Geometric approach to string analysis for biosequence classification

Boris Brimkov1,*

1Computational & Applied Mathematics, Rice University, Houston, TX 77005, USA
http://www.rice.edu

Summary

Tools that effectively analyze and compare sequences are of great importance in various areas of applied computational research, especially in the framework of molecular biology. In the present paper, we introduce simple geometric criteria based on the notion of string linearity and use them to compare DNA sequences of various organisms, as well as to distinguish them from random sequences. Several other theoretical and statistical results are outlined as well.

Our experiments reveal a substantial difference between biosequences and random sequences – the former having much higher deviation from linearity than the latter – as well as a general trend of increasing deviation from linearity between primitive and biologically complex organisms.

1 Introduction

The theory of words studies the structural properties of strings composed from letters of a given alphabet, and provides algorithms for solving diverse problems defined on strings. Among the most important motivations of the discipline is its relevance to computational biology, and more precisely, to the automated analysis of biosequences. This includes a great variety of problems whose portrayal is beyond the purposes of the present paper. Some avenues of the ongoing research are surveyed in [1, 2, 3]. In particular, an important task is identifying certain patterns, motifs, or biologically meaningful features in a given biosequence.

Typically, the considered problems are approached using combinatorial techniques such as combinatorial pattern matching and combinatorics on words. In this paper we instead use a geometric approach in an attempt to address questions that are important for understanding biological evolution. A few other geometric representations of biosequences have been considered (e.g., [4, 5]), often to aid with visualization or exhibit certain features in a sequence.

On the other hand, a number of past studies have attempted to address by quantitative means the question of what distinguishes biosequences from random sequences, and have met with varying levels of success. While by its very nature such a goal has been found quite elusive [6], there is substantial evidence in support of the argument that biosequences feature properties that are typical of random sequences (for example, near-total incompressibility [7]). Thus,

*To whom correspondence should be addressed. Email: boris.brimkov@rice.edu

biosequences are regarded as “slightly edited random sequences” [8], and modern proteins are believed to be “memorized” ancestral random polypeptides which have been slightly modified by the evolutionary selection process in order to optimize their stability under specific physiological conditions [9]. Biosequences are hardly distinguishable from their random permutations by many criteria, although the latter are clearly incongruous with living organisms [10, 11, 12]. While this may seem quite obvious from a biological point of view, there have also been numerous computational arguments that support this claim. For example, in [13] Pande et al. present results of mapping some protein sequences onto so-called Brownian bridges, which revealed a certain deviation from randomness. In another study, by estimating the differential entropy and context-free grammar complexity, Weiss et al. have shown that the complexity of large sets of non-homologous proteins is lower than the complexity of the corresponding sets of random strings by approximately 1% [8]. As a first major result of the present work, we introduce simple geometric criteria by which biosequences very strongly differ from random sequences of the same length. In view of the above-mentioned 1% difference demonstrated in [8], by “very strongly” we refer to differences in the order of several hundred percent, registered for 25 biosequences compared to random sequences over the same alphabet and length.

Furthermore, provided the widely adopted postulates of the theory of evolution and in view of the available theoretical and experimental results, it is natural to conjecture that in the evolutionary process of organisms from primitive to biologically complex, their corresponding biosequences have been evolving from random or close to random toward ones that feature increasing deviation from randomness. As a second major result, our experiments based on the introduced measures confirm this expectation. That is also in accordance with results suggesting that biosequences of proteins which are close in the genome are more similar than those of proteins far apart in the genome.

To this end, we use a discrete geometric approach. Given a string $s = s_1 \ldots s_m$ over an alphabet $X$ with $|X| = n$, we define an ordered set $L(s) = q_0 \ldots q_m$ of points which form a discrete monotone path in $\mathbb{Z}^n$. We then define the deviation of such a monotone path from linearity, and state some basic properties related to this notion.

After introducing several deviation measures, we apply them to the biosequences of a set of organisms that stand at different levels of the evolution scale (i.e., from primitive organisms such as microbes, through plants and reptiles, up to mammals). We also compare all these with random sequences of the same length. The obtained results demonstrate a substantial difference between biosequences and random sequences – the former being much further from linearity than the latter – as well as a general trend of increasing deviation from linearity between primitive and biologically complex organisms. Results of some other related experiments are outlined as well. For example, we compare the deviation from linearity of a biosequence to the maximal possible deviation over all of its permutations, in order to see if biosequences still have the potential for higher non-linearity after further evolution, or if their existing structures are already close to the maximum possible in terms of deviation from linearity.

