Dokkyo Journal of Medical Sciences
31(1) : 1～5, 2004

Originals

Similarity of Amino Acid Composition Based on Gene Assembly and Different Gene - size Distributions among the 11 Chromosomes in *Encephalitozoon cuniculi*

Teiji Okayasu[1], Yoshifumi Ebara[3] and Kenji Sorimachi[2]

[1] *Center for Medical Informatics and*
[2] *Department of Microbiology, Dokkyo University School of Medicine, Mibu, Tochigi 321 - 0293, Japan*
[3] *The student of Dokkyo University School of Medicine, who contributed to the present research in his small group study*

SUMMARY

The amino acid composition and the gene-size distribution based on the complete genome of *Encephalitozoon cuniculi* were analyzed. The amino acid compositions based on the 11 chromosomes of *Encephalitozoon cuniculi* were almost identical, whereas the distribution patterns of gene-size among the chromosomes slightly differed. Thus, the genome is constructed with gene assemblies which show similar amino acid compositions, and which do not have the restricted boundaries.

INTRODUCTION

We have shown that the basic pattern of cellular amino acid compositions is conserved from bacteria to human cells, while differences in cellular amino acid compositions seem to reflect biological evolution[1～4]. During these studies, the amino acid compositions based on the complete genome of Archaea[5～8] resembled those obtained from the amino acid analysis of cells, assuming just for calculation that all genes are expressed equally in a cell[4]. This coincidence has puzzled us for some time, because in cells each gene must be expressed differently. Recently, we showed that the amino acid composition based on a gene assembly encoding 3,000 – 7,000 amino acid residues represents the amino acid composition of the complete genome of *Saccharomyces cerevisiae*, and that the amino acid compositions of 16 chromosomes resemble each other[9]. This gene assembly is independent not only of the gene position but also of the gene size. Therefore, the "unit" consisting of a gene assembly does not have a definite restricted boundary. To generalize our previous result, *Encephalitozoon cuniculi*, consisting of 11 chromosomes of which complete genomes analyzed by other group[10], was examined. In addition, gene-size distribution was compared among the chromosomes.

Received July 25, 2003 ; accepted September 30, 2003
Reprint requests to : Kenji Sorimachi
Department of Microbiology, Dokkyo University School of Medicine, Mibu, Tochigi 321 - 0293, Japan

MATERIALS AND METHODS

The complete genome data of *Encephalitozoon cuniculi* was obtained from GenBank (http://www.genome.ad.jp), and data were analyzed using the Excel.

RESULTS AND DISCUSSION

The amino acid compositions based on the 11 chromosomes and the complete genome were almost identical, as shown in Fig. 1. This result is consistent with that obtained in our study on *Saccharomyces cerevisiae*[9]. In that study[9], the amino acid composition of a "unit" consisting of a gene assembly increased its similarity to that of the complete genome with increasing encoding number of the amino acid, and the amino acid composition

