Sequence complexity and DNA curvature

Andrei Gabrielian, Alexander Bolshoy*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room 8N-807, Bethesda, MD 20894, USA

Abstract

A linguistic complexity measure was applied to the complete genomes of HIV-1, Escherichia coli, Bacillus subtilis, Haemophilus influenzae, Mycoplasma genitalium, and to long human and yeast genomic fragments. Complexity values averaged over entire genomic sequences were compared, as were predicted average values of intrinsic DNA curvature. We found that both the most curved and the least complex fragments are located preferentially in non-coding parts of the genome. Analysis of location of the most curved and the simplest regions in bacteria showed that the low-complexity segments are preferentially located in close proximity to the highly curved sequences, which are, in turn, placed from 100 to 200 bases upstream to the start of the nearest coding sequence. We conclude that the parallel analysis of sequence complexity and DNA curvature might provide important information about sequence–structure–function relationship in genomes. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Complexity; DNA curvature; Nucleotide composition; Linguistic analysis

1. Introduction

Due to rapid advances in genomic sequencing, new and effective methods that allow one to predict details of genome functional organization have become increasingly important. Statistical analysis of bio-sequences aims at determining sequence patterns which have unexpected distributions or frequencies of occurrence (Konopka, 1994). Associating these patterns with particular types of structure or function enables one to make further predictions of biologically important sites in genomes. In this study we would like to bring attention to one such method, linguistic complexity, and to its relationship to DNA intrinsic structural features.

Functional organization of the genome can be viewed as a multilevel, superimposed hierarchy. Applying appropriate parameters, complexity may help in discovering compartmentalization of the genome into functional domains. This approach was used in the study of structure and complexity of yeast chromosome III (King, 1993). Ussery et al. performed a similar analysis to examine the Escherichia coli genome (Proceedings of the 25th FEBS meeting). In these and other studies, complexity mapping was shown to reveal occurrence of genes, regulatory domains, and patterns required for higher order genome organization.

2. Sequence complexity

The concept of genomic sequences as texts has been introduced a long time ago. In searching for the functional code, they were studied as linear texts, applying linguistic (Brendel et al., 1986; Gelfand, 1993; Popov...
et al., 1996) or information (Sibbald et al., 1989; Schmitt and Herzel, 1997) methods. One of the fundamental characteristics of any text is its complexity, which may be calculated either according to the Shannon entropy or Kolmogorov complexity methods.

Quite a few different measures of sequence complexity have been used to characterize biological sequences, as there is no unequivocal definition of sequence complexity. Nevertheless, there tends to be agreement on the definition of a 'low-complexity' sequence, which usually has a highly repetitive structure. Linguistic complexity was introduced by Trifonov (1990), and used by Popov et al. (1996) to compare biological sequences with natural language texts, and by Bolshoy et al. (1997) to demonstrate that a weak pattern can be significantly enhanced by subset selection according to the sequence complexity criterion.

We refer the reader who wishes to learn more about mathematical background and theoretical justifications of sequence complexity measures to the appropriate section in (Konopka, 1994).

3. DNA curvature

Intrinsic DNA curvature was shown to be important in many biological processes, in particular in the interaction of DNA with DNA-binding proteins.

To reconstruct the shape of a DNA molecule from a sequence, the nearest neighbor models are commonly used (Ulanovsky and Trifonov, 1987; De Santis et al., 1988; Bolshoy et al., 1991; Goodsell and Dickerson, 1994). According to such a model, local variations in DNA shape are described by angles between successive base pairs. There are several sets of these angles, derived by different groups from electrophoretic mobility data, or X-ray and NMR structures of DNA. Evidently, the resulting shape is dependent on the angles that were used in the stepwise transformations. In our study on the intrinsic curvature of promoters (Gabrielian et al., submitted), as well as in this study, we applied different algorithms and scales to assess relative levels of curvature (or bendability) of selected fragments. It was found that selection of particular angles was not critical, as our results were confirmed by all methods.

