ELSEVIER

Gene 205 (1997) 309–316

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species

J.R. Lobry *

CNRS UMR 5558-Laboratoire BGBP, Université Claude Bernard, 43 Bd. du 11-NOV-1918, F-69622 Villeurbanne, France

Accepted 26 June 1997

Abstract

The amino-acid composition of 23 490 proteins from 59 bacterial species was analyzed as a function of genomic G+C content. Observed amino-acid frequencies were compared with those expected from a neutral model assuming the absence of selection on average protein composition. Integral membrane proteins and non-integral membrane proteins were analyzed separately. The average deviation from this neutral model shows that there is a selective pressure increasing content in charged amino acids for non-integral membrane proteins, and content in hydrophobic amino acids for integral membrane proteins. Amino-acid frequencies were greatly influenced by genomic G+C content, but the influence was found to be often weaker than predicted. This may be evidence for a selective pressure, maintaining most amino-acid frequencies close to an optimal value. Concordance between the genetic code and protein composition is discussed in the light of this observation. © 1997 Elsevier Science B.V.

Keywords: Directional mutation pressure; Genetic code; Integral membrane proteins; Correspondence analysis

1. Introduction

The influence of the genomic G+C content on the average amino-acid composition of proteins was pioneered by Sueoka (1961). This work showed that, for 11 bacterial species whose DNA base composition varied from 35% to 72% G+C content, the amino-acid content of bulk protein preparation was dependent on genomic G+C content: out of 14 amino acids individually analyzed, there was a significant increase of Ala, Arg, Gly and Pro with G+C content, and there was a decrease of Ile, Lys, Tyr and Phe. G+C content has a strong influence, sufficient to modulate intracellular concentrations of tRNAs (Yamao et al., 1991), and is more influential than thermophilic adaptations (Filipski, 1990; Benachenhou-Lahfa et al., 1994).

Similar effects of the local G+C context on amino-acid content were reported for proteins from viruses (Bernardi and Bernardi, 1986; Karlin et al., 1990; Berkhout and van Hemert, 1994; Bronson and Anderson, 1994), mitochondria (Jukes and Bhushan, 1986; Jermiin et al., 1994), and eukaryotes (Bernardi and Bernardi, 1986; Hanai and Wada, 1988; Sueoka, 1988; D'Onofrio et al., 1991; Sueoka, 1992; Collins and Jukes, 1993), but no recent study has focused on the bacterial world. Bacteria are advantageous for the study of G+C influence on protein composition. First, there is an extreme wide variation, ranging from ~25% to ~75% G+C content, between different species of bacteria (Belozersky and Spirin, 1958; Sueoka, 1962), so that a wide range of the predictive variable is available. Second, the amount of intragenomic variability is, in contrast, very small (Rolfe and Meselson, 1959; Sueoka, 1959; Sueoka et al., 1959), so that to a first approximation, this effect can be neglected. Third, the within-species variability of G+C content is low (Brenner et al., 1972), so that G+C content polymorphism can also be neglected. Last, bacterial genomes are very small (10^6–10^7 bp), so that representative samples of many of their genomes are now available.

2. Materials and methods

2.1. Source of data

Data were from the international DDBJ/EMBL/GenBank databases (Benson et al., 1997; Stoesser et al., 1997; Tateno and Gojobori, 1997) structured under

* Tel: +33 4 72431287; Fax: +33 4 78892719; e-mail: lobry@biomserv.univ-lyon1.fr

PII S0378-1119(97)00403-4

ACNUC (Gouy et al., 1985) on 26 January 1997, including daily updates. Fifty-nine bacterial species (eubacteria + archaea) were selected, for which more than 100 kb of coding sequences were available (Table 1). In order to limit sampling bias due to multiple entries, the complete genomes of *H. influenzae* (Fleischmann et al., 1995), *M. genitalium* (Fraser et al., 1995), *M. pneumoniae* (Himmelreich et al., 1996), *M. jannaschii* (Bult et al., 1996), *Synechocystis* sp. (Kaneko et al., 1996), *E. coli* (F.R. Blattner unpublished, sequence accession number U00096), and the non-redundant *B. subtilis* database (Perrière et al., 1997) were used to extract coding sequences.

