Chapter 5

Segmentation of Genomic DNA

5.1 Introduction

As discussed in the previous Chapter, genomic DNA can be seen as a symbolic string in a 4-letter alphabet. Different parts of the genome have specific function and specific history, and since they are the result of evolution from a single ancestor, they thus bear the imprint of the processes and the times at which they were formed. Genomic DNA is patchy, namely, the composition and properties vary in different segments or patches along the DNA. How this organizational structure came about is a matter of speculation, but there is evidence that DNA probably also evolved in fits and starts, namely opportunistically.

DNA sequences are very heterogeneous [75, 104, 105, 131]. This is a consequence both of function—different parts of the DNA have different biological actions—and history—different parts of a genomic DNA evolved at different times in different environments. Thus the base composition of DNA is nonuniform along the chain with both biochemical and physical consequences; nucleotide densities and purine–pyrimidine (A or G and C or T) ratio etc. differ significantly in different portions of a DNA sequence.

With the genomes of a number of organisms being completely sequenced now, it has become possible to apply a variety of mathematical and statistical tools to analyze a wide variety of DNA sequences in order to study various features and underlying patterns. DNA, for the purpose of such analysis, is typically considered as a symbolic string of length \( N \) bases, the bases being
the nucleotides denoted $A, T, G$ and $C$. A prime concern in this direction is the examination of long-range correlations in DNA sequences. A number of studies have found evidence (with varying levels of certainty) for long-range fractal correlations \cite{8, 9, 28, 35, 53, 66, 72, 75, 84, 104, 105, 129, 130, 131, 136} although this is both controversial, and in the end, apparently of questionable biological significance \cite{9, 53, 73, 74}.

The patchy or "mosaic" structure of DNA has resulted from the expansion or modification of the DNA at different times, and can be seen in a variety of ways. Most of these techniques use some property or characteristic of a symbol string, and attempt to examine whether a string and its substrings differ significantly. As discussed in Chapter 1, the divergence of any property $P$ on segments of a string $S$ can be used to fragment a genomic string within a given analytic framework.

Finding structure within structure is also an aim of studies designed to explore fractal organization in the genome. This is usually evidenced in power-law distributions for characteristic quantities. Considerable recent work has been directed toward detecting and understanding power-laws and fractality as an underlying factor in genome organization \cite{9, 53, 84, 129}. The alleged differences between coding and noncoding DNA may be less significant than supposed \cite{28, 130}. Both noncoding and coding sequences have been seen to give some evidence of power-law correlations \cite{8, 131}. Genome-wide studies seem to show the existence of such correlations \cite{9, 53, 84, 129}. Lu et al. \cite{84} analyzed eight complete genomes of bacteria using some parametric methods and shown that fractal phenomena and long-range correlations prevail through the entire genome. Subsequent low frequency spectral analysis on thirteen microbial complete genomes by Vieira \cite{129} have shown the presence of such correlation, though the correlation length is found to be not extending to the entire DNA chain. Hao et al. \cite{53} used a simple deterministic method that generate a visual representation of long sequences or complete genomes and found fractal like patterns to be prevalent in all. Audit et al. \cite{9} attempt to find the cause of power-law correlations in genomic DNAs using wavelet analysis.

A measure that appears to have utility in both describing the statistical properties of DNA sequences and in having some correlation with biological aspects of the DNA is the Shannon entropy \cite{80}. A number of studies have focused on this quantity, and have made this the basis of uncovering the mosaic organization of DNA sequences \cite{35, 37, 47, 48, 78}. One approach,
generally termed "segmentation", attempts to find mutually distinctive portions of the DNA which are, however, homogeneous with respect to a given criterion. A number of different criteria can therefore be used to segment DNA [24, 35, 37, 48, 78, 110]. For instance, by using the Shannon entropy as a quantitative measure, the intention is therefore to fragment a given DNA string into substrings such that the intra-substring entropic variation is small, while the inter-substring entropic variation is large.

In this Chapter, we use the Jensen–Shannon entropic divergence [35, 48] to fragment several entire genomes. Our motivation is to examine the properties of the segmentation procedure, and to examine the statistical properties of the segments themselves. One feature which we wish to explore is whether the fragments obtained by entropic segmentation have evidence of scale invariance; numerous examples of similar processes appear to be scale invariant [38, 67]. Indeed, physical fragmentation also leads to a mass distribution which is fractal [4, 98]. We also propose here a new way of relabeling the domains which leads to a significant decomposition of the complexity of a genome. The implications of the relabeling seem to be significant with regard to evolutionary studies.

5.2 Segmentation through Entropic Divergence

The segmentation of a genomic sequence is accomplished in the following manner. Given two symbolic sequences built from an alphabet of $k$ symbols, the Jensen–Shannon divergence,

$$ J(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}) = H(\mathcal{F}) - \frac{n^{(1)}}{N} H(\mathcal{F}^{(1)}) - \frac{n^{(2)}}{N} H(\mathcal{F}^{(2)}), $$

is a measure of the compositional difference between them. Here $n^{(i)}$ and $\mathcal{F}^{(i)} = \{f_1^{(i)}, f_2^{(i)}, \ldots, f_k^{(i)}\}$ $i = 1, 2$ are the lengths and relative frequency vectors, respectively, characterizing the two sequences. By concatenating the sequences to get a single sequence of length $N = n^{(1)} + n^{(2)}$ with $\mathcal{F}$ the corresponding frequency vector, the corresponding Shannon entropy is

$$ H(\mathcal{F}) = - \sum_{i=1}^{k} f_i \log_2 f_i. $$

62
\( J \) may be generalized to study the divergence among \( m \) sequences; the process of segmentation suggested in \([35, 48]\) calculates the difference for \( m = 2 \) sequences.

