Correlation property of length sequences based on global structure of the complete genome

Zu-Guo Yu,[1,2]* V. V. Anh,[1] and Bin Wang[3]
[1]*Centre in Statistical Science and Industrial Mathematics, Queensland University of Technology,
GPO Box 2434, Brisbane Q 4001, Australia*
[2]*Department of Mathematics, Xiangtan University, Hunan 411105, People's Republic of China*[†]
[3]*Institute of Theoretical Physics, Academia Sinica, P.O. Box 2735, Beijing 100080, People's Republic of China*

(Received 26 June 2000; revised manuscript received 28 August 2000; published 20 December 2000)

This paper considers three kinds of length sequences of the complete genome. Detrended fluctuation analysis, spectral analysis, and the mean distance spanned within time L are used to discuss the correlation property of these sequences. The values of the exponents from these methods of these three kinds of length sequences of bacteria indicate that the long-range correlations exist in most of these sequences. The correlations have a rich variety of behaviors including the presence of anti-correlations. Furthermore, using the exponent γ, it is found that these correlations are all linear ($\gamma = 1.0 \pm 0.03$). It is also found that these sequences exhibit $1/f$ noise in some interval of frequency ($f > 1$). The length of this interval of frequency depends on the length of the sequence. The shape of the periodogram in $f > 1$ exhibits some periodicity. The period seems to depend on the length and the complexity of the length sequence.

DOI: 10.1103/PhysRevE.63.011903 PACS number(s): 87.14.Gg, 87.10.+e, 47.53.+n

I. INTRODUCTION

Recently, there has been considerable interest in the finding of long-range correlation (LRC) in DNA sequences [1–16]. Li *et al.* [1] found that the spectral density of a DNA sequence containing mostly introns shows $1/f^{\beta}$ behavior, which indicates the presence of LRC. The correlation properties of coding and noncoding DNA sequences were first studied by Peng *et al.* [2] in their fractal landscape or DNA walk model. The DNA walk defined in [2] is that the walker steps ''up'' if a pyrimidine (C or T) occurs at position i along the DNA chain, while the walker steps ''down'' if a purine (A or G) occurs at position i. Peng *et al.* [2] discovered that there exists LRC in noncoding DNA sequences while the coding sequences correspond to a regular random walk. By doing a more detailed analysis, Chatzidimitriou, Dreismann and Larhammar [5] concluded that both coding and noncoding sequences exhibit LRC. A subsequent work by Prabhu and Claverie [6] also substantially corroborates these results. If one considers more details by distinguishing C from T in pyrimidine, and A from G in purine (such as two- or three-dimensional DNA walk model [8] and maps given in [9]), then the presence of base correlation has been found even in coding sequences. In view of the controversy about the presence of correlation in all DNA or only in noncoding DNA, Buldyrev *et al.* [14] showed the LRC appears mainly in noncoding DNA using all the DNA sequences available. Alternatively, Voss [10], based on equal-symbol correlation, showed a power-law behavior for the sequences studied regardless of the percent of intron contents. Investigations based on different models seem to suggest different results, as they all look into only a certain aspect of the entire DNA sequence. It is therefore important to investigate the degree of correlations in a model-independent way.

Since the first complete genome of the free-living bacterium *Mycoplasma genitalium* was sequenced in 1995 [17], an ever-growing number of complete genomes has been deposited in public databases. The availability of complete genomes induces the possibility to ask some global questions on these sequences. The avoided and under-represented strings in some bacterial complete genomes have been discussed in [18–20]. A time series model of coding sequence in complete genome has also been proposed in [21]. Maria de Sousa Vieira [22] has done a low-frequency analysis of complete DNA of 13 microbial genomes and showed that its fractal behavior does not always prevail through the entire chain, and that the autocorrelation functions have a rich variety of behaviors including the presence of anti-correlations.

For the importance of the numbers, sizes and ordering of genes along the chromosome, one can refer to Part 5 of the famous book of Lewin (Ref. [23]). Hence one may ignore the composition of the four kinds of bases in coding and noncoding segments and only consider the rough structure of the complete genome or long DNA sequences. Provata and Almirantis [24] proposed a fractal Cantor pattern of DNA. They map coding segments to filled regions and noncoding segments to empty regions of a random Cantor set and then calculate the fractal dimension of the random fractal set. They found that the coding/noncoding partition in DNA sequences of lower organisms is homogeneous-like, while in the higher eucariotes the partition is fractal. This result seems too rough to distinguish bacteria because the fractal dimensions of bacteria they gave out are all the same. The classification and evolution relationship of bacteria is one of the most important problems in DNA research. Yu and Anh [25] proposed a time series model based on the global structure of the complete genome and considered three kinds of length sequences. After calculating the correlation dimensions and Hurst exponents, it was found that one can get more information from this model than that of the fractal Cantor pat-

*Corresponding author. Email address: yuzg@hotmail.com or z.yu@qut.edu.au

[†]Permanent corresponding address of Zu-Guo Yu.