Table 1
List of bacterial species in the dataset

Species	G+C%	Number of proteins	Number of amino acids
Clostridium botulinum	24.8	59	51 826
Borrelia burgdorferi	31.2	385	132 791
Mycoplasma genitalium	31.4	451	169 633
Methanococcus jannaschii	31.8	1 516	468 845
Staphylococcus aureus	32.5	391	148 501
Bacillus thuringiensis	35.7	156	134 674
Lactococcus lactis	35.8	294	121 671
Sulfolobus solfataricus	36.3	213	65 321
Enterococcus faecalis	37.6	100	35 161
Streptococcus pyogenes	37.7	159	82 050
Haemophilus influenzae	38.5	1 505	494 717
Actinobacillus pleuropneumoniae	39.7	75	38 549
Helicobacter pylori	40.1	87	38 916
Mycoplasma pneumoniae	40.3	657	237 009
Chlamydia trachomatis	40.4	154	53 321
Streptococcus pneumoniae	40.6	163	99 541
Coxiella burnetii	41.9	126	40 095
Bacillus sp.	42.9	107	50 092
Vibrio cholerae	43.0	201	77 174
Anabaena sp.	43.7	133	46 449
Bacillus subtilis	44.1	1 591	542 415
Yersinia enterocolitica	45.1	165	59 899
Salmonella enterica	47.3	164	111 395
Synechocystis sp.	48.3	2 908	1 016 507
Methanobacterium thermoautotrophicum	48.6	163	54 494
Bacillus stearothermophilus	49.7	158	68 143
Neisseria meningitidis	51.2	175	128 107
Escherichia coli	51.3	3 913	1 331 601
Salmonella typhimurium	52.6	604	222 317
Neisseria gonorrhoeae	53.0	191	97 406
Synechococcus sp.	54.6	184	65 251
Erwinia chrysanthemi	55.4	98	35 907
Corynebacterium glutamicum	55.7	121	51 660
Agrobacterium tumefaciens	56.3	254	84 683
Pseudomonas syringae	56.7	139	48 352
Klebsiella pneumoniae	56.7	241	93 937
Serratia marcescens	57.2	119	50 270
Mycobacterium leprae	59.3	590	190 055
Pseudomonas fluorescens	59.4	124	51 872
Rhizobium leguminosarum	59.5	159	57 765
Pseudomonas sp.	59.7	122	44 498
Pseudomonas putida	60.0	272	98 162
Sinorhizobium meliloti	61.5	282	109 622
Bradyrhizobium japonicum	63.0	149	53 471
Xanthomonas campestris	63.2	107	43 632
Pseudomonas aeruginosa	63.3	562	203 972
Ralstonia eutropha	64.1	128	51 480
Mycobacterium tuberculosis	65.1	1 447	547 457
Azotobacter vinelandii	65.2	134	51 251
Bordetella pertussis	65.3	92	52 323
Rhodobacter capsulatus	65.5	234	87 310
Paracoccus denitrificans	65.8	127	40 443
Rhodobacter sphaeroides	67.3	149	52 900
Thermus aquaticus	67.5	179	68 501
Myxococcus xanthus	68.7	88	39 989
Streptomyces hygroscopicus	69.3	42	40 271
Streptomyces lividans	70.9	101	40 745
Streptomyces coelicolor	71.1	165	63 708
Streptomyces griseus	71.7	117	40 588
Total		23 490	8 678 695

2.2. Classification of proteins into two groups

The most important factor underlying the between-protein composition variability is the contrast between integral membrane proteins (IMP), which are enriched in hydrophobic amino acids, and other proteins (cytoplasmic, periplasmic, exported), which are enriched in hydrophilic amino acids, at least in *E. coli* (Lobry and Gautier, 1994). For this reason, care was taken to analyze separately these two groups of proteins: the hydropathy GRAVY score (Kyte and Doolittle, 1982) was used to classify proteins into two groups: IMP (GRAVY > 0.45) and non-IMP (GRAVY < 0.45). In all species, most amino acids analyzed here (94 ± 3.5%) were from non-IMP. *Paracoccus denitrificans* was an outlier with a relatively low proportion of 76%.

2.3. Computation of amino-acid frequencies in proteins

Only sequences labelled as coding sequences (CDS) were used. Partial CDS were discarded. CDS with less than 300 bp were discarded, to remove small peptides with an atypical amino-acid composition (e.g., leader peptides) and sequences annotated as CDS but which could not correspond to actual CDS (Fickett, 1995). As a matter of comparison, the expected average reading-frame lengths in random sequences are about 40 bp and 200 bp for 25% and 75% G + C content, respectively (Oliver and Marín, 1996). CDS were translated into proteins, taking into account deviations from the standard genetic code when necessary. The initial amino acid was not removed. The estimated frequencies were computed as n_i/n, with n_i the number of amino acids of type i, and n the total number of amino acids, within the species, and the protein group, under consideration.