4. Methods

4.1. Linguistic complexity

The linguistic complexity measure (Trifonov, 1990) was introduced as a measure of the 'vocabulary richness' of a text. When a nucleotide sequence is studied as a text written in the four-letter alphabet, the repetitiveness of such a text, that is, the extensive repetition of some k-grams (words), can be calculated, and served as a measure of sequence complexity. Thus, the more complex a DNA sequence, the richer is its oligonucleotide vocabulary, whereas repetitious sequences have relatively lower complexities. We have recently improved the original algorithm described in (Trifonov 1990) without changing the essence of the linguistic complexity (LC) approach.

The meaning of LC may be better understood by regarding the presentation of a sequence as a tree of all subsequences of the given sequence. The most complex sequences have maximally balanced trees, while the measure of imbalance or tree asymmetry serves as a complexity measure. The number of nodes at the tree level i is equal to the actual vocabulary size of words with the length i in a given sequence; the number of nodes in the most balanced tree, which corresponds to the most complex sequence of length N, at the tree level i is either 4i or N−i+1, whichever is smaller. Complexity (C) of a sequence fragment (with a length RW) can be directly calculated as the product of vocabulary-usage measures (Ui):

\[ C = U_1 \times U_2 \times \ldots \times U_i \times \ldots \times U_n \]  

(1)

Vocabulary usage (Ui) for oligomers of a given size i can be defined as the ratio of the actual vocabulary size of a given sequence to the maximal possible vocabulary size for a sequence of that length. For example, U2 for the sequence ACGGGAAGCTGATTCCA = 14/16, as it contains 14 of 16 possible different dinucleotides; U3 for the same sequence = 15/15, and U4 = 14/14. For the sequence ACACACACACACA, U1 = 1/2; U2 = 2/16 = 0.125, as it has a simple vocabulary of only two dinucleotides; U3 for this sequence = 2/15, k-tuples with k from two to W considered, while W depends on RW. For RW values less than 18, W is equal to 3; for RW less than 67, W is equal to 4; for RW<260, W = 5; for RW<1029, W = 6, and so on. The value of C provides a measure of sequence complexity in the convenient range 0<C<1 for various DNA sequence fragments of a given length. This novel formula ((1)) is different from the previous LC measure in two respects: in the way vocabulary usage Ui is calculated, and because i is not in the range of 2 to N−1 but only up to W. This new limitation on the range of Ui makes the algorithm substantially more effective without loss of power.

The new refined version of a linguistic complexity program has been developed and implemented under UNIX.

4.2. Clustered representation of extreme regions of a variance

For long nucleotide sequences, such as whole
HIV genome

Linguistic complexity in various windows

![Linguistic complexity graph](image)

Fig. 1. Influence of a sliding window parameter. The linguistic complexity (LC) measure was applied to the HIV-1 genome with various sliding windows, that is, 50, 100, 200, 400 and 800 bases from top to bottom. Dashed curves present raw complexity data, while thicker solid lines are used to mark the simplest and most complex regions. Complexity maps from top to bottom refer to sliding-window sizes in the ascending order. All plots are shown on the same scale, while the average linguistic complexity of HIV-1 is different for different sliding windows, that is, 0.49, 0.45, 0.42, 0.40 and 0.38 from top to bottom.

bacterial genomes or entire chromosomal sequences, we have applied a simplified, ‘boxlike’ presentation of complexity profiles. In Fig. 1, the distribution of linguistic complexity along the entire HIV-1 genome (GenBank accession number K0455) is shown, while several sliding windows were applied. Dashed lines correspond to the raw results of complexity calculations, while solid lines show the mean complexity of the whole sequence for regions of low variance and ‘boxes’ define the most outstanding areas. The number and position of ‘boxed’ areas are dependent on following parameters: imposed thresholds, a smoothing window size, and a minimal width of the extremal region necessary to form a ‘box’. Such a threshold-box presentation helps to bring viewers’ attention to the areas of a maximal variance; however, it does not replace a complete detailed graph.

4.3. DNA curvature calculations

Prediction of DNA curvature was done using three different programs, with three different scales.

