

Intimate Evolution of Proteins

PROTEOME ATOMIC CONTENT CORRELATES WITH GENOME BASE COMPOSITION*

Received for publication, June 17, 2003, and in revised form, November 12, 2003
Published, JBC Papers in Press, November 29, 2003, DOI 10.1074/jbc.M306415200

Peggy Baudouin-Cornu^{‡§}, Katja Schuerer[¶], Philippe Marlière^{||}, and Dominique Thomas^{‡**}

From the [‡]Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique, 91 198 Gif sur Yvette Cedex, France, the [¶]Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France, and ^{||}Evologic SA, 89 rue Henri Rochefort, 91000 Evry, France

Discerning the significant relations that exist within and among genome sequences is a major step toward the modeling of biopolymer evolution. Here we report the systematic analysis of the atomic composition of proteins encoded by organisms representative of each kingdoms. Protein atomic contents are shown to vary largely among species, the larger variations being observed for the main architectural component of proteins, the carbon atom. These variations apply to the bulk proteins as well as to subsets of ortholog proteins. A pronounced correlation between proteome carbon content and genome base composition is further evidenced, with high G+C genome content being related to low protein carbon content. The generation of random proteomes and the examination of the canonical genetic code provide arguments for the hypothesis that natural selection might have driven genome base composition.

Comparative analyses of complete genome sequences were anticipated to reveal the molecular bases of biodiversity as well as to increase our comprehension of the constraints that shaped protein composition and structure during the natural history of living organisms. However, whereas comparative genomics highlighted genome structure plasticity and strengthened the role of lateral gene transfer during the natural history of living organisms (1, 2), less progress was accomplished in understanding adaptive evolution at a molecular level and how it contributes to changes in proteins.

Studies devoted to protein evolution mostly use comparative analyses of protein primary sequences and attempt to identify which rules govern the conservations, substitutions, and deletions of amino acids that are observed between proteins (3–5). By focusing on amino acid composition, such studies do not address the possibility that evolution of proteins, and more generally of biopolymers, might have been shaped at a more intimate level: the atomic level. It is worthwhile to note that, among the constitutive elements of proteins, carbon, nitrogen, and sulfur atoms are subject to geochemical cycles at the surface of the earth (6). As a consequence, large fluctuations of both form and abundance of elemental components of proteins are occurring in natural habitats. One would thus expect that specific molecular mechanisms might have evolved to allow

living organisms to respond to the elemental variations of the environment that they inevitably faced during their natural history. These adaptive processes undoubtedly left traces in the chemical composition of biopolymers, and it could be anticipated that the advent of complete genomic sequences, combined with the use of straightforward statistical methods, would open the way to retrieve such imprints in macromolecule sequences.

To approach this question, we previously investigated the atomic composition of protein subsets from both the bacteria *Escherichia coli* and the eukaryote *Saccharomyces cerevisiae* (7). This study allowed us to discern that, in both organisms, the atomic composition of several protein families has been constrained in response to specific selective pressures. A pronounced correlation between atomic compositions and metabolic roles were deciphered in enzymes involved in essential nutrient assimilation in these microbial organisms. For instance, carbon assimilatory enzymes were found to contain significantly less carbon atoms than the total proteins of *E. coli* and *S. cerevisiae* (7). Also, sulfur assimilatory enzymes of both organisms were shown to comprise less sulfur atoms than the bulk proteins. Such impoverishments of sulfur and carbon assimilatory enzymes in their respective element were interpreted as an imprint of variations in the nutritional availability of these elements during the natural history of *E. coli* and *S. cerevisiae*. These results thus suggested that the evolution of proteins could have been subjected to ecological constraints more than previously thought. In the next step, we wanted to know how broad the fluctuations of elemental protein composition might be, and more specifically, whether the atomic contents of proteins could display substantial variations among different organisms.

Here we report the first results of the systematic analysis of the atomic contents of the proteins encoded by organisms whose genome has been entirely sequenced. Complete protein data sets from 46 organisms, representative of the bacteria, archaea, and eukaryote kingdoms were compiled and analyzed using quantile distributions. The results show that the atomic composition of proteins, especially their carbon content, widely differs among species and may differ more between species than within proteomes. A strongly significant correlation between the average carbon content of proteomes and the DNA base composition of the genomes was moreover uncovered. Randomization methods and analysis of the genetic code further demonstrate that the extant genetic code relates DNA G+C usage with the carbon content of proteins.

EXPERIMENTAL PROCEDURES

Organisms Used for This Study—Five eukaryotic, ten archaeal, and thirty-one bacterial proteomes were downloaded from the GenBank™ FTP site (<ftp.ncbi.nih.gov/genbank/genomes/>). All of the proteins containing one or more undetermined amino acid (“X”) were discarded,

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ Supported by a grant from the Ministère de la Défense and the Leukemia and Lymphoma Society of America.

** To whom correspondence should be addressed: Centre de Génétique Moléculaire, CNRS, 91 198 Gif-sur-Yvette, France. Tel.: 33-1-69-82-32-33; Fax: 33-1-69-82-43-72; E-mail: thomas@cgm.cnrs-gif.fr.