2.4. Computation of genomic G + C content

The genomic G + C content was estimated directly from the total base count of coding sequences. These estimates were compared with those from buoyant density centrifugation or thermal denaturation midpoint

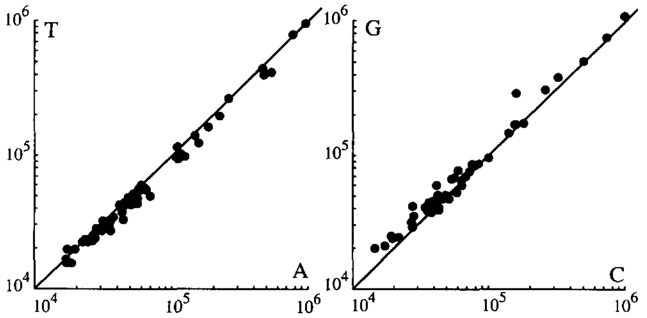

Fig. 1. Base counts in the dataset. Each point represents one bacterial species. The x-axis and y-axis are the total number of the indicated base in the coding sequences used in this study. The lines are the main diagonal ($y=x$). Points should be on this line if the simplifying assumption used in section 2.5 were perfectly true.

determinations of whole genomes when available. Values were very similar ($n=49$, $r^2=0.98$, slope$=1.04\pm0.04$) and never differed by more than 5% G+C, which is close to the within-species G+C content variability (Brenner et al., 1972).

Table 2
Average deviation between observed and expected amino-acid frequencies expressed in percent

Non-integral membrane proteins		Integral membrane proteins	
Arg	−4.56 (−5.06)	Arg	−6.75 (−7.23)
Ser	−3.14 (−2.72)	Pro	−3.30 (−2.75)
Pro	−2.96 (−2.67)	Ser	−2.78 (−2.27)
Cys	−2.04 (−1.89)	Cys	−2.21 (−2.02)
His	−0.99 (−0.85)	His	−1.63 (−1.49)
Thr	−0.70 (−0.69)	Thr	−0.89 (−0.90)
Trp	−0.48 (−0.52)	Gln	−0.66 (−0.80)
Tyr	−0.27 (−0.25)	Asp	−0.55 (−0.69)
Gly	+0.06 (−0.39)	Glu	−0.45 (−0.92)
Leu	+0.09 (+0.89)	Asn	−0.31 (−0.47)
Phe	+0.34 (+0.93)	Lys	−0.08 (−0.75)
Met	+0.66 (+0.59)	Tyr	−0.06 (−0.08)
Ile	+0.67 (+0.85)	Trp	+0.44 (+0.37)
Gln	+0.81 (+0.69)	Gly	+0.69 (+0.30)
Val	+0.89 (+0.92)	Met	+1.62 (+1.50)
Asn	+0.93 (+0.80)	Val	+2.32 (+2.51)
Ala	+1.99 (+1.95)	Ala	+2.53 (+2.44)
Lys	+2.30 (+1.68)	Ile	+3.38 (+3.43)
Asp	+2.53 (+2.41)	Phe	+3.49 (+4.04)
Glu	+3.40 (+2.95)	Leu	+4.64 (+5.29)

The value between parentheses is obtained when the assumption $P_A=P_T$ and $P_C=P_G$ is relaxed.

2.5. Computation of expected amino-acid frequencies

Let X_i be a random discrete variable whose value could be A, C, G or T, to denote the result of outcome number i in a random sampling experiment. Let $P_A=P(X_i=A)$, $P_C=P(X_i=C)$, $P_G=P(X_i=G)$, and $P_T=P(X_i=T)$, be the probabilities of obtaining each of the four bases. The way in which expected amino-acid frequencies were obtained will be better explained through an example. The probability for codon GAA was expressed as:

$$P(\mathrm{GAA})=P(X_1=G\cap X_2=\mathrm{A}\cap X_3=\mathrm{A})=P_G P_A P_A,$$

assuming the three events to be independent. Note that this assumption is restrictive in that it does not handle the case of coordinated mutation bias (that could result from CpG methylation or thymidine dimers for instance).

