Universal Rules Governing Genome Evolution Expressed by Linear Formulas

Kenji Sorimachi[*,1] and Teiji Okayasu[2]

[1]*Educational Support Center and* [2]*Center of Medical Informatics, Dokkyo Medical University, Mibu, Tochigi 321-0293, Japan*

Abstract: Codon evolution is influenced by an organelle-specific bias existing within cellular compartments. The discovery of evidence that codon evolution is controlled by different factors among nuclei, mitochondria and chloroplasts, indicates the existence of biases based on what we term "organelle biases". In plant chloroplasts and mitochondria, nucleotide substitutions, based on the correlations among nucleotides, were governed by the linear formula, $y = ax + b$, where y and x represent nucleotide contents, and a and b are constants. In animal mitochondria, with respect to the correlations between each nucleotide, only the correlations between purines (A versus G) or pyrimidines (C versus T) were almost linear, while the other correlations between purine and pyrimidine (A or G versus T or C) were not. However, nucleotide substitutions, based on the correlations of the same nucleotide between complete single-strand DNA and its coding or non-coding region, were always linear, $y = ax + b$, in any organelle. Linear correlations of the same nucleotide were also obtained between coding and non-coding region DNAs. The present results clearly reveal evidence that genome evolution is governed by universal linear formulas, although different rules are apparently observed in the different organelle fields formed by nuclei, mitochondria and chloroplasts. This report quantitatively demonstrates the existence of factors controlling genome evolution that are expressed by linear formulas.

INTRODUCTION

Traditional molecular biology methods, based on the amino acid sequences of proteins and the nucleotide sequences of nucleic acids, have contributed to our modern understanding of biology, including evolution and phylogeny. These approaches, however, are limited to the analysis of relatively small numbers of the same genes among different species. Recently, however, the development of *in silico*, bioinformatic approaches has enabled analyses of whole genome processes. By graphical presentation of the ratios of amino acids to the total number of amino acids, or those of nucleotides to the total number of nucleotides, we can readily draw conclusions from large data sets. Based on this concept, we revealed first that the genome is constructed homogeneously, with small units displaying similar codon usages [1, 2] and coding for similar amino acid compositions [2, 3]. We also showed that a gene assembly encoding 3,000-7,000 amino acid residues shows a similar amino acid composition to the complete genome. As this unit size is independent not only of genome size, but also of species, either single or multiple units can be compared not only in the same, but also in different genes or genomes [1-3]. According this principle, it is possible to directly compare different genomes consisting of different genes.

Although several basic hypotheses to explain codon formations have been proposed [4-6], no unified theory has been established. Nevertheless, it has been accepted that biased codon usage is a result of natural selection [7, 8] or mutational pressure [9-11]. We have previously shown, by simulation analysis based on random choices of nucleotide or amino acid residues, that codon formation was initially established under certain controls based on pre-determined amino acid sequences were formed in the absence of codons [12].

Chargaff's parity rules are universal for all replicating organisms [13-15], and the persistence of the second parity rule has been confirmed in a large number of genomes across various taxonomic groups [16]. These results indicate that nucleotide substitutions are strictly controlled by certain forces during biological evolution. However, these rules cannot reflect evolutionary differences among kingdoms. Recently, we showed that codon evolution, from bacteria to *Homo sapiens*, is governed by linear formulas [17], and that correlations among nucleotides are also governed by linear formulas in the non-coding region and in the organelle single-strand DNA, as well as in the coding region of nuclear DNA [17]. As nucleotide substitutions in nuclei are expressed by linear formulas, 64 codon usages can be naturally estimated from just one nucleotide content in the genome [17]. Based on these results, it is clear that nucleotide substitutions, which can be expressed by linear formulas, are controlled by certain rules. Therefore, the present study was designed to determine whether any nucleotide substitution is controlled by universal rules, not only in nuclei but also in mitochondria or chloroplasts.

METHODS

All genomic data were obtained from GenomeNet (http://www.genome.ad.jp) and codon usages were obtained from the Kazusa DNA Research Institute (http://www.kazusa.or.jp/codon). All gene and genome DNA sequences were

*Address correspondence to this author at the Educational Support Center, Dokkyo Medical University, Mibu, Tochigi 321-0293, Japan; Tel: +81-828-86-1111; Fax: +81-282-86-2157; E-mail: kenjis@dokkyomed.ac.jp

obtained from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih. gov/sites). Amino acid compositions and nucleotide contents were calculated on a personal computer, as described previously [17-19]. Nucleotide contents in the coding sequence, excluding RNA genes, were calculated from all genes contained in the complete mitochondrial and chloroplast genomes, and those in the non-coding sequence, including RNA genes, were obtained from the subtraction of nucleotides in the coding sequence from the complete mitochondrial and chloroplast single strand DNAs throughout this study. Genes located in both forward and reverse (without conversion) strands were used. Based on these procedures, a complete genome is assumed consisting of two huge molecules which represent a coding and a non-coding sequence. Figures calculated from nucleotide sequences exclude intra-strand deviations, and all organisms or organelles are compared with each other based on inter-species deviations. All calculations were performed using Microsoft Excel 2003. The amino acid order on the radar chart is based on the elution order of each amino acid from HPLC [20]. Statistical analysis was done with Excel 2003.

