doi:10.4149/gpb_2012_022

Short Communication

Scaling law: A global topological property of genetic sequences

Shou-Liang Bu

School of Physics and Optical Information Technology, Jiaying University, 514015 MeiZhou, People's Republic of China

Abstract. Here, we study the topological properties of genetic sequence by viewing the entire sequence as a whole. First, a systematic way of coding the gene-combinations in a genetic sequence is developed. Next, we apply the coding method to real genetic sequences, and find a scale-free power-law distribution for some particular kinds of gene-combinations. Furthermore, we also present a model to reproduce the observed scale-free feature, which is based on three generic mechanisms: 1) Growth mechanism – genetic sequences expands continuously by the addition of new codon; 2) Preferential replication mechanism – the newly added codon at every time step is a replication of one of existed condons, and the probability that a codon is replicated is proportional to its emerging times in existed sequence; and 3) Mutation mechanism – following (2) the newly added codon has a small probability to mutate. To our knowledge, no report has been published to study the genetic sequences in such way.

As is well-known, genes correspond to units of inheritance, and a gene is considered as a segment of nucleic acid which contains the information necessary to produce a protein (Pearson 2006). Further, a genetic sequence is a succession of four possible letters, i.e., *A*, *C*, *G*, and *T*, which represent the four nucleotide subunits of a DNA strand (RNA in some viruses).

Up to now, there have been many studies concerning local properties of a genetic sequence (Watson et al. 2004). Here, we focus on the global properties of a genetic sequence as a whole. First, a systematic coding method is developed here and applied to research various kinds of possible gene-combinations in genetic sequence. Then, we find that many real genetic sequences have a scale-free power-law distribution for some particular kinds of gene-combinations. Furthermore, to penetrate and reproduce the observed scale-free feature, we also present an evolution model of genetic sequences, which is based on three generic mechanisms. We believe that studies on integral properties must be able to help us to discover more secrets hidden in genetic sequences.

In order to explore the topological properties of genetic sequences, our first step is to code it in a systematic way. This means that we take every q-sequential letters starting from the first letter of a genetic sequence as one unit, which is called a '*vertex*' in the following, and label it with a specified nonnegative integer number ranging from 0 to $4^q - 1$, i.e., a quaternary number consisting of q-bits. For example, if one takes $q = 6$, then every 6-sequential letters is labeled by an integer number ranging from 0 to $4^6 - 1 = 4095$, as shown in Fig. 1. It should be noted that we omit to distinguish between coding and non-coding genes within the same sequence for temporal work, and leave it to further studies.

After coding a genetic sequence, we secondly study the frequency of each vertex (q-sequential letters) emerging in the underlying genetic sequence. We do not consider the correlations between neighboring vertices in the present work and define simply the degree k of a vertex, e.g., i, as the sum of times that this vertex emerges in the sequence. For example, if the 10^{th} vertex, *AAAAGG*, emerges twenty-four times in the underlying sequence, then its degree equals 24, denoted by $k_{10} = 24$.

Assuming that there are $n(k)$ vertices which all have degree k in a genetic sequence, then one gets $\sum_k n(k) = \sum_i k_i$. Defining the degree distribution $P(k)$ as the probability that an arbitrarily specified vertex has degree k, one has $P(k) = n(k)/\sum_k n(k)$. We now study the degree distribution $P(k)$ of real genetic sequences. In the following, we always take $q = 6$*, and this

Correspondence to: Shou-Liang Bu, School of Physics and Optical Information Technology, Jiaying University, 514015 MeiZhou, People's Republic of China
E-mail: slbu@ustc.edu

* In practice, other values of q could also be chosen according to the length of the underlying genetic sequences, e.g., $q = 5$ has been taken to code the sequences considered here and find that there is no clear laws in this case.

