Long-range correlations in prokaryotic chromosomes

D Swati
Department of Physics, MMV, Banaras Hindu University Varanasi 221005 India

Received 11 April 2011; revised 08 November 2011

One of the fascinating properties of the DNA sequences of prokaryotic and eukaryotic chromosomes is that they possess long-range order. Computational methods like spectral analysis, mutual information and DNA random walks have been used to probe long-range order via long-range correlations. This work attempts to show the advantage of using the Information Theoretic measure of mutual information for this purpose. A number \( \mu \) is found which indicates the existence of long-range order. \( \mu \) is the ratio between the value of mutual information function between two nucleotides of a DNA sequence separated by a large distance of 100 kilobases to the value expected from a randomized sequence of the same DNA. It is found that in spite of the constant shuffling of nucleotides due to insertion, deletion, inversion and recombination that occur during evolution, the chromosomal structure of prokaryotes is not always mosaic. While all archael chromosomes show mosaic structure and lack long-range order, a sizable fraction of the bacterial chromosomes do possess long-range order. A statistical multivariate analysis has been done to find which of the physical variables like genome size or GC\% affects the organization of the chromosome or correlates with the long-range order. The existence of long-range order in bacterial chromosomes could be directly correlated to the degree of gene strand bias shown by it. Firmicutes which have low GC content also have pronounced strand bias and show long-range correlations. It is observed that the occurrence of long-range order in bacteria is independent of genome size, but depends on its GC content and gene strand bias.

Keywords: Mutual information, Gene strand bias, Replication, Long-range order, Repeats, GC content, Prokaryotic chromosomes

The DNA molecule carries information since it is a repository of the genetic code. It may also be interpreted as a language written with four symbols, the nucleotides adenine (A), thymine (T), guanine (G) and cytosine (C). Though it appears as a quasi-random sequence, nature has hidden functions within it in the form of coding regions, promoters, regulators and transcription sites to cite a few examples. These contextually relevant regions may be compared with words or phrases in any language with DNA being written in an alphabet of four symbols \{A, T, G, C\}. The coding theorems of Information Theory were propounded by Claude Shannon\(^1\) to understand how to transmit messages across a noisy channel and which were later used to parse text, mining it to find hidden or encrypted messages. Similar to information content of a language, information content of the DNA sequence can also be analyzed and lead to a better understanding of the sequence itself.

This work is an attempt to probe, if long-range correlations exist in DNA sequence of prokaryotes and resolve an existing controversy. There is a school of thought which contends that requirements of biological functioning generate long-range correlations as evidenced by results of spectral analysis\(^2,3\) or results of using mutual information\(^4-6\), a measure from Information Theory. The other school contends that because of frequent shuffling within DNA sequence due to various evolutionary processes like insertion, deletion, inversion and recombination, structure of prokaryotic chromosomes is necessarily mosaic and, therefore, there can be no long-range order in prokaryotic chromosomes.

The idea is to find, if the occurrence of any nucleotide at a position on DNA sequence is influenced by the occurrence of another nucleotide at a given distance away from it. In other words, we are probing for correlations between occurrences of nucleotides at different positions of DNA sequence. There are several different ways of defining correlation functions in the literature\(^8\) and these vary from discipline to discipline. As Li\(^7\) states in a seminal review on this topic that there should be unanimity in definition of correlation function as well as in definition of “long-range” before long-range correlations can be discussed. He suggests the use of objective and system independent measure of mutual information \( M(k) \)\(^7\) which is discussed in ‘Methodology section’.

---

E-mail: swati@bhu.ac.in, swatid@gmail.com
Tel.: 919450823732

Abbreviations: CDS, coding sequences, GSB, gene strand bias, nt, nucleotides, Mb, megabase. SSR, short sequence repeats.
Mutual information function has been used to find coding and non-coding regions of DNA and repeat structures\textsuperscript{8,9} to identify possible homology between proteins\textsuperscript{10}, to compare and type bacterial strains\textsuperscript{11} and to evaluate gene expression data\textsuperscript{12}, to name a few of the biological applications of this versatile tool. In this work, the value of mutual information function at large base separation of 100 kb or 0.1 Mb is used to test long-range order in prokaryotic chromosomes. We name this indicator as $Mu$, which is the ratio of mutual information function at base separation of 0.1 Mb to the corresponding value for the randomized sequence of the same DNA molecule. This base separation is chosen because existence of long-range correlations across 100,000 bases is rare and unusual in prokaryotic chromosomes.

