does not identify regions or estimate local properties. The window scanning method is a powerful tool which can provide useful graphic summaries of the local properties of a sequence but it does not provide a method of determining significant departures from homogeneity.

In this paper, sequences will be assumed to have mosaic structure composed of segments which are homogeneous in composition but may differ from one another. Each segment can be classified into one of a finite number of states. The states represent an underlying structure to the sequence and will be assumed to evolve slowly according to a hidden Markov process. These assumptions will lead to a large and flexible class of models which can be used to produce graphic summaries, estimates of local properties and test statistics without the need for arbitrary specifications such as a window size.

Work on hidden Markov chain models can be traced back to Ott (1967) in the context of coding and information theory. The general form of the state-space models, as defined in equations (1) through (4) below, is due to Kitagawa (1987). Discrete-state, discrete-outcome versions of the state-space models will be developed here and used to reconstruct estimates of the underlying states. Parameter estimation is based on likelihood methods using an EM algorithm (Dempster *et al.*, 1977). Baum *et al.* (1970) describe an iterative technique for maximizing the likelihood in hidden Markov chain models and provide some of the key results later incorporated in Dempster *et al.* (1977). However, their approach does not involve explicit reconstruction of the underlying process.

3. The General State-Space Model. Consider a sequence of random variables $\{Y_i\colon i=1, \ldots, n\}$ with distributions determined by a corresponding sequence of unobservable states $\{s_i\}$. Denote the sequence of observed outcomes up to time t by $y^t=\{y_1, \ldots, y_t\}$ and similarly, the states $s^t=\{s_1, \ldots, s_t\}$. The probability distribution of an observation given the current state and past observations is called the *observation equation*, denoted $\Pr(y_t|s_t, y^{t-1})$. In the case of independent observations, this will be equal to $\Pr(y_t|s_t)$. The sequence of states represents an underlying structure which is not directly observed but can be inferred from the observations. The states are thought of as evolving slowly according to a set of system equations, denoted $\Pr(s_t|s^{t-1})$. These will be assumed to have the Markov property, so that $\Pr(s_t|s^{t-1})=\Pr(s_t|s_{t-1})$.

The problem addressed here is that of estimating the states given the sequence of observed outcomes. The observation and system equations are assumed to be completely specified. The marginal posterior distribution of the state at time t, $\Pr(s_t|y^n)$, will be called the smoothed estimate of s_t. An algorithm will be described for computing this distribution and related quantities. Graphic displays of the underlying state process will be produced by plotting the smoothed estimates against the sequence index t.

Filter: to begin, suppose that $\Pr(s_{t-1}|y^{t-1})$ is known. A prediction of the state at time t can be computed using the system equations. By the law of total probability, this is:

$$\Pr(s_t|y^{t-1})=\int \Pr(s_t|s_{t-1}, y^{t-1})\Pr(s_{t-1}|y^{t-1})\,\mathrm{d}s_{t-1}. \tag{1}$$

Next, the information in the current observation is incorporated by updating the predictive density. The filtering density is:

$$\Pr(s_t|y^t)=\frac{\Pr(y_t|s_t, y^{t-1})\Pr(s_t|y^{t-1})}{\Pr(y_t|y^{t-1})}, \tag{2}$$

by Bayes' theorem, where:

$$\Pr(y_t|y^{t-1})=\int \Pr(y_t|s_t, y^{t-1})\Pr(s_t|y^{t-1})\,\mathrm{d}s_t.$$

Smoother: the joint distribution of two successive states can be written, using the definition of conditional probability, as:

$$\begin{aligned}\Pr(s_t, s_{t+1}|y^n)&=\Pr(s_{t+1}|y^n)\Pr(s_t|s_{t+1}, y^n)\\&=\Pr(s_{t+1}|y^n)\Pr(s_t|s_{t+1}, y^t)\\&=\frac{\Pr(s_{t+1}|y^n)\Pr(s_t,s_{t+1}|\, y^t)}{\Pr(s_{t+1}|y^t)}\\&=\frac{\Pr(s_{t+1}|y^n)\Pr(s_{t+1}|s_t)\Pr(s_t|y^t)}{\Pr(s_{t+1}|y^t)}.\end{aligned} \tag{3}$$

The first equality follows from the conditional independence of s_t and y_i $(i>t)$ given s_{t+1}. The last two follow from the definition of conditional probability. The marginal distribution of the state at time t is obtained from the joint distribution by integration:

$$\begin{aligned}\Pr(s_t|y^n)&=\int \Pr(s_t, s_{t+1}|y^n)\,\mathrm{d}s_{t+1}\\&=\Pr(s_t|y^t)\int\frac{\Pr(s_{t+1}|y^n)\Pr(s_{t+1}|s_t)}{\Pr(s_{t+1}|y^t)}\,\mathrm{d}s_{t+1}.\end{aligned} \tag{4}$$

Thus, the smoothed estimate at time t can be expressed in terms of the system equations, quantities derived from filtering, and the smoothed estimate at time $t+1$.