RESULTS AND DISCUSSION

Codon Usages and Amino Acid Compositions

In the present study, 97 chloroplasts, 49 plant mitochondria and 105 animal mitochondria were examined (Additional data file 1). Plasmids coding for more than 3,000 amino acid residues showed similar amino acid compositions to their parental nuclei in bacteria and archaea [18], as shown in Fig. (**1**). This evidence clearly indicates that nucleotide substitutions in nuclei occurred synchronously between parental nuclei and their plasmids during biological evolution. This also suggests that plasmids are derived from their parental nuclei. On the other hand, plant cells have characteristic DNAs in nuclei, chloroplasts and mitochondria. In *Oryza sativa* (rice), codon usage patterns and amino acid compositions differ among the DNAs of the three organelles, as shown in Fig. (**2**). GC alternations in codon usages occurred synchronously among these three organelles. However, these differences do not mean that the codon evolution of the three genomes is governed by three different rules, because both chloroplasts [21] and mitochondria [22] are likely to be de-
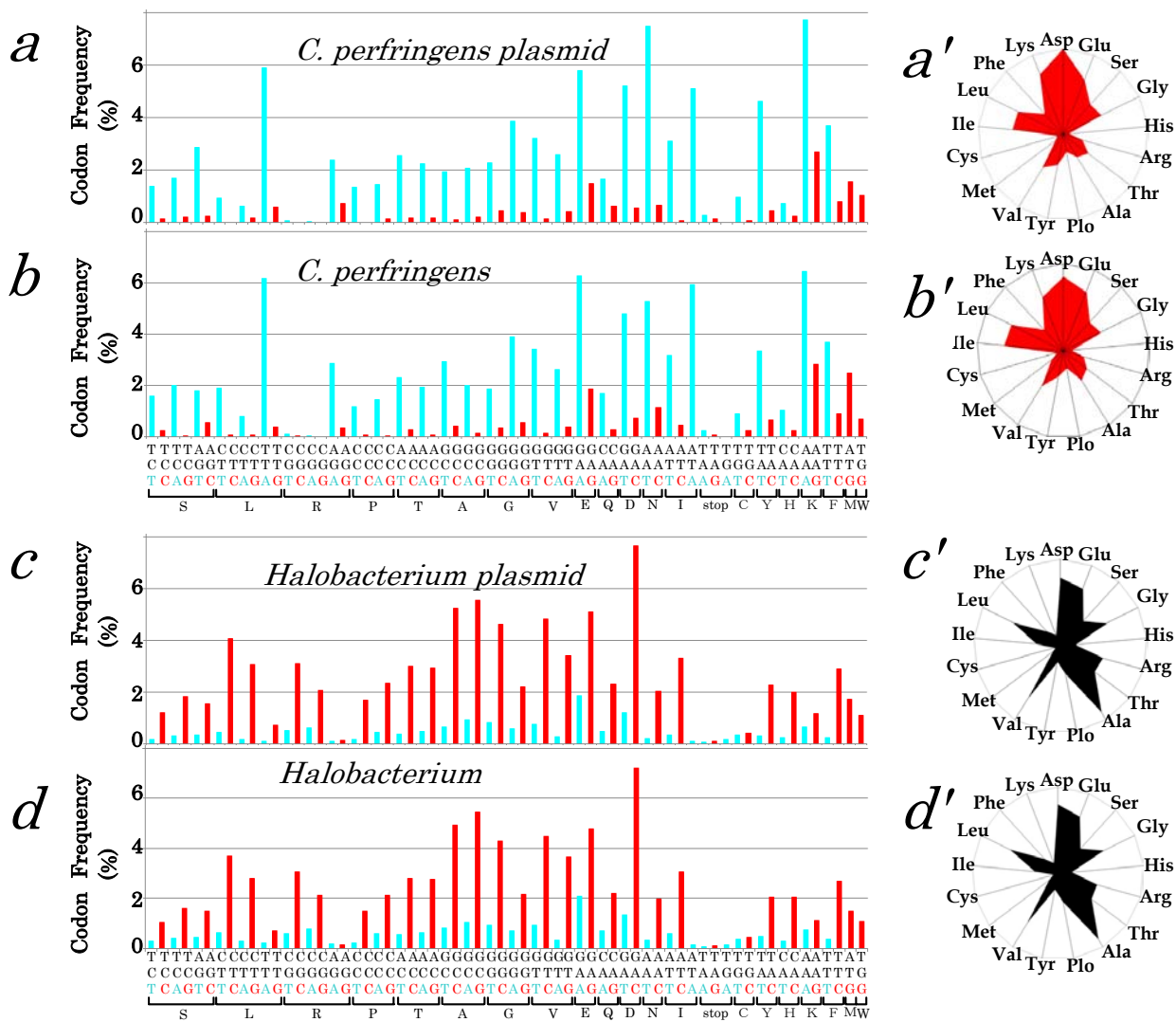

Fig. (1). Codon usage patterns and amino acid compositions. Codon usage (bar), a, b, c and d, and amino acid composition (radar chart), a', b', c' and d', are expressed as percentages of all codons and amino acids, respectively. *Clostridium perfringens* and *Halobacterium* are bacteria and archaea, respectively.
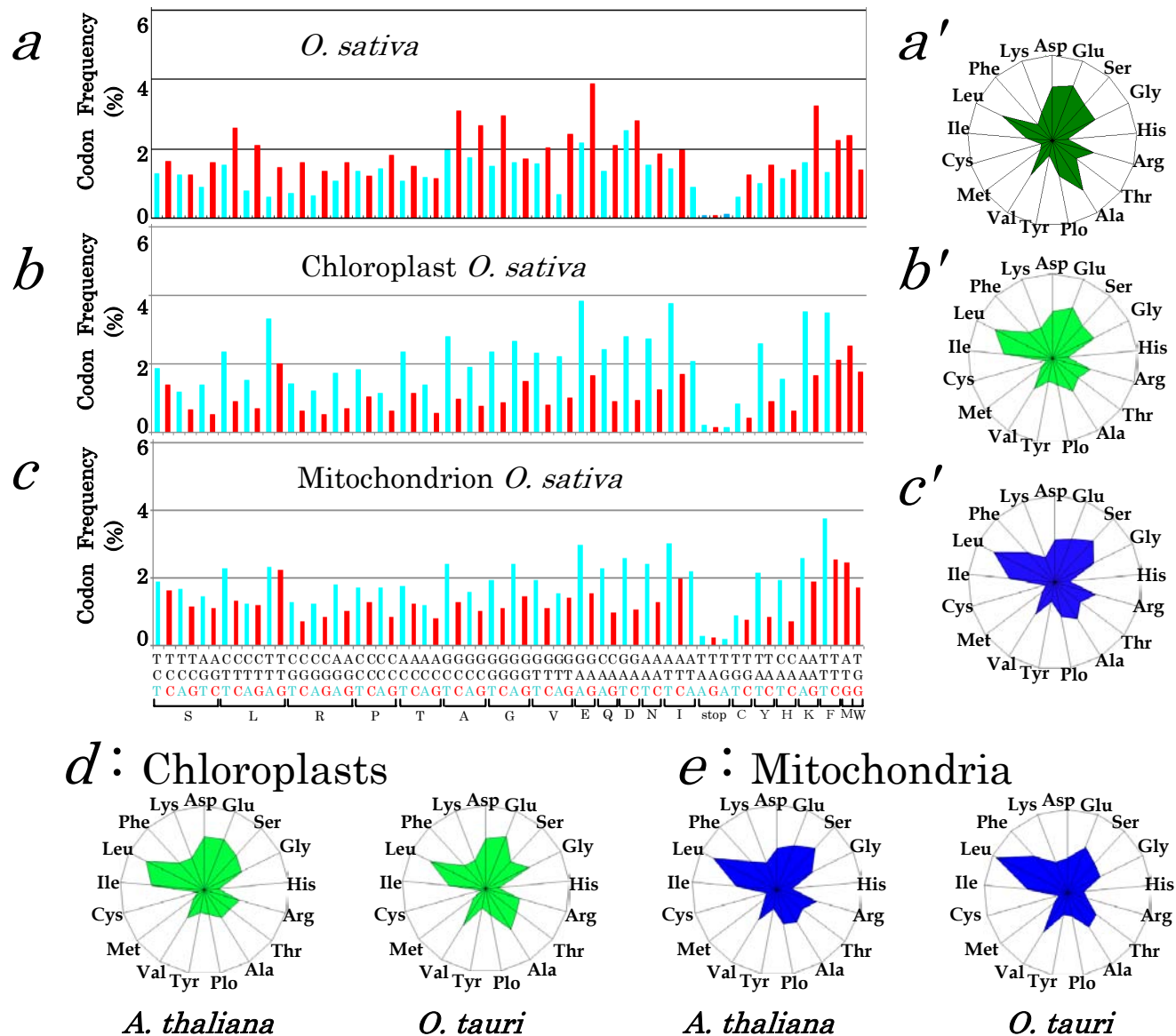
Fig. (2). Codon usage patterns and amino acid compositions. Codon usage (bar chart), a, b and c, and amino acid composition (radar chart), a', b', c', d and e, are expressed as percentages of all codons and amino acids, respectively.

rived from symbiotic relationships during evolution. The amino acid compositions in chloroplasts and mitochondria from two species, *Arabidopsis thaliana* and green algae (*Ostrecoccus tauri*), are shown in Fig. (**2d**) and (**2e**). Their patterns resemble each other, although high concentrations of Leu and Ile were observed in mitochondria (Fig. **2e**). In both chloroplasts and mitochondria, Ile concentrations reduced in parallel with increases in G or C content at the third codon position (Fig. **2**), as observed in the complete nuclear genomes of bacteria, archaea and eukaryotes [17].

