Bacterial genomes lacking long-range correlations may not be modeled by low-order Markov chains: The role of mixing statistics and frame shift of neighboring genes

Germinal Cocho a, Pedro Miramontes b,∗, Ricardo Mansilla c, Wentian Li d,∗

a Departamento de Sistemas Complejos, Instituto de Física, Universidad Nacional Autonoma de Mexico, Ciudad Universitaria, Mexico 04510, DF, Mexico
b Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, México 04510, DF, Mexico
c Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades, Universidad Nacional Autónoma de México, Ciudad Universitaria, Mexico 04510, DF, Mexico
d The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, NY, USA

A R T I C L E   I N F O

Article history:
Available online 30 August 2014

Keywords:
Bacterial genomes
Exponential correlation function
Markov model
Second largest eigenvalue
Hexamer
Periodicity of 10–11 bases
Heterogeneity
Codon positions

A B S T R A C T

We examine the relationship between exponential correlation functions and Markov models in a bacterial genome in detail. Despite the well known fact that Markov models generate sequences with correlation function that decays exponentially, simply constructed Markov models based on nearest-neighbor dimer (first-order), trimer (second-order), up to hexamer (fifth-order), and treating the DNA sequence as being homogeneous all fail to predict the value of exponential decay rate. Even reading-frame-specific Markov models (both first- and fifth-order) could not explain the fact that the exponential decay is very slow. Starting with the in-phase coding-DNA-sequence (CDS), we investigated correlation within a fixed-codon-position subsequence, and in artificially constructed sequences by packing CDSs with out-of-phase spacers, as well as altering CDS length distribution by imposing an upper limit. From these targeted analyses, we conclude that the correlation in the bacterial genomic sequence is mainly due to a mixing of heterogeneous statistics at different codon positions, and the decay of correlation is due to the possible out-of-phase between neighboring CDSs. There are also small contributions to the correlation from bases at the same codon position, as well as by non-coding sequences. These show that the seemingly simple exponential correlation functions in bacterial genome hide a complexity in correlation structure which is not suitable for a modeling by Markov chain in a homogeneous sequence. Other results include: use of the (absolute value) second largest eigenvalue to represent the 16 correlation functions and the prediction of a 10–11 base periodicity from the hexamer frequencies.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Long-range correlations often refer to a power-law correlation function, as versus short-range correlations referring in exponential correlation function. Many genomes, when a chromosome is treated as a sequence of symbols or numerical values, exhibit power-law long-range correlations (Li, 1997a; Buldyrev, 2006; Arneodo et al., 2011). More interestingly, the type of long-range correlations in genomes share similarity with the “1/f noise” time series (Li and Kaneko, 1992; Voss, 1992; Li et al., 1998; Li and Holste, 2005). Not all genomes exhibit power-law correlation functions, however – the bacteria genomes tend to exhibit 1/f² spectra (Li, 1997b) and exponential correlation functions (Bernaola-Galván et al., 2002).

There are many mathematical models of sequences with power-law correlations (Beran, 1994; Beran et al., 2014). Although there are attempts to propose a universal framework for all observed power-laws (Peterson et al., 2013), the mechanical model of any specific dataset with power-law distributions could be non-universal and not applicable to other datasets (Sornette, 2006). For example, many long-range correlations of complex genomes may be caused by large domains with differential base compositions, whose size follow a broad or even long-tailed distribution (Bernaola-Galván et al., 1996; Clay et al., 2001).

The range of mathematical models of sequences with exponential correlation function, on the other hand, is relatively narrow. Markov chains are almost always used as the generating model.
These naturally lead to the argument that bacterial genomes with exponential correlation functions should be modeled by first-order Markov models whose transition probabilities are obtained from the nearest neighbor bases.

In this paper, we will show that simple Markov models do not actually explain the empirical correlation function. On one hand, exponential correlation functions (modulated by the periodicity of three bases) are indeed observed in DNA sequences; on the other hand, we can also derive the Markov transition probabilities from the different frequencies. The decay rate expected from the constructed Markov model can be compared to the observed one.

To avoid any artifact introduced by collapsing four nucleotides into two symbols (either \{W = A or T, S = C or G\}, or \{R = A or G, Y = C or T\}) (for other attempts with the similar aim, see, e.g., Korotkov et al., 2003), we characterize the 4-nucleotide correlation by 16 correlation functions (which consist of 9 independent values under the assumption that the base compositions are given (Herzel and Grosse, 1995), or reduced to 10 by the approximate strand symmetry (Li, 1997a), or even to 1 as the exact strand symmetry would lead to a binary sequence which is known to have one independent correlation (Li, 1990).

The prediction on the decay exponent by the nearest-neighbor Markov model is made through the second largest eigenvalue (SLE), \(\lambda_2\) (the largest eigenvalue is equal to 1) of the transition matrix (see, e.g., Buldyrev, 2006). The propagation of this short-range correlation to longer distances is by multiplying the SLE again and again. Similarly, the observed correlation at a longer spacing can be viewed as a “transition” acting at a distance. Thus we can also use the SLE for such a “transition matrix” to characterize the 16 correlations. This idea is similar to the principal components used in Teitelman and Eeckman (1996).

