Research Article

The frequencies of amino acids encoded by genomes that utilize standard and nonstandard genetic codes

Jun Tsuji, Robert Nydza, Erika Wolcott, Erin Mannor, Brad Moran, Grant Hesson, Tygue Arvidson, Kelly Howe, Rachel Hayes, Manuel Ramirez, and Mike Way

Siena Heights University, 1247 East Siena Heights Drive, Adrian, Michigan 49221

Abstract. A number of genomes use genetic codes that are different from the standard genetic code. Since many of these genomes have recently been sequenced, we can now study the evolution of the genetic codes by examining the amino acid frequencies encoded by these genomes. We calculated the amino acid frequencies encoded by twenty-two genomes using codon usage frequencies tabulated from GenBank. By chi-square analysis, no significant differences were observed in the amino acid frequencies encoded by fourteen different genomes that use the standard genetic code. However, among eight genomes that employ nonstandard genetic codes, we found differences between three animal mitochondrial genomes (*Fasciola hepatica*, *Drosophila melanogaster*, and *Homo sapiens*) and the remaining 19 genomes. Of the genomes that were studied, we also observed a trend between the usage of an amino acid and its occurrence in the genetic code. Taken together, our study supports the hypothesis that the standard genetic code formed early in the development of modern life and evolved in different lineages to form several nonstandard genetic codes.

Introduction

The genetic code is a cipher used to decode the information in DNA and RNA for the synthesis of proteins (Crick, 1990). The genetic code consists of codons, triplets of RNA bases, and the amino acids encoded by the codons. Sixty-four codons can be constructed using different combinations of four RNA bases (uracil, adenine, guanine, and cytosine). Sixty-one codons encode information for the assembly of specific amino acids during protein synthesis (Nirenberg and Leder, 1964; Leder and Nirenberg, 1964a; Leder and Nirenberg, 1964b; Bernfield and Nirenberg, 1965; Brimacombe et al., 1965; Nirenberg et al., 1965; Trupin et al., 1965). Three codons, however, do not code for any amino acid and are used as stop codons to terminate protein synthesis (Caskey et al., 1968).

The genetic code is used by a wide range of genomes, including viral, mitochondrial, chloroplast, prokaryotic, and eukaryotic nuclear genomes. Some viruses, such as the Human immunodeficiency virus and the Influenza virus, possess RNA genomes, while others, like the Variola (smallpox) virus, use DNA as its genetic material (Tortora et al., 2007). Mitochondria and chloroplasts have small circular DNA genomes

Correspondence to: Jun Tsuji, Siena Heights University, 1247 East Siena Heights Drive, Adrian, MI 49221; phone (517) 263-0731; fax (517) 264-7704; e-mail: JTSUJI@sienaheights.edu

Bios 81(1) 22–31, 2010

similar to those of prokaryotes *Mycoplasma pneumoniae* and *Escherichia coli*, which are haploid (Anderson et al., 1981; Himmelreich et al., 1996; Sato et al., 1999; Welch, 2002). Eukaryotes, such as *Pyrenomonas salina* (alga), *Saccharomyces cerevisiae* (yeast), *Dictyostelium discoideum* (slime mold), *Arabidopsis thaliana* (plant), *Caenorabditis elegans* (nematode), *Drosophila melanogaster* (fly), *Xenopus laevis* (frog), and *Homo sapiens* (humans), also utilize the genetic code for their nuclear genomes, which are diploid, except for *X. laevis* which is tetraploid (Eschbach et al, 1991; Goffeau et al., 1996; TCESC, 1998; Adams et al., 2000; TAGI, 2000; IHGSG, 2001; Eichinger et al., 2005; Morin et al., 2006;). Since the genetic code is common to so many organisms, it became known as the universal or standard code.

Although the genetic code was once thought to be the same for all organisms, exceptions to the standard genetic code are now known. Mitochondria, a few prokaryotes, and a handful of eukaryotic nuclear genomes were found to use alternative genetic codes (Osawa et al., 1992). These nonstandard genetic codes differ from the standard code in only a few codons. The changes usually involve altering a sense codon to a stop signal or assigning an amino acid to a stop codon (Lewin, 2008). In vertebrate mitochondria, for example, AGA and AGG are used as stop codons instead of coding for arginine, UGA codes for tryptophan in place of a termination signal, and AUA corresponds to methionine instead of isoleucine (Barrell et al., 1979). In the prokaryote *Mycoplasma capricolum*, UGA is not used to terminate translation, but instead codes for tryptophan (Yamao et al., 1985). In the eukaryote *Tetrahymena thermophila*, UAA and UAG are used for glutamine in place of stop codons (Horowitz and Gorovsky, 1985). As a result of these kinds of reassignments, many nonstandard genetic codes have fewer stop codons than the standard version.

