Chapter 2

Markov Models of Segmentation

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. So far segmentation strategies have focused on characterizing homogeneity, and on generating optimal segments [118]. Entropic segmentation using a G+C based measure has been successful in identifying CpG islands and isochores from a large set of segments [84]. Similarly, homogeneous regions obtained from a HMM based segmentation were found to contain functionally important features [96]. Clearly, the study of compositional homogeneity is more relevant if such segments can be shown (say via annotation) to correspond to known biological features as can be done for isochore determination [35]. Similarly, gene identification programs rely on describing coding sequences via compositional measures [18], so in a sense these also segment genomes.

Other aspects of genome composition are however more subtle, and require more sensitive probes. For instance, dinucleotide frequency distributions have been discussed by Karlin et al. [73] 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. Secondly, the method of entropic segmentation has been reported to have limitations with identification of certain features that are discussed here. The isochore identification in human genome when tested using the approach of entropic segmentation indicated that only 41% of human genome is covered by isochores compared to earlier findings [31]. This is lower than proposed, and secondly, a four family isochore model (rather than five) fitted the empirical data. Limitation of the entropic segmentation method was reported
behind such inconsistencies [30]. Since simple entropic segmentation systematically loses GC-rich isochores because it fails to anticipate the considerable fluctuations that are highest in GC-rich DNA. In addition, a fraction of the complete set of isochores reported by Bernardi et al. [34] in all human chromosomes are poorly characterized. Splitting chromosomes into adjoining sets of isochores result in several of them having high heterogeneity as against the definitions set out for typical isochores (see section 4.2). Simple entropic segmentation has also been applied to the splice site recognition problem [10]. Here one incorporates coding position information to view DNA as composed of a 12 letter alphabet, but this method has not emerged as a reliable tool for border identification in prokaryotic genomes.

In the present chapter we introduce Markov models for genome segmentation. Our motivation in devising the present Markov model for segmentation (MMS) is to be able to incorporate higher order correlations which are better able to characterize inhomogeneities inherent in a given genomic sequence. The MMS method is presented in the next section, where we first describe the algorithm of zeroth order segmentation followed by the criteria for judging the statistical significance of the procedure within the model selection framework and hypothesis testing. In Section 2.2, the MMS is compared with zeroth-order model using a set of heterogeneous sequences. The method is applied 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. Furthermore, we also test the segment boundary sensitivity to fluctuations in sequence size.

2.1 Markov segmentation

Bernaola-Galvan et al. [11] 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, the method starts with locating the sequence position such that the adjacent subsequences are most distinct with respect to some predefined compositional measure, to a required degree of statistical significance. This process is repeated on the two resulting 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 Jensen-Shannon divergence [87] which is based on the Shannon
entropy and is a symmetric generalization of the Kullback-Leibler divergence (which is also
an useful information-theoretic divergence measure).

Consider a symbolic sequence $S$ of length $N$, constructed from an alphabet $A$ of size $\kappa$,

$$S \equiv \alpha_1 \alpha_2 \ldots \alpha_N,$$

where $\alpha \in A$, the subscript on $\alpha$ indicating the position in $S$. We describe the sequence as
deriving from a Markov chain of order $m$. The probability of this sequence in a $m$th order
Markov model is given by

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

where $P(w)$ is the probability of occurrence of the word (or subsequence) $w$ of length $m$
followed by any symbol $\alpha$ and $P(\alpha_i | w)$ is the transition probability from the word $w$ to the
symbol $\alpha_i$. Here we have estimated the initial probability $P(\alpha_1, \ldots, \alpha_m)$ from the corre­
sponding marginal probability [140]. In applications to DNA sequence analysis, $\kappa = 4$, and
$\alpha \in A = \{A, T, C, G\}$.

The Jensen-Shannon divergence between two subsequences, $S_1$ and $S_2$ that result from
the binary segmentation of a DNA sequence $S$ is given by [87, 11]

$$D(S_1, S_2) = H(S) - \pi_1 H(S_1) - \pi_2 H(S_2),$$

where $n_1$ and $n_2 = N - n_1$ are the lengths of subsequences, $S_1$ and $S_2$ respectively. $\pi_1$ and
$\pi_2$ are weight factors summing to 1. For segmentation analysis, weights proportional to the
length of the subsequences have been found to be most appropriate, $\pi_i = \frac{n_i}{N}$. $H(S)$ is the
Shannon entropy, given by

$$H(S) = -\sum_\alpha P(\alpha) \log_2 P(\alpha),$$

where $P(\alpha)$ denotes the probability of the nucleotide $\alpha$ and logarithms are taken in base 2.
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 $S$. Eq.
(2.3) defines the divergence in the zeroth order model.
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

\[
H^m(S) = - \sum_w \hat{P}(w) \sum_\alpha \hat{P}(\alpha|w) \log_2 \hat{P}(\alpha|w),
\]

