STATISTICAL ANALYSIS OF THE BASE COMPOSITION OF GENES USING DATA ON THE AMINO ACID COMPOSITION OF PROTEINS*

TOMOKO OHTA AND MOTOO KIMURA

National Institute of Genetics, Mishima, Japan

Received October 9, 1969

THE importance of DNA base composition per genome for the consideration of evolution and phylogeny was clearly shown by the work of SUEOKA (1961) who made an extensive compilation of the mean G-C (guanine and cytosine) content among diverse groups of organisms ranging from bacteria to vertebrates. SUEOKA (1962) and FREESE (1962) also proposed an evolutionary theory to account for the nature of the variation and heterogeneity of G-C content based on a statistical consideration of conversion between A-T (or T-A) and G-C (or C-G) pairs in a species.

From the standpoint of population genetics, molecular evolution consists of a series of base substitutions in the population, and there is growing evidence suggesting that this is *mainly* carried out by random fixation of selectively neutral mutants (KIMURA 1968, 1969; KING and JUKES 1969; CROW 1969).

For further understanding of the mechanism of evolution at the molecular level, it is desirable to know the frequency distribution of the base composition among the genes (cistrons).

In the present paper, we intend to analyze this distribution using data on amino acid composition of proteins derived from vertebrates.

It will be seen that this type of analysis poses many interesting problems not only in genetics and evolution but also in biometry.

STATISTICAL METHOD

In this section, we will describe a method of estimating the base composition of RNA which corresponds to a protein molecule, using published data on the amino acid composition. The standard RNA code table was used (see Table 1). Because of the degeneracy of the code, and especially because of the synonymy in the third position, a complete estimation is possible only for the first and second positions of the codon.

Let us designate the frequencies (in absolute numbers) of 20 kinds of amino acids in a molecule, by their three bracketed letters, such as [Ala], [Cys] and [Pro], respectively, denoting frequencies of alanine, cysteine and proline. Also, [Asn] and [Gln] denote the numbers of asparagine and glutamine, respectively. Let N be the total number of amino acids in a molecule. Also, let us denote by

* Contribution No. 745 from the National Institute of Genetics, Mishima, Shizuoka-ken, Japan. Aided in part by a Grant-in-Aid from the Ministry of Education, Japan.

TABLE 1

The standard RNA code dictionary
"Term." denotes chain-terminating codons

2 1	U	C	A	G	2 3
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term.	Term.	A
	Leu	Ser	Term.	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

U1, A1, C1 and G1 the frequencies of uracil, adenine, cytosine and guanine at the first position of the codon in the RNA molecule, and by U2, A2, C2 and G2, those at the second position.

Then, frequencies U2, A2 and G2 + C2 in the second position can be estimated *without error* as follows:

$$
\begin{aligned}
\mathrm{U2} &= \frac{1}{\mathrm{N}}\,([\mathrm{Phe}]+[\mathrm{Leu}]+[\mathrm{Ile}]+[\mathrm{Met}]+[\mathrm{Val}]) \\
\mathrm{A2} &= \frac{1}{\mathrm{N}}\,([\mathrm{Tyr}]+[\mathrm{His}]+[\mathrm{Gln}]+[\mathrm{Glu}]+[\mathrm{Asp}]+[\mathrm{Asn}]+[\mathrm{Lys}]) \qquad (1) \\
\mathrm{G2}+\mathrm{C2} &= \frac{1}{\mathrm{N}}\,([\mathrm{Ser}]+[\mathrm{Pro}]+[\mathrm{Thr}]+[\mathrm{Ala}]+[\mathrm{Cys}]+[\mathrm{Trp}]+ \\
&\qquad [\mathrm{Arg}]+[\mathrm{Gly}])
\end{aligned}
$$

In order to separate G2 from C2, an indirect method of estimation is required because serine having 6 synonymous codons contributes to both G2 and C2. Similarly, nucleotide frequencies in the first position have to be estimated indirectly, since leucine contributes both to U1 and C1, and, arginine contributes to both C1 and A1. The method of separation is as follows. In the case of serine, it corresponds to two groups of codons, one having UC and the other having AG in their first two positions. Accordingly, [Ser] can be divided into 2 parts, $[\mathrm{Ser}]_1$ and $[\mathrm{Ser}]_2$, and we assume that $[\mathrm{Ser}]_1$ is proportional to U1 × C2, and $[\mathrm{Ser}]_2$ is proportional to ½(A1 × G2). Namely, in this separation, we assume that frequencies in the first and second positions are statistically independent. Thus