The nonstandard genetic codes also differ from the standard code in the number of codons that code for a given amino acid. Since there are more codons than amino acids, most of the amino acids are represented by more than one codon. In the standard genetic code for instance, while only one codon (UGG) is used to assign tryptophan, any one of six codons (CGU, CGC, CGA, CGG, AGA, and AGG) can be used to designate arginine (Brimacombe et al., 1965; Nirenberg et al., 1965). These numbers may differ in nonstandard genetic codes, because certain codons have been reassigned from those in the standard code. In vertebrate mitochondria, for example, two codons (UGG and UGA), in place of one, code for tryptophan, while four (CGU, CGC, CGA, CGG), instead of six codons, encode arginine (Barrell et al., 1979). Many of the nonstandard genetic codes have more codons for serine and tryptophan and fewer codons for arginine as compared to the standard genetic code (Barrell et al., 1979; Yamao et al., 1985; Bessho et al., 1992; Santos and Tuite, 1995).

A relationship may exist between the number of codons for a given amino acid and the frequency of that amino acid in proteins (King and Jukes, 1969; Lewin, 2008). In general, amino acids that are encoded by 4 to 6 codons comprise a higher percentage of residues in proteins than those that are encoded by only 1 or 2 codons. For example, a survey of 207 random proteins revealed that alanine (4 codons), glycine (4 codons), leucine (6 codons) and serine (6 codons) occur among the highest frequencies, while methionine (1 codon) and tryptophan (1 codon) were the lowest. However, this trend is not absolute, for lysine (2 codons) and glutamate (2 codons) are present in proteins at higher frequencies than threonine (4 codons), proline (4 codons), and arginine (6 codons) although they are encoded by fewer codons (Klapper, 1977).

Amino acid frequencies in proteins have been analyzed to study the evolution of the genetic code (Brooks et al., 2002). Past studies suggest that the standard genetic code gradually expanded to its present size by the addition of amino acids (Brooks and Fresco, 2002; Jordan et al., 2005). Afterwards, mutations appeared to have occurred independently in different lines of evolution resulting in a number of nonstandard genetic codes (Osawa et al., 1992).

To our knowledge, little is known concerning the frequencies of amino acids encoded by genomes that utilize nonstandard genetic codes. With the recent availability of the complete genome sequences of a wide range of organisms and viruses, we can now calculate the frequencies of amino acids encoded by various genomes. In this study, we used amino acid frequencies to compare genomes that utilize nonstandard genetic codes to those that use the standard genetic code.

Materials and Methods

We constructed data sets using codon usage frequencies tabulated from GenBank (Nakamura et al., 2000). The first data set consisted of viral, chloroplast, prokaryotic, and eukaryotic nuclear genomes that utilize the standard genetic code: Human immunodeficiency virus, Influenza A virus, Variola virus, *Arabidopsis thaliana* chloroplast, *Mycoplasma pneumoniae*, *Escherichia coli* 536, *Pyrenomonas salina*, *Saccharomyces cerevisiae*, *Dictyostelium discoideum*, *Arabidopsis thaliana* nuclear, *Caenorabditis elegans*, *Drosophila melanogaster*, *Xenopus laevis*, and *Homo sapiens*. We then assembled a data set comprised of genomes that use nonstandard genetic codes. We selected the following as four representatives of nuclear genomes that employ nonstandard genetic codes: *Mycoplasma capricolum subsp. capricolum ATCC 27343* (UGA = Trp), *Candida albicans* (CUG = serine), *Tetrahymena thermophila* (UAA and UAG = glycine), and *Euplotes octocarinatus* (UGA = cysteine). We also selected four representatives of mitochondrial genomes that utilize nonstandard genetic codes: *Neurospora crassa* (AUA = methionine; CUU, CUC, CUA, and CUG = threonine; UGA = tryptophan), *Fasciola hepatica* (AAA = asparagine; AGA and AGG = serine; UGA = tryptophan; UAA = tyrosine), *Drosophila melanogaster* (AGA and AGG = serine; AUA = methionine; UGA = tryptophan), and *Homo sapiens* (AGA and AGG = stop; AUA = methionine; UGA = tryptophan).

We calculated amino acid frequencies (per 100 codons) using the genetic codes from the Codon Usage Database (Nakamura et al., 2000): *Mycoplasma capricolum* (Genetic code 4: mold, protozoan, coelenterate mitochondrial, and mycoplasma/spiroplasma), *Candida albicans* (Genetic code 12: alternative yeast), *Tetrahymena thermophila* (Genetic code 6: ciliate macronuclear and Dasycladacean), *Euplotes octocarinatus* (Genetic code 10: alternative ciliate macronuclear), *Neurospora crassa* mitochondrion (Genetic code 3: yeast mitochondrial), *Fasciola hepatica* mitochondrion (Genetic code 14: flatworm mitochondrial), *Drosophila melanogaster* mitochondrion (Genetic code 5: invertebrate mitochondrial), and *Homo sapiens* mitochondrion (Genetic code 2: vertebrate mitochondrial). Genetic code 1 (standard) was used for all others. The frequencies of the 20 amino acids encoded by each genome do not total 100, because the frequencies are based on the number of codons, some of which do not code for amino acids.

Since the amino acid frequencies are based on the number of codons, a form of nominal data, we chose to test for significant differences between genomes using nonparametric analysis. We used chi-square tests to look for differences in the distribution of the amino acid frequencies. A probability level ≤ 0.05 was considered significant.

