Amino Acids (2008) 34:661–668
DOI 10.1007/s00726-007-0024-3

ORIGINAL ARTICLE

Codon evolution is governed by linear formulas

K. Sorimachi · T. Okayasu

Received: 26 November 2007 / Accepted: 17 December 2007 / Published online: 8 January 2008

Abstract When nucleotide (G, C, T and A) contents were plotted against each nucleotide, their relationships were clearly expressed by a linear formula, $y = \alpha x + \beta$ in the coding and non-coding regions. This linear relationship was obtained from the complete single-stranded DNA. Similarly, nucleotide contents at all three codon positions were expressed by linear regression lines based on the content of each nucleotide. In addition, 64 codon usages were also expressed by linear formulas against nucleotide content. Thus, the nucleotide content not only in coding sequence but also in non-coding sequence can be expressed by a linear formula, $y = \alpha x + \beta$, in 145 organisms (112 bacteria, 15 archaea and 18 eukaryotes). Based on these results, the ratio of C/T, G/T, C/A or G/A one can essentially estimate all four nucleotide contents in the complete single-stranded DNA, and the determination of any ratio of two kinds of nucleotides can essentially estimate four nucleotide contents, nucleotide contents at the three different codon positions and codon distributions at 64 codons in the coding region. The maximum and minimum values of G content were $\sim$0.35 and $\sim$0.15, respectively, among various organisms examined. Codon evolution occurs according to linear formulas between these two values.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-007-0024-3) contains supplementary material, which is available to authorized users.

K. Sorimachi (✉)
Educational Support Center, Dokkyo Medical University,
Mibu, Tochigi 321-0293, Japan
e-mail: kenjis@dokkyomed.ac.jp

T. Okayasu
Center of Medical Informatics, Dokkyo Medical University,
Mibu, Tochigi 321-0293, Japan

Keywords Codon evolution · Nucleotide content · Linear empirical formula · Codon usage · Amino acid composition · Genome

Introduction

A half century ago, great scientific concepts regarding DNA structures were discovered: the helical double-stranded structure of DNA (Watson and Crick 1953) and Chargaff's parity rule, in which C/G, T/A and (C+T)/(A+G) ratios are one (Chargaff 1950). Chargaff's parity rule was later approximately proved in complete, single-stranded DNA (Rudner et al. 1968; Sueoka 1999). Chargaff's rules, which are universal for all replicating organisms, cannot reflect evolutionary differences based on different kingdoms. Therefore, the rules of genome construction remain to be discovered.

Changes in nucleotides or amino acid sequences have been applied to evolutionary research (Dayhoff et al. 1977; Sogin et al. 1986; Woese et al. 1990; Doolittle et al. 1994; Maizels and Weiner, 1994; DePouplana et al. 1998; Sakagami et al. 2006), on the assumption that amino acid sequence changes are linked to biological evolution. On the other hand, the basic pattern of cellular amino acid composition is conserved in various organisms from bacteria to mammalian cells (Sorimachi 1999; Sorimachi et al. 2000, 2001); and differences in cellular amino acid composition among organisms seem to reflect biological evolution. In addition, cellular amino acid compositions obtained experimentally resemble those conveniently calculated from a complete genome (Sorimachi et al. 2000, 2001). It was a puzzle to explain why these amino acid compositions obtained by two different methods, based on a protein mixture and an all gene assembly, resemble each other.

Quite recently, this puzzle was solved by the understanding that a genome is putatively constructed with gene assemblies forming similar amino acid compositions (Sorimachi and Okayasu 2003) and codon usages (Sorimachi and Okayasu 2004b) not only in prokaryotes but also in eukaryotes (Sorimachi and Okayasu 2004b; 2005). The amino acid composition based on the complete genome is mathematically represented by small gene assembly units coding 3,000–7,000 amino acid residues (Sorimachi et al. 2005), and equivalently the amino acid composition obtained experimentally from whole cells consisting of various proteins is represented by certain protein mixtures. A cellular amino acid composition based on phenotype eventually resembles the genomic amino acid composition based on genotype, as has been observed (Sorimachi et al. 2001a,b). Thus, the amino acid composition calculated from a complete genome reflects genomic structures. Using amino acid compositions, it is possible to compare among organisms not only the same genes but also the gene assemblies consisting of various different genes that represent the complete genome (Sorimachi and Okayasu 2003; 2005). Based on their complete genomes, bacteria are classifiable into two groups, 'S-type' represented by *Staphylococcus aureus* and 'E-type' represented by *Escherichia coli* (Sorimachi and Okayasu 2004a).

