

Multi-scale Complexity Analysis on the Sequence of *E. coli* Complete Genome

Jin Wang¹, Qidong Zhang¹

¹State Key Laboratory of Pharmaceutical Biotechnology, School of Life Science,
Nanjing University, Nanjing 210093

Kui Ren^{2,3}, Zhensu She²

²State Key Laboratory for Turbulence Research, Department of Mechanics and
Engineering Science, Beijing University, Beijing 100871

Abstract

We have analyzed the multi-scale density distribution of nucleotides from the complete *Escherichia coli* genome by applying the newly developed hierarchical structure theory for complex systems and the scaling method. A hierarchically similar density distribution of bases from this genome was obtained. Especially we have discovered that G, C density distribution that represent a strong H-bonding between the two DNA chains has a different similarity parameter compared to that of A, T density distribution, indicating the existence of a multi-scale self-organized structure in the genome sequence. The biological significance of these parameters is under investigation.

Keywords: complexity, hierarchical structure, genome, nucleotide, *Escherichia coli*

Introduction

The complete sequencing of human genome and other genome sequences from model organisms^[1,2] are generating a large amount of information in biology. Genomic information will answer many important questions, such as how organisms evolved, what mechanisms are involved in genesis, development and disease, and what the ultimate genetic instructions are to make a human being etc.. New laws of biology will be elucidated by the study on the DNA sequences of the complete genomes in which 4 nucleotides happen alternatively. The DNA sequence of the complete genome from an organism contains the information not only for producing all the proteins (genes) necessary for this organism, but also for assembling those proteins to construct the organism according to the specific time order and three-dimensional pattern. The study of decoding and regulating in biological systems from this kind of information has attracted many research work. For the one dimensional DNA sequence which consists of only 4 kinds of nucleotides, the most common way of investigation is statistical analysis^[3]. The algorithms for DNA sequence alignment and similarity search are developed for the study of phylogeny and evolution of many biological species^[4]. Other ways such as nonlinear theories and cryptology etc. were introduced to the investigation of sequence characterization and Junk DNA^[5,6]. While these investigations are mainly focused on gene sequences or other local segment of DNA sequence, other methods were proposed to measure the

correlation between nucleotides over long distances or over large scale along DNA chain, such as statistical physics analysis^[7,8], spectrum analysis^[9~11], wavelet analysis^[12,13], etc.. These studies concluded the existence of long correlation in DNA sequence^[14] though there are many controversial points^[15]. The similarity parameter in this study is also a type of quantitative description on the multi-scale correlation in DNA sequence.

Life system is a typical complex system, which is characterized as of both regularity and randomness^[16], for which the foundation at molecular level may be associated with the complexity of multi-scale density distribution of the 4 nucleotides along DNA chain. Thus these sequence data provide interesting information for the investigation on the complex functioning of organisms and for the approach of quantitative models. This will hopefully be of great help for obtaining a working frame for the origin and evolution of life as well as the regulation of biological processes. *Escherichia coli* was first isolated in 1922 and became a major model system for biochemical genetics, molecular biology, and biotechnology. Blattner et al. reported that they have sequenced the entire 4.6-megabase genome of K-12 strain of *E. coli* and provided a foldout with the arrangement of 4288 putative and known genes, operons, promoters, and protein binding sites^[17]. A newly developed theory of hierarchical structure for complex system and the scaling method^[18,19] were applied in *E. coli* genome and a hierarchically similar density distribution of bases and base combinations from this genome was obtained. Especially we have discovered that G, C density distribution that represent a strong H-bonding between the two DNA chains has a different similarity parameter compared to that of A, T density distribution. The biological significance of these parameters is under investigation. This work may imply a new method for the theoretical approach to the complexity of the structure of genomes and the self-organization of genomes. With the availability of more complete genome sequences, this will also facilitate studies of biological evolution.

