Zipf-Scaling Description in the DNA Sequences

Yasuo Yonezawa yonezawa@life01.dse.ibaraki.ac.jp
Hiroyuki Motohasi motohasi@class.dse.ibaraki.ac.jp
System Engineering Division, Graduate School of Science and Engineering,
Ibaraki University, 4-12-1 Naka-Narusawa, Hitachi 316-8511, Japan

1 Introduction

The background of this research is the point of view that the correlations and relationships of bases of nucleotide sequences are recorded of evolutionary history. In fact, the recent research for statistical analysis (called Linguistic test: Shannon's redundancy) of DNA sequences represent the 1/f fluctuation as power spectra results of Fourier transform at long DNA sequences [1]. Also, Zipf's test, one of linguistic tests, suggests the existence of simple rule into non-coding region called Junk [2]. The statistical analysis (related to Zipf's distribution and Homology of high utilized words) of the occurrence of particular nucleotide runs in DNA sequences of different species has been carried out. In this paper, we report that the adaptive DNA against Zipf's law is indicated on the dependency at homology rate ranking in utilized words in DNA sequences. In order to elucidate the more detail of statistical properties in non-coding DNA sequences, we examine the Zipf's distribution and homology analysis of high utilized word (DNA punctuated sequence length) for the DNA sequences of different species.

2 Materials and Methods

2.1 Materials

DNA data are collected through DDBJ (DNA DataBank of Japan) of Center for Information Biology at the National Institute of Genetics via Internet by Karashi program. The species of the analyzed DNA data are various as follows; *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *C. elegans*, *Penicillium chrysogenum* and *Emericella nidulans*. In addition, we classified coding and non-coding sequences in according to rule of specific code called ORF (Open Reading Frame).

2.2 Zipf's test procedure

Zipf's distribution is acquired by Zipf Plot's of words frequency versus these ranking. The evaluation of Zipf adaptability is carried out by angel of appearance frequency of words versus its ranking. The complete adaptability is indicated with the -1 at inclination angle of Zipf plot.

2.3 Word homology analysis

Homology rate of DNA words (words length=2,3,4,5,6,7) are calculate by method of Sankoff [3].

3 Results and Discussion

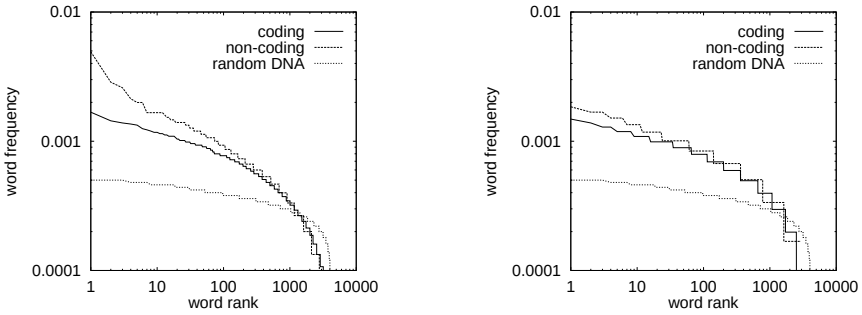
3.1 Zipf adaptability of DNA sequences

The results of the Zipf adaptability are shown in Fig. 1. These results present only typical analyzed output.

3.2 Homology Rate Profile of Higher Ranking DNA Word

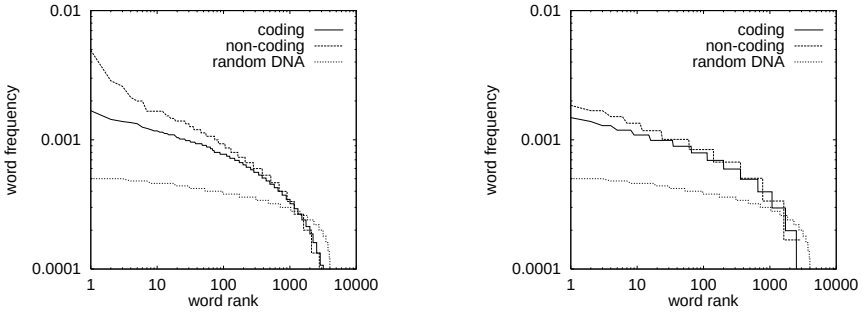
The results of homology rate in DNA word are shown in Fig. 2. Homology rate of Zipf adaptive and not adaptive DNA sequences indicated reversely results.

All correspondences should be to Yasuo Yonezawa.



(a) Zipf adaptive DNA : Saccharomyces cerevisiae (b) Zipf not adaptive DNA : Penicillium chrysogenum

Figure 1: Zipf's law analysis of DNA sequences in word length n=6. These figures are appeared in case of the Zipf adaptive (a), and Zipf not adaptive (b), respectively.



(a) Zipf adaptive DNA : Saccharomyces cerevisiae (b) Zipf not adaptive DNA : Penicillium chrysogenum

Figure 2: Homology of DNA words. These figures present the homology rate of higher ranking words in case of Zipf adaptive DNA and Zipf not adaptive DNA.

3.3 Conclusion

Based on these analysis of homology rate of Zipf adaptive DNA words and Zipf not adaptive words, we can classified two cases. In the first case of Zipf adaptive DNA sequences, the homology rate of higher ranking word (word length=6) indicated that non-coding DNA words were higher than coding DNA words. In the other case of Zipf not adaptive DNA sequences, however, the homology rate of higher ranking DNA words indicated that coding DNA words were higher than non-coding DNA words. Namely, the behavior of homology rate in Zipf adaptive were opposite to that of not adaptive DNA sequences. Furthermore, we are in progress analysis of several DNA sequences.

Acknowledgments

We would like to thank Professor Tutumi and Professor Agu for the support of our research environments. This research was supported by SVBL grant from Ibaraki University.

References

[1] Li, W., Marr, T.G., and Kaneko, K., Understanding long-range correlations in DNA sequences, *Physica D*, 75:392–416, 1994.

[2] Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M., and Stanley, H.E., Linguistic features of noncoding DNA sequences, *Physical Review Letters*, 73:3169–3172, 1994.

[3] Sankoff, D., Matching sequences under deletion/insertion constraints, *Proc. Natl. Acad. Sci. USA*, 69(1):4–6, 1972.