A genome constitutes coding and non-coding sequences and linkages of coding sequences and non-coding sequences are assumed to form two huge molecules, respectively. This coding sequence is constructed with putative small units with almost the same amino acid composition (Sorimachi and Okayasu 2003) and codon usage (Sorimachi and Okaysu 2004b). Thus, a small gene assembly unit is almost equivalent to a complete genome. The present study was designed to find certain rules that govern nucleotide alternations in biological evolution.

Materials and methods

Codon usage databases were obtained from the Kazusa DNA Research Institute (http://www.kazusa.or.jp/codon) and from GenomeNet (http://www.genome.ad.jp). In the present study, 145 organisms (Supplementary note) were examined. Nucleotide contents in the coding sequence excluding RNA genes were calculated from the all genes contained in the complete genome, and those in the non-coding sequence including RNA genes were obtained from the subtraction of nucleotides in coding sequence from the complete genome. A complete genome is assumed consisting of two huge molecules which represent a coding and a non-coding sequence. Under this assumption, figures based on nucleotide sequences exclude intra-strand deviations, and all organisms are compared with each other based on inter-species deviations (Sorimachi and Okayasu 2004a). The programs which estimate nucleotide contents in the coding sequence, nucleotide distributions at the three codon positions, 64 codons and amino acid compositions from just one nucleotide content were available on request. Statistical analysis was done with Excel 2003.

Results and discussion

Ratios of nucleotides

In *Ureaplasma urealyticum* (7.5×10^5 nt) and *Mycobacterium tuberculosis* (4.4×10^6 nt), the nucleotide contents and their ratios in the coding sequence were almost the same between the forward and reverse strands, and no significant difference was also observed in the non-coding sequence between the two strands (Table 1). Consistent results were obtained in *Arabidopsis thaliana* chromosome I (3.0×10^7 nt) (Table 1), although some difference was detected in *Encephalitozoon cuniculi* chromosome XI (2.7×10^5 nt), (data not shown). *U. urealyticum* was examined as the organism which has the smallest genomic size and *M. tuberculosis* was examined as the bacterium which has comparatively large genomic size among bacteria. *A. thaliana* was examined as the eukaryote which has been characterized earlier. Messenger RNA-synonymous strands with purine-rich clusters induce differences in nucleotide contents between the forward and reverse strands (Szybalski et al. 1966) and deviations from Chargaff's second parity rule (Nikolaou and Almirantis 2006; Bell and Forsdyke 2006). One of the reasons for this may be based on the small size of DNA segments examined (Sorimachi and Okaysu 2003; 2004a,b; Bell and Forsdyke 1999). On the other hand, symmetry was observed in long nucleotide sequences (Prabhu 1993) and in complete genomes (Qi and Cuticchia 2001). When the mutation rates are similar for complementary nucleotide substitutions between the two strands, nucleotide contents are the same between them (Sueoka 1995). These intra-strand differences may be based on mutational (Lobry and Sueoka 2002), replicational and transcriptional biases (McInerney 1998), or inversions and inverted transpositions (Albrecht-Beuhler 2006); however the reason for these differences has not yet been completely evaluated. The non-coding sequence complementally correlates with the coding sequence to satisfy Chargaff's rules. The total number of genes ($\sim$25,000) is quite similar between *Homo sapiens* (International Human Genome Sequencing Consortium 2001; Venter et al. 2001) and sea urchin (Sea Urchin Genome Sequencing Consortium 2006), although their complete genome sizes, 3.2×10^8 and 8.1×10^8 bp, respectively, differ substantially. In addition, $\sim$70% of genes were the same between these two species.

Table 1 Ratios of nucleotide, C/G, T/A and (C+T)/(A+G) in the complete single-strand DNA