The procedure to completely segment a given sequence into homogeneous subsequences (domains or patches) is as follows. A sequence of length \( N \) is partitioned into two subsequences of lengths \( n^{(1)} \) and \( n^{(2)} = N - n^{(1)} \) respectively, by varying the partition, namely all choices of \( n^{(1)} \) so as to maximize the divergence \( J \). This procedure is then carried out recursively.

One important issue needs to be addressed for the above procedure to be meaningful and self-consistent: How does one stop segmentation? Namely, can every sequence be partitioned into distinct subsequences? This issue has been addressed through an estimation of the statistical significance of the divergence measure. Thus, in order to determine whether the partitioning point that maximizes \( J \) is statistically significant or not, two potential subsegments are compared with those from random fluctuations. If \( J_{\text{max}} \) is the maximum value of \( J \) among all possible cutting points, the statistical significance of this maximum is determined by obtaining the probability of getting this value or less in a random sequence. The significance level is thus defined as

\[
S_{\text{max}}(x) = \text{Prob}\{J_{\text{max}} \leq x\}.
\]

An approximate analytic expression for the probability distribution of \( J_{\text{max}} \) has been found \([37, 48]\),

\[
S_{\text{max}}(x) = [F_\nu(\beta \cdot 2N \ln 2 \cdot x)]^{N_{\text{eff}}},
\]

where \( F_\nu \) is the \( \chi^2 \) distribution function with \( \nu = (k-1)(m-1) \) degrees of freedom, \( \beta \) is a scale factor largely independent of \( N \) and \( k \) and for each \( k \), \( N_{\text{eff}} = a \ln N + b \) (\( a, b \) are constants). The values of \( a, b \) and \( \beta \) are obtained from Monte Carlo simulations by fitting the empirical distributions to the above expression \([37, 48]\).

A sequence is segmented at a preassigned significance level \( s_0 \) as follows. If \( s_{\text{max}} \), determined as discussed above, exceeds \( s_0 \), the sequence is segmented at this point. The procedure is continued recursively for each of the two resulting segments. It is necessary to ensure that at each stage, the resulting subsequences maintain their distinction (vis-a-vis the Jensen–Shannon divergence) from their neighbours formed at the previous segmentation steps. The process is terminated when all segments thus obtained either have \( s_{\text{max}} \leq s_0 \),
or if a possible partition will lead to segments which are not compositionally
distinct from their neighbours.

5.2.1 Application and Results

The genomic DNA sequences studied here are represented by the usual 4-
symbol alphabet \( \{A, T, C, G\} \) and a 12-symbol alphabet \( \{A_i, T_i, C_i, G_i\}, i = 1, 2, 3 \), where the subscripts indicate the positions of bases within a codon [37]. The representative genomes taken from the GenBank are Thermoplasma acidophilum (archae bacteria, 1.56 Mbp), Campylobacter jejuni (eubacteria, 1.64 Mbp), Saccharomyces cerevisiae chromosome IV (eukaryota, 1.53 Mbp), Arabidopsis thaliana chromosome II (eukaryota, 1.5 Mbp) and human chromosome 22 (eukaryota, 1.52 Mbp). Since the chromosomes of A. thaliana and human are very long and incomplete, we take a long string of 1.5 Mbp which is completely specified for A. thaliana and a contig (gi|10880022|ref|NT_011522.1) of human chromosome 22.

These genomes were segmented as discussed in the preceding Section. We study the probability distribution of segment lengths obtained for the different genomes. Denote by \( n(\ell) \) as the number of segments of length \( \ell \), and consider the distribution

\[
F(\ell) = \sum_{\ell' = \ell}^{\infty} n(\ell') \approx \int_{\ell'}^{\infty} n(\ell') d\ell'
\]

(5.5)

which corresponds to the number of segments with length greater than or equal to \( \ell \). If the distribution of lengths follows a power law, namely \( n(\ell) \propto \ell^{-\alpha} \), \( F(\ell) \) also follows a power law, with exponent \( 1 - \alpha \).

Fig. 5.1–5.4 shows \( f(\ell) = \ell^{-1}F(\ell) \) for the above genomes using the 4-symbol alphabet, segmented at four different significance levels (99%, 90%, 90% and 85%). Note that for \( k = 4 \) and \( m = 2 \), the number of degrees of freedom \( \nu = 3 \), and from Monte-Carlo simulation, \( a = 2.44, b = -6.15 \) and \( \beta = 0.79 \) [48]. The number of segments obtained at 99% significance level are 516 for \( T. \) acidophilum, 654 for \( C. \) jejuni, 584 for \( S. \) cerevisiae, 1122 for \( A. \) thaliana and 1360 for human chromosome 22 contig. At a given significance level, a heterogeneous DNA sequence yields more segments than a relatively homogeneous one [35, 48]. As may be expected, therefore, the human chromosome has larger number of segments than a bacterial sequence of comparable length. Lowering the significance level increases the number
Figure 5.1: The patch length distribution $f(l)$ of homogeneous segments obtained by segmenting complete genome of *Thermoplasma acidophilum* represented by 4-symbol alphabet $A,T,C,G$ of bases, at 99%, 95%, 90% and 85% levels of statistical significance. The genomic sequence has been segmented using a recursive segmentation method as described in the text. Each distribution from below has been separated by a decade for clarity.

of segments obtained for an organism; similar trends observed as above for the five species can also be observed at other significance levels.

