Computers Chem. Vol. 16, No. 2, pp. 107–115, 1992
Printed in Great Britain

0097-8485/92 $5.00 + 0.00
Pergamon Press Ltd

HIDDEN MARKOV CHAINS AND THE ANALYSIS OF GENOME STRUCTURE*

GARY A. CHURCHILL
Biometrics Unit, Cornell University, Ithaca, NY 14853, U.S.A.

(Received 4 November 1991; in revised form 17 January 1992)

Abstract—In this paper, statistical methods based on a hidden Markov chain model are used to study the structure of some small complete genomes and a human genome segment. A variety of discrete compositional domains are discovered and their correlations with genome function are explored.

1. INTRODUCTION

The total genomic DNA of plants and higher animals shows little variation in total G + C content, with most species falling in the range 40–44%. However, when the DNA is fragmented, a wide range of intragenomic heterogeneity in G + C content is observed. Bernardi *et al.* (1985) proposed the isochore model of DNA sequence organization to explain these observations. The DNA sequence is viewed as a mosaic of large homogeneous segments ($>10^6$ bases) which differ from one another in composition. The banding patterns seen in stained metaphase chromosomes provide a striking example of this higher level of organization in the genomic DNA of eukaryotes (Holmquist, 1989). By contrast, the total G + C content in the genomes of bacterial species varies over a wide range, 25–75%. Intragenomic variation in G + C content is relatively small but is still in excess of random expectations. These observations led Elton (1974) to suggest a similar segmented genome model for prokaryotes.

With the increasing availability of large DNA sequences, it has become possible to study compositional heterogeneity directly. Ikemura *et al.* (1990) studied variations in the G + C content of human DNA by arranging sequences from the database into large linkage groups. They found patterns of G + C composition that are consistent with the isochore hypothesis and identified a potential isochore boundary. Fickett *et al.* (1991) investigated the statistical properties of human and *E. coli* sequences available in the current database. They found that the variation in composition exceeds expectations under a homogeneous stochastic model of the DNA sequence. However, they find that the local base composition tends to persist over large regions with stronger persistence in human sequences than in *E. coli* sequences. Pevzner *et al.* (1989) introduced a measure of heterogeneity for small words in genetic texts, the irregularity coefficient. They used this method to study the "zonal structure" of viral and *E. coli* sequences. In Kozhukhin & Pevzner (1991), this study was extended to include all large (>25 kb) sequences in the current database. They conclude that the major contribution to genome inhomogeneity is due to the uneven distribution of SS and WW dinucleotides. This heterogeneity cannot be explained by the isochore model because it occurs locally. These studies suggest that compositional heterogeneity is a general phenomenon that is present in a variety of distinct forms.

The statistical methods used in this work are based on a stochastic model, the hidden Markov chain (HMC) model, in which the DNA is composed of homogeneous segments belonging to a small number of distinct compositional classes (Churchill, 1989). The probability of observing a particular base at a given site on the DNA molecule depends on the type of segment it is in. An underlying organization of the DNA is assumed in which the switching from one segment to the next follows an HMC. The states of the hidden process indicate the type of segment and are not directly observable. Parameters which characterize the different states (e.g. G + C content and switching rates) are estimated using likelihood methods and a recursive smoothing algorithm is then applied to reconstruct the segments. The smoothing algorithm does not utilize a sliding window, thus sharp transitions of state are possible and both small and large segments can be detected in a single pass. Results of the smoothing are displayed graphically to provide a global summary of the organization of sequence composition. Furthermore, the model makes specific predictions about properties of the DNA sequence which are related to its base composition, e.g. distribution of restriction enzyme sites, and provides a framework within which hypotheses about genome structure can be formulated and tested.

* The preliminary version of this work was presented during the *Open Problems in Computational Molecular Biology Workshop*, Telluride Summer Research Center, Telluride, CO, 2–8 June 1991.

We illustrate the use of HMC methods with four examples of their application to nucleotide sequences obtained from the GenBank database (release 60, Burks *et al.* 1990). In the first example, the mitochondrial genomes of animals are shown to have a consistent heterogeneous pattern in the distribution of purines and pyrimidines. The purine content in one strand of the DNA is shown to be correlated with the coding sense of the strand and the type of product, protein or rRNA. In the second example, the genome of bacteriophage lambda is shown to also have a strand asymmetry which is associated with coding function. However, a simple two-state model provides an inadequate description of the compositional variation and a more general state-space model is suggested. In the third example, the two major transcriptional domains of simian virus 40 are found to correspond to two genome segments of distinct dinucleotide composition. In the final example, the human α-globin region is analyzed to detect variation in the distribution of CpG dinucleotides. The CpG-rich islands associated with genes in this region are sharply defined, suggesting the HMC methods could be used to screen databases and raw sequence data to search for genes.

2. HMC MODELS FOR DNA SEQUENCE DATA

The model

We will consider a single strand of the DNA molecule in 5′ to 3′ orientation. The base at position t of the sequence will be denoted by y_t and the entire sequence up to t by $y^t = \{y_1, \ldots, y_t\}$, where y_t will take values in the set {C, T, A, G}. Binary representations of the DNA such as the purine–pyrimidine (AG–CT) or the strong–weak hydrogen bonding (GC–AT) sequences will also be considered. The state of the sequence at position t will be denoted by s_t and the entire sequence of states up to t by $s^t = \{s_1, \ldots, s_t\}$. For clarity, we will describe the case of a binary sequence with two distinct states where both y_t and s_t will take values in the set $\{0, 1\}$.

