Statistical Science
1998, Vol. 13, No. 2, 142–162

Statistical Methods for DNA Sequence Segmentation

Jerome V. Braun and Hans-Georg Müller

Abstract. This article examines methods, issues and controversies that have arisen over the last decade in the effort to organize sequences of DNA base information into homogeneous segments. An array of different models and techniques have been considered and applied. We demonstrate that most approaches can be embedded into a suitable version of the multiple change-point problem, and we review the various methods in this light. We also propose and discuss a promising local segmentation method, namely, the application of split local polynomial fitting. The genome of bacteriophage λ serves as an example sequence throughout the paper.

Key words and phrases: Statistical genetics, change-point, hidden Markov chain, patchiness, quasideviance, split local polynomials, chromosome banding, bacteriophage λ.

1. INTRODUCTION

1.1 DNA Sequence Data

We consider methods of analyzing the deoxyribonucleic acid (DNA) sequences which are the basic information carriers of life. These sequences are long chainlike molecules composed of four nucleic acids, or bases. The bases are adenine (A), guanine (G), cytosine (C) and thymine (T); they are attached to a simple sugar–phosphate backbone, deoxyribose. A nucleotide is a base with its quota of backbone. DNA usually exists in a double strand with the two strands pairing by hydrogen bonding under the rule that A pairs with T and G pairs with C. The double strand is usually found twisted in the famous double helix structure. A single such molecule contains the instructions to generate a complete organism, such as a mouse or a man.

Triplets of bases code for amino acids, the building blocks of proteins (there are $4^3 = 64$ different such triplets) and are called *codons*. Three of the codons are "stop" codons which signal that the translation should stop. The remaining codons code for 20 amino acids. The central dogma of genetics is that DNA is transcribed into messenger ribonucleic acid (mRNA), which is translated into proteins, which form the building blocks of life.

Recent technological advances in the field of molecular genetics have generated a deluge of information—the sequences of entire genomes have been determined in many cases, and many more are expected (Pennini, 1997). Statistical analysis of DNA sequences is motivated by at least three areas of exploration, as Curnow and Kirkwood (1989) indicated:

1. Sequence data offer an extraordinarily fine view from which to extend the traditional methods of analysis of variation—for example, the analysis of variation between individuals can be brought to the level of even a single nucleotide difference.
2. Sequence data offer the opportunity to study the fine-tuning and organization of the genetic process—for example, the structure of genes may include elements such as protein binding sites or noncoding regions called introns; these may have characteristic physical and structural properties.
3. The comparison of sequences between species demands methods for determining similarities in evolution or function—for example, important sequences such as protein binding sites are conserved through evolution; their relationship can help describe the evolutionary relationships among widely different species.

Jerome V. Braun is staff member, Genentech, Inc., 1 DNA Way, South San Francisco, California 94080 (e-mail: braun@gene.com). Hans-Georg Müller is Professor, Department of Statistics, University of California, Davis, California 95616.

In earlier days, sequence information was obtained only for genes of known function. Nowadays large chunks of the genome are sequenced wholesale. The functions of many of these new sequences are unknown; most of the time scientists rely upon homology (similarity) with previously well studied small sequences. It is of interest to devise methods of describing and assessing sequences in ways which provide parsimonious, useful characterizations. One class of models for DNA sequences goes under the rubric of segmentation models. In such models, it is assumed that the sequence can be partitioned into a number of segments, where each segment has a certain degree of internal homogeneity.

1.2 Biological Basis for Segmentation Models

Early evidence of segmental genomic structure was provided by the phenomenon of chromosome banding. It was noticed early on that in the salivary glands of *Drosophila melanogaster* the chromosomes replicate many times and form what are called polytene chromosomes. Under the microscope one can see distinct banding patterns, which result from underlying physical or chemical structure. These bands are stable enough to make them useful for the identification of chromosomes and for genetic mapping.

By a variety of special staining techniques, similar banding patterns can be made to appear in the chromosomes of other organisms. Again under the microscope the chromosomes exhibit a pattern of light and dark transverse bands. For example, if the chromosomes are stained with Giemsa dye after protein denaturation, the so-called G-banding pattern appears. Bickmore and Sumner (1989) review the role of chromosome banding in elucidating the organization of the genome.

On the basis of a comparison of density gradient centrifuge data which showed the existence of differing DNA organization in both warm-blooded vertebrates and cold-blooded vertebrates, Bernardi et al. (1985) coined the term *isochore* to refer to large segments (greater than 300 kilobases (kb)) of DNA that belong to a "small number of classes characterized by different $[G + C]$ levels and by fairly homogeneous base compositions (at least in the 3 to 300 kb range)" and which may correspond to Giemsa- and reverse-banding patterns in mammalian chromosomes. See Ikemura, Wada and Aota (1990) for further discussion of isochores.

The neutral theory of evolution, which assumes that most mutations are neutral with respect to Darwinian selection, implies that most nucleotide changes occur by chance (Kimura, 1983). Compositional heterogeneity may give a toe-hold for natural selection to operate at the genome level. Holmquist (1989) postulates hierarchical selection on ever smaller functional units, noting that chromosome banding patterns appear to be evolutionarily stable; if there is a stable global structure which is maintained throughout evolution, this would indicate the existence of functional constraints on the ability of DNA sequences to mutate freely. Genomic structures which arise as instances of compositional constraints would not be subject to neutral mutations. Thus, understanding possible compositional constraints on sequence organization has implications for theories of evolution (Gillespie, 1991).

1.3 DNA Sequence Segmentation

We represent the observations along the sequence as $Y_1, \ldots, Y_n$, where Y takes on one of the values of the DNA alphabet (A, C, G or T). We further suppose that there are segments within which the observations follow the same or nearly the same distribution, and between which observations have different distributions. Interest may lie in describing the structure of the sequence, in detecting segments which are anomalous (in the sense that they are either mistakenly included in the sequence under consideration or perhaps derive from some other organizational scheme), or in comparing structures between sequences.

The nucleotides or the amino acids themselves could be further classified into various groups based on their physical and chemical properties. These groupings are termed alphabets; examples of some possible alphabets are listed in Tables 1 and 2. See Karlin, Ost and Blaisdell (1989) for a discussion of the use of these and other alphabets in the analysis of DNA sequences.

The DNA segmentation problem can be put into the framework of the multiple change-point problem for categorical data. In this approach, change-points correspond to the end points of the segments. The observations $Y_1, \ldots, Y_n$ are taken to be split into $R + 1$ contiguous segments by the values $0 = \tau_0 < \tau_1 < \cdots < \tau_R < \tau_{R+1} = 1$; the observations $Y_{[n\tau_{r-1}]+1}, \ldots, Y_{[n\tau_r]}$ are supposed to be identically distributed for each $\tau = 1, \ldots, R + 1$. The number

TABLE 1
Some statistical alphabets for DNA sequences

Pair	Alphabet
Purine versus pyrimidine	R (A or G); Y (C or T)
Heavy versus light	S (C or G); W (A or T)
Keto versus amino	K (T or G); M (A or C)

TABLE 2
Some statistical alphabets for amino acid sequences

Grouping	Alphabet
Chemical	Acidic (Asp, Glu); aliphatic (Ala, Gly, Ile, Leu, Val); amide (Asn, Glu); aromatic (Phe, Trp, Tyr); basic (Arg, His, Lys); hydroxyl (Ser, Thr); imino (Pro); sulfur (Cys, Met)
Charge	Acidic (Asp, Glu); basic (Arg, His, Lys); neutral (all others)
Hydrophobic	Hydrophobic (Ala, Ile, Leu, Met, Phe, Pro, Trp, Val); hydrophilic (Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Lys, Ser, Thr, Tyr)

R of change-points is usually unknown and needs to be determined.

The problem of statistically segmenting DNA sequence data has a history of about four decades. It has been known for at least that length of time that the sequence of bases does not follow a simple random assignment (Shapiro and Chargaff, 1960; Josse, Kaiser and Kornberg, 1961).

In Section 2, we give an overview of known segmentation methods for DNA sequence data. We also include a discussion of some recent controversies, summarized by the concepts of "long-range correlation" versus "patchiness." In Section 3, we discuss the embedding of the DNA segmentation problem in the multiple change-point framework. We suggest approaches to the DNA segmentation problem by considering multiple change-point methods, as well as introducing a new approach for local segmentation by split local polynomial fitting. In Section 4, we provide a brief discussion, including directions for further research.

The genome of the bacteriophage λ provides an example sequence which is used throughout the paper to give a feel for the behavior of the different models. We chose this genome because:

1. Its sequence was determined quite early (Sanger et al. 1982), and it has a long history of being used in demonstration of new numerical techniques.
2. It is a long enough sequence that statistical methods become believable; yet it is short enough so that graphs are not overwhelmingly compressed.
3. The sequence is a complete genome and is known biologically to have several large components.

Bacteriophage λ is a quite interesting organism in its own right. It is a virus which lives upon *Escherischia coli*, a common gut bacterium which is widely used in research and in the biotechnology industry. The bacteriophage binds to the cell and injects DNA into it, leaving behind an empty protein head. This DNA may either become integrated into the *E. coli* genome or drive construction of λ components and eventual cell lysis. Although the genome must be quite compact to fit into the protein head, not all of the genome is necessary for the last stage of construction. Thus, bacteriophage λ may be used to transfect *E. coli* cells with foreign DNA, by the device of replacing the unnecessary section with the genetic sequence of interest (of course, the section is quite necessary for the usual life cycle of the bacteriophage).

2. SEGMENTATION METHODS

2.1 Historical Attempts at Segmentation

Before the advent of the ability to sequence DNA accurately and in bulk, a variety of ingenious methods of determining sequence composition existed. These methods relied on indirect measurements using physical and chemical properties of the DNA as well as on direct measurements of the cell nucleus. An example is the method of density gradient centrifugation to determine molecular weight, density and heterogeneity of molecules of DNA, pioneered by Meselson, Stahl and Vinograd (1957). The use of density gradient centrifugation corresponds to the heavy–light alphabet in Table 1. In subsequent work, Skalka, Burgi and Hershey (1968) model the bacteriophage λ genome as a six-segment composition. They examine the density gradient graphs under the assumption that the graphs represent mixtures of homogeneous segments, and successively refine their analysis of the λ genome by selecting fragments from the visually determined modes of the mixture distribution. We may compare the results of their approach with the exact proportions of $G + C$ obtained from the sequence, as shown in Figure 1.

Other early approaches are discussed by Elton (1974), based on data from both density gradient centrifugation and melting curve techniques. In melting curve studies, the DNA is gradually denatured (the strands separated) by heating. The bonds of the G–C base-pairs are stronger, taking more heat to melt; thus the proportion of $G + C$ can be deduced. Elton investigates the variation of proportion G–C base-pairs with four statistical models: the simple random sequence model; a first-order Markov chain model; a "satellite DNA model"; and a segment model. Segment models appear to fit the data best in this study.