Patch length distributions in Fig. 5.1–5.4 show a characteristic scale separating two clear regimes of power–laws. This scale is nearly constant from archaeal genome to early eukaryotic and lies in the range of $\approx 1300$-$1500$ bp. Thus, $f(l) \propto l^{-\alpha}$, $i = 1, 2$, $\alpha_1$ is the exponent below the characteristic scale and $\alpha_2$ above. The exponents are obtained from a fit to the data, and $\alpha_1 < \alpha_2$. The values of $\alpha_1$ and $\alpha_2$ from Fig. 5.1–5.3 are given in Table I. The segment distributions of higher eukaryotes, for example *A. thaliana* and human (see Fig. 5.4) appear to be smooth and are conspicuous by the absence of a characteristic scale.
Figure 5.2: Same as in Fig. 5.1, for the bacterium *Campylobacter jejuni*.

Figure 5.3: Same as in Fig. 5.1, for *Saccharomyces cerevisiae* chromosome IV.
Table I. The scaling exponents \( \alpha_1 \) and \( \alpha_2 \) observed below and above the characteristic scale which partitions the patch length distribution into two power-law regimes for the genomes of three different representative organisms. The scaling exponents given correspond to the genomic sequences coded in \( k = 4 \) symbol alphabet and at four different significance levels as described in the text.

<table>
<thead>
<tr>
<th>Genome</th>
<th>99%</th>
<th>95%</th>
<th>90%</th>
<th>85%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \alpha_1 )</td>
<td>( \alpha_2 )</td>
<td>( \alpha_1 )</td>
<td>( \alpha_2 )</td>
</tr>
<tr>
<td>\textit{T. acidophilum}</td>
<td>1.12</td>
<td>2.76</td>
<td>1.18</td>
<td>2.85</td>
</tr>
<tr>
<td>\textit{C. jejuni}</td>
<td>1.1</td>
<td>2.81</td>
<td>1.18</td>
<td>3.1</td>
</tr>
<tr>
<td>\textit{S. cerevisiae}</td>
<td>1.18</td>
<td>2.37</td>
<td>1.25</td>
<td>2.65</td>
</tr>
</tbody>
</table>

Distribution profiles for the genomes encoded in 12-symbol alphabet and
Figure 5.5: The patch length distribution $f(\ell)$ as in Fig. 5.1 of five genomes, using a 12-symbol alphabet which takes into account both the base and the codon position $\{A_i, T_i, C_i, G_i\}, i = 1, 2, 3$ (see text). The segmentation has been done at 99%, 95% and 90% levels of statistical significance.

Segmented at 99%, 95% and 90% levels of significance are shown in Fig. 5.5(a)(d). For this code, the number of degrees of freedom is $\nu = 9$ [37, 48] and $a, b, \beta$ are 2.34, -3.69, 0.84 respectively.

The distributions show similar trends in bending profiles as for the 4-symbol alphabet, though the characteristic scales now show differences from archaean to early eukaryotes (Fig. 5.5(a)-(c)). The characteristic scale for $T. \text{acidophilum}$ and $C. \text{jejuni}$ is $\approx 800$ bp and for $S. \text{cerevisiae} \approx 1100$ bp. We find the two scaling regimes to be most clearly distinguished at 99% significance level. Segmentation of a nucleotide sequence coded in 12-symbol alphabet delineates the coding and noncoding regions; we find these characteristic lengths almost same as the average size of coding segments of the respective genomes. The power-law exponents $\alpha_1$ and $\alpha_2$ are given in Table II.
Table II. The scaling exponents $\alpha_1$ and $\alpha_2$ observed below and above the characteristic scale for the genomes of three different representative organisms. The scaling exponents given correspond to the genomic sequences coded in $k = 12$ symbol alphabet and at three different significance levels as described in the text.

<table>
<thead>
<tr>
<th>Genome</th>
<th>99%</th>
<th>95%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_1$</td>
<td>$\alpha_2$</td>
<td>$\alpha_1$</td>
</tr>
<tr>
<td><em>T. acidophilum</em></td>
<td>1.09</td>
<td>4.12</td>
<td>1.15</td>
</tr>
<tr>
<td><em>C. jejuni</em></td>
<td>1.08</td>
<td>4.13</td>
<td>1.13</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>1.22</td>
<td>3.81</td>
<td>1.26</td>
</tr>
</tbody>
</table>

The characteristic length which is observed as separating the two regimes of power-law behaviour does not appear to be an artifact either of the statistical significance level or of finite size (length). This is evident from our results shown for a wide range of significance levels in Fig. 5.1–5.3 and Fig. 5.5(a)–(c). The scaling exponents on either side of the characteristic length vary with the significance level used to segment a genome, but the length itself remains unchanged. A higher significance level results in reduced segmentation and thus larger patch sizes, while a lower significance level causes larger patches to segment further. Even when only half or three fourth of the genome is used, two scaling regimes result with the same characteristic length, as shown in Fig. 5.6 for *T. acidophilum* (90% significance level).
5.2.2 Discussion