Stop codons have to be taken into account in order that expected amino-acid frequencies sum to one. The conditional probability of obtaining codon GAA, knowing a priori that a stop codon (i.e. TAA, TAG or TGA) is not obtained in a coding sequence, is given by:

$$P(\mathrm{GAA}|\mathrm{not-stop})=\frac{P(\mathrm{GAA})}{P(\mathrm{not-stop})}$$

$$=\frac{P_G P_A P_A}{1-(P_T P_A P_A+P_T P_A P_G+P_T P_G P_A)}.$$

The amino acid Glu, for example, is encoded either by codon GAA either by codon GAG. Its probability is then:

$$P(\mathrm{Glu})=P(\mathrm{GAA}\cup \mathrm{GAG}|\mathrm{not-stop})$$

$$=P(\mathrm{GAA}|\mathrm{not-stop})+P(\mathrm{GAG}|\mathrm{not-stop}),$$

because these are mutually exclusive events. In a similar way, expected frequencies for all other amino acids also can be expressed as a function of the four parameters P_A, P_C, P_G and P_T.

To set the values of these parameters the simplifying assumption that $P_A=P_T$ and $P_C=P_G$ was used. This is an idealization of the statistical relationships termed parity rule type 2 (Sueoka, 1995), which is observed in many species, even when considering only coding sequences (Lobry, 1995). In the present dataset, this assumption appears reasonable (Fig. 1), except for the archaea *M. jannaschii*, which is relatively enriched in G, and the results are essentially the same when it is relaxed (Table 2). With this assumption, there is only one degree of freedom left, that is, only one predictive variable: the G+C content (P_C+P_G). This G+C content parameter was set to the value corresponding to each species, as computed in Section 2.4. In the absence of selection, the genomic G+C content is controlled by directional mutation pressure. Using instead the G+C content at the third codon position would be somewhat contradictory with the null hypothesis of the model for which all codon positions are equivalent because of the absence of selection.

2.6. Analysis of expected amino-acid frequencies

The expected amino-acid frequencies were computed as described above for different G+C content values, from 0% to 100% with an incremental step of 5%. The resulting table of predicted composition of proteins was analyzed with a multivariate method, correspondence analysis (Hill, 1974), which is similar to the usual principal component analysis method, except that it takes advantage of the χ^2 metric instead of using the Euclidian one.

3. Results

3.1. Expected influence of G+C content on amino-acid composition

The first and second factors of correspondence analysis were found to extract 82.0% and 17.8% of the initial variability, the first factorial map (Fig. 2) summing up most of the expected behavior of amino-acid frequencies when G+C content varies. Protein points in this map are disposed along an arch, a known feature in multivariate analysis, also called the 'horseshoe effect' (Hill and Gauch, 1980), suggesting that there was a grading of amino-acid frequencies, which is explained here by the underlying G+C gradient. Amino-acid points are clustered into eight groups that correspond to different curves of amino-acid frequencies as a function of G+C content (represented by solid line in Fig. 3). These eight groups were clustered into three main classes:

(1) Six amino acids whose predicted frequencies monotonously decrease with G+C content (Ile, Phe, Lys, Tyr, Asn, Leu).

(2) Ten amino acids whose frequencies increase and then decrease (Asp, Glu, Ser, Val, Thr, His, Gln, Cys, Met, Trp). Their frequencies are maximum for a G+C content of 50% (except for Met and Trp), which is in the middle of the biological range (25–75%), so the expected influence of G+C content is low (variations are less important close to a maximum). For Met and Trp, a small decrease and increase, respectively, are expected within the biological range of G+C content.

(3) Four amino acids (Gly, Pro, Ala, Arg) whose predicted frequencies monotonously increase with G+C content.

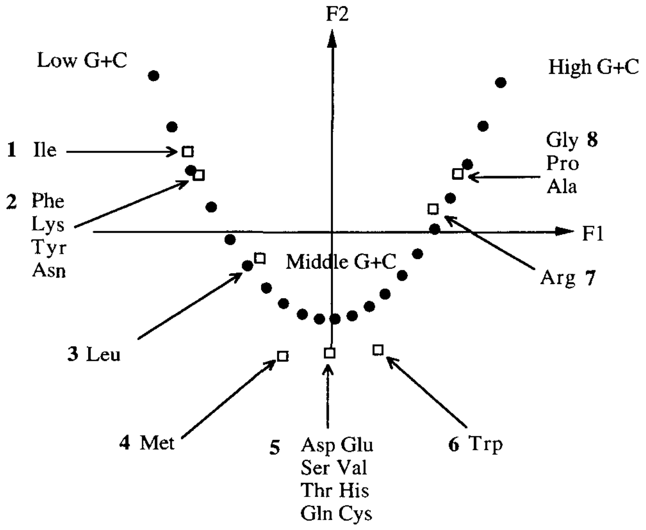

Fig. 2. First factorial map. Each point represents a protein with a theoretical amino-acid composition computed as explained in section 2.6. Squares represent one amino acid, or more when they share the same behaviour with respect to G+C content.