Correlations Between Nucleotide Contents in Plant Organelles

Correlations between two nucleotide species in the coding region were examined in chloroplasts and mitochondria in various plant species. The correlation between G and C was expressed by almost the same linear regression line be-

tween chloroplasts and mitochondria, as shown in Fig (**3**). However, these two lines based on chloroplasts and mitochondria significantly differed from that based on nuclei. Similarly, the correlation between G and either T or A, and that between C and either T or A, were also expressed by linear regression lines with almost the same slopes, but with different constant values on the y axis. For inverse correlations, consistent results were obtained (Additional data file 2). The equations and correlation coefficients obtained computationally are summarized in Table **1**. Almost the same slopes were obtained for the linear regression lines, but there were different constant values on the y axes. These results indicate that nucleotide substitution rates were almost the same among nuclei, mitochondria and chloroplasts, and that the original nucleotide contents differed among them. Thus, nucleotide alternations in the coding region differ between chloroplasts and mitochondria based on the slope and constant values.
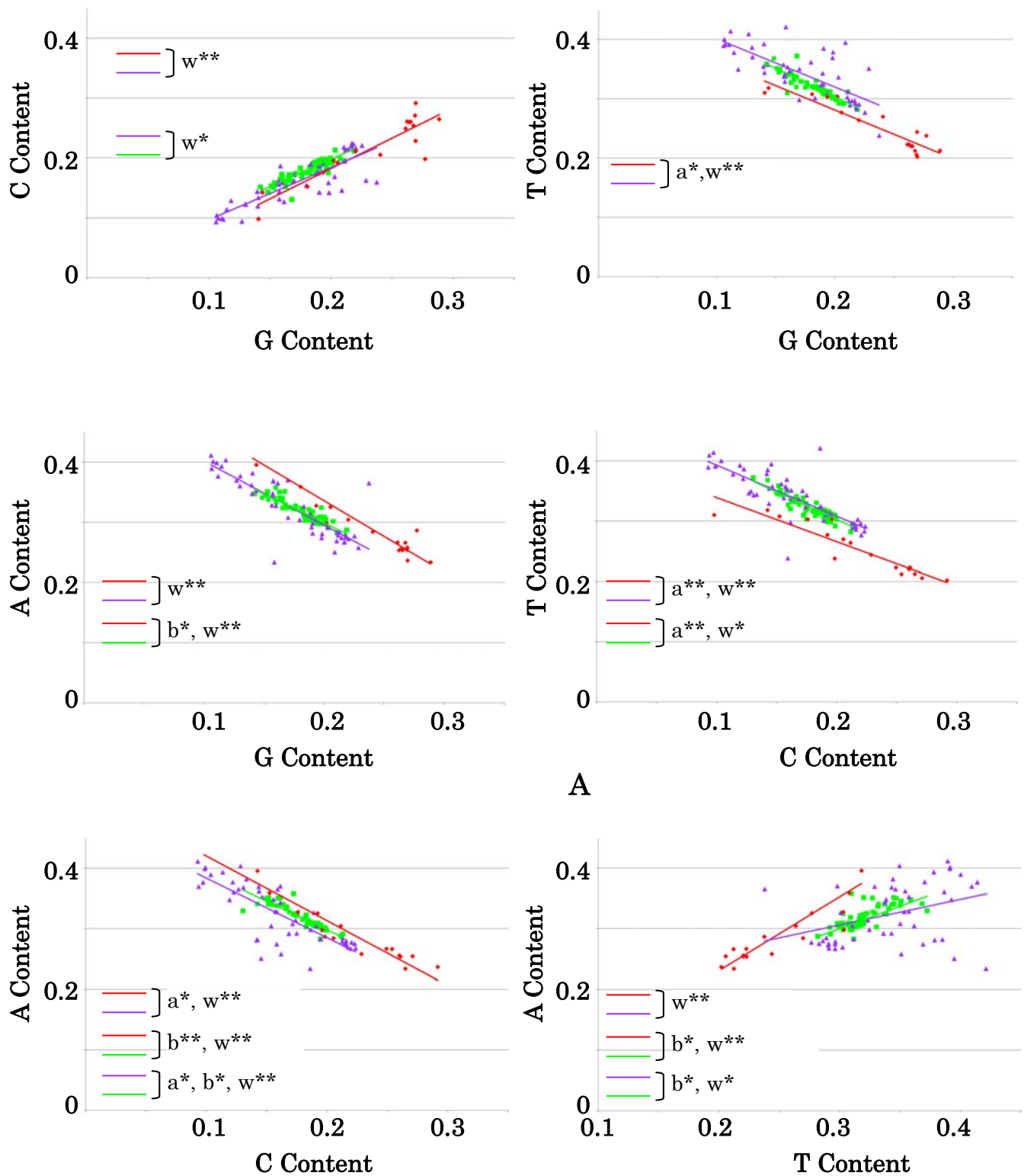
Fig. (3). Correlation of nucleotide contents plotted against G, C or T content in coding region in eukaryotic nuclei, chloroplasts and mito-chondria. Red, green and purple represent 18 eukaryotes, 97 chloroplasts and 49 plant mitochondria, respectively. Colored regression lines were computationally drawn. On the figure, a and b represent the slope and intercept, respectively, in y = ax + b, and w represents both factors. ** p < 0.01, * 0.05 > p > 0.01.

Table 1. Correlation of Nucleotide Contents. Values in Parentheses are Correlation Coefficients

Chloroplasts (99)	
G = 1.000G - 0.000 (1.00)	C = 1.000C + 0.000 (1.00)
C = 0.917G + 0.015 (0.89)	G = 0.856C + 0.046 (0.89)
T = - 0.950G + 0.493 (0.90)	T = - 0.912C + 0.486 (0.90)
A = - 0.967G + 0.492 (0.90)	A = - 0.944C + 0.487 (0.91)
A = 1.000A + 0.000 (1.00)	T = 1.000 + 0.000 (1.00)
G = - 0.835A + 0.447 (0.90)	G = - 0.858T + 0.457 (0.90)
C = - 0.873A + 0.458 (0.91)	C = - 0.882T + 0.464 (0.90)
T = 0.708A + 0.096 (0.72)	A = 0.740T + 0.078 (0.72)
Mitochondria (49)	
G = 1.000G - 0.000 (1.00)	C = 1.000C + 0.000 (1.00)
C = 0872G + 0.010 (0.85)	G = 0.834C + 0.039 (0.85)
T = - 0.811G + 0.482 (0.74)	T = - 0.847C + 0.478 (0.79)
A = - 1.061G + 0.508 (0.84)	A = - 0.987C + 0.483 (0.80)
A = 1.000A + 0.000 (1.00)	T = 1.000T + 0.000 (1.00)
G = - 0.669A + 0.390 (0.84)	G = - 0.681T + 0.406 (0.74)
C = - 0.650A + 0.372 (0.80)	C = - 0.744T + 0.415 (0.79)
T = 0.319A + 0.238 (0.37)	A = 0.425T + 0.178 (0.37)
Eukaryotes (18)*	
G = 1.000G - 0.000 (1.00)	C = 1.000C + 0.000 (1.00)
C = 1.013G - 0.020 (0.90)	G = 0.806C + 0.059 (0.90)
T = - 0.824G + 0.446 (0.93)	T = - 0.734C + 0.413 (0.93)
A = - 1.189G + 0.574 (0.95)	A = - 1.072C + 0.528 (0.96)
A = 1.000A + 0.000 (1.00)	T = 1.000T - 0.000 (1.00)
G = - 0.754A + 0.457 (0.95)	G = - 1.047T + 0.499 (0.93)
C = - 0.854A + 0.470 (0.96)	C = - 1.172T + 0.514 (0.93)
T = 0.609A + 0.073 (0.86)	A = 1.219T - 0.013 (0.86)

