Dokkyo Journal of Medical Sciences
32(2) : 101～106, 2005

Short Communication

Genomic Structure Consisting of Putative Units Coding Similar Amino Acid Composition : Synchronous Mutations in Biological Evolution

Kenji Sorimachi[1] and Teiji Okayasu[2]

[1] *Department of Microbiology and* [2] *Center of Medical Informatics,*
Dokkyo University School of Medicine, Mibu, Tochigi, 321-0293 Japan

SUMMARY

The complete *Methanobacterium thermoautotrophicum* genome consisting of 1,869 protein genes was divided automatically into 9 units consisting of 186 genes and one unit consisting of 195 genes, or into half size units consisting of 93 genes. The amino acid compositions based on each unit almost coincided with each other and with that based on the complete genome. Similarly, the codon usages based on each unit almost coincided with each other and with that based on the complete genome. In addition, not only the amino acid compositions but also the codon usages almost coincided with each other among the 16 *Saccharomyces cerevisiae* chromosomes. These results indicate that the genomic structure is constructed homogeneously with putative units coding similar amino acid composition. When both genomes were compared, serine, threonine, asparagine and lysine were increased but arginine, alanine, glycine, valine and isoleucine deceased equally among the *Saccharomyces cerevisiae* chromosomes. Thus, as the natural conclusion, mutations may occur synchronously within a genome during biological evolution to form new species.

INTRODUCTION

We have shown that the amino acid composition calculated from the complete genome consisting of different genes indicates a species specific pattern represented by a "star shape", assuming just for calculation that all genes are expressed equally in a cell[1]. Additionally, this basic pattern, also obtained from amino acid analyses of cells, was conserved in all organisms from bacterial to mammalian cells ; and further, that the differences in the patterns seem to reflect biological evolution[2]. These results suggest that the differences in amino acid composition based on the complete genome seem also to reflect biological evolution. Quite recently, we classified eubacteria into two new groups, "S-type" and "E-type", based on differences in amino acid composition calculated from complete genomes[3].

The amino acid compositions of the 16 *Saccharomyces cerevisiae* chromosomes were coincided with each other and with that calculated from the complete genome[4]. The consistent result was obtained from codon usages[5]. Thus, the eukaryotic genome seems constructed with chromosomes having similar amino acid compositions. The fact that codon usage changes between two species occurred homogeneously over the genome strongly suggests that mutations occurred synchronously in all genes during biological evolution. To demonstrate this fact as a general rule, the genome of *Methanobacterium thermoautotrophicum* has been investigated and compared with that of *Saccharomyces cerevisiae* based on not only the amino acid composition but also the codon usage in the present study. Additionally, this fact leads an answer whether the molecular phylogenetic trees obtained from different sin-

Received March 18, 2005 ; accepted May 9, 2005

Reprint requests to : Kenji Sorimachi
Department of Microbiology, Dokkyo University School of Medicine, Mibu, Tochigi 321-0293, Japan