3.2. Observed and expected average amino-acid frequencies

The observed and expected average amino-acid frequencies were generally of the same magnitude, but with noticeable deviations (Table 2). Both IMP and non-IMP were enriched in Ala, with a higher enrichment in IMP. Other excess amino acids differ between proteins: Val, Ile, Phe and Leu for IMP, and Lys, Asp and Glu for non-IMP. Avoided amino acids are the same in the two groups: Arg, Ser, the disulfide-bridge-forming Cys, and the α-helix breaker Pro.

3.3. Observed influence of G+C content on amino-acid composition

The most important absolute variations of amino-acid frequencies in the biological range (from 25% to 75%) were for Ala (+10.9%), Ile (−10.5%), Phe (−5.4%), and Gly (+5.3%) in IMP, and for Ala (+9.3%), Lys (−8.6%), Asn (−6.2%), Arg (+6.0%) and Ile (−5.8%) in non-IMP. For these amino acids, the absolute variation of their frequencies is similar to their average content in proteins so that their relative variation is close to 100%.

Since most proteins are non-IMP (cf., Section 2.2), the influence of genomic G+C content on amino-acid content for this group was compared with Sueoka's results (Sueoka, 1961) on bulk protein composition, and found to be always consistent (Table 3). The effect was found to be quantitatively even more important, especially for Lys. This difference cannot result from the fact that, in Sueoka (1961), amino-acid frequencies were expressed as a percentage of the amino-acid content over the sum of stable amino acids, because this should, on the contrary, slightly increase the absolute value of Sueoka's slopes, as compared to those obtained here. This difference may be attributed to the fact that proteins with a high intracellular concentration are overrepresented in bulk protein extract so that their particular