The probability distribution of y_t, conditional on the state s_t, will be binomial and can be written as

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

where

$$P_0 = \Pr(y_t = 1 | s_t = 0) \quad \text{and} \quad P_1 = \Pr(y_t = 1 | s_t = 1)$$

are assumed to be not equal. The segments of the sequence alternate between state 0 and state 1 according to a binary Markov process with transition probability matrix

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

The parameter λ defines the probability of switching from state 0 to state 1, i.e. $\lambda = \Pr(s_t = 1 | s_{t-1} = 0)$. The size of a state 0 segment will be geometrically distributed with mean $1/\lambda$. A similar interpretation applies to the parameter τ. Both λ and τ are thought of as being small ($< 10^{-3}$). The result is a sequence of hidden states which consists of long runs of all zeros or all ones. Segments of the sequence are identified with these runs.

In the general case, there are m observable outcomes (e.g. A, T, G, C). The probability of observing outcome i, given that the current state is k, is given by the parameter $P_{i,k} = \Pr(y_t = i | s_t = k)$. Successive outcomes may also be Markov dependent and the outcome transition probabilities are given by $P_{ij,k} = \Pr(y_{t+1} = j | y_t = i, s_t = k)$. The hidden process which defines the segments will switch according to a Markov chain on r states, with the $r \times r$ transition probability matrix denoted $\Lambda = [\lambda_{ij}]$. The number of free parameters required to specify a model with m outcomes and r states is given by $k = (m - 1)r + r(r - 1)$. If the outcomes are Markov dependent, $k = m(m - 1)r + r(r - 1)$. The one-state model ($r = 1$) is an important special case that includes the usual independent and Markov chain models for DNA sequences.

Parameter estimation

The model parameters are the outcome probabilities and state transition probabilities and will be denoted by θ. For example, in the two-state binary sequence model, $\theta = (P_0, P_1, \lambda, \tau)$. Typically, these values will not be known and must be estimated from the data. Churchill (1989) describes an Expectation-Maximization (EM) algorithm which computes a maximum likelihood estimate (MLE) of the model parameters, $\hat{\theta}$. This iterative optimization procedure operates by alternatively augmenting the observed data (E-step), and then maximizing the augmented data loglikelihood (M-step). In this implementation, the E-step constructs an estimate of the hidden states via the smoothing algorithm (next section) and the M-step is a trivial maximization.

Smoothing algorithm

The smoothing algorithm used to reconstruct the homogeneous segments of a sequence is described in detail by Churchill (1989). It requires that the model parameters be completely specified in advance. In the present work, all smoothed estimates are computed based on the MLE parameter values. The outcome of the procedure is, for each site in the sequence, the probability that the site belongs to state j conditional on the entire observed sequence. This quantity, denoted $\Pr(s_t = j | y^n)$, is called the *smoothed estimate* of s_t and can be plotted against the index t to provide a graphic summary of compositional structure. The conditional probabilities are computed by a recursive algorithm with two steps. The filtering step is a forward pass through the sequence with incorporates the information in all 5′ bases and the current base into the smoothed estimate. The smoothing step is a reverse pass through the sequence

which incorporates the information in 3′ bases into the smoothed estimate. This algorithm is related to the Kalman filter and is described in a general setting by Kitagawa (1987).

Model comparison

Given several candidate models, the model with the maximum value of the Bayesian information criterion (BIC),

$$\text{BIC} = l(\hat{\theta}) - \tfrac{1}{2}k \log n, \qquad (2)$$

is taken to be the best model. Here, $l(\hat{\theta})$ is the maximized loglikelihood, k is the number of free parameters in the model and n is the sequence length. Thus, the BIC value is a penalized form of the loglikelihood with a penalty that increases linearly with the number of parameters in the model. It is often more convenient to report changes in the value of BIC (ΔBIC) relative to a simple model. For example, the independent and equally-likely outcomes model has zero free parameters and the BIC is equal to the loglikelihood $l_0 = n \log(1/m)$. For large sequence lengths n, ΔBIC approximates the logarithm of the posterior odds ratio for the two models compared (Schwarz, 1978). The procedure of choosing the model with maximum BIC has been shown to be a consistent criterion for estimating the order of a Markov chain (Katz, 1981), a situation very similar to that considered here.

3. EXAMPLES OF GENOME ANALYSIS USING HMC MODELS

Animal mitochondrial DNAs

The mitochondrial genomes of animals are circular double-stranded DNA molecules ranging in size from 16 to 19 kb. Their functional organization is highly compact with most of the DNA involved in the coding of proteins or functional RNA molecules. Complete mitochondrial DNA (mtDNA) sequences have been determined for a variety of organisms including bovine (Anderson *et al.*, 1982), human (Anderson *et al.*, 1981), mouse (Bibb *et al.*, 1981), xenopus (Roe *et al.*, 1985) and drosophila (Clary & Wolstenholme, 1985). The mammalian and xenopus genomes share identical topographies with all coding regions in the same relative positions. The drosophila genome has a different gene order which can be related to the others by a small number of translocation and inversion events. All of the animal mtDNAs are AT-rich (bovine, 60.6%; human, 55.5%; mouse, 63.2%; and xenopus, 63.1%). The drosophila mtDNA has an exceptionally high A + T content of 78.6% and it has been suggested that selection has favored AT base-pairs at all locations where they are compatible with function (Clary & Wolstenholme, 1985).