except for *Arabidopsis thaliana* proteins in which only the X amino acids were discarded. The coding G+C contents were retrieved from Kazusa's Codon Usage Data Base (www.kazusa.or.jp/codon/), which is an extended WWW version of CUTG (Codon Usage Tabulated from GenBank™). The prokaryotic total G+C contents were found in the literature. The considered organisms are the following: *A. thaliana* (25,531 proteins; G+C coding = 44.17%), *Caenorhabditis elegans* (17,074 proteins; G+C coding = 42.59%), *Drosophila melanogaster* (14,335 proteins; G+C coding = 53.99%), *Homo sapiens* (24,493 proteins; G+C coding = 52.51%), *S. cerevisiae* (6,330 proteins; G+C coding = 39.72%), *Aeropyrum pernix* (2,694 proteins; G+C coding = 57.49%; G+C total = 56.3%), *Archaeoglobus fulgidus* (2,420 proteins; G+C coding = 49.37%; G+C total = 48.5%), *Halobacterium sp.* (2,058 proteins; G+C coding = 65.26%; G+C total = 65.9%), *Methanococcus jannaschii* (1,773 proteins; G+C coding = 31.94%; G+C total = 31.09%), *Methanobacterium thermoautotrophicum* (1,869 proteins; G+C total = 49.50%), *Pyrococcus abyssi* (1,765 proteins; G+C coding = 45.16%; G+C total = 44.7%), *Pyrococcus horikoshii* (2,038 proteins; G+C coding = 42.32%; G+C total = 42.0%), *Sulfolobus solfataricus* (2,977 proteins; G+C coding = 36.30%), *Thermoplasma volcanium* (1,499 proteins), *Thermoplasma acidophilum* (1,478 proteins; G+C coding = 47.38%; G+C total = 46%), *Aquifex aeolicus* (1,521 proteins; G+C coding = 53.58%; G+C total = 43.4%), *Bacillus halodurans* (4,066 proteins; G+C coding = 44.33%; G+C total = 43.7%), *Bacillus subtilis* (4,100 proteins; G+C coding = 44.35%; G+C total = 43.5%), *Borrelia burgdorferi* (1,606 proteins; G+C coding = 29.33%; G+C total = 28.6%), *Campylobacter jejuni* (1,654 proteins; G+C coding = 30.98%; G+C total = 30.6%), *Caulobacter crescentus* (3,767 proteins; G+C coding = 67.42%; G+C total = 67.2%), *Chlamidia muridarum* (898 proteins; G+C coding = 40.64%; G+C total = 40.3%), *Chlamidophila pneumoniae* (1,092 proteins; G+C coding = 41.29%; G+C total = 40.6%), *Deinococcus radiodurans* (3,089 proteins; G+C coding = 67.24%; G+C total = 66.6%), *E. coli* (4,289 proteins; G+C coding = 51.11%; G+C total = 50.8%), *Haemophilus influenzae* (1,666 proteins; G+C coding = 37.52%; G+C total = 38.1%), *Helicobacter pylori* (1,565 proteins; G+C coding = 40.45%; G+C total = 39%), *Lactococcus lactis* (2,266 proteins; G+C coding = 35.69%), *Mesorhizobium loti* (6,752 proteins; G+C coding = 57.29%; G+C total = 62.47%), *Mycobacterium leprae* (1,605 proteins; G+C coding = 59.63%; G+C total = 57.8%), *Mycobacterium tuberculosis* (4,137 proteins; G+C coding = 65.79%; G+C total = 65.6%), *Mycoplasma genitalium* (489 proteins; G+C coding = 31.74%), *Mycoplasma pneumoniae* (689 proteins; G+C coding = 41.06%; G+C total = 40.0%), *Mycoplasma pulmonis* (782 proteins; G+C coding = 26.38%), *Neisseria meningitidis* (2,081 proteins; G+C coding = 51.49%; G+C total = 51.5%), *Pasteurella meningitidis* (2,014 proteins; G+C coding = 37.81%; G+C total = 40.4%), *Pseudomonas aeruginosa* (5,565 proteins; G+C coding = 66.63%; G+C total = 66.6%), *Rickettsia prowazekii* (834 proteins; G+C coding = 30.60%; G+C total = 29.1%), *Staphylococcus aureus* (2,594 proteins; G+C coding = 32.90%), *Streptococcus pyogenes* (1,696 proteins; G+C coding = 40.99%; G+C total = 38.5%), *Synechocystis sp.* (3,168 proteins; G+C coding = 49.52%; G+C total = 49.5%), *Thermotoga maritima* (1,849 proteins; G+C coding = 46.45%; G+C total = 46.2%), *Treponema pallidum* (1,003 proteins; G+C coding = 52.52%; G+C total = 52.8%), *Ureaplasma urealyticum* (613 proteins; G+C coding = 26.55%; G+C total = 25.5%), *Vibrio cholerae* (3,822 proteins; G+C coding = 47.62%; G+C total = 47.48%), and *Xylella fastidiosa* (2,766 proteins; G+C coding = 53.78%; G+C total = 52.7%).

Orthologous Proteins—Sequences of proteins comprised within 25 different clusters of orthologous proteins (COGs)¹ spanning 43 genomes were retrieved (www.ncbi.nlm.nih.gov/COG/). The 25 used COGs consist of COGs corresponding to 11 ribosomal proteins (COG0522, COG0099, COG0081, COG0088, COG1841, COG1358, COG2075, COG2147, COG1890, COG1471, and COG2058) and 14 COGs corresponding to other abundant proteins (COG0495, COG0149, COG0126, COG0057, COG0167, COG0469, COG0492, COG0192, COG0191, COG0055, COG0450, COG1155, COG0174, and COG1156). Proteins belonging to these 25 COGs were clustered according to species (bacteria, 16 COGs; archaea, 20 COGs) or according to COGs, the corresponding protein carbon usages were calculated, and the resulting quantile distributions were displayed as for whole proteomes.

Random Proteomes—Random protein samples were first produced using the natural codon catalog from random DNAs displaying different G+C contents. For each G+C content value, five different protein samples were generated, and their resulting carbon quantile distribu-

tions were calculated. Each protein sample is composed of 2000 random proteins, the lengths of which are distributed among nine values ranging from 100 to 600 residues to mimic the distributions of real protein lengths. The codons were sorted out with the unique constraints of having $P_A = P_T$ and $P_G = P_C$, where P_X is the probability of sorting the base X out of the set {A, C, T, G}, and P_G is chosen to raise the wanted final G+C content. The stop codons were systematically discarded.

To analyze random protein samples produced with randomly generated genetic codes, 500 plausible alternative codes were first generated following the criteria of Haig and Hurst (8), which are: (i) to assign one of the 20 amino acids to each cluster of codons coding for one amino acid in real life and (ii) to keep the stop codons. From each randomly generated code, six randomly generated genomes coding for 2000 proteins of 280 residues were computed as described above and raised final G+C contents of 28.7, 37.5, 44.4, 50.9, 59.6, and 65.6%, respectively. The correlation factors r between those six G+C content values and the resulting carbon mean values were calculated for each code. The quantile distribution of the slope of the corresponding regression lines were plotted for all of the codes leading to a squared correlation factor of >0.9 .

Computerized DNA Mutations—Randomization essays were performed starting from the DNA sequence of an *E. coli* gene coding for a protein of length L . Repeatedly, one position of the DNA sequence was randomly chosen, and the corresponding base was proposed to be mutated into one randomly chosen different base. Every position in the DNA sequence had the same probability to be chosen. Similarly, replacement with each of the three remaining base was equiprobable. Only mutations that would strictly decrease the carbon content of the encoded protein were kept. The number of effective mutations, the GC value of the resulting DNA sequence, and the carbon content of the corresponding protein were checked every L fold proposals of mutation.