The DNA sequences are complex systems because they have structures at different length scales. From triplet or 3-base periodicity in coding regions as well as to promoters and regulatory sequences of few tens of bases each and coding sequences of lengths varying from (100-1000) bases, there are sub-sequences of varying length scales for each required biological function. The existence of structures at very large scale lengths results in long-range order\textsuperscript{13}.

Replication, transcription and regulation of gene expression are all processes that are instrumental in shaping the structure of genome. For instance, modality of replication is seen to correlate positively with the preferential loading of coding sequences (CDS) on the leading strand. Some bacteria, for example the Firmicutes show a marked preference for loading CDS on the leading strand which is known as gene strand bias\textsuperscript{14}. This is distinct from compositional bias or change in GC content between two strands of DNA molecule, although this is also a consequence of replication process. The two strands replicate differently and are, therefore, subject to different mutational pressure. Mrazek and Karlin\textsuperscript{15} have shown that since only one strand is involved in transcription, it is more prone to mutation.

The concept of strand asymmetry is used in two different contexts in the literature. The term usually refers to compositional bias, which points to a probable violation of Chargaff’s Second Parity rule in the individual strands of DNA molecule as a result of above-mentioned different mutation rates. The other usage refers to the preferential occurrence of CDS on one strand, usually the coding or sense strand, which is referred to as gene strand bias (GSB) in this work.

This preference for the leading strand is explained as possibly due to transcription and replication being co-directed to avoid head-on collision of DNA polymerase with RNA polymerase. This has been conclusively shown in a recent genome-wide study of wild and mutant strains of Bacillus subtilis using microarrays\textsuperscript{16}. Because both DNA and RNA polymerase proceed along the 5’ to 3’ direction, transcripts coded on lagging strand may get aborted due to head-on collision, whereas transcripts coded on leading strand may get stalled by co-oriented collisions and start again after a brief pause.

French\textsuperscript{17} in his landmark experiment has shown that only lagging strand transcripts are likely to get aborted by collision of RNA polymerase with DNA polymerase. There is a selective pressure for GSB to occur because of the possibility that truncated transcripts may get translated to toxic proteins which are deleterious for the growth of cells. Rocha and Danchin\textsuperscript{18} have shown that it is the essential genes like those involved in transcription and translation, which are predominantly piled on the leading strand and not highly expressed genes. More than 75% of B. subtilis essential genes and that of other Firmicutes show the same preference. E. coli and other Gamma proteo-bacteria do not show this pronounced bias. E. coli, for example has gene strand bias of $\sim 55\%$\textsuperscript{14,16}.

Earlier studies using spectral analysis or power spectra methods\textsuperscript{2,3,8} have shown that all bacterial chromosomes may show $\langle 1/f^\alpha \rangle$ behavior of power spectra with frequency $f$ with long-range order being indicated by value of $\alpha$ approximately equal to 1. These authors assumed that the bacteria having simple chromosomal structure with almost 90% plus of coding sequences would show long-range order. A study carried out by the present author using fast Fourier transforms (FFT) to calculate power spectra of a few prokaryotes like B. halodurans and C. tetani has shown similar behaviour with value of $\alpha$ obtained being close to 1.

Mutual Information is used as a tool to test this hypothesis of possible universal long-range order in prokaryotic chomosomes, since it could reveal hidden correlations, if any, and a large set of bacteria and as large a set as is possible with fully sequenced archaea is chosen for this purpose. One may expect that mutual information function, which is proportional to the weighted sum over squares of all correlation functions\textsuperscript{3} will be large in cases of pronounced GSB, because in this case the CDS are located co-linearly with respect to direction of transcription on one strand. For this
purpose, we have studied 291 prokaryotic chromosomes, 45 of them being archaean. The results show that Mu can be used as an indicator of long-range order. A large number of bacteria and archaean studied show no long-range order, but a sizeable fraction of bacteria (92 out of 246) studied do show long-range correlations. Statistical analysis is applied to the data set and has yielded an expected trend that size of genomes is directly proportional to their genomic GC content, as observed in the comprehensive review by Bentley and Parkhill. Statistical analysis also shows that a high value of Mu is associated with high AT%.

An experiment is devised where the real genomic sequence plus two artificially generated sequences of *Thermoanaerobacter tengcongensis* are analyzed using the code for computing Mutual Information. *T. tengcongensis* has 87% of its CDS on leading strand and also a long array of tandem repeats. It is found that coding sequences occurring co-directed for transcription and replication on one strand and not the repeats that give rise to the large value of Mu observed for this genome.

