Computational Biology and Chemistry 53 (2014) 15–25

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

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 Hamanidades, 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

ARTICLE INFO

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

ABSTRACT

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.

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 $1/f^2$ 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.

* Corresponding authors.
E-mail addresses: pmv@ciencias.unam.mx (P. Miramontes), wtli2012@gmail.com (W. Li).

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.005
1476-9271/

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 explained 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 dimer 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, λ_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 is complex, it represents a cyclic dynamics (e.g., Norris, 1997). A negative eigenvalue represents a periodicity-2, as the correlations behave like $(-|\lambda_2|)^d$ which oscillate between positive and negative values with even and odd distances. We are interested in whether the periodic components predicted this way by a Markov model is consistent with the observed ones in bacterial genome (such as the most dominant periodicity-3 component (Herzel and Grosse, 1997)).

Our seemingly simple task of fitting an exponential correlation function by a homogeneous statistical model, i.e., Markov model, is actually not simple at all. There have a been long history in applying Markov chain to DNA sequences (Garden, 1980; Fuchs, 1980; Blasidell, 1984; Avery and Henderson, 1999). Then it was realized that the three codon positions behave differently, leading to non-homogeneous, interpolated, interconnected Markov models (Tavaré and Song, 1989; Borodovsky and Peresetsky, 1994; Salzberg et al., 1998; Avery, 2002). Treating any source of inhomogeneity as hidden states, the hidden Markov model has also been applied to DNA (Durdin et al., 1998). However, most applications of Markov model refer to predictions (e.g. whether a region is protein-coding or not) and the predicted status is the hidden state (Krogh et al., 1994; Kulp et al., 1996).

We plan to show three unusual aspects of the correlation function in bacterial genomes in this paper. First, the correlation value is rather high due to a mixing of statistics at different codon positions. Second, the correlation delay is caused by a mixing of in-phase pattern with a CDS and out-of-phase between neighboring CDSs. The longer the distance, the more contribution from the inter-gene out-of-phase correlation, and the lower correlation value. This changing of mixture proportion is the main reason for the correlation decay. Third, the decay form of the correlation may not be intrinsically exponential. It is possible to construct certain distribution of CDS lengths which lead to non-exponential correlation functions such as linear function. Another potential source of correlation is related to the same codon position. We found that such correlation is not zero for the second, and for the third, codon position. This should be relevant to the weak correlation between amino acids in the coded protein sequences. Overall, Markov chains are not a good model for the correlation in bacterial genomes.

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/Bacteria/ (the file: Escherichia_coli_O127_H6_E2348_69_uid59343/NC_011601.gbk), or from EBI at http://www.sanger.ac.uk/resources/downloads/bacteria/escherichia-coli.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:

$$C_{\alpha,\beta}(d) = P_{\alpha,\beta}(d) - P_\alpha P_\beta \qquad \alpha, \beta = (\mathrm{A,C,G,T}), \quad d = 1, 2, \ldots \tag{1}$$

where $C_{\alpha,\beta}(d)$ is the joint probability of symbol α followed by symbol β d-bases to the right, and P_α (P_β) is the probability in finding symbol α (β) in the sequence. Of these 16 correlation functions, the strand symmetry leads to $C_{\alpha,\beta}(d) \approx C_{\beta',\alpha'}(d)$, where α' is the nucleotide that complement α (e.g. $\alpha = C, \alpha' = G$), and β' complement β. 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 $C_{\alpha,\beta}(d)$ 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

$$C_{\alpha\beta}(d) \sim \exp(-\gamma d) = \exp\left(-\frac{d}{d_0}\right), \tag{2}$$

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