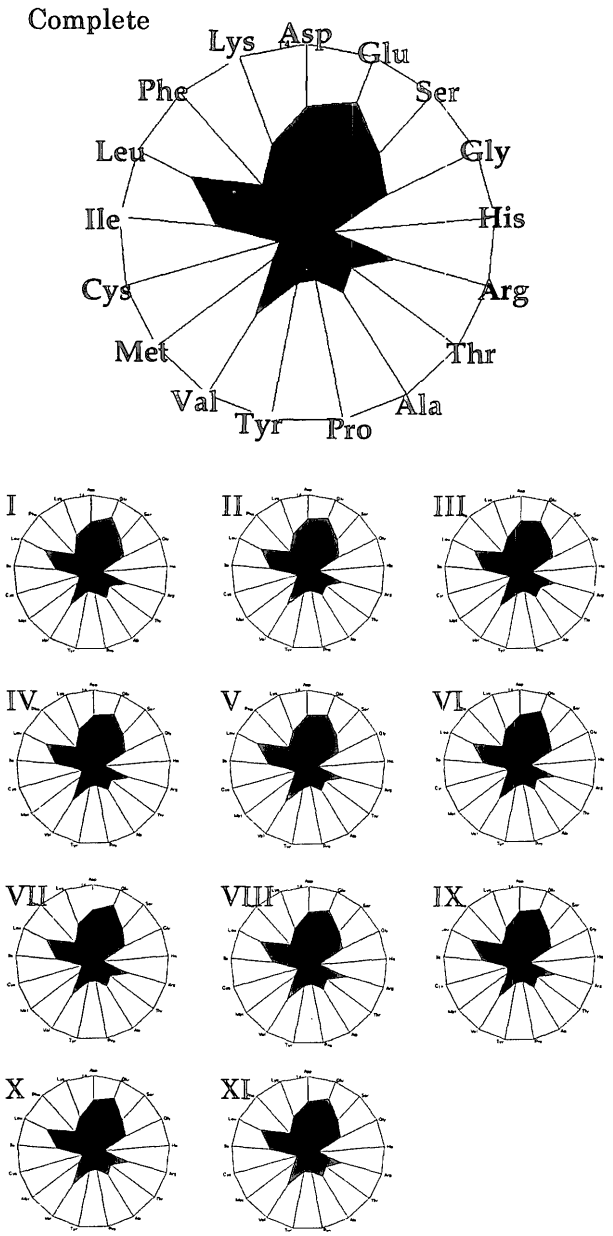

Fig. 1 Radar charts of amino acid compositions determined from genomic data of *Encephalitozoon cuniculi*. The values represent the percent of total amino acids.
Asn and Gln were calculated as Asp and Glu, respectively, and the order of amino acids on the radar chart was based on the elution order of amino acids from the HPLC.[1]

almost reached a constant pattern at 3,000 – 7,000 amino acid residues. Naturally, however, the amino acid sequence of each gene differs. The total numbers of amino acid residues encoded by each chromosome was calculated, as shown in Table 1, being 50,000 – 80,000 in the 11 chromosomes. As these numbers are much larger than the 3,000 – 7,000, the amino acid compositions based on each chromosome or on the complete genome were almost identical with each other.

In our earlier studies, tryptophan, which was decomposed in the amino acid analysis and whose concentrations were less than 1% of the total amino acids, was omitted from the calculation of amino acid compositions[4].

The concentrations of tryptophan contained in each chromosome in this study were almost the same (Table 1). In addition, among the 11 chromosomes the numbers of amino acid residues based on each average gene also coincided with each other. The numbers of amino acid residues encoded by an average size gene were 324 – 414, mainly being around 350.

The distribution of gene coding size was investigated in the 11 chromosomes, as shown in Fig. 2. Their distribution patterns slightly differed among the chromosomes, compared with a similarity in their amino acid compositions. Characteristic differences were observed in the small (less than 100 amino acid residues), or larger (more than 500 amino acid residues), and also in the existence of extremely large size (more than 1,000 amino acid residues) regions, as shown in Table 2.

On the other hand, the distribution of gene coding sizes based on the complete genome showed a smoothly decreasing curve with increasing encoding number (Fig. 2). Similar results were obtained from *Mycoplasma pneumoniae*[11]. Therefore, differences in the distribution of the sizes of the encoding genes are due to the small number of genes among the chromosomes.

The amino acid composition of small size genes encoding less than 100 amino acid residues differed from that based on the complete genome (data not shown), although that based on a gene assembly encoding more than 3,000 amino acid residues resembled the latter. However, even single genes encoding more than 3,000 amino acid residues show very similar amino acid compositions to that based on the complete genome of *Saccharomyces cerevisiae*[9]. Additionally, consistent data were obtained from mouse cDNAs and several bacterial genomes (unpublished data). Thus, the basic pattern of amino acid composition is determined by amino acid coding numbers, and this number is more than 3,000 amino acid residues in all organisms. Similarly, amino acid composition based on several proteins was almost identical with the cellular amino acid composition obtained from

Table 1 Numbers of amino acids and numbers of tryptophan based on each chromosome and the complete genome of *Encephalitozoon cuniculi.*