Besides the exponential decay of correlation, Markov model can also predict periodic components. Whenever a pair of eigenvalues are close to each other and both are complex, their real parts should be close to 1, and their imaginary parts should be close to \(\pi\). This makes the correlation function oscillate between positive and negative values with the same frequency. When the SLE is complex, the correlation function oscillates between positive and negative values with the frequency given by the imaginary part of the SLE.

2. A typical correlation function in bacterial genomes

We use the Escherichia coli genome as an illustration of autocorrelation function for a typical bacterial genome. We download the chromosomal sequence of the disease-causing (Enteropathogenic) strain of E. coli E2348/69 belonging to the phylogroup B2 (Iguchi et al., 2009) from ftp://ftp.ncbi.nih.gov/genomes/Bacterial (the file: Escherichia_coli_O127_H6_E2348_69_uidd59343/NC_011601.gbk), or from EBI at http://www.sanger.ac.uk/resources/downloads/bacteria/escherichiacoli.html (the FM180569 entry). The genome is circular with 4965553 bases, 4703 genes (including pseudogenes), of which 4554 are protein-coding genes with 1,411,554 amino acids.

The autocorrelation function measures the linear correlation between two types of nucleotides at two positions in the genome separated by a distance \(d\):

\[
C_{\alpha,\beta}(d) = P_{\alpha,\beta}(d) - P_{\alpha}P_{\beta}, \quad \alpha, \beta = \{A,C,G,T\}, \quad d = 1, 2, \ldots
\]

where \(C_{\alpha,\beta}(d)\) is the joint probability of symbol \(\alpha\) followed by symbol \(\beta\) \(d\)-bases to the right, and \(P_{\alpha}\) (\(P_{\beta}\)) is the probability in finding symbol \(\alpha\) (\(\beta\)) in the sequence. Of these 16 correlation functions, the strand symmetry leads to \(C_{\alpha,\beta}(d) = C_{\alpha',\beta'}(d)\), where \(\alpha'\) is the nucleotide that complement \(\alpha\) (e.g. \(\alpha' = C, \alpha' = G\)), and \(\beta'\) complement \(\beta\). Since we are mainly interested in examining the propagation of nearest neighbor correlations to intermediate distances, we limit \(d \leq 1000\) in this paper. Longer claimed periodicities such as the 117 kb spacing between evolutionarily conserved gene pairs (Wright et al., 2007), are not addressed here.

Fig. 1(A) shows the 16 correlation functions for distances smaller than 12, with complementary pairs in the same color (e.g. \(C_{AA}(d)\) and \(C_{TT}(d)\)). The periodicity of 3 is visible. The \(C_{AA}(d)\) function with \(x\) in log-scale (for yeast in Li, 1997a) and \(x-y\) in log-log scale (for Mycobacterium tuberculosis in Bernaola-Galván et al., 2002) have been shown before, and it is known that there are both positive and negative branches. Here we split these correlation in positive and negative (as well as close to zero) branches for each \((\alpha, \beta)\) pair (Fig. 1(B–D)), with the positive branch in semi-log scale (Fig. 1(B)).

It becomes clear from Fig. 1(B) that the positive branch of the correlation function decays exponentially. The \(C_{AA}(d) \approx C_{TT}(d)\) with \(d = 3, 6, \ldots\) represents the strongest correlation, followed by \(C_{AT}(d)\) (with \(d = 1, 4, 7, \ldots\)) and \(C_{TA}(d)\) (with \(d = 2, 5, \ldots\)). To quantify the exponential decay

\[
C_{\alpha,\beta}(d) \sim \exp(-\gamma d) = \exp \left(-\frac{d}{d_0}\right),
\]

we regress \(\log(C_{\alpha,\beta}(d))\) over distance \(d\). We obtained \(\gamma = 0.00147, 0.00158\) for \(C_{AA}\) and \(C_{TT}\), or \(d_0 = 678, 632\) bases; \(\gamma = 0.00154, 0.00127\) for \(C_{AA}\) and \(C_{TT}\) or \(d_0 = 650, 786\) bases (after removing the first few points). These results are comparable to the \(d_0\) value of 639 obtained in Bernaola-Galván et al. (2002). In the next section, we will examine whether simple Markov models can explain this decay rate.

3. First-order Markov model based on dimer frequencies

To construct a first-order Markov model, all 16 dimer types are counted. The first-order Markov transition probabilities are
Fig. 1. Base–base correlation functions for E. coli genomic sequence. (A) 16 correlation functions at distance ≤12. (B) The “positive branch” of the correlation functions (e.g. $C_{AA}(d)$ at $d = 3, 6, 9, \ldots$). (C) The “zero branch” of the correlation functions. (D) The “negative branch” of the correlation functions (e.g., $C_{GG}(d)$ at $d = 3, 6, 9, \ldots$).

