Markov models of genome segmentation

Vivek Thakur,[1] Rajeev K. Azad,[2] and Ram Ramaswamy[1,3]
[1]*Center for Computational Biology and Bioinformatics, School of Information Technology, Jawaharlal Nehru University, New Delhi 110 067, India*
[2]*Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*
[3]*School of Physical Sciences, Jawaharlal Nehru University, New Delhi 110 067, India*
(Received 2 March 2006; revised manuscript received 19 June 2006; published 17 January 2007)

We introduce Markov models for segmentation of symbolic sequences, extending a segmentation procedure based on the Jensen-Shannon divergence that has been introduced earlier. Higher-order Markov models are more sensitive to the details of local patterns and in application to genome analysis, this makes it possible to segment a sequence at positions that are biologically meaningful. We show the advantage of higher-order Markov-model-based segmentation procedures in detecting compositional inhomogeneity in chimeric DNA sequences constructed from genomes of diverse species, and in application to the *E. coli* K12 genome, boundaries of genomic islands, cryptic prophages, and horizontally acquired regions are accurately identified.

DOI: 10.1103/PhysRevE.75.011915 PACS number(s): 87.15.Cc

I. INTRODUCTION

The genome of an organism is a linear or circular DNA molecule and can be represented as a symbolic string in the alphabet {A,T,C,G}, corresponding to nucleotides adenine, thymine, cytosine, and guanine. At present the complete genomes of nearly 400 organisms are known [1], and this makes their detailed and extensive analysis possible both at an individual as well as at a comparative level [2]. Studies over the past two decades have revealed that the genome typically houses a variety of distinctive and diverse features, all of which taken together constitute the "blueprint" for an organism [3]. Some of these features are functionally important: protein coding sequences, regulatory sequences, operons, promoters, etc. Others are also of functional importance, but have evolutionary significance: horizontally transferred regions, duplications, prophages, etc. Beyond this, there are other genomic regions with specific structural properties such as CpG islands, isochores, etc.

It is now possible to apply statistical tools to uncover the underlying patterns of organization. Studies that identify and characterize compositional heterogeneities within genomes [4] yield a picture of genomes as *patchy*. Alternating regions of DNA of variable length are locally homogeneous with respect to specific statistical or biological properties, but on larger scales DNA is heterogeneous [5]. This heterogeneity in the distribution of statistical properties is supported by experimental evidence on DNA melting and density gradient centrifugation [6].

Base composition heterogeneity can sometimes be detected by elementary techniques such as a sliding window analysis. This procedure is sensitive to the size of the window [7,8] and the manner in which the analysis is carried out, apart from the fact that such methods apply only in the simplest of cases. More sophisticated approaches to detecting heterogeneity attempt to "segment" DNA into subsequences that are homogeneous with respect to a given criterion. By construction each segment will be distinct from its immediate neighbors, and the objective is to determine which criterion to use that results in the segments having biological significance. Tools that have been employed in segmentation algorithms include hidden Markov models (HMMs) and multiple change-point approaches, each of which has been validated to some degree in earlier work [9]. The use of HMMs in segmentation [10,11] involved the application of a first-order model to short DNA sequences of mitochondrial and phage genomes, which has later been extended to higher orders. Another study reported interesting biological applications of segmentation using different orders of HMMs [12]. A Bayesian multiple change-point approach [13,14] yields segments that are optimal in some sense, but these tend to be very short and have questionable biological significance.

In the present paper we introduce Markov models for genomic segmentation. These are a family of change-point methods that are sensitive to higher-order correlations in DNA sequences, and that generalize an entropic segmentation technique [15,16] proposed earlier. Zeroth-order Markov segmentation is identical to this earlier proposed method [15,16] wherein the Jensen-Shannon (JS) divergence of nucleotide distributions in subsequences is maximized. This technique has found an application in the characterization of varied aspects of genome organization. Applications of this method have been made to analyze varied aspects of genome organization such as determination of regions that are homogenous with respect to GC/AT or purine/pyrimidine composition [9,15,16]. The characterization of isochores [17,18] in eukaryotic chromosomes, delineation of protein-coding and noncoding regions [19], finding the CpG islands, and the detection of replication origin and terminus in bacterial genomes [8] are among several other instances that have been addressed by this methodology.

Do mathematical or statistical methods for genome segmentation uncover significant biological features? This is a question that has been posed since the earliest such methods were introduced, and so far segmentation strategies have focused on characterizing homogeneity, and on generating optimal segments [9]. Entropic segmentation using a G+C-based measure has been successful in identifying CpG islands and isochores from a large set of segments [8]. Similarly, homogeneous regions obtained from a HMM-based segmentation were found to be functionally important features [12]. Clearly, the study of compositional homogeneity is more relevant if such segments can be shown (say via

annotation) to known biological features as can be done, say, for isochore determination [20]. Similarly, gene identification programs rely on describing coding sequences via compositional measures [21]. However, other aspects of genome composition are more subtle, and may require more sensitive probes. For instance, dinucleotide frequency distributions have been discussed by Karlin *et al.* [4] in the context of genomic phylogenetic distances, and this is something that cannot be discovered through simpler measures such as the G+C content or the individual nucleotide distributions.

Our motivation in devising the present Markov model for segmentation (MMS) is therefore to incorporate higher-order correlations. These are able to characterize better the inhomogeneities inherent in a given genomic sequence. The MMS method is presented in the next section, where we also discuss criteria for judging the statistical significance of the procedure within the model selection framework. In Sec. III, the MMS is compared with segmentation using standard JS divergence in two applications. We first apply the method to chimeric sequences, namely, those that are artificially constructed; the constituent parts of these chimeric sequences have distinct evolutionary histories, and since the segment boundaries are known *a priori*, it is possible to judge the accuracy of the procedure. We further apply these methods to a complete prokaryote genome, *E. coli* K12, to examine the biological relevance of the partition points. The paper concludes with a discussion and summary in Sec. IV.

II. MARKOV SEGMENTATION

Bernaola-Galvan *et al.* [15] proposed a recursive segmentation method that fragments a DNA sequence into homogeneous components (sometimes also termed "patches") in a top-down fashion. For a given sequence, it starts with locating the sequence position such that the adjacent subsequences are most distinct with respect to some predefined compositional measure, and which satisfy statistical significance. This process is repeated on the resulting two subsequences and so on until further segmentation of sequence segments is not statistically significant. A measure that has been used frequently in the past to judge this distinctiveness is the JS divergence [22], which is based on Shannon entropy and is a symmetrized generalization of the Kullback-Leibler divergence, another information-theoretic divergence measure.