Unraveling the history of a given genome is a complex task, and one that needs a variety of different approaches. Identification of the different structural features within the DNA — the exons and introns, repetitive DNA, telomeres, isochores, for instance, is one objective. However, the DNA of any organism itself has a complicated evolutionary history with a variety of different selection pressures acting on it at different times. To uncover this aspect of DNA evolution, it may be essential to go beyond an analysis of the functional parts of the DNA, and segmentation study, like the one presented in this chapter, is one approach to understanding the genome organization.

The principle behind segmentation is simple: break up a complex object into its "constituent" parts, and thereby attempt to understand how the organization comes about in the first place.

This differs from the motivation, for instance, in studying the fragmentation of physical objects (rocks or gypsum molds of different shapes, for instance). Distribution profiles of masses of fragments so obtained have been
seen to follow a power–law behaviour [98]. There is, however, a characteris­
tic size so that the actual distribution of fragment masses \( m \) follows the law
\[ a m^{-b} \exp\left(-\frac{m}{m_0}\right), \]
where \( m_0 \) is the characteristic finite–size cutoff mass, and \( a \) and \( b \) are constants. Below \( m_0 \) the distribution is close to a power–law be­

haviour with exponent \( b \). Physical fragmentation is dictated by the breaking
of strong bonds between molecules that determine the structure and tensile
strengths of solids; the mass distribution follows a power–law for the entire
range of masses (minus the cutoff). Such fragments are held together in a
solid by the same type of adhesion so that a smaller fragment breaks in a
similar manner as a larger one.

DNA sequences are heterogeneous at various levels of description and
thus the fragmentation of DNA into entropically homogeneous segments is
different in principle. We have studied the differences in segment length dis­
tributions for the genomic DNAs of representative organisms spanning the
three classified kingdoms. The segments that a given sequence is divided into
are such that within a given domain, the composition is uniform (in terms
of the Shannon entropy); thus these domains could reflect the evolution of
a given genome. Based on this premise, one would expect that an organ­

ism further along the evolutionary tree will have a more complex genomic
organization.

This is borne out in our studies: for bacterial genomes, the domains ap­
pear to have a power–law distribution with evidence of two separate regimes
of scaling behaviour. Although we have analyzed all available complete

genomes of archaeabacteria and eubacteria, only representative data have been
presented here and compared with chromosome wide data for three eukary­
otes (again, not all the examples analyzed are presented here). Our results,
which are consistent across the kingdoms, show two regimes of power–law
scaling in the bacterial genomes as well as primitive eukaryotes like yeast.
Application of the segmentation algorithm to other bacterial genomes shows
similar features, displaying the general fractal organization of nucleotides
that make up such genomes. This is in contrast to the segment length distri­
bution of highly evolved eukaryotes like human or plant; here the distribution
shows a smooth transition across the entire range of segment lengths, thus
lacking distinct characteristic scales. The nucleotide composition of higher
eukaryotes is very complex; this is attributed to the abundance of noncoding
sequences or introns in DNA sequences of such organisms. Of as yet unde­
termined function, noncoding DNA may well be crucial vis-a-vis evolution.
Such regions are more prone to alteration by different evolutionary processes, e.g., duplication, mutation, insertion, deletion, etc. The fragments resulting from the segmentation algorithm carry the imprint of these different processes and it may be anticipated that the scaling features would be more complex than for bacterial sequences.

5.3 Domain Reorganization: Decomposing complexity

As noted above, the DNA sequences are generally heterogeneous. The base composition or that of a group, e.g. purine (A or G)/pyrimidine (C or T), strong (G or C)/weak (A or T) etc. are nonuniformly distributed along a DNA sequence. Coding regions which show periodicity three alternate with noncoding regions, the presence of high density CG dinucleotide regions (the so-called CpG islands) etc. Such heterogeneities lends a complex structure to DNA molecule. In comparison to English language which has 26 letter alphabet, a DNA sequence has only four but the complexity arises because of the length of sequence. DNA molecules are typically very long; human DNA, for instance, is estimated to contain \( \approx 3 \times 10^9 \) base pairs. Hence due to this length factor, there would be more combinations of letters and as a result more 'words'.

Long-range correlations in DNA [72, 75] have been attributed to the presence of complex heterogeneities in nucleotide sequences [35, 76]. Peng et al. have shown long-range correlation using a 'DNA Walk' model based on detrended fluctuation analysis and found such correlation to be intrinsic to intron-containing DNA sequences[104]. The sequence having long-range correlation shows a complex DNA walk landscape in comparison to one which lacks such correlation. The presence of complex heterogeneities result in hierarchical patterns in DNA described well by a 'domain within domain' type of structure first observed by Bernaola-Galván [35]. There have been some attempts to understand the biological implications of such complexity [9, 53, 73, 74], but very little is known as of now.