Recursive updating algorithm. The filtering and smoothing equations suggest an algorithm for computing $\Pr(s_t|y^n)$ and related quantities.

(1) Start with an initial prediction $\Pr(s_0)$.

(2) Use the starting value to obtain the predictive densities $\Pr(s_t|y^{t-1})$ and the filtered densities $\Pr(s_t|y^t)$.

(3) Using the quantities $\Pr(s_n|y^n)$ and $\Pr(s_n|y^{n-1})$ from the filtering step, compute $\Pr(s_{n-1}|y^n)$.

(4) Apply step 3 recursively to complete the smoothing.

The initial prediction is usually taken to be the stationary distribution of the state process or a fixed constant value. When the sequence is circular, it will be convenient to define $s_0 \equiv s_n$ and $y_0 \equiv y_n$. Using an appropriate starting value, the sequence should be filtered once to obtain $\Pr(s_n|y^n)$ and filtered again using this as the new starting value. The sequence can then be smoothed to obtain $\Pr(s_0|y^n)$ and smoothed again using this as the new starting value in step 3. This procedure for circular smoothing assumes that the sequence is sufficiently long that the effects of the first pass starting values on the values obtained at the other "end" of the sequence are negligible.

4. *The Discrete Case.* In applications to DNA, the observed outcomes correspond to the bases of the sequence. The states will be assumed to be fixed and finite in number and correspond to the different regions or segments of the DNA. The discrete state-space models described here are discrete in both the outcomes and the states. Models which admit a continuum of states are also possible within the same framework but will not be pursued here.

The simplest case is the two-state model for independent binary sequences. Consider a sequence of independent binary outcomes $y_t \in \{0, 1\}$ whose success probabilities depend on an underlying state $s_t \in \{0, 1\}$. The observation equations are binomial:

$$\Pr(y_t|s_t=j)=p_j^{y_t}(1-p_j)^{1-y_t}, j\in\{0, 1\}, \tag{5}$$

where $p_0=\Pr(y_t=1|s_t=0)$ and $p_1=\Pr(y_t=1|s_t=1)$ are known and assumed to be not equal. The state process alternates between state 0 and state 1 according to the system equations:

$$\Pr(s_t = j | s_{t-1} = i) = \lambda_{ij},\ i \in \{0, 1\}, j \in \{0, 1\}, \tag{6}$$

where $\sum_{j=0}^{1} \lambda_{ij} = 1$. The transition probability matrix for this binary Markov process will be denoted:

$$\Lambda = \begin{bmatrix} 1-\lambda & \lambda \\ \tau & 1-\tau \end{bmatrix}.$$

The probabilities λ and τ are known and assumed to be small, so that the states tend to persist. The sizes of the different regions will have geometric distributions with means equal to the reciprocals of the transition probabilities.

When the initial state is fixed at zero, $\Pr(s_0 = 0) = 1$, and transitions from state 1 to state 0 are not allowed, $\tau = 0$, the two-state model corresponds to the single change-point model described by Hinckley (1970). The process starts in state 0 and has, at each time point, a constant probability of switching to state 1. Once the process is in state 1, it persists. A Bayesian solution to problems of inferences concerning the location of the change-point is given by Smith (1975). If the prior distribution for the location of the change-point is taken to be geometric with parameter λ, his results are identical to those obtained by smoothing. The posterior density of a change point at time t is given by $\Pr(s_t = 0,\ s_{t+1} = 1 | y^n)$ and the probability of no change-point is $\Pr(s_n = 0 | y^n)$.

A more general version of the discrete state-space model admits a fixed, finite number of states and multinomial outcomes. Let $y_t = (y_{t,0}, \ldots, y_{t,m-1})$ be a vector whose components are all zero except for one equal to unity, indicating which of the m possible outcomes is observed. Each observation is associated with one of r states indicated by the vector $s_t = (s_{t,0}, \ldots, s_{t,r-1})$. The distribution of y_t given $s_t = k$ is multinomial. The parameter $p_{i,k}$ is the probability of observing outcome i when the current state is k, subject to the constraint $\sum_{i=0}^{m-1} p_{i,k} = 1$. The observation equations are:

$$\Pr(y_t | s_t = k) = \prod_{i=0}^{m-1} p_{i,k}^{y_{t,i}}. \tag{7}$$

It is also possible to allow for Markov dependence between the observed outcomes. In the case of first-order dependence, the probability of observing outcome j given that the previous outcome was i and the current state is k is $p_{ij,k}$, where $\sum_{j=0}^{m-1} p_{ij,k} = 1$. The system equations are:

$$\Pr(y_t | y_{t-1}, s_t = k) = \prod_{i=0}^{m-1} \prod_{j=0}^{m-1} p_{ij,k}^{y_{t-1,i} y_{t,j}}. \tag{8}$$