It is found that the indicator Mu, defined from the mutual information function which is a measure from Information Theory is an appropriate parameter for probing long-range correlations in prokaryotic chromosomes and may be used to quantify and categorize long-range order in their structure. There have been various studies of the replication mechanism in archaean, which catalogue the existence of possible multiple origins of replication. The existence of multiple origins of replication may be the cause of mosaic structure of archaean chromosomes which leads to a lack of long-range order in the archaean.

**Materials and Methodology**

The genomic data were downloaded from Genbank: http://www.ncbi.nlm.nih.gov/. GC content (the percentage of guanine and cytosine taken together relative to the total number of nucleotides in a given DNA sequence) was taken from the Comparative Genomics website. http://www2.unil.ch/comparativegenometrics/

The study involved 246 bacteria and 45 archaean from different phyla. The details of distribution over phyla are given Table 1. A large variation over both GC content and genome size was deliberately chosen as shown in Figs A1 and A2.

The variation of GC content was from 16.6% for *Candidatus Carsonella ruddi* to 74.9% for *Anaeromyxobacter dehalogenans*. The variation of genome size was from 0.16 Mb for *Candidatus Carsonella ruddi* to 9.1 Mb for

---

**Table 1—Distribution of prokaryotes with respect to phylum**

<table>
<thead>
<tr>
<th>Name of Phylum</th>
<th>Number of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Archaea</strong></td>
<td></td>
</tr>
<tr>
<td>Crenarchaeota</td>
<td>14</td>
</tr>
<tr>
<td>Euryarchaeota</td>
<td>29</td>
</tr>
<tr>
<td>Korarchaeota</td>
<td>01</td>
</tr>
<tr>
<td>Nanoarchaeota</td>
<td>01</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
</tr>
<tr>
<td><strong>Proteobacteria</strong></td>
<td></td>
</tr>
<tr>
<td>Alpha</td>
<td>29</td>
</tr>
<tr>
<td>Beta</td>
<td>09</td>
</tr>
<tr>
<td>Gamma</td>
<td>61</td>
</tr>
<tr>
<td>Delta</td>
<td>06</td>
</tr>
<tr>
<td>Epsilon</td>
<td>09</td>
</tr>
<tr>
<td><strong>Firmicutes</strong></td>
<td></td>
</tr>
<tr>
<td>Tenericutes</td>
<td>11</td>
</tr>
<tr>
<td><strong>Actinobacteria</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Cyanobacteria</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Spirochaetes</strong></td>
<td></td>
</tr>
<tr>
<td>Chlamydiae</td>
<td>09</td>
</tr>
<tr>
<td>Bacteroidetes-Chlorobi</td>
<td>08</td>
</tr>
<tr>
<td>Deinococcus-Thermus</td>
<td>05</td>
</tr>
<tr>
<td>Chloroflexi</td>
<td>01</td>
</tr>
<tr>
<td>Thermotogae</td>
<td>01</td>
</tr>
<tr>
<td>Aquificae</td>
<td>01</td>
</tr>
<tr>
<td>Fusobacteria</td>
<td>01</td>
</tr>
<tr>
<td>Total</td>
<td>246</td>
</tr>
</tbody>
</table>
Bradyrhizobium japonicum. The Table giving details of the data set will be made available by the author on request.

The two most important quantities related to information are entropy (self information contained in a random variable) and mutual information (the amount of information in common between two random variables). Applying Information Theory to DNA sequences we can define positional entropy $H(i)$ and $H(j)$ for the symbol (nucleotide) $\alpha$ occurring at position $i$ and symbol $\beta$ occurring at position $j$ independently and joint entropy $H(i, j)$ for symbol $\alpha$ occurring at position $i$ simultaneously with $\beta$ occurring at position $j$ as given below in Eqs (1), (2) and (3).

\[
H(i) = - \sum_{\alpha} p_\alpha(i) \ln p_\alpha(i) \quad \ldots (1)
\]

\[
H(j) = - \sum_{\beta} p_\beta(j) \ln p_\beta(j) \quad \ldots (2)
\]

\[
H(i, j) = - \sum_{\alpha} \sum_{\beta} p(\alpha; i | \beta; j) \ln p(\alpha; i | \beta; j) \quad \ldots (3)
\]

The mutual information $M(i, j)$ is defined as:

\[
M(i, j) = H(i) + H(j) - H(i, j)
\]