It is thus an important task to understand the information implicit in such heterogeneities. One way to achieve this goal is to segment a DNA sequence into compositionally homogeneous domains; recent applications suggest that
it may be possible to locate the coding/noncoding boundary through the Jensen-Shannon divergence [37]. As discussed in Section 5.2 stopping criteria determine the extent to which a sequence can be segmented. Alternate criteria for segmenting nucleotide sequences within a model selection framework has been suggested recently; Li [78] proposes a statistical tool called Bayesian information criterion (BIC) which imposes a minimum condition for the recursive segmentation to continue. This is a more stringent stopping criterion and can be used to detect biologically meaningful longer domains called isochores. In a separate work, Li [79] has shown that model selection scheme can also be used for detecting CpG island, finding the origin and terminus of replication in bacterial genomes, searching complex repeats in telomere sequences and finding borders of coding and noncoding regions.

In this Section I reexamine the segmentation procedures as proposed in [35, 37, 78] and [78] to examine a variety of genomic DNA sequences ranging from primitive bacteria to human. One demerit of these methods is that they present a local overview— the segmentation is done by maximizing the divergence $J$; it employs a search for local maximum and no attempt is made to look for global a maximum. The domains obtained by applying the segmentation procedures above are each heterogeneous with respect to neighbouring domains by definition. However, a given domain may be homogeneous with respect to some other domain on the genome. The domain picture provided by these methods is thus not complete: each domain may in reality be a part of a larger potential domain and a complete domain can be obtained by joining all such parts which are scattered at different locations in a genomic sequence. The number of reconstructed domains may in fact be lower than that obtained by following [78] and [35, 37].

The method that I propose here starts from the domains found from one of the above methods and builds up super domains by determining the inter-segmental homogeneity. Say that following the application of the primary segmentation procedure, the sequence $S$ is shown to be

$$S = S_1 S_2 S_3 \ldots S_i S_{i+1} \ldots S_n.$$  

The heterogeneity between $S_k$ and a subsequent segment, $S_{k+j}$, $k=1,2,\ldots, j=2,3,\ldots$ is determined in a similar fashion. The different segments are relabeled: If $S_k$ and $S_{k+j}$ cannot be distinguished at this level, then they are given the same label. This procedure is repeated for all segments until
no further relabeling is possible. This will be made clearer through a simple example below.

5.3.1 Segmentation using model selection

Segmentation based on a model selection [78] also uses $J$ as the entropic scissor, but presents a new stopping criterion for segmentation which offers a minimum condition for accepting a potential cut. This seems to be a stringent criterion and uses a model selection approach—whether a two random subsequence model is better than one random sequence model. This is determined by the Bayesian information criterion (BIC) [60, 109] which is defined as

$$BIC = -2 \log(\hat{L}) + \log(N)K + O(1) + \frac{1}{\sqrt{N}} + O\left(\frac{1}{N}\right),$$

where $\hat{L}$ is the maximum likelihood of the model, $N$ is the sample size and $K$ is the number of parameters in the model. BIC is well approximated by the first two terms; a given model is better if it has a larger integrated likelihood, and thus a smaller BIC.

The condition for segmentation to continue is $\Delta BIC < 0$, which from the above equation implies [78]

$$2NJ > (K_2 - K_1) \log(N),$$

where $K_1$ and $K_2$ are the number of free parameters of the models before and after the segmentation. This is thus the lower bound of the significance level. There is no upper bound and it can be preset by using a measure called segmentation strength [78],

$$s = \frac{2NJ - (K_2 - K_1) \log(N)}{(K_2 - K_1) \log(N)}.$$  \hspace{1cm} (5.8)

Eq. (5.7) is equivalent to $s = 0$. The results obtained in the subsequent sections are with respect to this segmentation method unless otherwise stated.

5.3.2 Insertion–deletion and heterogeneity

DNA sequences are prone to changes at the base level owing to different evolutionary processes—duplication, modification (mutation), insertion, deletion
etc... A number of studies have been devoted to understand the 'duplication-modification' process [71, 73, 81, 134] but less attention has been given to the process of 'insertion-deletion'.

To examine such processes, we perform the following simple theoretical experiment. A string of \( \sim 0.75 \text{ Mb} \) DNA of human chromosome 22 \( (\text{gi}10879979|\text{ref}NT_011521.1) \) is taken and random fragments of a bacterial sequence of 80 Kbp (taken from the genome of Ureaplasma urealyticum) are inserted at random positions in it. The bacterial sequence has been fragmented randomly into 403 pieces. It can be anticipated that the heterogeneity will increase because of such insertions. The number of fragments obtained after segmenting this sequence is 375; the original sequence gives 248 segments. To quantify the complexity of DNA sequences, we use a complexity measure introduced by Bernaola-Galván et al. called 'Sequence Compositional Complexity' (SCC) [36, 114], defined as,

\[
SCC = H(S) - \sum_{i=1}^{n} \frac{l_i}{L} H(S_i)
\]

where \( S \) denotes the whole sequence of length \( L \) and \( S_i \) is the \( i \)th domain of length \( l_i \).

\( SCC \) is a measure of the compositional complexity of a sequence and is independent of the length of sequence. This represents the maximum value of divergence for 1 to 2 segmentation process and corresponds to the optimum partition of a sequence. A relatively heterogeneous DNA sequence has a higher SCC. We computed the difference \( \Delta SCC = SCC' - SCC \) for domains obtained after and before the insertion for a number of genomes and found the compositional complexity to increase after insertion, i.e. \( \Delta SCC > 0 \) in all cases.