The state process is a Markov chain on the r states. Denoting the $r \times r$ matrix of state transition probabilities by $\Lambda = [\lambda_{ij}]$, the system equations can be written as:

$$\Pr(s_t|s_{t-1}) = \prod_{i=0}^{r-1} \prod_{j=0}^{r-1} (\lambda_{ij})^{s_{t,i}s_{t-1,j}}. \tag{9}$$

The recursive updating algorithm can be applied as follows to reconstruct the underlying state process. The general filtering and smoothing equations are replaced by their discrete-state analogs. The integrals are taken with respect to counting measure ds_t. The predictive densities [equation (1)] become:

$$\Pr(s_{t,j}|y^{t-1}) = \sum_{i=0}^{r-1} \lambda_{ij} \Pr(s_{t-1,i}|y^{t-1}), \tag{10}$$

and the filtered densities [equation (2)] are:

$$\Pr(s_{t,j}|y^t) = \frac{p_{y_{t,j}} \Pr(s_{t,j}|y^{t-1})}{\sum_{i=0}^{r-1} p_{y_{t,i}} \Pr(s_{t,i}|y^{t-1})}. \tag{11}$$

The joint distribution of adjacent states [equation (3)] is:

$$\Pr(s_{t,i}, s_{t+1,j}|y^n) = \frac{\Pr(s_{t+1,j}|y^n)\lambda_{ij}\Pr(s_t|y^t)}{\Pr(s_{t+1,j}|y^t)} \tag{12}$$

and the smoothed estimates [equation (4)] are:

$$\Pr(s_{t,i}|y^n) = \Pr(s_{t,i}|y^t) \sum_{j=0}^{r-1} \frac{\Pr(s_{t+1,j}|y^n)\lambda_{ij}}{\Pr(s_{t+1,j}|y^t)}. \tag{13}$$

5. *Parameter Estimation.* The recursive updating algorithm requires that the parameters of the observation and system equations be specified. Typically, these values will not be known and will have to be estimated from the data. In this section, a method of obtaining a maximum likelihood estimate (MLE) of the parameter vector $\phi = \{p, \Lambda\}$ will be described. The MLE, denoted $\hat{\phi}$, is that value of ϕ which maximizes the likelihood function $L(\phi) = \Pr(y^n|\phi)$. The likelihood function for state-space models can be conveniently computed using a formula based on the observation equations and the predictive densities

$$\Pr(y^n) = \prod_{t=1}^{n} \Pr(y_t|y^{t-1})$$

$$= \prod_{t=1}^{n} (\int \Pr(y_t|s_t, y^{t-1})\Pr(s_t|y^{t-1})\, ds_t), \tag{14}$$

where, for notational simplicity, dependence on ϕ has been suppressed.

The usual method of finding a MLE is to set the vector of first partial derivatives of the log-likelihood equal to zero and to solve the resulting set of likelihood equations. However, the likelihood equations for state-space models are not easily derived and do not yield analytic solutions (Baum *et al.*, 1970). We resort to seeking approximations to the MLE using an iterative procedure. The method used is a specialization of a general algorithm for estimation in an incomplete data context which was formalized by Dempster, Laird and Rubin (1977) and termed by them the EM algorithm.

In the context of state-space models, imagine that, at each time point, the current state s_t could be observed. The likelihood for this augmented-data problem is the product of the observation and state equations:

$$\Pr(y^n, s^n) = \prod_{t=1}^{n} \Pr(y_t|s_t, y^{t-1})\Pr(s_t|s^{t-1}). \tag{15}$$

In the general discrete case, the augmented-data likelihood for circular sequences ($s_0 \equiv s_n$) with independent outcomes is:

$$\Pr(y^n, s^n) = \prod_{t=1}^{n}\left[\prod_{i=0}^{m-1}\prod_{j=0}^{r-1} p_{i,j}^{y_{t,i}s_{t,j}}\right] \times \prod_{t=1}^{n}\left[\prod_{i=0}^{r-1}\prod_{j=0}^{r-1} \lambda_{ij}^{s_{t-1,i}s_{t,j}}\right]. \tag{16}$$

When the sequence is linear, the additional parameter $\pi = \Pr(s_1)$ is introduced. The augmented data likelihood is modified by taking the product over the system equations from $t=2$ to n and including the factor $\prod_{j=0}^{r-1}\pi_j^{s_{1,j}}$. For sequences with first-order dependence between outcomes, the observation equation term becomes:

$$\prod_{i=0}^{m-1}\prod_{j=0}^{m-1}\prod_{k=0}^{r-1} p_{ij,k}^{y_{t-1,i}y_{t,j}s_{t,k}}.$$