$$
\begin{aligned}
[\mathrm{Ser}]_1 &= \frac{2\mathrm{U1}\times\mathrm{C2}}{(2\mathrm{U1}\times\mathrm{C2})+(\mathrm{A1}\times\mathrm{G2})}\times[\mathrm{Ser}] \\
[\mathrm{Ser}]_2 &= [\mathrm{Ser}]-[\mathrm{Ser}]_1
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
[\mathrm{Leu}]_1 &= \frac{\mathrm{U1}}{\mathrm{U1}+2\mathrm{C1}}\times[\mathrm{Leu}] \qquad (2) \\
[\mathrm{Leu}]_2 &= [\mathrm{Leu}]-[\mathrm{Leu}]_1 \\
[\mathrm{Arg}]_1 &= \frac{2\mathrm{C1}}{2\mathrm{C1}+\mathrm{A1}}\times[\mathrm{Arg}] \\
[\mathrm{Arg}]_2 &= [\mathrm{Arg}]-[\mathrm{Arg}]_1
\end{aligned}
$$

With these separations all the other nucleotide frequencies can be estimated as follows:

$$
\begin{aligned}
\mathrm{A1} &= \frac{1}{\mathrm{N}}\,\{[\mathrm{Ile}]+[\mathrm{Met}]+[\mathrm{Thr}]+[\mathrm{Asn}]+[\mathrm{Lys}]+[\mathrm{Ser}]_2+[\mathrm{Arg}]_2\} \\
\mathrm{U1} &= \frac{1}{\mathrm{N}}\,\{[\mathrm{Phe}]+[\mathrm{Leu}]_1+[\mathrm{Ser}]_1+[\mathrm{Tyr}]+[\mathrm{Cys}]+[\mathrm{Trp}]\} \\
\mathrm{C1} &= \frac{1}{\mathrm{N}}\,\{[\mathrm{Leu}]_2+[\mathrm{Pro}]+[\mathrm{His}]+[\mathrm{Gln}]+[\mathrm{Arg}]_1\} \qquad (3) \\
\mathrm{G1} &= 1-\mathrm{A1}-\mathrm{U1}-\mathrm{C1} \\
\mathrm{C2} &= \frac{1}{\mathrm{N}}\,\{[\mathrm{Ser}]_1+[\mathrm{Pro}]+[\mathrm{Thr}]+[\mathrm{Ala}]\} \\
\mathrm{G2} &= 1-\mathrm{A2}-\mathrm{U2}-\mathrm{C2}
\end{aligned}
$$

TABLE 2

Estimated base frequencies (%) in the first and second positions of the RNA codons

Protein	Total number amino acids	Base frequency (percent) 1st position U	C	A	G	2nd position U	C	A	G
Cytochrome C—Human	104	13.202	13.251	41.816	31.713	22.115	17.053	42.308	18.524
Cytochrome B5—Bovine	85	15.706	16.118	25.823	42.353	20.000	20.453	44.706	14.842
Hemoglobin α—Human	141	18.963	23.330	20.119	37.589	28.369	33.004	29.078	9.549
Hemoglobin β—Human	146	15.826	23.735	18.658	41.781	30.822	22.333	32.192	14.654
Trypsinogen—Bovine	229	23.749	13.715	33.278	29.258	22.707	22.223	29.695	25.376
Ribonuclease—Bovine	124	23.074	14.451	35.056	27.419	16.935	29.616	37.904	15.546
Trypsin Inhibitor—Bovine	58	26.714	16.029	27.947	29.310	17.241	23.423	27.586	31.750
Corticotropin (ACTH)—Human	39	22.071	22.733	16.734	38.462	20.513	23.797	35.897	19.793
Growth Hormone—Human	188	26.382	20.571	22.727	30.319	29.787	20.655	35.106	14.451
Posterior Pituitary Peptide—Bovine	31	23.706	25.188	15.690	35.417	14.583	21.706	20.833	42.877
Parathyroid Hormone—Bovine	27	9.243	20.948	26.952	42.857	26.190	13.474	41.667	18.669
Lysozyme—Chicken	129	21.802	10.984	36.206	31.008	19.380	19.488	27.907	33.225
Haptoglobin 2α—Human	142	14.835	21.762	28.192	35.211	16.197	17.986	47.887	17.929
Bence Jones λ	213	22.861	20.736	23.069	33.333	18.310	35.382	28.638	17.669
Bence Jones κ	214	24.622	20.013	27.794	27.570	21.963	29.429	33.645	14.963
Elastase—Pig	240	19.735	19.579	27.770	32.917	25.000	23.055	26.250	25.695
Glyceraldehyde-3-Phosphate Dehydrogenase—Pig	332	13.975	13.373	31.688	40.964	28.916	22.957	30.723	17.404
	2516	19.592	17.995	28.192	34.221	23.490	23.864	32.909	19.737