Results

Using codon usage data, we calculated the frequencies of amino acids encoded by fourteen genomes that utilize the standard genetic code (Table 1). By chi-square analysis, no significant difference was observed in the distribution of the encoded amino acids between the DNA and RNA genomes, viral and nonviral genomes, nuclear and chloroplast genomes, eukaryotic and prokaryotic genomes, haploid and diploid genomes, haploid and tetraploid genomes, and diploid and tetraploid genomes ($P > 0.05$). Genome size and the complexity of the organism also did not significantly affect the frequencies of the encoded amino acids ($P > 0.05$).

We also calculated the frequencies of amino acids encoded by eight representative genomes that use nonstandard genetic codes (Table 2). Of the four mitochondrial genomes, the differences in the amino acid frequencies between *F. hepatica* and *N. crassa* ($\chi2 = 78.26$, df = 19, $P < 0.0001$), and also between *F. hepatica* and *H. sapiens* ($\chi2 = 49.99$, df = 19, $P = 0.0001$) were found to be significant. No significant differences were observed among the four non-mitochondrial

Table 1. Frequencies of amino acids (per 100 codons) encoded by genomes that utilize the standard genetic code.

Amino acid	HIV	Influenza virus	Variola virus	*A. thaliana* chloroplast	*M. pneumoniae*	*E. coli*	*P. salina*	*S. cerevisiae*	*D. discoideum*	*A. thaliana*	*C. elegans*	*D. melanogaster*	*X. laevis*	*H. sapiens*	Codons
Met	2.23	3.84	2.77	2.25	1.35	2.75	3.16	2.10	1.66	2.45	2.61	2.36	2.50	2.21	1
Trp	3.00	1.56	0.68	1.50	0.71	1.52	1.20	1.04	0.76	1.25	1.11	0.98	1.12	1.32	1
Phe	2.74	3.91	4.60	5.55	4.85	3.89	3.92	4.45	4.54	4.25	4.72	3.50	3.79	3.79	2
Tyr	2.66	2.39	5.30	3.39	3.16	2.87	3.24	3.36	3.42	2.83	3.12	2.92	3.02	2.74	2
His	2.52	1.50	2.07	2.10	1.60	2.29	0.98	2.14	1.79	2.25	2.33	2.70	2.52	2.59	2
Gln	5.03	3.81	2.27	3.38	5.97	4.45	3.16	3.94	5.13	3.46	4.17	5.19	4.63	4.64	2
Asn	5.38	4.99	6.97	4.61	6.10	3.94	3.24	6.05	10.52	4.32	4.84	4.73	4.29	3.60	2
Lys	5.69	5.55	7.46	5.80	6.93	4.37	6.63	7.27	7.27	6.35	6.34	5.63	6.49	5.62	2
Asp	3.95	4.84	6.56	4.17	5.17	5.11	5.19	5.78	5.13	5.38	5.28	5.22	5.29	4.70	2
Glu	6.74	7.76	5.28	5.74	5.48	5.69	4.14	6.48	5.77	6.65	6.53	6.36	7.09	6.84	2
Cys	2.27	1.70	2.12	1.23	0.48	1.17	1.74	1.29	1.44	1.77	2.03	1.85	2.13	2.31	2
Ile	6.39	6.63	9.34	7.92	4.95	5.99	7.46	6.50	8.09	5.26	6.05	4.88	5.02	4.42	3
Thr	5.95	6.25	6.24	5.21	6.70	5.43	7.22	5.87	6.06	5.12	5.82	5.63	5.38	5.31	4
Val	5.80	5.88	6.14	6.02	6.60	7.03	9.19	5.64	4.56	6.74	6.18	5.90	6.10	6.08	4
Pro	5.05	3.87	3.34	4.11	5.41	4.39	3.54	4.39	4.13	4.88	4.90	5.45	5.39	6.11	4
Ala	5.96	6.11	3.71	5.87	7.31	9.42	8.43	5.62	3.47	6.51	6.31	7.48	6.36	6.97	4
Gly	7.22	6.81	3.97	7.00	6.05	7.31	9.18	5.06	4.88	6.58	5.37	6.27	6.12	6.60	4
Ser	5.64	7.32	7.81	7.99	6.91	5.92	6.78	8.88	9.77	8.93	8.06	8.33	8.18	8.10	6
Leu	8.80	7.77	8.75	10.28	9.51	10.67	7.30	9.5	8.52	9.35	8.59	8.98	9.22	10.02	6
Arg	6.68	7.31	4.22	5.52	3.86	5.51	3.98	4.42	2.94	5.40	5.23	5.49	5.18	5.68	6

Table 2. Frequencies of amino acids (per 100 codons) encoded by genomes that use nonstandard genetic codes.