With $j = i + k$, we get the definition of Mutual Information function as given by Shannon as:

\[
M(k) = \sum_{\alpha} \sum_{\beta} p_{\alpha\beta}(k) \ln \left[ \frac{p_{\alpha\beta}(k)}{p_\alpha p_\beta} \right] \quad \ldots (4)
\]

where $p_\alpha = N_\alpha/N$, $p_\beta = N_\beta/N$ can be taken to be the independent frequency of occurrence of the two nucleotides (bases) $\alpha$ and $\alpha$ and $p_{\alpha\beta}$ ($k$) is the conditional probability of a base $\alpha$ (A, T, G, C) existing at site $i$ with base $\beta$ (A, T, G, C) at a site $j$. Here $n$ is the number of symbols in the DNA symbol string, also called the alphabet, $N_\alpha$ and $N_\beta$ are the number of nucleotides of type $\alpha$ and $\beta$ respectively and $N$ is the total number of nucleotides or size of the genome. The conventional units of $M(k)$ are bits and nats, if the logarithm is taken to base 2 and base $e$ (natural logarithm), respectively.

If $p_{\alpha\beta}(k) > p_\alpha p_\beta$ then the occurrence of bases $\alpha$ and $\beta$ is said to be positively correlated and the value of $M(k)$ as defined above is positive.

And if $p_{\alpha\beta}(k) = p_\alpha p_\beta$ then $M(k)$ is zero and it indicates that there is no correlation between occurrence of base $\alpha$ at site $i$ and base $\beta$ at site $k$ bases downstream from $i$. $M(k)$ is a measure of the total amount of correlation between all possible nucleotides at a separation $k$ apart. The value of $M(k)$ has to be corrected for finite-size of the sequence. The correction is:

\[
\sim (n-1)/2N
\]

where $n$ is the size of alphabet and $N$ is the size of genome, as defined above. The value for $n$ is 4 for DNA sequences and 20 for aminoacid sequences.

There is a minimum correlation in-built into the DNA sequence. This arises from the fact that when the first nucleotide is put at a site, say the first position on the sequence, then automatically that site or position is blocked for any other nucleotide. This is called Volume Exclusion Effect. If $Vex$ is taken as the symbol for Volume Exclusion Effect, then

\[
Vex \sim 1/N.
\]

$Mu$ is defined as:

\[
Mu = M(k = 100000)/Vex
\]

Before we can apply the information technique to the DNA symbol string, we have to digitize it. There are several ways to do this, but we have used the method of Projection Operators by Voss, whereby each nucleotide is assigned a number, say 1, 2, 3 and 4, wherever it occurs. With this procedure, the DNA symbol string becomes a string of four numbers and thereby amenable to computation.

The code has been written in C++ and can be used both for linear and circular DNA sequences. The input parameters are the word size or the size of the $n$-mer between which conditional probability $p_{\alpha\beta}$ is to be computed and stored to compute $M(k)$; $k$, the lag or separation between base positions and the initial position and the final position on the DNA sequence for computation of $M(k)$. The output is in the form of an array which gives the value of lag ($k$), the 16 terms of Eqn. (4) and their sum $M(k)$. This program may also be used to study frame-shifts and di-nucleotide di-nucleotide correlations. The statistical analysis and the histograms in Figs A1 and A2 as well as the scatter-plot generated in Fig. 2 were done using R-language.

\[
\text{SWATI: LONG-RANGE CORRELATIONS IN PROKARYOTIC CHROMOSOMES}
\]
Results and Discussion

In Fig. 1, M(k) for the genome of *Thermoanaerobacter tengcongensis* is plotted against the base separation k. The red curve shows the oscillations of the function due to the three base periodicity typical of coding regions with the superposed curve (in blue) being a plot of the data smoothened over a window size 3 to suppress the 3-periodicity. Two broad peaks at k = 11 and 21 are seen in the smoothened curve and they are signature of short sequence repeats (SSR) of the genome. The black dotted line shows the M(k) for the randomized sequence of the same genome with an almost constant value corresponding to ~Vex.