The paper is organized as follows. In the next section we introduce some technical notions and notations, including ones from the theory of words. In Section 3 we introduce the notions of string linearity and deviation from linearity, and study several related properties. In Sections 4 and 5, we present our experimental results and offer a short discussion. We conclude with final
remarks and open questions in Section 6.

2 Definitions and notations

By $|X|$ we denote the cardinality of set $X$ and by $xy$ the straight line segment with endpoints $x$ and $y$. By $d(x,y)$ we denote the Euclidean distance between points $x$ and $y$, and by $d(x,Y)$ the distance between point $x$ and set $Y$, i.e.,

$$d(x,Y) = \inf_{y \in Y} \{d(x,y)\}.$$ 

Given a list $T$ of nonnegative real numbers $t_1 \ldots t_\ell$ (not all of which equal 0), a normalization of $T$ is obtained by multiplying each value in $T$ by $\frac{100}{t_{\max}}$ where $t_{\max} = \max_{1 \leq i \leq \ell} \{t_i\}$.

In the literature, the terms word, sequence, and string are often used interchangeably. A sequence is often defined in mathematics as a function whose domain consists of a set of consecutive integers, and a string over $X$, where $X$ is a finite set, is often defined as a finite sequence $s$ of elements from $X$ ($X$ is also sometimes called the alphabet). The term word is frequently used as an abstraction of the other two terms. In biology, the prevalent term is biosequence; biosequences are built from the four letters A, T, C, G, and have finite length.

The theory of words is a central topic in theoretical computer science; see, for example, [14, 15]. Below, we recall a few basic notions and fix some denotations to be used in this paper.

In string $s = s_1 \ldots s_m$ over set $X$, $s_i$ is the $i$th term of $s$ ($1 \leq i \leq m$), which is some element of $X$. The number of elements in $s$ is called the length of $s$ and denoted $|s|$. If $|s| = 0$, we say that $s$ is the empty string, denoted by $\lambda$. We denote $r \geq 1$ consecutive repetitions of term $x$ in string $s$ by $x^r$.

If $s$ and $t$ are two strings, the string consisting of $s$ followed by $t$, written $st$, is called the concatenation of $s$ and $t$. A substring of a string $s$ is obtained by selecting some or all consecutive elements of $s$. More formally, a string $v$ is a substring of the string $s$ if there are strings $u$ and $w$ such that $s = uvw$ (where we may have $u = \lambda$ or $v = \lambda$).

Following the terminology introduced in [16, 9], given a string $s$, a subsequence of $s$ is any string $u$ which can be obtained by removing from $s$ one or more, not necessarily consecutive terms.

3 String geometrization

To present our approach, we conform to a digital geometry setting in which the considerations take place in the grid cell model. In this model, the regular orthogonal grid defines a partition of $\mathbb{R}^n$ into $n$-dimensional hypercubes (e.g., unit squares for $n = 2$, unit cubes for $n = 3$) also called $n$-cells or voxels. The $n$-cells are centered at the grid points and their edges are parallel to the coordinate axes. See [17] for more details.
Let $s = s_1 \ldots s_m$ be a string on an alphabet $X = \{x_1, \ldots, x_n\}$. We inductively construct an ordered set $L(s)$ of points $q_0, \ldots, q_m$ corresponding to string $s$ as follows.

We set $q_0$ to be the origin of the Cartesian coordinate system. Let $q_i = (q_{i,1}, \ldots, q_{i,n})$ be the $i^{th}$ element of $L$ for $0 \leq i < m$. If $s_{i+1} = x_j$ for some $1 \leq j \leq n$, then we set $q_{i+1} = (q_{i,1}, \ldots, q_{i,j} + 1, \ldots, q_{i,n})$. Thus, we obtain a monotone discrete path $L(s)$, in which the coordinates of a point are pairwise greater than or equal to the corresponding coordinates of any preceding point. Figure 1, left, gives an example of a string over the alphabet $\{x, y\}$ and its corresponding monotone discrete path.

Having such a discrete path constructed, one can study its geometric and combinatorial properties, which in turn can provide useful information about the original string $s$. Clearly, the properties and characteristics of monotone discrete paths, not necessarily representing strings, could be interesting in their own right.

Let $o$ be the origin and $p = (p_1, \ldots, p_n)$ be a point in $\mathbb{Z}^n$. Denote by $\mathbb{H}$ the set of all monotone discrete paths between $o$ and $p$. It is easy to see that

$$|\mathbb{H}| = \frac{(p_1 + \ldots + p_n)!}{p_1! \ldots p_n!}.$$ 

Each path $H \in \mathbb{H}$ consists of $1 + \sum_{i=1}^{n} p_i$ points, with initial point $o$ and terminal point $p$. If for every point $h$ in $H$, the voxel centered around $h$ intersects the line segment $\overline{op}$, we call $H$ a linear path. Accordingly, we call a string linear if its corresponding monotone path is linear.