Chromosome	Total No. of A. A.	Total No. of A. A.-Try	No. of Try	Ratio of Try (%)	Total gene No.	Ave. A. A No.
Complete	716,830	711,175	5,655	0.79	1,996	359
I	53,078	52,537	541	1.02	159	334
II	57,523	57,070	453	0.79	157	366
III	55,763	55,325	438	0.79	158	353
IV	61,075	60,589	486	0.80	172	355
V	61,082	60,633	449	0.74	172	355
VI	63,594	63,069	525	0.83	172	370
VII	67,318	66,834	484	0.72	188	358
VIII	68,373	67,804	569	0.83	211	324
IX	73,579	73,041	538	0.73	206	357
X	78,650	78,042	608	0.77	190	414
XI	76,795	76,231	564	0.73	211	364

Table 2 Comparison of the existence of small or large genes among *Encephalitozoon cuniculi* chromosomes and the complete genome.

Chromosome	X < 100	100 ≦ X < 500	500 ≦ X < 1,000	1,000 ≦ X
Complete	1.85	66.83	28.86	2.45
I	3.14	69.18	25.16	2.52
II	1.27	59.87	38.22	0.64
III	1.90	67.72	27.22	3.16
IV	2.33	72.67	21.51	3.49
V	2.33	68.60	27.33	1.74
VI	1.16	63.37	34.88	0.58
VII	1.60	70.74	23.94	3.72
VIII	1.90	73.46	24.17	0.47
IX	2.91	63.11	31.55	2.43
X	1.05	57.89	36.84	4.21
XI	0.95	67.77	27.49	3.79

X : the number encoded by a gene.
The values represent the percent of total genes in the indicated ranges.

the amino acid analysis of cells[4]. In both genomic amino acid analysis and cellular amino acid analysis, a small number of genes and proteins represented the amino acid compositions based on the complete genome and whole cells, respectively. Therefore, amino acid composition as obtained from the complete genome naturally resembles amino acid composition obtained from cells. This strongly suggests that the genome is constructed with a gene "unit" that encodes a similar amino acid composition throughout its structure. In addition, it indicates that biological evolution occurred coincidentally in every gene.