It is found that for all chromosomes whether prokaryotic or eukaryotic, there exists a maximum length for correlations within the coding regions. This length can be easily read from the semi-log plot of mutual information function M(k) versus base separation k. This length is of the order of k = 1000 nt or 1 kilobase (kb) scale. The value of M(k) is large within this region due to the occurrence of codons and the related codon usage bias which give rise to a three base periodicity. This 3-periodicity is seen in all cases studied and is a universal phenomenon among prokaryotes and small eukaryotes. A distribution of lengths for different sized CDS leads to a new length scale and it is found to be of the order of 10 kb. M(k) continues decreasing till base separation equals about 10 kb and for larger values of k tends towards the still lower value expected for the asymptotic value of mutual information at k = 100 kb or 0.1 Mb.

This is the usual behaviour of M (k), but we find that for Firmicutes (*T. tengcongensis* is a member of this phylum) and in a few atypical cases like for the Gamma proteo-bacterium *Xylella fastidiosa*, the *Ehrlichias* and *Bartonellas* of the phylum Alpha proteo-bacteria and the *Borrelias* of phylum Spirochaetes, the value of M(k) remains roughly constant up to k = 0.1 Mb and is unusually large. It is much more than ten-times the M(k = 0.1 Mb) value for a randomized sequence of the genome, which is approximately equal to Vex.

The reason for such long correlations is the existence of another scale in the chromosomal structure and this is what leads to long-range order. The strand bias because of a large number of co-oriented CDS occurring on the same strand leads to correlations over a length scale equal to several CDS. This length may differ in value for each chromosome, but is seen to be of the order of 0.1 Mb or 100 kb. M(k) as given in Eq. (4) is a sum over sixteen terms, each of which is proportional to the weighted square of the correlation function between a given pair of bases at two different positions along the DNA sequence, k bases apart; k may vary from 1 to any desired value. The maximum value of k is chosen to be 100 kb for reasons discussed in the ‘Methodology section’.

The variation of *Mu*, the indicator of asymptotic correlations with respect to volume exclusion factor Vex, genome size *Gs* (in Mb) and variation of GC content is studied and the results are shown as a scatter plot over all the parameters in Fig. 2, which gives their mutual dependences. *Mu* shows no clear dependence on genome size *Gs*, but does show a significant distribution with GC%. Table 2 shows that the Pearson correlation coefficient between *Mu* and GC% is clearly negative for bacteria, whereas for archaea it is equivocal since the p-value is high. Hence multi-variate analysis of their mutual dependence is done and the results are shown in Table 3. The result is a gamma distribution for variation of *Mu* with GC%. We have observed a positive correlation of genome size *Gs* to GC%. The larger bacteria tend to be GC rich, while the smaller endosymbionts and obligatory intercellular bacteria tend to be GC-poor. This agrees with the study of Bentley and Parkhill.

Figure 3 shows the variation of *Mu* for the prokaryotic data set with different symbols to distinguish the variation for bacteria from that for the archaea. As is clearly seen from this figure, some bacteria and all archaea studied show no long-range correlations. but, few of the bacteria (92 out of a total
of 246, or ~37%) show signature of long-range correlations. If \( M(k = 0.1 \text{ Mb}) \) is corrected for finite size effect and it comes out to be more than an order of magnitude larger than the \( M(k = 0.1 \text{ Mb}) \) for the corresponding randomized sequence, we take this as a first level indicator of long range order. Then we take the ratio as the definition for \( \mu \) as given in the ‘Methodology section’. Figure 3 shows that a large fraction of the total data set has value of \( \mu \) below 20, which is the value of \( \mu \) for the line shown. This value of indicator is more stringent and has been chosen to comply with the minimum value for any of the Firmicutes (a prototype of which is \( B. \) subtilis) that have ~80% of their CDS on the leading strand and which (with very few exceptions) have high values of \( \mu \). Firmicutes use two DNA polymerases Pol-C and DnaE polymerase with Pol-C being used for leading strand replication and DnaE for replication of the lagging strand\(^{26}\). These low GC, gram-positive bacteria that have evolved to have the majority of their CDS on the leading strand\(^{14}\) have their \( \mu \) values from 20 to a few hundreds.