2.2 Maximum Likelihood Estimation of Segments

The multiple segmentation problem in a sequence of independent Bernoulli variates is considered by

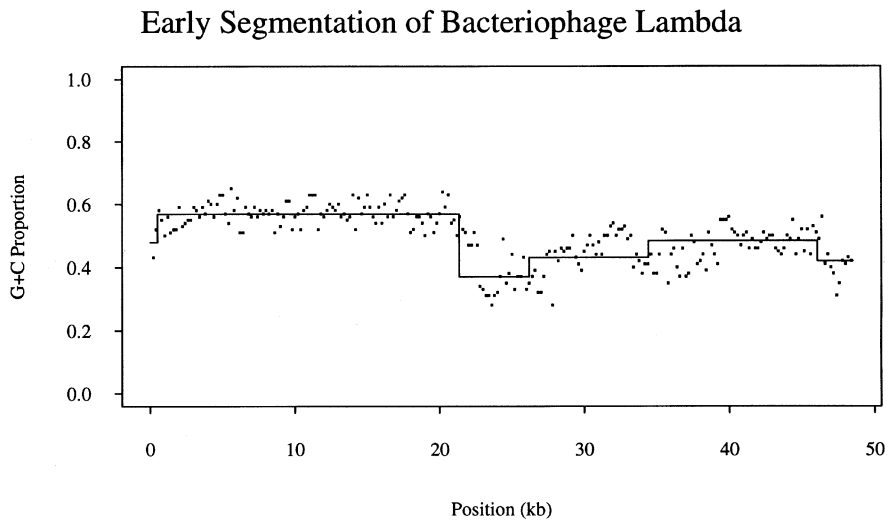

FIG. 1. *A comparison of the Skalka, Burgi and Hershey* (1968) *segmentation model for bacteriophage* λ: *fitted* $G + C$ *proportions* (*solid line*) *with the observed* $G + C$ *proportions in nonoverlapping* 200-*base windows* (*dots*). *The observed proportions are derived from the GenBank sequence LAMBDA.*

Fu and Curnow (1990). Their goal is to find statistical methods of determining secondary protein structure, but the approach applies to DNA sequence segmentation as well. They present an algorithm for computing the maximum likelihood estimate for the number of changed segments given a restriction on the minimum length of changed segments. We will outline their approach as applied to DNA sequence data.

In the most general form, denote the nucleotides as $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})$, $i = 1, \ldots, n$, where $Y_i = (1, 0, 0, 0)$ if the ith nucleotide is an A, and so forth. We model the distribution of the Y_i by a sequence of independent multinomial variables with probability vector $p = (p_1, p_2, p_3, p_4)$, so that

$$P[Y_i = y_i] = p_1^{y_{i1}} p_2^{y_{i2}} p_3^{y_{i3}} p_4^{y_{i4}}.$$

Further we allow p to take on only one of two values, so that one is considering two types of segments, "unchanged segments" ($p = \rho_0$) and "changed segments" ($p = \rho_1$).

For simplicity, define a new sequence c_i, $i = 1, \ldots, n$, where $c_i = 0$ for unchanged segments and $c_i = 1$ for changed segments. Let $P[Y_i = y_i|c_i]$ denote the probability mass function for Y_i given that the underlying sequence is unchanged ($c_i = 0$) or changed ($c_i = 1$). Allowing for a mixture of "changed" and "unchanged" segments, the likelihood function can be written as

$$L = \prod_{i=1}^{n} P[Y_i = y_i|c_i].$$

From this the likelihood ratio for the null hypothesis of no "changed segments" is

$$\ln LR = \ln \frac{\prod_{i=1}^{n} P[Y_i = y_i|c_i]}{\prod_{i=1}^{n} P[Y_i = y_i|c_i = 0]} = \sum_{k=1}^{l} \sum_{i=m_k}^{n_k} \ln\left(\frac{P[Y_i = y_i|c_i = 1]}{P[Y_i = y_i|c_i = 0]}\right),$$

where m_k, n_k, $k = 1, \ldots, l$, represent the beginning and ending of the kth changed segment out of a total of l changed segments. For an arbitrary segment $S = (Y_m, \ldots, Y_n)$ we define the function

$$f(S) = \sum_{i=m}^{n} \ln\left(\frac{P[Y_i = y_i|c_i = 1]}{P[Y_i = y_i|c_i = 0]}\right).$$

Then, representing the changed segments by S_k, $k = 1, \ldots, l$, the log-likelihood can be written as

$$\ln LR = \sum_{k=1}^{l} f(S_k).$$

This form shows that the log-likelihood ratio takes a simple form depending only upon the changed segments; the function f is a simple sum of the scores for each individual observation on the changed segments.

To compute the maximum likelihood solution, we might consider breaking up the sequence into all allowed configurations to find the segmentation which maximizes the log-likelihood ratio. Of course, this brute-force approach requires huge amounts of com-

puting and as a consequence will be impractical for sequences which are long or which contain many changed segments. Also, the log-likelihood ratio will be maximized for that segmentation which calls all runs of identical observations changed segments—here we see the reason for restricting the minimum length of the changed segments.

The dynamic programming algorithm of Bement and Waterman (1977) computes the global maximum with much less effort in the situation when the segment size is unrestricted (see Auger and Lawrence, 1989, for a use of the algorithm in a least-squares framework). Fu and Curnow devised a different algorithm for finding the global maximum which respects the minimum size of the changed segments. This algorithm can be described as follows. The determination of the l best segments (in terms of maximizing the log-likelihood) is performed sequentially. The best segment is found, then the best two segments are found and so on until the best l segments are found. The method of computing the best l segments from the best $l-1$ segments is simply described:

STEP 1. Find the best segment which does not overlap any two of the best $l-1$ segments.

STEP 2. Find the best splitting and expansion for each of the best $l-1$ segments.

STEP 3. Choose from Steps 1 and 2 the segment that provides the greater increase in the log-likelihood.

Given the restriction on the minimum size of the changed segments, the log-likelihood will eventually decrease with additional segments, or it will be impossible to add more segments. Of course, until that point, the log-likelihood must be increasing in l. The maximum likelihood estimate of the number of changed segments is the value of l which maximizes the log-likelihood. For a given value of l, the computational effort is at most $O(n^2)$.

Figure 2 shows the result of applying the original Fu–Curnow algorithm to bacteriophage λ. The bases were scored using the heavy–light alphabet, with $A, T = 1$ and $C, G = 0$, and the Bernoulli case was assumed. The probability vectors ρ_0 and ρ_1 were obtained by using the hidden Markov chain analysis in the next section and were taken as $\rho_0 = (0.5, 0.5)$ and $\rho_1 = (0.52, 0.48)$, with a slight abuse of notation. The likelihood was maximized when assuming four changed segments. Figure 3 shows the result of an alternative application of the Fu–Curnow algorithm to the same data. In this case, the nucleotides are assumed to follow independent multinomial distributions on each segment; initial probability vectors were obtained as before and were $\rho_0 = (0.21, 0.25, 0.25, 0.29)$ and $\rho_1 = (0.32, 0.21, 0.27, 0.20)$. The likelihood was maximized when assuming six changed segments.

The classical inferential theory for this estimation scheme is not well developed. Fu and Curnow use simulation to tabulate critical values for the number of changed segments. Key to the implementation of this approach is the fact that the probability vectors are assumed known. In practice, of course, we do not

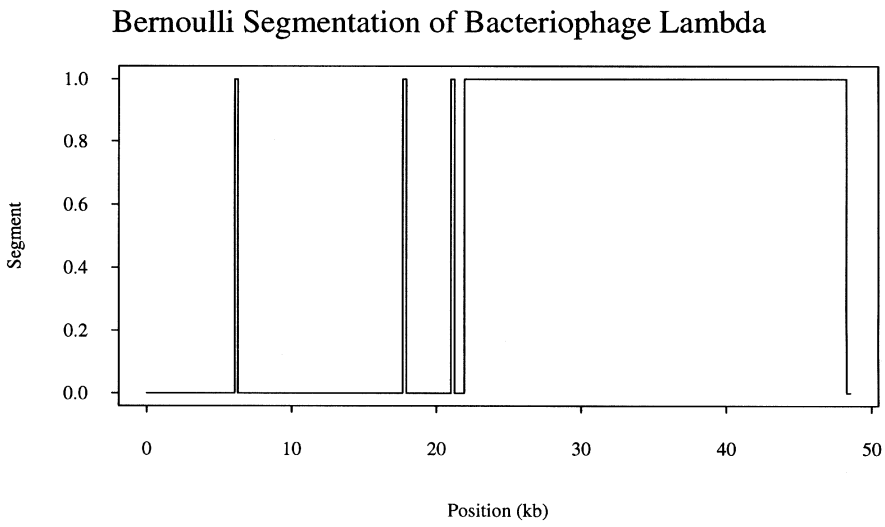

FIG. 2. *Results for the maximum likelihood method for finding changed segments using the Fu–Curnow algorithm for a binomial model. The probabilities represent independent nucleotide occurrences and were chosen as* $\rho_0 = (0.5, 0.5)$ *and* $\rho_1 = (0.52, 0.48)$ *by a hidden Markov chain analysis. A minimum segment size of* $\sqrt{n} = 220$ *was imposed.*

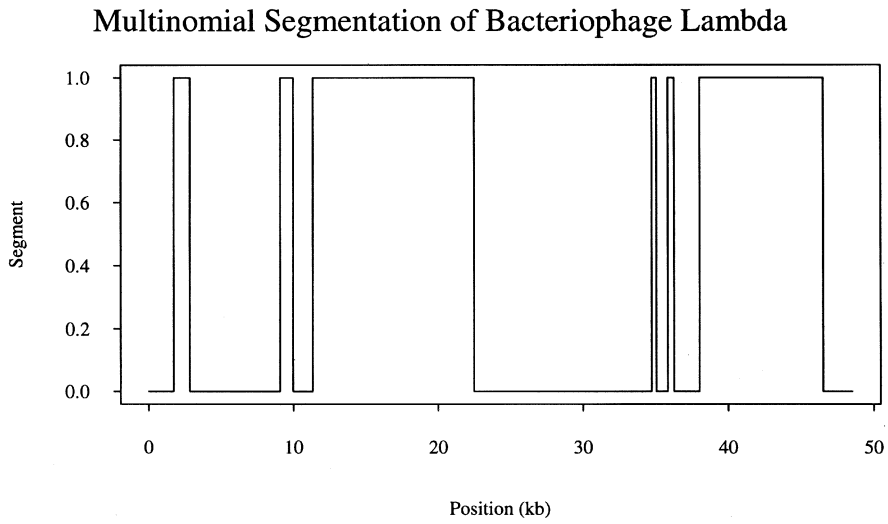

FIG. 3. *Results for the Fu and Curnow algorithm for a multinomial model, with* $\rho_0 = (0.21, 0.25, 0.25, 0.29)$ *and* $\rho_1 = (0.32, 0.21, 0.27, 0.2)$.

know these vectors; the scheme might be extended to some sort of joint estimation approach, where the unknown parameters are estimated iteratively, but this has not been studied. Some approximate approaches may be found in Wallenstein, Naus and Glaz (1994). Also see Auger and Lawrence (1989) for a discussion relating this problem to the segmented regression and multiple subsets problems, as well as some pragmatic comments about ad hoc use of the F-statistic.

2.3 Hidden Markov Chain Model

Churchill (1989, 1992) proposed a hidden Markov chain model to model segmentation of DNA sequences and to try to predict the locations of possible segments in mitochondrial and phage genomes. The hidden Markov model assumes that the different segments can be classified into a finite set of states, for example, CpG-rich or CpG-poor. In each state, the nucleotide data is assumed to follow a probability distribution, for example, a zero-order Markov chain. The states are assumed to switch from one to the other at random with low probability—since the states are unobserved and random in occurrence they form a hidden Markov chain. Note that under this model the lengths of segments follow geometric distributions with parameters given by the transition probabilities of the unobserved chain. A good introduction to the hidden Markov chain model is found in Rabiner (1989).

Suppose that there is a finite number r of states; these states may be described by r-vectors with $S = (1, 0, \ldots, 0)$ representing state 1 and so forth. Assume that the states underlying the observations, denoted by S_i, $i = 0, \ldots, n$, follow a Markov process with transition matrix $\Lambda = (\lambda_{jk})$. For example, with two states we may represent the Fu–Curnow situation of unchanged and changed segments.