RESULTS

Large Variations of Protein Atomic Contents in Firmicute Species—To analyze the atomic content of proteins encoded by entirely sequenced genomes, we used quantile distribution methods and stochastic ordering concepts (9, 10). For each organism, the quantile distributions of carbon, nitrogen, and sulfur atom frequencies within the residue side chains of the encoded proteins were calculated. The quantile $Q_A^S(x)$ is the fraction of proteins from the specie (S) in which the averaged number of atoms (A) per residue side chain is at most x . The medians (the $Q_A^{-1}(0.50)$ point) and the 80% quantile range (corresponding to the 0.10–0.90 quantile levels) are major statistical measurement of atomic contents (10).

As a first assessment of how the atomic composition of proteins fluctuates on a proteome-wide level, we chose the phylogenetically related group of firmicutes, a bacteria family that comprises the well known model organism *B. subtilis* and provides a large set of entirely sequenced genomes (11). The results of the analysis made with 11 different firmicute bacteria are depicted in Fig. 1. As previously noted for *E. coli* and *S. cerevisiae* (7), the quantile distributions indicate that, for each organism, the carbon, nitrogen, and sulfur contents of proteins follow bell-shaped distributions that are nearly Gaussian. Importantly, the results show that, depending on the atom considered, the atomic quantile distributions are differently scattered among species. Indeed, whereas quantile distributions for carbon, and to a lesser extent for sulfur, display substantial variations between species, the nitrogen quantile distributions are largely similar in all of the analyzed bacterial species. The statistical significance of the differences observed between carbon (or sulfur) distributions can be assessed by using normal z tests. A two-sided p value ($<10^{-9}$) confirms that, for instance, the differences observed between the *B. halodurans* and *S. aureus* carbon distributions are significant and likely not to occur by chance. A simple hypothesis accounting for the results reported in Fig. 1 would have been that the scattering of the quantile distributions seen among species reflects the variation of the number of the considered atom within the 20 canonical amino acids. The clustering of the

¹ The abbreviation used is: COG, cluster of orthologous proteins.

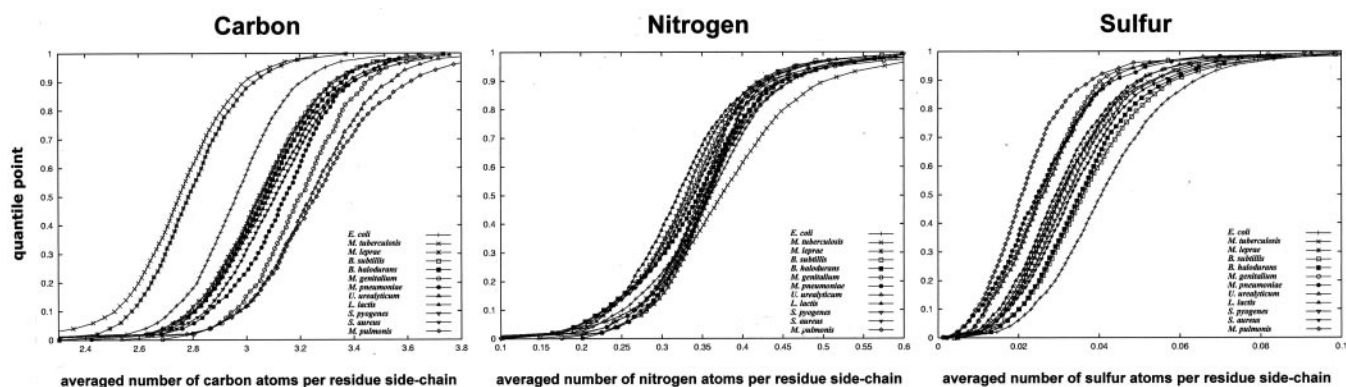


FIG. 1. Quantile representations of protein atomic content in the firmicute family of bacteria. For each proteome, the averaged number of carbon, nitrogen, or sulfur atoms found in residue side chains for each protein was calculated, and the totality of all these frequencies was described by a histogram. The quantile distributions are the cumulative representations of these histograms. For each protein sample, the quantiles were calculated so as to display the distribution by a 50-dot graph. The quantiles for *E. coli* were included as reference in each graph.

nitrogen, but not of the sulfur quantile distribution, rules out this hypothesis. Indeed, sulfur protein contents broadly vary between firmicute species, whereas amino acids contain at most one sulfur atom. In contrast, nitrogen protein contents are clustered, although either 0, 1, 2, or 3 nitrogen atoms are found in the side chains of canonical amino acids.

Altogether, the results displayed in Fig. 1 suggest that, at the atomic level, the composition of the proteins encoded by each genome broadly varies from one organism to another and overemphasize at a proteome-wide level the plasticity of protein atomic contents previously established for cyanobacterial light-harvesting proteins (12) and microbial assimilatory enzymes (7, 13).

Protein Carbon Contents Differ More between Species than within Each Proteome—Because the largest variations in protein atomic content were observed for carbon, which is the main architectural component of proteins, we focused our analysis on this atom (sulfur variation analyses will be reported elsewhere). Fig. 2 shows that the variations in protein carbon content observed for the firmicutes also hold for other bacterial species as well as for the archaea and eukaryote kingdoms. Remarkably, the carbon distributions for all kingdoms appear to be strictly ordered, with no (bacteria and archaea) or only few (eukaryotes) crosses between the plots, even at the outliers of the distributions. This indicates that although the mean values of protein carbon contents largely differ between species, the spread of each distribution is equivalent for all of the species analyzed. Because the overall dispersal of the carbon quantile distributions is larger than the spread of each quantile distribution (Fig. 2), the carbon variance of each species is smaller than would be the variance of a species composed of all the proteins of all organisms (data not shown). Therefore, the atomic composition of proteins appears to be more different between species than within each organism. The stochastic ordering of the carbon quantile distributions observed for both archaeal and bacterial species further demonstrates that at each level of protein carbon content, the quantile points are similarly ranked between species. Thus, whatever the origin of the observed variations in protein carbon content was, their cause seems to have changed uniformly the totality of the proteins encoded by each organism.