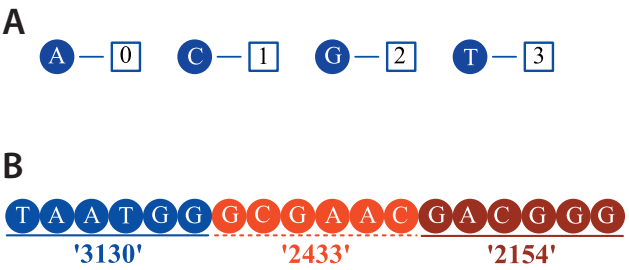

Figure 1. Schematic illustrations of our coding method. **A.** Setting the four letters *A*, *C*, *G*, and *T* to denote the quaternary number 0, 1, 2, and 3, respectively. **B.** Every *q*-sequential letters is taken as one 'vertex', three vertices and their labels specified to them corresponding to a portion of the chloroplast genetic sequence of *Solanum lycopersicum*, here $q = 6$ is taken.

implies that two genes correspond to a vertex. The log-log plots of the degree distributions corresponding to chloroplast genetic sequences of *Solanum lycopersicum*, *Chlorokybus atmophyticu*, *Ranunculus macranthus*, and *Arabidopsis thaliana* are presented in Fig. 2. It is found that they all obey a power-law, i.e., $P(k) \sim k^{-\gamma}$. The dashed lines have slopes $\gamma_{\text{Solanum}} = 3.2$ (A), $\gamma_{\text{Chlorokybus}} = 3.0$ (B), $\gamma_{\text{Ranunculus}} = 3.0$ (C), and $\gamma_{\text{Arabidopsis}} = 2.8$ (D), respectively. It should be emphasized that similar property has been found in the Internet, World Wide Web, and social networks, etc. (Barabási and Albert 1999; Wellman 2001; Albert and Barabási 2002).

In order to understand and reproduce the observed scale-free feature of genetic sequence, we now present a model which is based on three generic mechanisms: 1) Growth mechanism – genetic sequences expand continuously by the addition of new codons; 2) Preferential replication mechanism – the newly added codon is preferentially a replication of one existed gene, and a gene emerging more times in existed sequence is replicated more frequently; and 3) Mutation mechanism – following (2) the newly added codon has a small probability to mutate.

To incorporate the evolution processes of genetic sequences, starting with a small number (n_0) of codons, at