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:

\[
\hat{P}(w) = \frac{C(w)}{N-m},
\]

\[
\hat{P}(\alpha|w) = \frac{C(w\alpha)}{C(w)}, \text{ and}
\]

\[
C(w) = \sum_\beta C(w\beta),
\]

and the entropy function itself can be also be written as

\[
H^m(S) = - \sum_w \sum_\alpha \hat{P}(w\alpha) \log_2 \hat{P}(\alpha|w).
\]

It is simpler to use this latter expression in computation.

The generalized JS divergence is thus given as

\[
D^m(S_1, S_2) = H^m(S) - \pi_1 H^m(S_1) - \pi_2 H^m(S_2).
\]

Eq. (2.11) reduces to Eq. (2.3) 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 computation of JS divergence between all possible pairs of segmented subsequences, and the maximum over all partition points is \( D_{\text{max}} \). The sequence is segmented at this partition if and only if additional criteria such as statistical significance and minimal length are satisfied. A length 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—hypothesis testing and the model selection—have been used. These criteria are described in the following subsections.

The binary segmentation is applied recursively. Starting with the sequence \( S \), one obtains subsequences \( S_1 \) and \( S_2 \), 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.
2.1.1 Hypothesis Testing

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

$$s_{\text{max}}(x) = \text{Prob}\{D_{\text{max}} \leq x\}.$$  \hspace{1cm} (2.12)

For the zeroth order model, Grosse et al. [65] made the ansatz

$$s_{\text{max}}(x) = [F_{\nu}(\beta 2N(\ln 2)x)]^{N_{\text{e}}},$$  \hspace{1cm} (2.13)

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

For the higher order Markov models, we implemented the Monte-Carlo simulations suggested by Grosse et al. [65] to obtain an approximate analytic expression for the probability distribution of $D_{\text{max}}$ for Markov sources. The functional form, obtained in the form of chi-square distribution function with fitting parameters, was similar to that obtained by Grosse et al. [65],

$$s_{\text{max}}^{(m)}(x) = [F_{\nu}(\beta_{m} 2N(\ln 2)x)]^{N_{\text{e}}^{(m)}},$$  \hspace{1cm} (2.14)

with $N_{\text{e}}^{(m)} = a_{m}(\ln N) + b_{m}$. In the present work, we take the number of degrees of freedom, $\nu$ to be $4^{m+1} - 1$, which assumes that there are $4^m - 1$ marginal probability parameters and $4^{m+1} - 4^{m}$ transition probability parameters [9] (see also Billingsley [14], 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 determined by the free transition probability parameters alone. Then we have $\nu = 4^{m+1} - 4^{m}$ [14, 66, 45]. With either choice for $\nu$, the data can be fit to the empirical form, Eq. 2.13 with similar levels of accuracy.

In order to determine the parameters we used Monte-Carlo simulations wherein the lengths of the simulated sequences $N$ ranged from 500b to 1 Mb. Best-fit plots for the choice $\nu = 4^{m+1} - 1$ are shown in Figures 2.1 and 2.3 for 1st and 2nd order MMS, while the errors are shown in Figures 2.2 and 2.4. For 2nd order, 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 $2^{nd}$ order segmentation, $\beta_m$ is observed to have an additional correction which depends logarithmically on $N$, given by

$$\beta_m = c_m (\ln N) + d_m.$$  \hfill (2.15)

With $\nu = 4^{m+1} - 4^m$, we found the fitting of functional form to be essentially as good with comparable error (see Fig. (2.5). In practice this does not affect the actual segmentation results since the significance levels associated with the test statistic $x$ will be invariably same. The computations here use $\nu = 4^{m+1} - 1$ although, as discussed above, the results would be identical with the alternate choice of $\nu$. Tables 2.1, 2.2 give the values of the fitting parameters estimated by the Monte Carlo procedure for both choices of $\nu$.