The augmented-data likelihood is a member of the regular exponential family and simple closed-form solutions to the likelihood equations exist. For circular sequences with independent outcomes, the maximum likelihood parameter estimates are:

$$\hat{p}_{i,j} = \frac{\sum_{t=1}^{n} y_{t,i}s_{t,j}}{\sum_{t=1}^{n} s_{t,j}} \tag{17}$$

$$\hat{\lambda}_{ij} = \frac{\sum_{t=1}^{n} s_{t-1,i}\cdot s_{t,j}}{\sum_{t=1}^{n} s_{t-1,i}}. \tag{18}$$

When first-order Markov dependence between outcomes is present:

$$\hat{p}_{ij,k} = \frac{\sum_{t=1}^{n} y_{t=1,i} y_{t,j} s_{t,k}}{\sum_{t=1}^{n} y_{t-1,i} s_{t,k}}. \tag{19}$$

For linear sequences, the summations in equation (18) should run from $t = 2$ to n and the initial probability vector is estimated by $\pi = s_1$.

The EM algorithm is implemented as follows. Starting with an initial guess, $\phi^{(0)}$, of the parameter.

E-step. Run the recursive updating algorithm, with the current parameter estimate $\phi^{(p)}$. Estimate the states by their conditional expectations $E(s_t | y^n, \phi^{(p)})$.

M-step. Treating the estimated states as data, solve the augmented-data likelihood equations to obtain an updated estimate $\phi^{(p+1)}$.

The E-step and M-step are iterated until the estimate converges. Because the likelihood can be shown to be non-decreasing at each iteration, convergence to a local maximum or stable point is guaranteed for any bounded likelihood. If the likelihood surface is convex, convergence will always be to the global maximum.

6. *Examples.* The following series of examples illustrates the application of the methods described to actual DNA sequences. For the need of brevity, only minimal background information is provided. More details can be found in the cited references. In example A, the recursive updating algorithm is applied without formal parameter estimation to produce profile plots of G + C content in replication origins from yeast mtDNA. In example B, a change point model is used to isolate a potential isochore boundary region in a fragment of human X chromosomal DNA. Maximum likelihood estimates are used to produce a smoothed profile of the purine–pyrimidine distribution in two mammalian mtDNAs. The result is correlated with known functional features of these genomes. In the final example, smoothed estimates of A + T content in the phage lambda genome reveal a structure which is more complex than the simple mosaic structure implied by the model. An *ad-hoc* estimator of local composition is suggested. Graphic output was produced using the S system (Becker and Chambers 1984).

(*a*) *GC clusters in yeast mtDNA.* The mitochondrial genome of yeast is an 85 kb circular dsDNA molecule. Its sequence has been largely determined (de Zamaroczy and Bernardi 1986). The average G + C content of only 18% is one of the most extreme observed among naturally occurring DNA sequences. The genome appears to be composed of three types of segments distinguishable by their G + C content. The intergenic regions contain long stretches of DNA with less than 5% G + C content interspersed with short clusters of DNA with G + C content greater than 50%. The genes have moderate G + C contents ranging from 18 to 28%.

Respiration deficient mutants, called ρ^-, are formed by a deletion event in which most of the wild-type DNA (ρ^+) is lost. The small portion remaining is amplified, usually in tandem repeats, replicated and maintained in the mitochondrion. A subset of ρ^- mutants have the property that in matings with wild-type strains the ρ^- DNA frequently displaces the ρ^+ DNA in all of the diploid descendants of the zygotic cell. The sequences HS416 and HS3324 are examples of such hypersupressive ρ^- genomes. Both sequences contain one of several highly similar 300bp regions proposed to be primary origins of replication in wild-type mtDNA (Blanc and Dujon, 1980). The hypersuppressive phenotype is thought to result from the replicative superiority of tandem repeats of these origins. The sequences of two non-suppressive ρ^- mutants, *ori*°1 and *ori*°6, contain surrogate origins of replication with greatly reduced efficiencies (Goursot *et al.*, 1982).

Because these sequences contain no coding regions, a two-state model for circular binary sequences was used to compute smoothed estimates of the AT-rich and GC-rich states. The model parameters were set at $p_0 = 0.9$, $p_1 = 0.5$, $\lambda = 0.01$ and $\tau = 0.01$. The profiles of the hypersuppressive sequences (Figs 1a and b) reveal a similarity in general structure that spans over a 600bp region. The profiles of the non-suppressive sequences (Figs 1c and d) show different arrangements of GC-rich clusters. The minimal sequence requirements for maintenance of mtDNA in yeast have yet to be determined (Fangman, 1984) but the degree of replicative success in various ρ^- genomes may be related to the topological structure of the DNA resulting from the pattern of AT-rich and GC-rich sequences which they contain.

(*b*) *A human X chromosome fragment.* Xrep is a 2352bp human X chromosomal DNA fragment isolated because of its unusual ability to stimulate replication in bacterial plasmids (Riley *et al.*, 1987). It contains features similar to those found in eukaryotic viral origins of replication. Southern blot analysis shows that there are 2 or 3 non-identical copies of Xrep on the human X chromosome.