To further assess this point, the carbon contents of several ortholog proteins shared by archaeal and bacterial species were analyzed. Sequences of proteins comprised within different COGs (14) were retrieved from the COG data base, and the corresponding carbon contents were calculated and analyzed. Calculations made with 27 randomly chosen COGs first showed that the carbon contents of ortholog proteins vary among spe-

cies as the carbon content of whole proteomes (data not shown). Although only few data are available on the variations of protein abundance at the level of whole organisms, we tried to take into account the possible influence of protein abundance variations in our calculations. We used the recent study of Ghaemmaghami *et al.* (15), which reported the comprehensive analysis of the abundance of most proteins of *S. cerevisiae*, and we assumed that orthologs of abundant yeast proteins are probably abundant in their corresponding species. 25 COGs were selected as comprising one of the 75 most abundant proteins of *S. cerevisiae* and having representative in several bacteria or archaea species. The yeast proteins belonging to these COGs account for up to 17% of the expressed proteins in *S. cerevisiae*. Quantile distributions of carbon usage of proteins belonging to these 25 COGs were calculated for each species. The results (Fig. 3A) show that the carbon quantile distributions of abundant ortholog proteins are dispersed among species as the carbon quantile distributions calculated for all the proteins of each species. In addition, the quantile distributions of ortholog and total proteins are similarly ordered among species. Next, the carbon content of abundant ortholog proteins was calculated as gathered by COGs. Carbon quantile distributions were calculated for each of the 11 COGs having representatives in almost all of the 43 studied species and compared with the quantile distribution of the averaged carbon content of whole proteomes. As depicted in Fig. 3B, the slopes of the quantile distributions of carbon contents of ortholog proteins gathered by COGs are similar to the slope of the quantile distribution of the median carbon values ($Q_C^{-1}(0.5)$) of total proteins. This shows that, within a COG, the carbon content of ortholog proteins varies among species as the averaged carbon content of whole proteomes. Thus, taken together, the results of the analysis of COGs carbon contents establish that carbon content of ortholog proteins vary among species as the bulk proteins, independently of their abundance.

These results are further striking because the uncovered biases in protein carbon contents are quantitatively significant. Calculations indeed reveal that the amount of carbon atoms needed for protein synthesis may differ from more than 12% between species. This means that, depending on the organism, from 550 to 660 carbon atoms are, on average, used to build the residue side chains of a protein of 200 residues. We noticed that, according to the compilation of Neidhardt and Umberger (16), such a difference may account for more than 6% of all the carbon atoms fixed to construct a canonical bacterial cell such as *E. coli*.

Proteome Carbon Content Is Related to Genome Base Composition—Taken together, the above results argue that protein

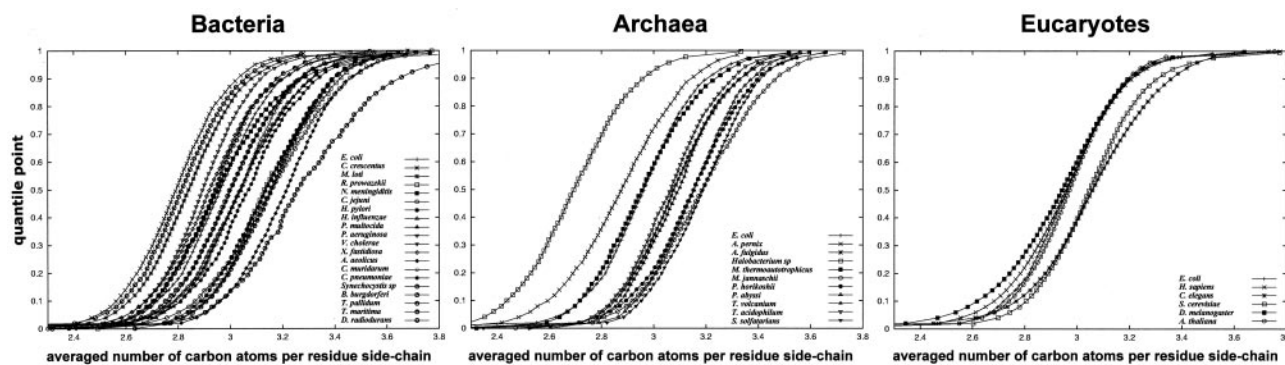


FIG. 2. Protein carbon usage in bacteria, archaea, and eukaryotes. The quantile distributions were calculated and are displayed as in Fig. 1. For bacteria, the carbon distribution plot does not include the firmicute distributions, which are displayed in Fig. 1. As in Fig. 1, the quantiles for *E. coli* were included in each graph.

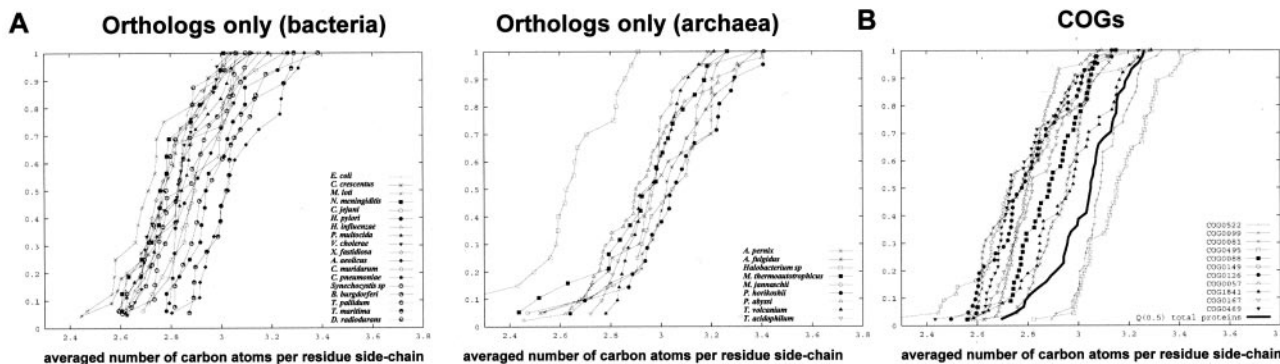


FIG. 3. A, carbon quantile distributions of abundant ortholog proteins shared by bacteria and archaea species. Several COGs (14) comprising one of the 75 more abundant proteins of *S. cerevisiae* were selected. The carbon usage of proteins belonging to these COGs was calculated for each indicated species (bacteria, 16 COGs; archaea, 20 COGs), and the corresponding quantile distributions are displayed. B, carbon contents of abundant ortholog proteins belonging to 11 different COGs were gathered by COG. The resulting COG carbon quantile distributions were displayed together with the quantile distribution of the median carbon contents ($Q_C^{-1}(0.5)$) of archaea and bacteria total proteins (plain line).

carbon content biases result from constraints that were superimposed to those acting on activity, folding, and stability of proteins and that are generally viewed as the primary determinants of protein evolution (17, 18). We thus searched for a plausible origin for these atomic composition biases. We first tried to establish whether variations of protein carbon contents result from the impoverishment or enrichment of particular amino acids. The amino acid contents of the proteins of organisms displaying various mean protein carbon contents were retrieved and compared. This comparison, made with seven prokaryotes, showed that organisms with high protein carbon contents contain, on average, more asparagine, lysine, and isoleucine residues (Fig. 4). Conversely, organisms with low protein carbon content encode proteins enriched in glycine and alanine residues. The differences are, however, weak, and the overall trend is rather that protein atomic variations result from variations in the usage of a large number of amino acids.