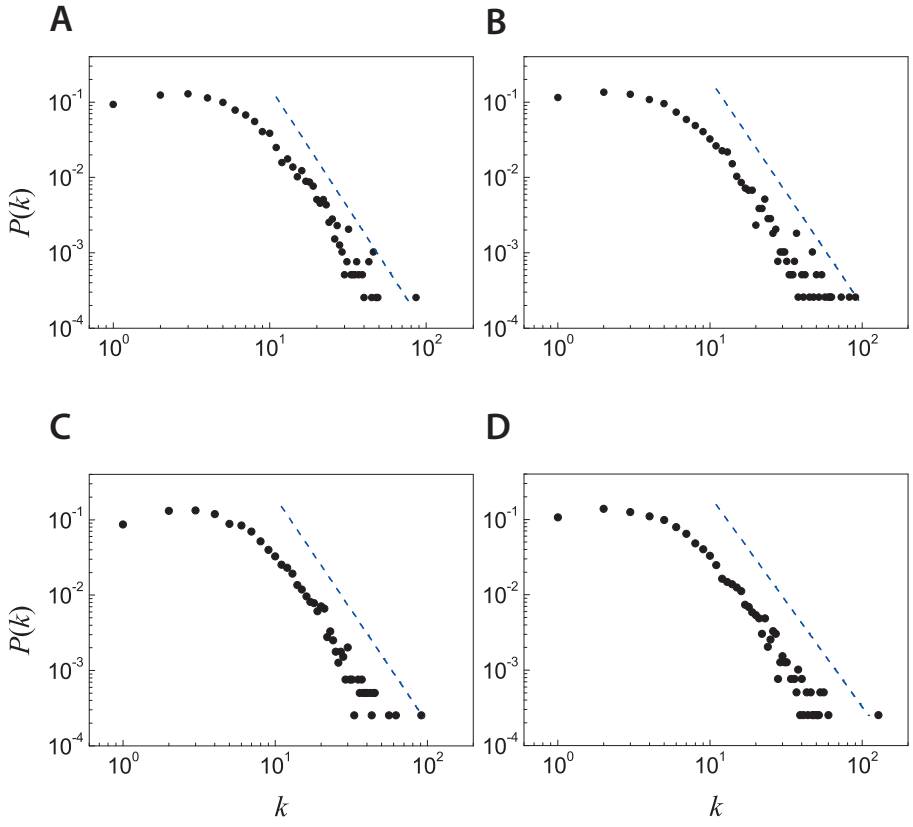

Figure 2. The log-log plots of the degree distributions $P(k)$ corresponding to chloroplast genetic sequences of *Solanum lycopersicum* (**A**), *Chlorokybus atmophyticu* (**B**), *Ranunculus macranthus* (**C**), and *Arabidopsis thaliana* (**D**). It is found that they all obey a power-law, i.e., $P(k) \sim k^{-\gamma}$. The dashed lines have slopes $\gamma_{\text{Solanum}} = 3.2$, $\gamma_{\text{Chlorokybus}} = 3.0$, $\gamma_{\text{Ranunculus}} = 3.0$, and $\gamma_{\text{Arabidopsis}} = 2.8$.