The system equations for the hidden chain are

$$P[S_i = s_i | S_{i-1} = s_{i-1}] = \prod_{j=1}^{r} \prod_{k=1}^{r} \lambda_{jk}^{s_{i,j} s_{i-1,k}}.$$

Now assume that the observations $Y_i = (Y_{i,1}, \ldots, Y_{i,4})$ follow a multinomial distribution which depends on the state, say,

$$P[Y_i = y_i | S_i = s] = \prod_{j=1}^{4} p_{s,j}^{y_{i,j}},$$

where with another slight abuse of notation $(p_{s,1}, \ldots, p_{s,4})$ is the multinomial parameter associated with state $S_i = s$. From this we obtain the system equations for the observations

$$P[Y_i = y_i | Y_{i-1} = y_{i-1},\ S_i = s] = \prod_{j=1}^{4} \prod_{k=1}^{4} p_{s,k}^{y_{i-1,j} y_{i,k}}.$$

The smoothing equations

$$P[S_i = s | Y_1, \ldots, Y_n]$$

can then be derived and plotted to indicate homogeneous regions in the sequence.

As Churchill points out, the unknown distribution of the states and the distributions on the states can be estimated from the data using the EM algorithm, and the Bayesian information criterion can be used to determine the number of states necessary. Parameter values can be estimated and displayed graphically, without the need for window

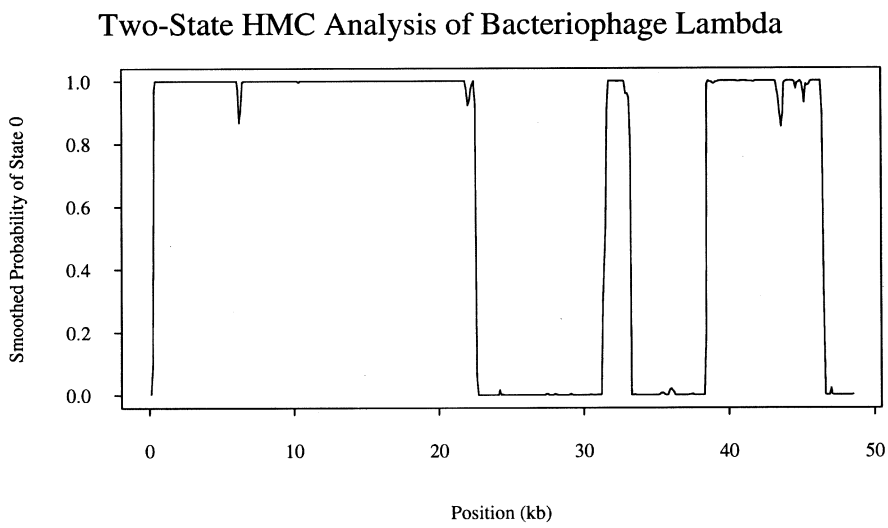

FIG. 4. *Two-state hidden Markov chain analysis for the bacteriophage λ sequence.*

size determination. This makes implementation very flexible and highly automatic. However, this method still forms a kind of smoothing, where the weightings of the data are not explicit and therefore not apparent to the user. An example of a two-state hidden Markov chain analysis of the bacteriophage λ sequence assuming independent multinomials on each segment is shown in Figure 4.

The use of hidden Markov chain models has expanded to the areas of chromosome hybridization data (Dupuis, 1994), of multiple DNA sequence alignment, where the models can be used to detect subtle sequence signals (Lawrence et al., 1993; Liu, Neuwald and Lawrence, 1995; Neuwald, Liu and Lawrence, 1995; Liu and Lawrence, 1996), and of protein modeling (Krogh et al., 1994). In the multiple alignment problem interest centers on determining whether small segments of DNA within several sequences are sufficiently similar to warrant declaring a match. For practical use the computational aspects are particularly important, and here the work of Krogh et al. in describing applications is quite useful.

2.4 Bayesian Approach

A theoretical advantage of the Bayesian approach to this estimation problem is that one can take advantage of the structure of the problem to marginalize over unknown parameters to obtain the global optimum. There are many early references for Bayesian methods for single change-points; examples are Smith (1975) and Raftery and Akman (1986). Methods for multiple change-points may be more immediately adaptable to sequence data. There are several approaches which are flexible enough. Product partition models (Hartigan, 1990; Barry and Hartigan, 1992) provide a method which is analytically tractable. Carlin, Gelfand and Smith (1992) explore hierarchical analysis of change-point problems; as an example they consider observations from a Markov chain with switching transition matrices. Stephens (1994) provides theory for binomial and regression models. Liu and Lawrence (1996) present a unified approach which directly includes segments, allows for incorporation of multiple sequences and addresses computation issues via the Gibbs sampler.

To be more precise, let $S = (S_1, \ldots, S_K)$ be the collection of sequences. Suppose that the number of segments in each sequence is Q, and Θ is the parameter vector for the sequence model. Let v_{kq} indicate the position of the last observation in the qth segment of the kth sequence, and let $V_k = (v_{k1}, \ldots, v_{kQ})$, where $v_{kQ} = n_k$ is the length of the kth sequence.

Now let $M_k = (m_{k1}, \ldots, m_{kQ})$, where m belongs to $\{1, \ldots, L\}$, indicating different types of segments. Let $S_k(i, j)$ denote the ith through the jth observations from the kth sequence. Under the assumption that observations within a segment are independent from observations in other segments (the Markov assumption), given the partition, the probability of observing that $S_k(1, j)$ has $q+1$ partitions with the last segment of type m is

$$\begin{aligned} &P\big(S_k(1, j)|\Theta,\ v_{k,q+1} = j,\ m_{k,q+1} = m\big) \\ &\quad = \sum_{v=1}^{j-1} \sum_{l=1}^{L} P\big(S_k(1, v)|\Theta,\ v_{kq} = v,\ m_{kq} = l\big) \\ &\qquad \cdot P\big(S_k v + 1,\ j|\Theta, m_{k,q+1} = m\big) \\ &\qquad \cdot P\big(m_{k,q+1} = m|m_{kq} = l\big), \end{aligned}$$

where $P(m_{k,q+1} = m | m_{kq} = l)$ is the probability of observing the segment from i to j of model type m.

As Liu and Lawrence point out, to implement Gibbs sampling would require sampling from $P(\Theta|M, V, R)$, where M and V represent all sequence segment models and partitions, respectively. If the global parameter Θ can be suitably decomposed, or approximately so, they show that the collapsing theorem of Liu (1994) can be applied. This allows integrating the parameter Θ out of the problem. With this simplification, Liu and Lawrence provide an extension of the dynamic programming approach of Auger and Lawrence which results in sampling from the joint distribution $P(V_k, M_k|R, V_{-k}, M_{-k})$ in reasonable time. Liu and Lawrence provide suggestions on model selection and also explore frequentist properties of the model; for further details on Bayesian model determination in similar situations see Green (1995).

2.5 Criticism and Controversy Related to Segmentation Models

On the biological side, while the emotional appeal of segmentation models is high, especially when considering such phenomena as chromosome banding, there is by no means universal agreement that genomes actually have segmented structure. Schweizer and Loidl (1987) provide an example of work with a complex computer model of the chemistry of DNA for C-banding patterns. (C-bands are produced by chemical treatment of the chromosome; the technique is especially useful to highlight the centromere and polymorphic bands.) They show that their mechanistic model provides a nonrandom distribution of C-bands without the need for an overall constraining optimality brought about by some sort of chromosome field, and so could theoretically explain the phenomenon of C-banding.

On the statistical side, one might wonder whether the effects of ignoring the known dependencies within DNA sequence data in favor of independence models is serious. It has been noted that independence models have good explanatory power in the protein sequence setting. For example, Krogh et al. found that the hidden Markov chain model agreed well with methods using three-dimensional structural information. Liu, Neuwald and Lawrence's (1995) work with the Gibbs sampler and multiple sequence alignment shows surprisingly good results even under the independence assumption.

A model which directly incorporates dependencies is the seventh-order Markov chain model considered by Scherer, McPeek and Speed (1994). This dependency model comes at the cost of requiring a huge number of parameters: since the model must estimate the proportions of octamers in use, there are 4^8 (= 65,536) parameters to estimate. The megabases of sequence information and the computing power available today make such an approach possible.

Ideally, though, one would like to incorporate information about the chemical and physical properties of the sequences. Amfoh, Shaw and Bonney (1994) show that models which include covariates and a dependency structure within a logistic regression model can adequately model mitochondrial DNA data. They used information about the structure and mutability of amino acids. This was not a segment model, but such models could be easily constructed within the same framework.

Other statistical models which are also not segment models *per se* but which may predict the occurrence of homogeneous stretches have been proposed. Among them are the walking Markov model and the long-range correlation model; both are shrouded in controversy.

2.6 Walking Markov Model

Fickett, Torney and Wolf (1992) examine the base composition of human and *E. coli* genomes. They analyze the phenomenon of strand symmetry (same number of occurrences of each base on each strand). They note the poor fit of homogeneous models, in particular the fit of Markov models using the Bayesian information criterion (BIC), and they identify a "large variance problem," that is, that there is less local homogeneity than necessary for most segment models in the literature which they studied. They are thus led to propose a new model: the walking Markov (WM) model.

The walking Markov model is a continuously varying stochastic process. The model is described as follows. A reflecting random walk on the interval $[1/3, 2/3]$ is denoted by W_i, $i = 1, \ldots, n$; in practice, we allow the W_i to advance in either direction by a small fixed distance. The W_i are assumed to control the distributions of the Y_i, $i = 1, \ldots, n$, by indexing first- or second-order Markov chains; that is, the observation Y_i is assumed to come from a Markov chain with transition matrix M_{W_i}, depending on one or two previous observations. The model parameters are estimated as follows. First, one tabulates quantities w_j, $j = 1, \ldots, J$, where J is the number of different observed $A + T$ fractions seen when running a 1,000-base window through the sequence of interest. The windows are separated by some fixed number of bases to reduce overlap.

For each observed $A + T$ fraction w_j, the windows which exhibit that fraction are collected. The BIC is used to select a first- or second-order Markov

chain M_w to model the data within the windows corresponding to w_j. (For example, if seven of the windows showed an $A + T$ fraction of 1/2, those windows provide a total of 7,000 observations for estimating the transition probabilities of the bases $M_{1/2}$.) Now to start the model, a w_0 is chosen in the range $[1/3, 2/3]$ under a uniform distribution; for each $i = 1, \ldots, n$, we perform a reflecting symmetric random walk in $[1/3, 2/3]$ with a small step size. The ith observation is chosen at random following the associated M_w. Fickett, Torney and Wolf claim the variance thus obtained is appropriate and the correlation structure is accurate. They also note that more complex models may be required: perhaps w should shift only every 100 bases or so; perhaps M_w should be a more complex Markov chain; perhaps the w should favor certain states. This approach embodies the philosophy that local gradual change is more realistic than infrequent large change.

This model is of course a version of the hidden Markov chain model. It has a large number of discrete underlying states, with an obligate jump to another state at each observation; in general the observations within each state follow a second-order Markov chain. Although the observation distributions are specified, the overall number of parameters seems too high to estimate any of them precisely. In Section 3.4 we discuss an alternative method of modeling a gradual change in composition, which at the same time allows for distinct jumps.