Preliminary analysis of Xrep base composition revealed heterogeneity in the hydrogen-bonding process. The overall G + C content is 57.8%. The sequence was divided into 19 segments of 120bp length and a 20th segment of 72bp. The chi-square statistic ($\chi^2_{19} = 77.9$, $p \ll 0.0001$) clearly indicates heterogeneity. Plots of the local base composition (Fig. 2a) using a range of window sizes suggest that the GC content of the 5′ end is near 50% and the GC content of the 3′ end is near 70% with the change occurring somewhere in the region from base 1500 to base 1800.

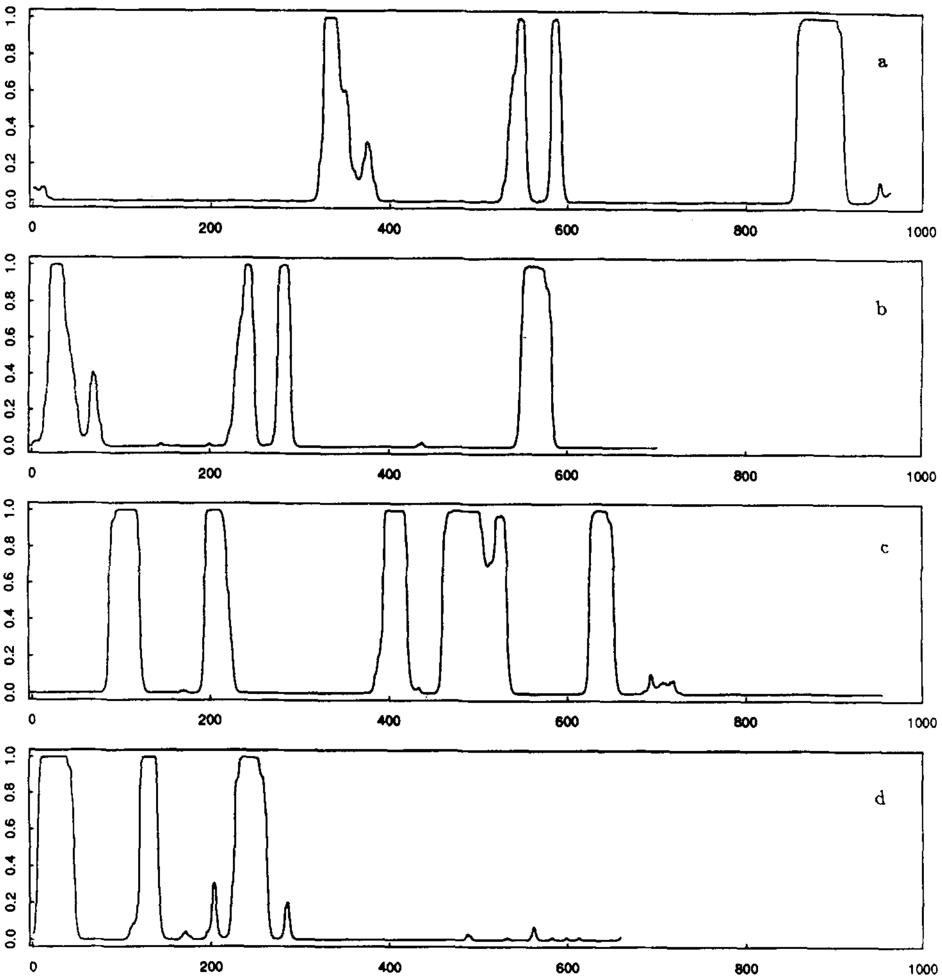

Figure 1. Yeast mtDNA GC-clusters. Smoothed estimates were computed for the hydrogen-bonding sequences of yeast mitochondrial petite genomes: HS3324 (a); HS416 (b); ori°1 (c); ori°6 (d). Posterior state probabilities $\Pr(s_t = 1|y^n)$ are plotted against the sequence index t. Peaks indicate GC-rich regions.

A change-point model was fit to the binary sequence using parameters $p_0 = 0.5$, $p_1 = 0.7$ and $\lambda = 0.0005$. The smoothed estimate of the state process (Fig. 2b) shows a sharp transition from state 0 to state 1 in the neighborhood of base 1600. The posterior density of the change-point (Fig. 2c) is highly concentrated with mode at base 1590 and a 90% maximum posterior density region extending from base 1571 to 1632.