In applying formula (2) for separating $[\mathrm{Ser}]_1$ from $[\mathrm{Ser}]_2$ and so forth, we take arbitrary starting values, $A1_0, U1_0, C1_0, G1_0, C2_0$ and $G2_0$, for the corresponding frequencies A1, U1, etc. Here, $A1_0 + U1_0 + C1_0 + G1_0 = 1$ and $C2_0 + G2_0 = 1 - A2 - U2$. Using these initial values for base frequencies, the first estimates for $[\mathrm{Ser}]_1$, $[\mathrm{Ser}]_2$ . . . etc. are obtained using formula (2). Then the first estimates for base frequencies, A1, U1, . . . G2, are obtained by formula (3). Next, using these estimates, the second cycle of the same procedure gives the second set of estimates. With the help of the IBM 360 computer, the process is repeated many times until the maximum of the differences in the absolute values of the consecutive estimates of 6 frequencies is less than 0.00001.

ANALYSIS AND DISCUSSION

Seventeen vertebrate proteins were chosen from data compiled by DAYHOFF and ECK (1969). To avoid repetition, closely related proteins were not included. For example, once human hemoglobin α-chain was chosen, other vertebrate hemoglobin α-chains were not included in the analysis. Table 2 shows the results of the analysis. From this analysis, it turned out that the mean G-C content at the first position of the codon ($\bar{p}_1$) is 52.22% and that at the second position ($\bar{p}_2$) is 43.60%. Figure 1 illustrates the distribution of G-C content among genes

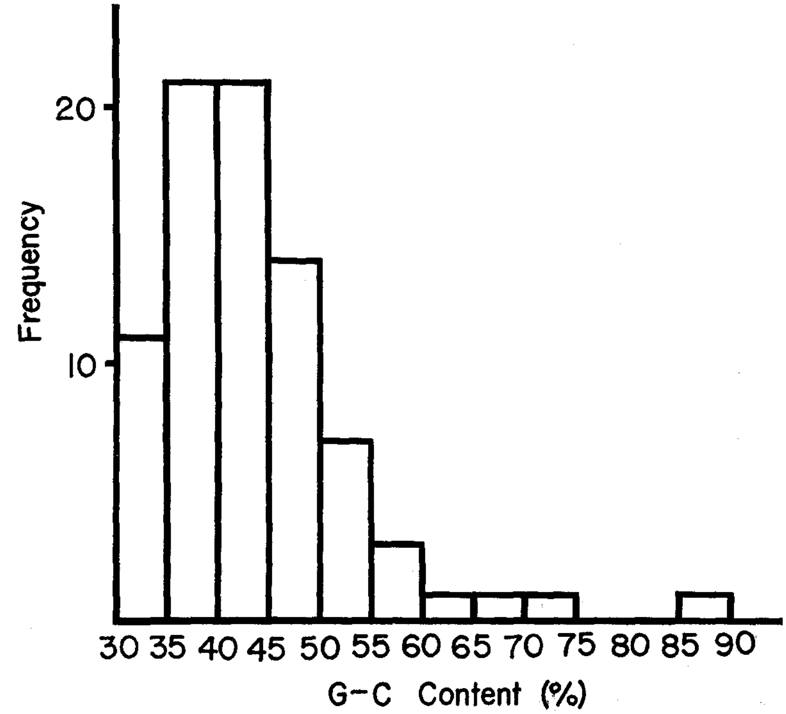

FIGURE 1.—Frequency distribution of G-C content among vertebrate genes with respect to the second position of codons.