Theory

Based on the theoretical model for the analysis on hierarchical structure in turbulence system proposed by She & Leveque^[18~21], we have derived a theoretical model for the analysis on hierarchical structure in complete genomes. First, we define

$\rho_L = \frac{1}{L} \sum_{i=1}^L \delta_{u_i, A}$ as the variant of density fluctuation of certain nucleotide (say, A)

over scale L (the number of nucleotides in a window is L), where $\delta_{u_i, A}$ is Kronecher symbol, if $u_i = A$, then $\delta_{u_i, A} = 1$, or $\delta_{u_i, A} = 0$. We then get the p th-order moment

of ρ_L

$$S_p(L) = \int (\rho_L)^p P(\rho_L) d(\rho_L), \quad (1)$$

where $P(\rho_L)$ is the distribution of probability density of ρ_L . A set of $P(\rho_L)$ will be defined with different scale (L) for a complete genome sequence. Next, we introduce $F_p(L)$ to describe the p th-order hierarchical structure

$$F_p(L) = \frac{S_{p+1}(L)}{S_p(L)} = \frac{\int (\rho_L)^{p+1} P(\rho_L) d(\rho_L)}{\int (\rho_L)^p P(\rho_L) d(\rho_L)}, \quad (2)$$

which just has a dimension of density and is a function of p and L . $F_p(L)$ represents the intensity of the p th-order density fluctuation over scale L .

Scaling exponent ζ_p is so defined as $S_p(L) \propto L^{\zeta_p}$, which represents the variation of p th-order fluctuation with changing scale L . We introduce the principle of relative scaling here. It has been indicated^[22] that the p th-order moment of certain physical variant has an exponential relation to its q th-order moment

$$S_p(L) \propto S_q(L)^{\zeta_{p,q}}, \quad (3)$$

when scaling $S_p(L) \propto L^{\zeta_p}$ was not satisfied. This phenomenon means that, in the case that scaling is not satisfied, different orders of moment of physical variants have the same deviation from the exponential scaling while the relative relation of dependence between them keeps unchanged. This kind of relativism is named extended self-similarity (ESS), and $\zeta_{p,q}$ is called relative scaling exponent.

Next, according to an important assumption that the hierarchical structure over different scale is of statistically self-similarity (hidden symmetry), for the multi-scale fluctuation structure in a self-organized state,

$$\frac{F_{p+1}(L)}{F_\infty(L)} = A_p \left(\frac{F_p(L)}{F_\infty(L)} \right)^\beta \quad (4)$$

in which, $0 < \beta < 1$, is a constant independent of p and L , called intermittent parameter. A_p relies on p , and is independent of L . $F_\infty(L) = \lim_{p \rightarrow \infty} F_p(L)$. $F_\infty(L)$

characterizes the most intermittent structures.

Hierarchical similarity^[20] indicates that the difference between the multi-scale fluctuation intensities of different hierarchies has recursion invariance of hierarchical similarity, implying that **high intensity can be deduced from low intensity**. Hierarchical similarity indicates a self-organization of the system. The biological interpretation of the parameter β have not been very clear. Basically, hierarchical similarity parameter β indicates the extent of similarity between different scales. When $\beta \rightarrow 1$, there is a strict similarity. $\beta \rightarrow 0$ means the existence of only the most intermittent structure, whereas $\beta \neq 0,1$ indicates the variation of the intensity of fluctuation

structure over different scales. When $\beta < 1$, the difference between intensities of adjacent high hierarchies decreases with increasing order of hierarchy. For high-order moments p , the most intermittent structure characterizes the system.

From (4), $\frac{F_p(L)}{F_1(L)}$ and $\frac{F_{p+1}(L)}{F_2(L)}$ are calculated directly from data. Then plot $\text{Log}_{10} \frac{F_p(L)}{F_1(L)}$ versus $\text{Log}_{10} \frac{F_{p+1}(L)}{F_2(L)}$, slope β can be obtained by a least square fitting. This manipulation is called β -test^[21]. If a multi-scale signal can satisfy the linear relation with high confidence, it is hierarchically similar. Note that the scale exponents here are determined through ESS (Extended Self-Similarity).

Results and Discussion

The DNA sequence of the complete genome of *E.coli* (K-12) was retrieved from GenBank from a mirror site (<http://www.cbi.pku.edu.cn/>). The genome has a total of 4639221 bp (base pairs) nucleotides. The average interval between genes is 118bp. Coding area accounts for 87.8% of the genome DNA sequence. Let N be the sum of the number of one of the 4 nucleotides A, G, C, T or combinations of the 4 characters over scale L (a DNA segment of length L), so that the correspond density in this window scale is N/L , i.e., ρ_L . Then move the window along the whole sequence and normalize those items to get the density distributions of 4 nucleotides and their combination items respectively. For different scale L , a set of density distributions can be obtained by the same way. Figure 1 shows the density distribution of A (adenosine) over DNA sequence 3000 to 7000 when $L=100$. Density distribution fluctuations were also observed for G, C, T and their combinations over different scales. In the following analysis, we have chosen the range of scale L to be 100-8000bps in which relative scaling and hierarchical similarity were satisfied.