Binary purine–pyrimidine sequences derived from the L-strands of each genome were used to compute model comparison statistics and MLEs (Table 1). In every case, the largest value of ΔBIC is found for the two-state model. Furthermore, the consistency of the estimated parameters across all five genomes suggests a common organization at the level of purine composition, despite the wide divergence in A + T content. For bovine mtDNA, the states are characterized by average purine contents of 45.3% (state 0) and 55.4% (state 1). The expected size of a state 0 region is $1/\lambda \approx 11$ kb and that of a state 1 region is $1/\tau \approx 2$ kb. Similar interpretations apply to the other genomes.

The smoothing algorithm was applied to identify the genome segments involved. The sequence profiles (Fig. 1) reveal large segments with fairly sharp transitions. The general pattern of mtDNA organization in these animals is a single purine-rich region that contains the sense-strand rRNA genes and a single pyrimidine-rich region that contains the sense-strands of most protein encoding genes. The organization of the drosophila sequence provides a remarkable example of the consistency of this pattern. The profile shows a greater number of segments but the functional correlations are identical. Using the major coding regions as landmarks, we see that the sense-strands in rRNA encoding genes are purine-rich while the sense-strands of protein encoding genes are pyrimidine-rich.

In order to explain the pyrimidine excess in the sense-strand of protein encoding genes, a statistical analysis of base composition by reading frame was carried out (results not shown). The most significant contributing factor was found to be the predominance of T (>40%) in the second coding position. This in turn is explained by the high proportion of

Table 1. mtDNA model selection and parameter estimation

Species	ΔBIC			$\hat{P}_0$	$\hat{P}_1$	$\hat{\lambda} \times 10^4$	$\hat{\tau} \times 10^4$
	One-state	Two-state	Three-state				
Bovine	27.01	47.61	24.32	0.453	0.554	0.93	5.04
Human	112.42	142.34	111.44	0.425	0.525	1.15	6.10
Mouse	26.90	49.31	15.38	0.454	0.542	0.90	4.34
Xenopus	36.92	57.77	25.61	0.446	0.527	3.11	9.59
Drosophila	0.27	30.23	7.77	0.452	0.544	3.72	5.95

The L-strand sequences of five animal mtDNAs were converted to binary indicators (0 = pyrimidine and 1 = purine). One-state, two-state and three-state HMC models with independent outcomes were fitted to these sequences by maximum likelihood. The model comparison statistics shown are changes in BIC relative to a baseline model with independent and equally-likely outcomes, $\text{BIC}_0 = -n \log 2$. Parameter estimates are shown for the two-state model.