**Bacteria**

The Firmicutes and Gamma proteo-bacteria have been studied extensively. Examples of Firmicutes are Bacilli, Clostridia, Lactobacilli, Staphylococci and Streptococci, among which are the pathogenic Bacilli

---

### Table 2—Results of Pearson’s correlation test for \( \mu \), Vex, Gs with GC%

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Data set</th>
<th>p-Value</th>
<th>95% Confidence interval</th>
<th>Correlation coefficient</th>
<th>Comments on correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>Prokarya</td>
<td>2.36e-07</td>
<td>-0.3758, -0.1627</td>
<td>-0.2726</td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>6.15e-08</td>
<td>-0.3948, -0.1645</td>
<td>-0.2837</td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>Archaea</td>
<td>0.05*</td>
<td>0.0054, 0.544</td>
<td>0.2984</td>
<td>* p-Value high</td>
</tr>
<tr>
<td>Vex</td>
<td>Prokarya</td>
<td>3.83e-08</td>
<td>-0.4158, -0.2080</td>
<td>-0.3154</td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>3.92e-09</td>
<td>-0.4679, -0.2506</td>
<td>-0.3643</td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>Archaea</td>
<td>0.849*</td>
<td>-0.286, 0.343</td>
<td>0.032</td>
<td>* p-Value high</td>
</tr>
<tr>
<td>Gs</td>
<td>Prokarya</td>
<td>5.21e-12</td>
<td>0.2879, 0.4833</td>
<td>0.3900</td>
<td>Positive</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>5.932e-13</td>
<td>0.3310, 0.5338</td>
<td>0.4380</td>
<td>Positive</td>
</tr>
<tr>
<td></td>
<td>Archaea</td>
<td>0.994*</td>
<td>-0.2946, 0.2924</td>
<td>-0.0360</td>
<td>* p-Value high</td>
</tr>
</tbody>
</table>

* p-values for Archaeal data are too high in each case.
like *B. anthracis* and other highly pathogenic bacteria like *C. perfringens* and *C. tetani*. For instance *C. tetani*, not only has more than 85% CDS on the leading strand, but a large number of gene duplications and variable number of tandem repeats or VNTR as well. A large number of VNTR is also a feature of the chromosome of the plant pathogen *Xylella fastidiosa* and bovine pathogen *Ehrlichia ruminantium* that causes heart-water disease. *Borrelia burgdorferi* chromosome has a large number of CDS (~70%) on one strand, but no tandem repeat structure.

The marine bacterium *Prochlorococcus marinus* which is ecologically important in the carbon cycle of our planet has been studied by Rocap *et al.*. They performed comparative genomic study of the strains of this phytoplankton and divided into two clades-HL (high light adaptative) that can only grow near the surface of sea and LL (low light adaptative), whose ecological niche is at depths of the ocean in the range of 100 m. Their photosynthetic proteins not only differ from that of the *Synechococcus* spp., but among each other. Swati has shown that by using mutual information as a genome signature, one can clearly distinguish between the different strains of HL-type and strains of LL-type. The strains MIT 9303 and MIT 9313, both of the LL-type have large value of \( \text{Mu} \) and both have GC% of 50. Whereas all other five strains of HL-type examined show low values of \( \text{Mu} \).

The p-value for intercepts is less than 2.0e-16 in all cases, except the last row, where it is 2.6e-13.

\[ f(x) = \frac{\lambda^t x^{t-1} \exp(-\lambda x)}{\lambda(t)} \]

where \( x > 0 \). The mean of Gamma Distribution is (\( \lambda t \)) and dispersion is (\( \lambda t^2 \))

In this study, \( x \) is the base separation \( k \) and \( \lambda \) and \( t \) are parameters given in Table 3 above.

The Mutual Information function \( M(k) \) is given by

\[ M(k) = \frac{\lambda^t k^{t-1} \exp(-\lambda k)}{\lambda(t)} \]

where \( \lambda = 0.0003 \) and \( t = 0.0022 \) for bacteria.

![Figure 3](image_url) — Plot of \( \text{Mu} \) versus GC% [Bacteria, black dots; Archaea, red squares, \( \text{Mu} = 20 \)- Gray line]