$P_{\alpha \rightarrow \beta} = P_{\alpha}(d = 1)/P_{\alpha}$. For E. coli, we have calculated the transition probabilities as:

$$
\begin{pmatrix}
A & T & C & G \\
0.0733/0.2474 & 0.0672/0.2474 & 0.0553/0.2474 & 0.0517/0.2474 \\
0.0459/0.2469 & 0.0733/0.2469 & 0.0577/0.2469 & 0.0700/0.2469 \\
0.0703/0.2526 & 0.0512/0.2526 & 0.0581/0.2526 & 0.0731/0.2526 \\
0.0580/0.2531 & 0.0553/0.2531 & 0.0815/0.2531 & 0.0583/0.2531 \\
\end{pmatrix}
$$

Note that in general, on the same strand $P_{\alpha \rightarrow \beta}$ is not equal to $P_{\beta \rightarrow \alpha}$ (Provata et al., 2014). The largest transition probability for a given nucleotide (maximum value per row) is marked in bold. Eq. 3 shows that $A/T/C/G$ is more likely (with probabilities around 0.3) to be followed by $A/T/G/C$. The least likely transitions are $A \rightarrow G$, $T \rightarrow A$, $C \rightarrow T$, and $G \rightarrow T$ (with probabilities around 0.2). Fig. 2(left) shows the Markov transitions with transition probability larger than 0.25. Note that strand symmetry at the dimer frequency (e.g. $P_{AT} \approx P_{TA}$) does not lead to a similar symmetry in transition probabilities (e.g. $P_{AT} \neq P_{TA}$).
The four eigenvalues of the transition matrix in Eq. (3) are: $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)=(1, 0.0675 \pm 0.0639i, -0.0818)$. The R package eigen is used for this calculation. The largest eigenvalue $\lambda_1 = 1$ leads to a constant term, to be subtracted out in the correlation function, and the second largest eigenvalue determines the decay rate of the exponential correlation (Li, 1987). Complex eigenvalues lead to periodic components in the correlation function by the angle formed by the complex eigenvalue pair in the plane (Li, 1987). For our transition matrix, $|\lambda_2| = |\lambda_3| = 0.093$ and angle $\theta_2 = \theta_3 = 43.4^\circ$, which corresponds to a periodicity of 360/43.4 = 8.3 bases.

Since the correlation function of a sequence generated by this transition matrix will be proportional to $|\lambda_2|^d = \exp(-\log(1/|\lambda_2|)d)$, we predict the exponentially decay rate $\gamma = 2.376$ and the characteristic length $d_0 = 0.42$ bases. Comparing this $d_0$ with the observed value of $\sim 600$, Eq. (3) completely fails to estimate the parameter in the exponential correlation function observed from the same sequence. An even weaker periodic component is anticipated by the eigenvalue $\lambda_4$, with $\gamma = 2.5$ or $d_0 = 0.4$ bases, and periodicity of two.

There is a more direct way to check the validity of the first-order Markov model: by comparing the observed and expected correlation at $d = 2$. The expected correlation is simply through the product of the transition matrix:

$$\hat{C}_{\alpha\beta}(d = 2) = P_\alpha \sum_{\alpha'} P_{\alpha'\rightarrow \alpha} P_{\alpha\rightarrow \beta} - P_\alpha P_\beta$$

(4)

The observed-over-expected ratio (O/E) is:

$$\frac{O}{E} = \frac{C_{\alpha\beta}(d = 2)}{C_{\alpha\beta}(d = 2)} = \begin{pmatrix}
0.017/0.0002 & -0.0033/0.0012 & 0.0072/(−0.0011) & −0.0057/(−0.0036) \\
0.005/(−0.0018) & 0.0019/0.00023 & −0.00032/0.00084 & −0.0066/0.0007 \\
−0.0068/0.00073 & −0.0056/(−0.00037) & −0.0057/0.00075 & 0.018/(−0.0011) \\
0.00007/0.00086 & 0.007/(−0.0011) & −0.0012/(−0.00053) & −0.0059/0.00075
\end{pmatrix}$$

(5)

\[
\begin{pmatrix}
A & T & C & G \\
A & 8.9 & −2.7 & −6.8 & 15.8 \\
T & −2.8 & 8.1 & 0.38 & −9.3 \\
C & −9.3 & 15 & −7.7 & −16.5 \\
G & 0.08 & −6.4 & 2.3 & −7.8
\end{pmatrix}
\]
This shows that the first-order Markov chain could not even predict
the next-nearest-neighbor correlation correctly: the TC, GA corre-
lation is over-estimated, and others are under-estimated, with half
of them having the wrong sign. Similar O/E can be determined at
d = 3, 4, ..., and the inconsistency is even more severe (results not
shown).

4. Two halves of the genome

Circular bacterial genomes are naturally partitioned into two
halves at the points of replication origin (Ori) and replication ter-
minus (Ter). During replication, there are four strands serving as
template: two leading strand templates in 3’ to 5’ direction, one on
the left side or Ori, one on the right; and two lagging strand tem-
plates in the 5’ to 3’ direction. The copied strands as a conse-
quence of replication are two leading strands from 5’ to 3’, and two
lagging strands from 3’ to 5’.

The *E. coli* sequence downloaded starts from 5’ at position 1
to 3’ at position 4965553. The Ter is half-genome away from the
replication origin, at 1739942, and the Ori is given at the seg-
ment of (4222363, 4223074). We estimate the Markov transition
probabilities separately for right leading strand (1–1739942, and
4223074–4965553), and left leading strand (1739942, 4222363).

For the latter, the downloaded sequence in the region is reverse-
complemented.