Any sequence into which parts of sequences of different base compositions get inserted will now have a greater tendency to fragment; this is a consequence of built in complexity as a result of this process. To have a feeling of the relative increase in the number of domains, we now employ simultaneously the other process also- 'deletion' keeping all such processes random and constraining the sequence length to remain the same. This is done as
follows. After every insertion, a fragment of random length is deleted from a random position (the range of lengths being deleted is kept same as that of the 'inserts'). The 'inserted-deleted' sequence yields 450 segments upon segmentation (the original one gives 248 segments). Obviously this shows such processes cause the heterogeneity to build up. The above experiment is repeated by adding instead of one insertion at a time, any number of bacterial segments ranging from 0 to 10 chosen randomly and then deleting any number of segments again ranging from 0 to 10. We carried out more such experiments for different random realizations and the results are found to substantiate what has been said above.

5.3.3 Patch relabeling

I now describe the relabeling procedure in detail. Assume sequence S to be partitioned into n segments labeled $S_1, S_2, S_3, \ldots, S_n$. Such a segmentation technique proceeds by executing 1 to 2 segmentation as explained above and the segments obtained are heterogeneous relative to their neighbours. Hence a more reasonable picture of homogeneity can only be formed by searching for a segment, all other segments (except the neighbours) which when joined together results in a homogeneous sequence. Such a merger of segments renders the splitting points insignificant and no new split point is significant for the resulting sequence. The segmentation methods partition a sequence, we suggest rejoining but in a new perspective. It thus goes a step beyond where the previous methods halt. It brings forth a global view of dispersion of domains vis-a-vis the local one generated by 1 to 2 segmentation process. An optimal realization of this can be got by following this procedure: for segment $S_3$, check if it is homogeneous with $S_1$ (there is no need to check with $S_2$ as it is heterogeneous with $S_3$ as per the definition). Join $S_1$ and $S_3$ and segment, if not segmented (homogeneous) then label it as $S_1$, otherwise as before ($S_3$). Now check similarly for $S_4$ with $S_2$, if homogeneous, change its label to $S_2$, otherwise check with $S_1$ only if $S_3$ is not changed to $S_1$. Check $S_5$ with labels before the fourth segment, but now instead of checking with a single segment, check it with all segments of same label joined together (e.g., if $S_3$ is not $S_1$, check with $S_3$, but if $S_3$ is now changed to $S_1$, join $S_1, S_3$ and $S_5$ and check for the splitting point, if not segmented label fifth segment also as $S_1$, if segmented proceed to label $S_2$ and check and so on. Since in the segmentation process, neighbours of a segment are heterogeneous with it, it should be
ensured that no two consecutive segments have the same label. Thus for every next segment $S_i$, check with the labels \{ $S_{i-2}$, $S_{i-3}$, ..., $S_1$ \} step by step but now globally (joining all segments with same label together and thus checking with the segment of interest) until $S_i$ is found to be homogeneous with the combination of segments of same label; in this way finally we obtain a symbolic representation of domains with number of labels much fewer than $n$ and all the segments with same label when joined are homogeneous at a preset level of significance.

An example is given below.

The complete genome of a bacterium *Ureaplasma urealyticum* (751719 bp) and a contig of human chromosome 22 (gi | 10879979 | ref | NT_011521.1 |, 767357 bp) represented by base alphabet \{ A, T, C, G \} are segmented at the lower bound of the stopping criterion (Eq. (5.7)). The number of segments obtained is 86 for the bacterium and 248 for human.

Following the above procedure, the symbolic representation for the bacterium can be written as-

```
S_1 S_2 S_3 S_4 S_5 S_3 S_1 S_2 S_1 S_6 S_4 S_1 S_7 S_2 S_1 S_6 S_4 S_8 S_3 S_4 S_9 S_{10} S_4 S_9 S_{11} S_6 S_4 S_{10} S_6 S_{10} S_6 S_{11} S_6 S_7 S_6 S_{11} S_7 S_3 S_{11} S_3 S_{10} S_6 S_3 S_3 S_9 S_{11} S_{10} S_4 S_4 S_{11} S_{10} S_{13} S_4 S_{13} S_9 S_{11} S_4 S_6 S_4 S_{15} S_1 S_9 S_4 S_{16} S_9 S_{17} S_{15} S_6 S_{17} S_7 S_{17} S_1 S_{17} S_8 S_{16} S_{14}
```

The 86 segments have been labeled by only 17 symbols.