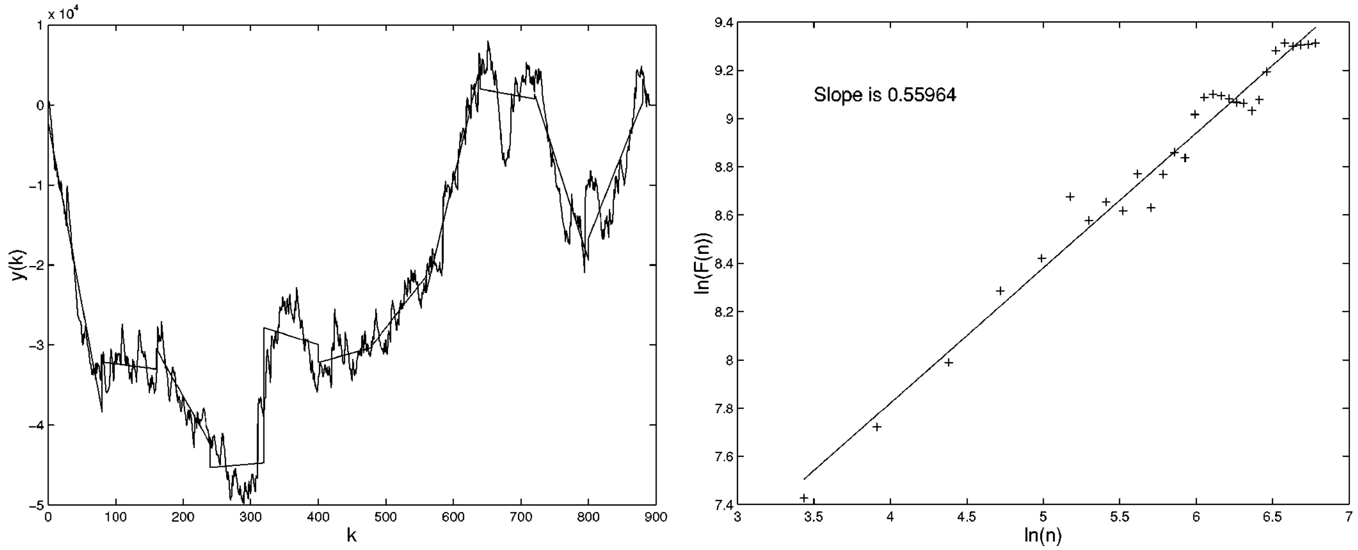

FIG. 1. An example to show how to do detrended fluctuation analysis. (Left) To get the sequence $y_n(k)$. (Right) To get the exponent α using least-square linear fit.

tern. Some results on the classification and evolution relationship of bacteria were found in [25]. Naturally it is desirable to know if there exists LRC in these length sequences. The quantification of these correlations could give insight to the role of the ordering of genes on the chromosome, which is far from irrelevant for gene function. We attempt to answer this question in this paper.

Viewing from the level of structure, the complete genome of an organism is made up of coding and noncoding segments. Here the length of a coding/noncoding segment means the number of its bases. Based on the lengths of coding/noncoding segments in the complete genome, we can get three kinds of integer sequences by the following ways.

(i) First we order all lengths of coding and noncoding segments according to the order of coding and noncoding segments in the complete genome, then replace the lengths of noncoding segments by their negative numbers. This allows to distinguish lengths of coding and noncoding segments. This integer sequence is named *whole length sequence*.

(ii) We order all lengths of coding segments according to the order of coding segments in the complete genome. We name this integer sequence *coding length sequence*.

(iii) We order all lengths of noncoding segments according to the order of noncoding segments in the complete genome. This integer sequence is named *noncoding length sequence*.

We can now view these three kinds of integer sequences as time series. In the following, we will investigate the correlation property through *Detrended Fluctuation Analysis* (DFA) [26] and spectral analysis.

II. DETRENDED FLUCTUATION ANALYSIS AND SPECTRAL ANALYSIS

We denote a time series as $X(t), t=1, \ldots, N$. First the time series is integrated as $y(k)=\Sigma_{t=1}^{k}[X(t)-X_{ave}]$, where X_{ave} is the average over the whole time period. Next, the integrated time series is divided into boxes of equal length, n. In each box of length n, a least-squares line is fit to the data, representing the trend in that box. The y coordinate of the straight line segments is denoted by $y_n(k)$. We then detrend the integrated time series, $y(k)$, by subtracting the local trend, $y_n(k)$, in each box. The root-mean-square fluctuation of this integrated and detrended time series is calculated as

$$F(n)=\sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k)-y_n(k)]^2}. \tag{1}$$

Typically, $F(n)$ will increase with box size n. A linear relationship on a double log graph indicates the presence of scaling

$$F(n)\propto n^{\alpha}. \tag{2}$$

Under such conditions, the fluctuations can be characterized by the scaling exponent α, the slope of the line relating $\ln F(n)$ to $\ln n$. For uncorrelated data, the integrated value $y(k)$ corresponds to a random walk, and therefore, $\alpha=0.5$. A value of $0.5<\alpha<1.0$ indicates the presence of LRC so that a large interval is more likely to be followed by a large interval and *vice versa*. In contrast, $0<\alpha<0.5$ indicates a different type of power-law correlations such that large and small values of time series are more likely to alternate. For examples, we give the DFA of the coding length sequence of *A. aeolicus* in Fig. 1.