The Markov transition matrices for the right and left leading
strand are:

\[
P_{\alpha \rightarrow \beta (\text{right-leading})} = \begin{pmatrix}
A & T & C & G \\
A & 0.297 & 0.273 & 0.216 & 0.214 \\
T & 0.184 & 0.293 & 0.225 & 0.298 \\
C & 0.272 & 0.207 & 0.222 & 0.299 \\
G & 0.229 & 0.220 & 0.311 & 0.240 \\
\end{pmatrix}
\]

\[
P_{\alpha \rightarrow \beta (\text{left-leading})} = \begin{pmatrix}
A & T & C & G \\
A & 0.301 & 0.188 & 0.269 & 0.242 \\
T & 0.270 & 0.300 & 0.204 & 0.230 \\
C & 0.217 & 0.229 & 0.220 & 0.334 \\
G & 0.198 & 0.284 & 0.280 & 0.238 \\
\end{pmatrix}
\] (6)

The strand symmetry between C and G, A and T, is broken when
only the half of the genome is examined (Lobry, 1996; Sánchez and
José, 2002), e.g., with G% higher than C% on the leading strand. The
left and right leading strands, however, may or may not differ. For E.
coli, the base composition of the two halves of the leading strands is
very similar. Eq. (6) shows that the Markov transition probabilities
have minor differences. The set of eigenvalues of the two matrices
in Eq. (6) are similar to each other, as well as similar to that of Eq. (3).
The reverse complement operation does not affect the eigenvalues.
Due to this similarity of eigenvalues, we will ignore the partition of
the two halves from now on.

5. Second largest eigenvalue function

If the absolute value of the second largest eigenvalue (SLE)
used to characterize the nearest neighbor correlation is propagated
to longer distances by the Markov model, it can also be used to
characterize the observed correlation at a distance. To this end, we
convert a correlation function to a transition matrix (function):

\[
P_{\alpha \rightarrow \beta (d)} = \frac{C_{\alpha \beta (d)} + P_{\alpha}P_{\beta}}{P_{\alpha}}
\]

and the (absolute value) of the second largest eigenvalue of \(P_{\alpha \rightarrow \beta (d)}\)
can be determined at all distances.

Fig. 3 shows the SLE as a function of \(d\) in semi-log scale. It repro-
duce almost all observations from the 16 correlation functions, such
as, (1) the existence of two branches; (2) the exponential decay of
the higher branch; (3) the periodicity of 3. Also, it provides extra
advantages such as, the “negative branch” in Fig. 1 becomes the
lower (but positive) branch in Fig. 3, so it can be fitted by an expo-
nential function. The regression line for the upper branch has a
slope of \(-0.0006\) (i.e. SLE \(\sim \exp(-0.0006d)\)). The regression lines
for \(\log_{10}(\text{SLE})\) over \(d\) for the lower branch has slope \(-0.0008\). We
can easily compare the actually SLE with those in a randomized
sequence, representing chance event (the red dots in Fig. 3).
To see how miserably wrong the first-order Markov model is, we propagate the correlation by multiplying a SLE for each added base in the distance. The gray line in Fig. 3 shows the result: which already drop to below insignificant level after 5 steps. The decay of correlation due to propagation in Markov model looks almost like a vertical line.

6. Higher-order Markov models

Higher-order Markov chains have been applied to fit DNA sequence data before (Blasidell, 1984; Raftery and Tavaré, 1994; Usatenko and Yampol’skii, 2003; Bejerano, 2004; Wang and Hannenhalli, 2005; Menéndez et al., 2011). For a true long-range correlated DNA sequence, the best-fit order of a Markov chain increases with the sequence length (Papapetrou and Kugiumtzis, 2014). We are less worried about the parsimony issue (the higher the order, the more number of parameters need to be estimated), but more concerned about whether the predicted exponential decay rate of the correlation function is consistent with the data.

The second-order Markov model is used to model the transition probability from a dimer (α₁α₂) to the next base β: $P_{α₁α₂→β}$. In order to carry out a similar eigenvalue calculation, we need to make the matrix square (though there exist alternative methods such as the non-negative matrix factorization (Lee and Seung, 1999)), thus we re-write the transition matrix as a dimer-to-dimer one (overlapping): $P_{α₁α₂→β_1β_2}$ (non-zero only when $β_1 = α₂$). This formulation is a Markov model on the de Bruijn graph (Heath and Pati, 2007; Pati, 2008), with each node as a dimer, which is widely used in the de novo assembly of short reads from the next generation sequencing (Pevzner et al., 2001; Zerbino and Birney, 2008; Compeau et al., 2011; Pell et al., 2012). The transition probabilities can be estimated from the trimer frequencies: $P_{α₁α₂→β}$ = $P_{α₁α₂β_1}/P_{α₁α₂β_2}$.

The 16 eigenvalues from the dimer-to-dimer transition matrix obtained from the E. coli genome are: (1, −0.132 ± 0.208i, 0.187 ± 0.106i, −0.057 ± 0.184i, −0.094, 0.057 ± 0.026i, 0.010 ± 0.058i, −0.007 ± 0.051i, −0.043, 0.025). The magnitude of the second largest eigenvalue is $|λ_2| = 0.2466$, corresponding to $γ = 1.4$, $d₀ = 0.71$. This prediction is still far from adequate. However, interestingly, the second/third largest eigenvalue anticipate a periodicity of 2.94, very close to the actual periodicity of three. If we move further down the eigenvalue lists, $|λ_4| = |λ_5| = 0.2149$ with periodicity of 12.2 bases; $|λ_6| = |λ_7| = 0.1927$ with periodicity of 3.36 bases.