Consider a symbolic sequence $\mathcal{S}$ of length N, constructed from an alphabet $\mathcal{A}$ of size κ,

$$\mathcal{S} \equiv \alpha_1 \alpha_2 \cdots \alpha_N, \tag{1}$$

where $\alpha \in \mathcal{A}$ and the subscript to α indicates the position in $\mathcal{S}$. We describe the sequence as deriving from a Markov chain of order m. The probability of this sequence in an mth-order Markov model is given by

$$P(\mathcal{S}) = P(\alpha_1, \ldots, \alpha_m) \prod_{i=m+1}^{N} P(\alpha_i | w = \alpha_{i-m} \alpha_{i-m+1} \cdots \alpha_{i-1}), \tag{2}$$

where $P(w)$ is the probability of occurrence of the word (or subsequence) w of length m followed by any symbol α and $P(\alpha_i|w)$ is the transition probability from the word w to the symbol α_i. Here we have estimated the initial probability $P(\alpha_1, \ldots, \alpha_m)$ from the corresponding marginal probability [23]. In applications to DNA sequence analysis, $\kappa=4$, and $\alpha \in \mathcal{A}=\{\mathrm{A},\mathrm{T},\mathrm{C},\mathrm{G}\}$.

The JS divergence between two subsequences $\mathcal{S}_1$ and $\mathcal{S}_2$ resulting from the binary segmentation of a DNA sequence S is given by [15,22]

$$D(\mathcal{S}_1, \mathcal{S}_2) = H(\mathcal{S}) - \pi_1 H(\mathcal{S}_1) - \pi_2 H(\mathcal{S}_2), \tag{3}$$

where n_1 and $n_2=N-n_1$ are the lengths of subsequences $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. π_1 and π_2 are weight factors summing to 1. For segmentation analysis, weights proportional to the length of the subsequences were found to be most appropriate, $\pi_i = \frac{n_i}{N}$. $H(\mathcal{S})$ defines the Shannon entropy, given by

$$H(\mathcal{S}) = -\sum_{\alpha} P(\alpha) \log_2 P(\alpha), \tag{4}$$

where $P(\alpha)$ denotes the probability of the nucleotide α. The maximum likelihood estimate of this parameter is simply $\hat{P}(\alpha)=C(\alpha)/N$, where C is the count of the nucleotide in the sequence. Implicit in the above measure of divergence is the assumption of the independence of occurrence of each nucleotide in $\mathcal{S}$.

The JS divergence measure can be easily generalized to account for the short-range interdependence of nucleotides. Considering the sequence to be generated by a Markov source of order m, the entropy function for the sequence is given by

$$H^m(\mathcal{S}) = -\sum_{w} \hat{P}(w) \sum_{\alpha} \hat{P}(\alpha|w) \log_2 \hat{P}(\alpha|w), \tag{5}$$

where the first summation is over all possible distinct m-mers, w. The estimates of the marginal probability $\hat{P}(w)$, and the transition probability $\hat{P}(\alpha|w)$ are obtained from the counts of the oligonucleotides: $\hat{P}(w)=C(w\cdot)/(N-m)$, $\hat{P}(\alpha|w)=C(w\alpha)/C(w\cdot)$, and $C(w\cdot)=\Sigma_\beta C(w\beta)$. It is easy to see that the entropy function can also be written as $H^m(\mathcal{S}) = -\Sigma_w \Sigma_\alpha \hat{P}(w\alpha) \log_2 \hat{P}(\alpha|w)$, which was used in computing the value of the entropy function.

The generalized JS divergence is thus given as

$$D^m(\mathcal{S}_1, \mathcal{S}_2) = H^m(\mathcal{S}) - \pi_1 H^m(\mathcal{S}_1) - \pi_2 H^m(\mathcal{S}_2). \tag{6}$$

Equation (6) reduces to Eq. (3) when $m=0$. The above expression for the JS divergence can be further generalized to consider partitioning into any number of subsequences. The procedure of segmentation involves the computation of JS divergence between all possible pairs of segmented subsequences, and the maximum over all partition points is $D_{\max}$. The sequence is segmented at this partition if and only if additional criteria such as statistical significance and minimal length are satisfied. A lower cutoff of 15 bp or more is typically applied so as to avoid obtaining numerous small segments of questionable significance. The former problem of judging statistical significance is a more serious issue, and two different criteria—the hypothesis testing and the model

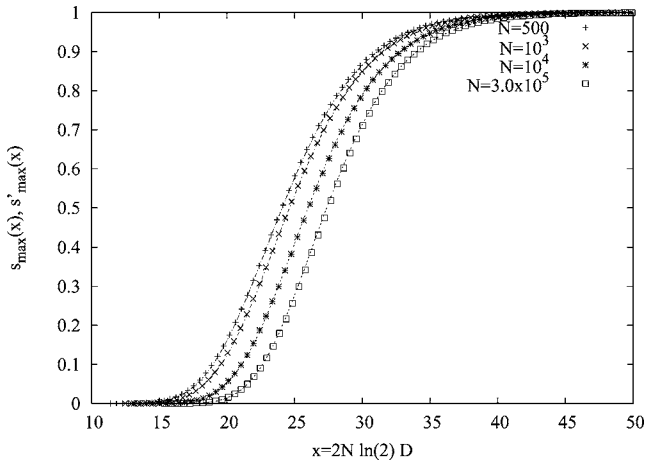

FIG. 1. Histogram, $s_{\max}(x)$, of $x=2N(\ln 2)D_{\max}$ and their finite-size approximations, $s'_{\max}(x)$, for the first-order Markov model.

selection—have been used. These criteria are described in the following subsections.

The binary segmentation is applied recursively. Starting with the sequence $\mathcal{S}$, one obtains subsequences $\mathcal{S}_1$ and $\mathcal{S}_2$, to each of which the segmentation is applied, and so on. The procedure is continued until further segmentation fails to be statistically significant by any of the applied criteria.