These results suggest that the Xrep fragment may contain a boundary region between two isochores. It has been suggested that the coordination of DNA

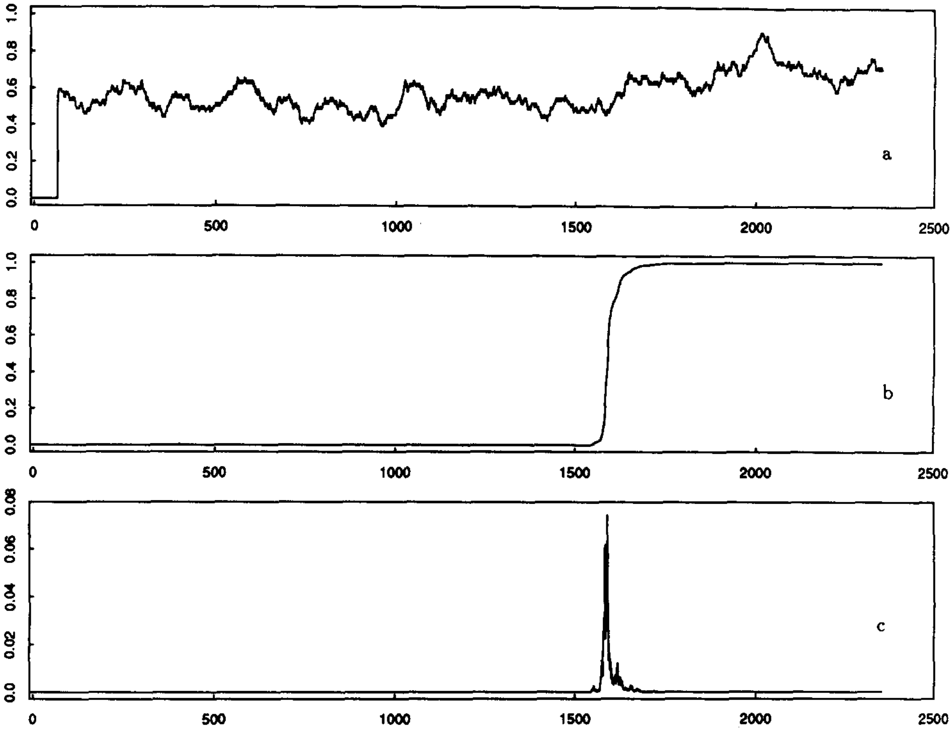

Figure 2. Xrep change-point model. The GC content averaged over a window of size 64 (a); smoothed estimates of the hydrogen bonding process (b); the posterior density of the change point (c); are plotted against the sequence index.

replication may be related to compositional compartmentalization in eukaryotic genomes (Bernardi *et al.*, 1985). Isolation of other potential replication origins and mapping of isochore boundaries may help reveal the mechanisms underlying this regulation.

(*c*) *Mammalian mitochondrial DNAs.* The sequence organization of mammalian mitochondrial genomes is very different from that of yeast. The small ($\approx$16kb) circular molecules are extremely compact, encoding several essential polypeptides as well as rRNAs and tRNAs. There are none or very few non-coding bases between genes and the only non-coding region is the D-loop region which acts as an origin for DNA replication. The human mtDNA sequence (Anderson *et al.*, 1981) is 16596bp in size with an overall AT content of 55.63%. The L-strand, which contains the sense sequence of the rRNAs, most tRNAs and mRNAs, has a purine content of 44.06%. The sequence of mouse mitochondrial DNA (Bibb *et al.*, 1981) is 16295bp in size and largely homologous with human mtDNA. Its AT content is 63.26%, higher than that of human mtDNA, and the purine content of the L-strand is 46.88%. Both sequences were obtained from the GenBank database.

Two-state independent base models were fit by maximum likelihood to the purine–pyrimidine processes of these sequences. The MLEs obtained for human mtDNA were $\hat{p}_0=0.425$, $\hat{p}_1=0.525$, $\hat{\lambda}=0.000115$ and $\hat{\tau}=0.000610$. The MLEs for mouse mtDNA were $\hat{p}_0=0.454$, $\hat{p}_1=0.541$, $\hat{\lambda}=0.000090$ and $\hat{\tau}=0.000434$. Plots of the smoothed estimates (Figs 3a and b) reveal in both sequences a single region of excess purine content which corresponds very closely with the 12s and 16s rRNA coding sequences.

