Vestiges of Primordial Words in Base Sequences of Modern Genomes

Nobuyuki Uchikoga
uchiko@genji.c.u-tokyo.ac.jp

Akira Suyama
suyama@dna.c.u-tokyo.ac.jp

Department of Life Sciences, The University of Tokyo
3-8-1 Komaba, Meguro-ku Tokyo 153, Japan

Abstract

Base sequences of genomes were found to consist of a restricted set of homologous short segments when examining base sequences of several genomes classified into different branches of the phylogenetic tree of life. The homologous segments were longer than the triplet codons encoding amino acids, and were observed in both coding and noncoding regions more frequently than the random expectation level. They are thus likely to be the vestiges of the primordial words arisen in the primeval Earth eons ago.

The prebiotic machinery for self-replication undergoes nonenzymatic synthesis of nucleic acids, yielding only short nucleic acid chains compared with enzymatic synthesis of the modern machinery. It is therefore natural to consider how nucleic acid chains long enough to encode functional polypeptide chains arose in the primordial soup. To answer this question S. Ohno has proposed a model of the primordial coding sequences composed of repeats of base oligomers [1]. He demonstrated that the repeats of base oligomers can extend their length in the nonenzymatic self-replication process. Especially when the number of bases in the oligomeric unit is not a multiple of the codon length, the repeats can acquire the coding capacity of polypeptide chains with a long periodicity and a strong resistance to random mutations in evolution.

His model is very attractive. However, the universality of his model still remains open to question because he derived the model from base sequences of some genes with unusually regular oligomeric repeats. We have thus examined the universality of the model by analyzing base sequences of several genomes classified into different branches of the phylogenetic tree.

Figure 1 shows the distribution of the number of homologous segments in the base sequence encoding human phosphoglycerate kinase pseudogene. The sequence is not composed of regular repeats of base oligomers, and thus seems to be almost random. However, the base sequences in both coding and noncoding regions showed statistically significant deviations in the number of homologous segments from random base sequences of the same base compositions. The deviations imply that the base sequences are composed of a restricted set of homologous base segments of less than ten bases long.

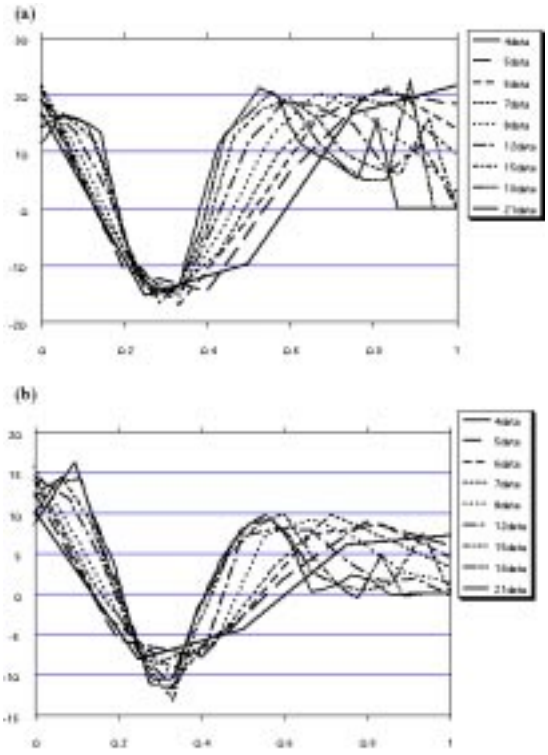

Figure 1: The distributions of the number of homologous segments of 4-21 bases long in the coding region of human phosphoglycerate kinase pseudogene (a) and in the noncoding region surrounding it (b). Let $N(r)$ be the observed number of homologous segments of n bases long with r bases matched and $N_{random}(r)$ be the number expected for random sequences. The deviation from random sequences is given by $d(r) = [N(r) - N_{random}(r)] / N_{random}(r)^{1/2}$.

The large deviations were also observed among both coding and noncoding base sequences of other genomes classified into different branches of the phylogenetic tree. Therefore, it is concluded that the model of base oligomer repeats as the primordial coding sequences is universal and can apply to both coding and noncoding base sequences of the present living organisms. The homologous short base segments may thus spread over entire genomes as the vestiges of the primordial words arisen in the primeval Earth eons ago.

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] S. Ohno "Repeats of base oligomers as the primordial coding sequences of the primeval Earth and their vestiges in modern genes," *J. Mol. Evol.*, 20, 313-321, 1984.