The results of generalized linear regression model using Gamma distribution \( \text{Mu}, \text{Vex}, \text{Gs} \) with GC% are shown in Table 3 below.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Data set</th>
<th>Intercept ± Std. error</th>
<th>Parameter ± Std. error</th>
<th>p-Value for parameter</th>
<th>Dispersion parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Mu} )</td>
<td>Prokarya</td>
<td>0.0022 ± 0.0004</td>
<td>0.00002 ± 0.000002</td>
<td>1.29e-06</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>0.0022 ± 0.0004</td>
<td>0.00034 ± 0.00003</td>
<td>3.24e-06</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>Archaea</td>
<td>-0.0037 ± 0.0002</td>
<td>+0.0002 ± 0.0032</td>
<td>0.046*</td>
<td>0.047</td>
</tr>
<tr>
<td>( \text{Vex} )</td>
<td>Prokarya</td>
<td>0.0193 ± 0.0006</td>
<td>0.0145 ± 0.0030</td>
<td>3.93e-09</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>0.0195 ± 0.0006</td>
<td>0.0059 ± 0.0001</td>
<td>1.2e-08</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>Archaea</td>
<td>0.0219 ± 0.001</td>
<td>-0.0003 ± 0.0001</td>
<td>0.849*</td>
<td>0.052</td>
</tr>
<tr>
<td>( \text{Gs} )</td>
<td>Prokarya</td>
<td>0.0389 ± 0.0007</td>
<td>-0.0014 ± 0.0002</td>
<td>1.27e-13</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>0.0273 ± 0.0008</td>
<td>-0.0016 ± 0.0002</td>
<td>7.89e-13</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>Archaea</td>
<td>0.0022 ± 0.0002</td>
<td>0.00058 ± 0.0008</td>
<td>0.994*</td>
<td>0.052</td>
</tr>
</tbody>
</table>

*\( \text{p} \)-value for Archaeal data is too high in each case.

The Gamma distribution is given by: \( f(x) = \frac{\lambda^t x^{t-1} \exp(-\lambda x)}{\lambda(t)} \)

where \( \lambda > 0 \). The mean of Gamma Distribution is (\( \lambda t \)) and dispersion is (\( \lambda t^2 \))

In this study, \( x \) is the base separation \( k \) and \( \lambda \) and \( t \) are parameters given in Table 3 above.

The Mutual Information function \( M(k) \) is given by

\[ M(k) = \frac{\lambda^t k^{t-1} \exp(-\lambda k)}{\lambda(t)} \]

where \( \lambda = 0.0003 \) and \( t = 0.0022 \) for bacteria.
computational time for finding long-range correlations for 1 Mb or more for these large chromosomes is prohibitive, the reason for their possessing long-range order could not be investigated.

Archaea

All the archaea have values of $Mu$ less than 20 and thus show a lack of long-range order. Their genome sequences show mosaic structure and the clear strand asymmetry due to replication is lacking. It is conjectured that this is due to the possible multiple origins of replications in archaea. A large number of origins of replication would lead to mosaic structure. S. solfataricus is reported to have three origins of replication. The recent comprehensive review by Barry and Bell shows that though the replication machinery of the archaea resembles the eukaryotic one on a simpler scale, but there is no common prototype. Nikolaou and Almirantis have correlated compositional bias with strand asymmetry due to replication and transcription. They also observed that the archaeal genomes sequenced up to 2005 show neither pronounced compositional bias nor strand asymmetry. The study presented in this paper corroborates this, since a lack of pronounced strand asymmetry is directly correlated with low values of $Mu$.

An in silico experiment

A computational case study has been devised to observe the effect of large number of tandem repeats on the organization of a chromosome and long-range order. Thermoanaerobacter tengcongensis, a Firmicute is taken as the genome to be studied. It is an extreme thermophile with a genome size of 2.689445 Mb and a GC% of 37.6. It has a sizable fraction of repeats comprising 9.1% of the DNA sequence with those in non-coding being 0.5% and those in coding regions being 8.6% of the genome. These include simple repeats of size 25-30 bases with several hundred copies as well as complex repeat copies of CDS of transposases. There are two regions in the chromosome from 2,326,770 to 2,331,441 and 2,537,291 to 2,555,096 where these repeats occur in tandem. Two virtual sequences have been generated, one of which having the repeat region from 2.32 to 2.56 Mb randomized and rest of the chromosome left untouched. This is labeled virtual sequence I. Virtual sequence II is constructed from leaving the repeat region untouched and randomizing the rest of the chromosome. M(k) is then plotted against base separation k for these two virtual sequences along with the real and randomized sequences. The results are shown in Fig. 4.

It is observed that the effect of randomizing the repeat regions leaving the rest of the chromosome intact (virtual sequence I) does not affect the value of M(k) at 0.1 Mb, but lowers the information content in the coding regions somewhat. This is because the tandem repeats are in the CDS of transposases. But preserving the selected repeat region and randomizing the rest (virtual sequence II) lowers the M(k) value drastically, where beyond the coding regions, it is almost equal to that due to a randomized sequence. It may be concluded that the repeats in coding regions have some effect on the information content, but the factor leading to large value of $Mu$ is the location of 86.7% of the CDS on the leading strand with respect to replication. Therefore, this in silico experiment proves that the preponderance of genes occurring on one strand or GSB is responsible for the high value of Mutual Information at 0.1 Mb and, therefore, $Mu$ for this genome. Further evidence is provided by Table 4,
where the GSB (in CDS% on one strand) is given along with the value of $Mu$ for a genome.