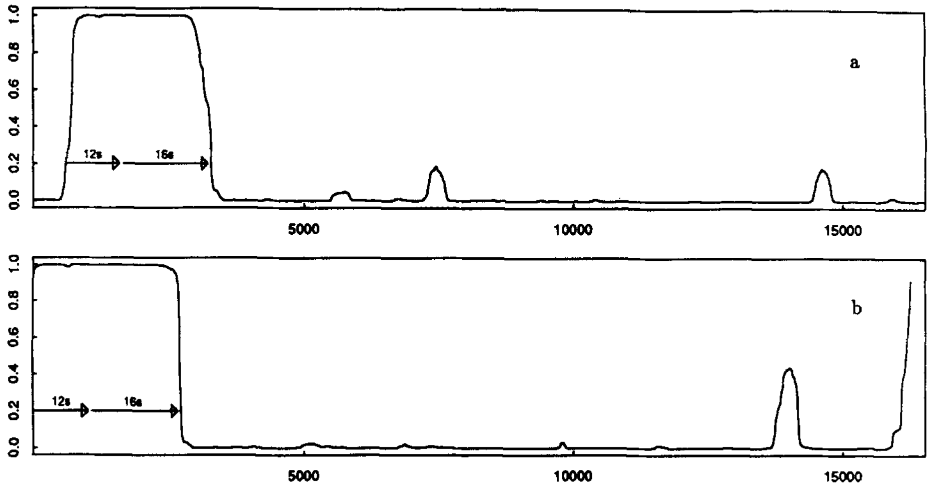

Figure 3. Mammalian mitochondrial DNA. The two-state smoothed estimates from maximum likelihood fits to the purine–pyrimidine processes of human (a) and mouse (b) mtDNA are plotted against the sequence index. Ribosomal RNA coding sequences are indicated by arrows.

The fact that the overall A + T contents of these two genomes have diverged significantly while the pattern of purine–pyrimidine composition has remained similar suggests that the pattern is linked to functional constraints on the molecules. The ribosomal RNAs encoded by the purine-rich region serve a direct functional role and it is perhaps not surprising that purine-content is conserved. The remainder of the genome encodes primarily messenger RNAs. The conservation of overall purine content there suggets that it is an important property of mRNAs in mammalian mitochondrial systems.

(*d*) *Bacteriophage lambda.* Skalka *et al.* (1967) used density gradient centrifugation techniques to determine the G + C of fragments from bacteriophage λ DNA fractionated in various ways. They concluded that the genome is composed of six segments with different G + C contents ranging from 37 to 57%. The segments were determined to be reasonably homogeneous internally with sharp boundaries between segments. Their results are summarized in Fig. 4a.

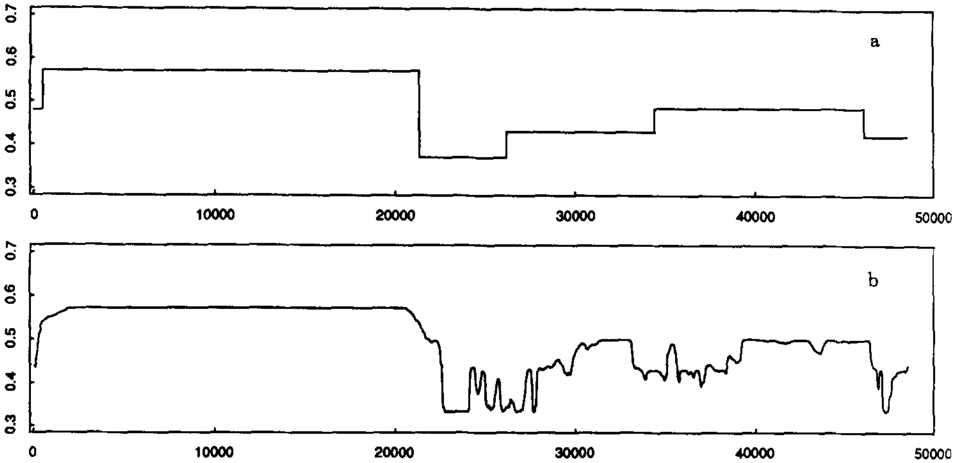

Figure 4. Bacteriophage lambda. Plots of the local proportions of A+T based on the six segment model of lambda composition proposed by Skalka *et al.* (1967) (a) and on a four-state model fit by maximum likelihood (b) are shown.

The sequence of the complete λ genome has been determined (Sanger *et al.*, 1982) and was obtained from the GenBank database. Likelihoods were computed for the four-base sequence using models with independent or first-order dependent outcomes and up to 6 underlying states were fit to this sequence. Using a Bayesian information criterion (Schwarz 1978), the three-state first-order dependent model was found to provide the best description of the data.

In order to make comparisons with the results of Skalka *et al.*, the maximum likelihood smoothed estimates were computed on the binary hydrogen bonding process. The local composition was estimated as:

$$\hat{\pi}_t = \sum_{i=0}^{r-1} \hat{p}_i \Pr(s_t = i | y^n). \tag{20}$$

The smoothed estimates of local composition based on a four-state model are shown in Fig. 4b. The agreement is in general very good. The left molecular half is largely homogeneous with ~55% G+C content and the right half is heterogeneous with an average G+C content of ~45%. Compositional fluctuations evident in the right molecular half suggest that a mosaic model may not provide the best description of the structure of phage lambda. A model which admits a continuum of states will perhaps provide a better fit.

8. Conclusions. In the near future, large scale projects are likely to be undertaken to determine the DNA sequences of entire genomes. It will become possible to study the relationships between DNA primary structure and the global organization of entire chromosomes. Methods will be needed to extract underlying features from large amounts of primary sequence data. Stochastic models can provide tools for summarizing and extracting the major features from complex data and will play an important role in understanding of DNA structure and function.

