Nucleotide Sequence of a Gene Encodes Characteristic Features of an Organism as Well as an Amino Acid Sequence

Hiroshi Nakashima [1] knishika@genes.nig.ac.jp

Motonori Ota [2] naka@kenroku.kanazawa-u.ac.jp

Ken Nishikawa [2] mota@genes.nig.ac.jp

Tatsuo Ooi [3]

[1] School of health sciences, Faculty of medicine, Kanazawa University
5-11-80 Kodatsuno, Kanazawa 920-0942, Japan.

[2] Center for Information Biology, National Institute of Genetics
1111, Mishima, Shizuoka 411-8540, Japan.

[3] Kyoto Women's University, Kitahiyoshi-cho 35, Higashiyama-ku, Kyoto 605, Japan.

Recently, we have reported that protein coding nucleotide sequences of human, yeast (*Saccharomyces cerevisiae*) and *Escherichia coli* have different feature in the frequency of occurrence of dinucleotides [1]. The genes of human are completely separated in the dinucleotide composition space from those of *E. coli*, and those of yeast sit in-between. This result holds for the genes which encodes homologous proteins. For example, the human gene which encodes H+-transporting ATP synthase a chain, which has 72% amino acid identical to that of *E. coli* is separated from the corresponding *E. coli* gene in the dinucleotide space. This result indicated that a nucleotide sequence of a gene encodes information not only for an amino acid sequence but also for the characteristic features of an organism. As an extension of our previous study, the genes from nine complete genomes; *Escherichia coli, Haemophilus influenzae, Helicobacter pylori, Mycoplasma genitalium, Mycoplasma pneumoniae, Synechocystis* sp., *Methanococcus jannaschii, Archaeoglobus fulgidus, and Saccharomyces cerevisiae* have been investigated [2].

The dinucleotide composition was significantly different between the organisms. The distribution of genes from an organism was clustered around its center in the dinucleotide composition space. The genes from closely related organisms such as Gram-negative bacteria, mycoplasma species and eukaryotes showed some overlaps in the space. The genes from nine complete genomes together with those from human were discriminated into respective clusters with 80% accuracy using the dinucleotide composition alone. The composition data estimated from a whole genome was close to that obtained from genes, indicating that the characteristic feature of dinucleotides holds not only for protein coding regions but also noncoding regions. When a dendrogram was constructed from the disposition of the clusters in the dinucleotide space, it resembled the real phylogenetic tree. Thus, the distinct feature observed in the dinucleotide composition would reflect the phylogenetic relationship of organisms. The cause of the deviation of a whole genome at the dinucleotide level that characterizes individual organisms by 80% confidence is discussed.

References

[1] Nakashima, H., Nishikawa, K., and Ooi, T., Differences in dinucleotide frequencies of human, yeast, and *Escherichia coli* genes, *DNA Res.*, 4:185–192, 1997.

[2] Nakashima, H., Ota, M., Nishikawa, K., and Ooi, T., Gene from nine genomes are separated into their organisms in the dinucleotide composition space, *DNA Res.*, in press.