Now we analyze the time series using the quantity $M(L)$, the mean distance a walker spanned within time L. Dunki and Ambuhl [27,28] used this quantity to discuss the scaling property in temporal patterns of schizophrenia. Denote

$$W(j):=\sum_{t=1}^{j}[X(t)-X_{ave}], \tag{3}$$

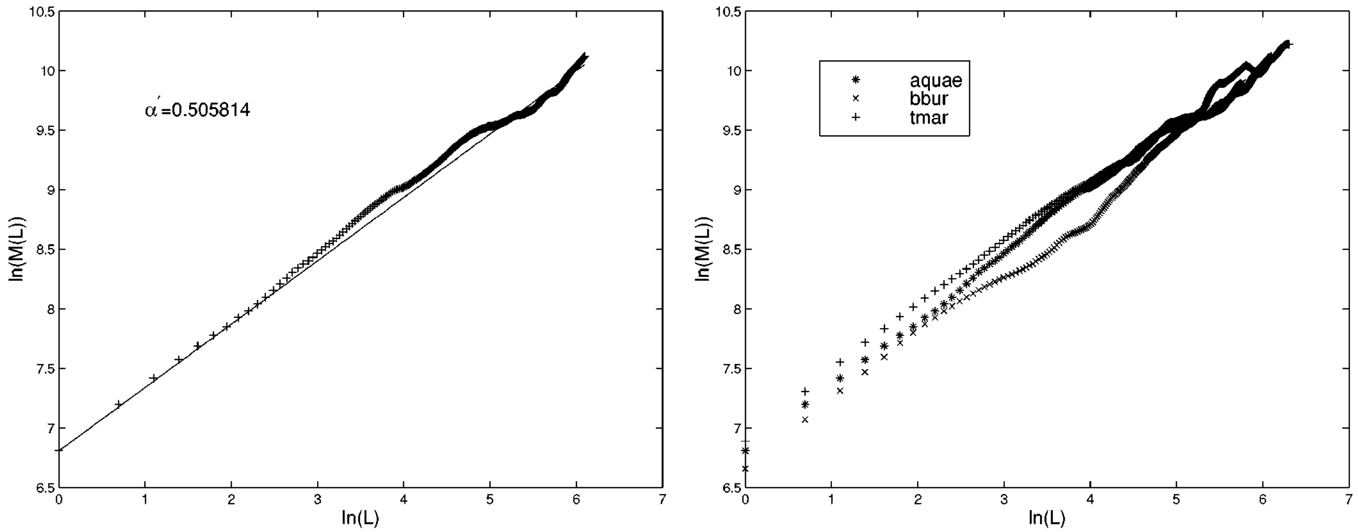

FIG. 2. (Left) To get the exponent α' using least-square linear fit. (Right) The analysis of coding length sequences of three bacteria using mean distance a walker spanned within time L.

from which we get the walks

$$M(L) := \langle |W(j) - W(j+L)| \rangle_j, \tag{4}$$

where $\langle\ \rangle_j$ denotes the average over j, and $j=1,\ldots,N-L$. The time shift L typically varies from $1,\ldots,N/2$. From a physics viewpoint, $M(L)$ might be thought of as the variance evolution of a random walker's total displacement mapped from the time series $X(t)$. $M(L)$ may be assessed for LRC [29] [e.g., $M(L) \propto L^{\alpha'}$, $\alpha' = 1/2$ corresponding to the random case]. We give some examples to estimate the scale parameter α' in Fig. 2.

Dunki *et al.* [28] proposed the following scale which seems to perform better than the scale α'. The definition

$$W'(j) := \sum_{t=1}^{j} |X(t) - X_{ave}| \tag{5}$$

leads to

$$M'(L) := \langle |W'(j) - W'(j+L)| \rangle_j. \tag{6}$$

Analyses of test time series showed that Eq. (6) are more robust against distortion or discretization of the corresponding amplitudes $X(t)$ than Eq. (4). From the ln(L) vs $\ln(M'(L))$ plane, we find the relation

$$M'(L) \propto L^{\gamma}. \tag{7}$$

The exponent γ measures only the presence of nonlinear correlations and remains equal to unity for all sequences with only linear correlations. We carried out this kind of analysis on coding length sequences of *A. aeolicus, B. burgdorferi* and *T. maritima*. The results are reported in the left figure of Fig. 3.

We also consider the discrete Fourier transform [30] of the time series $X(t)$, $t=1,\ldots,N$ defined by