The third-order Markov model is analyzed similarly, with $P_{α₁α₂α₃→β} = P_{α₁α₂β_1}/P_{α₁α₂β_2}$. This can be estimated by the k-mer frequency for k=4. The Markov transition matrix is a 64-by-64 matrix for trimer-to-(overlapping) trimer transition: $P_{α₁α₂α₃→β_1β_2β_3}$ (non-zero only when $α₃ = β_2$). The largest eight eigenvalues are (1, −0.215 ± 0.434i, 0.369, 0.295 ± 0.133i, −0.129 ± 0.267i). The second largest eigenvalue has $|λ_2| = 0.475$ ($γ = 0.744$, $d₀ = 1.34$). The complex eigenvalue pair for $λ₂, λ₃$ leads to a periodicity of 2.08 bases.

For 4th-order Markov model, $P_{α₁α₂α₃α₄→β} = P_{α₁α₂α₃β_1}/P_{α₁α₂α₃β_2}$. The Markov transition matrix has the 256-by-256 dimension (tetramer to tetramer). The value of the second largest eigenvalue is increased to $|λ₂| = |λ₃| = 0.499$, corresponding to $γ = 0.695$ and $d₀ = 1.44$. Despite the disappointment of the prediction of exponential decay rate, the periodicities predicted remain close to the observed value: 3.1 bases for $λ₂, λ₃$, and 2.7 bases for $λ₄, λ₅$.

Probably the most important higher-order Markov model to be examined is the 5th-order, using the hexamer frequencies, transition from pentamer to pentamer. This model is important because it covers the size of two codons. The difference between nearest neighbor codon–codon correlation and hexamer statistics is that the former is “in frame”, whereas the moving hexamer window would average over all three reading frames. The dimension of the Markov transition matrix is 1024-by-1024. The top eigenvalues are (1, −0.24 ± 0.49i, 0.43 ± 0.29i, 0.49, −0.27 ± 0.36i, 0.062 ± 0.45i, −0.44 ± 0.062i, 0.41, −0.029 ± 0.38i, −0.29 ± 0.22i, −0.18 ± 0.30i, 0.35, ...).

The second/third largest eigenvalue is $|λ_2| = |λ₃| = 0.546$ ($γ = 0.605$, $d₀ = 1.65$), with an oscillating component of 3.09 bases. The 4th/5th largest eigenvalue is $|λ₄| = |λ₅| = 0.521$, with the oscillating component of 10.49 bases. These results show that although 5th-order Markov model may not capture the exponential decay rate of the correlation, it anticipates the periodicity-3 and periodicity-10 as the two strongest components.

The prediction of the ~10.5 periodicity based on hexamer frequencies is surprising. It is well known that the periodicity of 10/10.5/11 is the second strongest periodic signal in DNA sequences (Trifonov and Sussman, 1980; Trifonov, 1998; Herzel et al., 1999; Kravatskaya et al., 2011), and it can be both related to a periodicity in protein sequence (Herzel et al., 1999) or a nucleosome positioning signal (Trifonov and Sussman, 1980; Salih et al., 2008, 2014; Sosa et al., 2013).

For a pentamer which can make a Markov transition to four other pentamers, following the transition with the highest probability leads to the most likely path. All such paths after a transient will either move into a fixed point (AAAAA →AAAAA), periodicity-3 (CAGCA → AGCAG → CGACG → CAGCA, or (GCA)ₙ) (CGGCG → GCGCG → GCGGC → CGGGC, or (GCG)ₙ), or periodicity-6 ((GCCAGC)ₙ). Interestingly, if we follow the transition with the second largest probability, most paths are attracted to a cycle with periodicity of 11 (ATTGA → TGGAA → GAAAT → AAATT → AATTC → ATTCC → TCCTA → CCATT → CATTG → ATTGA).

7. Frame-specific Markov models for CDS

Bacterial genomes such as E. coli contain large proportion of protein-coding regions. For example, 1,411,554 × 3 = 4,234,662 bases (or 85.3%) of the E. coli genome are occupied by coding sequences. It has been proposed for long time that an extension of the Markov model to be a frame-specific one (called non-uniform Markov models in Borodovsky et al., 1986) could fit the data better.

Instead of the natural genomic sequence, we patch all coding-DNA-sequences (CDS) in E. coli, in phase and converted to the same strand if necessary, into a single CDS sequence. Pseudogenes,
non-coding sequences are removed. From this single CDS sequence, three frame-specific Markov transition matrices are obtained as:

\[
\begin{align*}
P_{\alpha_1 \to \beta_2} &= \begin{pmatrix} A & T & C & \color{red} G \\ 0.332 & 0.345 & 0.212 & 0.112 \\ 0.194 & 0.415 & 0.215 & 0.177 \\ 0.277 & 0.325 & 0.182 & 0.216 \\ 0.315 & 0.202 & 0.273 & 0.210 \end{pmatrix} \\
P_{\alpha_2 \to \beta_3} &= \begin{pmatrix} A & T & C & \color{red} G \\ 0.313 & 0.272 & 0.213 & 0.201 \\ 0.111 & 0.273 & 0.221 & 0.395 \\ 0.198 & 0.180 & 0.274 & 0.348 \\ 0.086 & 0.331 & 0.398 & 0.185 \end{pmatrix} \\
P_{\alpha_3 \to \beta_1} &= \begin{pmatrix} A & T & C & \color{red} G \\ 0.289 & 0.151 & 0.252 & 0.307 \\ 0.243 & 0.171 & 0.214 & 0.373 \\ 0.266 & 0.166 & 0.189 & 0.378 \\ 0.230 & 0.140 & 0.304 & 0.326 \end{pmatrix}
\end{align*}
\]