	G	C	T	A	C/G	T/A	(C+T)/(A+G)
Ureaplasma urealyticum							
Coding[a]							
Forward	0.143	0.115	0.343	0.399	0.803	0.861	0.845
Reverse	0.141	0.116	0.341	0.402	0.825	0.850	0.844
Both	0.142	0.115	0.342	0.400	0.813	0.856	0.844
Non-coding[b]							
Forward	0.117	0.135	0.398	0.349	1.156	1.140	1.144
Reverse	0.121	0.133	0.395	0.351	1.097	1.125	1.118
Both	0.119	0.134	0.397	0.350	1.125	1.132	1.130
Complete							
Single strand	0.129	0.126	0.372	0.373	0.970	0.997	0.990
Mycobacterium tuberculosis							
Coding[a]							
Forward	0.335	0.322	0.174	0.168	0.960	1.036	0.985
Reverse	0.336	0.322	0.173	0.169	0.959	1.023	0.981
Both	0.336	0.322	0.174	0.169	0.960	1.030	0.983
Non-coding[b]							
Forward	0.322	0.333	0.170	0.175	1.036	0.975	1.014
Reverse	0.322	0.333	0.171	0.174	1.036	0.984	1.017
Both	0.322	0.333	0.171	0.174	1.036	0.980	1.016
Complete							
Single strand	0.327	0.329	0.172	0.172	1.004	1.000	1.002
Arabidopsis thaliana							
Coding[a]							
Forward	0.239	0.203	0.272	0.287	0.849	0.948	0.903
Reverse	0.238	0.202	0.272	0.287	0.850	0.946	0.902
Both	0.239	0.203	0.272	0.287	0.849	0.947	0.903
Non-coding[b]							
Forward	0.170	0.176	0.328	0.326	1.036	1.006	1.016
Reverse	0.169	0.176	0.328	0.326	1.038	1.006	1.017
Both	0.170	0.176	0.328	0.326	1.037	1.006	1.017
Complete							
Single strand	0.179	0.180	0.320	0.321	1.003	0.999	1.000

[a] RNA genes were removed

[b] RNA genes were included

Thus, evolutionary differences between these two species may be based on differences in the non-coding sequence.

Nucleotide contents against G content

Plotting C, T and A contents against G content, these relationships were clearly expressed by a linear formula, $y = \alpha x + \beta$, in the coding and non-coding regions. Further, linear correlations were obtained from other nucleotide combinations (Fig. 1). Quite recently, linear relationships between nucleotide contents in complete single-stranded DNA consisting of coding and non-coding sequences were obtained from the huge number of bacteria, archaea, eukaryotes and viruses consisting of double-stranded DNA (Mitchell and Bridge 2006). Based on linear formulas presented in Table 2, determination of any ratio of two kinds of nucleotide can essentially estimate the contents of all four nucleotides. In addition, according to Chargaff's rules, G = C, T = A, T = -C +0.5 and A = -G + 0.5; thus, the ratio of C/T, G/T, C/A or G/A can essentially estimate all four nucleotide contents.

Nucleotide contents in the coding regions

C, T and A contents were plotted against G content in 145 organisms (Fig. 2**a**). Regression lines based on different nucleotide combinations are shown in Electronic Supplementary Material Fig. 1. Each regression line and its correlation coefficient based on 145 organisms are summarized in Table 2; those based on different kingdoms are summarized in Supplementary Table 1. Every correlation coefficient was close to one. C and A versus G content were expressed by linear formulas with almost the same slope ($\sim$1.3) and showing a conflicting sign, while C and A versus T content were expressed by formulas with the same slope, but showing reversed signs. The summations of intersections representing nucleotide contents were one, and those of the slopes were zero; G + C + A + T = 1. Similarly, correlations of G and T versus C or A content were expressed by linear formulas with the reciprocal of $\sim$1.3 as the slope. As the two linear formulas representing G and C contents intersect at a G content of 0.36 (Fig. 2**b**, left panel and Table 2), the ratio of C to G was less than one below a G content of 0.36. In fact, the ratio of C to G of many organisms with a G content of less than 0.36 was less than one (Fig. 2**b**, middle panel). Similarly, as the two linear formulas representing A and T contents intersect at a G content of 0.32 (Fig. 2**b** right panel), the ratio of T to A of many organisms with a G content of less than 0.32 is less than one. Furthermore, when the two lines, (A + G) and (T + C), intersect at a G content of 0.34, the ratio of (T + C) to (A + G) is less than one below a G content of 0.34 (Fig.2**c**). Thus, the ratios of C to G, T to A and (T + C) to (A +G) strongly depend on the G content; these values approach one when the G content approaches the value where the two lines intercross. These relationships can explain Szybalski's rule, which indicates that (T+C)/(A+G) < 1 in coding sequence (Szybalski et al. 1966).