4.3.1. CURVATURE program

This program calculates the three-dimensional path of DNA molecules and estimates the curvature of the path of the axis (Shigelman et al., 1993). Dinucleotide wedge angles of Bolshoy et al. (1991) and twist angles of Kabsch et al. (1982) were used for all calculations. In earlier work, CURVATURE was successfully used for realistic prediction of DNA shape (for example, Suka et al., 1993). The new version of this software runs under UNIX OS.

4.3.2. Helical asymmetry of bendability distribution

It has been shown (Gabrielian et al., 1996) that DNA sites known to be curved frequently possess periodic distribution of bendable and rigid fragments. Helical asymmetry is not a geometric parameter; it can be considered as a measure of ‘compatibility’ of DNA sequence with a curved conformation. Sequence-dependent bendability data were taken from (Gabrielian and Pongor, 1996). The ‘consensus’ scale was obtained by averaging the bendability scales of Satchwell et al. (1986) and Brukner et al. (1995) derived from statistics of nucleosome positioning and DNaseI cutting frequencies. The program for calculating helical asymmetry is available on the Web site http://www.icgeb.trieste.it/dna.
4.3.3. BEND program

The program utilizes a simplified approach for calculating the amplitude of curvature, without the prior calculation of three-dimensional structure of DNA. The set of roll, tilt, and twist angles of De Santis et al. (1990) was used in ANSI C modification of the publicly available version of the BEND program (Goodsell and Dickerson, 1994).

5. Results and discussion

5.1. Comparison of complexity of biological and random sequences

One of the most important tests of every complexity measure should be comparison between complexity of a natural text, which is supposed to carry redundancy properties (Trifonov, 1990; Konopka, 1990; Popov et al., 1996), and its randomized counterpart. To perform such a test, the above-mentioned sequence complexity measure LC was applied to the standard (reference) HIV-1 genome, to a shuffled HIV-1 genome, to a long E. coli fragment starting at position 19,901 and having a length of 60000 nt, and to its randomized counterpart. Randomization was performed as a first-order shuffling, i.e., with preservation of nucleotide composition only. Complexity was calculated in sliding windows with length 120 and 200 nt. The results are summarized in Table 1. As we expected, original genomic sequences show substantially lower average complexity than randomized sequences with the same composition (i.e., shuffled HIV-1 genome and shuffled E. coli fragment). We observed an increment of complexity as a result of sequence randomization approximately equal to one standard deviation.

We use the terms ‘random’ and ‘randomization’ as they are used in the literature. However, to avoid misunderstandings, we would like to emphasize that so-called random sequences are usually obtained by shuffling corresponding biological sequence, using a generator of pseudo-random numbers. Such a randomized sequence does not necessarily possess the highest possible complexity for a defined nucleotide composition.

The most complex sequence must be carefully designed, and thus will not be really random. It is therefore incorrect to assume that the most complex regions of the genome (of the HIV-1 genome, for example) are the most ‘random’ regions. Usually, it is much easier to provide a reasonable explanation for a nature of ‘low complexity’ of a certain region.

5.2. Complexity and a sliding window parameter

Table 1 shows that linguistic complexity is sensitive to window size. Fig. 1 illustrates that, presenting results of HIV-1 genome mapping with several sliding windows. To facilitate visualization of similarities and differences, the complexity graphs are plotted one under the other. The simplest and the most complex regions were found for every sliding window individually, and sometimes they are placed rather distinctly, depending on the window size. However, certain features show up in all graphs: for example, a simple site always appears close to nucleotide position 5000. The three upper curves, corresponding to smaller windows, depict a region around position 6100 as the simplest, and the region approximately at position 6900 as the most complex. The two lowest curves, corresponding to larger windows of 400 and 800 bases, present a ‘bird’s-eye view’ on an overall tendency: the gag-pol region is simpler than the env region. It is very tempting to correlate distribution of complexity along the HIV-1 genome with distribution of its variability, but this interesting issue is beyond the scope of this study.