* Data related to eukaryotes are referred from an earlier report of ours [17].

Animal Mitochondria

The codon usages in human nuclei and mitochondria are shown in Fig. (**4**). The patterns differ between the two organelles. As presented in Fig. (**2**), codon usage patterns are closely linked to amino acid compositions. Amino acid compositions, based on various animal mitochondria, are shown in Fig. (**4**). Extremely high Leu, Ile and Thr contents were observed in human (*H. sapiens*) mitochondria. In our previous studies [1-3], a gene assembly coding for 3,000 - 7,000 amino acid residues was shown to have almost the same amino acid composition as the complete genome, even though the coding size may differ.

In vertebrate mitochondria, extremely high Leu, Ile and Thr concentrations were observed in salamander, reptile (*Eurycea cirrigera*), bird (*Gallus gallus*) and fish (*Latimeria chalumnae* and *Diodon holocunthus*). Comparing the shapes of the radar charts, based on Pro, Ala and Thr concentrations, that of the ancient fish, the coelacanth (*L. chalumnae*), more closely resembles those of salamander and bird compared than those of other fish (*D. holocunthus*). Those of salamander and bird resemble each other, consistent with the close phylogeny of these species. These results are consistent with the already established phylogenetic concept.

In invertebrate mitochondria, high Phe content, as well as high Leu and Ile concentrations, were observed in Daphnia (*Daphnia pulex*), insect (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*) (Fig. **4**). These characteristic amino acid composition patterns have never been observed in the nuclear DNA of any organism previously examined [17, 18], nor in any plant organelle, as described in the present study (Fig. **2**). However, Phe and Ser concentrations were low and, in *Dictyostelium discoideum*, the Ile concentration was higher than the Leu concentration. The shape of the radar chart based on Lys, Asp, Glu and Ser in *D. discoideum* differs completely from those of the other invertebrates. This is consistent with differences in coding gene numbers in mitochondria between *D. discoideum* and the other invertebrates. Thus, it seems that codon evolution in animal mitochondria is governed by different rules to those that exist in the nucleus or in plant organelles (Figs. **2** and **3**). In nuclear DNA, the abundance of some amino acids was correlated to G + C contents [23], and additionally the amino acid composition coded in the complete genome can be estimated by just a single nucleotide content [19]. Furthermore, it has been shown that codon frequencies are strongly correlated with some amino acid concentrations in nuclear DNA based on the complete genome [24].

Correlations Between Nucleotides in Animal Mitochondria

To evaluate the rule that controls codon evolution in animal mitochondria, correlations between nucleotide contents were examined in the coding region. The G content was plotted against C, T and A contents, C content was plotted against T and A content, and T content was plotted against A content (Fig. **5**). On these graphical plots, the vertebrate and invertebrate groups were apparently isolated into two different distributions without correlation, and only correlations between C and T and between G and A were seen (linear regression lines not shown) (Fig. **5**). Thus, nucleotide substitutions between purines or pyrimidines seem to be controlled, whereas those between purine and pyrimidine seem not to be controlled. When plant mitochondrial data were simultaneously plotted over animal mitochondrial data, they were localized in the invertebrate group. On the basis of the correlations between C and T and between G and A, three distribution groups, having different regression lines, were obtained. Thus, codon evolution among plant, vertebrate and invertebrate mitochondria took place under different conditions. However, plant and vertebrate groups are distributed within the invertebrate group rather than in three independent groups. Thus, based on the present results, it seems that invertebrate mitochondria possess a mixture of different characteristics.
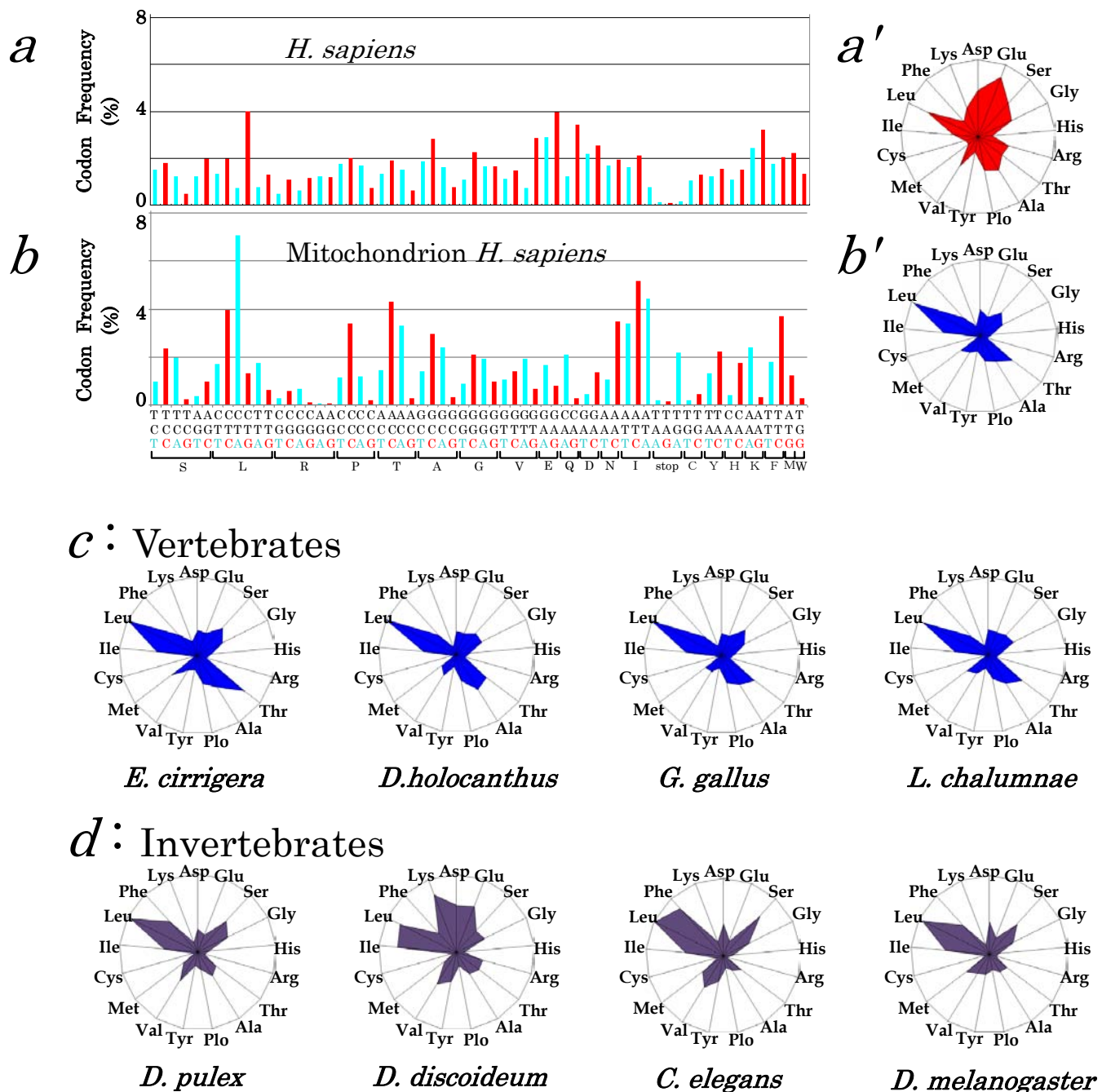
Fig. (4). Codon usage patterns and amino acid compositions. Codon usage (bar) and amino acid composition (radar chart) are expressed as percentages of all codons and amino acids, respectively.