(8)

The second- to third-codon-position transition matrix in Eq. (8) is similar to the first-order Markov transition matrix without considering the frame (Eq. (3)). Eq. (8) shows that the nucleotide G is preferred in the first position, A and T are preferred in the second position, as previously reported in Gutiérrez et al. (1996). The third- to first-codon-position transition matrix should be related to the intercodon dimer frequency studied in De Amicis and Marchetti (2000), Fadie et al. (2001), and Sánchez (2011, 2013). The Markov transitions with transition probabilities larger than 0.25 are drawn in Fig. 2 (right).

The three sets of eigenvalues of these three Markov transition matrices are: (1, 0.1 ± 0.004i, -0.061), (1, -0.22, 0.13 ± 0.031i), and (1, -0.084, 0.047, 0.012). The three second largest eigenvalues are 0.13, 0.22, and 0.084, with the geometric mean being 0.13 (corresponding to \( \gamma = 2 \) and \( d_0 = 0.5 \)). This is only slightly larger than the second largest eigenvalue without considering reading frames.

To utilize the codon information as much as possible, we construct a fifth-order, frame-specific Markov model: \( P_{\text{smert} \to \beta_1} \), \( P_{\text{smert} \to \beta_2} \), and \( P_{\text{smert} \to \beta_3} \). These transition matrices are actually singular, as there are forbidden hexamers at specific frames. To avoid this problem, we add a small amount (0.1) to the zero hexamer counts.

For \( P_{\text{smert} \to \beta_1} \), \(|\lambda_2| = 0.498 \) associated with a periodicity of 3.27. The next largest eigenvalue is \(|\lambda_4| = 0.49 \) associated with a periodicity of 10.77. For the second codon position \( P_{\text{smert} \to \beta_2} \), \(|\lambda_2| = 0.526 \) associated with periodicity of 10.25. For the third codon position \( P_{\text{smert} \to \beta_3} \), \(|\lambda_2| = 0.535 \) associated with a periodicity of 11.9. These results show that the codon positions 2 and 3 are more related to the periodicity of 10–11, and codon position 1 is more related to periodicity of 3.

To summarize all Markov models we have investigated, Table 1 shows the SLE and the corresponding periodicities, as well as the next largest eigenvalues plus the corresponding periodicities, for the model analyzed. There are several trends revealed. One, the \(|\lambda_2| \) increases with the order of the Markov model. It is understandable as the more information known about the context, the more certain we are able to predict the base, thus the higher the correlation. Second, periodicity-3 becomes a dominant periodic component when the Markov order is more than 2. Periodicity-10/11 becomes a dominant component when the Markov order is more than 4–5. Third, frame-specific correlation (as measured by \(|\lambda_2| \)) in the first-order Markov chain can be stronger than frame-averaged one. In fact, the correlation is the strongest from codon position 2 to position 3. This conclusion is not obvious from \(|\lambda_2| \) for the fifth-order Markov chain, as \(|\lambda_2| = 0.546 > 0.498, 0.526, 0.536 \). However, when top eigenvalues beyond \(|\lambda_2| \) are accumulated, the conclusion becomes clear (result not shown).

8. Within-frame correlation in CDS subsequences

A key question on the source of the correlation is: if we take away the frames, use bases from only a particular codon position in the CDS sequence, do we still see correlation? We construct three subsequences, for bases on codon position-1 only, 2 only, and 3 only (called RF1, RF2, RF3). CDSs from different genes are joined into a single sequence. Fig. 4 shows \( C_{\text{AA}} \), \( C_{\text{CT}} \), \( C_{\text{CC}} \), \( C_{\text{CD}} \) in three sub-sequences (orange: codon-position-1, green: codon-position-2, and blue: codon-position-3). Also shown are the correlation function for a shuffled sequence as the baseline. The frame-averaged \( C_{\text{AA}}, C_{\text{CT}}, C_{\text{CC}}, C_{\text{CD}} \) are larger (same as Fig. 1). Note that the nearest neighbor in the sequence from a particular frame/codon-position is equivalent to distance = 3 in the genomic sequence.

Careful examination of Fig. 4 shows that the \( C_{\text{AA}}, C_{\text{CT}} \) in RF2, RF3, and \( C_{\text{CC}}, C_{\text{CD}} \) in RF3 exhibit excess correlation compared to the random sequence, up to certain distances. The correlation in RF1 is closer to noise. There are also various cyclic patterns from a peak to another peak at a longer distance.

To characterize all 16 correlation functions in the three subsequences, we convert correlation function to transition probability matrices and calculate the second largest eigenvalue as a function of the distance (Fig. 5). Again, the SLE function for RF1 is not distinguishable from a random sequence, however, those for RF2, RF3 are clearly statistically significant. The SLE function from the genomic sequence at distance \( d = 3 \) is higher than the SLE function at other distances (low branch) is also shown as a comparison.