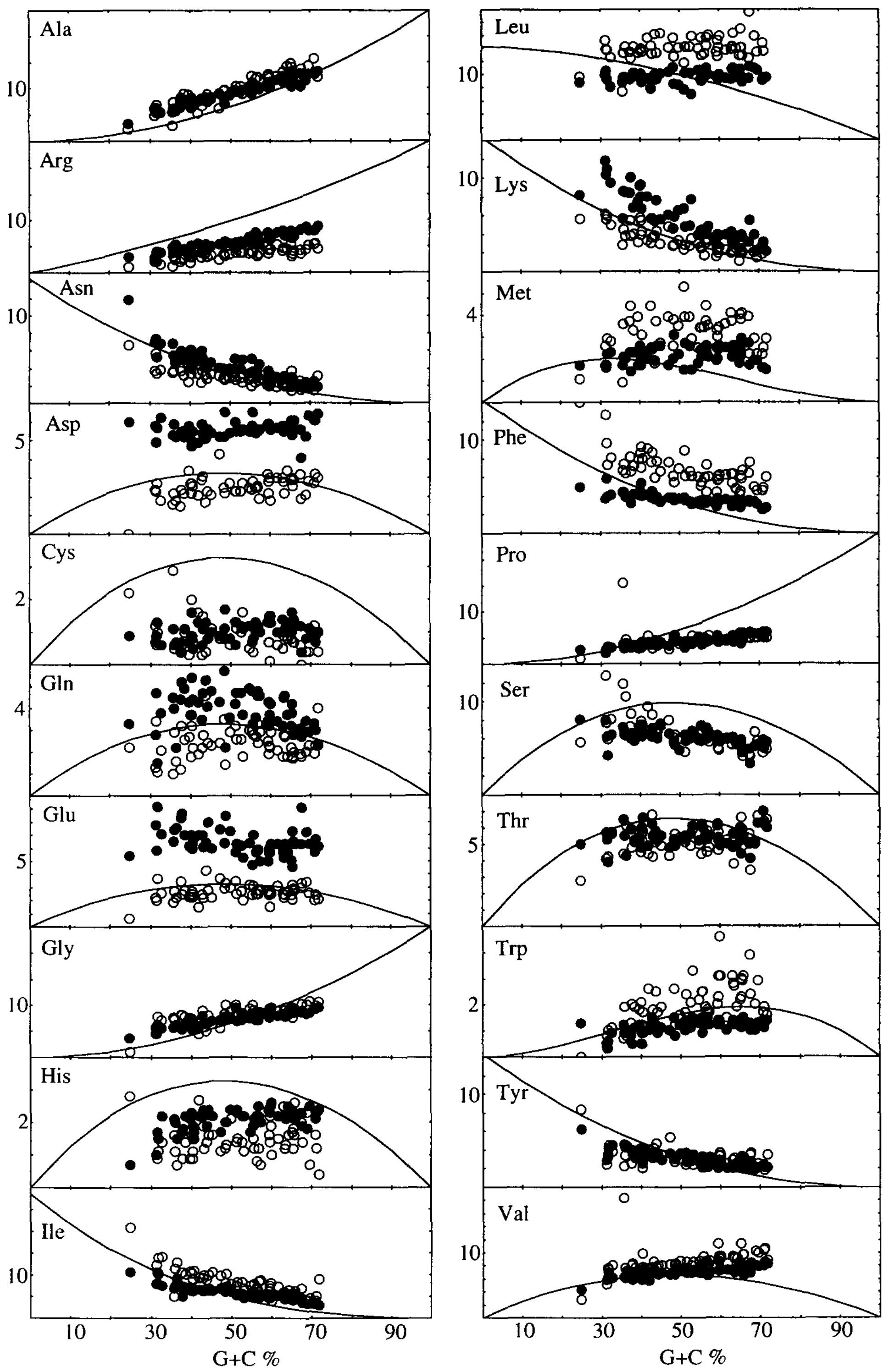

Fig. 3. Influence of G+C content on average amino-acid composition. Each point represents one bacterial species. Black points, non-integral membrane proteins; white points, integral membrane proteins. The y-axes are the amino-acid frequencies in percent, the x-axes the genomic G+C content in percent. The solid lines represent the amino-acid frequencies expected under the neutral model described in section 2.5.

amino-acid content may be of great influence: in *E. coli*, this bias is the second factor accounting for protein content variability, after the opposition between IMP and non-IMP (Lobry and Gautier, 1994).

The influence of genomic G+C content was found to be qualitatively the same for IMP and non-IMP (Table 3). However, the absolute values of the regression line slopes were different (at a significance level of 5%)

Table 3
Trends of the variation of amino-acid frequencies with G+C content

Group[a]	aa	Expected[b]	Non-IMP proteins[c]	IMP proteins[d]	Sueoka[e]
1	Ile	-0.239	-0.115±0.014	-0.210±0.032	-0.098
2	Tyr	-0.135	-0.053±0.010	-0.043±0.018	-0.047
	Asn	-0.135	-0.124±0.018	-0.065±0.014	
	Lys	-0.135	-0.171±0.024	-0.087±0.014	-0.084
	Phe	-0.135	-0.038±0.008	-0.107±0.026	-0.040
3	Leu	-0.148	+0.017±0.018	+0.025±0.034	-0.006
4	Met	-0.032	+0.004±0.008	+0.010±0.018	-0.024
5	Glu	-0.006	-0.029±0.020	+0.006±0.012	
	Asp	-0.006	+0.008±0.010	+0.020±0.012	
	Val	-0.012	+0.047±0.012	+0.075±0.020	+0.008
	His	-0.006	+0.016±0.006	+0.003±0.010	-0.010
	Thr	-0.012	+0.003±0.012	+0.018±0.016	0.000
	Ser	-0.019	-0.039±0.016	-0.082±0.024	-0.017
	Cys	-0.006	+0.008±0.006	-0.007±0.010	
	Gln	-0.006	-0.016±0.018	-0.002±0.014	
6	Trp	+0.027	+0.013±0.006	+0.041±0.014	
7	Arg	+0.248	+0.120±0.012	+0.069±0.018	+0.089
8	Pro	+0.254	+0.070±0.008	+0.051±0.014	+0.024
	Ala	+0.254	+0.186±0.014	+0.217±0.024	+0.164
	Gly	+0.254	+0.093±0.012	+0.105±0.026	+0.051

[a] Amino-acid groups are defined by a similar expected behaviour of their concentration with G+C content (Fig. 2). The horizontal lines delimit the three main amino-acid classes defined in section 3.1.