A. Hypothesis testing

In the hypothesis testing approach, the statistical significance $s_{\max}(x)$ of a binary segmentation is determined by the probability of obtaining a maximal divergence of $D_{\max}$ or less for random sequences of equivalent size, namely,

$$s_{\max}(x) = \text{Prob}\{D_{\max} \leqslant x\}. \tag{7}$$

For the zeroth-order model, Grosse *et al.* [24] made the ansatz

$$s_{\max}(x) = [F_\nu(\beta 2N(\ln 2)x)]^{N_e}, \tag{8}$$

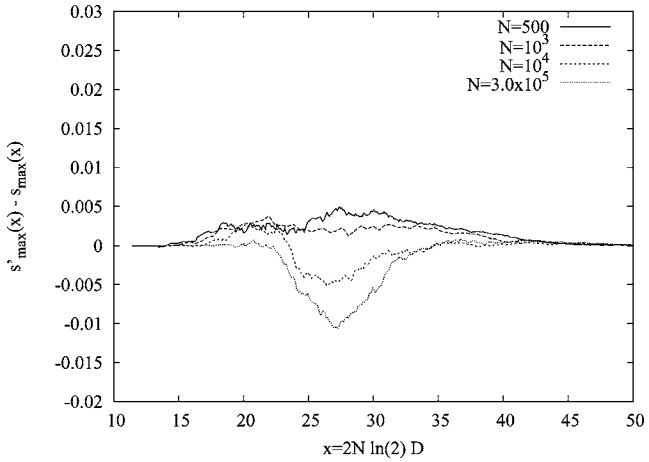

FIG. 2. Difference between $s'_{\max}(x)$ and $s_{\max}(x)$ for the first-order Markov model, which is a measure of the error associated with the approximation.

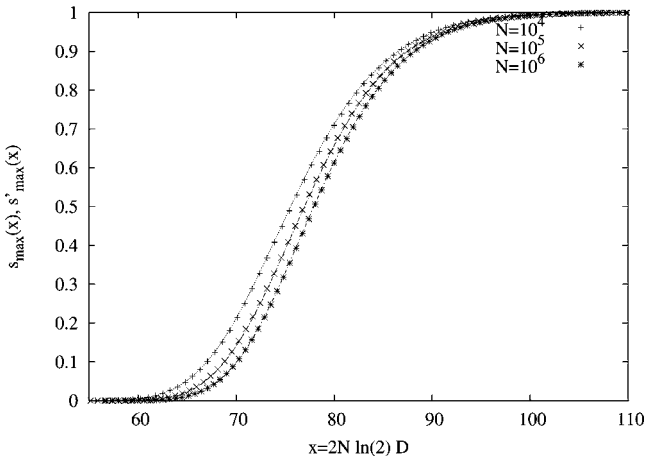

FIG. 3. As in Fig. 1 but for the second order.

where F_ν is the χ^2 distribution with $\nu=(k-1)$ degrees of freedom. β is the scaling factor independent of N and N_e is an effective length given by $N_e=a(\ln N)+b$. The parameters a, b, and β are obtained by fitting the theoretical distribution to the empirical distribution of $D_{\max}$ obtained through simulations.

For the higher-order Markov models, we implemented the Monte Carlo simulations suggested by Grosse *et al.* [24] to obtain an approximate analytic expression for the probability distribution of $D_{\max}$ for Markov sources. The functional form, obtained in the form of a chi-square distribution function with fitting parameters, was similar to that obtained by Grosse *et al.* [24],

$$s^{(m)}_{\max}(x) = [F_\nu(\beta_m 2N(\ln 2)x)]^{N^{(m)}_e}, \tag{9}$$

with $N^{(m)}_e=a_m(\ln N)+b_m$. In the present work, we take the number of degrees of freedom ν to be $4^{m+1}-1$. This assumes that there are 4^m-1 marginal probability parameters and $4^{m+1}-4^m$ transition probability parameters [25] (see also Billingsley [26], p. 14). Alternately, if we consider that we have an ergodic Markov chain process, which tends to converge to a solution irrespective of the choice of initial probability parameters, the number of degrees of freedom will be

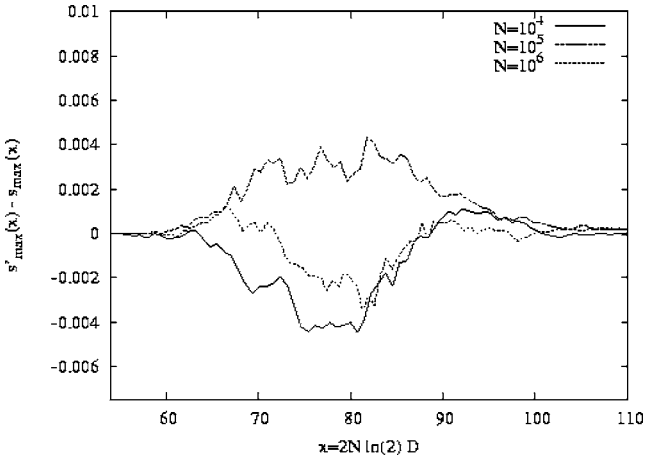

FIG. 4. As in Fig. 2 but for the second order.

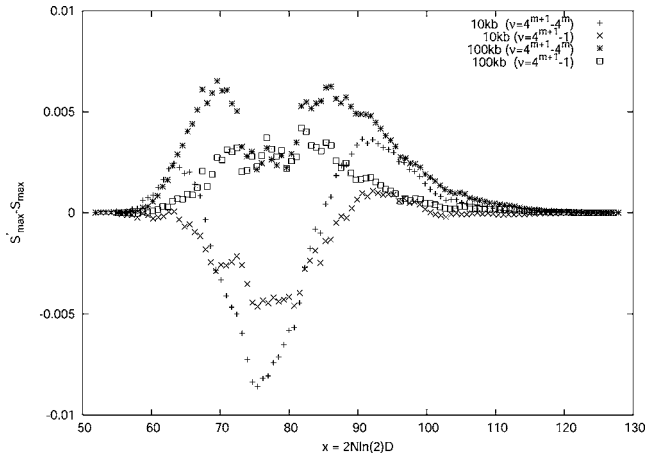

FIG. 5. Error in the numerical approximation of empirical distribution for different degrees of freedom in second-order MMS.

determined by the free transition probability parameters alone. Then we have $\nu=4^{m+1}-4^m$ [26–28]. With either choice for ν, the data can be fit to the empirical form, Eq. (8), with similar levels of accuracy.

We used Monte Carlo simulations wherein the lengths of the simulated sequences N ranged from 500 b to 1 Mb. Best-fit plots for the choice $\nu=4^{m+1}-1$ are shown in Figs. 1 and 3 for first- and second-order MMS, while the errors are shown in Figs. 2 and 4. For second order, a simulation of sequences smaller than 10 Kb was avoided to obtain a better set of truly random sequences. This gave an empirical distribution to which the appropriate form was fit. For second-order segmentation, β_m is observed to have an additional correction, which depends logarithmically to N, given by

$$\beta_m = c_m(\ln N) + d_m. \tag{10}$$

With $\nu=4^{m+1}-4^m$, we found the fitting of functional form to be essentially as good with comparable error (see Fig. 5). In practice though, this does not affect the actual segmentation results since the significance levels associated with the test statistic x will be invariably the same. The computations of this paper use $\nu=4^{m+1}-1$ although, as discussed above, the results would be identical with the alternate choice of ν. Tables I and II give the values of the fitting parameters estimated by the Monte Carlo procedure for both choices of ν.