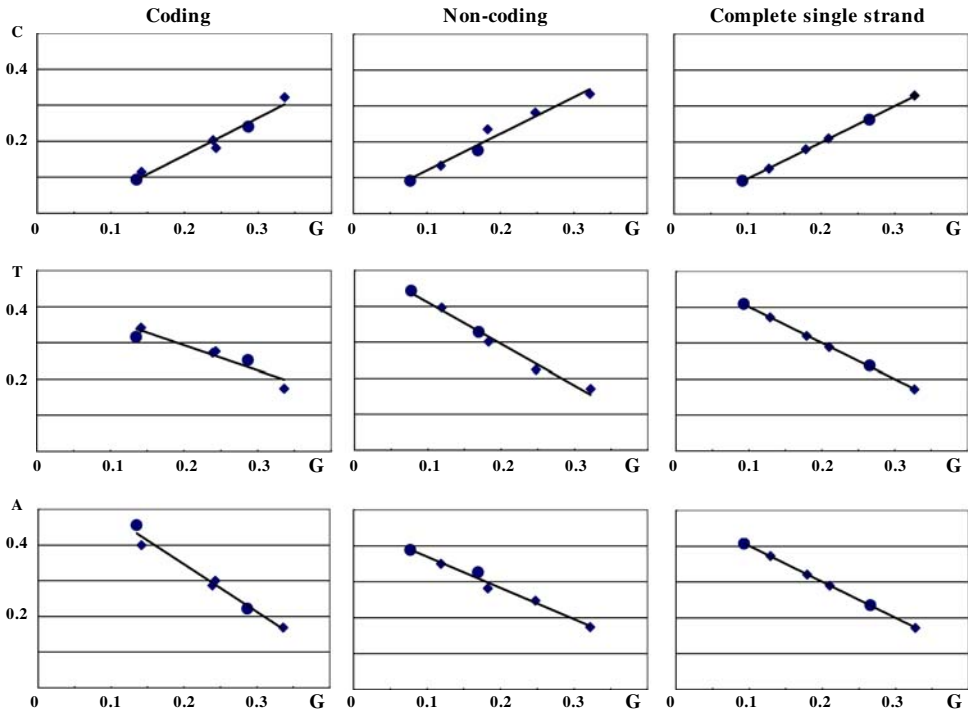

Fig. 1 Nucleotide contents versus G content in the complete single-strand DNA. *Closed diamond* and *circle* represent bacteria (*Ureaplasma urealyticum, Mycobacterium tuberculosis, Treponema pallidum* and *Carboxydothermus hydrogenoformans*), and eukaryotes (*Arabidopsis thaliana* and *Plasmodium falciparum*), respectively. All regression lines were drawn computationally

Nucleotide contents at different codon positions

The four nucleotide frequencies in human (Zhang and Chou 1993; 1996) and *E. coli* (Zhang and Chou 1994) genes were graphically presented by a point in a three-dimensional space. Meanwhile, the similar codon usage approach was also used to analyze the HIV proteins (Chou and Zhang 1992) and anti-sense proteins (Chou et al. 1996). On the other hand G, C, T and A contents at all three codon positions were expressed by linear regression lines based on the content of each nucleotide (Electronic Supplementary Material Fig. 2*a–d*). Nucleotide contents varied very considerably at the third codon position, while only small changes were observed at the second codon position. This relates to synonymous codons, which are degenerate

Table 2 Regression lines representing nucleotide contents in the coding region based on 145 organisms

G	=	1.00G	−	0.00 (1.00)	C	=	1.00C	+ 0.00 (1.00)
C	=	1.24G	−	0.09 (0.95)	G	=	0.73C	+ 0.09 (0.95)
T	=	−0.95G	+	0.49 (0.97)	T	=	−0.72C	+ 0.41 (0.96)
A	=	−1.30G	+	0.60 (0.97)	A	=	−1.01C	+ 0.50 (0.98)
Σ		0		1.00	Σ		0	1.00
A	=	1.00A	+	0.00 (1.00)	T	=	1.00T	− 0.00(1.00)
G	=	−0.72A	+	0.45 (0.97)	G	=	−1.00T	+ 0.50(0.97)
C	=	−0.96A	+	0.49 (0.98)	C	=	−1.29T	+ 0.55(0.96)
T	=	0.68A	+	0.06 (0.94)[a]	A	=	1.29T	− 0.55(0.94)[a]
Σ		0		1.00	Σ		0	1.00

The value in parentheses is a correlation coefficient obtained from 145 organisms