Amino acid	*N. crassa* mitochondrion	*F. hepatica* mitochondrion	*D. melanogaster* mitochondrion	*H. sapiens* mitochondrion	*M. capricolum*	*C. albicans*	*E. octocarinatus*	*T. thermophila*
Met	5.78	3.00	5.40	5.65	1.72	1.82	2.91	1.92
Trp	1.36	3.61	2.82	2.43	1.02	1.09	1.10	0.77
Phe	6.20	11.04	9.06	5.45	5.21	4.53	4.29	4.46
Tyr	4.29	4.65	4.27	3.52	4.06	3.58	3.94	3.73
His	1.95	1.61	2.39	2.14	1.30	2.02	1.69	1.52
Gln	2.77	0.79	1.85	2.33	3.76	4.24	3.63	6.93
Asn	5.38	2.28	5.08	4.49	8.35	6.07	4.88	7.15
Lys	6.14	1.26	2.20	2.67	10.79	6.73	8.68	9.21
Asp	3.99	2.11	1.92	1.79	5.63	5.71	7.05	5.48
Glu	4.71	2.18	1.98	2.44	6.05	6.17	7.33	6.93
Cys	0.76	3.30	0.89	0.59	0.59	1.12	2.05	1.82
Ile	5.21	4.93	10.57	8.52	9.99	6.85	7.06	7.37
Thr	9.88	2.46	5.51	9.28	5.30	6.67	5.22	5.16
Val	5.84	12.89	5.20	4.98	5.43	5.83	6.16	5.14
Pro	4.12	2.87	3.90	5.88	2.46	4.64	3.46	3.16
Ala	6.25	3.58	4.97	7.04	4.52	5.73	5.23	5.26
Gly	6.00	9.01	6.16	5.76	4.34	5.54	6.08	4.69
Ser	7.77	9.58	8.43	6.78	6.75	9.03	6.67	7.13
Leu	7.27	16.50	15.41	16.33	9.94	8.80	8.20	8.38
Arg	4.09	2.02	1.66	1.59	2.54	3.65	4.10	3.61

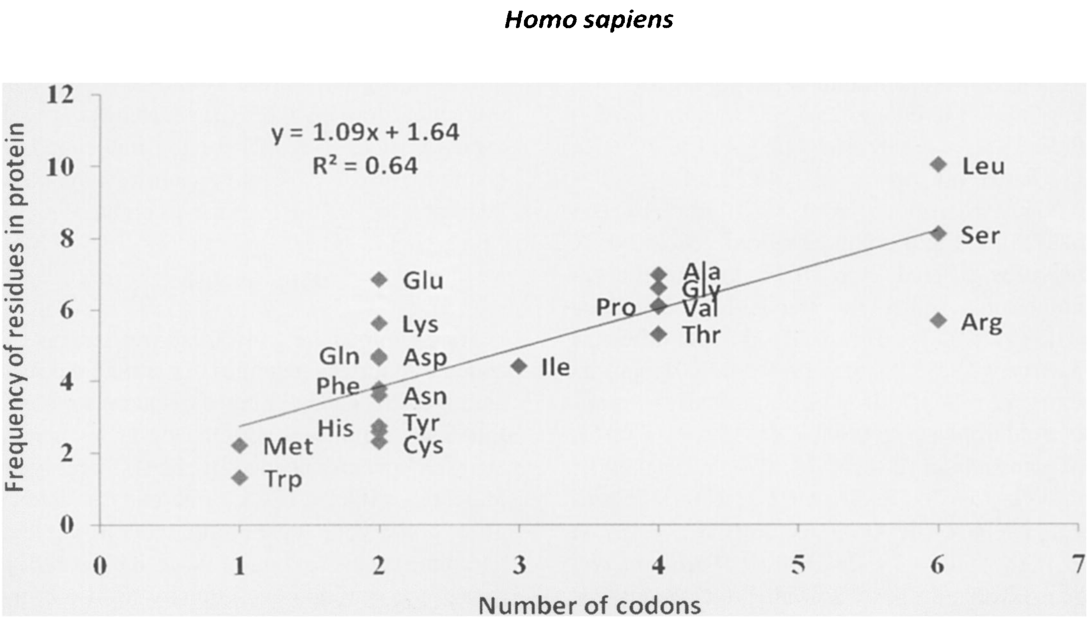

Figure 1. The frequency of amino acids in proteins as a function of the number of codons for that amino acid. Each diamond represents one of twenty amino acids encoded by the *H. sapiens* nuclear genome. The equation of the linear trend line and the coefficient of determination (R^2) are shown.

genomes (P > 0.05). However, the amino acid frequencies encoded by three of the mitochondrial genomes were found to be significantly different from those of the non-mitochondrial genomes. The distribution of amino acids encoded by the mitochondrial genome of *H. sapiens* was observed to be significantly different from those of the genomes of *M. capricolum* (χ2 = 39.10, df = 19, P = 0.0042), *E. octocarinatus* (χ2 = 33.17, df = 19, P = 0.0229), and *T. thermophila* (χ2 = 41.57, df = 19, P = 0.0020). Significant differences were also noted between the mitochondrial genome of *F. hepatica* and all four of the non-mitochondrial genomes (*M. capricolum*: χ2 = 139.08, df = 19, P < 0.0001; *C. albicans*: χ2 = 86.89, df = 19, P < 0.0001; *E. octocarinatus*: χ2 = 103.63, df = 19, P < 0.0001; *T. thermophila* (χ2 = 149.01, df = 19, P < 0.0001). The distribution of the amino acids encoded by the mitochondrial genome of *D. melanogaster* also significantly differed from those of the four of the non-mitochondrial genomes (*M. capricolum*: χ2 = 34.68, df = 19, P = 0.0152; *C. albicans*: χ2 = 33.15, df = 19, P = 0.0230; *E. octocarinatus*: χ2 = 34.69, df = 19, P = 0.0152; and *T. thermophila* (χ2 = 42.32, df = 19, P = 0.0016).