5.3. Comparison of linguistic complexity of coding and non-coding regions

Fig. 2 shows histograms of linguistic complexity in coding and non-coding regions of E. coli. The original annotation of the E. coli genome (GenBank accession number U00096) was used to divide the whole sequence into coding and non-coding parts. Two datasets, consisting of 300 coding and non-coding fragments of E. coli, were compiled by randomly cutting 300 nt pieces off the corresponding parts of the genome. After dataset compilation, LC was calculated using a sliding window of 100 nt in every fragment. Histograms of averaged LC100 clearly show different distributions between coding and non-coding E. coli sequences. Although not every non-coding sequence is extremely simple, the fraction of simpler non-coding fragments is significantly larger than among coding sequences. Thus, our more detailed analysis of E. coli sequence is consistent with the results of Konopka (1990) and Trifonov (1990). A similar analysis was applied to datasets of human introns and exons with rather similar results (data not shown), although the
5.4. Comparison of linguistic complexity of genomes

The genomes of higher organisms are clearly distinguished from those of lower organisms by the amount of 'silent' non-coding DNA. In higher organisms only an estimated 5% of DNA encodes proteins. There are other biological messages coded at different levels of genomic organization (see, e.g., Trifonov, 1989, 1996). Nevertheless, the huge amounts of so-called 'junk DNA' remain *terra incognita*, and often it is desirable to separate this DNA from fragments potentially possessing biologic function. Sequence analysis methods, such as complexity calculations, may help in the search for hidden biological messages, by cumulating the statistics for known subsets of genomic sequences. Transcription regulation and genome packing are two examples of known functions that require presence of specific sequence patterns.

We applied the linguistic complexity algorithm to six entire genomes and to a long fragment of the human genome, and calculated mean complexity values. It is evident that the direct comparison of average complexity values or any other parameters, averaged over whole genomes, may serve only as the first crude approximation in this complex analysis. More accurate approaches might include the comparison of distributions of the specified feature (e.g., complexity or curvature) along the whole genomes or, ultimately, in the specific parts of genomes (coding, non-coding, promoter regions, etc.). Nevertheless, even the comparison of average values presented in Table 2 is interesting. It is no surprise that *E. coli* and *B. subtilis* have the highest average complexity, but it is certainly intriguing that the parasite *M. genitalium* has, on average, the least complex genome. Another interesting observation from Table 2 is that the *E. coli* genome with the highest average complexity has quite low average curvature, whereas the low-complexity *M. genitalium* genome has the highest mean curvature. We will discuss the relationship between curvature and sequence complexity in more detail below.

5.5. Comparison of linguistic complexity and curvature

Fig. 2 shows that *E. coli* intergenic regions have, on average, lower linguistic complexity than protein coding sequences (CDS). In our recent study of DNA cur-
Table 2
Linguistic complexity and curvature of selected genomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Average LC (Sliding window 100 nt)</th>
<th>Average curvature in nucleosome units (Arc siz 21 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>0.55 ± 0.09</td>
<td>0.18 ± 0.11</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>0.53 ± 0.09</td>
<td>0.21 ± 0.14</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>0.51 ± 0.10</td>
<td>0.23 ± 0.12</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (chromosome IV)</td>
<td>0.51 ± 0.11</td>
<td>0.23 ± 0.12</td>
</tr>
<tr>
<td>Homo sapiens (T-cell receptor)</td>
<td>0.46 ± 0.11</td>
<td>0.22 ± 0.12</td>
</tr>
<tr>
<td>Homo sapiens (introns)</td>
<td>0.46 ± 0.10</td>
<td>0.20 ± 0.12</td>
</tr>
<tr>
<td>Homo sapiens (exons)</td>
<td>0.46 ± 0.10</td>
<td>0.18 ± 0.12</td>
</tr>
<tr>
<td>HIV-1</td>
<td>0.45 ± 0.10</td>
<td>0.17 ± 0.10</td>
</tr>
<tr>
<td>Mycoplasma genitalium</td>
<td>0.41 ± 0.10</td>
<td>0.20 ± 0.12</td>
</tr>
</tbody>
</table>

* Genomic mean values of linguistic complexity (LC) and DNA curvature ± standard deviations are shown.

Curvature in promoter regions (Gabrielian et al., submitted), we showed that E. coli non-coding regions are predicted to be more curved than coding areas. Therefore, non-coding regions of E. coli are both simpler and more curved than coding sequences. Since both curvature and sequence complexity were demonstrated as related to functional organization of genome, we became interested in studying the relationship between them.