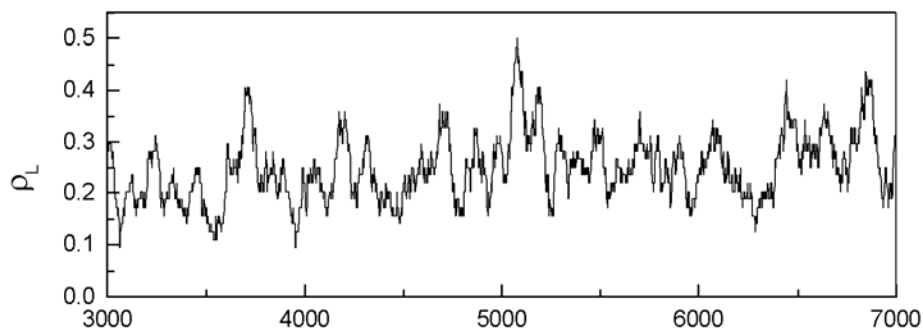


Fig.1 Density distribution of nucleotide A (adenine) over *E.coli* genome sequence 3000-7000 at scale $L=100$.

The data produced were put into hierarchical structure analysis over the complete *E.coli* genome DNA sequence by using the theoretical model described in last section. Figure 2 shows the result for relative scaling analysis on nucleotide A

over the complete *E.coli* genome sequence. Fig.2(a) gives the representative results of relative scaling of 5th and 6th to 3rd-order respectively, that shows relative scaling is satisfied, indicating a hierarchical similarity. Fig.2(b) shows the relationship of scaling exponent to the order. The non-linear relationship indicates that the fluctuation of the intensity of A density is fractal, and has a non-Gaussian statistical property. The relative scaling exponents (relative to 3rd-order) of density fluctuation of the 4 kinds of bases (A, T, G, C) were listed in table 1.

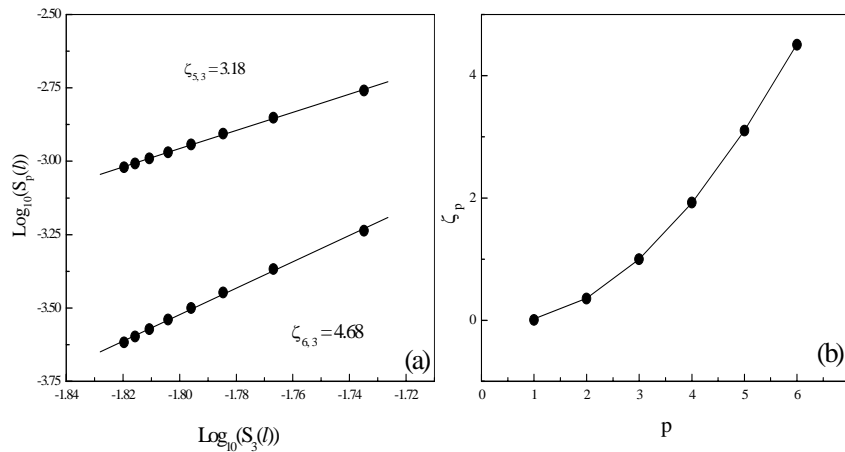


Fig.2 Relative scaling analysis on adenine (A) over *E.coli* genome sequence.
 (a) Relative scaling of 5th to 3rd-order with the relative exponent to be 3.18 and 6th to 3rd-order with the relative exponent to be 4.68;
 (b) Plot of scaling exponent to the order.

Figure 3 gives the result for β -test on the density distribution of nucleotides A and C over *E.coli* genome. The linear fitting with a slope $\beta = 0.931$ for A gives a high confidence, indicating a precise hierarchical similarity. Similar results were obtained from the hierarchical structure analysis of other bases G and T. Table 2 gives β -value of all the 4 bases in *E.coli* genome from β -test.