$P < 10^{-66}$

[a] $t = 31.930$

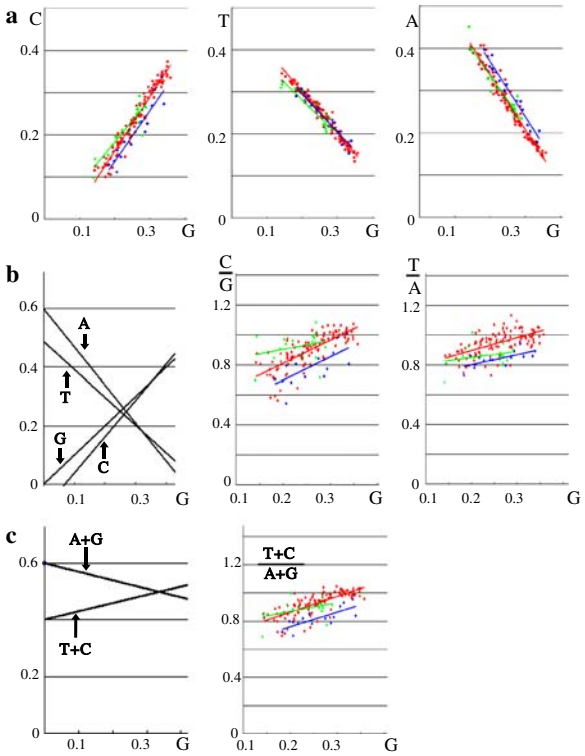

Fig. 2 Nucleotide contents and nucleotide ratios against G content. **a**, C, T and A contents in the coding sequence were plotted against the G content in the coding sequence of 145 organisms. **b,** Linear formulas representing A, T, and C contents were based on the regression lines drawn in figure 1a. **c,** Linear formulas, (A + G) and (T + C), were obtained from the 4 linear formulas drawn in figure 2**b.** *Red, blue* and *green* represent bacteria, archaea and eukaryotes, respectively. *Colored* regression lines were computationally drawn

Table 3 Regression lines representing nucleotide contents at the three codon positions based on 145 organisms

First letter						Second letter						Third letter					
G																	
g_1	=	0.257G	+	0.049	(0.93)	g_2	=	0.161G	+	0.016	(0.95)	g_3	=	0.583G	−	0.065	(0.99)
c_1	=	0.305G	−	0.009	(0.89)	c_2	=	0.177G	+	0.029	(0.84)	c_3	=	0.762G	−	0.106	(0.95)
t_1	=	−0.200G	+	0.109	(0.94)	t_2	=	−0.074G	+	0.119	(0.59)[a]	t_3	=	−0.674G	+	0.259	(0.97)
a_1	=	−0.362G	+	0.185	(0.93)	a_2	=	−0.264G	+	0.169	(0.92)	a_3	=	−0.670G	+	0.245	(0.96)
$\sum$		0		0.333		$\sum$		0		0.333		$\sum$		0		0.333	
C																	
g_1	=	0.183C	+	0.072	(0.87)	g_2	=	0.120C	+	0.029	(0.92)	g_3	=	0.428C	−	0.016	(0.95)
c_1	=	0.250C	+	0.011	(0.95)	c_2	=	0.148C	+	0.040	(0.91)	c_3	=	0.602C	−	0.051	(0.98)
t_1	=	−0.149C	+	0.093	(0.92)	t_2	=	−0.068C	+	0.116	(0.70)	t_2	=	−0.503C	+	0.203	(0.94)
a_1	=	−0.284C	+	0.158	(0.95)	a_1	=	−0.200C	+	0.148	(0.91)	a_3	=	−0.527C	+	0.196	(0.98)
$\sum$		0		0.333		$\sum$		0		0.333		$\sum$		0		0.333	
T																	
g_1	=	−0.253T	+	0.176	(0.89)	g_2	=	−0.158T	+	0.096	(0.91)	g_3	=	−0586T	+	0.227	(0.97)
c_1	=	−0.305T	+	0.143	(0.87)	c_2	=	−0180T	+	0.118	(0.83)	c_3	=	−0.803T	+	0.285	(0.98)
t_1	=	0.205T	+	0.008	(0.94)	t_2	=	0.085T	+	0.079	(0.66)[b]	t_3	=	0.710T	−	0.087	(0.99)
a_1	=	0.353T	+	0.006	(0.88)	a_2	=	0.252T	+	0.040	(0.86)	a_3	=	0.680T	−	0.092	(0.95)
$\sum$		0		0.333		$\sum$		0		0.333		$\sum$		0		0.333	
A																	
g_1	=	−0.183A	+	0.164	(0.89)	g_2	=	−0.120A	+	0.090	(0.95)	g_3	=	−0.420A	+	0.196	(0.95)
c_1	=	−0.245A	+	0.134	(0.96)	c_2	=	−0.144A	+	0.113	(0.91)	c_3	=	0.570A	+	0.242	(0.96)
t_1	=	0.144A	+	0.019	(0.91)	t_2	=	0.060A	+	0.084	(0.64)[c]	t_3	=	0.477A	−	0.042	(0.91)
a_1	=	0.284A	+	0.016	(0.98)	a_2	=	0.203A	+	0.047	(0.95)	a_3	=	0.513A	−	0.063	(0.98)
$\sum$		0		0.333		$\sum$		0		0.333		$\sum$		0		0.333	