It is easy to see that given a line segment $\overline{op}$, there is at least one linear path from $o$ to $p$; also, if $H$ is a linear path from $o$ to $p$, then $d(h, \overline{op}) \leq \sqrt{n} \quad \forall h \in H$.

Next we define some string characteristics that are instrumental to the experimental studies presented in the subsequent sections.

Let $s$ be a string and $L(s) = q_0 \ldots q_m$ be its corresponding monotone path. We define the maximum deviation of $s$ from linearity as

$$mdv(s) = \max_{i=0}^{m} \{d(q_i, \overline{q_0q_m})\},$$
and average deviation of $s$ from linearity as

$$adv(s) = \frac{\sum_{i=0}^{m} d(q_i, q_{0q_{m}})}{m + 1}.$$ 

**Remark 1** Note that when $n = 4$ (which is the case for biosequences), the $adv$ and $mdv$ of a linear string are at most 1.

The third characteristic of a string $s$ will be called the **number of local maxima** of $s$ and denoted $nlm(s)$. Formally, by a local maximum we mean a point $q_i \in L$ for which $d(q_i, q_{0q_{m}})$ is greater than $d(q_{i-1}, q_{0q_{m}})$ and $d(q_{i+1}, q_{0q_{m}})$. However, regarding the usual applications of extrema of discrete functions, in particular in view of our own purposes, counting all such maxima does not seem to be very relevant. Instead, local maxima can be counted only if they “stand out” compared to other, “indistinguishable” local extrema, which differ very little from neighboring points. Thus, we adopt the notion of number of local maxima as “method dependent.” Specifically, our choice of method is the one provided by [18]. The $nlm$ measure may be useful to distinguish between strings which have the same $adv$ and $mdv$, as illustrated in Fig. 1, right.

### 3.1 Maximal string deviation over all permutations

As we will see in the following sections, biosequences deviate from linearity significantly more than random sequences. It is also interesting to compare the deviation of a given biosequence to the maximal possible deviation over all of its permutations. This will reveal whether the evolutionary process has driven biosequences to reach a deviation from linearity that is close to the maximal possible over the same content of nucleotide bases. In this section we develop a technical tool for such a study. To this end, we solve a combinatorial problem which could also be of independent interest.

We will refer to the largest $mdv$ attained by a permutation of string $s$ as $maxmdv(s)$. That is,

$$maxmdv(s) = \max \{mdv(s') : s' \text{ is a permutation of } s\}.$$ 

We have the following theorem.

**Theorem 1** Given a string $s$ over an alphabet $X = \{x_1, \ldots, x_n\}$ in which letter $x_i$ appears $a_i$ times, $1 \leq i \leq n$,

$$maxmdv(s) = \max_\mathbb{P} \left\{ \sqrt{\frac{\sum_{i \in N} a_i^2 (\sum_{j \in Z} a_j^2)}{\sum_{k=1}^n a_k^2}} \right\},$$

(1)

where $\mathbb{P}$ is the set of all partitions of $\{1, 2, \ldots, n\}$ into two disjoint, nonempty sets $N$ and $Z$. If $P = \{N^*, Z^*\} \in \mathbb{P}$ is a partition for which $maxmdv(s)$ is attained, then for every permutation $s'$ of $s$ whose corresponding discrete monotone path contains the point $p = (p_1, p_2, \ldots, p_n)$, where $p_i = a_i$ if $i \in N^*$ and $p_i = 0$ if $i \in Z^*$, $mdv(s') = maxmdv(s)$.
entire genomes. We assumed that excerpts of this size are sufficiently large to minimize the effect of the non-stationarity of genomic sequences, though future experiments could select samples from the genome-scale repository and type of organisms we consider are typical of comparative analysis studies in molecular biology (see, e.g., [22]). We took the biosequences from the genome-scale repository and browser Ensembl Genomes, which is managed by the European Bioinformatics Institute. For all strings ending with $x_1$, and all strings ending with $x_1^{a_1}$.

4 Deviation from linearity of random sequences and biosequences: Experimental study

4.1 General description of experimental procedures

The notion of string linearity furnishes an easily implementable tool to compare the biosequences of various organisms. It is reasonable to conjecture that biologically complex organisms have highly structured DNA whose corresponding monotone path strongly deviates from a straight line, while primitive organisms have less structured DNA, whose corresponding monotone path is closer to a straight line. Moreover, a completely random sequence over the alphabet $\{A, T, C, G\}$ has no structure, and therefore its corresponding monotone path can be expected to be much closer to a straight line.