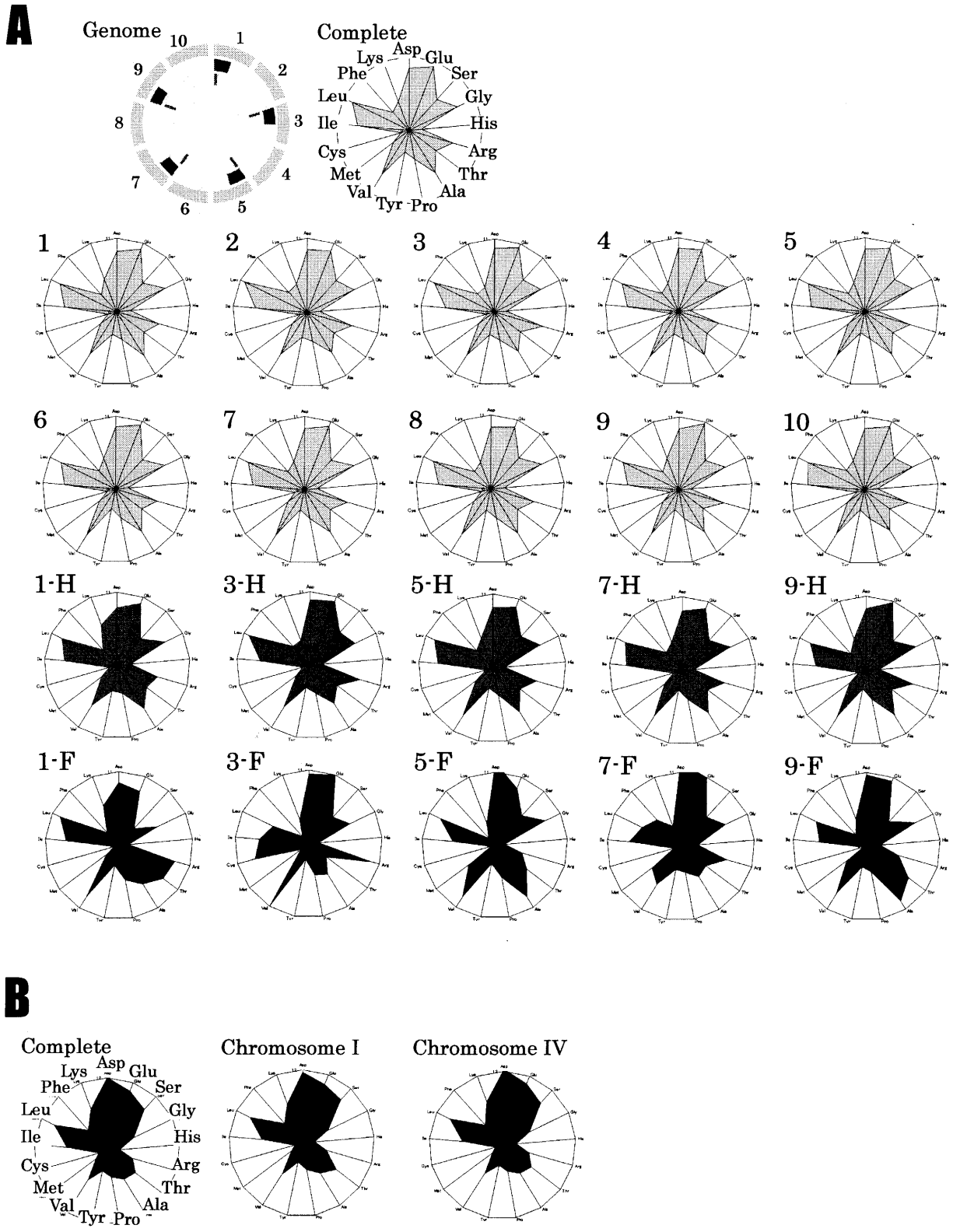

Fig. 1 Radar charts of amino acid compositions calculated from various units of the complete genome of *Methanobacterium thermoautotrophicum* and *Saccharomyces cerevisiae*. **A**, the complete *Methanobacterium thermoautotrophicum* genome consisting of 1,869 protein genes [12] was divided into 10 or 20 units. Ten units (1 − 10) ; based on 186 and 195 genes, half size units (1-H − 9-H) ; based on 93 genes, single genes (1-F − 9-F) ; based on the first single gene of each unit. Glutamine and asparagine were calculated as glutamic acid and asparatic acid, respectively, and tryptophan which is less than 1 % was omitted in the radar charts[2]. **B**, *Saccharomyces cerevisiae*.

gle genes are consistent with each other. Precise molecular phylogenetic trees, for example, cytochrome C[6], small subunit ribosomal protein[7~9] and t-RNA[10,11] were obtained from changes in amino acid or nucleotide sequences, based on a molecular clock.

MATERIALS AND METHODS

Data and calculations. The complete genome of *Methanobacterium thermoautotrophicum* which is constructed with 1,869[12] protein genes, was investigated. The complete genome was divided into several units according to their order as shown on the data sheet of GenomeNet (http://www.genome.ad.jp). To calculate amino acid compositions, all amino acids coded in a certain unit consisting of a number of genes, or in a complete genome, were added automatically, and the numbers of each amino acid were divided by the total number of amino acids. The values are presented by the percent of total amino acids, and the order of amino acids has been previously described[2]. The source of *Saccharomyces crevisiae* have been previously described[4].

RESULTS AND DISCUSSION

Amino acid composition. The complete genome consisting of 1,869 genes was divided into 9 units consisting of 186 genes each and one unit consisting of 195 genes (Fig. 1A, 1－10), or into half size units consisting of 93 genes (Fig. 1A, 1-H－9-H). The amino acid compositions calculated from all genes contained in each unit almost coincided with each other. Thus, it is certain that the genomic structure is homogeneously constructed with putative units with similar amino acid compositions regarding the open reading frame, although each gene has a clearly different amino acid sequence. On the other hand, amino acid compositions based on single genes differed from each other and from that based on the complete genome, but seem to roughly resemble each other (Fig. 1A, 1-H－9-F). Therefore, all proteins may show basically a rough "star shape", which may represent one of characteristics on Earth[2].