Correlations Between the Same Nucleotides

Based on GC and AT skews, mitochondria were classified into three groups, and their DNA deviated from Chargaff's second parity rule, while chloroplast genomes shared the patterns of bacterial genomes [25]. Deviations from Chargaff's second parity rule were attributed to the small size of the mitochondrial DNA. Plotting nucleotide contents from the same mitochondrial DNA revealed no correlation between pyrimidine and purine, although linear regression lines were obtained between purines and pyrimidines (Fig. **5**). However, plotting the contents of the same nucleotide between coding and non-coding regions produced linear regression lines in any organelle, including mitochondria (Fig. **6** and Additional data file 3). These results indicate that any nucleotide substitution is governed by linear formulas. The slope and constant values in the regression line, $y = ax + b$ are shown in Table **2**.
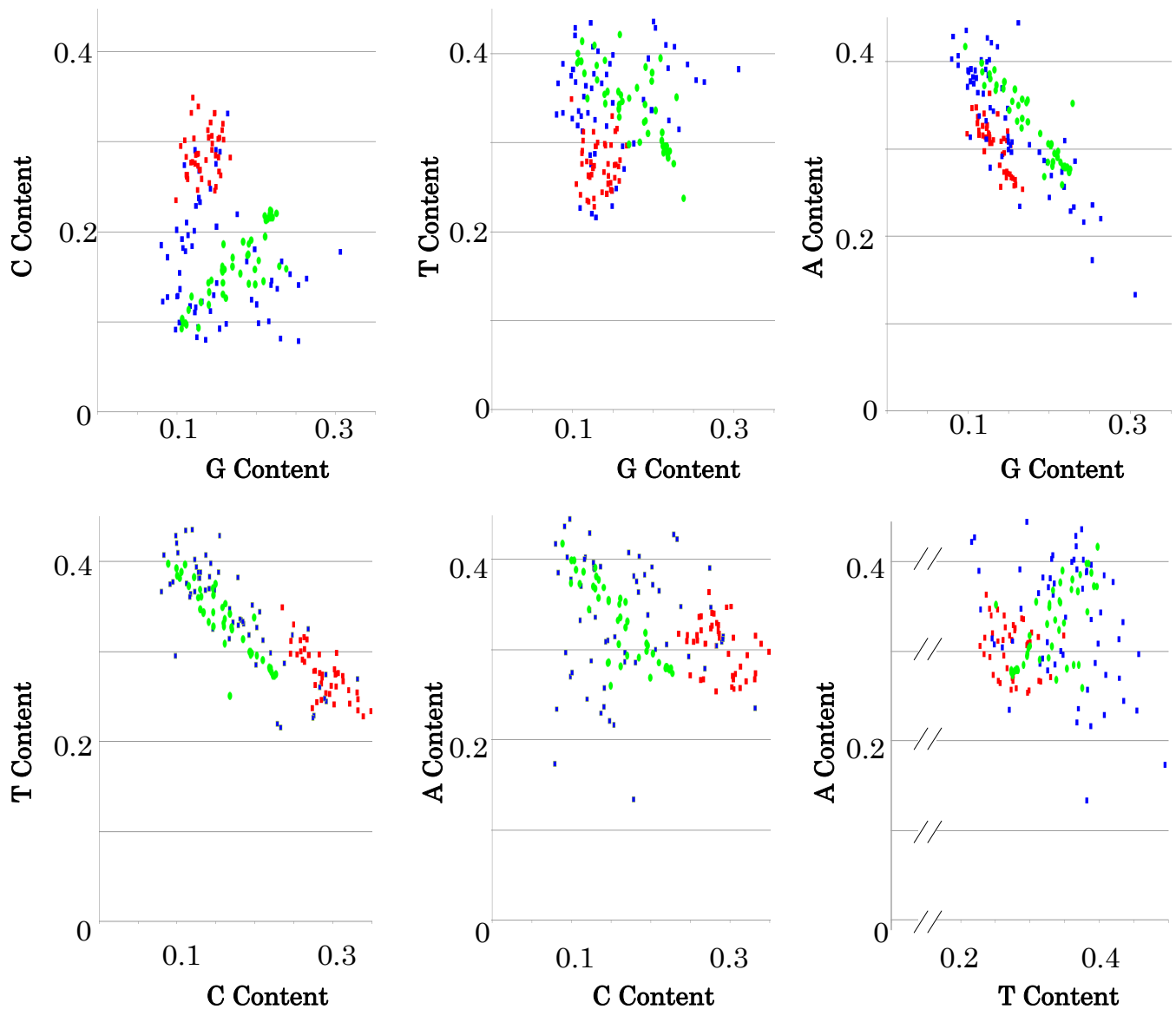
Fig. (5). Correlation of nucleotide contents plotted against G, C or T content in coding region in vertebrate and invertebrate mitochondria. Red and blue represent 45 vertebrate and 60 invertebrate mitochondria, respectively. Plant mitochondria (green) were simultaneously plotted.

Plotting the content of the same nucleotide in an organelle single strand against that in coding or non-coding region DNA, in chloroplasts and mitochondria (used as comparison), produced linear regression lines. Two lines, based on each organelle, always crossed at almost the same nucleotide concentrations, and significantly differed from each other (Fig. **7** and Additional data file 4). These results indicate that nucleotide substitutions between chloroplasts and mitochondria are controlled by different rules, even in the same plant. The slope and constant values of the regression lines are shown in Table **3**. In nuclear DNA, the correlations among nucleotide contents in the organelle, coding and non-coding region DNAs were linear [17]. Plotting the content of the same nucleotide in an organelle single strand against that in coding or non-coding region DNA revealed correlations expressed by linear regression lines (Fig. **8** and Table **4**). Similarly, linear regression lines were obtained when comparing nucleotide contents in the organelle, coding and non-coding

DNAs, in both chloroplasts and plant mitochondria (data not shown).