To test this hypothesis, we compare the deviation from linearity of the biosequences of 25 organisms with varying biological complexity, as well as that of random sequences. The number and type of organisms we consider are typical of comparative analysis studies in molecular biology (see, e.g., [22]). We took the biosequences from the genome-scale repository and browser Ensembl Genomes, which is managed by the European Bioinformatics Institute. For each organism, we processed relatively short substrings and subsequences of DNA in FASTA format, selected randomly from an excerpt of the genome containing about two million letters. We assumed that excerpts of this size are sufficiently large to minimize the effect of the non-stationarity of genomic sequences, though future experiments could select samples from the entire genomes.
We first studied the effects of string size on the proposed linearity measures and then selected a suitable string size for more extensive experiments. Our results in the following section demonstrate that a comparison of samples of a certain reasonable size will classify organisms in the same relative order as a comparison of much larger samples. Moreover, fixing the length of the samples facilitates their comparison, since the (absolute) deviation from linearity of a string is generally dependent on its length. Since the genomes of different organisms have different lengths, and some genomes are still uncharted or studied only partially, a direct comparison of the linearity of entire genomes is not feasible.

Recall also that we distinguish between substrings and subsequences of a given string \( s \), the former being segments of consecutive terms while the latter being ordered subsets of not necessarily consecutive terms of \( s \). Typically, experimental research involving biosequences is based on processing families of substrings rather than subsequences. Only recently, Apostolico and Cunial attempted to assess (“perhaps for the first time,” as these authors believe) the structure and randomness of polypeptides in terms of subsequences satisfying certain conditions [16, 9]. As the results obtained therein seem interesting and promising, we performed all our experiments both on substrings and subsequences. The similarities and differences within both frameworks are presented and discussed in the following sections.

4.2 Effects of substring and subsequence length

We investigated how the length of a biosequence affects its deviation from linearity, and ascertained that deviation increases with the size of the string. However, the rate of increase seems to be independent from the type of organism, and therefore an organism’s deviation from linearity relative to the deviation of other organisms is independent of the length of the biosequences, as long as the length is constant across organisms. These claims are supported by Figure 2, which shows the absolute and normalized maximum and average deviations from linearity of different organisms measured for substrings and subsequences of increasing length.

The top four graphs display the \( \text{adv} \) and \( \text{mdv} \) computed for substrings, and the bottom four graphs display the \( \text{adv} \) and \( \text{mdv} \) computed for subsequences. The left four graphs display the absolute \( \text{adv} \) and \( \text{mdv} \) and the right four graphs display the normalized \( \text{adv} \) and \( \text{mdv} \) for the graphs on their left. Note that the graphs in the two columns are essentially the same, where the normalized graphs in the right column are the result of “pulling up” the left sides of the graphs in the left column.

For all graphs, the substrings and subsequences are taken randomly from the 26 sources listed in Table 1. For the top four graphs, the corresponding linearity measures were computed for substrings of length \( 10,000 \times k \) for \( 1 \leq k \leq 20 \); for each of these lengths, the linearity measures were computed for 200 different substrings taken randomly from each of the 26 sources, and the average values of the 200 trials were plotted. For the bottom four graphs, the corresponding linearity measures were computed for subsequences of length \( 10,000 \times (2k-1) \) for \( 1 \leq k \leq 10 \); for each of these lengths, the linearity measures were computed for 100 different subsequences taken randomly for each of the 26 sources, and the average values of the 100 trials were plotted.\(^1\)

\(^1\)Fewer lengths and trials were used for subsequences because adding more “resolution” to the almost com-
Figure 2: Absolute and normalized $adv$ and $mdv$ measured for substrings and subsequences of increasing length. These graphs show that $adv$ and $mdv$ are essentially independent of length, since the lines remain in the same relative positions as length varies. Note also that the lowest line in each graph, (which stands out noticeably when the samples are substrings) corresponds to the random sequence.
From Figure 2, it can be seen that \( \text{adv} \) and \( \text{mdv} \) measured from subsequences are more independent of length than \( \text{adv} \) and \( \text{mdv} \) measured from substrings, since there is less crossing between lines in the bottom four graphs. However, even when measured from substrings, \( \text{adv} \) and \( \text{mdv} \) are principally independent of length, since for any of the lengths examined the organisms are more or less in the same relative position compared to the other organisms.