For human, the symbolic representation is obtained as—
The 248 original segments are now labeled by only 53 distinct symbols. It is interesting to observe that symbols like $S_1$ in the bacterium or $S_2$ in human are dispersed throughout the genome. The symbolic representation shows the pattern of distribution of parts of domains. Unfolding of such a pattern can lead to some important biological implications. A striking feature that comes immediately into picture is an evidence of 'insertions', one of the crucial factors in evolution. Expansion-modification and insertion-deletion [73] are thought to play major role in evolution. The former ensures duplication accompanied by point mutations in genomes and the latter results in insertion of a part of chromosome inside a nucleotide sequence or deletion of base pairs from a nucleotide sequence. An initially homogeneous sequence may thus become heterogeneous by insertions/deletions that consistently accompanies evolution and pieces of a homogeneous sequence can be widely displaced.
5.3.4 Measuring the complexity

The decomposition of complexity is quantified in terms of measures like Bayesian information criterion (BIC) and Sequence compositional complexity (SCC). As explained earlier, BIC is used for model selection and a model is better than the other if it has a lower BIC. Even though it can be anticipated from the symbolic representation that the complexity gets reduced remarkably, it would be interesting to see whether this model is better than the earlier one. For example, for Ureaplasma urealyticum, originally 86 domains have been obtained and BIC can be expressed as,

\[
BIC = -2 \log(\hat{L}) + \log(N)(86 \times 3 + 85) \\
= -2 \log(\hat{L}) + \log(N) \times 343, 
\]

(5.10)

where \( K = 343 \) free parameters correspond to \( 86 \times 3 \) base compositions and 85 borders. After following the procedure described above, we are left with only 17 domains and we have

\[
BIC' = -2 \log(\hat{L}') + \log(N)(17 \times 3 + 85) \\
= -2 \log(\hat{L}') + \log(N) \times 136. 
\]

(5.11)

The maximum likelihood can be expressed as

\[
L(p_\alpha) = \prod_\alpha p_\alpha^{N_\alpha},
\]

(5.12)

where \( \{p_\alpha\} \) and \( \{N_\alpha\} \) are the base composition parameter and the base counts respectively corresponding to the alphabet \( \{\alpha = A,T,G,C\} \). \( L \) in fact would be larger than \( L' \) because it uses a more accurate measurement of base composition. Hence the first term of BIC is smaller than that of BIC'. But this decrease is actually much less than the corresponding decrease in second term of BIC' with respect to that of BIC which shows that the saving on model complexity is large and BIC' becomes smaller than BIC.

For U. urealyticum and human DNA, \( \Delta BIC = BIC' - BIC \) are -1709 and -4884 respectively which shows that the reconstructed symbolic representation model is better than the original one (we use the lower bound i.e. \( \Delta BIC < 0 \) for determining the statistical significance). As another example, we found \( \Delta BIC \) for Thermoplasma acidophilum (archaebacteria,
1564906 bp) and another contig of human chromosome 22 (gi | 10880022 | ref | NT_011522.1 |, 1528072 bp) to be −2808 and −10420 respectively.

To quantify the decrease in complexity in terms of SCC (Eq. (5.9)), the difference $\Delta SCC = SCC' - SCC$ for domains obtained after and before relabeling is computed for a number of genomes: we find $\Delta SCC < 0$. The above two measures substantiates our proposal of decomposing the complexity of genomic sequences in a new perspective.

### 5.3.5 A model of host–parasite interaction

One possible application of the above procedure is in modeling host–parasite interactions. A more heterogeneous DNA sequence may have its evolutionary history starting from a relatively homogeneous predecessor. Depending on the various evolutionary processes acting over a long time, as discussed in previous sections of this Chapter, the complexity and heterogeneity, in general, increases. In pursuance of our studies above, we present here simple computational experiments which mimic a situation where a homogeneous human DNA takes inside fragments from homogeneous segments of bacterial genomes and thus changes to a relatively heterogeneous sequence.

It is well known that viruses and bacteria have participated in the so-called lateral gene transfer, genes or DNA segments can be inserted into a genome from another. Does this simple process underlie the basic heterogeneity of genomic sequences as is probed by the entropic divergence?

In the simplest model, we consider two very different DNA sequences corresponding to host and parasite respectively, and have the parasite insert $N$ DNA fragments into the host DNA. If the segmentation procedure is able to detect the differences, then there will be $2N + 1$ different segments. By attempting to reduce complexity as discussed above, we should be able to show that there are indeed only two different sequences.

An experimental implementation of the above idea is as follows. We take the longest reconstructed homogeneous patch (eight segments join here to make one homogeneous domain of length 100139 bp) of the human genome segmented and symbolically represented as above. A homogeneous segment of length 17584 bp from the genome of *Ureaplasma urealyticum* is taken and fragmented into three pieces of 5000, 7000 and 5584 bp each and inserted at three different locations inside the human segment. Now in the new sequence (of total length 100139 + 17584 bp), the three fragments of bacterial
genome are at locations 10000 – 15000, 50000 – 57000 and 92000 – 97584 bp respectively. Thus there are seven distinct segments in all (3 bacterial plus 4 human).

This sequence is segmented and the borders are found to be at locations 9984, 15000, 49751, 50060, 56978, 91636 and 97575 bp. The borders almost match with the real ones (above) except that there is one extra border (49751 & 50060 are very close and correspond to 50000 as in the real one). For the seven segments, the symbolic representation of the domains is obtained is: $S_1S_2S_1S_2S_1S_2S_1$. The four $S_1$ segments concatenated retrieves the human segment and the three $S_2$'s together reconstruct the original unfragmented bacterial sequence.

The example presented is simple and the inserted fragments are well separated, but the above procedure works well even with variations in the scheme of insertion: by inserting the fragments obtained from two different bacterial genomes (Ureaplasma urealyticum and Thermoplasma acidophilum) now fragmented randomly and inserted at random locations inside the human segment. The two bacterial segments of length 17584 bp and 13084 bp of U. urealyticum and T. acidophilum respectively were fragmented into three random parts each and inserted randomly. The 'inserts' and their positions in the human segment are shown in Fig. 5.7(a). The total number of mutually distinct patches inside the human sequence is 13.