2.7 Long-Range Correlation Model

There has been a running debate over the last few years about the presence of long-range correlation in DNA sequences. In an effort to understand the correlation structure of DNA sequences, Peng et al. (1992) studied the following model. They used the purine–pyrimidine alphabet to score a number of DNA sequences. Specifically, for a given DNA sequence, use a score function $g(\cdot)$ to obtain the scores,

$$g(Y_i) = \begin{cases} +1, & \text{if the } i\text{th position is an } A \text{ or } G, \\ -1, & \text{if the } i\text{th position is a } C \text{ or } T. \end{cases}$$

Then the cumulative sum

$$S_k = \sum_{i=1}^{k} g(Y_i), \quad k = 1, \ldots, n,$$

defines what is called the *DNA walk*. An example of a similar walk is shown in Figure 5, using the heavy–light alphabet (C or G versus A or T) instead of the purine–pyrimidine alphabet. These graphs appear to display a certain amount of self-similarity at different scales, leading Peng et al. to believe that there might be some fractal properties encoded in the sequences.

We can define a scan statistic

$$S_k(L) = \sum_{i=k}^{k+L-1} g(Y_i), \quad k = 1, \ldots, n - L + 1,$$

where $1 \le L \le n$ is a given window size. Thus, for a given window size L, we have a finite population $\{S_k(L)\}_{k=1}^{n-L+1}$ of $n-L+1$ partial walks. The statistic of interest is the variance of this finite population,

$$F^2(L) = \frac{1}{n - L + 1} \sum_{j=1}^{n-L+1} \big(S_j(L) - E[S_j(L)]\big)^2.$$

The quantity $F(L)$ is called the fluctuation.

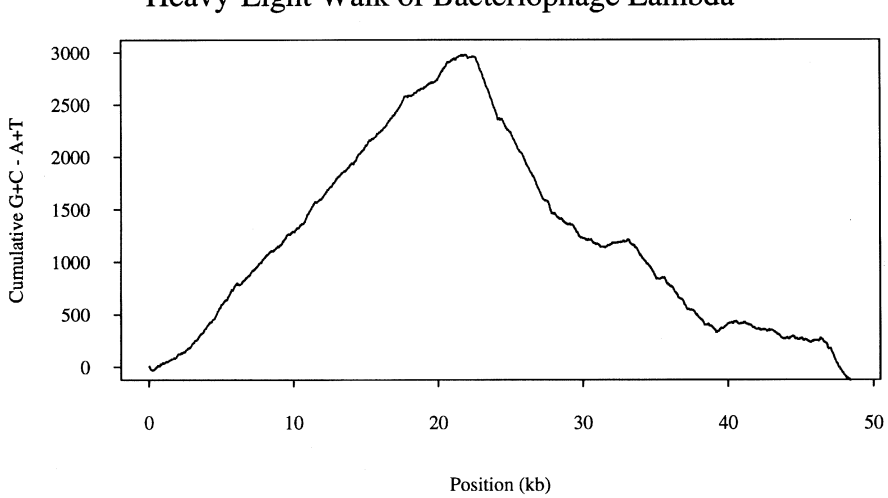

FIG. 5. *The DNA random walk for bacteriophage λ, when selecting the heavy–light alphabet,* $(G + C)$–$(A + T)$.

Under the condition of stationarity of the sequence, we have

$$E[F^2(L)] = LR(0) + 2\sum_{j=1}^{L-1}(L-j)R(j),$$

where $R(j) = \mathrm{cov}[Y_i, Y_{i+j}]$ is the autocovariance function of the sequence. If there is no covariance between the Y_i, then, for $j > 0$, $R(j) = 0$ and

$$E[F^2(L)] = LR(0),$$

so that we expect the ln-fluctuation is linear in $\ln L$ with slope 0.5. (The same behavior holds for the case of covariance up to a specified distance m_0.) If the covariance is decaying geometrically, so that $R(j) \sim \alpha^j$, where $0 < \alpha < 1$, we have what is known as a short-memory process, and

$$\begin{aligned} E[F^2(L)] &= LR(0) + 2\sum_{j=1}^{L-1}(L-j)R(j) \\ &= c_0 L + \sum_{j=1}^{L-1} L c_j \alpha^j \sim L. \end{aligned}$$

Therefore, we again expect the ln-fluctuation to be linear in $\ln L$ with slope 0.5. If the covariance is decaying in an inverse polynomial fashion, so that $R(j) \sim j^{2d-1}$, where $d \in [0, 1/2]$, then we have what is known as a long-memory process (Brockwell and Davis, 1991). In this case we expect the ln-fluctuation to be linear in $\ln L$ with slope $d + 1/2$ greater than 0.50 if $d \neq 0$.

Peng et al. use a minimum–maximum partitioning of the sequence data to remove strand bias. Comparing a sample of intron-containing sequences having average slope of 0.61 with a sample of intronless sequences having average slope 0.50 gave a statistically significant difference. Their conclusion was that there seems to be evidence for long-range correlation in intron-rich sequences, while intronless sequences do not show long-range correlation. These findings were greeted with skepticism by some and admiration by others (Maddox, 1992). Prabhu and Claverle (1992) reanalyze the data, as well as additional sequences, and find no particular association between introns and slopes on fluctuation plots. They conclude that the min–max partitioning scheme introduces biases into the estimation of the fluctuation.

2.8 Long-Range Correlation versus Patchiness

Nee (1992) noted that genomes have known structural patterns and that these patterns may have been responsible for the observation of Peng et al. (1992). Karlin and Brendel (1993) explored the implications of a segmented structure or "patchiness" for the fluctuation statistic. They consider the case where each segment is independently distributed, for simplicity, and the distribution differs between segments. Then the classical partitioning of variance gives

$$\begin{aligned} E[F^2(L)] &\doteq \mathrm{var}[S_k(L)] \\ &= E\big[\mathrm{var}[S_k(L)|\text{Segment}]\big] \\ &\quad + \mathrm{var}\big[E[S_k(L)|\text{Segment}]\big] \\ &= L(2p_1q_1 + 2p_2q_2) + L^2(p_1 - p_2)^2 \\ &= c_1 L + c_2 L^2. \end{aligned}$$

With this in mind, we expect $\ln F(L)$ to be curvilinear in $\ln L$. This contrasts with the linear forms which were proposed by Peng et al.

Karlin and Brendel note that if there are only two patch types, and the probability of falling into either patch is 1/2, then each Y_i can be thought of as a transformed Bernoulli random variable. That is, $Y_i = 2X - 1$, where X is a Bernoulli random variable. Then in patch j, $j = 1, 2$, we have $E[Y_i|\text{Segment } j] = p_j - q_j$ and $\mathrm{var}[Y_i|\text{Segment } j] = 4p_jq_j$, so that

$$\begin{aligned} E[F^2(L)] &= E\big[L\,\mathrm{var}[Y_i|\text{Segment}]\big] \\ &\quad + \mathrm{var}\big[LE[Y_i|\text{Segment}]\big] \\ &\doteq L(2p_1q_1 + 2p_2q_2) + L^2(p_1 - p_2)^2 \\ &= c_1 L + c_2 L^2. \end{aligned}$$

Only in the case when $p_1 = p_2$ will the L^2-term vanish.

Upon reexamination, the fluctuation plots for the sequences studied by Peng et al. were found to be nonlinear. They could be modeled pretty well by a straight line up to a window width of about $L = 50$, but the fit breaks down after that. (It is interesting to speculate whether or not the curvature could tell us something about the composition of the patches.)

The work of Peng et al. is extended by Voss (1992) by using a mapping which does not introduce correlations. Voss seems to find evidence for long-range correlation ($1/f$-noise), from which he infers a fractal scaling present in the genome. Buldyrev et al. (1993) dispute this finding, by showing that these observations could also have arisen from a segmented organization. Voss (1993) defends his previous work. He points out that the long-range correlation would be expected to create segments of strand bias, although he does not elaborate on this idea.

Our conclusion is that, since DNA sequences are known to show structural differences, the results of Karlin and Brendel's analysis suggest there is

no reason to reject the simpler model of mosaic structure and patchiness in favor of the long-range correlation model, which is very hard to interpret biologically. We also note the questions about the usefulness of the power-law model which are put forward by Avnir, Biham, Lidar and Malcai (1998).

3. CHANGE-POINT METHODS FOR SEGMENTATION

In the following, we discuss multiple and single change-point methods which have potential for the segmentation of DNA sequences. These general schemes could be applied to the particular case of DNA sequences by postulating binomial or multinomial models for the observations under the independence assumption. Alternatively, the observations could be assumed to come from some other distribution in conjunction with a scoring method.

The basic form of the single change-point problem is as follows. Let $Y_1, \ldots, Y_n$ be ordered random variables and suppose that the distribution of the Y_i changes (at most once) to a different distribution after some number $[n\tau_1]$ of observations, where $0 < \tau_1 < 1$. The multiple change-point case introduces the possibility of more than one change in distribution, say R changes. That is, each grouping of data $Y_{[n\tau_{r-1}]+1}, \ldots, Y_{[n\tau_r]}$, $r = 0, \ldots, R$, is supposed to arise from a different distribution, where for convenience we take $0 = \tau_0 < \tau_1 < \cdots < \tau_R < \tau_{R+1} = 1$. For reviews of theoretical issues connected with change-points see Zacks (1983), Wolfe and Schechtman (1984) and, more recently, Bhattacharya (1994). For modern developments in theory and application see the monograph *Change-Point Problems* edited by Carlstein, Müller and Siegmund (1994).

For applications to DNA sequences, the specific case of the retrospective, multiple change-point problem is of interest. The multiple change-point problem can be broken down into the problem of determining how many change-points exist in a sequence and of determining the locations of these change-points. One can distinguish among three types of approaches, to solve these two problems simultaneously or sequentially, which are relevant for DNA segmentation:

(a) iterated use of tests for single change-points, thus finding a sequence of change-points and associated segments;
(b) minimization of a global objective function to find an arbitrary number of arbitrarily located change-points;
(c) using exceedance of a "local" objective function above a threshold to declare a change-point and to find the associated segments.

There is a need to develop change-point methods which can deal with dependent data in a reasonable way. The Markov chain method allows for dependence and is very flexible; the price to be paid for such flexibility is the explosion in the number of parameters which must be estimated. In this regard sparsely parametrized Markov chain models are of interest, as are data-driven Markov chains. The maximum likelihood method for estimating segments can be extended to include the Markov case as well. Some recent results in the use of logistic or other regression models which incorporate dependency in the form of a covariance between observations may be useful here (e.g., Amfoh, Shaw and Banney, 1994). The following discussion is based on the assumption of independent data.

3.1 Binary Segmentation

To fix ideas, suppose that we are in possession of a likelihood ratio test statistic T for the test of the null hypothesis of no change-point. Assume that we reject the null hypothesis for values of T which are too large. If the null hypothesis of no change-point is rejected for the data at a predetermined significance level α, we declare a change-point at that value $n\tau_{1,1}$ which maximizes T. On each segment, we now apply the test statistic only to that segment. For example, if we reject at the significance level α on both segments, we would obtain two more change-points $n\hat{\tau}_{2,1}$ and $n\hat{\tau}_{2,2}$. We continue in this fashion until the null hypothesis cannot be rejected anymore on any of the current subsegments. This process defines the binary segmentation method.

The algorithm above which continually refines change-points by further splitting subsequences is discussed in Scott and Knott (1974) for the normally distributed case, where connections with cluster analysis are also noted, and a simple conservative method for splitting data into groups is discussed. The upper bound on the probability of splitting into l homogeneous groups and then getting at least one significant split is

$$\alpha^* = 1 - (1 - \alpha)^l,$$

where α is the significance level which is set for the test statistic used to split the groups.