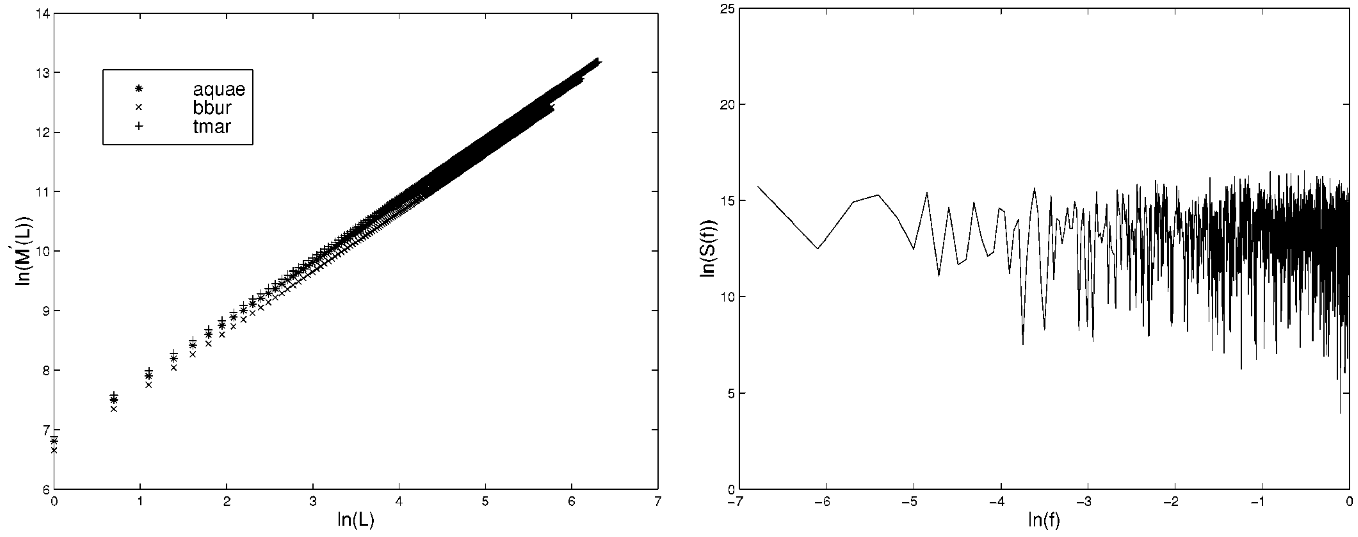

FIG. 3. (Left) Estimate the scale γ. (Right) An example of spectral analysis of low frequencies $f<1$.

TABLE I. α_{whole}, α_{cod}, and α_{noncod} of 21 bacteria.

Bacteria	Category	α_{whole}	α_{cod}	α_{noncod}
Rhizobium sp. NGR234	Proteobacteria	0.24759	0.11158	0.34342
Mycoplasma genitalium	Gram-positive Eubacteria	0.37003	0.25374	0.18111
Chlamydia trachomatis	Chlamydia	0.42251	0.37043	0.49373
Thermotoga maritima	Hyperthermophilic bacteria	0.43314	0.47659	0.49279
Mycoplasma pneumoniae	Gram-positive Eubacteria	0.44304	0.45208	0.49922
Pyrococcus abyssi	Archaebacteria	0.48568	0.39271	0.42884
Helicobacter pylori J99	Proteobacteria	0.48770	0.43562	0.42089
Helicobacter pylori 26695	Proteobacteria	0.49538	0.37608	0.41374
Haemophilus influenzae	Proteobacteria	0.49771	0.42432	0.53013
Rickettsia prowazekii	Proteobacteria	0.49950	0.33089	0.51923
Chlamydia pneumoniae	Chlamydia	0.53982	0.53615	0.38085
Methanococcus jannaschii	Archaebacteria	0.54516	0.58380	0.34482
M. tuberculosis	Gram-positive Eubacteria	0.55621	0.57479	0.52949
Aeropyrum pernix	Archaebacteria	0.57817	0.63248	0.44829
Bacillus subtilis	Gram-positive Eubacteria	0.58047	0.59221	0.54480
Borrelia burgdorferi	Spirochaete	0.58258	0.53687	0.51815
Archaeoglobus fulgidus	Archaebacteria	0.59558	0.59025	0.46596
Aquifex aeolicus	Hyperthermophilic bacteria	0.59558	0.55964	0.43141
Escherichia coli	Proteobacteria	0.60469	0.62011	0.52000
M. thermoautotrophicum	Archaebacteria	0.62055	0.64567	0.38825
Treponema pallidum	Spirochaete	0.67964	0.70297	0.60914

$$\hat{X}(f)=N^{-(1/2)}\sum_{t=0}^{N-1}X(t+1)e^{-2\pi ift}, \tag{8}$$

then

$$S(f)=|\hat{X}(f)|^{2} \tag{9}$$

is the *power spectrum of* $X(t)$. In recent studies, it has been found [31] that many natural phenomena lead to the power spectrum of the form $1/f^{\beta}$. This kind of dependence was named $1/f$ noise, in contrast to white noise $S(f)=\text{const}$, i.e., $\beta=0$. Let the frequency f take k values $f_k=k/N$, $k=1,\ldots,N$. From the ln(f) vs ln($S(f)$) graph, the existence of $1/f^{\beta}$ does not seem apparent. For example, we give the figure of the coding length sequence of *A. aeolicus* on the right of Fig. 3.

When we use the least squares line to fit data, we need to consider the errors. If the data are $\{(x_i,y_i)\}_{i=1}^{n}$, we can define the *coefficient of linear correlation* as [32]

$$r=\frac{\Sigma_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{(\Sigma_{i=1}^{n}(x_i-\overline{x})^2\Sigma_{i=1}^{n}(y_i-\overline{y})^2}}, \tag{10}$$

where $\overline{x}$ and $\overline{y}$ are the average of the values $\{x_i\}_{i=1}^{n}$ and $\{y_i\}_{i=1}^{n}$ respectively. If $r=\pm 1$, then the points lie exactly on a straight line; that is, there is a perfect linear relationship between x and y. If $r=0$, there is no linear relationship. The quantity r measures the strength of linear relationships between x and y. The values of r in figures of obtaining exponents α, α', and β are 0.987 685, 0.994 993 9, and 3.791 8E-03 respectively. Hence we can see the data given by exponents α and α' are more convincing than that given by exponent β.

III. DATA AND RESULTS

More than 21 bacterial complete genomes are now available in public databases. There are five Archaebacteria: *Archaeoglobus fulgidus* (aful), *Pyrococcus abyssi* (pabyssi), *Methanococcus jannaschii* (mjan), *Aeropyrum pernix* (aero) and *Methanobacterium thermoautotrophicum* (mthe); four Gram-positive Eubacteria: *Mycobacterium tuberculosis* (mtub), *Mycoplasma pneumoniae* (mpneu), *Mycoplasma genitalium* (mgen), and *Bacillus subtilis* (bsub). The others are Gram-negative Eubacteria. These consist of two Hyperthermophilic bacteria: *Aquifex aeolicus* (aquae) and *Thermotoga maritima* (tmar); six Proteobacteria: *Rhizobium sp. NGR234* (pNGR234), *Escherichia coli* (ecoli), *Haemophilus influenzae* (hinf), *Helicobacter pylori J99* (hpyl99), *Helicobacter pylori 26695* (hpyl) and *Rickettsia prowazekii* (rpxx); two Chlamydia: *Chlamydia trachomatis* (ctra) and *Chlamydia pneumoniae* (cpneu), and two Spirochaete: *Borrelia burgdorferi* (bbur) and *Treponema pallidum* (tpal).