Clearly, as the length of substrings and subsequences approaches 0, the measures of deviation from linearity will approach 0, and will be slightly more unstable and unreliable. From the normalized graphs, we notice that the slight initial fluctuations first disappear around substrings and subsequences of length 50,000. For this reason, we carried out our further experiments (which involve more trials) with substrings and subsequences of this length.

4.3 Description of computational procedure

We used version R2011a of Matlab to carry out our computations. The built-in Matlab functions \texttt{randi} and \texttt{randseq} were applied to make random selections and generate random sequences with independent symbols and uniform distribution; additionally, we used a function called \texttt{peakfinder}, provided by [18], to find local extrema. Our computational process can be broken up into the following components: selection of samples, computation of linearity measures, and compilation and normalization of data.

4.4 Selection of samples

To select a random substring \( b_1 \ldots b_m \) from a larger string \( a_1 \ldots a_n \), we pick a random integer \( i \) from the interval \([1, n - m + 1]\) using Matlab’s \texttt{randi} function; then, we set \( b_j = a_{i+j-1} \) for \( 1 \leq j \leq m \). In this manner, we select substrings from each of the organisms and store the substrings, which all have length \( m \), in an array. We also store a random sequence of length \( m \) over the alphabet \( \{A, T, C, G\} \), generated using Matlab’s \texttt{randseq} function, in the array.

Similarly, to select a random subsequence \( b_1 \ldots b_m \) from a larger string \( a_1 \ldots a_n \), we first obtain a list \( C \) of \( m \) distinct random integers ranging from 1 to \( n \). We sort the elements of this list in increasing order, and obtain \( C = c_1 \ldots c_m \) where \( 1 \leq c_i < c_{i+1} \leq n \) for \( 1 \leq i < m \). We then set \( b_i = a_{c_i} \) for \( 1 \leq i \leq m \). In this manner, we select subsequences of length \( m \) from each of the organisms and store them in an array, along with a random sequence of length \( m \).

Due to the motivation given earlier that fluctuations in the linearity measures disappear in samples of length 50,000, we selected substrings and subsequences of length 50,000 from the larger excerpts of genomes.

4.5 Computation of linearity measures

In order to calculate the maximum and average deviation from linearity for a given string \( S \) of length \( m \), we first count the number \( x_l \) of letters \( l \in \{A, T, C, G\} \) in \( S \) (where \( \sum x_l = m \)).
Let $S_t$ be the substring composed of the first $t$ letters in $S$, and $t_l$ be the number of letters $l \in \{A,T,C,G\}$ in the string $S_t$ (where $\sum t_l = t$). We compute the distance $D(S_t)$ from point $(t_A, t_T, t_C, t_G)$ to the line passing through the origin and $(x_A, x_T, x_C, x_G)$ for $1 \leq t \leq m$ with the formula

$$D(S_t) = \sqrt{(x_A\mu - t_A)^2 + (x_T\mu - t_T)^2 + (x_C\mu - t_C)^2 + (x_G\mu - t_G)^2}$$

where $\mu = \frac{x_A t_A + x_T t_T + x_C t_C + x_G t_G}{x_A^2 + x_T^2 + x_C^2 + x_G^2}$.

We then store the distances $D(S_t)$, $1 \leq t \leq m$ in a list, and find the average and maximum values of this list using elementary techniques, as well as the number of local maxima using the `peakfinder` function.

### 4.6 Compilation and normalization of data

After we have obtained an array of 26 samples (25 biosequences and one random sequence), we calculate the linearity measures $adv$, $mdv$, and $nlm$ for each sample in the array and end up with a $26 \times 3$ array. We repeat this procedure 1,000 times with different random samples, find the average and standard deviation over the 1,000 trials, and normalize the average for each linearity measure in order to better see the relationships between organisms. Thus, we obtain a table of the normalized means which allows us to easily compare organisms based on the three linearity criteria. We repeat the whole procedure twice – once where the samples are substrings and once where they are subsequences. These results are presented in Table 1.

We also computed two measures of statistical dispersion, the coefficient of variation (CV) and the quartile coefficient of variation (QCV). The CV of a data sample is the standard deviation divided by the mean, and the QCV is $(Q_3 - Q_1)/(Q_3 + Q_1)$ where $Q_1$ and $Q_3$ are the first and third quartiles, respectively. The table displaying the CV and QCV between trials is omitted due to space restrictions, but we summarize its contents in Section 5.

### 5 Discussion

In this section we provide further details about our experimental work and discuss the obtained results. We comment on results which are obvious to our unarmed eye, hoping that other interesting conclusions could also be drawn by those with higher expertise in biological sciences.