We then compared the amino acid frequencies encoded by the eight genomes that use nonstandard genetic codes to those of the fourteen genomes that employ the standard genetic code. We found that three of the mitochondrial genomes differed significantly from some or all of the genomes that utilize the standard genetic code. The mitochondrial genome of *H. sapiens*, for instance, differed significantly from half of the genomes that use the standard genetic code (HIV: χ2 = 30.61, df = 19, P = 0.0445; Variola virus: χ2 = 34.16, df = 19, P = 0.0176; *M. pneumonia*: χ2 = 37.60, df = 19, P = 0.0066; *P. salina*: χ2 = 33.10, df = 19, P = 0.0233; *D. discoideum*: χ2 = 40.96, df = 19, P = 0.0023; *D. melanogaster*: χ2 = 31.13, df = 19, P = 0.0390; *H. sapiens*: χ2 = 30.28, df = 19, P = 0.0473). On the other hand, the mitochondrial genome of *D. melanogaster* differed significantly from all but the *A. thaliana* chloroplast genome (HIV: χ2 = 43.02, df = 19, P = 0.0012; Influenza A virus: χ2 = 34.76, df = 19, P = 0.0149; Variola virus: χ2 = 32.96, df = 19, P = 0.0243; *A. thaliana* chloroplast: χ2 = 21.35, df = 19, P = 0.3177; *M. pneumonia*: χ2 = 47.04, df = 19, P = 0.0003; *E. coli*: χ2 = 30.81, df = 19, P = 0.0424; *P. salina*: χ2 = 37.67, df = 19, P = 0.0060; *S. cerevisiae*: $\chi2$ = 32.31, df = 19, P = 0.0288; *D. discoideum*: $\chi2$ = 40.29, df = 19, P = 0.0030; *A. thaliana*: $\chi2$ = 34.12, df = 19, P = 0.0177; *C. elegans*: $\chi2$ = 32.01, df = 19, P = 0.0311; *D. melanogaster*: $\chi2$ = 42.01, df = 19, P = 0.0017; *X. laevis*: $\chi2$ = 46.39, df = 19, P = 0.0050; *H. sapiens*: $\chi2$ = 40.70, df = 19, P = 0.0026), while the mitochondrial genome of *F. hepatica* differed from all fourteen of the genomes that utilize the standard genetic code (HIV: $\chi2$ = 90.76, df = 19, P < 0.0001; Influenza A virus: $\chi2$ = 86.99, df = 19, P < 0.0001; Variola virus: $\chi2$ = 86.07, df = 19, P < 0.0001; *A. thaliana* chloroplast: $\chi2$ = 60.67, df = 19, P = 0.0022; *M. pneumonia*: $\chi2$ = 107.78, df = 19, P < 0.0001; *E. coli*: $\chi2$ = 70.58, df = 19, P = 0.0001; *P. salina*: $\chi2$ = 70.43, df = 19, P = 0.0001; *S. cerevisiae*: $\chi2$ = 88.99, df = 19, P < 0.0001; *D. discoideum*: $\chi2$ = 120.00, df = 19, P < 0.0001; *A. thaliana*: $\chi2$ = 72.23, df = 19, P = 0.0002; *C. elegans*: $\chi2$ = 80.43, df = 19, P < 0.0001; *D. melanogaster*: $\chi2$ = 88.94, df = 19, P < 0.0001; *X. laevis*: $\chi2$ = 86.76, df = 19, P < 0.0001; *H. sapiens*: $\chi2$ = 79.72, df = 19, P < 0.0001).

To examine the relationship between the number of codons for a given amino acid and the frequency of that amino acid in proteins, we constructed graphs for the *H. sapiens* nuclear genome (Figure 1), as well as for the twenty-one other genomes that were studied (not shown). In general, there is a tendency for amino acids that are represented by more codons to be more frequent in proteins, although the strength of this relationship, as evidenced by the slopes of the linear trend lines and the coefficients of determination (R^2), varied considerably between genomes. Among the fourteen genomes that use the standard genetic code, the slopes of the linear trend lines ranged from a high of 1.09 (*H. sapiens*) to a low of 0.64 (Variola virus) with an average of 0.91 ± 0.04 (mean ± standard error), while the coefficients of determination varied from 0.64 (*H. sapiens*) to 0.14 (*D. discoideum*) with an average of 0.45 ± 0.04. Of the eight genomes that employ nonstandard codes, the slopes of the linear trend lines ranged from 1.72 (*H. sapiens* mitochondrion) to 0.43 (*M. capricolum*) with averages of 1.25 ± 0.23 for the mitochondrial genomes, 0.58 ± 0.09 for the non-mitochondrial genomes, and 0.91 ± 0.13 for all eight genomes. The coefficients of determination for this group varied from 0.42 (*H. sapiens* mitochondrion) to 0.05 (*M. capricolum*) with averages of 0.33 ± 0.03 for the mitochondrial genomes, 0.17 ± 0.07 for the non-mitochondrial genomes, and 0.25 ± 0.05 for all eight.