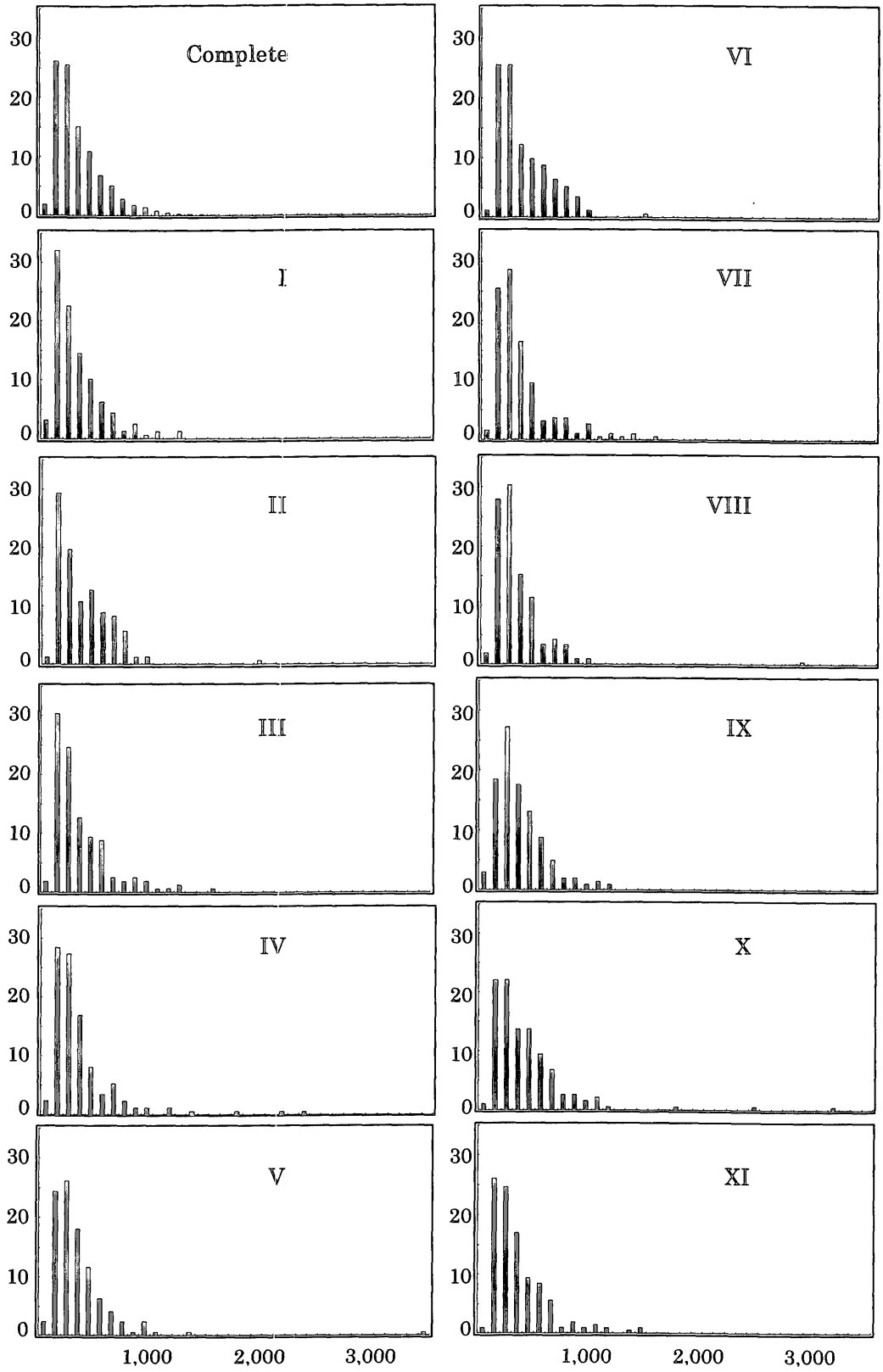

Fig. 2 Distribution patterns of gene sizes in *Encephalitozoon cucniculi* chromosomes and the complete genome. The bars represent the percent of total gene numbers contained in each 100 amino acid residues range.

REFERENCES

1) Okayasu T, Ikeda M, Akimoto K, et al : The amino acid composition of mammalian and bacterial cells. Amino Acids, **13** : 379-391, 1997.
2) Sorimachi K : Evolutionary changes reflected by the cellular amino acid composition. Amino Acids, **17** : 207-226, 1999.
3) Sorimachi K, Okayasu T, Akimoto K, et al : Conservation of the basic pattern of cellular amino acid composition during biological evolution in plants. Amino Acids, **18** : 193-197, 2000.
4) Sorimachi K, Itoh T, Kawarabayasi Y, et al : Conservation of the basic pattern of cellular amino acid composition of archaeobacteria during biological evoultion and the putative amino acid composition of primitive life forms. Amino Acids, **21** : 393-399, 2001.
5) Bult CJ, White O, Olsen GJ, et al : Complete genome sequence of the methanogenic archaon, *Methanococcus jannaschii*. Science, **273** : 1058-1073, 1996.
6) Smith DR, Doucette-Stamm LA, Deloughery C, et al : Complete genome sequence of *Methanobacterium thermoautotrophicum* △H : functional analysis and comparative genomics. J. Bacteriol., **179** : 7135-7155, 1997.
7) Klenk H-P, Clayton RA, Tomb J-F, et al : The complete genomic sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature, **390** : 364-370, 1997.
8) Kawarabayasi Y, Sawada M, Horikawa H, et al : Complete sequence and gene organization of the genome of a hyper-thermophilic archaobacterium, *Pyrococcus horikoshii* OT3. DNA Res., **5** : 55-76, 1998.
9) Sorimachi K and Okayasu T : Gene assembly consisting of small units with similar amino acid composition in the *Saccharomyces cerevisiae* genome. Mycoscience, **44** : 415-417, 2003.
10) Katinka MD, Duprat S, Cornillot E, et al : Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature, **414** : 450-453, 2001.
11) Himmelreich R, Hilbert H, Plagens H, et al : Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acid Res., **24** : 4420-4449, 1996.