with respect to the second position of codons. In constructing the histogram, data listed by SMITH (1966) were also included, in addition to those given by DAYHOFF and ECK (1969). Also, the variances of G-C content in the first and second positions are, respectively, $s^2_{p_1} = 0.00552$ and $s^2_{p_2} = 0.00495$. Note that their expected values, as calculated by $\bar{p}(1 - \bar{p})/\tilde{N}$ in which $\tilde{N}$ is the harmonic mean of N, turned out to be 0.0024. According to CHARGAFF (1955), the mean G-C content of vertebrate DNA is about 42% (see also SUEOKA 1962). It is interesting that $\bar{p}_1$ is considerably higher than this, although $\bar{p}_2$ is very close to it. The present analysis also reveals an interesting fact that frequencies of A and U, and, also those of G and C are not generally equal. This should mean that the base composition of an informational strand and that of its complementary strand are not generally equal. In fact, Table 2 shows that adenine content is almost always higher than uracil content. A χ^2 test for the hypothesis A2 = U2, using 17 proteins in Table 2, shows that deviation from it is highly significant. This conclusion must be especially valid for the second position of the codon, because, as shown in formulas (1), A2 and U2 can be estimated without error, and furthermore, as seen from Table 2, A2 > U2 was observed in all the vertebrate proteins analyzed in this paper. The tendency of A > U may also be inferred for the third position by comparing U + C *vs.* A + G, as tested by the ratio {[His]+[Asn]+[Asp]} : {[Gln]+[Glu]+[Lys]} which gave 331:400. Since the analysis was done at the level of messenger RNA, the informational strand in vertebrate DNA must contain more thymine than adenine. If the base composition is held at equilibrium by a balance between mutations of 4 bases in 12 directions, their mutation rates must be different. FITCH (1967) found a nonrandom character to base replacement by analyzing human hemoglobin mutants and also evolutionary amino acid substitutions in cytochrome c in various organisms. He found excess of guanine → adenine direction. This seems to agree with our result at the second position.

A difference of base composition between 2 complementary strands can also be seen clearly from Table 3, which is the result of analysis of silkworm (*Bombyx mori*) fibroin using data of SHIMIZU, FUKUDA and KIRIMURA (1957). They listed the amino acid composition of silkworm fibroin in weight percentage. This was converted to molar percentage and then analyzed by our statistical method. This protein has the characteristic that it consists mainly of alanine and glycine, so that its guanine content is very high. The base compositions in the first and the second positions of the codon are also quite different for this protein. The first column of Table 3 shows the base composition obtained by chemical analysis of the corresponding messenger RNA (data from S. AKUNE by personal communication). The difference between the result obtained by chemical analysis and those obtained by statistical analysis is probably due to the base composition at the third position, provided that the result of chemical analysis is accurate.

Finally we tried to reconstruct the amino acid composition in the same 17 enzymes as before, using the estimated base composition and assuming random combination of bases. The expected frequencies were calculated by multiplying base composition at the first position by that of the second position. For example, frequency of alanine is calculated by G1 × C2. When the synonymous codons are separated also at the third position, the ratio (U + C) *vs.* (A + G) at that position was multiplied. The amino acid frequencies thus obtained were corrected to make 100% in total, since there are three terminating codons, UAA, UAG and UGA.

TABLE 3

*Base composition (%) in the messenger RNA of silkworm fibroin as obtained by direct chemical analysis (*AKUNE*, personal communication) and as estimated by statistical analysis of the amino acid composition of fibroin*

RNA bases	By chemical analysis of messenger RNA	By statistical analysis of protein 1st position of codon	 2nd position of codon
G	49.4	79.7	47.8
C	12.5	0.8	41.1
U	11.6	16.6	3.6
A	26.5	2.9	7.5

TABLE 4

Amino acid composition of 17 proteins listed in Table 2

The expected frequencies were calculated assuming random combination of bases in the codon.

Amino acid	Observed frequency (%)	Expected frequency (%)
Serine	8.585	7.549
Glycine	8.267	7.087
Alanine	7.870	8.569
Valine	7.631	8.435
Leucine	7.273	7.077
Lysine	6.995	5.327
Threonine	6.121	7.059
Aspartic acid	5.445	5.350
Glutamic acid	5.008	6.466
Asparagine	4.928	4.408
Proline	4.491	4.506
Isoleucine	3.975	5.047
Glutamine	3.895	3.400
Tyrosine	3.855	3.063
Arginine	3.776	6.921
Phenylalanine	3.299	2.814
Cystein	3.021	1.837
Histidine	2.782	2.816
Tryptophan	1.471	1.110
Methionine	1.312	1.786

Table 4 gives the expected and the observed amino acid frequencies. The overall agreement between them is quite good, although for arginine and cysteine the agreement is less satisfactory. A similar comparison was made by KIMURA (1968) and more recently by KING and JUKES (1969). In their comparison, the observed frequency of arginine is much lower than expected. This tendency is also noticed in our Table 4, although the difference is not so marked. For a possible explanation of this discrepancy, readers may refer to KING and JUKES (1969).

At any rate, the good overall agreement between the observed and expected amino acid composition of proteins suggests, as pointed out by KING and JUKES (1969) and also by CROW (1969) and KIMURA (1969), a random nature of amino acid substitution in protein evolution, since, if it had occurred exclusively by natural selection of rare advantageous mutations, the frequency of a particular amino acid would not depend on whether it has many synonymous codons or only one. Actually, nonrandomness was pointed out already by many investigators. This should not mean that amino acid substitution in evolution is completely at random and that natural selection is not important.