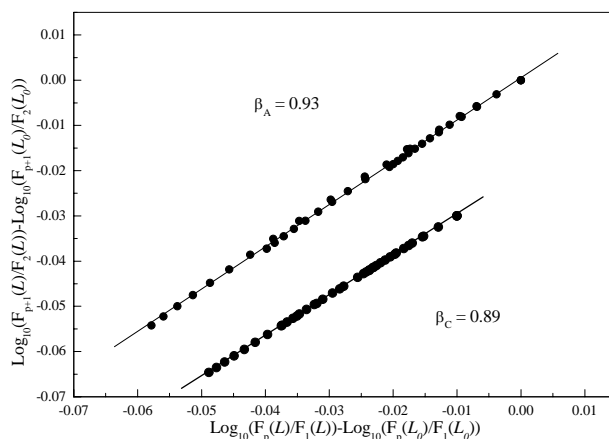


Fig 3 β -test of the multi-scale hierarchical analysis of density distribution of A and C over

E.coli genome sequence. The hierarchical structure parameters are β_A and β_C for A and C respectively, and indicated near their fitting plots.

Table1 Relative scaling exponents of base density distribution of *E.coli* genomic sequence*

	$\zeta_{2,3}$	$\zeta_{3,3}$	$\zeta_{4,3}$	$\zeta_{5,3}$	$\zeta_{6,3}$
A	0.35	1	1.95	3.18	4.68
T	0.35	1	1.95	3.17	4.67
G	0.36	1	1.89	3.00	4.31
C	0.36	1	1.91	3.05	4.41

* Relative to 3rd order.

Table 2 β -value of hierarchical structure analysis on the base density distribution of *E.coli* genomic sequence

	A	T	G	C
β	0.93	0.94	0.90	0.89

From the analysis above, we may divide A, T, C and G into 2 groups, A/T and G/C, according to their values of hierarchical parameter from density distributions in genome. The differences inside the groups are obviously less than between the groups. Because G/C-rich region represents gene-rich region, the difference between A/T and G/C shows the difference between coding region and non-coding region to some extent, i.e., non-coding region is strongly correlated in the long-range^[10]. As G pairs with C by 3 hydrogen bonds and A pairs with T by 2 hydrogen bonds, the difference in the chemical mechanism and function results in the difference in their structure of self-organization. This grouping phenomenon is perhaps a reasonable representation of the symmetry of DNA double-strand as bases are complement. The analysis of hierarchical structure on the 4 bases of *E.coli* genome sequence shows that hierarchical similarity is valid on the scale of the typical size inside a gene (100bp) to common size of gene cluster (8000bp), indicating the possibility of existence of complex structure of gene distribution in genome.

In fact, hierarchical similarity reflects strong base compositional bias in DNA sequence^[23~25]. This bias makes the statistical properties of sequence far from Gauss. We believe that base compositional bias has strong correlation with biological origin and evolution, though the reason why there appears base compositional bias and the biological significance of this bias are not well known^[23,25].

We conclude from the analysis above that *E.coli* genome sequence has strong intermittent multi-scale structure. In this sequence, we found an extended self-similarity on the scale from inside a gene (100bp) to a group of genes (8000bp) when certain physical variants were put into investigation. Hierarchical similarity was also satisfied for the sequence. This indicates that the base density distribution has

evolved into certain statistical self-organized state during the long history of evolution. The statistical state can be depicted quantitatively by parameter of hierarchical similarity. We find that β -value of *E.coli* genome is much complicate than that of other natural phenomenon of complex such as turbulence. This implies in some senses that biological sequences assume a more apparent diversity in the multi-scale structure. Life system can be regarded as a set of numerous events happened at low probability. As the theory of hierarchical structure characterizes specifically the multi-scale properties of events with low probability and high fluctuation, it is possible for the theory to serve as a quantitative model for representing the complete biological information of organisms. As non-coding regions are responsible for some fundamental biological functions, such as replication, cleavage, recombination and the maintenance of the stability of chromosomes, it retains a high abundance and low amount of information by natural selection in order to prevent or reduce the errors brought to the regions. Parameters of hierarchical structure may serve as characteristic measure of complex self-organization of genome, and can be used to study systemically the hierarchical structures of sequences from genomes of different species. For the high compositional bias of bases in genomes, this theory can also be used to study the hierarchical structure of coding region and non-coding region.