A powerful data analytic tool for studying compositional heterogeneity in DNA has been described. It provides results that are satisfying both theoretically and in applications. Mosaic structure appears to be widespread among naturally occurring DNA sequences and results obtained so far with discrete-state models confirm this. The framework of the general state-space model will allow further development of models which can capture more general heterogeneous structure as well.

The presence of compositional heterogeneity reflects constraints on genomes which are poorly understood at this time. These constraints appear to affect both coding and non-coding sequences in a wide range of organisms, suggesting a physiological role for non-coding DNA (Bernardi and Bernardi, 1986). Compositional features may affect the structure and stability of DNA at the chromatin and chromosomal levels and may play a role in the modulation of basic genome functions. Investigation of the evolutionary causes and the functional and structural implications of this phenomenon is a problem worthy of close attention.

LITERATURE

Anderson, S., A. T. Bankier, B. G. Barrell, M. H. L. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden and I. G. Young. 1981. "Sequence and Organization of the Human Mitochondrial Genome." *Nature* **290**, 457–464.

Baum, L. E., T. Petrie, G. Soules, N. Weiss. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabalistic Functions of Markov Chains." *Ann. Math. Statist.* **41**, 164–171.

Becker, R. A. and J. M. Chambers. 1984. *S—An Interactive Environment for Data Analysis.* Belmont, CA: Wadsworth.

Bernardi, G. and G. Bernardi. 1986. "Compositional Constraints and Genome Evolution." *J. Molec. Evol.* **24**, 1–11.

———, G., B. Olofsson, J. Filipski, M. Zerial, G. Cuny, M. Meunier-Rotival, F. Rodier. 1985. "The Mosaic Genome of Warm Blooded Vertebrates." *Science* **228**, 953–957.

Bibb, M. J., R. A. Van Etten, C. T. Wright, M. W. Walberg, D. A. Clayton. 1981. "Sequence and Gene Organization of Mouse Mitochondrial DNA." *Cell* **26**, 167–180.

Blanc, H. and B. Dujon. 1980. "Replicator Regions of the Yeast Mitochondrial DNA Responsible for Suppressiveness." *Proc. Natn. Acad. Sci. U.S.A.* **77**, 3942–3946.

de Zamaroczy, M., G. Bernardi. 1986. "The Primary Structure of the Mitochondrial Genome of *Saccharomyces cerevisiae*—a review." *Gene* **47**, 155–177.

Elton, R. A. 1974. "Theoretical Models for Heterogeneity of Base Composition in DNA." *J. Theor. Biol.* **45**, 533–553.

Dempster, A. P., N. M. Laird, D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *J. R. Statist. Soc.* **B39**, 1–22.

Fangman, W. L. and B. Dujon. 1984. "Yeast Mitochondrial Genomes Consisting of Only AT Base Pairs Replicate and Exhibit Suppressiveness." *Proc. Natn. Acad. Sci. U.S.A.* **81**, 7156–7160.

Goursot, R., M. Mangin, G. Bernardi. 1982. "Surrogate Origins of Replication in the Mitochondrial Genomes of *ori*° Petite Mutants of Yeast." *EMBO J.* **1**, 705–711.

Hinckley, D. V. 1970. "Inference About the Change Point in a Sequence of Random Variables." *Biometrika* **57**, 1–17.

Kitagawa, G. 1987. "Non-Gaussian State-Space Modeling of Nonstationary Time Series." *J. Am. Statist. Assoc.* **82**, 1032–1041.

Ott, G. 1967. "Compact Encoding of Stationary Markov Sources." *IEEE Trans. Inf. Theor.* **IT-13**, 82–86.

Riley, D. E., R. Reeves, S. M. Gartler. 1986. "Xrep, a Plasmid-Stimulating X Chromosomal Sequence Bearing Similarities to the BK Virus Replication Origin and Viral Enhancers." *Nucl. Acids Res.* **14**, 9407–9423.

Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen. 1982. "Nucleotide Sequence of Bacteriophage λ DNA." *J. Molec. Biol.* **162**, 729–773.

Schwarz, G. 1978. "Estimating the Dimension of a Model." *Ann. Statist.* **6**, 461–464.

Skalka, A., E. Burgi, A. D. Hershey. 1968. "Segmental Distribution of Nucleotides in the DNA of Bacteriophage Lambda." *J. Molec. Biol.* **34**, 1–16.

Smith, A. F. M. 1975. "A Baysean Approach to Inference About a Change Point in a Sequence of Random Variables." *Biometrika* **62**, 407–416.

Staden, R. 1984. "Graphic Methods to Determine the Function of Nucleic Acid Sequences." *Nucl. Acids Res.* **12**, 521–538.

Sueoka, N. 1959. "A Statistical Analysis of Deoxyribonucleic Acid Distribution in Density Gradient Centrifugation." *Proc. Natn. Acad. Sci. U.S.A.* **45**, 1480–1490.

Received for publication 16 June 1988