Table 2.1: 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.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$a_m$</th>
<th>$b_m$</th>
<th>$c_m$</th>
<th>$d_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.44</td>
<td>-6.15</td>
<td>0.0</td>
<td>0.80</td>
</tr>
<tr>
<td>1</td>
<td>1.557</td>
<td>-2.195</td>
<td>0.0</td>
<td>0.946</td>
</tr>
<tr>
<td>2</td>
<td>1.130</td>
<td>-2.447</td>
<td>0.0023</td>
<td>1.025</td>
</tr>
</tbody>
</table>

Table 2.2: The values of parameters for the choice $\nu = 4^{m+1} - 4^m$ estimated from Monte-Carlo simulations fit to the theoretical distribution of $D_{max}$. $m$ denotes the order of Markov model.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$a_m$</th>
<th>$b_m$</th>
<th>$c_m$</th>
<th>$d_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.543</td>
<td>-4.77</td>
<td>0.0</td>
<td>0.848</td>
</tr>
<tr>
<td>2</td>
<td>2.39</td>
<td>-7.66</td>
<td>0.0029</td>
<td>0.841</td>
</tr>
</tbody>
</table>
Figure 2.1: Cumulative distribution, $s_{\text{max}}(x)$, of $x = 2N(\ln 2)D_{\text{max}}$ and their finite size approximations, $s'_{\text{max}}(x)$, for the first order Markov model.

Figure 2.2: Difference between $s'_{\text{max}}(x)$ and $s_{\text{max}}(x)$ for the first order Markov model, which is a measure of the error associated with the approximation, Eq. 2.14.
Figure 2.3: Cumulative distribution, $s_{\text{max}}(x)$, of $x = 2N(\ln 2)D_{\text{max}}$ and their finite size approximations, $s'_{\text{max}}(x)$, for the second order Markov model.

Figure 2.4: Difference between $s'_{\text{max}}(x)$ and $s_{\text{max}}(x)$ for the second order Markov model.
2.1.2 Model Selection Framework

In the model selection framework [85], models of the DNA sequence before and after a putative binary segmentation are compared. Prior to segmentation, a single random-sequence model is used to describe the DNA sequence \( S \), and after the segmentation a two random-sequence model is used to describe the resulting subsequences.

Whenever the two-random-subsequence model is found to be superior (as determined by a separate criterion; see below) to a single-random-sequence model, segmentation is performed. The selection is primarily governed by two factors: the ability of the model to fit the data, and the complexity of the model itself. A balance is sought between these two factors to avoid both overfitting or underfitting. Among the different criteria used for model selection, the Bayesian Information Criterion (BIC) [71, 120, 110] is defined by

\[
BIC \approx -2 \log(\hat{L}) + K \log N \tag{2.16}
\]

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 integrated likelihood, and thus a smaller value of BIC.

Figure 2.5: Error in numerical approximation of empirical distribution for different degrees of freedom in second order MMS.
Li [85] has shown that the expression of JS divergence appears in obtaining the difference between BIC of the candidate models. While this was done for zeroth order model \((m = 0)\), it can be easily generalized: considering \(m^{th}\) order model, the likelihood of the sequence before segmentation is

\[
\hat{L} = \hat{P}(\alpha_1 \ldots \alpha_m) \prod_w \prod_\alpha \hat{P}(\alpha|w)^{C(\alpha w)}
\]

\[
= \hat{P}(\alpha_1 \ldots \alpha_m) \prod_w \prod_\alpha \hat{P}(\alpha|w)^{(N-m)\hat{P}(w)\hat{P}(\alpha|w)}
\]

Taking logarithms, one obtains

\[
\log \hat{L} = \log \hat{P}(\alpha_1 \ldots \alpha_m) + \sum_w \sum_\alpha (N - m)
\times \hat{P}(w)\hat{P}(\alpha|w) \log \hat{P}(\alpha|w),
\]

which further simplifies to

\[
\log \hat{L} = \log \hat{P}(\alpha_1 \ldots \alpha_m) - (N - m)H^m(S).
\]

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 \(\log \hat{L}(1)\) and \(\log \hat{L}(2)\) for the two subsequences. The change in the log–likelihood is

\[
\Delta \hat{L} = \log \hat{L}(1) + \log \hat{L}(2) - \log(\hat{L}).
\]

It can be easily seen that asymptotically for large \(N\),

\[
\frac{\Delta \hat{L}}{N} = D^m(S_1, S_2).
\]

For segmentation to be accepted, \(\Delta BIC < 0\), which leads to the condition,

\[
2ND^m(S_1, S_2) > (K_2 - K_1) \log N.
\]

The parameters \(K_1\) and \(K_2\) for models before and after segmentation are \(4^{m+1} - 1\) and \(2 \cdot 4^{m+1} - 1\). Here, although the results obtained by use of \(\Delta BIC\) are comparable to those obtained using hypothesis testing [85, 84], it should be noted that the use of \(\Delta BIC\) as a criterion for deciding the change point is itself somewhat controversial [78] and thus requires caution.
2.2 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 [84] using zeroth order segmentation. As a consequence in this section we carry out an assessment of MMS using sets of specifically constructed heterogeneous sequences as described below.