Fig. 5 shows a big drop of SLE from the first distance \( (d = 3) \) to the second distance \( (d = 6) \). The SLE for RF1, RF2, RF3 are 0.0399, 0.0388, 0.147 at \( d = 3 \), but 0.008, 0.0292, 0.0216 at \( d = 6 \). The SLE becomes relatively high again at \( d = 12: 0.016, 0.031, 0.033 \) at RF1, RF2, RF3. These values are 1.4, 2, 1.4 times of the SLE from randomized sequence. Further oscillations of the peak locations can be seen from Fig. 5, sometimes after 9 bases, other times after 12 bases. The existence of longer periodicity after periodicity-3 signal being removed (here by examining only the RF1, RF2, RF3 subsequences) has been studied before (Korotkov et al., 1997; Cohanim et al., 2006; Cohanim and Haran, 2009; Kravatskaya et al., 2011).

One possible mechanism for forcing a codon-codon correlation is the avoidance of certain motif for ribosomal RNA binding site, called Shine-Dalgarno sequence (Shine and Dalgarno, 1974). It has been shown that the presence of Shine–Dalgarno-like sequence in mRNA causes translational pause with deleterious consequences (Li et al., 2012). Other possible underlying biology of longer-range correlation and periodicity include the nucleosome positioning within exons (Andersson et al., 2009), and periodicity/repeats in protein sequences (Marcotte et al., 1998; Korotkova et al., 1999; Heger and Holmm, 2000; Rajathei and Selvaraj, 2013).

Another way to remove the effect of three reading frames is to check the amino-amine correlation in the coded protein sequence (Suryapavan et al., 2006). We use the annotation of the E. coli genome sequence, and count all amino-acid dimers. The few occurrence of the 21st amino acid, Selenocysteine (U) (Stadman, 1996), are removed. From the resulting 20-by-20 transition matrix, the
The second largest eigenvalue (its absolute value) and the corresponding periodicity (if it the eigenvalue is complex or negative) for various Markov models examined. The next largest eigenvalue (which can be $|\lambda_4|$ or $|\lambda_3|$) and its corresponding periodicity is also shown. If the eigenvalue is real positive, a “–” is listed.

| Model                     | Dimension | $|\lambda_2|$ | Periodicity | Next eigenvalue | Periodicity |
|--------------------------|-----------|---------------|-------------|----------------|-------------|
| Genomic sequence         | 4-by-4    | 0.0929        | 8.3         | 0.082          | 2           |
| 1st order Markov (dimer) | 16-by-16  | 0.247         | 2.9         | 0.0215         | 12.2        |
| 2nd order Markov (trimer)| 64-by-64  | 0.475         | 3.1         | 0.369          | –           |
| 4th order Markov (pentamer)| 256-by-256| 0.499       | 3.1         | 0.425          | 2.7         |
| 5th order Markov (hexamer)| 1024-by-1024| 0.546     | 3.1         | 0.521          | 10.5        |
| CDS                      |           |               |             |                |             |
| 1st order, pos3-to-pos1  | 4-by-4    | 0.084         | 2           | 0.047          | –           |
| 1st order, pos1-to-pos2  | 4-by-4    | 0.131         | 8.97        | 0.0609         | 2           |
| 1st order, pos2-to-pos3  | 4-by-4    | 0.217         | 2           | 0.135          | 26.8        |
| 5th order, pos3-to-pos1  | 1024-by-1024| 0.498    | 3.3         | 0.490          | 10.8        |
| 5th order, pos1-to-pos2  | 1024-by-1024| 0.526    | 10.3        | 0.487          | 2           |
| 5th order, pos2-to-pos3  | 1024-by-1024| 0.535    | 11.9        | 0.519          | 2.8         |

The second largest eigenvalue is $|\lambda_2| = 0.0495$. This value is consistent with the within-frame correlation at $d = 3$ in Fig. 5.

9. Mixing out-of-phase CDSs in an artificial sequence

The CDS, where each individual base is globally in-phase, is still not a good representation of the genomic sequence, where different genes may be out-of-phase or be on different strands/direction. We use two similar procedures to construct artificial sequences to closely resemble the genomic sequence but removing the correlation within a CDS.

In the first simulation, we keep each codon intact, whereas shuffle the codons within a CDS, and introducing possible frame shift between neighboring CDSs. This procedure follows these steps: (1) for any CDS, for the first half, a 43% chance to use the reverse complement sequence, whereas in the second half, a 57% chance. These

Fig. 4. The correlation $C_{AA}$, $C_{TT}$, $C_{CC}$, $C_{GG}$ for the three codon positions subsequences (RF1 in orange, RF2 in green, RF3 in blue). The corresponding correlation for the genomic sequence is plotted as a comparison (same as Fig. 1(B)). The correlation on the shuffled random sequences are shown in gray. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)
numbers are obtained from the *E. coli* gene annotation (2) for any CDS, shuffle the codons; (3) insert a spacer of size of 1, or 2, or 3 bases, chosen at random, between two neighboring CDSs.

In the second simulation, the steps (1), (3) are the same as above. The step (2) does not preserve individual codon, but preserve phase information. We shuffle the subsequence concatenated from all bases from codon position-1. Same is done for the subsequence from codon position-2, and that from codon position-3. Then these shuffled subsequences are combined at their corresponding codon position.