Assuming that codon evolution is based on stoichiometric reactions of DNA with certain small molecules, numerous small molecules that change with evolutionary time should exist in each organelle to control time-dependent nucleotide alternations. Such a small molecule has not been identified. It has been reported that evolution is based on neutral mutations [26, 27]. In addition, Sueoka [28] proposed that mutations are not random but occur with a directionality toward higher or lower G + C content of DNA. Asymmetrical directional mutation pressure was observed in mammalian mitochondria [29]. If codon evolution is based on just spontaneous mutation, codon evolution should be governed by the same rules among the nucleus, chloroplast and mitochondrion. However, there are different rules in different organelles. Rapid mutations were observed
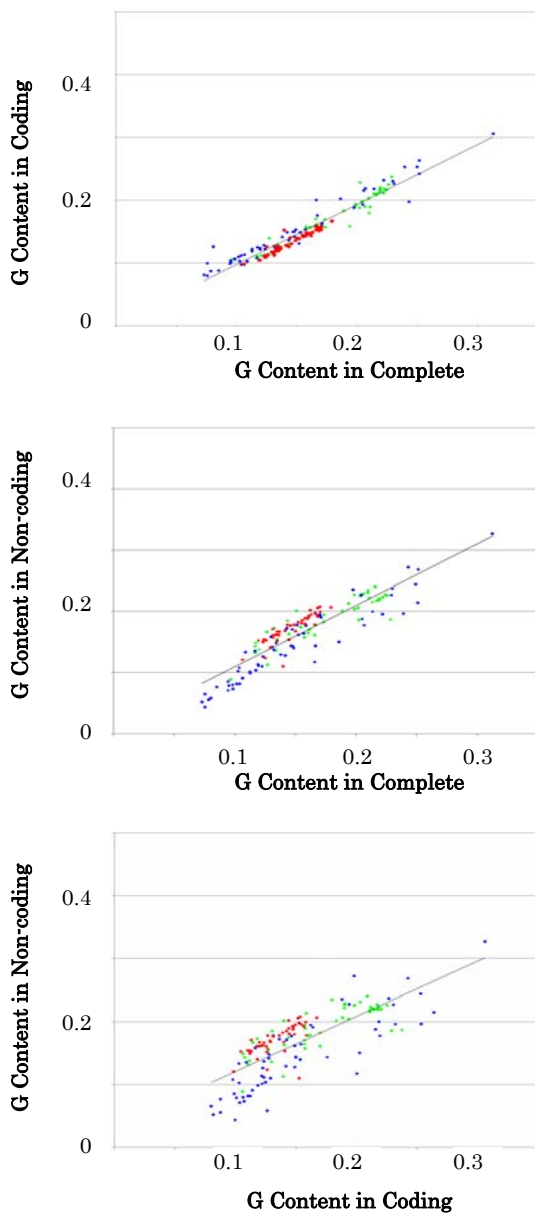
Fig. (6). Correlation of G contents compared between the complete single strand and its codinfg or non-coding regions. Red, blue and green represent vertebrate, invertebrate and plant mitochondria, respectively.

in animal mitochondria [30]. As codon evolution occurs synchronously over the genome [1, 2, 19], the effect on DNA should be steady and homogeneous. To satisfy these conditions, an "organelle specific field" existing inside each organelle has been proposed in this study. The fact that completely different codon usages can show almost the same amino acid compositions (Fig. **9**) strongly supports the existence an "organelle biases". Thus, all organisms show a similar "star-shape" in their cellular [31-33] and genomic [1, 17, 19] amino acid compositions on the radar chart. With respect to the same nucleotide, the correlations between the single strand and its coding or non-coding regions are linear, even in mitochondria. These results clearly indicate that coding and non-coding regions are strongly linked to each other in nucleotide alternations.

Table 2.   Correlation of Nucleotide Contents. Values of Slope, Constant and Correlation Coefficients of Regression Lines Based on Vertebrate, Invertebrate and Plant Mitochondria