Conclusions

The results of this study lead to the inference that high values of $Mu$ are indicative of strand asymmetry or GSB. Since the scale of the long-range correlations in this case is 0.1Mb or 100,000 nt, this can be taken as an indication of long-range order. Table 4 has data from the whole genome papers and Rocha and Dankin. The domain of Archaea can be typified by low values of $Mu$ or an absence of GSB. Firmicutes, which utilize two different DNA polymerases, DnaE and Pol-C for replication and have advantage of having the majority of all their genes on the leading strand, have high values of $Mu$. Gamma proteo-bacteria and other proteo-bacteria (save a few exceptions) have low values of $Mu$. Gamma proteo-bacteria and other proteo-bacteria, Actinobacteria and other groups studied (see Table 1) have either DnaE (homodimer DnaE1/DnaE1) or DnaE (heterodimer DnaE1/DnaE2) polymerases as discussed in. And it may be speculated that replication mechanisms of these bacteria do not give them the advantage of avoiding the probability of aborted transcripts due to headlong collision of DNA polymerase and RNA polymerase and so distribution of their genes with respect to either strand is not as biased as Firmicutes. The exceptions like Ehrlichias and Bartonellas from group of Alpha proteo-bacteria have very large number of repeats, similar to Xylella fastidiosa from Gamma proteo-bacteria discussed earlier. For the group of Spirochaetes, it is observed that linear chromosomes of Borrelia all show high values of $Mu$, but Leptospira chromosomes do not. In case of in silico study of T. tengcongensis, (Firmicute), the effect of repeats is negligible as discussed. The occurrence of repeats cannot be directly correlated to large value of $Mu$, signifying long-range correlations. Long-range order will depend on the type and copy number of repeats, which vary from genome to genome.

It may be stated that selection of cases is biased towards Firmicutes (60 cases) and Gamma proteo-bacteria (61 cases, see Table 1). This is partly due to a large number of completely sequenced genomes of these two phyla as well as an outcome of trying to find the exceptions that strayed from the general trend, those that would not give predictable high value of $Mu$ for Firmicutes or low values of $Mu$ for Gamma proteo-bacteria. There are a few exceptions in the phylum Firmicutes. Some of Streptococci do not have large value of $Mu$. It may be due to large number of recombination events, making their genomes mosaic and there is no significant bias in the gene distribution. Mycoplasmas, which belong to Tenericutes, do not give a high value of $Mu$. It has been reported that their genomes show neither compositional bias nor gene strand bias.

All prokaryotic chromosomes will exhibit correlations due to the length scale corresponding to the distribution over CDS lengths, but an additional long-range correlation length is required for long-range order in the structure of chromosomes to emerge. This additional correlation may be due to concatenation of majority of co-transcribing genes on the same strand (GSB) or to a very large number of copies of tandem repeats occurring together on a scale of 100 kb or more. It may be concluded that pronounced (70% or more) GSB leads to long-range order. The converse might not always hold true, since arrays of tandem repeats on a global scale of the genome do generate long-range order.

The results of statistical analysis give a gamma distribution of $Mu$ with respect to GC%. The gamma distribution is often used to find life-expectancy of a population, say for insurance purposes. Each observation amounts to a time-interval between events and just as there are few, but not negligible beneficiaries of an insurance policy, there are also a few but not negligible number of bacteria showing high values of $Mu$ or long-range order. There is a clear negative correlation of $Mu$ with GC%. It has been catalogued that AT-rich genomes are prone to have large number of repeats.

It is observed that genome size $Gs$ has a positive correlation with GC%. There is no clear correlation of $Mu$ to genome size. This behaviour is as expected. The high value of asymptotic or long-range M(k) is generated by inherent structure of the genome and not its size. The data for archaea and the subsequent analysis show no correlation to GC% and the very low value of $Mu$ for each archaea points to the fact that organization of archaean genomes is very different from bacterial genomes. This is borne out by the markedly different variation of each of the parameters with bacterial genomes. This may be seen in Table 2 for the Pearson correlation coefficients.
Acknowledgement

The author wishes to thank Prof. C K Mitra, Department of Biochemistry, Hyderabad University, for carefully going over the manuscript and making helpful comments and valuable suggestions.

References