B. Model selection framework

In the model selection framework [29] models of the DNA sequence before and after a putative binary segmentation are compared. Before the segmentation, a single-random-sequence model is used to describe the DNA sequence $\mathcal{S}$, and after the segmentation a two-random-sequence model is used to describe the resulting subsequences.

TABLE I. The values of parameters for the choice $\nu=4^{m+1}-1$ estimated from Monte Carlo simulations fit to the theoretical distribution of $D_{\max}$. m denotes the order of Markov model.

m	a_m	b_m	c_m	d_m
0	2.44	−6.15	0.0	0.80
1	1.557	−2.195	0.0	0.946
2	1.130	−2.447	0.0023	1.025

TABLE II. Same as in Table I but for $\nu=4^{m+1}-4^m$.

m	a_m	b_m	c_m	d_m
1	2.543	−4.77	0.0	0.848
2	2.39	−7.66	0.0029	0.841

Whenever the two-random-subsequence model is found to be superior [as determined by separate criterion (see below)] to a single-random-sequence model, segmentation is performed. The selection primarily is governed by two factors: the model's ability to fit the data and the complexity of the model. A balance is sought between these two factors to avoid both the overfitting or underfitting. Among the different criteria used for model selection, the Bayesian information criterion ($\mathcal{C}_{\mathrm{BIC}}$) [30–32] is defined by

$$\mathcal{C}_{\mathrm{BIC}} \approx -2\ln(\hat{L}) + K\ln N, \tag{11}$$

where $\hat{L}$ is the maximum likelihood, K is the number of parameters in the model, and N is the number of data points. In theory, a superior model has a larger intergrated likelihood and thus smaller value of $\mathcal{C}_{\mathrm{BIC}}$.

Li [29] has shown that the expression of JS divergence appears in obtaining the difference between $\mathcal{C}_{\mathrm{BIC}}$ of the candidate models. While this was done for the zeroth-order model ($m=0$), it can be easily generalized: considering the mth-order model, the likelihood of the sequence before segmentation is

$$\hat{L} = \hat{P}(\alpha_1 \cdots \alpha_m)\prod_w \prod_\alpha \hat{P}(\alpha|w)^{C(w\alpha)}$$
$$= \hat{P}(\alpha_1 \cdots \alpha_m)\prod_w \prod_\alpha \hat{P}(\alpha|w)^{(N-m)\hat{P}(w)\hat{P}(\alpha|w)}. \tag{12}$$

Taking logarithms, one obtains

$$\ln \hat{L} = \ln \hat{P}(\alpha_1 \cdots \alpha_m) + \sum_w \sum_\alpha (N-m)$$
$$\times \hat{P}(w)\hat{P}(\alpha|w)\ln \hat{P}(\alpha|w), \tag{13}$$

which further simplifies to

$$\ln \hat{L} = \ln \hat{P}(\alpha_1 \cdots \alpha_m) - (N-m)H^m(\mathcal{S}). \tag{14}$$

After segmentation, the likelihood for the model is the product of the likelihood of first subsequence $\hat{L}(1)$ and that of second subsequence $\hat{L}(2)$. As shown above one can similarly obtain $\ln \hat{L}(1)$ and $\ln \hat{L}(2)$ for the two subsequences. The *change* in the logarithm likelihood is

$$\Delta\hat{L} = \ln \hat{L}(1) + \ln \hat{L}(2) - \ln(\hat{L}). \tag{15}$$

It can be easily seen that asymptotically for large N,

TABLE III. Five genome pairs used to construct chimeric sequences. The difference in the GC% as well as the dinucleotide relative abundance $\delta*$ values are listed.

Pair	Genome *A*	Genome *B*	$\Delta(G+C)\%$	δ^*
I	*B. fragalis*	*A. marginale*	5.4	112.45
II	*B. subtilis*	*B. fragalis*	2.0	85.53
III	*Halobacterium sp.*	*D. radiodurans*	1.0	224.0
IV	*B. subtilis*	*P. aeruginosa*	25.0	94.0
V	*B. subtilis*	*M. tuberculosis*	23.6	134.0

$$\frac{\Delta\hat{L}}{N} = D^m(\mathcal{S}_1,\mathcal{S}_2). \tag{16}$$

For segmentation to be accepted, $\Delta\mathcal{C}_{\mathrm{BIC}}<0$, which leads to the condition

$$2ND^m(\mathcal{S}_1,\mathcal{S}_2) > (K_2-K_1)\ln N. \tag{17}$$

The parameters K_1 and K_2 for models before and after segmentation are $4^{m+1}-1$ and $2\times4^{m+1}-1$. Here, although the results obtained by use of $\Delta\mathcal{C}_{\mathrm{BIC}}$ are comparable [8,29], it should be noted that the use of $\Delta\mathcal{C}_{\mathrm{BIC}}$ as a criterion for deciding the change point has been questioned in earlier work [33].

III. APPLICATION AND ASSESSMENT OF MODELS

In the present case there is no existing benchmark against which we can standardize the results of MMS. Studies that have explored the association of biological features with homogeneous segments have been limited to coding (noncoding) boundaries, CpG islands, and isochores [8]. As a consequence, in this section we carry out an assessment of MMS using sets of specifically constructed heterogeneous sequences. These chimeric sequences are described in the next subsection. We also applied our methods to analyze a complete genome to uncovering a known set of biological features, and these results are described in Sec. III B.

A. Dataset of sequence constructs

Here we apply the above Markov models of segmentation to sequences of known heterogeneity to assess the accuracy and efficiency of the present procedure. The strategy for a quantitative assessment of segmentation is based on the ability of the method to detect *known* boundaries. For this purpose *chimeric* sequences are constructed from genomes of a set of distantly related organisms that are known to differ in their compositional organization. The dinucleotide relative abundance δ^* [34–36], which has been widely reported as a measure of the genomic signature in prokaryotes as well as eukaryotes, is taken as the discriminator: fragments of genomic DNA from closely related organisms have similar δ^*, in contrast to distantly related species [34–36]. In simple chimeras, we take two fragments of equal or unequal lengths. These are generated by the concatenation of a pair of subsequences c_A and c_B of sizes l_A and l_B from genomes *A* and *B*. We considered five pairs of prokaryotic genomes (see Table III). These range from nearly identical to very different GC content.