Acknowledgement: Project supported by NSF of China (No. 79970116, 39970184).

References:

1. Dunham L, Shimizu N, Roe B A, et al. The DNA sequence of human chromosome 22. *Nature*, 1999, 402: 489-495
2. Gerald M R, Mark D Y, Jennifer R W et al. Comparative Genomics of the Eukaryotes. *Science*, 2000, 287: 2204-2215
3. Reddy BV, Pandit M W, A. statistical analytical approach to decipher information from biological sequences: application to murine splice-site analysis and prediction. *J Biomol Struct Dyn.*, 1995, 12(4):785-801
4. WANG Ning, CHEN Runsheng Comparative analysis of phylogeny based on intron and exon. *Chinese Science Bulletin*, 1999, 44(19):2095-2102 (In Chinese)
5. Gtaziano P, Marcella A, Cecilia S. Linguistic Analysis of Nucleotide Sequences: Algorithms for Pattern Recognition and Analysis of Codon Strategy. *Method in Enzymology*, 1996, 266: 281-294
6. Flam F. Hints of a language in Junk DNA. *Science*, 1994, 266: 1320
7. Peng C K, Buldyrev S V, Goldberger A L et al. Long-range correlations in nucleotide sequences. *Nature*, 1992, 356:168 ~170
8. Lu X, Sun Z, Chen H, et al. Characterizing self-similarity in bacteria DNA sequences. *Phys. Rev. Series E*, 1998, 58: 3578-3584
9. Voss R F. Evolution of Long-Range Fractal Correlation and 1/f Noise in DNA Base sequences. *Phys. Rev. Lett.*, 1992, 68:3805~3808
10. Buldyrev S V, Goldberger A L, Havlin S, et al. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. Series E*, 1995,

51:5084~5091

11. Li W, Marr T and Kaneko K. Understanding long-range correlations in DNA sequences. *Physica Series D*, 1994, 75:392~416
12. Arneodo A, Bacry C, Graves V, et al. Characterizing Long-range Correlation in DNA sequences from Wavelet Analysis. *Phys. Rev. Lett.*, 1995, 74:3293~3296
13. Arneodo A, D'Aubenton-Carafa Y, Audit B, et al. What can we learn with wavelets about DNA sequences. *Physica Series A*, 1998, 249:439~448
14. Li W. The study of correlation structures of DNA sequences: a critical review. *Computers Chem.*, 1997, 21(4):257~271
15. Maddox J. Long-range correlations within DNA. *Nature*, 1992, 358: 103~103
16. Grassberger P. Toward a quantitative theory of self-generated complexity. *Inter J Theor Phys*, 1986, 25(9): 907~938
17. Blattner F R, Plunkett III G, Bloch A C, et al. The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, 1997, 277: 1453~1462
18. She Z S and Leveque E. Universal scaling laws in fully developed turbulence. *Phys. Rev. Lett.*, 1994, 72: 336~339
19. She Z S and Waymire E. Quantized energy cascade and Log-Poisson statistics in fully developed turbulence. *Phys. Rev. Lett.*, 1995, 74: 262~265
20. She Z S, Hierarchical Structures and scaling in turbulence. *Lecture Notes in Physics*, 1997, 491: 28~52
21. She Z S, Ren K, Lewis G S, et al. Scaling and structures in turbulent Couette-Taylor flow. *Phys. Rev. Series E*, 2001, (In press)
22. Bershadskii A. Generalized scaling in nonscaling diffusion. *Physica Series A*, 2000, 278: 497~503
23. Bernaola-Galvan P, Oliver J L, Roman-Roldan R. Decomposition of DNA sequence Complexity. *Phys. Rev. Lett.*, 1999, 83:3336~3339
24. Gautier C. Compositional bias in DNA. *Curr. Opin. Genet. Dev.*, 2000, 10:656~661
25. Roman-Roldan R, Bernaola-Galvan P, Oliver J L. Sequence Compositional Complexity of DNA through Entropic Segmentation Method. *Phys. Rev. Lett.*, 1998, 80:1344~1347