Discussion

We examined the amino acid frequencies encoded by fourteen genomes that utilize the standard genetic code. Using chi-square tests, we found no significant differences in the amino acid frequencies encoded by the different types of genomes (DNA, RNA, viral, nuclear, chloroplast, prokaryotic, eukaryotic, haploid, diploid, tetraploid). This finding is most likely a reflection of the common evolutionary history of the fourteen genomes. If modern life had more than one origin, then we would expect two or more very different types of genetic codes, since there is no obvious biological or chemical reason why any particular codon should code for any given amino acid. (Brown, 2007). Instead, the genetic code is virtually the same in all organisms, strongly suggesting that the various genomes shared a common past. The genetic code probably evolved very early in the history of modern life, since the translation of nucleotide sequences into amino acids is a vital process required for the growth and replication of all living cells, organelles, and viruses. The essential nature of translation would ensure that the genetic code would be inherited intact by all generations as well as maintained during speciation. Since the amino acid-coding abilities of the fourteen genomes are extremely well conserved, there may be little tolerance for change to the existing standard genetic code.

Although the standard genetic code is well conserved, alternative genetic codes, nevertheless, do exist. We examined the amino acid frequencies encoded by genomes that employ eight different nonstandard genetic codes and found significant differences between the three animal mitochondrial genomes (*F. hepatica*, *D. melanogaster*, and *H. sapiens*) and the remaining nineteen genomes. This difference may be due largely in part to the close proximity of the mitochondrial genome to the electron transport chain of cellular respiration. Due to faulty electron transfer, highly-reactive oxygen free radicals are often generated, which can damage the mitochondrial genome (Allen and Raven, 1996). Since mitochondria lack many of the repair systems used to correct damaged DNA, mitochondria can accumulate mutations at a higher rate than their nuclear counterparts (Krishnan, et al., 2007). Some of these mitochondrial DNA mutations appear to have altered the codon assignments and provided variations in the genetic code that were selected for during the course of evolution. In human mitochondria, for example, the genetic code has been simplified by replacing two codons that had different designations (UGG = tryptophan and UGA = stop) with a pair that has the same meaning (UGG & UGA = tryptophan) (Lewin, 2008). Mitochondria may also be more tolerant of mutations that affect their genetic code than the nucleus because animal mitochondrial genomes only code for thirteen proteins (Anderson et al, 1981). Since most mitochondrial proteins are encoded by nuclear genes, changes in the mitochondrial genetic codes would only affect a small fraction of all mitochondrial proteins. Those codons that are reassigned would most likely involve substituting one amino acid for a similar amino acid, since altered protein and mitochondrial function can be lethal.

Of the four mitochondrial genomes, we observed differences between the fungal (*N. crassa*) mitochondrial genome and the three animal mitochondrial genomes. This finding may be related to the differences in the life spans between fungi and animals. The mitochondrial genomes of long-lived species, like animals, are subject to the mutagenic effects of oxygen free radicals for longer periods of time than short-lived species, such as fungi. This may explain the finding that mammals have accumulated more changes in their mitochondrial DNA than yeast (Clark-Walker, 1991). Animals and fungi also differ in the size and organization of their mitochondrial genomes. Animal mitochondrial genomes are relatively small, ranging in size from 16 to 17 kb, and lack introns. In contrast, fungal mitochondrial genomes are much larger, ranging from 19 to 100 kb, and contain large genes interrupted with introns (Lewin, 2008). While mutations to animal mitochondrial genomes would alter exons, random mutations to the fungal mitochondrial genome would be distributed among exons and introns. Since introns do not code for amino acids and are removed from mRNA prior to translation, these sequences may help to absorb mutations that would otherwise affect coding regions. As a result, fungal mitochondria would then incur fewer changes to their genetic codes than animal mitochondria.

We also examined the relationship between the number of codons for a given amino acid and the frequency of that amino acid in proteins. By constructing graphs for each of the twenty-two genomes that were studied, we observed a tendency for amino acids that are represented by more codons to be more frequent in proteins. This trend may be a reflection of an organism's amino acid requirements (Swire et al., 2005). While some of the codon assignments may be the result of mutation and genetic drift, most are probably the results of natural selection for specific amino acids. Mutations that alter codon assignments may be favored by natural selection if those changes somehow enhanced fitness. For instance, an organism may benefit if a codon is reassigned so that it codes for an amino acid that is limited in supply, is energetically less expensive to synthesize, or can be used to build more efficient protein structures. If there are many of these kinds of reassignments, then more useful amino acids will accumulate more codons, and a trend will develop between the usage of an amino acid and its occurrence in the genetic code (Swire et al., 2005).