We calculate scales α,α',β of low frequencies ($f<1$) and γ of three kinds of length sequences of the above 21 bacteria. The estimated results are given in Table I (where we denote by α_{whole}, α_{cod}, and α_{noncod} the scales of DFA of the whole, coding and noncoding length sequences, from top to bottom, in the increasing order of the value of α_{whole}), Table II [where we denote by α'_{whole}, α'_{cod}, and α'_{noncod} the scales of $M(L)$ of the whole, coding and noncoding length

TABLE II. α'_{whole}, α'_{cod}, and α'_{noncod} of 21 bacteria.

Bacteria	Category	α'_{whole}	α'_{cod}	α'_{noncod}
Rhizobium sp. NGR234	Proteobacteria	0.17021	0.11223	0.28573
Chlamydia trachomatis	Chlamydia	0.172340	0.23801	0.66431
M. tuberculosis	Gram-positive Eubacteria	0.20185	0.18451	0.43716
Mycoplasma genitalium	Gram-positive Eubacteria	0.21632	0.25185	0.25971
Escherichia coli	Proteobacteria	0.25837	0.24567	0.62126
Pyrococcus abyssi	Archaebacteria	0.29809	0.18061	0.48169
Bacillus subtilis	Gram-positive Eubacteria	0.36791	0.46816	0.55325
Mycoplasma pneumoniae	Gram-positive Eubacteria	0.37148	0.46475	0.46829
Chlamydia pneumoniae	Chlamydia	0.37216	0.26939	0.50734
Rickettsia prowazekii	Proteobacteria	0.41040	0.23109	0.50930
Archaeoglobus fulgidus	Archaebacteria	0.43149	0.35370	0.60835
Helicobacter pylori 26695	Proteobacteria	0.44082	0.38500	0.39325
Haemophilus influenzae	Proteobacteria	0.46121	0.44842	0.34842
Aeropyrum pernix	Archaebacteria	0.46203	0.45520	0.24850
M. thermoautotrophicum	Archaebacteria	0.48038	0.48870	0.36249
Thermotoga maritima	Hyperthermophilic bacteria	0.49453	0.50457	0.27005
Aquifex aeolicus	Hyperthermophilic bacteria	0.50237	0.50582	0.31488
Helicobacter pylori J99	Proteobacteria	0.54547	0.50999	0.48640
Treponema pallidum	Spirochaete	0.56357	0.56808	0.65350
Borrelia burgdorferi	Spirochaete	0.61186	0.58016	0.61772
Methanococcus jannaschii	Archaebacteria	0.72726	0.73384	0.33780

sequences, from top to bottom, in the increasing order of the value of α'_{whole}] and Table III (where we denote by β_{whole}, β_{cod}, and β_{noncod} the scales of spectral analysis of the whole, coding and noncoding length sequences, from top to bottom, in the decreasing order of the value of β_{whole}; we denote by γ_{whole}, γ_{cod}, and γ_{noncod} the scales of γ of the whole, coding and noncoding length sequences).

From the right figure of Fig. 3 it is seen that $S(f)$ does not display clear power-law $1/f$ dependence on the frequencies when $f<1$. Although the meaning of region $f>1$ of the power spectrum is not clear, whether $S(f)$ displays perfect power-law $1/f$ in this region is important. When one considers the electrical characteristics of polysilicon emitter bipolar transistors, for high frequency analog applications the transistor $1/f$ noise is also an important parameter since it can degrade the spectral purity of the circuit [33]. There is also some evidence that $1/f$ noise spectral density in the low and in the high current region have a different physical origion (the reader can refer to Ref. [33] and reference therein). We want to know if there is another region of frequencies in which $S(f)$ displays perfect power-law $1/f$ dependence on the frequencies. We carried out the spectral analysis for $f>1$, and found that $S(f)$ displays almost a perfect power-law $1/f$ dependence on the frequencies in some interval:

$$S(f)\propto\frac{1}{f^{\beta}}. \tag{11}$$

We give the results for coding length sequences of *M. genitalium, A. fulgidus, A. aeolicus* and *E. coli* (their lengths are 303, 1538, 891 and 3034 respectively) in Fig. 4, where we take k values $f_k=3k$ $(k=1,\ldots,1000)$ of the frequency f. From Fig. 4, it is seen that the length of the interval of frequency in which $S(f)$ displays almost a perfect power-law $1/f$ depends on the length of the length sequence. The shorter sequence corresponds to the larger interval.

From Fig. 4, one can see that the power spectrum exhibits some kind of periodicity. But the period seems to depend on the length of the sequence. We also guess that the period also depends on the complexity of the sequence. To support this conjecture, we got a promoter DNA sequence from the gene bank, then replaced A by -2, C by -1, G by 1 and T by 2 (this map is given in [9]); so we obtained a sequence on alphabet $\{-2,-1,1,2\}$. Then a subsequence was obtained with the length the same as the coding length sequences of *A. aeolicus, A. fulgidus* and *M. genitalium* (their lengths are 891, 1538 and 303 respectively). A comparison is given in Fig. 5, but the results are not clear-cut.