Various traits of the analyzed species were next inspected and did not suggest a direct explanation for the origin of the biases. However, we noticed that the ordering of the carbon distributions was somehow related to phylogeny and taxonomy. This pointed us at the guanine plus cytosine composition of genomic DNAs, which displays both large interspecific and low intraspecific variations as the above reported protein carbon biases. Indeed, DNA G+C content is known to vary from ~25 to 75% between species, with closely related species generally having similar DNA G+C contents (19). We thus examined whether the average protein carbon content might be correlated to DNA G+C content. As shown in Fig. 5A, a strong correlation between the genome nucleotide composition and the mean value of protein carbon contents was uncovered. The

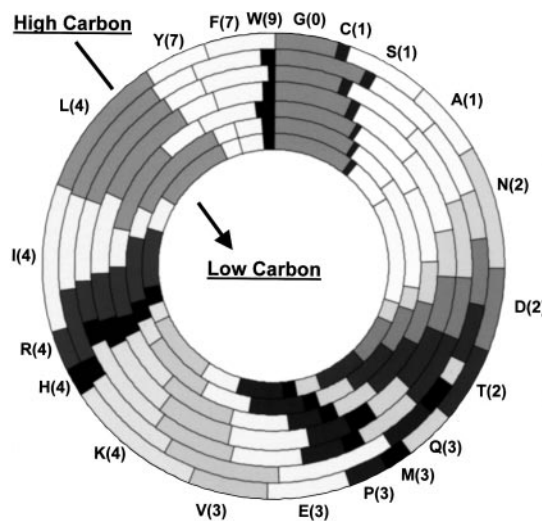


FIG. 4. Amino acid composition of proteomes with different carbon contents. The average amino acid content of proteins encoded by seven different prokaryotic species exhibiting different protein carbon contents was calculated from the Kazusa's Codon Usage Data base (www.kazusa.or.jp/codon/) and depicted as rings. Species (*U. urealyticum*, *T. maritima*, *T. pallidum*, *M. tuberculosis*, *M. pneumoniae*, *D. radiodurans*, and *B. subtilis*) were ordered outside to inside according to their averaged protein carbon contents. The amino acids are depicted as single-letter codes together with the number of carbon atoms found in their side chains.

results are statistically highly significant, with the median of the carbon quantile distributions being correlated with both the G+C content of the coding DNA ($p < 10^{-16}$, 44 species) and

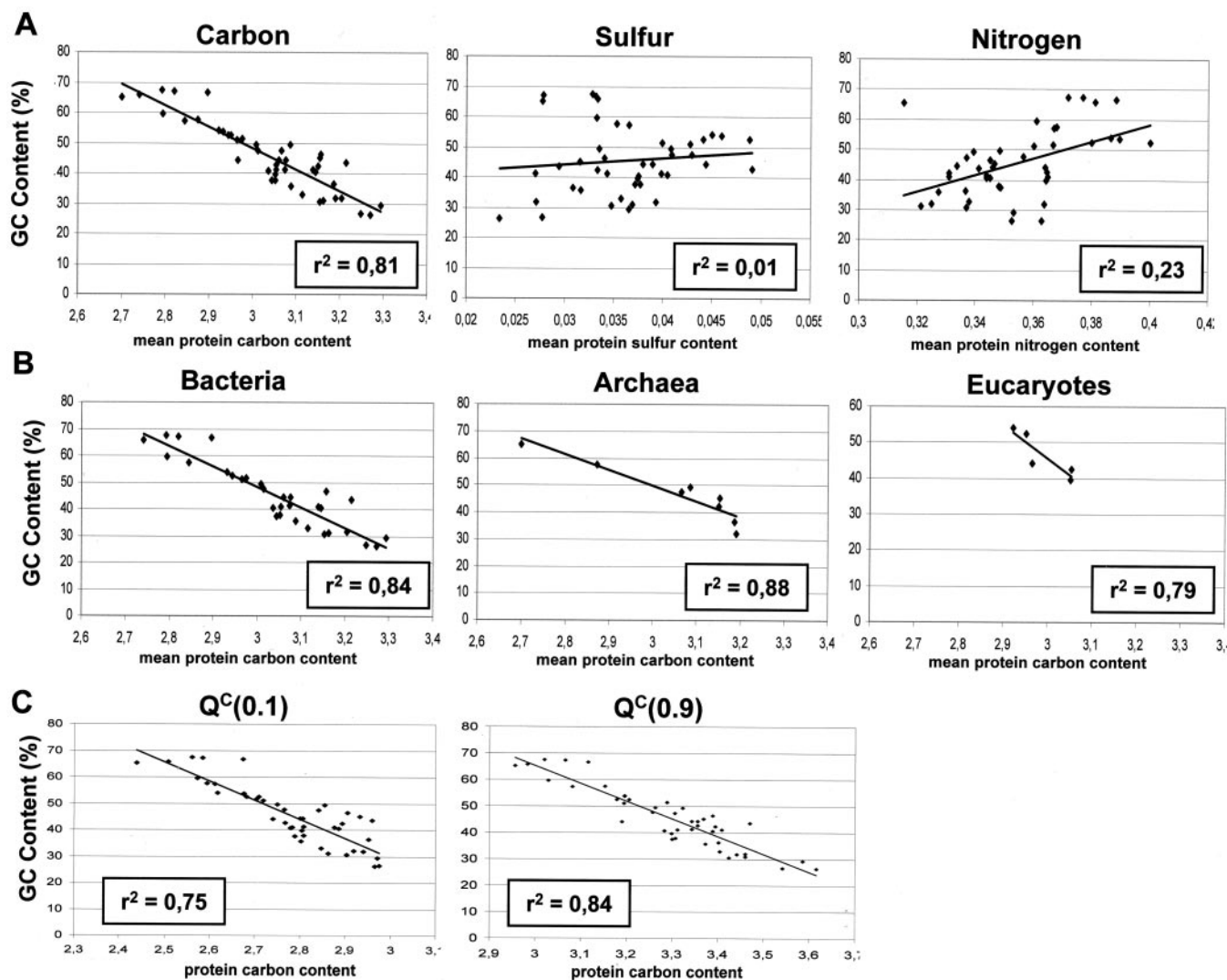


FIG. 5. **Correlation between G+C content and protein carbon content.** A, mean protein carbon content decreases with increasing G+C content of genome DNA, whereas no correlation is observed between either mean sulfur or mean nitrogen content of proteins and the G+C content of genome DNA. B, the correlation between G+C content and the mean protein carbon content applies in the three lineages: bacteria, archaea, and eukaryotes. C, regressions calculated for the $Q_C^{-1}(0.10)$ and $Q_C^{-1}(0.90)$ show that the correlation between genome G+C contents and protein carbon contents apply for at least 80% of the total proteins encoded by each species.