2.2.1 Dataset of sequence constructs

We apply the above Markov models of segmentation to sequences of known heterogeneity in order to assess the accuracy and efficiency of the 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 which are known to differ in their compositional organization. The dinucleotide relative abundance $\delta^*$ [75, 76, 77] 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 $\delta^*$, in contrast to distantly related species [75, 76, 77]. In the simplest case, we take two fragments of equal or unequal lengths. These are generated by concatenation of 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 2.3). 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 Kb to 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

$$SN = \frac{n_p}{n'},$$

and $d$ was taken to be $0.05L$. Our results for segmentation using the hypothesis testing based criterion are summarized in Table 2.4.
Table 2.3: Five genome pairs used to construct chimeric sequences. The difference in the GC% as well as the dinucleotide relative abundance δ* values are listed.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Genome A</th>
<th>Genome B</th>
<th>Δ(G+C)%</th>
<th>δ*</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>B. fragilis</td>
<td>A. marginale</td>
<td>5.4</td>
<td>112.45</td>
</tr>
<tr>
<td>II</td>
<td>B. subtilis</td>
<td>B. fragilis</td>
<td>2.0</td>
<td>85.53</td>
</tr>
<tr>
<td>III</td>
<td>Halobacterium sp.</td>
<td>D. radiodurans</td>
<td>1.0</td>
<td>224.0</td>
</tr>
<tr>
<td>IV</td>
<td>B. subtilis</td>
<td>P. aeruginosa</td>
<td>25.0</td>
<td>94.0</td>
</tr>
<tr>
<td>V</td>
<td>B. subtilis</td>
<td>M. tuberculli</td>
<td>23.6</td>
<td>134.0</td>
</tr>
</tbody>
</table>

We find that in general the sensitivity increases with the order of the Markov model, and for sufficiently long sequences the 2nd 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 2.4), where the GC content differs by a large amount, suggesting that this is one of the important determinant of the performance of the model. The improvement in sensitivity with sequence length is subtle for higher order models, but our principal 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 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 ≤ 40 Kb proved difficult to segment accurately. Segmentation of sequences within the model selection framework gave comparable results: sensitivity improves with increase in Markov order as well as sequence size (Table 2.5).

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 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
Table 2.4: 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.

<table>
<thead>
<tr>
<th>Pair</th>
<th>2Kb</th>
<th>40Kb</th>
<th>200Kb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0&lt;sup&gt;th&lt;/sup&gt;</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
</tr>
<tr>
<td>I</td>
<td>0.34</td>
<td>0.41</td>
<td>0.35</td>
</tr>
<tr>
<td>II</td>
<td>0.22</td>
<td>0.36</td>
<td>0.26</td>
</tr>
<tr>
<td>III</td>
<td>0.14</td>
<td>0.5</td>
<td>0.55</td>
</tr>
<tr>
<td>IV</td>
<td>0.87</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>V</td>
<td>0.92</td>
<td>0.88</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 2.5: Sensitivity of the Markov segmentation of chimeras composed of segments of equal size using Bayesian Information criterion.

<table>
<thead>
<tr>
<th>Pair</th>
<th>2Kb</th>
<th>40Kb</th>
<th>200Kb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0&lt;sup&gt;th&lt;/sup&gt;</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
</tr>
<tr>
<td>I</td>
<td>0.34</td>
<td>0.09</td>
<td>0.0</td>
</tr>
<tr>
<td>II</td>
<td>0.24</td>
<td>0.04</td>
<td>0.0</td>
</tr>
<tr>
<td>III</td>
<td>0.16</td>
<td>0.12</td>
<td>0.0</td>
</tr>
<tr>
<td>IV</td>
<td>0.88</td>
<td>0.8</td>
<td>0.01</td>
</tr>
<tr>
<td>V</td>
<td>0.88</td>
<td>0.78</td>
<td>0.02</td>
</tr>
</tbody>
</table>

performance improves with increase in $|l_B - l_A|$ (see Tables 2.6 and 2.7).