#### 5.1 General observations and comments

The first two columns of Table 1 give scientific and common names for the organisms which we have examined. The specific strains of Chlamydia, Tuberculosis, Gingivalis, and Streptococcus are Nigg, CCDC5180, W83, and ND03, respectively. All the DNA we processed was from the
<table>
<thead>
<tr>
<th>Scientific Name</th>
<th>Common Name</th>
<th>Substrings</th>
<th>Subsequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>adv</td>
<td>mdv</td>
</tr>
<tr>
<td>Random Sequence</td>
<td>Random</td>
<td>12.8</td>
<td>12.8</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>Human</td>
<td>88.9</td>
<td>89.9</td>
</tr>
<tr>
<td>H. neanderthalensis</td>
<td>Neanderthal</td>
<td>85.5</td>
<td>85.1</td>
</tr>
<tr>
<td>G. gorilla</td>
<td>Gorilla</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>P. troglodytes</td>
<td>Chimp</td>
<td>81.2</td>
<td>81.0</td>
</tr>
<tr>
<td>C. familiaris</td>
<td>Dog</td>
<td>71.8</td>
<td>74.3</td>
</tr>
<tr>
<td>G. gallus</td>
<td>Chicken</td>
<td>59.1</td>
<td>57.2</td>
</tr>
<tr>
<td>C. jacchus</td>
<td>Marmoset</td>
<td>55.0</td>
<td>55.8</td>
</tr>
<tr>
<td>R. norvegicus</td>
<td>Rat</td>
<td>68.2</td>
<td>68.0</td>
</tr>
<tr>
<td>M. musculus</td>
<td>Mouse</td>
<td>62.5</td>
<td>64.9</td>
</tr>
<tr>
<td>O. anatinus</td>
<td>Platypus</td>
<td>41.2</td>
<td>41.7</td>
</tr>
<tr>
<td>A. carolinensis</td>
<td>Lizard</td>
<td>31.6</td>
<td>32.3</td>
</tr>
<tr>
<td>D. rerio</td>
<td>Zebrafish</td>
<td>59.6</td>
<td>58.5</td>
</tr>
<tr>
<td>O. latipes</td>
<td>Medaka fish</td>
<td>46.6</td>
<td>47.5</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Fruit fly</td>
<td>49.1</td>
<td>49.3</td>
</tr>
<tr>
<td>C. intestinalis</td>
<td>Sea squirt</td>
<td>32.8</td>
<td>32.2</td>
</tr>
<tr>
<td>C. elegans</td>
<td>Nematode</td>
<td>51.6</td>
<td>52.4</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>Yeast</td>
<td>39.4</td>
<td>39.8</td>
</tr>
<tr>
<td>C. muridarum</td>
<td>Chlamydia</td>
<td>29.7</td>
<td>29.3</td>
</tr>
<tr>
<td>M. tuberculosis</td>
<td>Tuberculosis</td>
<td>33.8</td>
<td>34.0</td>
</tr>
<tr>
<td>P. gingivalis</td>
<td>Gingivalis</td>
<td>50.0</td>
<td>48.1</td>
</tr>
<tr>
<td>S. thermophilus</td>
<td>Streptococcus</td>
<td>37.2</td>
<td>36.5</td>
</tr>
<tr>
<td>O. sativa</td>
<td>Rice</td>
<td>73.8</td>
<td>76.9</td>
</tr>
<tr>
<td>Z. mays</td>
<td>Corn</td>
<td>76.1</td>
<td>80.3</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>Cress</td>
<td>44.8</td>
<td>44.4</td>
</tr>
<tr>
<td>G. max</td>
<td>Soybean</td>
<td>52.2</td>
<td>53.0</td>
</tr>
</tbody>
</table>

Maximum pre-normalized value: 554.5 1100.8 21.9 739.5 1563.6 22.4
first chromosomes of the organisms, except for the fruit fly and the yeast, where the DNA was taken from chromosomes 2L and 4, respectively.

Columns 3-5 of Table 1 show the results of our experiments when the $adv$, $mdv$, and $nlm$ were measured for substrings taken from the 25 biosequences and one random sequence. The last three columns show the results of computing the measures on subsequences instead of substrings. All measures were computed on samples of length 50,000 and are the average of 1,000 trials.

Note that for most species, the normalized average $adv$ and $mdv$ measured from substrings differ by less than 1%; only for four organisms this difference is more than 2%, but no more than 4.2%. When measured from subsequences, the difference between $adv$ and $mdv$ is not always small; although for most species it is less than 2%, for 6 organisms the difference is more than 4%, and for a few – as much as 10% or 20%. Also, we obtained that the CV and QCV of $adv$ and $mdv$ differ by .02 or less for most organisms, and never by more than 0.06; this is true of substring as well as subsequence samples. These similarities between $adv$ and $mdv$ are also supported by Figure 2, which shows that the $adv$ graphs are nearly identical in appearance to the $mdv$ graphs – where the difference is slightly more noticeable in the cases where $adv$ and $mdv$ are computed over subsequences.