the G+C content of the total DNA ($p < 10^{-15}$, 40 species). In addition, regressions were calculated for each of the three kingdoms, bacteria, archaea, and eukaryotes, confirming that this relation applies for every lineages (Fig. 5B). In contrast, there is no correlation between protein sulfur content and genome composition, although protein sulfur distributions widely vary among species (Fig. 5A). To further assess that the correlation observed between the median quantile point of the carbon distributions and the G+C content indeed applies to the whole proteome and was not a result of extreme carbon usage in particular proteins or protein families, we used the 80% quantile ranges. Regressions between both the $Q_C^{-1}(0.10)$ and the $Q_C^{-1}(0.90)$ and the G+C DNA content were calculated, and as for the mean carbon content, highly significant correlations were found at both values (Fig. 5C). Finally, we compared the carbon content of ortholog proteins belonging to several COGs and their relationship to the G+C content of the genomes that encode them. As shown in Table I, the relationships between low protein carbon content and high DNA G+C content hold as well for ortholog proteins. All of these results therefore provide strong evidence that the carbon content of proteins is strongly related to the base composition of their encoding genome.

Proteome Carbon Content and Genome Base Composition Relationship Is Imprinted in the Extant Genetic Code—That proteome atomic composition is so closely entwined with genome base composition also suggests that this relation is imprinted in the structure of the canonical genetic code. To test the hypothesis that the genetic code indeed relates low protein carbon content to high DNA G+C content, we first followed a randomization approach. Using the natural codon sets, random protein samples were produced from random DNAs with different G+C contents (see “Experimental Procedures”), and the resulting protein carbon quantile distributions were calculated. This shows that all the random protein samples produced from random DNAs having the same G+C content display the same carbon quantile distributions (Fig. 6A). Moreover, the random protein carbon distributions are stochastically ordered according to the G+C content of the randomly generated DNAs. As shown in Fig. 6A, we found a strong correlation between nucleotide composition of random genomes and the median quantile value $Q_C^{-1}(0.5)$ of carbon contents of the corresponding proteins. On the contrary, protein sulfur content distributions of random proteomes do not correlate with the base composition of random DNAs (data not shown) as observed for real organisms.

To further analyze how the extant genetic code relates low

TABLE I
Variations, among six organisms displaying different DNA G+C contents, of the amount of carbon required for the building of several ortholog proteins

Ortholog proteins are defined according to the COGs established by Tatusov *et al.* (14). "Average carbon" is the averaged number of carbon atoms found per residue side chain, and "total carbon" is the total number of carbon atoms used to build each protein.

CDG		0012	0018	0030	0051	0081	0149	0575
Protein		Predicted GTPase	Arginyl-tRNA synthetase	Dimethyladenosine transferase	Ribosomal protein S10	Ribosomal protein L1	Triose-phosphate isomerase	CDP-diglyceride synthetase
<i>E. coli</i> gene		<i>vehF</i>	<i>argS</i>	<i>ksgA</i>	<i>rpsJ</i>	<i>rplA</i>	<i>tplA</i>	<i>cdsA</i>
<i>C. crescentus</i> (G + C = 67.42%)	residues	366	600	258	102	229	253	275
	average carbon	2.78	2.89	2.66	3.04	2.56	2.46	2.95
	total carbon	1752	2938	1204	514	1046	1130	1365
<i>M. tuberculosis</i> (G + C = 65.79%)	residues	357	550	317	101	235	261	305
	average carbon	2.74	2.79	2.79	2.99	2.63	2.64	2.81
	total carbon	1695	2634	1520	504	1089	1213	1474
<i>M. leprae</i> (G + C = 59.53%)	residues	356	550	306	101	235	261	312
	average carbon	2.76	2.78	2.87	3.00	2.56	2.68	2.83
	total carbon	1696	2632	1493	505	1096	1223	1508
<i>E. coli</i> (G + C = 51.11%)	residues	363	577	273	103	234	255	249
	average carbon	2.84	2.98	2.97	2.99	2.63	2.65	3.25
	total carbon	1760	2874	1357	514	1081	1187	1307
<i>M. pneumoniae</i> (G + C = 41.06%)	residues	362	537	263	108	226	244	395
	average carbon	3.06	3.28	3.20	3.10	2.84	2.99	3.36
	total carbon	1834	2834	1368	551	1095	1218	2121
<i>M. genitalium</i> (G + C = 31.74%)	residues	367	537	259	106	226	244	374
	average carbon	3.10	3.31	3.34	3.13	2.92	3.03	3.33
	total carbon	1871	2852	1385	544	1112	1228	1995

proteome carbon content to high G+C genome content, we next examined, within the current codon catalog, all of the single mutations that replace one amino acid with another, and we determined how single mutations that lower or increase amino acid carbon content in turn modify the G+C content of corresponding codons. Calculations (Fig. 6B) demonstrate that among the single mutations that lower amino acid carbon content, 49.05% do increase G+C codon content, 35.22% do not change the G+C codon content, and only 15.72% decrease the G+C codon content (as expected, the exact opposite trend was measured for mutations increasing amino acid carbon content with 49.05% lowering and 15.72% increasing the G+C codon content, respectively). Randomization assays show how the accumulation of mutations lowering the carbon content of a protein in turn increases the G+C content of its encoding gene (Fig. 6C).

Finally, we wondered whether the relationship for which we showed evidence above is specific to the extant genetic code. We therefore analyzed random protein samples produced with randomly generated codes. 500 alternative codes were generated according to the criteria of Haig and Hurst (8). From each randomly generated code, six randomly generated genomes coding for 2000 proteins of 280 residues and raising six different G+C contents were computed. The correlation factors r between those six G+C content values and the resulting carbon mean values were calculated for each code. The quantile distribution of the slopes of the corresponding regression lines were plotted for all of the codes leading to a squared correlation factor of >0.9 . As shown in Fig. 6D, the random reassignments of the 20 amino acids to the codon sets observed in the canonical genetic code show that less than 6% of randomly generated codes allow the protein carbon content to be negatively correlated with DNA G+C content and to vary with an amplitude equal or larger than what is observed with the canonical genetic code. Therefore, the ability of relating low protein carbon content to high G+C DNA content appears to be a property shared by only few genetic codes that could be generated from the codon sets found in the canonical code.