Consistency results for this procedure for estimating the true change-points $\tau_1, \ldots, \tau_R$ under mild conditions, assuming that the $\tau_1, \ldots, \tau_R$ are fixed, have been obtained by Vostrikova (1981); under the more difficult condition that the $\tau_1, \ldots, \tau_R$

may asymptotically move closer together, Venkatraman (1992) has achieved similar results. An example of Venkatraman demonstrates that if some of the R change-points move together such that their distances decrease at a rate $n^{-1/2}$ as n increases, the binary segmentation method can be inconsistent, even if the jump sizes increase with n.

A modified version of this algorithm parallels the idea of forward stepwise regression, as pointed out by Christensen and Rudemo (1996). Their algorithm includes the possibility of removing a change-point which was previously included and of retesting individual change-points at an increased significance level in analogy to forward stepwise regression.

This estimation procedure is intuitive and simple, and the algorithm for carrying out the estimation is straightforward. However, the theory of inference for this method can be quite involved. For the most part, these methods depend on simulation in order to determine overall levels, although asymptotic results are available for classical distributions in the independent observation case.

3.2 Global Segmentation

The idea in global segmentation is to minimize some target criterion over every possible allowable partition of the data into $R + 1$ contiguous segments. A useful target criterion is the deviance function, which naturally extends the idea of least-squares estimation to the exponential family case. The quasideviance may be used as a target criterion when only the mean–variance structure is specified; this leads to a minimum quasideviance approach to change-point estimation. A refinement may be made by adding a penalty function to the target function. Such a penalty function is usually a function of the number of change-points. Use of the Schwarz criterion for change in normal means was developed by Yao (1988); this was extended to the estimation of a step function in Gaussian noise by Yao and Au (1989), and naturally generalizes to the quasideviance setting; for more details see Braun and Müller (1998).

This estimation method is straightforward in principle. The details of implementation are more difficult, if large data sets are involved. Theoretical results on asymptotic properties are available for the change-point problem in a wide variety of settings for independent observations. For example, the multinomial case can be treated within the quasideviance framework; then results are available on consistency of estimators for location of change-points, for number of change-points and for other parameters, as well as asymptotic distribution theory for all estimates.

Such a procedure seems in principle to require one to check all possible partitions of the data into $R+1$ contiguous blocks. The number of possible partitions of the data increases dramatically with the sample size n and number of change-points R and is given by

$$\binom{n-1}{R}.$$

As already noted, this approach is impractical for the large sequences found in molecular biology databases.

The Auger–Lawrence dynamic programming algorithm is designed to compute the global minimum in just the situation described above, and consists of two main steps:

STEP 1. Compute the target criterion for all possible segments.

STEP 2. Sequentially compute the optimal partition for $2, \ldots, R+1$ segments using the information obtained in Step 1.

The algorithm requires roughly $O(n^2T(n))$ operations, where n is the length of the sequence and $T(n)$ is the overhead associated with computing the target criterion in the first step. This method is guaranteed to find the global minimum. It may be possible in some situations to achieve greater efficiency in the overhead. In any case, if direct computation of the global minimum is necessary, then this algorithm or a similar one should be used.

To cut the computational effort even further, ad hoc algorithms may be devised. The following provides an example:

STEP 1. Begin with no change-points.

STEP 2. Determine a change-point which gives the maximum reduction in the target criterion.

STEP 3. Split each segment, and take as the next change-point a point determining a split which produces the maximum reduction in the target criterion.

STEP 4. Repeat Step 3 $R - 2$ more times to obtain R change-points.

This method shares with the binary segmentation procedure the property of maintaining the original splits; it may be easier to accept a method where

two large segments are further refined rather than where the original large segments are erased and differing change-points are introduced.

Computationally, this hierarchical approach requires $O(n)$ operations. However, it is not guaranteed to find the global minimum and in fact may fail badly. Variants which allow certain changes on previously established segments may be of interest. For instance, after a new change-point has been identified, a loop can be added which iterates through the current change-points and repositions them within their respective segments with respect to the global minimum deviance target. Such computational methods are probably useful mainly for screening or for very large datasets.

Ultimately, there is a tradeoff to be made between the sophistication of the model and the computational effort needed to calculate results. We note that this problem is combinatorial in nature, and similar problems have been successfully tackled in the engineering and physics literature by related methods, such as simulated annealing, or "artificial intelligence" methods, such as the taboo search (Cvijovic and Klinowski, 1995).

3.3 Scan Statistics

In a brief digression, we touch upon a widely used method for detecting regions of biological interest in DNA sequences, namely, the method of scan statistics. The method may also be thought of as a visualization method for sequence segmentation.

In scanning techniques, the letters of an alphabet of interest are numerically scored, for example, by assigning $+1$ to G or C and 0 to A or T, and a moving average is plotted (see Staden, 1984). An example of this approach for bacteriophage λ is shown in Figure 6, where at least two regimes of differing $G + C$ proportions can be discerned.

Formally, we take $Y_1, \ldots, Y_n$ as defined above and define a mapping g from an alphabet of interest into the real numbers. Next, we consider the partial sum S_j of L observations of $g(Y_i)$ starting at position j in the sequence, where $j = 1, \ldots, n - L + 1$, that is,

$$S_j = \sum_{i=j}^{j+L} g(Y_i).$$

The quantities S_j are evaluated locally at each point of the sequence and are known as L-scan statistics. Intuitively, high values of S_j represent locations in the genome where there is a concentration of observations with high values of g. The stochastic properties of L-scans have been derived by Karlin, Dembo and Kawabata (1990) and Karlin and Altschul (1990). This involves a study of the deviation from the independent sequence case with the aim to pinpoint areas of statistically significant deviation, and includes the Markov-dependent case. For review and further development of these approaches, see Karlin and Dembo (1992) and Karlin and Brendel (1992).

The running window approach is a graphical presentation of the L-scan approach. This leads naturally to the questions of optimal assignment of scores to the alphabet and selection of window size so that important features will be detected visually. These issues are explored in Tajima (1991) under the independence assumption by means of simulation studies and elementary probabilistic reasoning. Also of interest in this connection is the work of

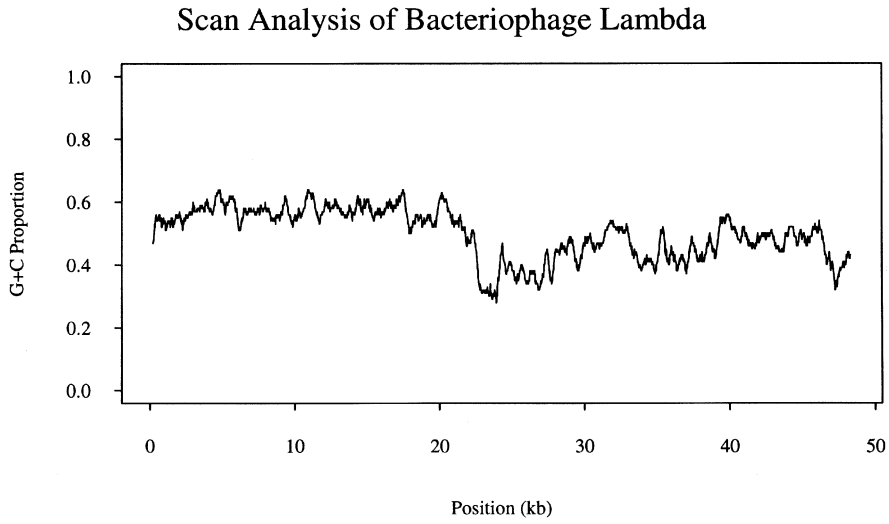

FIG. 6. *A typical scan analysis of bacteriophage λ. A moving average $G + C$ content with a 400-base window centered at the current position is used to highlight areas of increased $G + C$ content.*

Stoffer, Tyler and McDougall (1993), who derive the optimal score to emphasize periodicities in the sequence data, assuming a stationary time-dependent structure.

In general, the L-scan statistics correspond to smoothing the data, and the window selection problem corresponds to the familiar bandwidth selection problem in the smoothing literature. Such L-scan statistics may provide a pointer to locations in the genome where there are interesting things. These methods are not designed to estimate change-points, but the transformed data can also be used as input to methods for detecting change-points.

3.4 Locally Weighted Split Polynomial Regression

Kernel smoothing and similar nonparametric regression methods allow modeling smooth variation of probabilities and distribution parameters within segments, while suitably modified versions also allow for the local detection of change-points; see Müller (1992) for a kernel method and Müller (1993) for an analogous proposal based on local polynomial fitting. For specific versions, the rate of convergence of estimated change-point locations is shown to be n^{-1} in Müller and Song (1997). Under the more restrictive assumption of Gaussian observations, a similar result was obtained by Loader (1996). Adapting methods developed in Fan, Heckman and Wand (1995) and refined in Fan and Gijbels (1996), these ideas can be extended to localized quasilikelihood models. In the following we propose a framework for the application of split local polynomial fitting to the DNA segmentation problem.

In general, we could estimate a smooth mean regression function $g(x) = E[Y|X = x]$ from data (X_i, Y_i) by locally fitting polynomial functions. The fitting procedure employs kernel weights generated by $K_b(z) = (1/b)K(z/b)$. Here K is a nonnegative kernel or weight function with compact support; for instance, $K(x) = (1 - x^2)_+^\mu$, where $\mu = 0$ (rectangular kernel), $\mu = 1$ (Epanechnikov kernel), $\mu = 2$ (biquadratic kernel) or $\mu > 2$. Allocating the weight $K_b(X_i - x)$ to the observation (X_i, Y_i) in the quasilikelihood Q, the local polynomial kernel estimator $\hat{g}$ of a mean regression function g is given by

$$\hat{g}(x; p, b) = \tilde{\beta}_0(x),$$

where

$$\begin{aligned}\tilde{\beta}(x) &= (\tilde{\beta}_0, \ldots, \tilde{\beta}_p)'(x) \\ &= \arg\max_{\beta \in \Re^{p+1}} \sum_{i=1}^{n} Q\big((\beta_0 + \cdots + \beta_p(X_i - x)^p), Y_i\big) \\ &\qquad \cdot w_i K_b(X_i - x)\end{aligned}$$

and the w_i denote additional case weights which are sometimes useful.

The weighted least squares approach is a special case, where

$$\begin{aligned}\tilde{\beta}(x) = \arg\min_{\beta \in \Re^{p+1}} &\sum_{i=1}^{n} \{\beta_0 + \cdots + \beta_p(X_i - x)^p - Y_i\}^2 \\ &\cdot w_i K_b(X_i - x).\end{aligned}$$

In this prescription for local polynomial smoothing, the smoothness of the function estimate $\hat{g}$ is inherited from the smoothess of the kernel function K. To incorporate segmentation, assume that $S_1, S_2, \ldots, S_l$ are l given segments.

The idea is to fit a smooth function within the segments, but to allow for jump discontinuities at the endpoints of the segments. For the local polynomial modeling approach, this is easily implemented by using the modified target criterion

$$\begin{aligned}\hat{\beta}(x) = \arg\min_{\beta \in \Re^{p+1}} &\sum_{j=1}^{l} \sum_{i=1}^{n} 1_{\{x \in S_l,\, X_i \in S_l\}} \\ &\cdot \{\beta_0 + \cdots + \beta_p(X_i - x)^p - Y_i\}^2 \\ &\cdot w_i K_b(X_i - x),\end{aligned}$$

in the least squares case and an analogous criterion in the quasilikelihood case. This means that the function values are fitted for each segment separately, not taking any observations from other segments into account. The local polynomial method automatically provides for the necessary adjustments near the endpoints; for analogous implementations with kernel methods, such adjustments can be explicitly implemented by using boundary kernels.