We generate ensembles of $n'=500$ chimeras from each pair of genomes $\{A,B\}$ with $l_A=l_B$. These are subjected to segmentation using Markov models of order $m=0,1,2$. The accuracy of segmentation is judged by how closely the actual boundaries are identified. The chimeric sequences C are of length L varying between 2 and 200 Kb. The *sensitivity* of the segmentation model is determined by the number of successes, i.e., the number of chimeras, n_p, segmented at the midpoint in the very *first step* of recursion with an allowed error d, i.e., $l'=l_A\pm d$. This is given by

$$S_N = \frac{n_p}{n'}, \tag{18}$$

and d was taken to be $0.05L$. Our results for segmentation using the hypothesis-testing-based criterion are summarized in Table IV.

TABLE IV. Sensitivity of the Markov segmentation of chimeras composed of segments of equal size using hypothesis testing. Pairs I–V are those listed in Table II.

	2 Kb			40 Kb			200 Kb		
Pair	Zeroth	First	Second	Zeroth	First	Second	Zeroth	First	Second
I	0.34	0.41	0.35	0.63	0.83	0.92	0.68	0.91	0.98
II	0.22	0.36	0.26	0.40	0.67	0.78	0.46	0.81	0.91
III	0.14	0.5	0.55	0.25	0.92	0.88	0.27	0.98	0.97
IV	0.87	0.88	0.88	0.98	0.97	0.98	0.99	0.98	0.98
V	0.92	0.88	0.86	0.99	0.99	0.99	1.0	1.0	1.0

TABLE V. As in Table IV, but using Bayesian information criterion for segmentation.

	2 Kb			40 Kb			200 Kb		
Pair	Zeroth	First	Second	Zeroth	First	Second	Zeroth	First	Second
I	0.34	0.09	0.0	0.59	0.85	0.85	0.73	0.92	0.97
II	0.24	0.04	0.0	0.48	0.69	0.5	0.44	0.82	0.9
III	0.16	0.12	0.0	0.2	0.92	0.93	0.3	0.98	0.98
IV	0.88	0.8	0.01	0.98	0.98	0.97	0.99	0.99	0.98
V	0.88	0.78	0.02	1.0	1.0	1.0	1.0	1.0	1.0

We find that in general the sensitivity increases with the order of the Markov model, and for sufficiently long sequences the second-order model is clearly the most sensitive. There is, however, a complicated dependence on the degree of relatedness of the genomes. Exceptionally high sensitivity is associated with even the zeroth-order model for pairs IV and V (Table IV), where the GC content differs by a large amount, suggesting that this is one of the important determinants of the performance of the model. The improvement in sensitivity with sequence length is subtle for higher-order models, but our prinicipal observation is that higher-order Markov models of segmentation are particularly successful in cases where the distantly related genomes have similar GC content. For sequences of length greater than 200 Kb the sensitivity values were over 0.9.

Sensitivity is also observed to rise with an increase in sequence size for all orders of segmentation model. Longer sequences allow for a better model construction (in terms of having sufficient statistics to estimate the parameters). In contrast, smaller chimeric sequences of length $\leqslant$40 Kb proved difficult to segment accurately. Segmentation of sequences within the model-selection framework gave comparable results: sensitivity improves with an increase in Markov order as well as sequence size (Table V).

The models were also assessed on mosaics with $l_A \neq l_B$, which corresponds to the naturally occurring scenario since segments are typically unequal in size. For fixed length of the smaller fragment and $L=l_A+l_B$, the performance of each of the methods improves with an increase in L. However, the performance also depends upon the absolute length of the smaller segment, l_A: so long as the smaller segment is longer than a threshold size, the performance improves with an increase in $|l_B-l_A|$ (see Tables VI and VII).

We further test Markov segmentation for complete recursive segmentation using *complex* chimeric sequences constructed as follows. Five selected genome sequences are segmented using zeroth-, first-, and second-order models as discussed in Sec. II. Those segments which are common to all orders of the segmentation are deemed homogeneous to all orders. Variable numbers of such homogeneous segments from these five genomes were randomly assembled into super sequences to construct two complex chimeric sequences, $\hat{C}_1$ and $\hat{C}_2$. These were then subject to recursive segmentation to examine whether the segment structure could be reconstructed. We measure the performance in terms of both sensitivity and specificity.

Results for chimeric sequence $\hat{C}_1$ are shown graphically in Figs. 6 and 7 at confidence levels 0.99 and 0.95, respectively, which demonstrate for the higher orders of MMS a better correspondence of partition points to the boundaries. A similar trend is observed for chimeric sequence $\hat{C}_2$, shown in Figs. 8 and 9. If n_{cp} is the number of correctly predicted boundaries, n_p is the total number predicted, and n_k is the number of boundaries actually present, then the sensitivity is

$$S_N = \frac{n_{cp}}{n_k}, \tag{19}$$

and the specificity is

$$S_P = \frac{n_{cp}}{n_p}. \tag{20}$$

The observed set of partition points suggests that each order of MMS performed fairly well in detecting the existing boundaries but varied in their accuracy of prediction. The zeroth order had very poor accuracy, particularly at a 0.95 confidence level, while the second-order model was much more accurate at both 0.99 and 0.95 confidence levels (Table VIII).

TABLE VI. Sensitivity of Markov segmentation of chimeras constructed from genomes with nearly identical GC composition. Indicated in the top row are l_A+l_B, with fixed l_A. Pair labels are those indicated in Table III.

	1 Kb+1 Kb			1 Kb+9 Kb			1 Kb+39 Kb		
Pair	Zeroth	First	Second	Zeroth	First	Second	Zeroth	First	Second
I	0.34	0.41	0.35	0.4	0.5	0.5	0.35	0.43	0.47
II	0.22	0.36	0.26	0.27	0.34	0.37	0.28	0.35	0.33

TABLE VII. As in Table VI, but for l_A=20 Kb.

	20 Kb+20 Kb			20 Kb+180 Kb		
Pair	Zeroth	First	Second	Zeroth	First	Second
I	0.63	0.83	0.92	0.89	0.98	1.0
II	0.4	0.67	0.78	0.81	0.95	0.97

B. Segmentation of a complete genome

When a complete genome is subject to Markov segmentation, even for a high significance level, there can be a very large number of segments. The second-order MMS gives, for example, over 1000 segments for the *E. coli* K12 genome. Investigation of the biological importance of each segment is infeasible as complete annotation of every genomic feature is not available to date.