DISCUSSION

This study reveals the remarkable plasticity of the elemental composition of proteins, a striking but previously overlooked feature of these biopolymers. Indeed, large and widespread differences between protein atomic compositions are observed among species. This holds mainly for two elements, sulfur and carbon, the larger variations being measured for the later atom. Indeed, protein carbon contents differ more between species than within species. Moreover, these differences are not due to biases found in specific protein subclasses but apply to the entire proteomes. The carbon contents of ortholog proteins vary as the carbon contents of total proteins, and these variations are observed for abundant proteins as well. This indicates that constraints different from those acting on activity, specificity, folding, and stability came into play during the evolution of protein atomic structures. These results generalize at a proteome-wide level the variation of protein atomic contents already established for several subsets of enzymes such as cyanobacterial light-harvesting proteins (12) and microbial assimilatory enzymes (7). Such a protein evolution feature escaped previous attention because most of the studies devoted to protein evolution focused on amino acid composition. Indeed, protein carbon content differences do not result in biases of one or few amino acids but modify the usage of most of them. It is worthwhile to note that such protein carbon biases are of quantitative significance because they could account for more than 6% of the carbon required to construct a microbial cell, therefore strengthening the hypothesis of nutritional constraints as shaping protein structures.

A highly significant correlation between protein carbon content and the base composition of the genomes is moreover evidenced, with low protein carbon content being correlated with high DNA G+C content. This correlation, found in bacteria, archaea, and eukaryote kingdoms, does not arise from extreme carbon usage in peculiar proteins or protein families but applies to entire proteomes. The analysis of several ortholog proteins, belonging to different COGs, shows that the

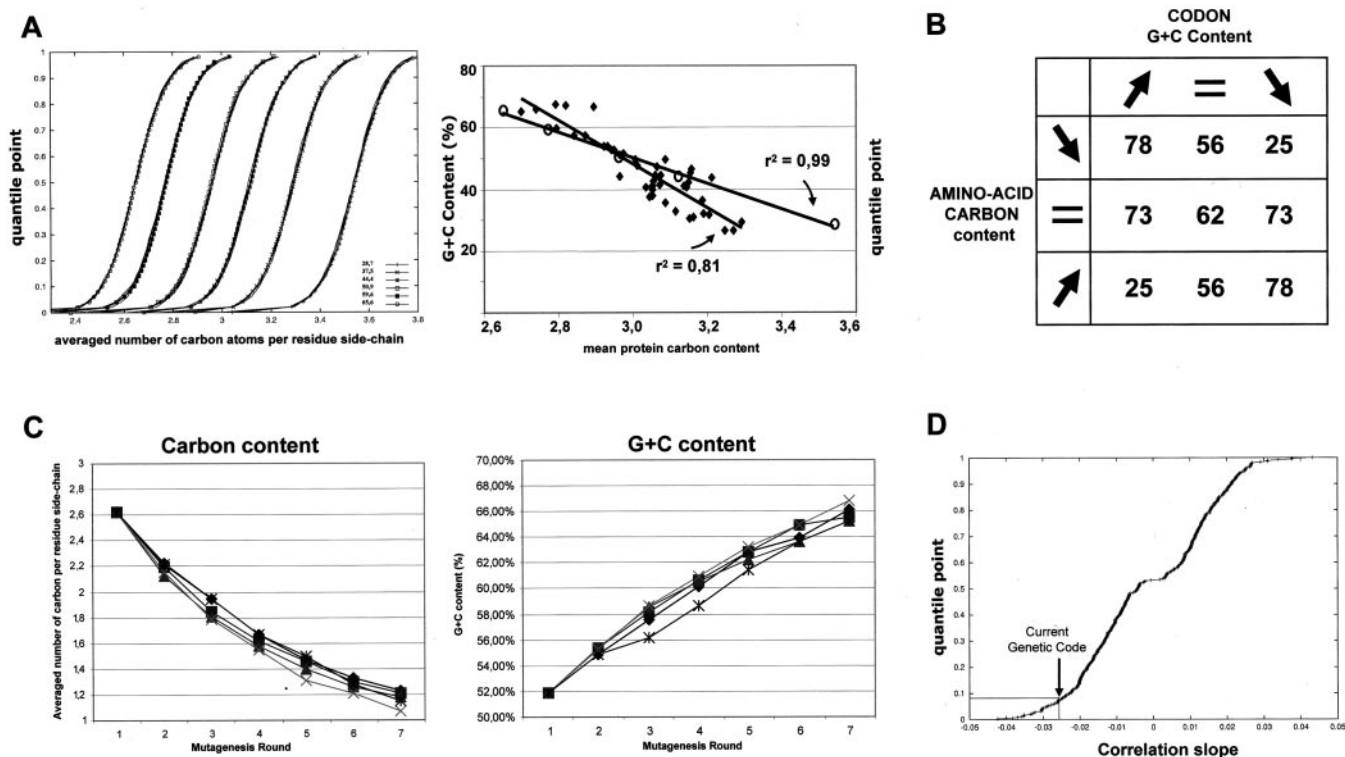


FIG. 6. A, DNA G+C content correlates with protein carbon content in random samples generated using the canonical genetic code. For each G+C value, five different random protein samples were produced, each comprising 2000 proteins. The carbon content of the protein data sets was next analyzed by the quantile method, and the corresponding distributions were plotted. For each G+C content value, the carbon quantile distributions of five different randomly generated protein samples are plotted (left panel). Correlation between G+C content and protein carbon content in the case of real proteomes (diamonds) and of randomly generated protein samples (circles, right panel). B, distribution, in the two-dimensional space of the codon G+C content changing and the corresponding amino acids carbon content changing, of all the single point mutations transforming one amino acid into one another. Up and down arrows indicate increase and decrease, respectively, of amino acid carbon content and codon G+C content. C, simulation of the evolution of protein carbon and gene G+C contents. The simulation was done using the *E. coli* ribosomal protein L1 (234 amino acids) as a starting point and submitting its corresponding gene to computerized mutations. Five independent simulations, each represented by one curve, were performed. For each simulation, up to 1638 mutations were proposed, and only those that strictly decreased the protein carbon content were kept. Carbon and G+C contents were checked every 234 proposed mutations (one mutagenesis round). D, analysis of G+C content and protein carbon content correlation with 500 randomly generated genetic codes. The random genetic codes were generated using the method of Haig and Hurst (8). For each randomly generated genetic code, six different genomes were randomly generated to have different G+C value contents. The correlation factors r between those six G+C content values and the resulting carbon mean values were next calculated. The graph displays the quantile distribution of the regression line slopes for all the codes leading to a squared correlation factor of >0.9 . The position of the canonical genetic code is indicated. Only the codes that are situated at the left of the canonical genetic code within the quantile distribution allow protein carbon content to vary in response to G+C content with an amplitude equal or larger than what is observed with the canonical genetic code.

relationships between low protein carbon content and high DNA G+C content hold as well for ortholog proteins. The correlation between DNA G+C and protein carbon contents strongly sustains the possibility that the composition of one class of biopolymers might have been determined by the other through the canonical genetic code. At the same time, this strong relation poses the conundrum of the flow of causality: DNA composition dictating protein atomic composition or the converse?

