Some Common Statistical Features of Complete DNA Base Sequence

Nobuyuki Nakajima [1]

nakajima@hit.ipc.ibaraki.ac.jp

Masahiro Agu [2]

agu@dmt.ibaraki.ac.jp

Hideo Akabane [3]

akabane@dmt.ibaraki.ac.jp

[1] Graduate School of Science and Engineering, Ibaraki University

[2,3] Department of Media and Telecommunications Engineering, Ibaraki University

4-12-1 Nakanarusawa-cho, Hitachi, Ibaraki 316, Japan

Abstract

It has been reported that 1/f-like spectra exist in the DNA base sequences. But the meaning of their existence has not been clarified enough still nowadays. Fortunately, a few DNA base sequences have been determined completely in some organism for these three years(Saccharomyces cerevisiae,Haemophilus influenzae, Mycoplasma genitalium,Cyanobacterium Synechocystis etc). In this paper, we show with the use of these entire genomes that most of the complete genomes indicate the common statistical features such as the existence of 1/f-like fluctuation and direction inversion asymmetry of one dimensional G-C base concentration series.

1 Introduction

Several noises which exist in the nature are classified into Lorentzian and 1/f noise. The generation mechanism is understood in the case of Lorentzian noise by the theory of Brown motion, but 1/f noise is not well understood though it exists universally in natural phenomena. For these five years, it has been found that 1/f fluctuations exist in DNA base density series. For what reason 1/f fluctuations exist in DNA base sequence? So, using the fact that some of DNA base sequences of particular microorganism are determined completely, we try to extract their common statistical features of those genomes.

2 Data & Methods

We used the five organisms, S.cerevisiae strain S288C, from ftp.ebi.ac.uk (directory /pub/databases/yeast). H.influenzae strain KW20 and M.genitalium, from TIGR Microbial Database(MDB). Synechocystis sp. Strain PCC6803, from The KAZUSA DNA Research Institute (KDRI). E.coli strain K-12, from E.coli Genome Center(University of Wisconsin). We computed the power specta of base density fluctuation and base molecular weight fluctuation The base density of a unit series with the interval w was defined by[1],

$$Dens(i) = \sum_{j=iw}^{(i+1)w-1} x(j) \tag{1}$$

and its power spectrum was calculated by the equation,

$$S(f) = \frac{1}{N^2} \left| \sum_{i=0}^{N-1} Dens(i)e^{-i2\pi f j} \right|^2 \tag{2}$$

where f is frequency and $f = k/N\{k = 1, 2, 3, , , \}$. Further more, in order to see direction inversion asymmetry of one dimensional G-C base concentration series, we used a fourth order correlation function:

$$\phi(j) = \frac{\left| \overline{\{f(i) - \overline{f(i)}\} \times \{f(i+j) - \overline{f(i+j)}\}^3} - \overline{\{f(i) - \overline{f(i)}\}^3 \times \{f(i+j) - \overline{f(i+j)}\}} \right|}{\overline{f(i) - \overline{f(i)}}^4} \qquad (3)$$

where $f(i,j)$ is $Dens(i,j)$ and j is distance between the unit series.

3 Result & Discussion

The typical results are summarized in the figures bellow.
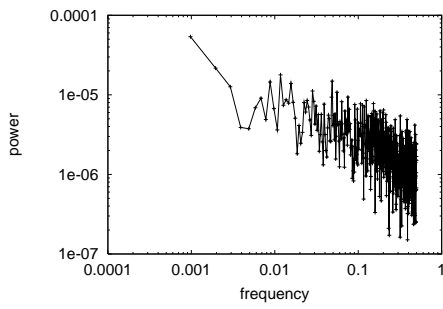

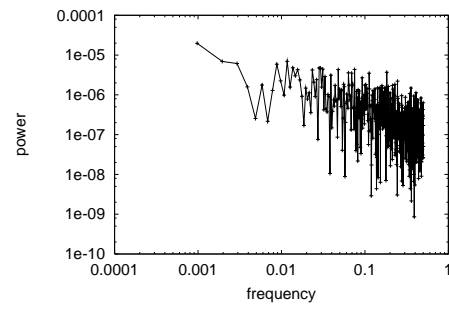
Figure 1: Power spectrum of base density

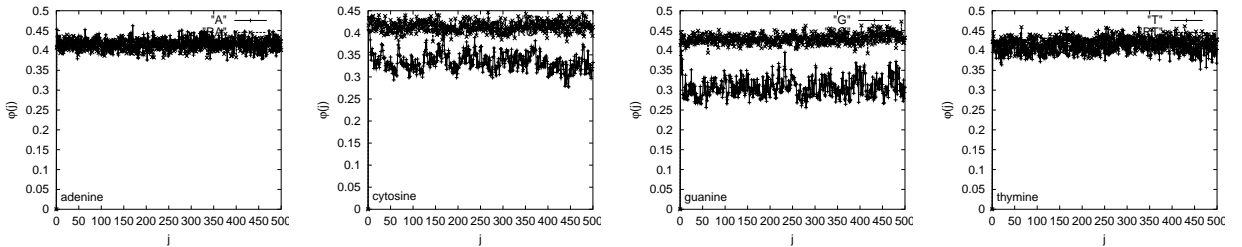Figure 2: Power spectrum of molecular weight

Figure 3: Direction reversal asymmetry of each base

The Fig.1 shows respectively the f^{-1} power spectrum of the base density fluctuation of H.influenzae strain KW20 and Fig.2 that of its molecular weight series fluctuation. The Fig.3 shows the direction inversion asymmetry of the G-C base concentration series of S.cerevisiae strain S288C, chromosome I, compared with the value ($\simeq 0.4$) of $\phi(j)$ of the artificial random base sequence. No inversion asymmetry was found in the A-T base concentration series, as was also shown in Fig.3.

References

[1] Wentian Li, Thomas G Marr and Kunihiko Kaneko, "Understanding Long-range correlation in DNA sequences" *Physica D*, pp392-416, 1994.

[2] Richard F. Voss, "Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequence," *Phys. Rev. Lett*, 68:3805-3808, 1992.