We further test Markov segmentation for complete recursive segmentation using complex chimeric sequences constructed as follows. Five selected genome sequences are segmented using 0<sup>th</sup>, 1<sup>st</sup> and 2<sup>nd</sup> order models as discussed in Section 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.
Table 2.6: 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 2.3.

<table>
<thead>
<tr>
<th>Pair</th>
<th>1Kb +1Kb</th>
<th>1Kb +9Kb</th>
<th>1Kb +39Kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>0th</td>
<td>0.34 0.41 0.35</td>
<td>0.4 0.5 0.5</td>
<td>0.35 0.43 0.47</td>
</tr>
<tr>
<td>1st</td>
<td>0.22 0.36 0.26</td>
<td>0.27 0.34 0.37</td>
<td>0.28 0.35 0.33</td>
</tr>
</tbody>
</table>

Table 2.7: 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 \( l_A = 20 \) Kb.

<table>
<thead>
<tr>
<th>Pair</th>
<th>20Kb +20Kb</th>
<th>20Kb +180Kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>0th</td>
<td>0.63 0.83 0.92</td>
<td>0.89 0.98 1.0</td>
</tr>
<tr>
<td>1st</td>
<td>0.4 0.67 0.78</td>
<td>0.81 0.95 0.97</td>
</tr>
</tbody>
</table>

Results for chimeric sequence \( \hat{C}_1 \) is shown graphically in Fig. 2.6 and 2.7 at confidence levels 0.99 and 0.95 respectively, which demonstrates 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 Fig. 2.8 and 2.9.

If \( n_{cp} \) is the number of correctly predicted boundaries, \( n_p \) the total number predicted, and \( n_k \) the number of boundaries actually present, then the sensitivity is

\[
SN = \frac{n_{cp}}{n_k}.
\]  \hfill (2.24)

and the specificity is

\[
SP = \frac{n_{cp}}{n_p}.
\]  \hfill (2.25)

The observed set of partition points suggests that all three orders of MMS tested here performed fairly well in detecting the existing boundaries but varied in their accuracy of prediction. The zeroth order had very poor accuracy, particularly at 0.95 confidence level,
Figure 2.6: Segments obtained from Markov segmentation of the complex chimeric sequence \( \hat{C}_1 \) at 0.99 significance level. The sequence comprises of 5 segments from five different genomes. The GC content of each segment is indicated on the ordinate. As can be seen the 2\(^{nd}\) order method gives essentially exact results.

Figure 2.7: Segments obtained from Markov segmentation of the complex chimeric sequence \( \hat{C}_1 \) at 0.95 significance level. The sequence comprises of 5 segments from five different genomes, and the GC content of each segment is indicated on the ordinate. Again the second order is exact even at this level.
Figure 2.8: Segments obtained from Markov segmentation of the complex chimeric sequence \( \hat{C}_2 \) at 0.99 significance level. The sequence comprises of 13 segments from five different genomes. The above pattern is observed here too, namely that second order MMS is exact.

Figure 2.9: Segments obtained from Markov segmentation of the complex chimeric sequence \( \hat{C}_2 \) at 0.95 significance level. The sequence comprises of 13 segments from five different genomes.
while the 2nd order model was more accurate at both 0.99 and 0.95 confidence levels. (Table 2.8).

Table 2.8: Sensitivity (SN) and specificity (SP) of methods tested on mosaic constructs (see text) from *Anaplasma marginale* (GC=49.8%), *Bacteroides fragilis 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.

<table>
<thead>
<tr>
<th></th>
<th>( S_0 )</th>
<th>0.99</th>
<th>0.95</th>
<th>0.99</th>
<th>0.95</th>
<th>0.99</th>
<th>0.95</th>
<th>0.99</th>
<th>0.95</th>
<th>0.99</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( m )</td>
<td>( 0^{th} )</td>
<td>( 1^{st} )</td>
<td>( 2^{nd} )</td>
<td>( 0^{th} )</td>
<td>( 1^{st} )</td>
<td>( 2^{nd} )</td>
<td>( 0^{th} )</td>
<td>( 1^{st} )</td>
<td>( 2^{nd} )</td>
<td>( 0^{th} )</td>
</tr>
<tr>
<td>( S_N )</td>
<td>0.75</td>
<td>1.00</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( S_P )</td>
<td>0.60</td>
<td>0.80</td>
<td>1.00</td>
<td>0.50</td>
<td>0.80</td>
<td>1.00</td>
<td>0.86</td>
<td>0.80</td>
<td>0.92</td>
<td>0.57</td>
<td>0.60</td>
</tr>
</tbody>
</table>