DNA curvature could be viewed as synchronous repetition of deviations from linearity. This repetitiveness is a consequence of the fact that curvature is a cooperative long-range feature (one can tell that curvature is a result of local bends, arranged according to the period of B-DNA). The repetitiveness of local bends does not necessarily mean repetitiveness of the underlying sequence; however, very curved DNA fragments are usually very repetitive. Well-known examples of intrinsically curved sequences are kinetoplast (Marini et al., 1982; Diekmann and Wang, 1985) and minisatellite DNA (Fitzgerald et al., 1994), which could also serve as good examples of low-complexity sequences because of phased repeating segments. On the other hand, the repeated sequences that result in curved DNA (i.e., those whose period is close to 10.5 bp) constitute only a minor part of all possible repeating sequences and their arrangements. Therefore, from a purely statistical point of view, one should not expect significant overlap between two sets of genomic sequences: the least complex and the most curved.

Following calculation of LC distribution with a sliding window parameter equal to 100 nt, we prepared, for every genome, complete sets of all non-overlapping pieces of 100 nt with mean linguistic complexity not exceeding 0.25. Further, we calculated curvature for the simplest pieces of all selected genomes (see Table 3). It became apparent that, for E. coli, B. subtilis and H. influenzae, the majority of low-complexity segments are curved above the average genome level, whereas this relationship does not hold for eukaryotic sequences. Fig. 3 illustrates this relationship for the B. subtilis genome, in which 68% of the fragments with low complexity (LC<0.25) are more curved than the mean curvature of 0.21 nucleosomal units. In Fig. 3,

Table 3
Correlation between complexity and curvature for the non-overlapping fragments with the lowest complexity

<table>
<thead>
<tr>
<th>Genome</th>
<th>Number of fragments (size 100 nt) with LC &lt; 0.25</th>
<th>With curvature greater than average and with LC &lt; 0.25</th>
<th>With curvature greater than average + STD with LC &lt; 0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>361 from 46390</td>
<td>225 (62%)</td>
<td>115 (32%)</td>
</tr>
<tr>
<td>B. subtilis</td>
<td>769 from 42147</td>
<td>524 (68%)</td>
<td>268 (35%)</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>348 from 18300</td>
<td>229 (66%)</td>
<td>114 (33%)</td>
</tr>
<tr>
<td>S. cerevisiae (chromosome IV)</td>
<td>500 from 15205</td>
<td>239 (48%)</td>
<td>88 (18%)</td>
</tr>
<tr>
<td>Homo sapiens (T-cell receptor)</td>
<td>210 from 7152</td>
<td>101 (48%)</td>
<td>51 (24%)</td>
</tr>
<tr>
<td>M. genitalium</td>
<td>200 from 5800</td>
<td>113 (57%)</td>
<td>62 (31%)</td>
</tr>
</tbody>
</table>

* The genomes and long T-cell receptor sequences were mapped by the linguistic complexity measure with the sliding window equal to 100 nt. Further, the complexity values were sorted, and, among the lowest values, those were chosen that provided non-overlapping fragments.
Fig. 3. Complexity of the most curved fragments and curvature of the least complex fragments of the *Bacillus subtilis* genome. All non-overlapping 100 nt fragments possessing average values of linguistic complexity lower than 0.25 were selected from the *Bacillus subtilis* genome. Independently, all non-overlapping fragments possessing average values of DNA curvature greater than 0.35 nucleosome units were selected. Mean curvature values of all selected fragments were plotted against mean complexity values of the same fragments. Average genome complexity (0.53) and average curvature (0.21) values are shown as well.

we also present data illustrating the complexity of the most curved fragments. These results show that the overlap between two sequence sets is quite small: there are only few segments that are both very curved and have minimal complexity. Nevertheless, there is a pronounced tendency for curved sequences to be less complex than the average genomic sequence. Interestingly, low-complexity sequences also have a marked tendency to be more curved than the ‘average’ sequence from the same genome.