IV. DISCUSSION AND CONCLUSIONS

Although the existence of the archaebacterial urkingdom has been accepted by many biologists, the classification of bacteria is still a matter of controversy [34]. The evolutionary relationship of the three primary kingdoms (i.e., archeabacteria, eubacteria and eukaryote) is another crucial problem that remains unresolved [34].

From Table I, we can roughly divide bacteria into two classes, one class with α_{whole} less than 0.5, and the other with α_{whole} greater than 0.5. All Archaebacteria belong to the same class except *Pyrococcus abyssi*. All Proteobacteria

TABLE III. β_{whole}, β_{cod}, and β_{noncod}; γ_{whole}, γ_{cod}, and γ_{noncod} of 21 bacteria.

Bacteria	β_{whole}	β_{cod}	β_{noncod}	γ_{whole}	γ_{cod}	γ_{noncod}
M. genitalium	0.05880	0.02030	-0.00708	1.00017	0.99698	1.01652
H. pylori 26695	0.05026	-0.01412	0.01196	0.99902	1.00057	0.99538
M. jannaschii	0.04850	-0.02640	-0.12547	0.99727	0.99079	0.99767
C. pneumoniae	0.04405	0.01071	-0.01906	0.99998	1.00099	0.99348
A. aeolicus	0.03152	0.00811	-0.00115	1.00441	0.99816	0.99870
H. pylori J99	0.01968	0.04512	-0.05815	0.99867	0.99926	0.99349
T. maritima	0.00737	-0.02656	0.01965	0.99726	0.99524	0.98866
C. trachomatis	0.00256	-0.05829	-0.02549	0.99767	1.00211	0.98553
R. sp. NGR234	0.00230	0.04048	-0.10905	1.00570	0.99612	1.01515
M. thermoauto.	-0.00217	-0.11916	0.02079	1.00479	1.00171	1.00063
T. pallidum	-0.00422	-0.02902	0.09510	1.01009	1.01532	1.00222
M. pneumoniae	-0.01137	0.03437	-0.05573	0.98820	0.98783	0.97260
P. abyssi	-0.01589	-0.04242	0.00071	0.99888	0.99816	0.99293
E. coli	-0.01917	-0.05513	0.01772	0.99856	1.00197	0.98938
M. tuberculosis	-0.02653	-0.05653	-0.02698	1.00062	0.99974	1.00801
A. pernix	-0.03882	0.01648	-0.09395	1.00298	1.00407	1.00286
B. burgdorferi	-0.04420	-0.05189	-0.10710	0.99287	0.99792	1.03206
R. prowazekii	-0.04884	-0.12438	-0.07581	1.00284	0.99043	0.99991
H. influenzae	-0.05338	-0.04853	-0.04341	0.99798	1.00248	0.98684
A. fulgidus	-0.06372	-0.08130	-0.00881	1.00347	1.00610	0.98219
B. subtilis	-0.06887	-0.17231	-0.02380	0.99629	1.00853	0.98666

belong to the same class except *E. coli*; in particular, the closest Proteobacteria Helicobacter pylori 26695 and Helicobacter pylori J99 group with each other. In the class with $\alpha_{whole}<0.5$, we have $\alpha_{cod}<\alpha_{noncod}$ except *H. pylori J99* and *M. genitalium*; but in the other class we have $\alpha_{cod}>\alpha_{noncod}$.

Using the exponent α', we can also divide bacteria into two class as in Table II. In one class, $\alpha'_{cod}<\alpha'_{noncod}$. In another class, we have $\alpha'_{cod}>\alpha'_{noncod}$ except *Treponema pallidum* and *Borrelia burgdorferi*. Two Hyperthermophilic bacteria *Aquifex aeolicus* and *Thermotoga maritima* group with each other.

From Tables I and II, we can see the similar rules as above if we use the exponents α_{cod} and α'_{cod}. This follows the fact that the coding sequences occupy the main part of space of the DNA chain of bacteria. This coincides with the conclusion of Ref. [25].

Although, from Table III, we can see the values of all β are not far from 0. From Figs. 1, 2, and 3, one can see exponents α and α' are more convincing than the exponent β because the error of estimating α and α' using the least-squares linear fit is much less than that of the exponent β (the values of r in figures of obtaining exponents α, α', and β are 0.987 685, 0.994 993 9, and 3.791 8E-03 respectively).