|		Slope	Constant	r
G	Complete:Coding	0.954	0.003	0.97
	Complete:Non-coding	1.003	0.009	0.90
	Coding:Non-coding	0.875	0.034	0.78
C	Complete:Coding	1.084	- 0.012	0.99
	Complete:Non-coding	0.811	0.026	0.93
	Coding:Non-coding	0.690	0.047	0.86
T	Complete:Coding	1.009	0.005	0.97
	Complete:Non-coding	0.934	0.009	0.89
	Coding:Non-coding	0.789	0.052	0.77
A	Complete:Coding	1.070	- 0.030	0.98
	Complete:Non-coding	0.824	0.071	0.89
	Coding:Non-coding	0.662	0.129	0.78
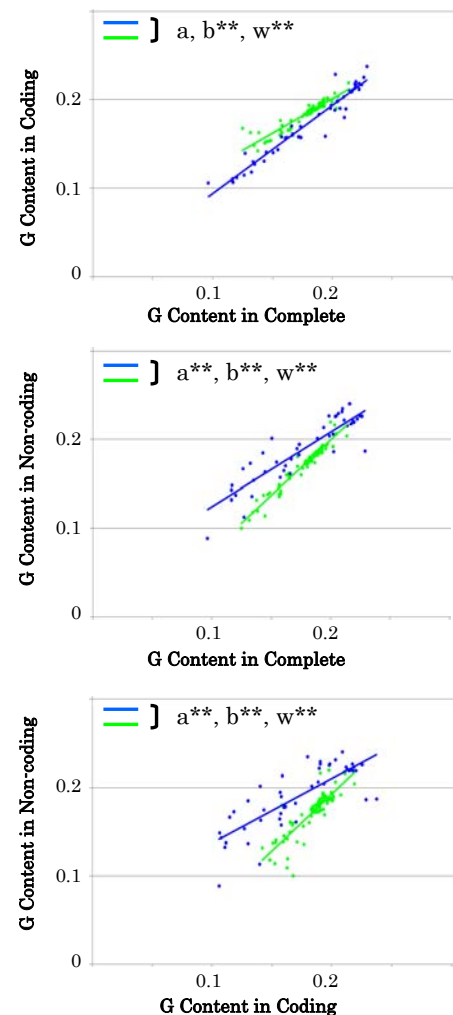
Fig. (7). Correlation of G contents compared between the complete single strand and its coding or non-coding regions. Blue and green represent plant mitochondria and chloroplasts, respectively. On the figure, a and b represent the slope and intercept, respectively, in y = ax + b, and w represents both factors. ** p < 0.01, * 0.05 > p > 0.01.
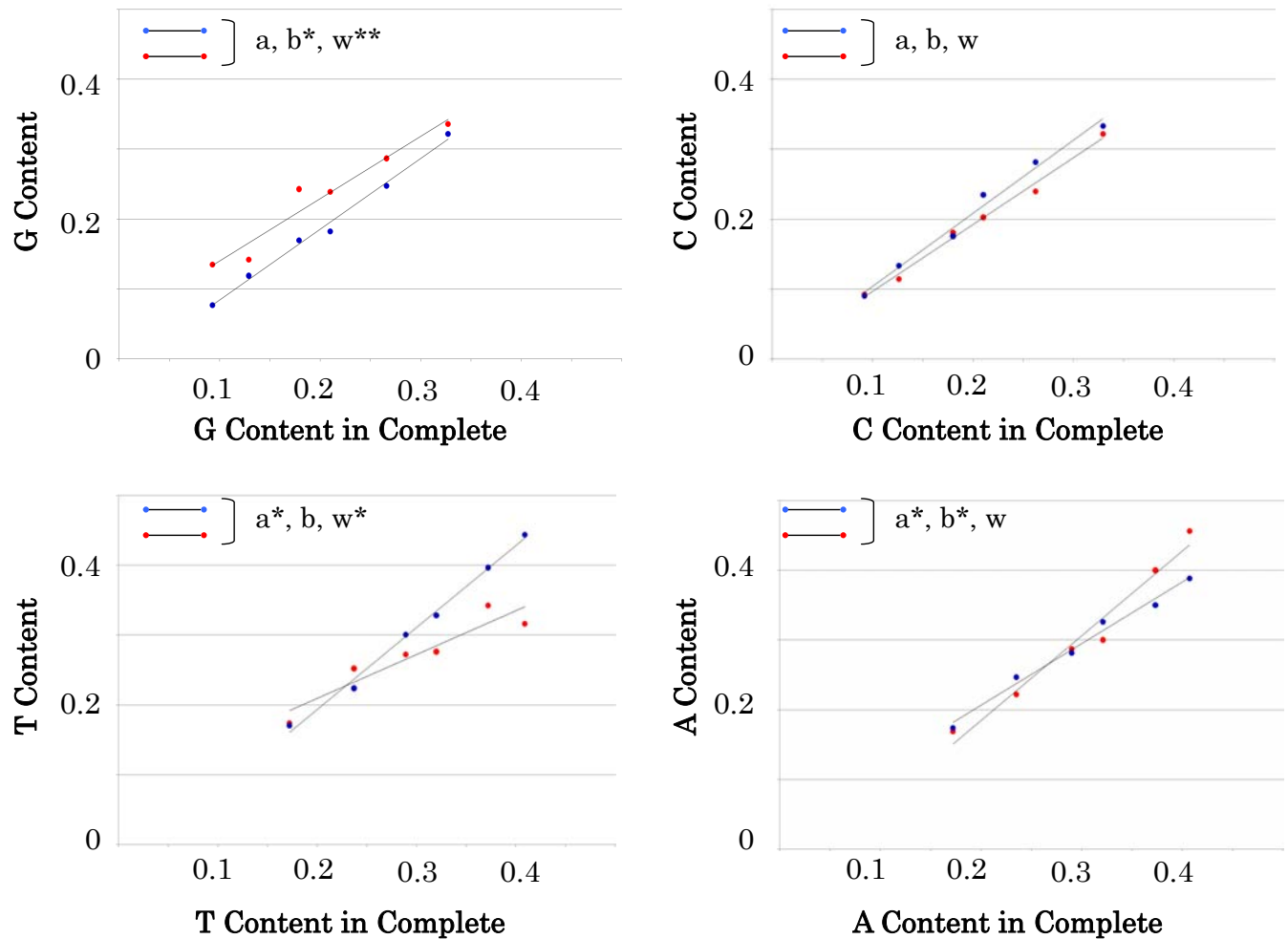
Fig. (8). Correlation of nucleotide contents compared between the complete single strand and its coding or non-coding regions. Red and blue represent correlations between the nuclear complete single strand and its coding or non-coding regions, respectively. Original data are cited from our previous study [17]. On the figure, a and b represent the slope and intercept, respectively, in $y = ax + b$, and w represents both factors. ** $p < 0.01$, * $0.05 > p > 0.01$.

Table 3.   Correlation of Nucleotide Contents. Values of Slope, Constant and Correlation Coefficients of Regression Lines Based on Chloroplasts and Mitochondria

|		Chloroplasts			Mitochondria		
		Slope	Constant	r	Slope	Constant	r
	Complete:Coding	0.781	0.045	0.93	0.993	- 0.005	0.97
G	Complete:Non-coding	1.247	- 0.050	0.97	0.839	0.040	0.90
	Coding:Non-coding	1.302	- 0.067	0.85	0.728	0.065	0.80
	Complete:Coding	0.809	0.037	0.96	0.955	0.002	0.98
C	Complete:Non-coding	1.236	- 0.046	0.97	0.956	0.014	0.97
	Coding:Non-coding	1.326	- 0.064	0.88	0.941	0.022	0.93
	Complete:Coding	0.831	0.051	0.93	0.984	0.018	0.93
T	Complete:Non-coding	1.181	- 0.054	0.96	0.685	0.088	0.83
	Coding:Non-coding	1.130	- 0.035	0.82	0.498	0.142	0.63
	Complete:Coding	0.804	0.058	0.96	1.078	- 0.027	0.98
A	Complete:Non-coding	1.244	- 0.073	0.97	0.840	0.049	0.92
	Coding:Non-coding	1.348	- 0.100	0.89	0.689	0.099	0.84
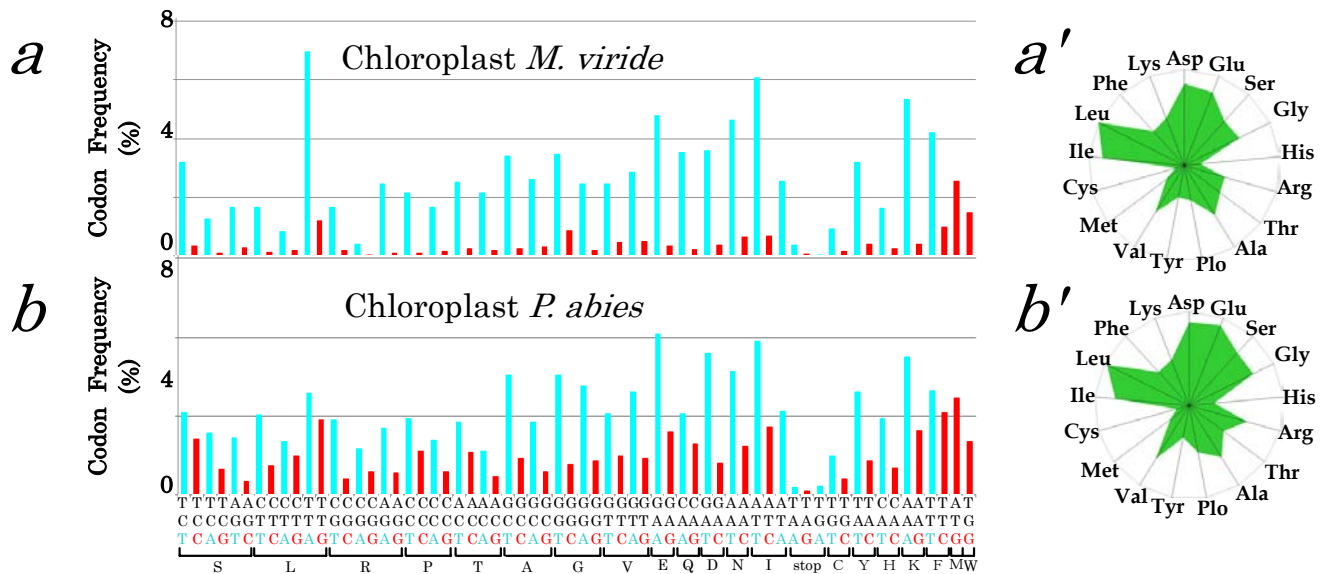
Fig. (9). Codon usage patterns and amino acid compositions of chloroplasts. Codon usage (bar), a and b, and amino acid composition (radar chart), a', and b', are expressed as percentages of all codons and amino acids, respectively.