The SLE function of the CDS-mixing-and-codon-shuffled sequence and in-frame-base-shuffled sequence are shown in Fig. 6 (in blue and in green). The SLE of the genomic sequence is also shown as a comparison (in black). The two different simulations lead to very little difference except distances 1 and 2. It is expected as the first simulation preserve codons, whereas the second simulation does not. We can also see that the main feature of the correlation function in the original sequence is somewhat reproduced, i.e., the slow-decaying of an exponential function. Some detail are not reproduced, however, for example, the slopes for the artificial sequence are \(-0.00051\) or \(-0.00052\) (when \(y = \log_{10}\) transformed) whereas that for the original sequence is \(-0.00049\). Also, the lower branch of the correlation function (at distances which are not multiple of 3’s) for the artificial sequence is higher than that of the original sequence.

There are a few differences between the genomic sequence and our artificial sequences. Due to the small spacer size (1, 2, 3) the non-coding sequence in the artificial sequences is almost nonexistent whereas close to 15% of the original sequence are non-coding. The in-frame correlation between bases at the same codon position is destroyed in artificial sequences but present in the original sequence. Also, whether a CDS is on the direct or opposite strand in the artificial sequence is randomly chosen. The fact that the SLE function of the genomic and artificial sequences is not identical, though similar, indicates that non-coding sequences, in-frame correlation, etc. do contribute to the overall correlation function, in their small way.

In artificial sequences, the length distribution of the CDSs is preserved. What happens if the lengths are restricted by an upper limit? We show another SLE function in Fig. 6 (in light blue) when a CDS after codon shuffling is cut at its first 200 codons (or 600 bases). The two branches of SLE function decay faster than exponential – actually the decay is linear. The two merge at distance of 600. There is a very simple explanation: if the distance is much lower than 600, most base–base pairs are counted within a (in-frame) CDS, with some across the two neighboring CDSs. On the other hand, when the distance is longer than 600 bases, no counts are from an in-frame CDS and all from neighboring CDSs which may or may not be in-frame (as the spacer can be 1, 2, or 3). Then the periodicity-3 and two-branch pattern is destroyed.

These shuffled sequences enable us to understand the main cause of the slow decay correlation in the genomic sequence. When the distance is small (e.g., \(d = 6\)), as we count base–base pairs separated by 6 bases by moving position one at the time, we are mixing the three reading frames with very different base compositions. This mixing of heterogeneous statistics can easily lead to correlation (Karlin et al., 1993; Hassler and Thadewald, 2003), and a change in mixing ratio may change the correlation value or even the conclusion (Li, 2008). The following one-line formula shows the effect of mixing a base composition \(p_{A1}\) with weight \(w\) and a base composition \(p_{A2}\) with weight 1 \(- w\), where no correlation exists in each subsequence: \(p_{A1}w + p_{A2}^2(1 - w) - (p_{A1}w + p_{A2}(1 - w))^2 = (p_{A1} - p_{A2})^2w(1 - w) \neq 0\) (if \(p_{A1} \neq p_{A2}\)).

Such mixing-three-reading frame caused correlation can extend to very long distance without decay if the three reading frames continuously cover the whole genome. When such phase is interrupted by the spacer with length 1 or 2, longer distances are not equivalent to short distances. A longer distance is more likely to take base-base-pair counts from across two CDSs. These counts are less likely to be in the \(p_{A1}\) same codon positions, which weaken the overall correlation. As distance becomes longer, these type of counts contribute more, and correlation decay linearly. As we can see, this description is very different from a Markov chain which propagates short-range correlation to longer distances.

10. Conclusions

We have shown that the seemingly simple exponential correlation function in a typical bacterial genome sequence, *E. coli*,
has complicated causes. The correlation is not due to a propagation of nearest neighbor correlation (first-order Markov model) through multiplying it geometrically with the distance. Nor can it be explained by some high-order Markov models up to fifth-order, or three frame-specific Markov models across codon positions. The slow decaying correlation first requires the sampling and mixing of counts from three different codon positions which may have very different base compositions. Then it needs neighboring coding-DNA-sequences (CDS) to be out-of-phase. The length distribution of all CDSs, the weak correlation between in-phase bases on the same codon position, the non-coding sequences, all contribute to the extent to the value as well as the shape of the correlation function. Our simulation shows that the correlation function in bacterial genomes needs not to be intrinsically exponential. It can be non-decaying if all CDSs are in-frame and on the same strand. It can also decay linearly instead of exponentially if all CDSs have similar lengths.

Besides the above main conclusion, there are also a few minor conclusions from this study. The in-frame base–base correlation (for subsequences in CDS) almost does not exist in the first codon position, but some weak correlation may be detected at the second or the third codon position. The existence of longer periodocities when the periodicity-3 is completely removed by focusing on a particular codon position in CDS. The usefulness in using second largest eigenvalue to characterize the correlation function. The prediction of periodicity 10–11 when only the hexamer information is used.

Acknowledgements

We thank two reviewers for valuable comments and suggestions. W.L. thanks the financial support from the Robert S Boas Center for Genomics and Human Genetics, and the study was partially funded by project PAPIIT-UNAM IN107414.

References

Bissell, D.M., 1984. The mosaic nature of the 3′ untranslated regions of human mRNA. Cell 36, 467–475.