The amino acid composition based on the complete *Saccharomyces cerevisiae* genome almost coincided with those of chromosome I (smallest) and chromosome IV (largest)[4].

The present study indicates that gene units consisting of more than 93 genes coding more than 18,000 amino acid residues in *Methanobacterium thermoautotrophicum* show similar amino acid compositions. The minimum putative unit size was mathematically a 3,000 － 7,000 amino acid residue coding size[13] based on a multinomial equation developed by Hochberg and Tamhane[14]. Thus, genomic structure is constructed with putative units with similar amino acid compositions.

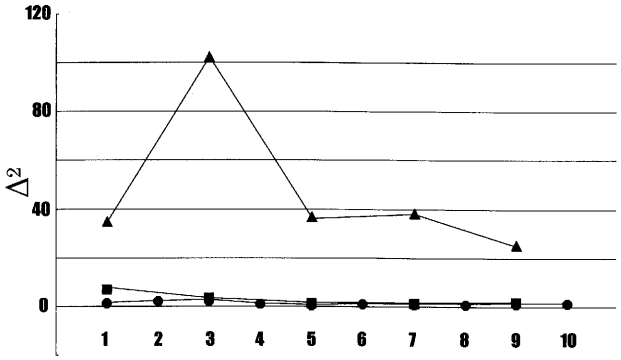

Fig. 2 Quantitative comparison of amino acid compositions between the complete and divided genomes. The vertical axis represents the summation of difference squares based on the subtraction of each amino acid composition of the complete genome from that of the certain unit. The horizontal axis represents the number of unit.

Quantitative comparison. To evaluate a similarity of amino acid compositions, the summation of difference squares based on the subtraction of each amino acid composition of the complete genome from that of the certain unit were calculated. The values calculated from each unit consisting of 186 or 195 genes and the complete genome were almost zero and those based on the half size unit consisting of 93 genes were also almost zero (Fig. 2). On the other hand, the values based on the single genes and the complete genome varied between 25 and 100. The amino acid compositions based on the units consisting of more than 93 genes coincide quantitatively with each other.

Codon usage comparison. The homogeneity of the genomic structure is also examined from 64 codon usages. Particularly, as many amino acids are coded by degenerated codons, large variations might be observed even among gene units of large numbers. Although codon usages based on 10 single genes showed large variations (Fig. 3A), the variations greatly reduced using 10 large units (Fig. 3B). These results indicate that codon usages are equal among 10 divided genomes. Similarly,