The value in parentheses is a correlation coefficient

[a] $t = 8.682$, $P < 10^{-14}$

[b] $t = 10.531$, $P < 10^{-18}$

[c] $t = 10.009$, $P < 10^{-17}$

codons. The ratios of C and G, T and A and (C + T) and (A + G) at the first codon position were < 1, while that of C and G at the second codon position was > 1 (Electronic Supplementary Material Fig. 3). The ratios of T and A and (C + T) and (A + G) at the second codon position were between 0.5 and 1.5, and those of C and G, T and A and (C + T) and (A + G) at the third codon position were between 0.5 and 2.0 (Electronic Supplementary Material Fig. 3). Thus, the Szybalski's rule cannot apply to the second and third codon positions. Moderate changes were observed at the first codon position. The regression lines and correlation coefficients are summarized in Table 3 and Electronic Supplementary Material Table 2*a–c*. The correlation coefficients were very close to one, except that of T content at the second codon position. This indicates that essentially any nucleotide content can be used to estimate nucleotide contents at the three codon positions. The summations of intersections representing nucleotide contents at the first, second and third codon positions were 1/3, and those of the slopes were zero (Table 3). Nucleotide alternations are strictly governed by these conditions regarding the intersection and slope in biological evolution. In addition, G and T were expressed by regression lines with almost the same slopes, but with the opposite sign (Table 3 and Electronic Supplementary Material Table 2*a–c*). A similar relationship between C and A was obtained.

Sixty-four codon usages

Regression lines representing 64 codons against certain nucleotide were computationally calculated from 145 organisms (data not shown), and 64 codon usages were estimated by 64 linear formulas. When the 0.189 G content of *Staphylococcus aureus* was applied to the formula, the 64 codons were estimated (Fig. 3**a**). This codon usage pattern resembled that calculated from the complete genome (Fig. 3**b**). Furthermore, the estimated amino acid composition resembled that directly calculated from the complete genome (Figs. 3**a'** and **b'**). Radar charts have been used to illustrate the difference in amino acid composition for predicting protein subcellular localization (Chou and Elrod 1999). Also, radar charts were applied in a different manner to show the subsite coupling for the cleavable peptides by HIV protease (Chou 1993). Radar charts were applied in a different manner to show