The first mode (DNA composition dictating protein atomic composition) is in line with the neutral theory of evolution (20) and more specifically with the original view of Suekoea (3), according to which genome composition biases are due to differences between the forward and backward mutation rates of the GC and AT pairs (21). Although the molecular basis of biased AT/GC pressure is unknown, it was suggested to act uniformly on DNA molecules (19). Biased mutation pressure was proposed to result in turn in amino acid composition biases in proteins (4, 21, 22). In this model, any modification in DNA base usage will gradually lead, along generations, to either frugal or wasteful carbon fixation. It has been noticed early that the smallest amino acids were indeed encoded by codons comprising the G and C bases only (23), yet the close relation-

ship that exists between protein carbon and DNA G+C contents might have been overlooked.

The second mode (protein composition dictating DNA composition) fits with a more selectionist view of evolution and with the previous proposals that metabolic flows and geochemical budgets might be constraints that were imprinted on protein evolution (7, 24). In a simple model, the adaptation to nutritional resources scarce in carbon could result in the fixation of mutations trimming the protein carbon content, which in turn lead to a progressive enrichment of genomes in GC bases. Examination of the structure of the genetic code provides further arguments for such a model. Indeed, calculations and randomization assays show that, given the structure of the canonical codon catalog, the successive accumulation of mutations lowering the carbon content of proteins will in turn lead to a genomic enrichment in GC bases. The fact that compositional constraints might come into play during protein evolution has been previously delineated in the case of highly expressed proteins in cyanobacteria; the sulfur-oligotrophic *Calothrix* Sp PCC7601 encodes a sulfur-depleted version of its most abundant protein (phycocyanin) that is specifically expressed under sulfur limitation (12). Likewise, upon cadmium exposition, yeast cells save sulfur by reducing the expression of sulfur

rich-proteins, such as several abundant glycolytic enzymes that are replaced by sulfur-depleted isoenzymes (13). Taken together, our results suggest that DNA G+C content biases might be a direct consequence of the optimization of protein atomic contents in response to carbon availability in natural habitats. Noteworthy, G+C contents are known to vary only weakly in metazoan species with the vertebrate genomes that show a quite uniform average G+C content, ranging from 40 to 45% (21, 25). This could be accounted for by the fact that vertebrates are expected to be subjected to specific elemental nutritional constraints less than micro-organisms. Also, free living bacteria tend to have genomes with significantly higher G+C contents, and thus to express proteins constructed with less carbon atoms, than pathogen or symbiont bacteria, which rely on nutritional resources provided by their hosts and thus are less subjected to environmental carbon limitations (26). Experimental confirmations of the hypothesis that genome base composition is driven by nutritional constraints shaping protein atomic contents would provide a new way to explore the origin of genome composition biases, a question that lies at the heart of current molecular evolutionary debates.

Acknowledgment—We are especially grateful to Bruno Sargueil for fruitful discussions and suggestions.

REFERENCES

- Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000) *Nature* **405**, 299–304
- Koonin, E. V., Aravind, L., and Kondrashov, A. S. (2000) *Cell* **101**, 573–576
- Sueoka, N. (1961) *Cold Spring Harbor Symp. Quant. Biol.* **26**, 35–43
- Singer, G. A. C., and Hickey, D. A. (2000) *Mol. Biol. Evol.* **17**, 1581–1588
- Thorne, J. L. (2000) *Curr. Opin. Genet. Dev.* **10**, 602–605
- Brock, D., and Madigan, M. T. (1991) *Biology of Microorganisms*, 6th Ed., pp. 634–648, Prentice-Hall International, London
- Baudouin-Cornu, P., Surdin-Kerjan, Y., Marliere, P., and Thomas, D. (2001) *Science* **293**, 297–300
- Haig, D., and Hurst, L. D. (1991) *J. Mol. Evol.* **33**, 412–417
- Karlin, S., Blaisdell, B. E., and Bucher, P. (1992) *Protein Eng.* **5**, 729–738
- Karlin, S., and Brendel, V. (1992) *Science* **257**, 39–49
- Woese, C. R. (1987) *Microbiol. Rev.* **51**, 221–271
- Mazel, D., and Marlière, P. (1989) *Nature* **341**, 245–248
- Fauchon, M., Lagniel, G., Aude, J. C., Lombardia, L., Soularue, P., Petat, C., Marguerie, G., Sentenac, A., Werner, M., and Labarre, J. (2002) *Mol. Cell* **9**, 713–723
- Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997) *Science* **278**, 631–637
- Ghaemmghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., De-phoure, N., O'Shea, E. K., and Weissman, J. S. (2003) *Nature* **425**, 737–741
- Neidhardt, F. C., and Umberger, H. E. (1996) in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* (Neidhardt, F. C., Curtiss, R., III, Ingramham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., Reznikoff, W. S., Riley, M., Schaechter, M., and Umberger, H. E., eds) pp. 13–16, American Society for Microbiology Press, Washington, D. C.
- Nei, M. (1987) *Molecular Evolutionary Genetics*, Columbia University Press, New York
- Tourasse, N. J., and Li, W. H. (2000) *Mol. Biol. Evol.* **17**, 656–664
- Muto, A., and Osawa, S. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84**, 166–169
- Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, UK
- Sueoka, N. (1962) *Proc. Natl. Acad. Sci. U. S. A.* **48**, 582–592
- Lobry, J. R. (1997) *Gene (Amst.)* **205**, 309–316
- Woese, C. R. (1965) *Proc. Natl. Acad. Sci. U. S. A.* **54**, 71–75
- Akashi, H., and Gojobori, T. (2002) *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3695–3700
- Bernardi, G., and Bernardi, G. (1985) *J. Mol. Evol.* **22**, 363–365
- Rocha, E. P., and Danchin, A. (2002) *Trends Genet* **18**, 291–294