[b] Slope of the regression line of amino-acid concentration versus genomic G+C content when observed values are replaced by predicted values.

[c] Observed slope of the regression line of amino-acid concentration versus genomic G+C content for 59 bacterial species. Values are given ±1.96 SD. IMP stands for Integral Membrane Proteins.

[d] As in previous column but *B. thuringiensis*, outlier for Val and Pro in Fig. 3, was removed from the analysis.

[e] Results obtained by Sueoka (1961) from bulk protein preparations.

for some amino acids, showing that some regression lines between the two groups were not parallel, as one would have expected if the amino-acid sensitivity to G+C directional mutation pressure had been the same between the two groups. The amino acids that are relatively more affected by G+C directional mutation pressure are Ile, Ala, Phe, Ser and Val in IMP, and Lys, Asn and Arg in non-IMP.

3.4. Comparison of observed and predicted trends

There was a general qualitative agreement between expected and observed trends: the frequencies of amino acids of class 1 tended to increase with G+C content, those of class 2 to be relatively constant, and those of class 3 to decrease (Table 3), but there were quantitative differences. For the 12 amino acids that are expected to be influenced by G+C content (all but group 5), with the exceptions of Asn and Lys that fit the expected trend well, the absolute value of the slope of the regression line for the observed data was always lower than expected, with the extreme case of Leu and Met that are constant, whereas they are expected to decrease (Fig. 3). Hence, for most amino acids expected to be sensitive to the G+C directional mutation pressure, the effect was observed, but with a magnitude smaller than expected from the neutral model: observed slopes values are, on average, half of the expected slope values. For the eight amino acids that are expected to be independent of G+C content (group 5), small but significant variations were observed for Val and Ser.

4. Discussion

The general concordance between the structure of the genetic code and the amino-acid content of proteins can be interpreted in two different ways. It could be interpreted as an evidence that the genetic code evolved to its definitive form because this form best matches the amino-acid composition required by proteins in living material (MacKay, 1967). Another interpretation is that most amino acids in proteins are neutral so that a significant proportion of the present amino acids have arisen by random mutation and drift: amino acids will be present in rough accordance with their numbers of synonymous codons, weighted by base frequencies (King and Jukes, 1969).

When, for an amino acid, the concordance is not perfect, this deviation is usually explained as the results of selection, at the amino-acid level, against the genetic code (Jukes et al., 1975). In a similar way, the most important deviations observed here can be interpreted easily in terms of selective pressure at the amino-acid level thanks to the separation of proteins into two groups. Non-IMP are enriched in charged amino acids to increase their solubility, and IMP are enriched in hydrophobic amino acids to stay in the membrane. Moreover, in both groups, Cys is counter-selected to avoid the formation of unrequired disulfide bridges, and Pro is counter-selected because of its drastic effects on the 3D structure of proteins. The reason why Arg and Ser are avoided is less clear.

When, for an amino acid, there is a good concordance, these is no way to choose between the two hypotheses from the sole comparison of average expected and observed frequencies. However, for the 12 amino acids whose frequencies are expected to be affected by G+C content, some information can be obtained from the comparison of the observed and expected trends. Out of those, 10 amino acids do not follow the predicted trend with the expected magnitude. This could be interpreted as the result of a selective pressure to maintain their average content in proteins around an optimal concentration. This selective pressure would counteract the G+C directional mutation pressure, so that the observed trend is less than expected under purely neutral conditions. Then, if this interpretation is correct, it would mean that the genetic code was selected to match the average optimal concentration of these amino acids. For the eight remaining amino acids, no influence of G+C directional mutation pressure is expected, so that the study of the influence of G+C content on average amino-acid composition of proteins is not informative with respect to this problem.

Acknowledgement

This work was supported in part by a grant from Ministère de l'Enseignement Supérieur et de la Recherche (ACC-SV 9503006). Thanks are due to many participants of the ISME 1997 Symposium for stimulating discussion, to an anonymous referee and to Oliver Clay for constructive suggestions.

References

Belozersky, A.N., Spirin, A.S., 1958. A correlation between the compositions of deoxyribonucleic and ribonucleic acids. Nature 182, 111–112.

Benachenhou-Lahfa, N., Labedan, B., Forterre, P., PCR-mediated cloning and sequencing of the gene encoding glutamate dehydrogenase from the archaeon *Sulfolobus shibatae*: identification of putative amino-acid signatures for extremophilic adaptation. 1994. Gene 140, 17–24.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., 1997. GenBank. Nucleic Acids Res. 25, 1–6.