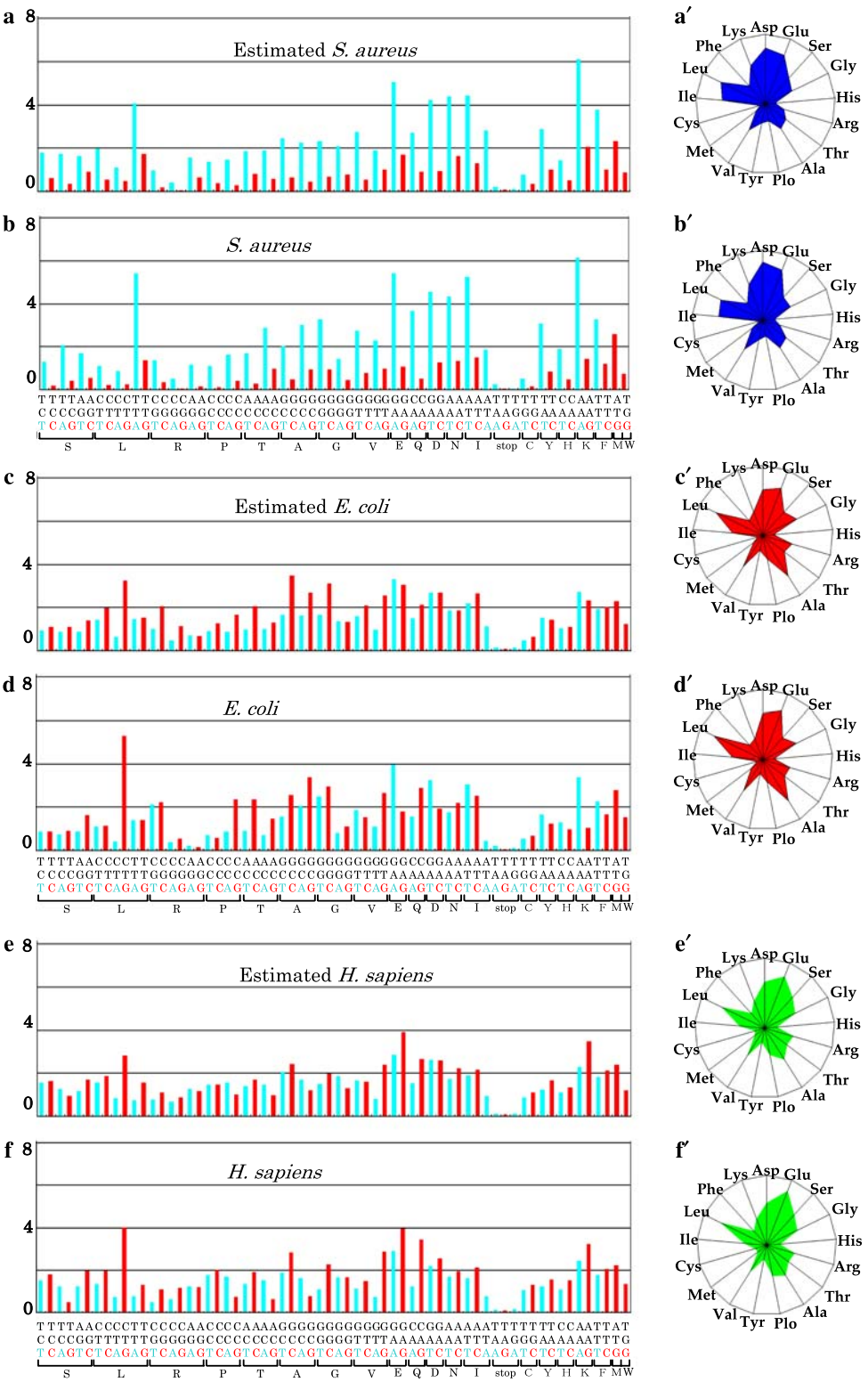

Fig. 3 Codon usage patterns and amino acid compositions. Codon usages (**a, c** and **e**) and amino acid compositions (**a', c'** and **e'**) were estimated by our programs, based on empirical formulas (Supplementary Methods Information), and codon usages (**b, d** and **f**) and amino acid composition (**b', d'** and **f'**) were calculated from the complete genome (Sorimachi *et al*. 2001). Codon usage (*bar*) and amino acid composition (*radar chart*) were expressed by percent of total codons and amino acids, respectively

interaction of HIV protease and proteins (Chou 1993). Consistent results were obtained from *E. coli* (Figs. 3**c**, **d**, **c'** and **d'** When the linear formulas based on these 145 organisms were applied to *H. sapiens*, whose 0.264 G content was similar to that of *E. coli*, at 0.273, the amino acid composition of *H. sapiens* resembled that of *E. coli* (data not shown). However, applying linear formulas based on eukaryotes to *H. sapiens*, the estimated amino acid composition resembled that based on the complete genome (Figs. 3**e'** and **f'**). Additionally, many regression lines

differ among bacteria, archaea and eukaryotes. Thus, biological evolution differs among different kingdoms. Some amino acid compositions were related to G+C contents in cells, but other amino acid compositions did not, as shown by experimental (Sueoka 1961) and theoretical (Lobry 1997) investigations. This means that the amino acid composition is not expressed by the G+C content. Thus, our present study is the first in which the amino acid composition and codon usages of organisms were estimated from the content of just one nucleotide. All nucleotide alternations, including Chargaff's rules, are correlated with each other by linear formulas, and these alternations occur synchronously (Sorimachi and Okayasu 2003; 2004b).