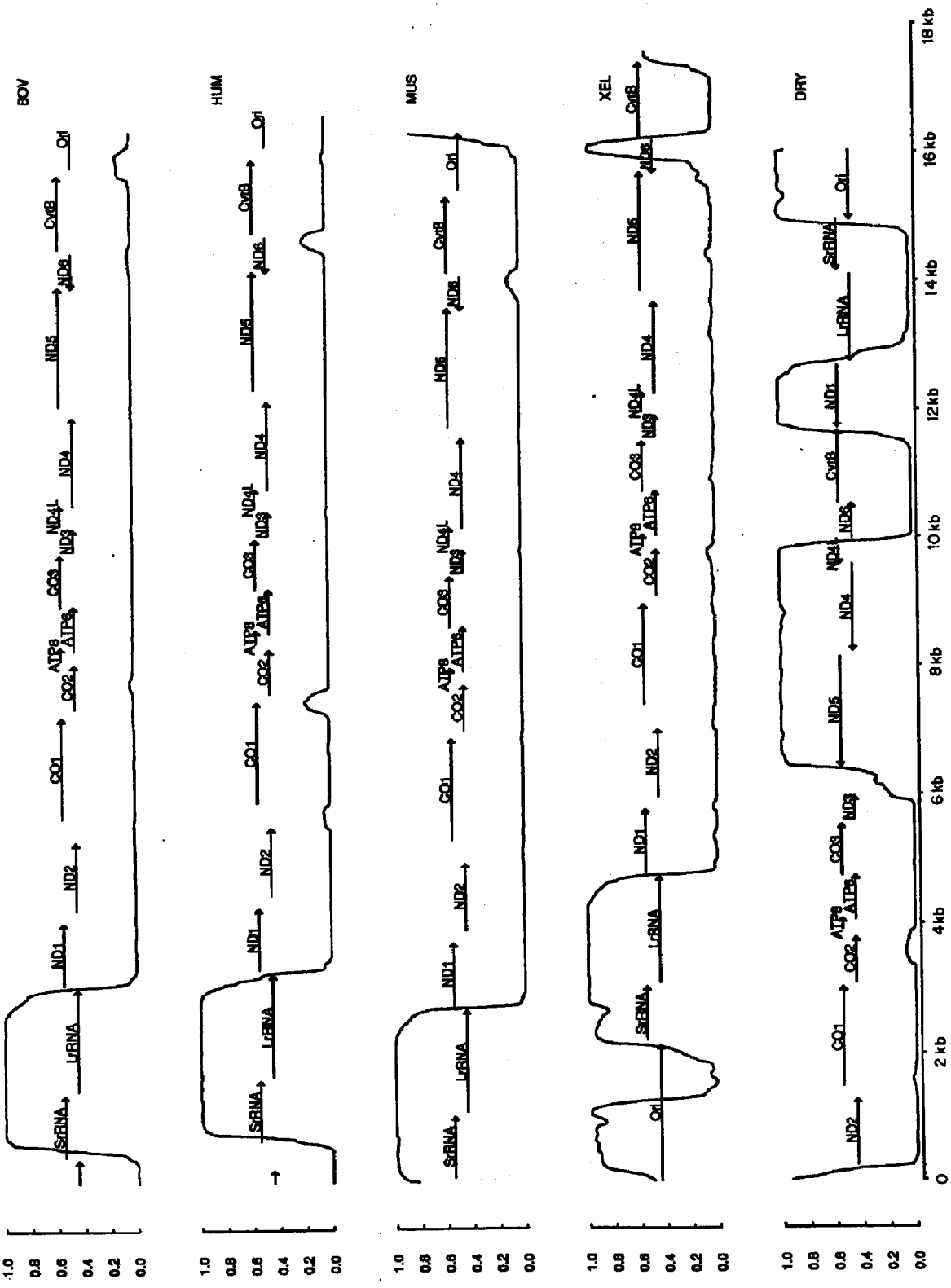

Fig. 1—*Legend opposite.*

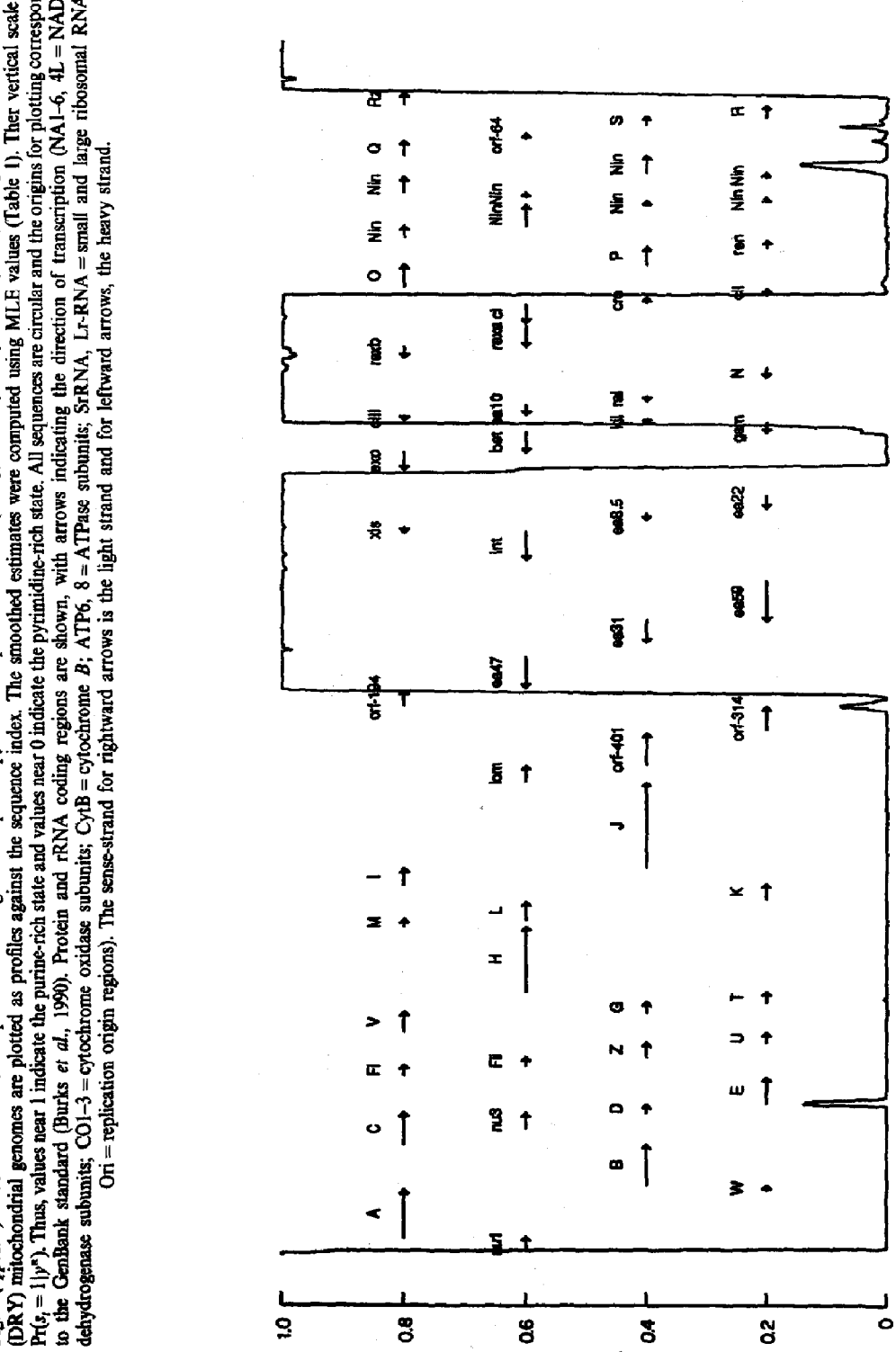

Fig. 1. *(opposite)* Smoothed estimates computed from the light strand purine–pyrimidine sequences of bovine (BOV), human (HUM), mouse (MUS), xenopus (XEL) and drosophila (DRY) mitochondrial genomes are plotted as profiles against the sequence index. The smoothed estimates were computed using MLE values (Table 1). Ther vertical scale is $\Pr(s_t = 1|y^n)$. Thus, values near 1 indicate the purine-rich state and values near 0 indicate the pyrimidine-rich state. All sequences are circular and the origins for plotting correspond to the GenBank standard (Burks *et al.*, 1990). Protein and rRNA coding regions are shown, with arrows indicating the direction of transcription (NA1–6, 4L = NADH dehydrogenase subunits; CO1–3 = cytochrome oxidase subunits; CytB = cytochrome *B*; ATP6, 8 = ATPase subunits; SrRNA, Lr-RNA = small and large ribosomal RNAs; Ori = replication origin regions). The sense-strand for rightward arrows is the light strand and for leftward arrows, the heavy strand.

Fig. 2. Smoothed estimates computed from the four-outcome sequence of bacteriophage lambda are plotted as a profile against the sequence index. A model with independent outcomes and two hidden states is assumed, with parameter values as in Table 2. The vertical scale is $\Pr(s_t = 1|y^n)$. Protein encoding regions are shown with arrows to indicate the direction of transcription. Identifiers correspond to GenBank release 60 (Burks *et al.*, 1990).

the hydophobic amino-acids leucine and isoleucine in mitochondrial encoded proteins. Little or no pyrimidine excess is seen in the third coding position. Further calculations confirmed that the amino-acid composition of mitochondrial proteins is sufficient to produce the observed strand asymmetry.

The mitochondrial DNA sequences were also studied using the strong–weak (S–W) alphabet. The BIC indicates S–W heterogeneity in the drosophila and xenopus genomes. The xenopus sequence profile (not shown) reveals an *AT*-rich segment in the replication origin region. The drosophila sequence profile (not shown) shows rapid and partial state changes that suggest the segmented genome model is inadequate to describe the heterogeneity of S–W bases in this genome.

Bacteriophage lambda