This method provides segmented smooth fits once the segments have been determined and a smoothing parameter b is provided. In practice, one needs to estimate the segments, usually by determining first a suitable smoothing parameter or bandwidth and then estimating the endpoints. One common principle to locate the segments is to choose those locations as endpoints of segments where the resulting jumps in the function g are maximized. Assume that one considers just two segments S^+ and S^-, which have a common endpoint at t, that is, $S^- = (-\infty, t)$ and $S^+ = [t, \infty)$. We then define the function

$$\hat{\Delta}(t) = \hat{\beta}_0^+(t) - \hat{\beta}_0^-(t),$$

where $\hat{\beta}^\pm(t) = (\hat{\beta}_0^{\ \pm}(t), \ldots, \hat{\beta}_p^{\ \pm}(t))'$ and

$$\begin{aligned}\hat{\beta}^\pm(t) = \arg\min_{\beta \in \Re^{p+1}} &\sum_{i=1}^{n} 1_{\{X_i \in S^\pm\}} \\ &\cdot \big\{\beta_0 + \cdots + \beta_p(X_i - t)^p - Y_i\big\}^2 \\ &\cdot w_i K_b(X_i - t).\end{aligned}$$

The change-point estimate is then $\hat{\tau} = \arg\max \hat{\Delta}(t)$. Denote the locations of the ordered maxima of $\hat{\Delta}(\cdot)$ by $\hat{\tau}_1, \hat{\tau}_2, \ldots$, that is, $\Delta(\hat{\tau}_1) \geq \Delta(\hat{\tau}_2) \geq \Delta(\hat{\tau}_3) \geq \cdots$. Choosing l change-points, the $(l+1)$ segments generated by the sequence $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_l$ are then $\hat{S}_1 = (-\infty, \hat{\tau}_{(1)})$, $\hat{S}_2 = [\hat{\tau}_{(1)}, \hat{\tau}_{(2)}), \ldots, \hat{S}_l = [\hat{\tau}_{(l-1)}, \hat{\tau}_{(l)})$ and $\hat{S}_{l+1} = [\hat{\tau}_{(l)}, \infty)$, where $(\hat{\tau}_{(1)}, \ldots, \hat{\tau}_{(l)})$ are the order statistics of $(\hat{\tau}_1, \ldots, \hat{\tau}_l)$. The estimated function is obtained as $\hat{g}(x) = \hat{\beta}_0(x)$, where in the definition of $\hat{\beta}(x)$ above the segments S_i are replaced with estimated segments $\hat{S}_i$. This estimate is adapted to the estimated segmentation by declaring the estimated change-points as endpoints of the support.

Theoretical results for this kind of procedure show that the rate of convergence for estimated change-points is $O_p(n^{-1})$. Invariance principles hold for the fixed jump sizes when using smooth boundary kernels, at the cost of slightly slower rates for estimated change-points; they hold for split local polynomial change-point estimators only for the case of contiguous jump sizes. That means that one assumes that asymptotically the jump sizes converge to zero; this is also known as the case of a *faint signal* or the small jumps case. These asymptotic properties are explored in Müller (1992) and Müller and Song (1997).

3.5 Local Nonparametric Segmentation in Action

A big advantage of the local split polynomial fitting method, compared with global schemes, is the fact that everything is based on local computations, and efficient algorithms for local polynomial or kernel fitting require only $O(n)$ multiplications. Bandwidth choice and selection of the number of change-points increase the computational effort again. Nevertheless, the local methods scale up to larger numbers of bases much better than the global methods, which are hampered by the requirement to solve high-dimensional optimization problems.

Even if one eventually prefers global, for instance Bayesian, segmentation methods, the locally based nonparametric approaches provide easy-to-compute starting configurations for global segmentation schemes. Since global schemes typically depend on iteration algorithms, the computational ease as well as quality of an initial segmentation can be a crucial ingredient for the success of a global method.

For the practical implementation of local nonparametric segmentation algorithms, certain auxiliary parameters need to be determined. For the implementation with split local polynomials, these are the order of the local polynomial to be fitted and the bandwidth. As for the order of the polynomial, for virtually all applications fitting of local first-order polynomials, that is, choice of $p = 1$, corresponding to the fitting of local lines, will suffice. A harder problem is the choice of the bandwidth, which can be coupled with the problem of finding the number of change-points l.

One approach is to base these choices on an estimate of the prediction error such as that provided by the leave-one-out cross-validation sum of squares,

$$\mathrm{CV}(l, b) = \sum_{i=1}^{n} (\hat{g}^{(-i)}(X_i) - Y_i)^2 w_i,$$

which is to be minimized with respect to l and b. Other strategies which are worth exploring include two-step procedures, selecting first a bandwidth based on pilot methods and then the number of change-points via cross-validation. The problem of choosing the number of segments can alternatively be phrased as the problem of finding a critical threshold Δ_0, such that change-points occur at all t where

$$\left|\hat{\Delta}(t)\right| \geq \Delta_0.$$

We note that within the segments, the mean regression function $\hat{g}$ is allowed to vary smoothly in the local method, but is required to be constant in the global segmentation schemes. This increased flexibility is an advantage of the local method. The disadvantage of the local methods is that they depend heavily on the choice of a smoothing parameter b as well as of a critical determination of number of change-points, or alternatively, of the critical jump size value Δ_0. This drawback is shared with iterated testing and stepwise regression methods, where levels of the iterated tests must be specified which then also have a strong impact on the number of segments estimated. If Δ_0 is chosen too small, spurious jumps which are caused by noise in the data will be erroneously detected and the number of segments will be overestimated; the opposite happens if Δ_0 is chosen too large. These choices become much more tricky for dependent data (see Lombard and Hart, 1994). However, at a minimum the outcomes of such local approaches could be used as starting points for the more elaborate global methods.

After determining the change-points, the usual L^p-convergence properties of the curve estimates on the segments are preserved due to fast convergence of the estimated change-point locations, as demonstrated in Müller (1992). In several limited simulations (not shown) with the binomial, exponential and Gaussian cases, this local segmentation method performed well overall. While it is well known that cross-validation gives consistent (if not sometimes

overly variable) bandwidth choices, the asymptotic properties of the corresponding selector of the number of change-points need to be further investigated. A competitor is the determination of the multiplicity of change-points with the Schwarz criterion (Yao, 1988).

Demonstrating these local segmentation ideas with the bacteriophage λ sequence, we first binned the data into nonoverlapping windows of length 200, leading to 242 such bins. The bandwidth was chosen as 5,000 (the unit is base-pair here). Both cross-validation and a version of the pilot method (Müller, 1985) gave bandwidths which were considered too small. The function $\hat{\Delta}(t)$ under these specifications and the cross-validation sum of squares in dependency on the number of change-points are shown in Figure 7. Constraints were imposed for change-points to have a minimum distance of 5,000 from one another as well as from the ends of the data. Under this constraint, the order of the local maxima of the function $\Delta(t)$ was as follows (in units of 1,000 base pairs): $\hat{\tau}_1 = 22.6$; $\hat{\tau}_2 = 33.2$; $\hat{\tau}_3 = 39.2$; $\hat{\tau}_4 = 6.0$, $\hat{\tau}_5 = 27.8$; $\hat{\tau}_6 = 17.6$; and $\hat{\tau}_7 = 11.4$.

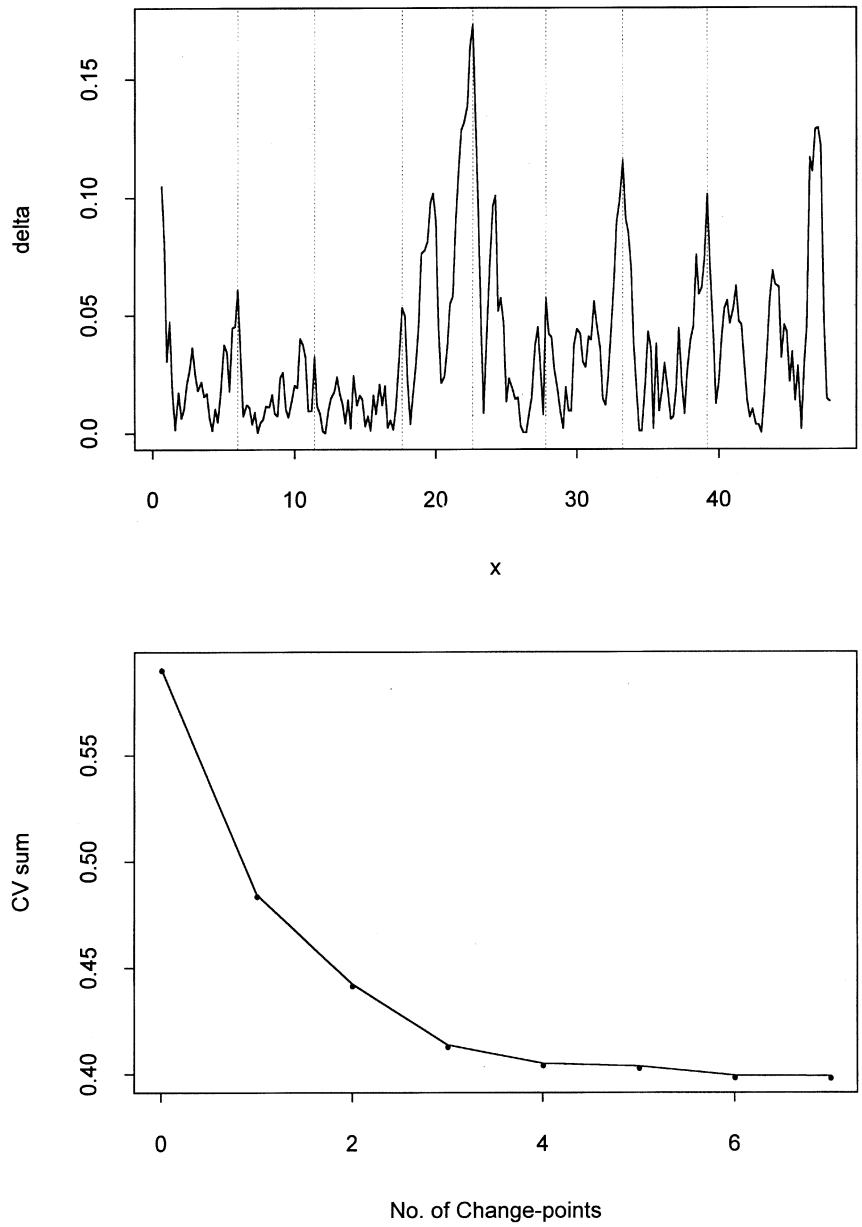

FIG. 7. (*Top*) *the function* $\hat{\Delta}(t)$ *indicating where change-points might be located in the bacteriophage* λ *sequence*; (*bottom*) *the cross-validation sum of squares as a function of the number of change-points.*