Table 4. Correlation of Nucleotide Contents. Values of Slope, Constant and Correlation Coefficients of Regression Lines Based on Nuclear DNA

		Slope	Constant	r
	Complete:Coding	0.892	0.051	0.97
G	Complete:Non-coding	1.013	-0.017	1.00
	Coding:Non-coding	1.075	-0.061	0.97
	Complete:Coding	0.957	0.001	1.00
C	Complete:Non-coding	1.043	0.000	0.99
	Coding:Non-coding	1.073	0.002	0.98
	Complete:Coding	0.625	0.085	0.94
T	Complete:Non-coding	1.173	-0.041	1.00
	Coding:Non-coding	1.624	-0.131	0.92
	Complete:Coding	1.212	-0.058	0.98
A	Complete:Non-coding	0.881	0.031	0.99
	Coding:Non-coding	0.693	0.083	0.96

CONCLUSION

All nucleotide substitutions in the whole genome are wholly governed by a universal rule expressed by linear formulas based on the same nucleotide in any organelle. All nucleotide substitutions that occur in biological evolution and the substitutions in coding and non-coding regions are strictly linked to each other.

The apparent great divergence to *Homo sapiens* from bacteria is expressed by linear formulas with small turbulences based on the complete genome in biological evolution [17]. Chargaff's parity rules governing all organisms, expressed by the linear formulas G = C, T = A, T = -C + 0.5 and A = -G + 0.5, completely exclude these turbulences.

Thus, biological evolution seems to be observed as a result of mere nucleotide substitutions based on simple mathematical principles, while natural selection affects species preservation after nucleotide alternations. Therefore, it seems hard to elucidate a direct correlation between genotypic and phenotypic alternations in biological evolution.

ADDITIONAL DATA FILES

The following additional data files are available with the online version of this paper. Additional data file 1 is a list of species examined in the present study. Additional data file 2 shows correlations between different nucleotides in eukaryotes, chloroplasts and plant mitochondria. Additional data file 3 shows correlations between the same nucleotides in a complete single strand and its coding or non-coding regions, in vertebrate, invertebrate and plant mitochondria. Additional data file 4 shows between the same nucleotides in a complete single strand and its coding or non-coding regions, in plant mitochondria and chloroplasts.

REFERENCES

[1] Sorimachi K, Okayasu T. An evaluation of evolutionary theories based on genomic structures in *Saccharomyces cerevisiae* and *Encephalitozoon cuniculi*. Mycoscience 2004; 45: 345-50.

[2] Sorimachi K, Okayasu T. Genome structure is homogeneous based on codon usages. Curr Top Pept Protein Res 2007; 8: 19-24.

[3] Sorimachi K, Okayasu T. Gene assembly consisting of small units with similar amino acid composition in the *Saccharomyces cerevisiae* genome. Mycosicence 2003; 44: 415-7.

[4] Crick FHC. The origin of the genetic code. J Mol Biol 1968; 38: 367-79.

[5] Woese CR. Order in the genetic code. Proc Natl Acad Sci USA 1965; 54: 71-5.

[6] Wong JT-F. A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 1975; 72: 1909-12.

[7] Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 1986; 24: 28-38.

[8] Akashi H. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics 1994; 136: 927-35.

[9]   Jukes TH, Bhushar V. Silent nucleotide substitution and G content of some mitochondria and bacterial genes. J Mol Evol 1986; 24: 39-44.

[10] Osawa S, Ohama T, Yamano F, *et al.* Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. Proc Natl Acad Sci USA 1988; 85: 1124-8.

[11] Sueoka N. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 1988; 85: 2653-7.

[12] Sorimachi K, Okayasu T. Mathematical proof of the chronological precedence of protein formation over codon formation. Curr Top Pept Protein Res 2007; 8: 25-34.

[13] Chargaff E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experimentia 1950; VI: 201-9.

[14] Rudner R, Karkas JD, Chargaff. E. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. Proc Natl Acad Sci USA 1968; 60: 921-2.

[15] Sueoka N. Two aspects of DNA base composition: G+C content and translation-coupled deviation intra-strand rule of A=T and G=C. J Mol Evol 1999; 49: 49-62.

[16] Mitchell D, Bridge R. A test of Chargaff's second rule. Biochem Biophys Res Commun 2006; 340: 90-4.

[17] Sorimachi K, Okayas T. Codon evolution is governed by linear formulas. Amino Acids 2008; 34: 661-8.

[18] Sorimachi K, Okayasu T. Classification of Eubacteria based on their complete genome: where does Mycoplasmataseae belong? Proc R Soc Lond B 2004; 271(Suppl): S127-S130.

[19] Okayasu T, Sorimachi K. Organisms can essentially be classified according to two codon patterns. Amino Acids 2008; doi:10.1007/s00726-008-0059-0.

[20] Okayasu T, Ikeda M, Akimoto K, Sorimachi K. The amino acid composition of mammalian and bacterial cells. Amino Acids 1997; 13: 379-91.

[21] Raven JA, Allen JF. Genomics and chloroplast evolution: What did cyanobacteria do for plants? Gen Biol 2003; 4: 209-15.

[22] Gray MW, Burger G, Lang BF. Mitochondrial evolution. Science 1999; 283: 1476-81.

[23] Sueoka N. Correlation between base composition of deoxyribonucleic acid and amino acid composition of proteins. Proc Natl Acad Sci USA 1961; 47: 1141-9.

[24] Lobry JR. Influence of genomic G + C content on average amino-acid composition of proteins from 59 bacterial species. Gene 1997; 205: 309-16.

[25] Nikolaou C, Almirantis Y. Deviations from Chargaff's second parity rule in organellar DNA. Insights into the evolution of organellar genomes. Gene 2006; 381: 34-41.

[26] Kimura M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge Univ. Press 1983.

[27] Van Nimwegen E, Crutchfield JP, Huynen M. Neutral evolution of mutational robustness. Proc Natl Acad Sci USA 1999; 96: 9716-20.

[28] Sueoka N. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 1988; 85: 2653-7.

[29] Reyes A, Gissi C, Pesole G, Saccone C. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol Biol Evol 1998; 15: 957-66.

[30] Brown WM, George M, Wilson AC. Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci USA 1988; 76: 1967-71.

[31] Sorimachi K. Evolutionary changes reflected by the cellular amino acid composition. Amino Acids 1999; 17: 207-26.

[32] Sorimachi K, Okayasu T, Akimoto K, Niwa A. Conservation of the basic pattern of cellular amino acid composition during biological evolution in plants. Amino Acids 2000; 18: 193-6.

[33] Sorimachi K, Itoh T, Kawarabayasi Y, Okayasu T, Akimoto K, Niwa A. Conservation of the basic pattern of cellular amino acid composition during biological evolution and the putative amino acid composition of primitive life forms. Amino Acids 2001; 21: 393-9.