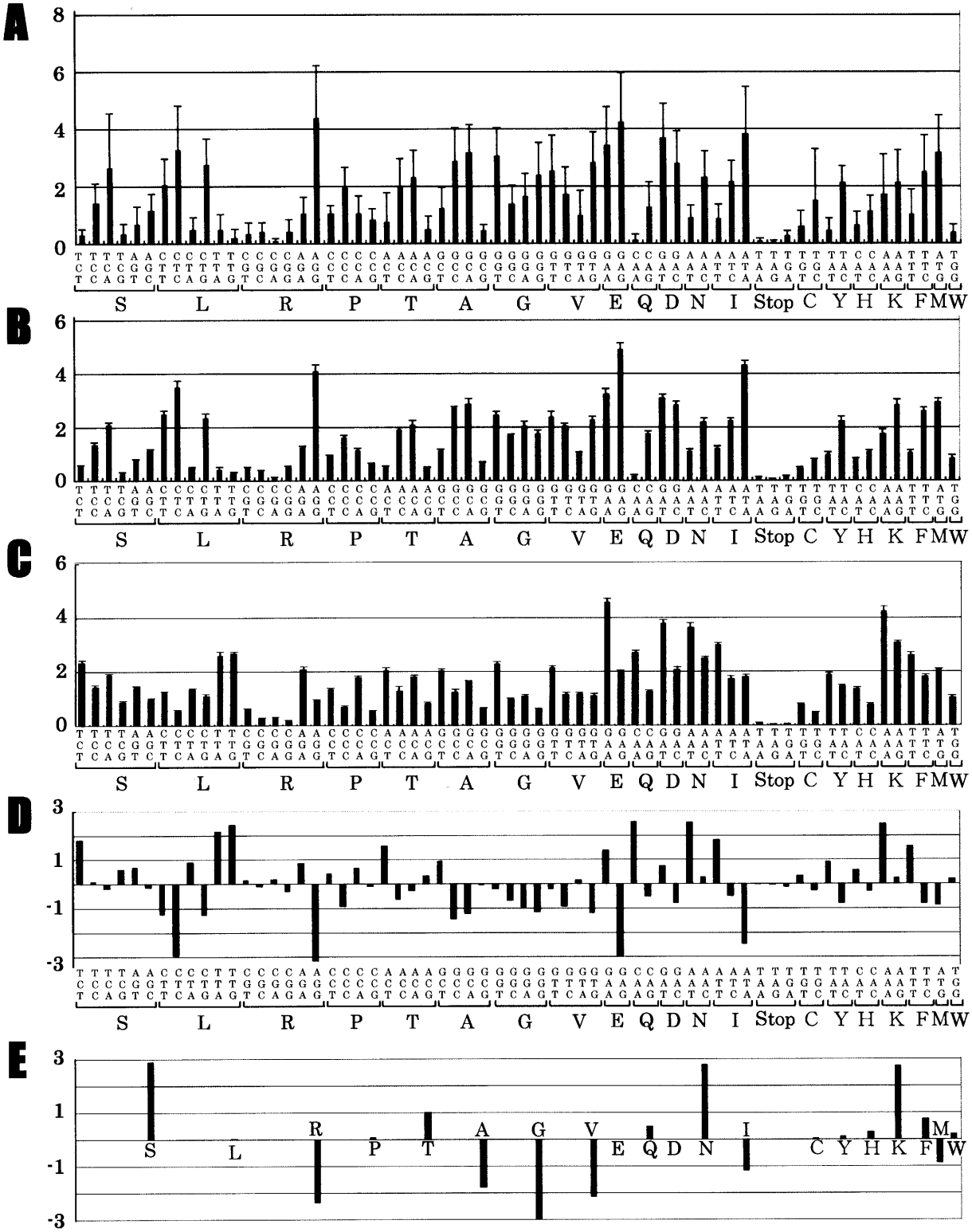

Fig. 3 Codon usages in the various units. **A**, the first *Methanobacterium thermoautotrophicum* genes of 10 units were used. **B**, 10 units consisting of 186 or 195 *Methanobacterium theromoautotrophicum* genes were used and the values ± S.D. **C**, 16 *Saccharomyces cerevisiae* chromosomes were used and the values ± S.D. **D**, the values of Fig. 3**B** were subtracted from those of Fig. 3**C**. **E**, the values of each codon were summed up within the degenerated codons. The values are the percentage of total codons. S, serine ; L, leucine ; R, arginine ; P, proline ; T, threonine ; A, alanine ; G, glycine ; V, valine ; E, glutaimic acid ; Q, glutamine ; D, asparatic acid ; N, asparagine ; I, isoleucine ; C, cystine ; Y, tyrosine ; H, histidine ; K, lysine ; F, phenylalanine ; M, methionine ; W, tryptophan.

the codon usages are equal among the 16 *Saccharomyces cerevisiae* chromosomes (Fig. 3C). Evidently, mutations are strictly controlled by amino acid composition within a unit coding of at least more than 3,000 amino acid residues ; though as yet we have no explanation for this intracellular force. On the other hand, based on single gene analyses, it has been generally thought that mutations occur randomly[15]. We suggested that differences in the cellular amino acid composition seem to reflect biological evolution[2]. Additionally, eubacteria were classified into two groups, "S-type' and "E-type", based on differences in amino acid compositions calculated from the complete genome[3]. Although *Saccharomyces cerevisiae* seems not to be evolved from *Methanobacterium thermoautotrophicum* during biological evolution, the latter codon usages were conveniently subtracted from the former codon usages (Fig. 3D). Comparing these two species, certain codons increased but other codons decreased in the degenerated codons. Therefore, the values of each codon were summed up within the degenerated codons to evaluate the direction of amino acid composition changes between two species (Fig. 3E). In *Saccharomyces cerevisiae*, serine, threonine, asparagine and lysine compositions significantly increased, whereas arginine, alanine, glycine, valine and isoleucine compositions decreased. The increase in serine and threonine compositions might induce an increase in protein functions involving phosphorylation, and that in lysine might be due to appearance of histones which form the nucleosome in eukaryotes. On the other hand, the decrease in alanine, glycine, valine and isoleucine compositions, which might strongly contribute molecular aggregation by a hydrophobic force during life formation, seems to reflect their reduced functions after life formation[2]. The reduction of arginine might be due to compensation for the increased positive charge based on lysine increase, although arginine is also an important component in histones.