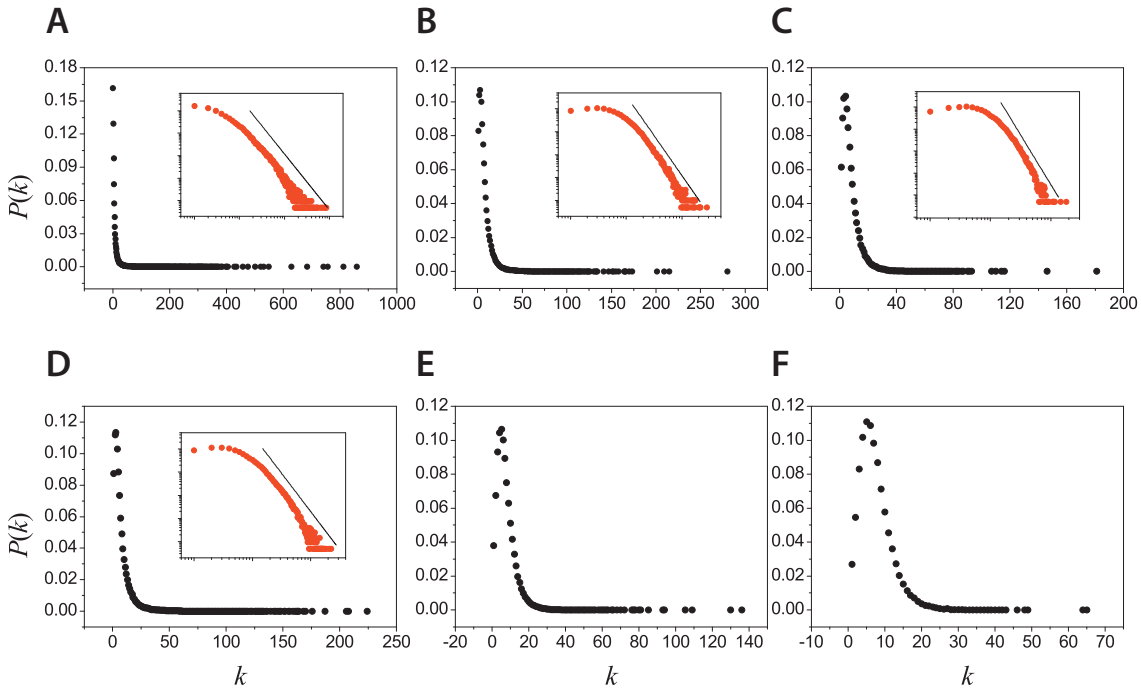

Figure 3. The degree distributions of genetic sequences generated from the model with $n_0 = 1$, $\Lambda = 0.1$ (**A**), $n_0 = 10$, $\Lambda = 0.1$ (**B**), $n_0 = 20$, $\Lambda = 0.1$ (**C**), $n_0 = 1$, $\Lambda = 0.15$ (**D**), $n_0 = 10$, $\Lambda = 0.15$ (**E**), and $n_0 = 20$, $\Lambda = 0.15$ (**F**). The insets in A–D are the corresponding log-log plots, and the solid lines have slopes $\gamma = 2.5$, 3.5, 4.5 and 3.3, respectively. In addition, $t + n_0 = 6 \times 10^4$ for A–F, and the computations are carried out over 100 independent starting configurations. n_0, the initial number of genes; Λ, the mutation probability.

every time step we add a codon to already existed sequence. Next, to incorporate the preferential replication mechanism, we assume that the probability Γ_i that the newly added codon replicates the ith codon is proportional to the emerging times f_i of that codon, so that $\Gamma_i = f_i/\Sigma_j f_j$. Finally, to incorporate the mutation mechanism, at every time step, we let each letter of the newly added codon mutate with probability Λ. After t time steps, the model leads to a complex genetic sequence with $t + n_0$ genes. The scale-free feature emerges for $n_0 << t$ and $\Lambda << 1$. In Fig. 3, we show the degree distributions of genetic sequences generated from the model with $n_0 = 1$, $\Lambda = 0.1$ (A), $n_0 = 10$, $\Lambda = 0.1$ (B), $n_0 = 20$, $\Lambda = 0.1$ (C), $n_0 = 1$, $\Lambda = 0.15$ (D), $n_0 = 10$, $\Lambda = 0.15$ (E), and $n_0 = 20$, $\Lambda = 0.15$ (F), respectively. In addition, the generated sequences contain $t + n_0 = 6 \times 10^4$ codons for all the cases. The insets in Fig. 3A–D are the log-log plots of the corresponding degree distributions. Here, all computations are carried out over 100 independent starting configurations. In addition, if n_0 is not far less than t or one slowly increases the mutation probability Λ, the scaling law gradually disappears. In practice, by comparing results from real sequences and the model, possible topological feature could reflect and provide us useful information concerning the evolutionary history of various genetic sequences.

In summary, a systematic way of coding the gene-combinations in a genetic sequence is developed and applied to several real sequences here. It is found that scale-free power-law distributions for some particular gene-combinations emerge in many real genetic sequences. We also propose a model to reproduce the observed power-law feature. Our aim, on the other hand, is to attract jade by throwing a brick. We believe that any correct theory which is devoted to the origin and evolutionary history of genetic sequences should reasonably expound and reproduce the observed global topological properties for the studied sequence.

Acknowledgements. We acknowledge financial support of this work by Natural Science Foundation of MeiZhou Science and Technology Bureau and Jiaying University under the grant No. 2010KJA13.

References

Albert R., Barabási A.-L. (2002): Statistical mechanics of complex networks. Rev. Mod. Phys. **74,** 47–97 http://dx.doi.org/10.1103/RevModPhys.74.47

Barabási A.-L., Albert R.(1999): Emergence of scaling in random networks. Science **286,** 509–512

http://dx.doi.org/10.1126/science.286.5439.509
Pearson H. (2006): What is a gen. Nature **441,** 399–401
http://dx.doi.org/10.1038/441398a
Watson J. D., Baker T. A., Bell S. P., Gann A., Levine M., Losick R. (2004): Molecular Biology of the Gene. 5th ed., Peason Benjamin Cummings, Cold Spring Harbor Laboratory Press
Wellman B. (2001): Computer networks as social networks. Science 293, 2031–2034
http://dx.doi.org/10.1126/science.1065547

Received: September 26, 2011
Final version accepted: March 26, 2012