The complete genome of the bacteriophage lambda is a double-stranded circular DNA molecule of 48,502 base-pairs (Sanger *et al.*, 1982). The genome is compact with very little non-coding DNA and several overlapping reading frames. Markov chain analyses of the lambda sequence have suggested that there is a long-range dependence between neighboring bases (Tavare & Giddings, 1989; Churchill, 1988). The compositional heterogeneity of lambda DNA was observed by Skalka *et al.* (1968) and studied using statistical methods by Pevzner *et al.* (1989) and Churchill (1988, 1989). An HMC analysis is carried out here to identify regions which might correspond to distinct genome modules.

HMC models with independent and first-order dependent outcomes and up to four hidden states were fitted to the four-base sequence of bacteriophage lambda. According to the ΔBIC criterion (Table 2), the best model has three states and first-order dependent bases. However, examination of the smoothed profiles suggests that the discrete state model is not an adequate description of heterogeneity in lambda. Similar analyses were carried out for binary AT–GC and AG–TC sequences as well as three-outcome sequence with indicators for SS, WW and mixed dinucleotides. In every case, the left half of the lambda genome appears as a single homogeneous segment and the right half fails to show distinct segments. We conclude that the compositional variation in lambda does not fall into a small number of distinct states. Thus, it will be necessary to consider more general state-space models before an adequate characterization of this genome can be obtained.

Table 2. Model comparison for the viral sequences

			ΔBIC	
r	*o*	*k*	Lambda	SV40
1	0	3	30.4	78.1
2	0	8	516.8	107.3
3	0	15	569.3	105.8
4	0	24	582.8	80.6
1	1	12	462.1	296.8
2	1	26	934.2	303.4
3	1	42	944.8	266.2
4	1	60	920.5	207.0

Models with independent outcomes ($o = 0$) and first-order Markov dependent outcomes ($o = 1$) and up to $r = 4$ hidden states were fitted to the four-outcome sequences of bacteriophage lambda and SV40. The number of free parameters is shown under the column labeled k. Model comparison statistics were computed by maximum likelihood. The baseline model for relative changes in BIC is the independent and equally-likely model with $BIC_0 = -n \log 4$.

Table 3. Parameter estimates for the bacteriophage lambda two-state model

	Outcome probabilities			
	T	C	A	G
State 0	0.2078	0.2475	0.2464	0.2983
State 1	0.3235	0.2085	0.2697	0.1984

MLEs for the two-state independent outcomes model fit to the four-outcome sequence of bacteriophage lambda.