Our data also indicate that the base arrangement is not wholly at random. The variances of G-C content between genes at the first and the second position of codons as estimated in the present paper are significantly larger than their expected values, indicating some nonrandomness in the occurrence of purines and pyrimidines. A similar tendency was noticed by KIMURA (1961) for data obtained by SUEOKA (1959, 1961) on the variance of G-C content among DNA molecules within an organism. For example, in calf thymus DNA, the observed variance is some 36 times as large as that expected from complete randomness. The harmonic mean of the number of nucleotide pairs per molecule is about 10^4 in this case. The ratio between observed and expected variances becomes less when the molecules are split into pieces of about one-tenth in size by ultrasonic vibration. KIMURA (1961) tried to explain such results by assuming the existence of repeating sequences. But it now appears more probable that such discrepancy is due to intrinsic differences in mean G-C content between genes caused by natural selection acting on gene function.

SUEOKA's work also shows that the variance is larger in higher than in lower organisms. The famous work of JOSSE, KAISER and KORNBERG (1961) on nearest neighbor analysis clearly shows nonrandom arrangement between adjacent base sequences in various organisms. SMITH (1969) recently evaluated the information density of DNA for several organisms and obtained a higher density for higher organisms.

Considering all these facts, we should conclude that although the majority of the base substitutions are selectively neutral or nearly neutral in the evolution of the species, natural selection is essential to bring out order in genetic constitution (see also KIMURA 1969).

SUMMARY

A statistical method was developed to estimate base composition of genes (cistrons) from amino acid composition of proteins. The method enables us to obtain base frequencies in the first and second positions of codons through an iterative process with a computer. However, with respect to the second position of codons, the G-C (guanine and cytosine) content can be estimated directly and *without error* by a simple formula. Using this method, 17 vertebrate proteins were analyzed to obtain the distribution of base frequencies among genes in the verte-

brate genome. The average G-C content turned out to be about 52% for the first position and 43% for the second position of codons. Their variances are higher than expected. It was also noticed that adenine content is almost always higher than uracil content, indicating different base composition between two complementary strands of vertebrate DNA. The problem of randomness in base arrangement in relation to molecular evolution by random drift and selection was discussed.

LITERATURE CITED

CHARGAFF, E., 1955 Isolation and composition of deoxypentose nucleic acids and of the corresponding nucleoproteins. pp. 307–372. In: *The Nucleic Acids 1*. Academic Press, New York.

CROW, J. F., 1969 Molecular genetics and population genetics. Proc. 12th Intern. Congr. Genetics **3**: 105–113.

DAYHOFF, M. O. and R. V. ECK, 1969 *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, Maryland.

FITCH, W. M., 1967 Evidence suggesting a nonrandom character to nucleotide replacements in naturally occurring mutations. J. Mol. Biol. **26**: 499–508.

FREESE, E., 1962 On the evolution of the base composition of DNA. J. Theoret. Biol. **3**: 82–101.

JOSSE, J., A. D. KAISER and A. KORNBERG, 1961 Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J. Biol. Chem. **236**: 864–875.

KIMURA, M., 1961 Natural selection as the process of accumulating genetic information in adaptive evolution. Genet. Res. **2**: 127–140. ——, 1968 Evolutionary rate at the molecular level. Nature **217**: 624–626. ——, 1969 The rate of molecular evolution considered from the standpoint of population genetics. Proc. Natl. Acad. Sci. U.S. **63**: 1181–1188.

KING, J. L. and T. H. JUKES, 1969 Non-Darwinian evolution: Random fixation of selectively neutral mutations. Science **164**: 788–798.

SHIMIZU, M., N. FUKUDA and J. KIRIMURA, 1957 Silk proteins. pp. 317–377. In: *Protein Chemistry* 5. (in Japanese) Kyoritsu Shuppan, Tokyo.

SMITH, M. H., 1966 The amino acid composition of proteins. J. Theoret. Biol. **13**: 261–282.

SMITH, T. F., 1969 The genetic code, information density, and evolution. Mathematical Biosciences **4**: 179–187.

SUEOKA, N., 1959 A statistical analysis of deoxyribonucleic acid in density gradient centrifugation. Proc. Natl. Acad. Sci. U.S. **45**: 1480–1490. ——, 1961 Variation and heterogeneity of base composition of deoxyribonucleic acids: A compilation of old and new data. J. Mol. Biol. **3**: 31–40. ——, 1962 On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. U.S. **48**: 582–592.