We also observed that the relationship between the number of codons for a particular amino acid and the amino acid frequency was stronger, as indicated by the coefficients of determination, among some genomes than others. On average, we found that the genomes that utilize the standard genetic code have larger coefficients of determination than those that use non-standard codes. This finding is consistent with the hypothesis that the nonstandard genetic codes evolved from the standard code, since codon reassignments alter the numbers of codons for each amino acid without a proportional change in the amino acid frequency. Interestingly, we observed that more complex organisms tend to have higher coefficients of determinations than less developed species. This suggests that the amino acid requirements of organisms may vary depending on their relative complexity, perhaps influenced by the types of proteins encoded by each genome.

Although the genetic code was once thought to be frozen in time, evidence collected over the last thirty years strongly suggests that the genetic code is in a state of evolution. We studied the amino acid frequencies encoded by various genomes and interpreted them as a record of evolutionary change. Our study suggests that the standard genetic code formed early in the development of modern life and may have evolved to reflect different amino acid requirements. Our findings also support the hypothesis that the nonstandard genetic codes evolved from the standard code.

Literature Cited

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D., Scheeler, F., Shen, H., Shue, B.C., Sidén-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., Woodage, T., Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., and J.C. Venter. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**:2185–2195.

Allen, J.F. and J.A. Raven. 1996. Free-radical-induced mutation vs. redox regulation: Costs and benefits of genes in organelles. *Journal of Molecular Evolution* **42:** 482–492.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R., and I.G. Young. 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290:**457–465.

Barrell, B.G., Bankier, A.T., and J. Drouin. 1979. A different genetic code in human mitochondria. *Nature* **282:**189–194.

Bernfield, M.R. and M.W. Nirenberg. 1965. The nucleotide sequences of multiple codewords for phenylalanine, serine, leucine, and proline. *Science* **147:**479–484.

Bessho, Y., Ohama, T., and S. Osawa. 1992. Planarian mitochondria. II. The unique genetic code as deduced from cytochrome c oxidase subunit I gene sequences. *Journal of Molecular Evolution* **34:**331–335.

Brimacombe, R., Trupin, J., Nirenberg, M., Leder, P., Bernfield, M., and T. Jaouni. 1965. Nucleotide sequences of synonym codons for arginine, valine, cysteine, and alanine. *Proceedings of the National Academy of Sciences USA* **54:**954–960.

Brook, D.J. and J.R. Fresco. 2002. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. *Molecular & Cellular Proteomics* **1:**125–131.

Brooks, D.J., Fresco, J.R., Lesk, A.M., and M. Singh. 2002. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. *Molecular Biology and Evolution* **19:**1645–1655.

Brown, T.A. 2007. Genomes 3. Garland Science, New York.

Caskey, C.T., Tompkins, R., Scolnick, E., Caryk, T., and M. Nirenberg. 1968. Sequential translation of trinucleotide codons for the initiation and termination of protein synthesis. *Science* **162:**135–138.

Clark-Walker, G.D. 1991. Contrasting mutation rates in mitochondrial and nuclear genes of yeast versus animals. *Current Genetics* **20:**195–198.

Crick, F. 1990. What mad pursuit: a personal view of scientific discovery. Basic Books, New York.

Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M.A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., Tunggal, B., Kummerfeld, S., Madera, M., Konfortov, B.A., Rivero, F., Bankier, A.T., Lehmann, R., Hamlin, N., Davies, R., Gaudet, .P, Fey, P., Pilcher, K., Chen, G., Saunders, D., Sodergren, E., Davis, P., Kerhornou, A., Nie, X., Hall, N., Anjard, C., Hemphill, L., Bason, N., Farbrother, P., Desany, B., Just, E., Morio, T., Rost, R., Churcher, C., Cooper, J., Haydock, S., van Driessche, N., Cronin, A., Goodhead, I., Muzny, D., Mourier, T., Pain, A., Lu, M., Harper, D., Lindsay, R., Hauser, H., James, K., Quiles, M., Madan Babu, M., Saito, T., Buchrieser, C., Wardroper, A., Felder, M., Thangavelu, M., Johnson, D., Knights, A., Loulseged, H., Mungall, K., Oliver, K., Price, C., Quail, M.A., Urushihara, H., Hernandez, J., Rabbinowitsch, E., Steffen, D., Sanders, M., Ma, J., Kohara, Y., Sharp, S., Simmonds, M., Spiegler, S., Tivey, A., Sugano, S., White, B., Walker, D., Woodward, J., Winckler, T., Tanaka, Y., Shaulsky, G., Schleicher, M., Weinstock, G., Rosenthal, A., Cox, E.C., Chisholm, R.L., Gibbs, R., Loomis, W.F., Platzer, M., Kay, R.R., Williams, J., Dear, P.H., Noegel, A.A., Barrell, B., and A. Kuspa. 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435**:43–57.