The two-state model fitted to the four-base sequence of lambda illustrates the pattern. Estimated state transition probabilities are $\lambda = 0.000115$ and $\tau = 0.000224$ and the estimated base compositions are summarized in Table 3. A two-state profile of the smoothed estimates (Fig. 2) shows a general correspondence between the segments identified by HMC analysis and the direction of transcription of the genes. The reversal around 32 kb reflects the poor fit of the two-state model in the right half of lambda.

Simian virus 40 (SV40)

The SV40 genome is a circular double-stranded DNA molecule of 5243 base-pairs (e.g. Reddy *et al.*, 1978). The expression of SV40 genes is temporally regulated by two major transcripts, early and late. Transcripts are generated in opposite orientations outward from the replication origin region. The early transcript produces the major and minor T-antigens and the late transcript produces other viral proteins.

As in the previous example, a number of HMC models were fitted to the SV40 four-outcome sequence. As indicated by the ΔBIC statistics (Table 2), the best model has two states and first-order dependent bases. Parameter estimates for this model are shown in Table 4. A plot of the smoothed estimates reveals two large segments which correspond to the early transcript (state 1) and to the replication origin and late transcript (state 0). These regions are not identified by any of the independent outcome models nor are they apparent from models fitted to binary AT–GC or AG–TC sequences. Thus, the states reflect regions of distinct dinucleotide composition. The CpG dinucleotide is rare in both states but especially in state 1 as is reflected in the outcome transition probabilities ($P_{CG,0} = 0.0435$,

Table 4. Parameter estimates for the SV40 two-state model

	State 0				State 1			
	T	C	A	G	T	C	A	G
T	0.3339	0.1346	0.1938	0.3377	0.3968	0.2385	0.2125	0.1522
C	0.3973	0.2152	0.3440	0.0435	0.3317	0.2339	0.4289	0.0055
A	0.1976	0.2018	0.3572	0.2433	0.2669	0.1815	0.3427	0.2089
G	0.2190	0.2276	0.2505	0.3030	0.3074	0.3162	0.1867	0.1897

MLEs for the two-state first-order model fit to the four-outcome sequence of SV40.

$P_{CG,1} = 0.0055$). Other notable differences between the two states are due the frequenceis of TG, GG and GA dinucleotides.

Human α-globin region

The dinucleotide CpG is rare in vertebrate DNA, occurring at only 10–20% of its expected frequency. The CpG dinucleotide is also the unique site for methylation of vertebrate DNA and, thus, is potentially important for the regulation of gene expression. Protein encoding regions are often found to be associated with 5′ and/or 3′ regions of elevated CpG content (Bird, 1986). These CpG-rich islands provide easily recognizable landmarks for locating genes in large DNA sequences. HMC models can be adapted to locate CpG-rich islands and any other types of regions characterized by an excess or deficit of specific nucleotide patterns. Thus, HMC methods could be used to screen databases or raw sequence data to locate regions of potential interest. This method of screeing would be especially robust to sequencing errors.

We have examined the distribution of CpG dinucleotides in a 12.5 kb region of the human α-globin cluster, which include the functional genes α1 and α2 and the pseudogene ψα1. The four-base DNA sequence was converted to a binary sequence with 1 indicating the C position of a CpG dinucleotide and 0s elsewhere. The binary sequence is modeled as a first-order Markov chain with two hidden states, "normal" and "CpG-rich". The CpG dinucleotide cannot self-overlap, thus it is not possible to observe two consecutive 1s in this sequence and the transition from outcome 1 to outcome 0 will occur with probability 1 in both states. The outcome transition probabilities from 0 to 1 are estimated from the data.

The ΔBIC statistics for models with two and three states relative to the one-state model are 180.77 and 157.12, respectively, clearly indicating that a two-state model is best. The rate of CpG dinucleotides is 0.0202 for state 0 (normal DNA) and 0.1166 for state 1 (CpG-rich DNA). The estimated state transition probabilities are $\lambda = 0.000245$ (switching into the CpG-rich state) and $\tau = 0.001171$ (switching out of the CpG-rich state). Thus, the pattern is long segments of low CpG content interrupted by short segments of relatively high CpG content.

The profile plot (Fig. 3) reveals three CpG-rich islands with fairly sharp boundaries. The first is located 2 kb upstream from the pseudogene. The other two begin in the 5′ regions of the functional genes and contain at least the first two exons. In order to contrast the HMC approach with a standard sliding-window analysis, the proportions of G + C and CpG over windows of size 256 base-pairs were plotted (Fig. 4). The window size was selected to give the best visual resolution of the CpG-rich islands.