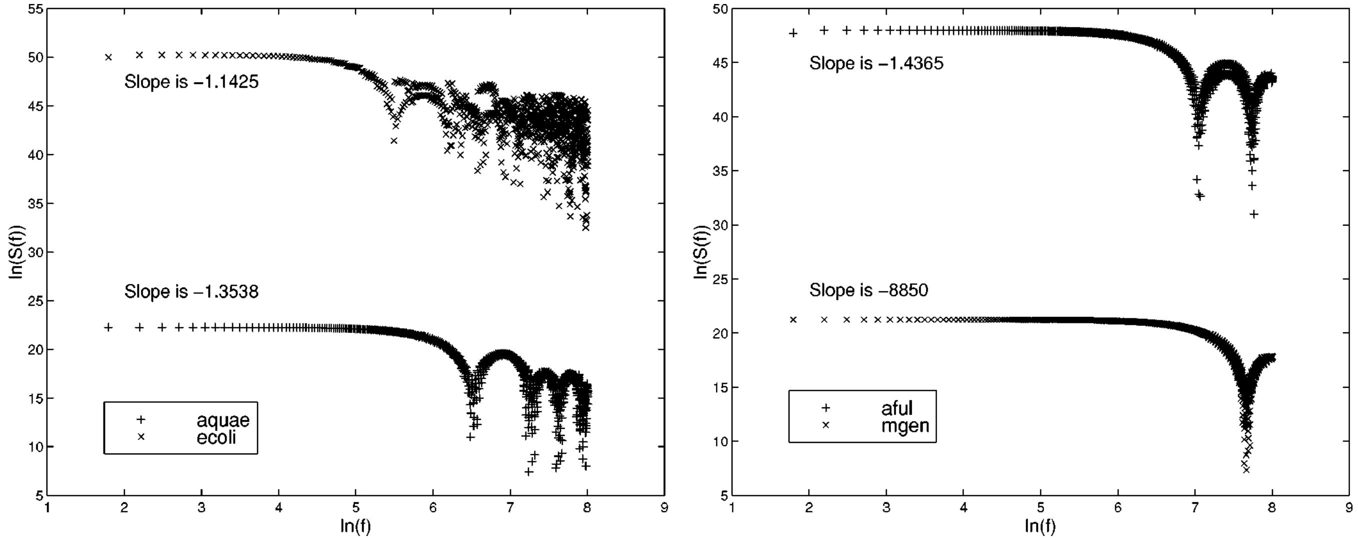

FIG. 4. There exists $1/f$ noise in the interval of $f>1$.

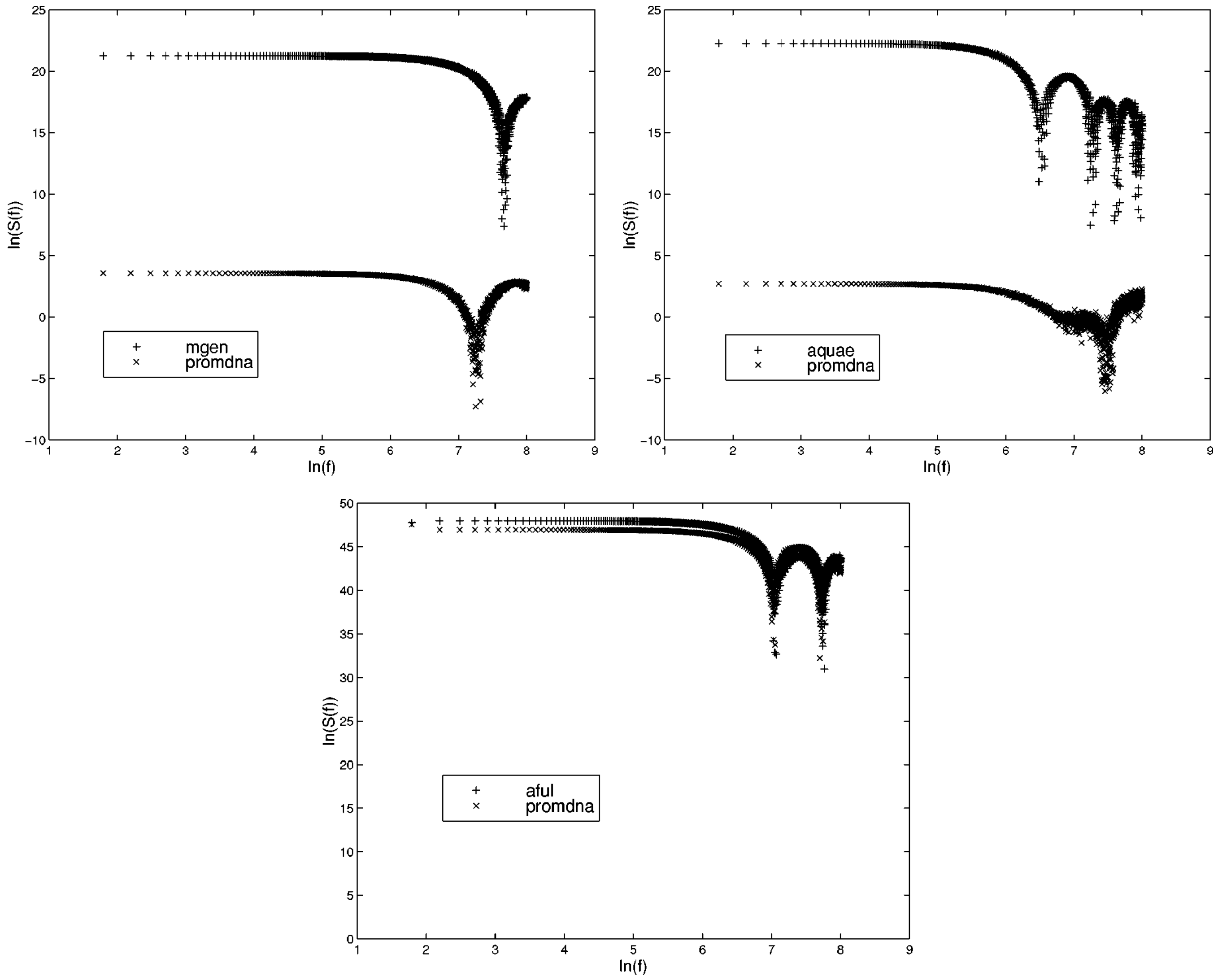

FIG. 5. Compare the power spectral of length sequences and DNA sequences when $f>1$.