CONCLUSION

All existing organisms are present as a result of mutations during biological evolution, indicating that their open reading frames in a genome must differ from those of primitive life forms. If all mutations within certain units consisting of a number of genes occur separately among the units, the amino acid compositions based on these units naturally differ unit by unit throughout the genome. However, in the present study not only the amino acid compositions but also the codon usages based on units consisting of a number of genes coincided with each other among units throughout the genome. In addition, amino acid compositions based on large single genes resemble those based on the complete genome[4]. Thus, even single genes have similar characteristics as the unit consisting of a number of genes, but this is hidden by large variations based on each small gene. This evidence clearly indicates that the same amino acid composition changes due to mutations occurred synchronously, not only among units but also single genes throughout the genome during biological evolution to form new species. Inversely, this means that primitive life forms might also be constructed with gene assembly units coding similar amino acid compositions.

REFERENCES

1) Sorimachi K., Itoh T, Kawarabayasi Y, et al : Conservation of the basic pattern of cellular amino acid composition of archaeo bacteria during biological evolution and the putative amino acid composition of primitive life forms. Amino Acids, **21** : 393-399, 2001.

2) Sorimachi K. : Evolutionary changes reflected by the cellular amino acid composition. Amino Acids, **17** : 207-226, 1999.

3) Sorimachi K, Okayasu T. : Classification of eubacteria based on their complete genome : where does Mycoplasmataceae belong ? Proc. R. Soc. Lond. B (Suppl.), **271** : S127-S130, 2004.

4) Sorimachi K, Okayasu, T. : Gene assembly consisting of small units with similar amino acid composition in the *Saccharomyces cerevisiae* genome. Mycoscience, **44** : 415-417, 2003.

5) Sorimachi K., Okayasu T. : An evaluation of evolutionary theories based on genomic structures in *Saccharomyces cerevisiae* and *Encephalitozoon cuniculi*. Mycoscience, **45** : 345-350, 2004.

6) Dayhoff MO, Park CM, McLaughlin PJ. : Building a phylogenetic tree : cytochrome C, in "Atlas of protein sequence and structure" National Biomedical Foundation, Washington D.C., Vol. 5, pp 7-16, 1972.

7) Doolittle WF, Brown JR. : Tempo, mode, the progenote, and the universal root. Proc. Natl. Acad. Sci. USA, **91** : 6721-6728, 1994.

8) Sogin ML, Elwood HJ, Gunderson JH. : Evolutionary diversity of eukaryotic small subunit rRNA genes, Proc. Natl. Acad. Sci. USA, **83** : 1383-1387, 1986.

9) Woese CR, Kandler O, Wheelis ML. : Towards a natural system of organisms : Proposal for the domains archaea, bacteria, and eucarya. Proc. Natl. Acad. Sci. USA, **87** : 4576-4579, 1990.

10) DePouplana LR, Turner RJ, Steer BA,et al : Genetic code origins : tRNAs older than their synthetases ? Proc. Natl. Acad. Sci. USA, **95** : 11295-11300, 1998.

11) Maizels N, Weiner AM. : Phylogeny from function : Evidence from the molecular fossil record that tRNA originated in replication, not translation. Proc. Natl. Acad. Sci. USA, **91** : 6729-6734, 1994.

12) Smith R.D., Doucette-Stamm LA, Deloughery C, et al : Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH : functional analysis and comparative genomics. J. Bacteriol., **197** : 7135-7155, 1997.

13) Sorimachi K, Okayasu T, Ebara Y,et al : Mathematical proof of genomic amino acid composition homogeneity based on putative small units. Dokkyo J. Med. Sci., **32** : 99-100, 2005.

14) Hochberg Y, Tamhane AC. : Multiple comparison procedures, in "Probability and Mathematical Static's". ed. by Hochberg Y, Tamhane AC , John Wiley & Sons, New York, pp 274-309, 1987.

15) King JL, Jukes TH. : Non-Darwinian evolution. Most evolutionary change in proteins may be due to neutral mutations and genetic drift. Science, **164** : 788-798, 1989.