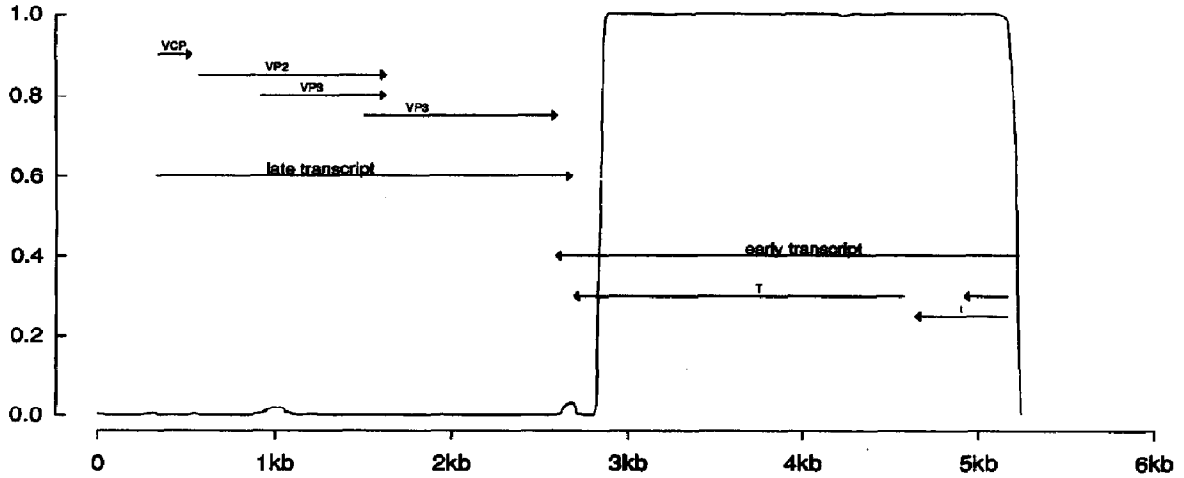

Fig. 3. Smoothed estimates computed from the four-outcome sequence of SV40 are plotted as a profile against the sequence index. A model with Markov-dependent outcomes and two hidden states is assumed with parameter values as in Table 3. The vertical scale is $\Pr(s_t = 1|y^n)$. The major transcripts and their protein products are shown with arrows to indicate the direction of transcription.

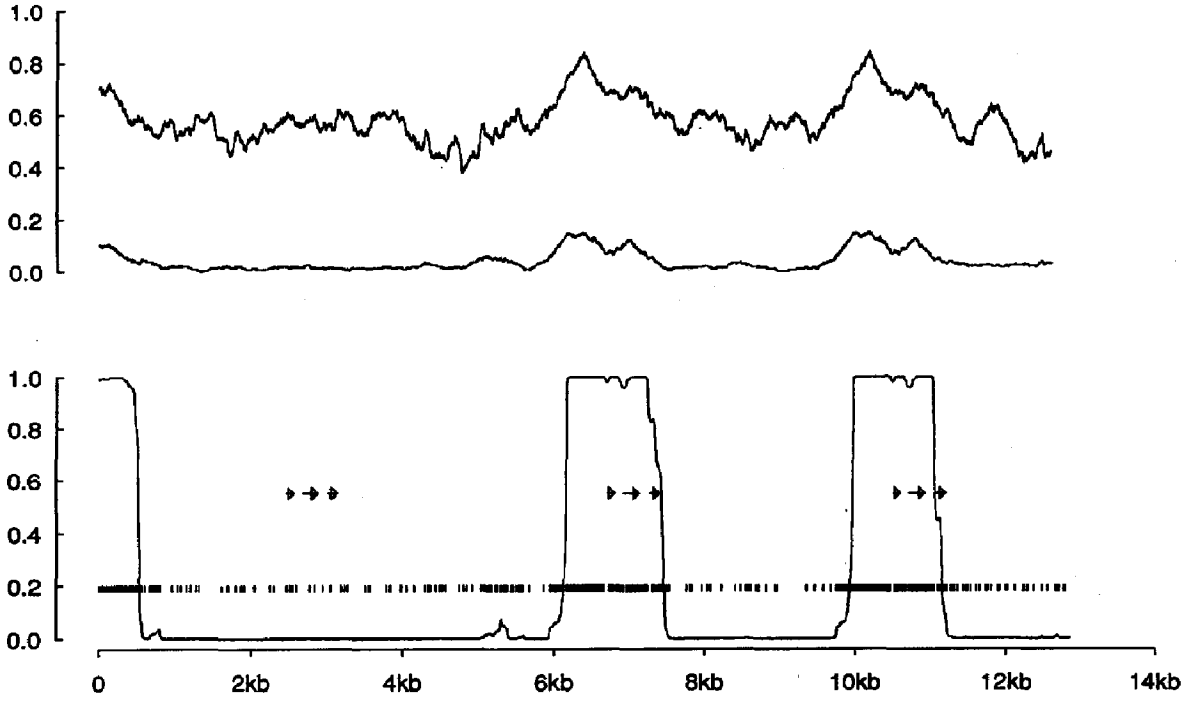

Fig. 4. The upper figure shows the proportion of G + C and the proportion of CpG in overlapping windows of 256 base-pairs plotted as a profile against the sequence index. The vertical scale is the proportion. The lower figure shows a profile of the smoothed estimates. The vertical scale is $\Pr(s_i = 1 | y^n)$. Thus, values near 1 indicate the CpG-rich state. Exons of the pseudogene and the active genes are shown with arrows to indicate the direction of transcription. The location of CpG dinucleotides are indicated by vertical hatch marks.

Smaller window sizes produced noisier plots and larger window sizes tended to wash-out the features. Although the sliding-window method is sufficient to detect the CpG-rich islands in this example, it has several limitations: different window sizes must be tested; the boundaries are not sharply defined; estimates of CpG content and feature size are not readily available; it is not possible to make formal comparisons among alternative models.