References

Albrecht-Beuhler G (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transcriptions. Proc Natl Acad Sci USA 103:7828–17833

Bell SJ, Forsdyke R (1999) Accounting units in DNA. J Theor Biol 197:51–6

Bell SJ, Forsdyke R (2006) Deviations from Chargaff's second parity rule correlate with direction of transcription. J Theor Biol 197:63–76

Chargaff E (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia VI:201–209

Chou K-C (1993) A vectorized sequence-coupling model for predicting HIV protease cleaveage sites in proteins. J Biol Chem 268:16938–16948

Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Eng 12:107–118

Chou KC, Zhang CT (1992) Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res Hum Retroviruses 8:1967–1976

Chou KC, Zhang CT, Elrod DW (1996) Do antisense proteins exist? J Protein Chem 15:59–61

Dayhoff MO, ParkCM, McLaughlin PJ (1977) Building a phylogenetic trees: cytochrome C. In Atlas of protein sequence and structure. Vol. 5, National Biomedical Fundation, Washington, D.C., pp 7–16

DePouplana L, Turner RJ, Steer BA, Schimmel P (1998) Genetic code origins: tRNAs older than their synthetases? Proc Natl Acad Sci USA 95:11295–11300

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2128

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lobry JR (1997) Influence of genomic G + C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205:309–316

Lobry JR, Sueoka N (2002) Asymmetric directional mutation pressures in bacteria. Genome Biol. 3: research0058.1–0058.14

Maizels N, Weiner M (1994) Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. Proc Natl Acad Sci USA 91:6729–6734

McInerney JO (1998) Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc Natl Acad Sci USA 95:10698–10703

Mitchell D, Bridge R (2006) A test of Chargaff's second rule. Biochem Biophys Res Commun 340:90–94

Nikolaou C, Almirantis Y (2006) Deviations from Chargaff's second parity rule in organelle DNA insights into the evolution of organelle genomes. Gene 381:34–41

Prabhu VV (1993) Symmetry observations in long nucleotide sequences. Nucleic Acids Res 21:2797–2800

Qi D, Cuticchia J (2001) Compositional symmetries in complete genomes. Bioinformatics 17:557–559

Rudner R, Karkas JD Chargaff.E (1968) Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proc Natl Acad Sci USA 60:921–922

Sakagami M, Nakayama T, Hashimoto T, et al (2006) Phylogeny of the centrohelida inferred from SSU rRNA, tubulin, and actin genes. J Mol Evol 61:765–775

Sea urchin genome sequencing consortium (2006) The genome of the sea urchin *Strongylocentrotus purpurantus*. Science 314:941–952

Sogin ML, Elwood HJ, Gunderson H (1986) Evolutionary diversity of eukaryotic small subunit rRNA genes. Proc Natl Acad Sci USA 83:1383–1387

Sorimachi K (1999) Evolutionary changes reflected by the cellular amino acid composition. Amino Acids 17:207–226

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

Sorimachi K, Okayasu T (2004a) Classification of eubacteria based on their complete genome: where does Mycoplasmataseae belong? Proc Biol Sci 271:S127–S130

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

Sorimachi K, Okayasu T (2005) Simulation analysis of genomic amino acid composition homogeneity based on putative small units. In Proceedings, the 9th world multi-conference on systemics, cybernetics and informatics, Orlando, Florida, USA, Vol. VI:190–196

Sorimachi K, Okayasu T, Akimoto K, et al (2000) Conservation of the basic pattern of cellular amino acid composition during biological evolution in plants. Amino Acids 18:193–196

Sorimachi K, Itoh T, Kawarabayasi Y, et al (2001) 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 21:393–399

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

Sueoka N (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of proteins. Proc Natl Acad Sci USA 47:1141–1149

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325

Sueoka N (1999) 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 49:49–62

Szybalski, Kubinski H, Sheldrick P (1966) Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis. Cold Spring Harb Symp Quant Biol 31:123–127

Zhang C-T, Chou K-C (1993) Graphic analysis of codon usage strategy in1490 human proteins. J Protein Chem 12:329–335

Zhang C-T, Chou K-C (1994) A graphic approach to analyzing codon usage in 1562 *Echerichia coli* protein coding sequences. J Mol Biol 238:1–8

Zhang C-T, Chou K-C (1996) An analysis of base frequencies in the anti-sense strands corresponding to the 180 human protein coding sequences. Amino Acids 10:253–262

Venter JC, et al (2001) The sequence of the human genome. Science 291:304–1351

Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribonucleic acid. Nature 171:964–967

Woese CR (1998) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859