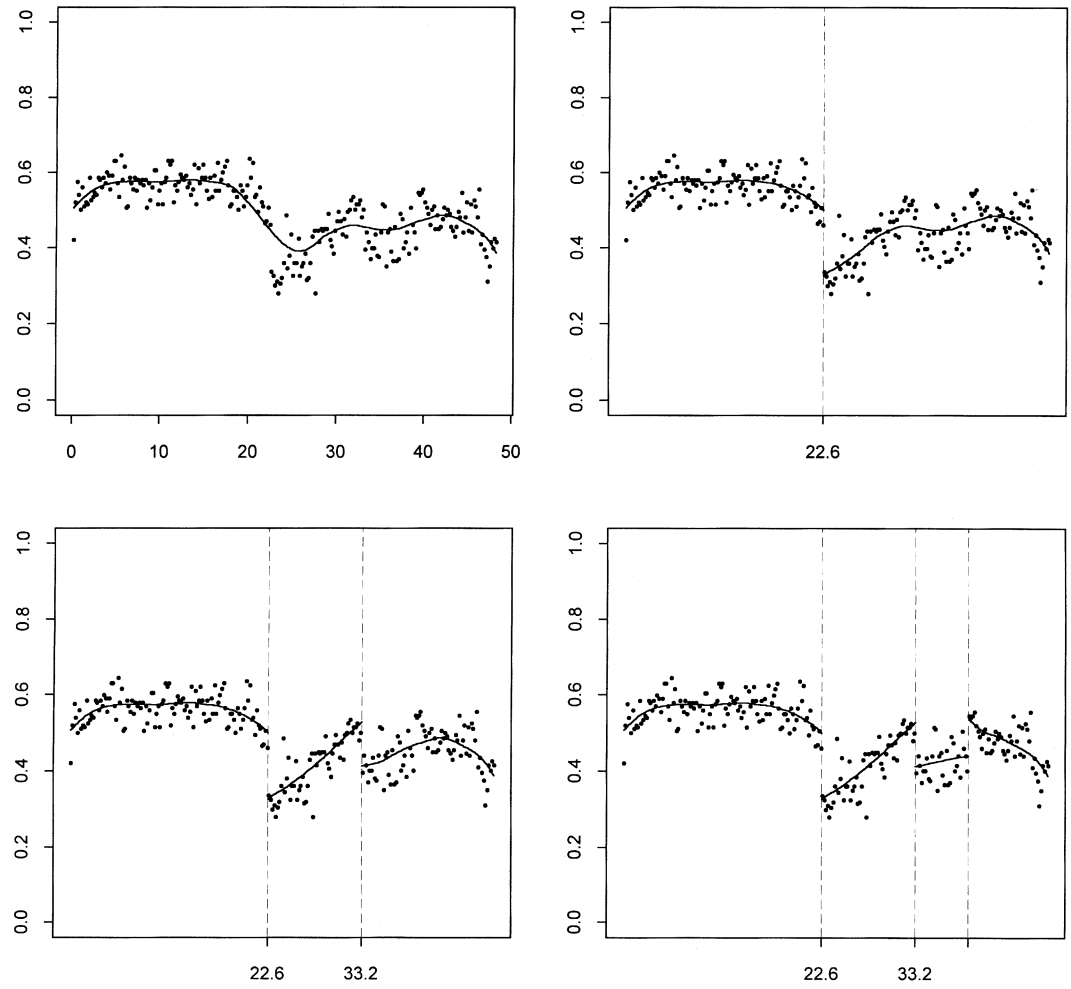

FIG. 8. *Choosing bandwidth* 5,000, *fits with split local polynomials, assuming no* (*upper left*), *one* (*upper right*), *two* (*lower left*) *and three* (*lower right*) *change-points for the bacteriophage* λ *sequence.*

The fits which were obtained when assuming 0–7 change-points are shown in Figures 8 and 9. From the cross-validation plot in Figure 7 as well as from visual inspection of the various fits in Figures 8 and 9, it appears that assuming about three change-points leads to a reasonable fit for these data.

4. DISCUSSION

We have observed that segmentation models are useful for modeling DNA sequence data. They are soundly underpinned by biological theory, which demonstrates and predicts such segments within the genome—so far, results from other models such as the long-range dependence model or the walking Markov model remain somewhat mysterious. Applications of segmentation models are locating introns, sequence alignment, decomposition of long sequences into homogeneous pieces and evolutionary studies. The various segmentation models which have been used are considered within the unifying framework of the multiple change-point problem. Besides reviewing the segmentation techniques that have been used successfully, we briefly discussed a promising new visualization method: local segmentation by nonparametric regression, which in addition accommodates drift within segments.

As noted by Geyer (1995), we should distinguish between the various concerns involved in the analysis of DNA sequence. First, we have the underlying biological theory, which encompasses the physical and chemical composition of the sequence, as well as its evolutionary history. This theory may or may not make definite predictions about the structure of the sequence, but in any case should be respected by the statistical model we choose for analysis. Second, we have the stochastic model of the sequence. Independence models may be appropriate when we

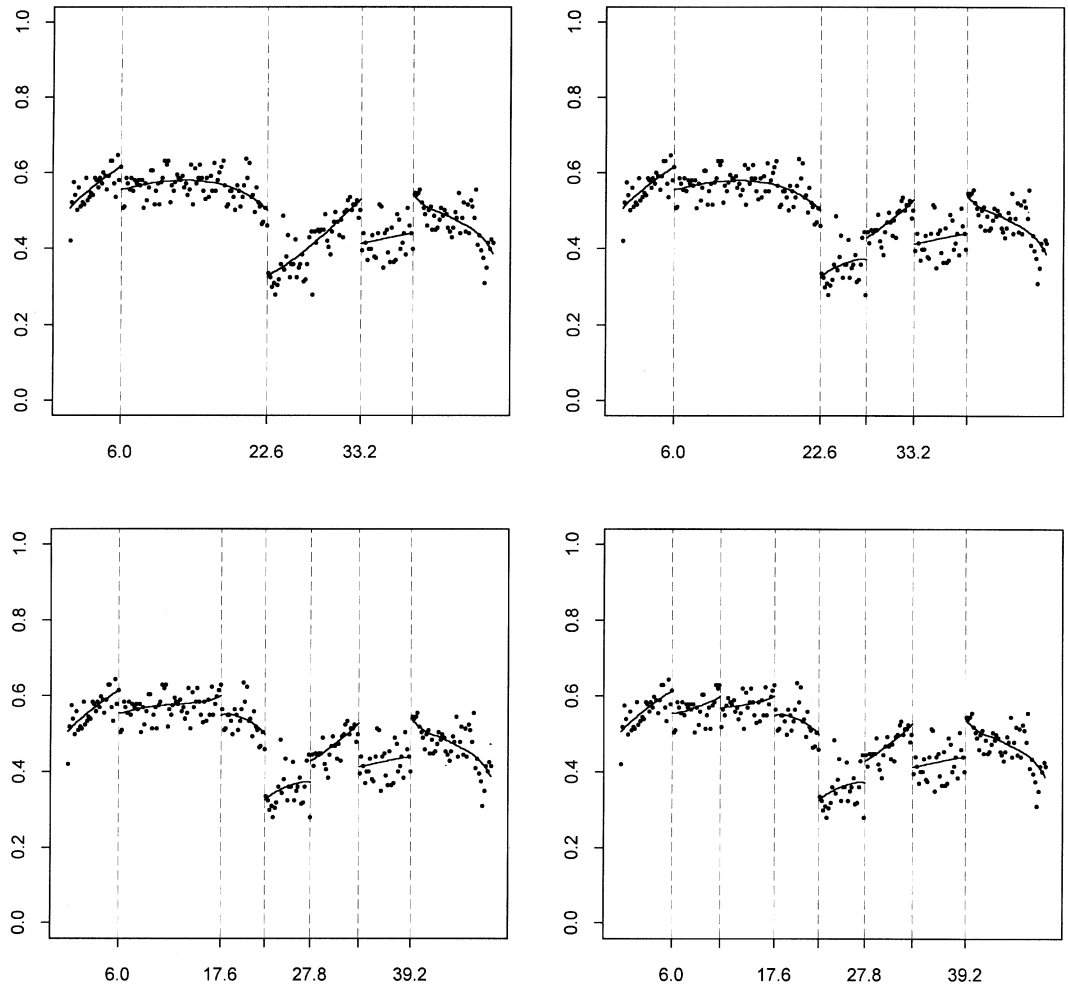

FIG. 9. *Split local polynomial fits for the bacteriophage λ sequence, assuming four (upper left), five (upper right), six (lower left) and seven (lower right) change-points.*

are interested in structural properties, and may adequately reflect biological functional constraints. More general models which incorporate dependence structures based on the underlying chemistry and physics are of interest but have not been well developed to date. Third, we have the statistical theory of inference and its associated mathematical machinery. Fourth, and finally, we have the practical problem of actually computing the quantities of interest from the given sequence data.

The method of Fu and Curnow is thus usable when interest centers on the structural properties of the sequence. The problem of choosing l changed segments is seen as a special case of the problem of choosing $R = 2l$ suitably restricted change-points. The model requires prior knowledge of the parameters of the underlying distributions; the restriction on segment length must be evaluated within the scope of its intended application. As usual with more complicated models, the classical inferential theory may be difficult. Also, work needs to be done to determine the sensitivity of this approach to misspecification of those parameters, as well as to allow estimation of the parameters from the data, perhaps by an iterative approach. Dynamic programming approaches provide $O(n^2)$ or faster methods for computing the locations of the segments.

The hidden Markov chain model seems overall to be suitably flexible, finding practical use in many settings. It is readily customized to capture underlying biological theory, and the estimation theory is particularly straightforward. It has been noted that this very flexibility may also be a weak point, requiring large data sets for successful use. Also, there is little classical theory addressing model selection or inference. Dynamic programming techniques provide fast computation of smoothing distributions given the parameters and the data, but

estimation of the parameters requires a method such as the EM algorithm, which may fail to find the global optimum.

The framework put forward by Liu and Lawrence provides a general setting for such problems. It can incorporate biological information such as the existence of secondary protein structure. The theory of inference is straightforward and, under some conditions, it may be possible to marginalize over nuisance parameters. Computational issues center around implementation of simulation methods for computing posterior distributions, methods for sensitivity analysis and optimizing the effort needed.

Local segmentation by split local polynomial fitting provides a formal extension of scan statistics for visualizing regions of interest. The walking Markov model contains the idea of a smoothly varying change in the underlying structure, but does not seem as parsimonious. Local segmentation provides for estimation of location and size of jump points, and naturally implements the idea of gradual change within segments. The method, as are most such methods, is analyzed under the assumption of independent data. No further restrictions are necessary; computation is local and therefore fast and scales up to large data sets; and a fairly comprehensive asymptotic theory is in place, although there remain many open questions. The relatively fast implementation of the local nonparametric methods allows visual selection of a bandwidth or window width and thresholds, and altogether this makes the local nonparametric method a very attractive alternative to the global schemes. Global optimization schemes with very large sequences will require further research in improving their computational properties, as will extensions to dependent data and use of covariate information.

There is a large body of theory and methods for estimation and inference in the multiple change-point problem which may be taken advantage of when contemplating the DNA segmentation problem. For all such methods, there are some practical and numerical issues to be resolved before large-scale applications are possible—in particular, methods with global target criteria need to be implemented with efficient programming approaches. It is mandatory that the statistical model chosen reflects the underlying biological process and its associated chemical and physical constraints, at least to the extent necessary to obtain valid predictions and interpretations.

We note that even though DNA sequence data are known to show dependence of all sorts, the independence assumption is so attractive theoretically and practically that it is usually made, with the understanding that it allows capturing the relevant functional constraints to a large extent. It seems that the necessary infrastructure is in place for development of both theory and application of multiple change-point models for dependent data, both within and beyond the established hidden Markov chain models. Such developments will be immediately usable in the DNA segmentation problem.

We thus expect that the multiple change-point problem, especially for dependent data, and in particular the local approaches, will continue to be an important research area. It is of interest to develop tests and diagnostics regarding the issue of whether change-points associated with discontinuities in an otherwise smooth function are present (for a proposal in this direction, see Müller and Stadtmüller, 1997) and to find reliable and computationally feasible estimation procedures for the number of change-points present. The potential applications range far beyond DNA sequence data to other sequence data problems in biology and in science and technology in general.