Our strategy has been to verify whether segment boundaries correspond to the loci of known biological features, and thereby validate the segmentation procedure. Genomic regions that arise from evolutionary events—such as the insertion of foreign DNA—are among the most suitable candidates for examining the relation between homogeneity and biological features. Such horizontally transferred regions frequently differ in compositional measures from the surrounding "native" genome with variation in the codon usage bias, positional G+C composition, etc. [37]. Similarly, the duplications of a particular sequence involve insertions of the same at newer locations in the genome and thus are most likely to result in a marked change in the compositional measure at the boundaries. A number of such features have been classified, and beyond horizontal transfer, there can be so-called genomic islands, prophages, IS elements, etc. These may even have overlapping boundaries, and here we assess the performance of MMS applied to *E. coli* K12 in locating the boundaries of such features.

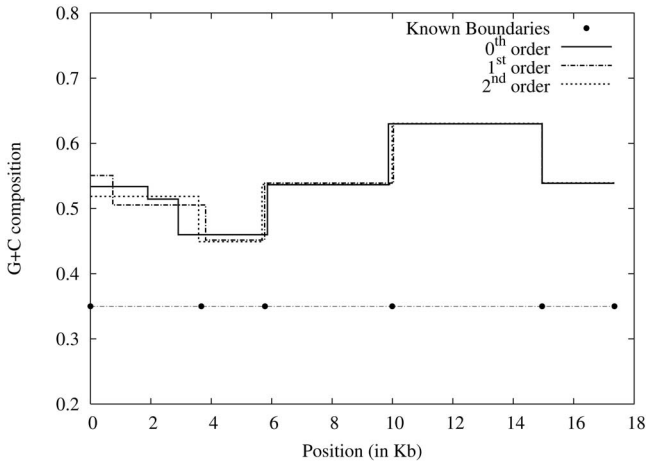

FIG. 6. Segments obtained from Markov segmentation of the complex chimeric sequence $\hat{C}_1$ at a 0.99 significance level. The sequences is comprised of five segments from five different genomes. The GC content of each segment is indicated. As can be seen, the second-order method gives essentially exact results.

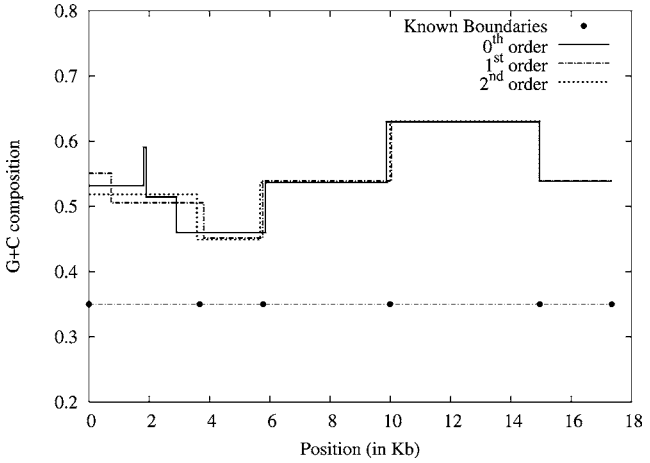

FIG. 7. As in Fig. 6 but for a 0.95 significance level. Again the second order is exact even at this level.

E. coli K12 [38] is arguably the most well-studied prokaryotic genome so far. As for any other organism, *E. coli* K12 genome evolution has been dynamic and several underlying structures that result from evolutionary events such as duplication, deletion, or insertion have been observed [39]. Quantitatively, the amount of acquired DNA is estimated to be about 12.8% [37] and there have been attempts to identify such elements, many of them appearing as large mosaic regions known as "genomic islands" [40–42]. Such genomic islands can be identified, for instance, via the program TRNACC [42]. To locate laterally acquired regions, the "enteric" server [43] was used to identify subsequences that are uniquely present in *E. coli* K12 and sometimes also in its closer relatives but absent in other species of *Enterobacteriaceae*. This genome is also known to contain many repetitive elements such as Rhs, REP, LDR, etc. and a set of repeating sequences of size ~5 Kb was identified through standard bioinformatics tools [44,45]. Coordinates of several examples of such features were obtained, and representatives from each category were chosen for this test.

The performance of MMS in identifying the specified boundaries was evaluated quantitatively; boundary predic-

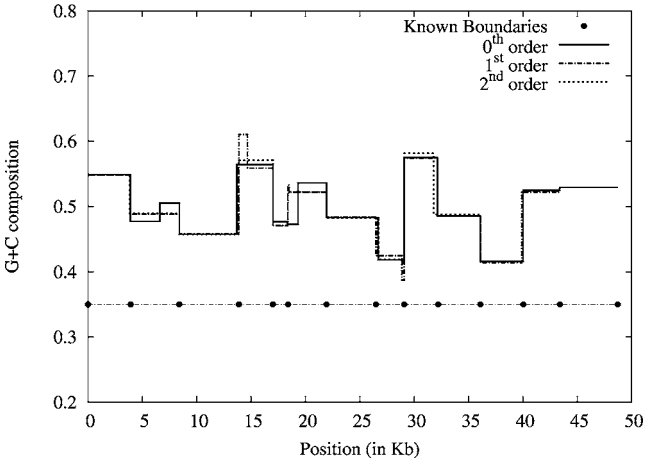

FIG. 8. As in Fig. 6 but for the complex chimeric sequence $\hat{C}_2$ comprised of 13 segments. The above pattern is observed here too.

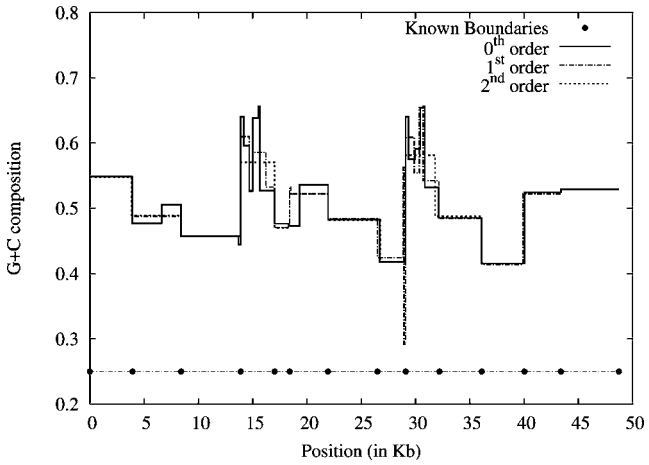

FIG. 9. As in Fig. 8, for a significance level of 0.95.