Eschbach, S., Hoffmann, C. J., Maier, U.G., Sitte, P., and P. Hansmann. 1991. A eukaryotic genome of 660 kb: electrophoretic karyotype of nucleomorph and cell nucleus of the cryptomonad alga, *Pyrenomonas salina*. *Nucleic Acid Research* **19**:1779–1781.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and S.G. Oliver. 1996. Life with 6000 genes. *Science* **274**:563–567.

Himmelreich, R. Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acid Research* **24**:4420–4449.

Horowitz, S. and M.A. Gorovsky. 1985. An unusual genetic code in nuclear genes of *Tetrahymena*. *Proceedings of the National Academy of Sciences USA* **82**:2452–2455.

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

Jordan, I.K., Kondrashov, F.A., Adzhubei, I.A., Wolf, Y.I., Koonin, E.V., Kondrashov, A.S., and S. Sunyaev. 2005. A universal trend of amino acid gain and loss in protein evolution. *Nature* **433**:633–638.

King, J.L. and T.H. Jukes. 1969. Non-Darwinian evolution. *Science* **164**:788–798.

Klapper, M.H. 1977. The independent distribution of amino acid near neighbor pairs into polypeptides. *Biochemical and Biophysical Research Communications* **78**:1018–1024.

Krishnan, K.J., Greaves, L.C., Reeves, A.K., and D. Turnbull. 2007. The ageing mitochondrial genome. *Nucleic Acid Research* **35**:7399–7405.

Leder, P. and M. Nirenberg. 1964a. Nucleotide sequence of a valine RNA codeword. *Proceedings of the National Academy of Sciences USA* **52**:420–427.

Leder, P. and M. Nirenberg. 1964b. On the nucleotide sequence of a cysteine and a leucine RNA codeword. *Proceedings of the National Academy of Sciences USA* **52**:1521–1529.

Lewin, B. 2008. Genes IX. Jones and Bartlett Publishers, Sudbury, MA.

Morin, R.D., Chang, E., Petrescu, A., Liao, N., Griffith, M., Kirkpatrick, R., Butterfield, Y.S., Young, A.C., Stott, J., Barber, S., Babakaiff, R., Dickson, M.C., Matsuo, C., Wong, D., Yang, G.S., Smailus, D.E., Wetherby, K.D., Kwong, P.N., Grimwood, J., Brinkley 3rd, C.P., Brown-John, M., Reddix-Dugue, N.D., Mayo, M., Schmutz, J., Beland, J., Park, M., Gibson, S., Olson, T., Bouffard, G.G., Tsai, M., Featherstone, R., Chand, S., Siddiqui, A.S., Jang, W., Lee, E., Klein, S.L., Blakesley, R.W., Zeeberg, B.R., Narasimhan, S., Weinstein, J.N., Pennacchio, C.P., Myers, R.M., Green, E.D., Wagner, L., Gerhard, D.S., Marra, M.A., Jones, S., and R.A. Holt. 2006. Sequencing and analysis of 10, 967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling. *Genome Research* **16**:796–803.

Nakamura, Y., Gojobori, T., and T. Ikemura. 2000. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. *Nucleic Acid Research* **28**:292.

Nirenberg, M.W. and P. Leder. 1964. The effect of trinucleotides upon the binding of sRNA to ribosomes. *Science* **145**:1399–1407.

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F., and C. O'Neal. 1965. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences USA* **53**:1161–1168.

Osawa, S., Jukes, T.H., Watanabe, K., and A. Muto. 1992. Recent evidence for evolution of the genetic code. *Microbiological Reviews* **56**:229–264.

Santos, M.A.S. and M.F. Tuite. 1995. The CUG codon is decoded *in vivo* as serine and not leucine in *Candida albicans*. *Nucleic Acid Research* **23**:1481–1486.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and S. Tabata. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research* **6**:283–290.

Swire, J., Judson, O.P., and A. Burt. 2005. Mitochondrial genetic codes evolve to match amino acid requirements of proteins. *Journal of Molecular Evolution* **60**:128–139.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**:796–815.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**:2012–2018.

Tortora, G.J., Funke, B.R., and C.L. Case. 2007. Microbiology: an introduction, 9th edition. Pearson Benjamin Cummings, New York.

Trupin, J.S., Rottman, R.M., Brimacombe R.L., Leder, P., Bernfield, M.R., and M.W. Nirenberg. 1965. On the nucleotide sequences of degenerate codeword sets for isoleucine, tyrosine, asparagine, and lysine. *Proceedings of the National Academy of Sciences USA* **53**:807–811.

Welch, R.A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S.-R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G.F., Rose, D.J., Zhou, S., Schwartz, D.C., Perna, N.T., Mobley, H.L.T., Donnenberg, M.S., and F.R. Blattner. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences USA* **99**:17020–17024.

Yamao, F., Muto, A., Kawauchi, Y., Iwami, M., Iwagami, S., Azumi, Y., and S. Osawa. 1985. UGA is read as tryptophan in *Mycoplasma capricolum*. *Proceedings of the National Academy of Sciences USA* **82**:2306–2309.

Received 16 January 2009; accepted 3 April 2009.