5.6. Location of the most curved and least complex sequences in genomes

Since we have found a statistically proven tendency for non-coding regions to contain both curved sequences and low-complexity sequences, we attempted to find out whether there are any preferred positions for those ‘extreme’ sequences and whether there are any preferred distances between the most curved and least complex sequences. To answer these questions, we calculated distances from each curved and each low-complexity fragment to the closest start of coding region (CDS), annotated in the corresponding GenBank file. Histograms of these distances for three entire genomes and chromosome IV of *Saccharomyces cerevisiae* are presented in Fig. 4. For comparison, we calculated distances from randomly distributed points along corresponding genomes. The distribution of these distances appeared close to normal, with no preferred positioning in respect to CDS. Contrary to that, the histograms for actual extreme fragments show that the most curved as well as the least complex fragments are found preferentially in the range from −100 to −300 nt upstream of the start of translation, a preference that is statistically significant for *E. coli*, *B. subtilis*, and *M. genitalium*. The histogram for curved fragments in chromosome IV of *Saccharomyces cerevisiae* (Fig. 4A) has the least pronounced asymmetry, although the histogram for low-complexity fragments is very similar to that of above-mentioned bacterial genomes. Interestingly, the *Haemophilus influenzae* genome does not exhibit any significant fluctuation from the random distribution, neither for the most curved nor for the low-complexity fragments. To verify the results of curvature calculations that were obtained with the CURVATURE program (Shigelman et al., 1993), we applied two other popular programs:
Fig. 4. Histograms of distances from the most curved and least complex fragments to the nearest coding sequences. Non-overlapping 100 nt fragments, covering 3% of the entire genome, were selected according to the LC or DNA curvature criteria. Distances from the most curved fragment to the closest CDS are shown as diamonds, distances from low-complexity fragments as circles. Distances from randomly distributed points to the closest CDS are shown as solid line. Negative values correspond to fragments located upstream to an appropriate CDS. A: *S. cerevisiae* (chromosome IV); B: *E. coli*; C: *B. subtilis*; D: *M. genitalium*. 
Bacillus subtilis

M. genitalium

Fig. 4 (continued)
Fig. 5. Histograms of distances from the most curved fragments, calculated by three different curvature/bendability programs, to the nearest coding sequences. Three datasets of the most curved 100 nt fragments, covering 3% of the entire genome, were prepared using the programs CURVATURE (red asterisks), BEND (crosses in magenta), and HELIX (purple empty circles). Filled circles correspond to the dataset of simpler fragments (as in Fig. 4). Distances from randomly distributed points to the closest CDS are shown as solid green line. A: *H. influenzae*; B: *M. genitalium*. 
HELIX (Gabrielian and Pongor, 1996) with the consensus scale and BEND (Goodsell and Dickerson, 1994) with the parameters of De Santis et al. (1990) to all genomic sequences under study. The results for two genomes are presented in the Fig. 5. Results produced with different algorithms are consistent: all methods show specific distribution for three bacterial genomes but not for H. influenzae. At this stage, we have no reasonable hypothesis to explain this striking difference.

The unexpected conservation in distances of low-complexity and curved sequences from the nearest start of coding sequence might shed new light on hidden rules of genome structure-function relationships. Oligonucleotide repeats were found responsible for DNA adopting a variety of conformations (Cox and Mirkin, 1997). Some of these conformations might be easily distorted, some might offer unique determinants for DNA-protein contacts and DNA packing. The mapping of potential alternative DNA structures in genomes might help in revealing the common principles of genome structure and functioning.

We plan to expand this work studying more genomes, performing detailed analysis of the types of repeats present in low-complexity regions and investigating relative positioning of 'extreme' segments in respect to known biologically important genomic sites.

Acknowledgements

The authors would like to express their gratitude to Drs. A. Konopka, E.N. Trifonov, and D. Landsman for their invaluable suggestions during the preparation of the manuscript. Dr. Birgit An der Lan provided excellent editorial assistance. A.B. was supported by the Danish National Research Foundation and NCBI Scientific Visitors Program at the National Library of Medicine, NIH.

References

Gabrielian, Landsman and Bolshoy, submitted.
Suka, N., Shinozaka, Y., Saijoh, Y., Saijoh, H., Ohtomo, K.,