tion to within 400 bases was deemed true, and anything above that was termed a false prediction. [This threshold is itself estimated from the frequency distribution of errors in the prediction of change points for an ensemble of chimeric sequences of different size (data not shown here).] Results are presented in Table IX for a total of 12 features with 24 borders. The second-order Markov model locates 22 of the borders to reasonable accuracy, while the first-order model finds 20, and the zeroth-order model misses an additional three. We obtained an additional set of 40 boundaries of the putative horizontally transferred regions to test the statistical significance of our finding. We applied the sign test for large samples as follows. The entries in the error column (Table IX) were subtracted from the cutoff length (=400 bases) and for each model order, if this value was positive, it was considered a success, otherwise a failure. Let p be the probability of correctly detecting the features' borders by a method. The null hypothesis H_0 and the alternative hypothesis H_1 is set as—H_0: $p=0.5$, the success is merely due to chance; H_1: $p>0.5$, the success is indeed significant. Assuming the sampling distribution of the statistics to be a normal distribution with mean $\mu=Np$ and standard deviation $\sigma=\sqrt{Np(1-p)}$ ($N=64$ and $p=0.5$), the p values, which are defined as the largest significance level at which null hypothesis is rejected, for each of the model order were obtained. The p value for the zeroth-order model with 34 successes (out of 64) was 0.6497 while for the first- and second-order models with successes of 44 and 50, respectively, p values were >0.99. Considering the significance thresholds of 0.95 and 0.99, which are frequently used in arriving at a decision, the null hypothesis H_0 was rejected in the case of model orders 1 and 2 while this could not be rejected for model order 0. We thus infer that the higher-order segmentation method is indeed more effective in locating the features' borders compared to the conventional zeroth-order segmentation method.

As evident from the results, the segmentation methods in general are sensitive to the boundaries laid down by evolutionary events. The Markov model's performance, in particular, the second-order model, measured in terms of sensitivity, is somewhat better than the others. In specific cases, though, higher-order Markov models perform comparably or even worse than the most simple zeroth-order model; this might be a consequence of the inadequate representation of specific words due to the small size of subsequences flanking a particular change point. This was also evident from the simulation involving the segmentation of chimeric sequences. Further studies of the higher-order Markov models are currently underway.

IV. SUMMARY AND DISCUSSION

A number of segmentation methods have been developed in recent years with the aim of computationally dissecting a given genomic sequence into portions that correspond to specific structural and functional units. Different approaches to this general problem have given some insight into genome organization. The entropic segmentation method has considerably helped in uncovering the underlying structures of genomes [8]. By incorporating higher-order Markov models, the present work retains the simplicity of the JS divergence-based approach while allowing the basic model to be more sophisticated.

One of the drawbacks of the "1–2 segmentation" proposed by Bernaola-Galvan *et al.* [15] is that the boundaries obtained in the initial steps of the segmentation procedure are retained in the subsequent steps of recursive segmentation although they may no longer be significant at later stages where local heterogeneities are measured. Segmenting a given nonstationary DNA sequence into two "supposedly stationary" subsequences is at most an approximate approach as the subsequences may in fact be nonstationary and thus an estimation of the probability parameters from these subsequences is not completely a valid approach. However we believe that at later stages of recursive segmentation this effect is minimized. While obtaining an "optimal" segmenta-

TABLE VIII. Sensitivity (S_N) and specificity (S_P) of methods tested on mosaic constructs (see text) from *Anaplasma marginale* (GC=49.8), *Bacteroides fragalis NCTC 9343* (GC=44), *Escherichia coli* K12 (GC=50), *Thermotoga maritima MSB8* (GC=45), and *Treponema pallidum subsp. pallidum str. Nichols* (GC=52). The constructs are denoted by $\hat{C}_1$ and $\hat{C}_2$ and the S_0 denotes the threshold significance level.

	$\hat{C}_1$						$\hat{C}_2$					
S_0	0.99			0.95			0.99			0.95		
m	Zeroth	First	Second	Zeroth	First	Second	Zeroth	First	Second	Zeroth	First	Second
S_N	0.75	1.00	1.00	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S_P	0.60	0.80	1.00	0.50	0.80	1.00	0.86	0.80	0.92	0.57	0.60	0.92

TABLE IX. Segmentation at significance S_0=0.99 of the *E. coli* K12 genome to detect known genomic features. Columns 3–5 are the errors in prediction of the feature boundary. Those that meet the criterion of being below the specified threshold are shown in boldface.

Label	Position of known features	Error in prediction of feature boundary		
		Zeroth order	First order	Second order
Category: Putative horizontally acquired regions				
H1 (Start)	764372	**136**	**115**	**116**
H1 (End)	770608	**7**	**0**	**152**
H2 (Start)	2338419	2473	1072	**97**
H2 (End)	2342886	**13**	**44**	**12**
H3 (Start)	2478412	**282**	**217**	**366**
H3 (End)	2493555	1549	**131**	**31**
Category: Duplication				
D1 (Start)	223625	**396**	**202**	**189**
D1 (End)	228880	**106**	**115**	**106**
D2 (Start)	4033409	**363**	**243**	**371**
D2 (End)	4038664	**146**	**138**	**131**
D3 (Start)	c2729507	2332	429	**374**
D3 (End)	c2724086	**59**	**12**	**24**
Category: Genome islands				
aspV (Start)	237008	**36**	**131**	**3**
aspV (End)	239419	**127**	**77**	**55**
thrW (Start)	262171	2572	**89**	**90**
thrW (End)	302055	**246**	**247**	**248**
argU (Start)	564023	**286**	**90**	**14**
argU (End)	585323	957	**173**	1063
Category: Cryptic prophages				
CP4-6 (Start)	262182	583	**100**	**101**
CP4-6 (End)	296489	**100**	**118**	**26**
Pe-14 (Start)	1194346	**370**	**250**	**340**
Pe-14 (End)	1210646	1200	1200	4000
Rac (Start)	1409949	**27**	**356**	**353**
Rac (End)	1433008	**48**	828	**316**

tion of a nonstationary sequence is challenging, there have been some attempts in recent years to get an optimal compositional partitioning of DNA sequences using probabilistic models, mainly, the hidden Markov models (HMMs). The HMM-based methods aim to find the most likely path of hidden states that underlie a given DNA sequence using dynamic programming algorithms; while this approach is promising and has been successfully applied in gene identification, it has yet to significantly accomplish deciphering other regions of biological significance. The recently developed HMM-driven Bayesian method by Boys and Henderson [46] generates segments having genes in the same direction of transcription. Nicholas *et al.* [12] obtained similar results using their HMM approach. Another approach that generates an optimal segmentation using a dynamic programming method gives many very short sequences (sometimes of just one or two nucleotides) whose significance is often questionable [14]. Notwithstanding the drawbacks of the rather approximate approach of Bernaola-Galvan *et al.* [15], it has been extensively used in deciphering a number of functional or structural features in genome sequences. The use of dinucleotide or trinucleotide frequencies as a statistical determinant of sequence features makes it far more effective than the conventional approach as the application to sequence constructs as well as real genome sequence confirms.