ACKNOWLEDGMENTS

We thank two anonymous referees for many valuable comments which led to substantial improvements in the presentation. One referee pointed out several additional references and points of view which helped to broaden the scope of this article. This research was supported in part by NSF Grants DMS-94-04906 and DMS-96-25984 and by NSA Grant MDA 904-96-10026.

REFERENCES

AMFOH, K. K., SHAW, R. F. and BONNEY, G. E. (1994). The use of logistic models for the analysis of codon frequencies of DNA sequences in terms of explanatory variables. *Biometrics* **50** 1054–1063.

AUGER, I. E. and LAWRENCE, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology* **51** 39–54.

AVNIR, D., BIHAM, O., LIDAR, D. and MALCAI, O. (1998). Is the geometry of Nature fractal? *Science* **279** 39–40.

BARRY, D. and HARTIGAN, J. A. (1992). Product partition models for change-point models. *Ann. Statist.* **20** 260–279.

BEMENT, T. R. and WATERMAN, M. S. (1977). Locating maximum variance segments in sequential data. *Mathematical Geology* **9** 55–61.

BERNARDI, G., OLOFSSON, B., FILIPSKI, J., ZERIAL, M., SALINAS, J., CUNY, G., MEUNIER-ROTIVAL, M. and RODIER, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228** 953–958.

BHATTACHARYA, P. K. (1994). Some aspects of change-point analysis. In *Change-Point Problems* (E. Carlstein, H.-G. Müller and D. Siegmund, eds.) 28–56. IMS, Hayward, CA.

BICKMORE, W. and SUMNER, A. T. (1989). Mammalian chromosome banding—an expression of genome organization. *Trends in Genetics* **5** 144–148.

BRAUN, J. V. and MÜLLER, H. G. (1998). Quasi-likelihood fitting of multiple change-points, with application to DNA segmentation. Technical report, Univ. California, Davis.

BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series: Theory and Methods*. Springer, New York.

BULDYREV, S. V., GOLDBERGER, A. L., HAVLIN, S., PENG, C.-K., SIMONS, M., SCIORTINO, F. and STANLEY, H. E. (1993). Comment. *Phys. Rev. Lett.* **71** 1776.

CARLIN, B. P., GELFAND, A. E. and SMITH, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *J. Roy. Statist. Soc. Ser. B* **41** 389–405.

CARLSTEIN, E., MÜLLER, H.-G. and SIEGMUND, D., eds. (1994). *Change-Point Problems*. IMS Hayward, CA.

CHRISTENSEN, J. and RUDEMO, M. (1996). Multiple change-point analysis of disease incidence rates. *Preventive Veterinary Medicine* **26** 53–76.

CHURCHILL, G. A. (1989). Stochastic models for heterogenous DNA sequences. *Bulletin of Mathematical Biology* **51** 79–94.

CHURCHILL, G. A. (1992). Hidden Markov chains and the analysis of genome structure. *Computers in Chemistry* **16** 107–115.

CURNOW, R. N. and KIRKWOOD, T. B. L. (1989). Statistical analysis of deoxyribonucleic acid sequence data—a review. *J. Roy. Statist. Soc. Ser. B* **152** 199–220.

CVIJOVIC, D. and KLINOWSKI, J. (1995). Taboo search—an approach to the multiple minima problem. *Science* **267** 664–666.

DUPUIS, J. (1994). Change-point problem in determination of identity-by-descent. Technical Report 1, Stanford Univ.

ELTON, R. A. (1974). Theoretical models for heterogeneity of base composition in DNA. *Journal of Theoretical Biology* **45** 533–553.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling.* Chapman and Hall, London.

FAN, J., HECKMAN, N. E. and WAND, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90** 141–150.

FICKETT, J. W., TORNEY, D. C. and WOLF, D. R. (1992). Base compositional structure of genomes. *Genomics* **13** 1056–1064.

FU, Y.-X. and CURNOW, R. N. (1990). Maximum likelihood estimation of multiple change points. *Biometrika* **77** 563–573.

GEYER, C. J. (1995). Comment on "Bayesian computation and stochastic systems," by J. Besag, P. Green, D. Higdon and K. Mengerson. *Statist. Sci.* **10** 46–48.

GILLESPIE, J. H. (1991). *The Causes of Molecular Evolution.* Oxford Univ. Press.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **41** 389–405.

HARTIGAN, J. A. (1990). Partition models. *Comm. Statist. Theory Methods* **19** 2745–2756.

HOLMQUIST, G. P. (1989). Evolution of chromosome bands: Molecular ecology of noncoding DNA. *Journal of Molecular Evolution* **28** 469–486.

IKEMURA, T., WADA, K. and AOTA, S. (1990). Giant G+C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions. *Genomics* **8** 207–216.

JOSSE, J., KAISER, A. D. and KORNBERG, A. (1961). Enzymatic synthesis of deoxyribonucleic acid. VII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *Journal of Biological Chemistry* **236** 864–875.

KARLIN, S. and ALTSCHUL, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat. Acad. Sci. U.S.A.* **87** 2264–2268.

KARLIN, S. and BRENDEL, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science* **257** 39–49.

KARLIN, S. and BRENDEL, V. (1993). Patchiness and correlations in DNA sequences. *Science* **259** 677–680.

KARLIN, S. and DEMBO, A. (1992). Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. in Appl. Probab.* **24** 113–140.

KARLIN, S., DEMBO, A. and KAWABATA, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Statist.* **18** 571–581.

KARLIN, S., OST, F. and BLAISDELL, B. E. (1989). Patterns in DNA and amino acid sequences and their statistical significance. In *Mathematical Methods for DNA Sequences* (M. S. Waterman, ed.) 133–158. CRC Press, Boca Raton, FL.

KIMURA, M. (1983). *The Neutral Allele Theory of Molecular Evolution.* Cambridge Univ. Press.

KROGH, A., BROWN, M., MIAN, I. S., SJÖLANDER, K. and HAUSSLER, D. (1994). Hidden Markov models in computational biology: application to protein modeling. *Journal of Molecular Biology* **235** 1501–1531.

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 alignments. *Science* **262** 208–214.

LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966.

LIU, J. S. and LAWRENCE, C. E. (1996). Unified Gibbs method for biological sequence analysis. In *Proceedings of the Biometrics Section* 194–199. Amer. Statist. Assoc., Alexandria, VA.

LIU, J. S., NEUWALD, A. F. and LAWRENCE, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist. Assoc.* **90** 1–15.

LOADER, C. R. (1996). Change point estimation using nonparametric regression. *Ann. Statist.* **24** 1667–1678.

LOMBARD, F. and HART, J. D. (1994). The analysis of change-point data with dependent errors. In *Change-Point Problems* (E. Carlstein, H.-G. Müller and D. Siegmund, eds.) 194–209. IMS, Hayward, CA.

MADDOX, J. (1992). Long-range correlations within DNA. *Nature* **358** 103.

MESELSON, M., STAHL, F. W. and VINOGRAD, J. (1957). Equilibrium sedimentation of macromolecules in density gradients. *Proc. Nat. Acad. Sci. U.S.A.* **43** 581–588.

MÜLLER, H. G. (1985). Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statist. Decisions Suppl.* **2** 193–206.

MÜLLER, H. G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.* **20** 737–761.

MÜLLER, H. G. (1993). Comment on "Local regression: automatic kernel carpentry," by T. Hastie and C. Loader. *Statist. Sci.* **8** 134–139.

MÜLLER, H. G. and SONG, K. S. (1997). A two-stage procedure for change-point detection in nonparametric regression. *Statist. Probab. Lett.* **34** 323–335.

MÜLLER, H. G. and STADTMÜLLER, U. (1997). Discontinuous versus smooth regression. Technical report, Univ. California, Davis.

NEE, S. (1992). Uncorrelated DNA walks. *Nature* **357** 450.

NEUWALD, A. F., LIU, J. S. and LAWRENCE, C. E. (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science* **4** 1618–1632.

PENG, C. K., BULDYREV, S. V., GOLDBERGER, A. L., HAVLIN, S., SCIORTINO, F., SIMONS, M. and STANLEY, H. E. (1992). Long-

range correlation in nucleotide sequences. *Nature* **356** 168–270.

PENNINI, E. (1997). Microbial genomes come tumbling in. *Science* **277** 1433.

PRABHU, V. V. and CLAVERLE, J.-M. (1992). Correlations in intronless DNA. *Nature* **359** 782.

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–286.

RAFTERY, A. E. and AKMAN, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73** 85–89.

SANGER, F., COULSON, A. R., HONG, G. F., HILL, D. F. and PETERSEN, G. B. (1982). Nucleotide sequence of bacteriophage lambda DNA. *Journal of Molecular Biology* **162** 729–773.

SCHERER, S., MCPEEK, M. S. and SPEED, T. P. (1994). Atypical regions in large genomic DNA sequences. *Proc. Nat. Acad. Sci. U.S.A.* **91** 7134–7138.

SCHWEIZER, D. and LOIDL, J. (1987). A model for heterochromatin dispersion and the evolution of C-band patterns. *Chromosomes Today* **9** 61–74.

SCOTT, A. J. and KNOTT, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* **30** 507–512.

SHAPIRO, H. S. and CHARGAFF, E. (1960). Studies on the nucleotide arrangement in deoxyribonucleic acid. IV. Patterns of nucleotide sequence in the deoxyribonucleic acid of rye germ and its fractions. *Biochimica et Biophysica Acta* **39** 68–82.

SKALKA, A., BURGI, E. and HERSHEY, A. D. (1968). Segmental distribution of nucleotides in the DNA of bacteriophage lambda. *Journal of Molecular Biology* **34** 1–16.

SMITH, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika* **62** 407–416.

STADEN, R. (1984). Graphical methods to determine the function of nucleic acid sequences. *Nucleic Acids Research* **12** 521–538.

STEPHENS, D. A. (1994). Bayesian retrospective multiple change-point identification. *J. Roy. Statist. Soc. Ser. B* **43** 159–178.

STOFFER, D. S., TYLER, D. E. and MCDOUGALL, A. J. (1993). Spectral analysis for categorical time series: scaling and the spectral envelope. *Biometrika* **80** 611–622.

TAJIMA, F. (1991). Determination of window size for analyzing DNA sequences. *Journal of Molecular Evolution* **33** 470–473.

VENKATRAMAN, E. S. (1992). Consistency results in multiple change-point situations. Technical report, Dept. Statistics, Stanford Univ.

VOSS, R. F. (1992). Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* **68** 3805–3808.

VOSS, R. F. (1993). Comment. *Phys. Rev. Lett.* **71** 1777.

VOSTRIKOVA, L. J. (1981). Detecting "disorder" in multidimensional random processes. *Soviet Math. Dokl.* **24** 55–59.

WALLENSTEIN, S., NAUS, J. and GLAZ, J. (1994). Power of the scan statistic in detecting a changed segment in a Bernoulli sequence. *Biometrika* **81** 595–601.

WOLFE, D. A. and SCHECHTMAN, E. (1984). Nonparametric statistical procedures for the changepoint problem. *J. Statist. Plann. Inference* **9** 389–396.

YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189.

YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A.* **51** 370–381.

ZACKS, S. (1983). Survey of classical and Bayesian approaches to the change-point problem: fixed sample and sequential procedures of testing and estimation. In *Recent Advances in Statistics* (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.) 245–269. Academic Press, New York.