The ‘inserted’ human sequence is segmented and we now found the segments to be greater than 13. Obviously with such random insertions, the complexity of the sequence increases and we now need to increase the segmentation strength $s$ from lowest threshold ($s = 0$) to a value where we obtain most of the borders of inserts and the number of segments are close to 13. The segmentation method with varying values of $s$ is shown to locate a number of features in DNA sequences [79]. We find such a reasonable compromise at $s = 0.2$ and the human sequence gets partitioned into 18 segments as shown in Fig. 5.7(b). The procedure to obtain symbolic representation of domains is executed at the same $s$ and Fig. 5.7(b) shows the segments which have been identified as human or the bacterial. We have been able to retrieve 84% of human, 85% of U. urealyticum and 67% of T. acidophilum.

We obtained some more realizations of the above experiment and found that with our approach, we are able to show the essence of some processes perpetually active in evolution.

While this procedure appears to work quite generally, it should be added
Figure 5.7: (a) Fragments of two bacterial sequences *T. acidophilum* (T) and *U. urealyticum* (U) are inserted randomly inside a human sequence (H) (see text) and (b) The sequence as in (a) is segmented and followed by the procedure of symbolic representation of domains as described in the text in order to retrieve the original configuration as in (a).

that we have found it difficult to retrieve small inserts as well as a small proportion of inserts that accidentally match the composition of the host sequence.

### 5.3.6 Patch length distribution

As in Section 5.2.1, we have shown the differences in patch length distribution by segmenting the whole genomes of organisms spanning the three classified kingdoms using hypothesis testing approach. Patch length distributions for archaea, eubacteria and primitive eukaryotes like yeast show a characteristic scale separating two clear regimes of power-laws which becomes smooth transition for highly evolved eukaryotes like plant or human. The features like power-laws and scaling observed in bacterial and primitive eukaryotic genomes are evident at wide range of significance levels used to segment the genome (we observed up to 99% level of significance), but for much higher levels like 99.9999% (which is almost equivalent to BIC), where one gets fewer patches, it is difficult to differentiate such features clearly. Hypothesis testing framework provides flexibility in choosing lower and upper bounds of significance level. The domains obtained at a particular level of significance
Figure 5.8: The patch length distribution at a significance level of 90% for complete genome of *Thermoplasma Acidophilum*. The genome is segmented and distribution obtained as described in the text. The distribution titled ‘Before’ is for segments obtained after the recursive segmentation is complete. ‘After’ stands for the distribution of domains obtained after joining parts of the domains scattered at different locations in the genomic sequence determined via a symbolic representation as explained in the text.

may be segmented at a lower significance level. Thus by going down the significance level, one can observe domains within domains. We use significance levels which yield reasonable number of domains appropriate for distribution analysis.

The whole genome of *T. acidophilum* and the human chromosome 22 contig (gi|10880022|ref|NT_011522.1|) are segmented at 90% significance level. Then the symbolic representations are obtained and thus the longer homogeneous patches by joining all segments of the same label. Fig. 5.8 shows the patch length distributions \( f(l) \) of *T. acidophilum* before and after the symbolic representation. For ‘before’, the distribution shows transition from one power-law region to another at a characteristic scale \( \sim 1300-1500 \) bp. For ‘after’, while the two power-law regimes are retained, the characteristic scale increases.

For human DNA, the distribution is like a smooth transition for ‘before’ (Fig. 5.9). Distribution of domains after the symbolic representation shows
the signs of appearance of a power-law regime (see 'after' in Fig. 5.9).

We also notice that scaling features are less distinct for distribution of *T. acidophilum* at 99.9999% level (128 segments) with respect to domain distribution obtained after the relabeling at 90% significance level (88 segments) even though the two distributions result from comparable number of domains obtained.
5.3.7 Discussion

The present approach provides a novel method for the unraveling of the compositional heterogeneity of a DNA sequence. The recursive segmentation methods [37, 78, 114] result in segments which are compositionally different with respect to neighbours. This local view can be made more global by examining the intersegmental relationship. Segments formed through one process of segmentation need not be distinct from those that are not their immediate neighbors; this leads to a novel means of reducing the apparent complexity of a sequence.

Insertion-deletion is a prime factor in evolution. A homogeneous sequence may become heterogeneous by the insertions of sequences from other parts of a chromosome and also the deletion process that consistently works on in evolution. This has been shown through some simple experiments. The number of segments obtained from a sequence subjected to such processes is found to increase considerably in comparison to the original sequence. The presence of parts of a domain at different positions in a chromosome sequence is a strong pointer substantiating the ‘insertion theory’. It can be seen that the domain of $S_2$ type is dispersed in parts almost from beginning to end in the human chromosome contig segmented above. Such observations can help to understand some of the processes involved in molecular evolution.

Our present studies can be augmented and reinforced through the examination of sequence homology, for example through the similarities that can be discovered through a program such as BLAST. It will be interesting to determine whether or not the different segments that are discovered to share the same entropy (as discovered via the J-S divergence) do indeed share more biological features such as function or ancestry. Such an analysis is more involved and tedious, and some initial explorations have been initiated.