Typically, distributions whose coefficients of variation are less than one are regarded as low-variance. The variance between the 1,000 trials was less than 1 for all organisms and all linearity measures – with the variance of trials sampling subsequences being several times lower than those sampling substrings. In particular, the range over different organisms of CV and QCV for $adv$ measured from substrings was [0.24, 0.59] and [0.16, 0.41], respectively. For subsequences these ranges were [0.04, 0.25] and [0.03, 0.17], respectively.

We suppose that when subsequences rather than substrings are analyzed, some of the structure of the DNA is lost and the results are not as precise. However, as mentioned above, subsequences are slightly more independent of the length of the sample and have lower variance. When substrings are analyzed the entire structure of the DNA is preserved and taken into consideration, so linearity measures computed over subsequences seem to better reflect the complexity of the organism. As can be seen from the fifth and eighth columns of Table 1, the $nlm$ measure is useful for distinguishing between random sequences and biosequences, but not as good for classifying organisms in terms of biological complexity.

5.2 Distinction between biosequences and random sequences

Our first important conclusion is the distinction between the linearity of biosequences and random sequences. All of our experiments show that biosequences have a higher average and maximum deviation from linearity than random sequences. When substrings are analyzed, this difference is quite large. In particular, in Table 1, the normalized $adv$ and $mdv$ for random sequences are both 12.8, whereas the normalized $adv$ and $mdv$ for the organism with the smallest deviation are 29.3 and 29.7, respectively. When subsequences are analyzed, the difference is not as substantial, but random sequences still have a smaller average and maximum deviation from linearity than any of the organisms considered.

This conclusion is also supported by Figure 2, where in the top four graphs (which are computed over substrings), the line positioned visibly below the others is the one representing the random sequence. In the bottom four graphs (which are computed over subsequences), the lowest line is again the one representing the random sequence, but the difference is not as significant.

The criterion of \( nlm \) presents an even more sizeable difference between random sequences and biosequences. Again, the difference is larger when substrings are considered: the normalized \( nlm \) for random sequences is 100.0, whereas the normalized \( nlm \) for the organism with the largest number of local maxima is 39.8. Computed over subsequences, the normalized \( nlm \) for random sequences is again 100.0, and the largest number of local maxima for an organism is 96.2.

### 5.3 Gradient between primitive and biologically complex organisms

Our experiments also support the hypothesis that the sequences of primitive organisms are closer in linearity to random sequences than the sequences of biologically complex organisms. As most primitive organisms, we consider bacteria and microscopic organisms; we consider plants the next most evolved organisms, followed by fish, reptiles, and other egg-laying vertebrates. Finally, we consider mammals and primates as organisms at the top of the evolutionary ladder. We expected that the graded change in the magnitude of deviation from linearity of different organisms would be in accordance with the aforementioned classification of their biological complexity. Indeed, our experiments support this expectation.

In particular, when measured over substrings, the Human, Neanderthal, Gorilla, and Chimpanzee have the highest \( adv \) and \( mdv \); the bacterium Chlamydia has the smallest \( adv \) and \( mdv \) after the random sequence. The other organisms with the lowest deviations from linearity are two other bacteria, the yeast, sea squirt, and lizard. In the mid-low range are organisms like the fruit fly, medaka fish, and soybean plant, and in the mid-high range are organisms like the zebrafish, chicken, and mouse. Another interesting observation is that primates have a lower \( nlm \) than other animals. Aside from that, the \( nlm \) measure is not as useful as the other measures for classifying organisms by biological complexity; a weak inverse relationship exists between \( nlm \) and biological complexity, especially when \( nlm \) is measured from substrings.

When subsequences are considered some of the structure of the DNA is inadvertently lost, so the results are not as consistent with expectation. However, many similar trends can be seen in the last three columns of Table 1 as well. Primates generally have the highest deviation from linearity, while bacteria like Tuberculosis and Gingivalis have among the lowest deviations from linearity. Other organisms of medium biological complexity span the intermediate values of the linearity measures.

In all considerations, some anomalies and incongruences with expectation are manifested. For example, when measured over substrings, the \( adv \) and \( mdv \) of rice and corn are relatively high – higher, for example, than the \( adv \) and \( mdv \) of the mouse and rat.
Table 2: The ratio $\frac{\text{maxmdv}}{\text{mdv}}$ computed for substrings of length $\{1, \ldots, 10\} \times 10^5$.