Berkhout, B., van Hemert, F.J., 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res. 22, 1705–1711.

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 1–11.

Brenner, D.J., Fanning, G.R., Skerman, F.J., Falkow, S., 1972. Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. J. Bacteriol. 109, 953–965.

Bronson, E.C., Anderson, J.N., 1994. Nucleotide composition as a driving force in the evolution of retroviruses. J. Mol. Evol. 38, 506–532.

Bult, C.J., et al., 1996. Complete genome sequence of the methanogenic archeon, *Methanococcus jannaschii*. Science 273, 1058–1073.

Collins, D.W., Jukes, T.H., 1993. Relationship between G+C in silent sites of codons and amino acid composition of human proteins. J. Mol. Evol. 36, 201–213.

D'Onofrio, G., Mouchiroud, D., Aòssani, B., Gautier, C., Bernardi, G., 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol. 32, 504–510.

Fickett, J.W., 1995. ORFs and genes: how strong a connection? J. Comp. Biol. 2, 117–123.

Filipski, J., 1990. Evolution of DNA sequence. Contributions of mutational bias and selection to the origin of chromosomal compartments. In Obe, G. (Ed.), Advances in Mutagenesis Research 2, Springer, Berlin, pp. 1–54.

Fleischmann, R.D., et al., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

Fraser, C.M., et al., 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397–403.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., di Paola, G., 1985. ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Appl. Biosci. 3, 167–172.

Hanai, R., Wada, A., 1988. The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. J. Mol. Evol. 27, 321–325.

Hill, M.O., 1974. Correspondence analysis: a neglected multivariate method. Appl. Stat. 23, 340–353.

Hill, M.O., Gauch, H.G., 1980. Decentered correspondence analysis: an improved ordination technique. Vegetatio 42, 47–58.

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

Jermiin, L.S., Graur, D., Lowe, R.M., Crozier, R.H., 1994. Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome *b* genes. J. Mol. Evol. 39, 160–173.

Jukes, T.H., Holmquist, R., Moise, H., 1975. Amino acid composition of proteins: selection against the genetic code. Science 189, 50–51.

Jukes, T.H., Bhushan, V., 1986. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J. Mol. Evol. 24, 39–44.

Kaneko, T., et al., 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* PCC6803. II. Sequence determination of the entire genome and assisgnment of potential-coding regions. DNA Res. 3, 109–136.

Karlin, S., Blaisdell, B.E., Schachtel, G.A., 1990. Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypotheses. J. Virol. 64, 4264–4273.

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

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Lobry, J.R., Gautier, C., 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. 22, 3174–3180.

Lobry, J.R., 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. 40, 326–330; 41, 680.

MacKay, A.L., 1967. Optimization of the genetic code. Nature 216, 159–160.

Oliver, J.L., Marín, A., 1996. A relationship between GC content and coding-sequence length. J. Mol. Evol. 43, 216–223.

Perrière, G., Moszer, I., Gojobori, T., 1997. The NRSub database: update 1997. Nucleic Acids Res. 25, 53–56.

Rolfe, R., Meselson, M., 1959. The relative homogeneity of microbial DNA. Proc. Natl. Acad. Sci. USA 45, 1039–1043.

Stoesser, G., Sterk, P., Tuli, M.A., Stoehr, P.J., Cameron, G.N., 1997. The EMBL nucleotide sequence database. Nucleic Acids Res. 25, 7–13.

Sueoka, N., 1959. A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proc. Natl. Acad. Sci. USA 45, 1480–1490.

Sueoka, N., Marmur, J., Doty, P., 1959. Heterogeneity in deoxyribonucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine–cytosine. Nature 183, 1427–1431.

Sueoka, N., 1961. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc. Natl. Acad. Sci. USA 47, 1141–1149.

Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA 48, 582–592.

Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85, 2653–2657.

Sueoka, N., 1992. Directional mutation pressure, selective constraints, and genetic equilibria. J. Mol. Evol. 34, 95–114.

Sueoka, N., 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40, 318–325; 42, 323.

Tateno, Y., Gojobori, T., 1997. DNA data bank of Japan in the age of information biology. Nucleic Acids Res. 25, 14–17.

Yamao, F., Andachi, Y., Muto, A., Ikemura, T., Osawa, S., 1991. Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. Nucleic Acids Res. 22, 6119–6122.