A proper test of segmentation strategies is made difficult by the paucity of biological reference data with accurately characterized segmental structure. The few sequences that are known to be embedded with segments such as isochores in the major histocompatibility complex (MHC) sequence or experimentally confirmed CpG islands can be successfully analyzed with any order model since these features are compositionally fairly simple. We constructed chimeric sequences in order to demonstrate the basic methodology of Markov segmentation, and then applied this to the *E. coli* K12 genome to detect known features with biological significance. Higher-order Markov models provide sensitivity not only to overall base composition, but also to higher-order organizational aspects such as di-, tri-, or oligonucleotide usage. As in gene identification problems, there may be

a trade-off between model order and model sensitivity. In future work we hope to study these aspects of Markov segmentation.

ACKNOWLEDGMENT

This research was supported by the Department of BioTechnology and the University Grants Commission, New Delhi. We thank Pedro Bernaola-Galvan for helpful correspondence in the implementation of Monte Carlo simulations. We also thank Andrew Lynn, Dhwani Desai, Anchal Vishnoi, Alex Mitrophanov, and Shai Fine for useful discussions. R.R. would like to acknowledge the hospitality of the Institute for Advanced Study, Princeton, where this work was begun.

[1] GenBank (http://www.ncbi.nlm.nih.gov/).
[2] D. Graur and W.-H. Li, *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, 2000).
[3] T. A. Brown, *Genomes* (BIOS Scientific Publishers, Oxford, 2002).
[4] S. Karlin, A. M. Campbell, and J. Mrazek, Annu. Rev. Genet. **32**, 185 (1998).
[5] N. Sueoka, Proc. Natl. Acad. Sci. U.S.A. **48**, 582 (1962).
[6] R. A. Elton, J. Theor. Biol. **45**, 533 (1974).
[7] F. Tajima, J. Mol. Evol. **33**, 470 (1991).
[8] W. Li, P. Bernaola-Galvan, F. Haghighi, and I. Grosse, Comput. Chem. (Oxford) **26**, 491 (2002).
[9] J. V. Braun and H.-G. Müller, Stat. Sci. **13**, 2 (1998).
[10] G. A. Churchill, Bull. Math. Biol. **51**, 79 (1989).
[11] G. A. Churchill, Comput. Chem. (Oxford) **16**, 107 (1992).
[12] P. Nicolas *et al.*, Nucleic Acids Res. **30**, 1418 (2002).
[13] J. Liu and C. E. Lawrence, Bioinformatics **15**, 38 (1999).
[14] V. E. Ramensky, V. J. Makeev, M. A. Roytberg, and V. G. Tumanyan, J. Comput. Biol. **7**, 215 (2000).
[15] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).
[16] J. L. Oliver, R. Román-Roldán, Javier Pérez, and Pedro Bernaola-Galván, Bioinformatics **15**, 974 (1999).
[17] J. L. Oliver, P. Bernaola-Galván, P. Carpena, and R. Román Roldán, Gene **276**, 47 (2001).
[18] J. L. Oliver, P. Carpena, M. Hackenberg, and P. Bernaola-Galván, Nucleic Acids Res. **32**, W287 (2004).
[19] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).
[20] G. Cuny, P. Soriano, G. Macaya and G. Bernardi, Eur. J. Biochem. **115**, 227 (1981).
[21] M. Borodovsky and J. McIninch, Comput. Chem. (Oxford) **17**, 123 (1993).
[22] J. Lin, IEEE Trans. Inf. Theory **37**, 145 (1991).
[23] Note that an accurate estimation of marginal distributions requires sufficiently long stationary segments. In the initial stages of the algorithm, which we discuss in this section, the segments are long but are possibly nonstationary and may require further segmentation. However, as the segmentation proceeds, by the end of the procedure the segments are shorter and are more likely to be stationary.
[24] I. Grosse *et al.*, Phys. Rev. E **65**, 041905 (2002).
[25] I. Ben-Gal *et al.*, Bioinformatics **21**, 2657 (2005).
[26] P. Billingsley, Ann. Math. Stat. **32**, 12 (1961).
[27] P. Guttorp, *Stochastic Modeling of Scientific Data* (Chapman & Hall, London, 1995).
[28] Y. Ephraim and N. Merhav, IEEE Trans. Inf. Theory **48**, 1518 (2002).
[29] W. Li, Phys. Rev. Lett. **86**, 5815 (2001).
[30] H. Jeffreys, *Theory of Probability* (Clarendon Press, Oxford, 1961).
[31] G. Schwartz, Ann. Stat. **6**, 461 (1978).
[32] A. E. Raftery, in *Sociological Methodology*, edited by P. V. Marsden (Blackwell, Oxford, 1995), p. 185.
[33] R. E. Kass and A. E. Raftery, J. Am. Stat. Assoc. **90**, 773 (1995).
[34] S. Karlin and C. Burge, J. Comput. Biol. **11**, 283 (1995).
[35] S. Karlin and I. Ladunga, Proc. Natl. Acad. Sci. U.S.A. **91**, 12832 (1994).
[36] S. Karlin, J. Mrázek, and A. M. Campbell, J. Bacteriol. **179**, 3899 (1997).
[37] H. Ochman, J. G. Lawrence, and E. A. Groisman, Nature (London) **405**, 299 (2000).
[38] F. R. Blattner *et al.*, Science **277**, 1453 (1997).
[39] H. Ochman and I. B. Jones, EMBO J. **19**, 6637 (2000).
[40] Y. Mantri and K. P. Williams, Nucleic Acids Res. **32**, D55 (2004).
[41] W. Hsiao, I. Wan, S. J. Jones, and F. S. L. Brinkman, Bioinformatics **19**, 418 (2003).
[42] H.-Y Ou *et al.*, Nucleic Acids Res. **34**, e3 (2006).
[43] L. Florea *et al.*, Nucleic Acids Res. **31**, 13 (2003).
[44] S. Kurtz *et al.*, Adv. Genome Biol. **5**, R12 (2004).
[45] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, J. Mol. Biol. **215**, 403 (1990).
[46] R. J. Boys and D. A. Henderson, Biometrics **60**, 573 (2004).