<table>
<thead>
<tr>
<th>Organism</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>17.3</td>
<td>15.0</td>
<td>12.6</td>
<td>13.2</td>
<td>15.6</td>
<td>10.5</td>
<td>13.1</td>
<td>11.4</td>
<td>14.3</td>
<td>12.2</td>
</tr>
<tr>
<td>Gorilla</td>
<td>12.7</td>
<td>20.1</td>
<td>15.2</td>
<td>12.0</td>
<td>11.8</td>
<td>13.5</td>
<td>13.0</td>
<td>11.2</td>
<td>10.6</td>
<td>10.9</td>
</tr>
<tr>
<td>Mouse</td>
<td>19.9</td>
<td>28.1</td>
<td>30.8</td>
<td>40.2</td>
<td>37.9</td>
<td>41.7</td>
<td>45.2</td>
<td>57.1</td>
<td>46.7</td>
<td>50.0</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>28.2</td>
<td>24.9</td>
<td>27.7</td>
<td>32.7</td>
<td>31.7</td>
<td>28.6</td>
<td>32.8</td>
<td>37.0</td>
<td>26.6</td>
<td>40.8</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>44.6</td>
<td>64.8</td>
<td>66.5</td>
<td>84.6</td>
<td>88.1</td>
<td>96.5</td>
<td>105.1</td>
<td>116.3</td>
<td>122.5</td>
<td>144.9</td>
</tr>
<tr>
<td>Random Sequence</td>
<td>126.7</td>
<td>164.3</td>
<td>205.7</td>
<td>260.2</td>
<td>280.7</td>
<td>304.7</td>
<td>325.1</td>
<td>336.0</td>
<td>351.2</td>
<td>397.0</td>
</tr>
</tbody>
</table>

5.4 How far are biosequences from maximal string deviation over all permutations

As already mentioned, it would be interesting to know how much a biosequence $s$ deviates from linearity compared to the maximum deviation over all of its permutations (recall that the latter quantity is labeled $\text{maxmdv}(s)$). In view of the trend demonstrated in the previous sections, we want to find out if biosequences still have the potential for higher deviation from linearity after further evolution, or if their existing structures are already close to the maximum possible in terms of deviation from linearity.

Using the tools developed in Section 3.1 and especially Corollary 1, we computed the ratio $\frac{\text{maxmdv}}{\text{mdv}}$ for a number of biosequences. Taking the deviation from linearity of a biosequence as a measure of its structure, we found that all organisms still have the capacity for further structural increase, although the currently existing structures are not vastly smaller than the maximum possible. Naturally, the computed ratios differ between organisms, with primitive organisms having a greater ratio than complex organisms. Table 3 gives $\frac{\text{maxmdv}}{\text{mdv}}$ of several organisms computed for substrings of length $\{1, \ldots, 10\} \times 10^5$.

As this particular investigation is not the primary focus of the paper, we only include a cursory overview of these results, in order to convey a sense of the discrepancy between $\text{mdv}$ and $\text{maxmdv}$. Further and more detailed investigation will be the subject of future work.

6 Concluding remarks

In this paper we introduced a geometric approach for string analysis based on the notions of string linearity and deviation from linearity. Our experiments showed that, unlike some other criteria, ours strongly separate random sequences from biosequences, as well as primitive from biologically complex organisms. These results are in accordance with certain earlier interpretations that biosequences have been evolving towards energy minimization in physical terms, as well as of lowering their information complexity [9, 13].

As the proposed quantitative measures seem to be quite robust and reliable in practice, important future tasks are seen in performing systematic extensive experiments on a larger set of biosequences, their interpretation and deeper analysis from a biological point of view, and comparison with results obtained by other approaches. In addition to a more extensive study of
the general trends exhibited in the present work, possible future tasks can pursue understanding the meaning and functions (from a biological point of view) of biosequence locations where deviation from linearity achieves local maxima or minima. Certain anomalies from the general trends featured by the experiments could also be addressed.

Our study was meant to be a pilot one rather than exhaustive. It was intended to provide initial tests of the proposed approach and to serve as a prelude to a more complete interdisciplinary investigation, involving specialists who are better equipped to carry out large-scale experiments, interpret the results, and apply them – for instance, in the construction of phylogenetic trees. As a final remark, we believe that the introduced approach could be applied to other areas where string comparison is relevant, like in comparing the structure of natural languages or diverse encoding schemes.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1450681. Some of the presented results have been reported at the 8th International Conference on Practical Applications of Computational Biology & Bioinformatics, held in Salamanca, June 2014 [23].

References