### 2.2.2 Boundary Sensitivity

In addition to validating the Markov segmentation, another aspect which needs consideration is the robustness of the segment boundaries to variation in the overall size of a sequence or an intermediate segment. In recursive segmentation, the boundary at any step is determined by maximization of the JS divergence values obtained from all possible segmentation points since the position of a boundary primarily depends upon the distribution of nucleotides. In order that the boundaries are reliable they should be insensitive to the size of input sequences, at least within certain limits. Here we carry out an empirical analysis to gauge this reliability.

Test sequences of length few Mbs in size from human genome were subjected to Markov segmentation. Three representative sequences were selected as follows: one sequence had a single maximum in the JS divergence, the second had two of almost similar height, and the third had more than two comparable maxima. A set of subsequences were constructed by trimming each of the three original sequences from either end or both ends keeping their first boundary intact. For example, the first sequence was of size \( \approx 2.35 \) Mb with its first boundary located at 0.999 Mb, generating a pair of segments denoted by their respective
size in the format [0.999 Mb & 1.35 Mb]. If the segment on the right is trimmed to a size of 0.1Mb, the new sequence should be comprised of segments of size 0.999 Mb and 0.1 Mb. Similarly several subsequences were generated (Figure 2.10) and were segmented using second-order Markov segmentation till the first recursion and the location of new maxima was compared with the original ones.

As may be expected, sequences showed some variation in their sensitivity to the original boundary with change in size. Results for the sequences with a single peak in the JS divergence were most robust to the shortening of sequence; new boundaries emerged when one of the segments was shortened several times of its original size (Figure 2.10). On the other hand for the sequence having multiple peaks in the JS divergence, as the size of the paired segments became more skewed, newer boundaries were observed which actually coincided with the boundaries from the subsequent step(s) of segmentation. This shows that a change in sequence length has the propensity to affect the boundary in sequences with multiple maxima. However, the newer boundaries when formed will in most cases actually converge to one of the maxima from subsequent steps.

2.3 Discussion

One of the drawbacks of the ‘1 to 2 segmentation’ proposed by Bernaola-Galvan et al. [11] 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 non-stationary DNA sequence into two supposedly stationary subsequences is at most an approximate approach as the subsequences may in fact be non-stationary and thus 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’ segmentation of a non-stationary sequence is challenging, there have been some attempts in recent years to get an optimal compositional partitioning of DNA sequences using probabilistic models, mainly 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 is yet to accomplish significantly in deciphering other re-
Figure 2.10: Effect of variation of sequence size on the first boundary. Four different sub­sequences of a 2.349Mb long sequence of human genome (solid line in figure (a)) were generated by trimming the distal ends of segments on left and/or right of the boundary (indicated by dashed portion), displayed schematically on the top of each figure. The vertical bar denotes the first boundary at 0.999Mb obtained by second order Markov segmentation. (a) The com­plete sequence has single JS divergence maxima at 0.999Mb generating a pair of segments [0.999Mb & 1.35Mb]. The JS divergence peaks observed in the second recursion have been indicated by horizontal arrows. Each subsequence was subject to second order Markov seg­mentation once. In the remaining figures (b-e) the JS divergence of sub-sequences [0.999Mb & 0.999Mb], [0.999Mb & 0.1Mb], [0.5Mb & 1.35Mb], and [0.0675Mb & 1.35Mb] has been plotted w.r.t. sequence position (in Mb).

regions of biological significance. Recently developed HMM driven Bayesian method by Boys and Henderson [19] generates segments having genes in the same direction of transcription.
Nicolas et al. [96] 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 [111]. Notwithstanding the drawbacks of the rather approximate approach of Bernaola-Galvan et al. [11], 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 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 which 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 analysed 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 we apply this to the genome sequence to detect known features with biological significance. This will be discussed in the next Chapter. Higher order Markov models provide sensitivity not only to overall base composition, but also to higher order organizational aspects such as di-, tri-, or oligo-nucleotide usage. As in gene identification problems, there may be a trade-off between model order and model sensitivity, and in future work we hope to study these aspects of Markov segmentation.