4. DISCUSSION

One of the fundamental questions of modern biology concerns the nature and extent of compositional variation and its relation to the organization of structure and function in genomic DNA. Since random mutation processes would tend to homogenize DNA, it is reasonable to suppose that some constraints are active in creating and maintaining compositional variation. Bernardi & Bernardi (1986) suggested that compositional constraints which affect both coding and non-coding sequences have resulted from selective pressures, perhaps at the level of chromosome structure, and that these constraints represent an important subset of the total constraints acting on the evolution of a genome. Analysis of heterogeneity could provide clues about the nature of compositional constraints and different levels of organization in large genomes. An alternative explanation for compositional heterogeneity is that the pattern of mutations varies across regions of the genome. Understanding the nature of this variation could provide significant insights into the process of point mutations and their role in genomic evolution.

Fickett *et al.* (1991) conclude from their studies that multidomain models are inadequate to describe the observed variation and suggest that models with continuous variation in local G + C content should be considered. Kozhukhin & Pevzner (1991) also raise doubts about the generality of the large homogeneity domains in DNA sequences. These doubts are confirmed by the bacteriophage lambda example above. However, the HMC model is readily generalized to include continuous variation. Some computational problems involved with fitting such models are currently being studied.

The idea that different functional domains of DNA can be distinguished by their statistical characteristics is not new (e.g. Smith *et al.*, 1983). However, the statistical approach to interpretation of DNA sequences has met with only limited success. One problem may be that statistical methods are not well suited for the detection of unique features that are often biologically important. In their proper domain, statistical methods can provide very power-

ful descriptive and inferential tools. When we begin to look at DNA on a global scale, we can expect to overlook some of the important details that are essential to DNA function. What we gain is a new perspective, a view of the higher levels of organization which exist in genomic DNA. Presently it appears that the genomic DNA of eukaryotes is hierachically organized at several different levels but our understanding of this organization is limited. The patterns of DNA sequence organization which will eventually be observed in the complete nuclear genomes of eukaryotes (and also in prokaryotic genomes) may be very different from the patterns found in the examples presented here, but the HMC methods will allow us to observe and analyze these patterns.

Program availability—The computer programs used to produce the examples presented in this paper are available from the author upon request.

Acknowledgments—The author is grateful to P. Pevzner, A. Konopka and E. Trifonov for helpful discussions. Partial support for this work was provided by USDA Hatch Program Grant No. NYC-151411.

REFERENCES

Anderson S., Bankier A. T., Barrell B. G., de Bruijn M. H. L., Coulson A. R., Drouin J., Eperon I. C., Nierlich D. P., Roe B. A., Sanger F., Schreier P. H., Smith A. J. H., Staden R. & Young I. G. (1981) *Nature* **290,** 457.

Anderson S., de Bruijn M. H. L., Coulson A. R., Eperon I. C., Sanger F. & Young I. G. (1982) *J. Mol. Evol.* **156,** 683.

Bernardi G. & Bernardi G. (1986) *J. Mol. Evol.* **24,** 1.

Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M. & Rodier F. (1985) *Science* **228,** 953.

Bibb M. J., Van Etten R. A., Wright C. T., Walberg M. W. & Clayton D. A. (1981) *Cell* **26,** 167.

Bird A. P. (1986) *Nature* **321,** 209.

Burks C. *et al.* (1990) *Methods Enzymol.* **183,** 1.

Churchill G. A. (1988) Ph.D. Thesis, Univ. of Washington, Seattle.

Churchill G. A. (1989) *Bull. Math. Biol.* **51,** 79.

Clary D. O. & Wolstenholme D. R. (1985) *J. Mol. Evol.* **22,** 252.

Elton R. A. (1974) *J. Theor. Biol.* **45,** 533.

Fickett J. W., Torney D. C. & Wolf D. R. (1991) Base compositional structure of genomes. Unpublished manuscript.

Holmquist G. P. (1989) *J. Mol. Evol.* **28,** 469.

Ikemura T., Wada K-N. & Aota S.-I. (1990) *Genomics* **8,** 207.

Katz R. W. (1981) *Technometrics* **23,** 243.

Kitagawa G. (1987) *J. Am. Stat. Assoc.* **82,** 1032.

Kozhukhin C. G. & Pevzner P. A. (1991) *CABIOS* **7,** 39.

Pevzner P. A., Bordovsky M. Y. & Mironov A. A. (1989) *J. Biomol. Struct. Dyn.* **6,** 1027.

Reddy V. B., Thimmappaya B., Dhar R., Subramanian K. N., Zain S., Pan J., Ghosh P. K., Celma M. L. & Weissman S. M. (1978) *Science* **200,** 494.

Roe B. A., Ma D. P., Wilson R. K. & Wong J. F. H. (1985) *J. Biol. Chem.* **260,** 9759.

Sanger F., Coulson A. R., Hong G. F., Hill D. F. & Peterson G. B. (1982) *Nucleic Acids Res.* **14,** 9407.

Schwarz G. (1978) *Ann. Stat.* **6,** 461.

Skalka A., Burgi E. Hershey A. D. (1968) *J. Mol. Biol.* **34,** 1.

Smith T. F., Waterman M. S. & Sadler J. R. (1983) *Nucleic Acids Res.* **11,** 2205.

Tavare S. & Giddings B. W. (1989) In *Mathematical Methods for DNA Sequences* (Edited by Waterman M. S.). CRC Press, Boca Raton, RL.