From Tables I and II, we can see most values of α and α' are not equal to 0.5, hence we can conclude that most of these length sequences exhibit long-range correlations. We can also see the correlations have a rich variety of behaviors including the presence of anti-correlations. Hence the length sequences have the same character as the DNA sequences [22]. Furthermore, from Table III, we get $\gamma=1.0\pm0.03$. Hence we can conclude that the long-range correlations that exist in most length sequences are linear.

We find in an interval of frequency $(f>1)$, $S(f)$ displays perfect power-law $1/f$ dependence on the frequencies (see Fig. 4):

$$S(f)\propto\frac{1}{f^{\beta}}.$$

The length of the interval of frequency in which $S(f)$ displays almost a perfect power-law $1/f$ depends on the length of the length sequence. The shorter sequence corresponds to the larger interval. The shape of the graph of power spectrum in $f>1$ also exhibits some kind of periodicity. The period seems to depend on the length and the complexity of the length sequence.

ACKNOWLEDGMENTS

Z.G.Y. and B.W. would like to express their thanks to Professor Bai-lin Hao of Institute of Theoretical Physics of Chinese Academy of Science for introducing them into this field and continuous encouragement. They also want to thank Dr. Guo-Yi Chen of ITP for useful discussions. Research was partially supported by Postdoctoral Research Support Grant No. 9900658 of QUT.

[1] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992); W. Li, T. Marr, and K. Kaneko, Physica D **75**, 392 (1994).

[2] C. K. Peng, S. Buldyrev, A. L. Goldberg, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[3] J. Maddox, Nature (London) **358**, 103 (1992).

[4] S. Nee, Nature (London) **357**, 450 (1992).

[5] C. A. Chatzidimitriou-Dreismann and D. Larhammar, Nature (London) **361**, 212 (1993).

[6] V. V. Prabhu and J. M. Claverie, Nature (London) **359**, 782 (1992).

[7] S. Karlin and V. Brendel, Science **259**, 677 (1993).

[8] L. Luo, W. Lee, L. Jia, F. Ji, and L. Tsai, Phys. Rev. E **58**, 861 (1998).

[9] Z.-G. Yu and G.-Y. Chen, Commun. Theor. Phys. **33**, 673 (2000).

[10] (a) R. Voss, Phys. Rev. Lett. **68**, 3805 (1992); (b) Fractals **2**, 1 (1994).

[11] H. E. Stanley, S. V. Buldyrev, A. L. Goldberg, Z. D. Goldberg, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C. K. Peng, and M. Simons, Physica A **205**, 214 (1994).

[12] H. Herzel, W. Ebeling, and A. O. Schmitt, Phys. Rev. E **50**, 5061 (1994).

[13] P. Allegrini, M. Barbi, P. Grigolini, and B. J. West, Phys. Rev. E **52**, 5281 (1995).

[14] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **51**, 5084 (1995).

[15] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[16] A. K. Mohanty and A. V. S. S. Narayana Rao, Phys. Rev. Lett. **84**, 1832 (2000).

[17] C. M. Fraser *et al.*, Science **270**, 397 (1995).

[18] Z.-G. Yu, B.-L. Hao, H.-M. Xie, and G.-Y. Chen, Chaos Solitons Fractals **11**, 2215 (2000).

[19] B.-L. Hao, H.-C. Lee, and S.-Y. Zhang, Chaos Solitons Fractals **11**, 825 (2000).

[20] B.-L. Hao, H.-M. Xie, Z.-G. Yu, and G.-Y. Chen, ‘‘Avoided strings in bacterial complete genomes and a related combinatorial problem,’’ Ann. Combinatorics (to be published).

[21] Z.-G. Yu and B. Wang, Chaos Solitons Fractals **12**, 519 (2001).

[22] M. de Sousa Vieira, Phys. Rev. E **60**, 5932 (1999).

[23] B. Lewin, *Genes VI* (Oxford University Press, Oxford, 1997).

[24] A. Provata and Y. Almirantis, Fractals **8**, 15 (2000).

[25] Z.-G. Yu and V. Anh, ‘‘Time series model based on global structure of complete genome,’’ Chaos Solitons Fractals (to be published).

[26] A. L. Goldberger, C. K. Peng, J. Hausdorff, J. Mietus, S. Havlin, and H. E. Stanley, in *Fractal Geometry in Biological Systems*, edited by P. M. Iannaccone and M. Khokha (CRC, Boca Raton, FL, 1996), pp. 249–266.

[27] R. M. Dunki and B. Ambuhl, Physica A **230**, 544 (1996).

[28] R. M. Dunki, E. Keller, P. F. Meier, B. Ambuhl, Physica A **276**, 596 (2000).

[29] C. K. Peng, J. E. Mietus, J. M. Hausdorff, S. Havlin, H. E. Stanley, and A. L. Goldberger, Phys. Rev. Lett. **70**, 1343 (1993).

[30] R. H. Shumway, *Applied Statistical Time Series Analysis* (Prentice Hall, Englewood Cliffs, NJ, 1988).

[31] F. N. H. Robinson, *Noise and Fluctuations* (Clarendon, Oxford, 1974).

[32] R. D. Remington and M. A. Schork, *Statistics with Applications to the Biological and Health Sciences*, 2nd ed. (Prentice Hall, Englewood Cliffs, NJ, 1985).

[33] E. Simoen, S. Decoutere, and A. Cuthbertson, IEEE Trans. Electron Devices **43**, 2261 (1996).

[34] N. Iwabe *et al.*, Proc. Natl. Acad. Sci. U.S.A. **86**, 9355 (1989).