ENTROPY OF THE GENETIC INFORMATION AND EVOLUTION

MASAMI HASEGAWA and TAKA-AKI YANO

Dept. of Biophysics and Biochemistry, Faculty of Science, University of Tokyo, Hongo, Tokyo, Japan

Abstract. The entropy of the amino acid sequences coded by DNA is considered as a measure of diversity or variety of proteins, and is taken as a measure of evolution. The DNA or m-RNA sequence is considered as a stationary second-order Markov chain composed of four kinds of bases. Because of the biased nature of the genetic code table, increase of entropy of amino acid sequences is possible with biased nucleotide sequence. Thus the biased DNA base composition and the extreme rarity of the base doublet C_pG of higher organisms are explained. It is expected that the amino acid composition was highly biased at the days of the origin of the genetic code table, and the more frequent amino acids have tended to get rarer, and the rarer ones more frequent. This tendency is observed in the evolution of hemoglobin, cytochrome C, fibrinopeptide, immunoglobulin and lysozyme, and protein as a whole.

1. Entropy of Amino Acid Sequence in Protein

In a pair of papers, Zuckerkandl *et al.* (Zuckerkandl *et al.*, 1971; Vogel and Zuckerkandl, 1971) have suggested that, on the level of amino acid substitution during the evolution of proteins (cytochrome c and globins), the more frequent amino acids tend on the average to get rarer, and rarer ones more frequent, and hence the entropy of the amino acid sequence in protein tends to increase. Recently, the present authors (Yano *et al.*, 1973) have extensively studied on this subject.

The probable phylogenetic trees of several protein families are constructed, and the ancestral sequences are estimated from the contemporary protein sequences by the principle of minimum number of mutations. The protein families used are hemoglobin, cytochrome c. fibrinopeptide, immunoglobulin, lysozyme, toxin, insulin, proinsulin, growth hormone, virus coat protein, ferredoxin and calcitonin (Dayhoff, 1972). Along each branch of these trees, we count the number of the accepted point mutations, F_{ij}, from the *j*th amino acid to the *i*th, and the number of invariants, F_{ii}. The transition probability matrix of amino acid substitution, **L**, is obtained by normalizing the matrix **F**;

$$L_{ij}=F_{ij}/\sum_k F_{kj}.$$

The transition matrices are calculated for each protein family and for protein as a whole, and the latter is shown in Figure 1.

Dayhoff *et al.* (1969) have previously obtained the transition probability matrix by another way. Their method, however, does not take account of the direction of time explicitly, and, therefore, the origin of the asymmetry of the matrix is artificial and not obscure. On the other hand, our method takes the direction of time into account explicitly, and the asymmetrical matrix is automatically constructed.

Let's denote amino acid frequencies by a column vector **p**. If we multiply the initial vector $\mathbf{p}_0$, which represents amino acid frequencies of a contemporary protein family

Genetic code table

2nd→

		U	C	A	G		
1st↓	U	Phe Leu	Ser	Tyr term	Cys term Trp	U C A G	3rd↓
	C	Leu	Pro	His Gln	Arg	U C A G	
	A	Ile Met	Thr	Asn Lys	Ser Arg	U C A G	
	G	Val	Ala	Asp Glu	Gly	U C A G	

or protein as a whole, by the corresponding transition probability matrix **M**, the resulting vector **p** will be the amino acid frequencies after an arbitrary evolutionary interval,

$$\mathbf{p}=\mathbf{M}\cdot\mathbf{p}_0. \qquad (1)$$

The entropy of amino acid sequence in protein, H_{AA}, is defined by

$$H_{\mathrm{AA}}=-\sum_{i=1}^{20} p_i \log_2 p_i \quad \text{(bits)}. \qquad (2)$$

The entropy H_{AA} is a measure of the degree of evenness of amino acid distribution; i.e., the larger this value, the more evenly amino acids are distributed.

Figure 2 shows the behavior of entropy H_{AA} as a function of times of operations of **M** on the contemporary amino acid frequencies $\mathbf{p}_0$. Increasing trend of entropy is observed for every protein family as far as we have examined and protein as a whole. This trend contradicts Ohta and Kimura (1971) suggestion that the amino acid composition of contemporary organisms represents quasi-equilibrium states of substitution among selectively neutral or nearly neutral mutations. This contradiction comes from the difference of our transition matrix from Ohta and Kimura, which is counted by Dayhoff *et al.* (1969). The latter contains some artificial assumption on the origin of the asymmetry of matrix, while ours do not.

Zuckerkandl *et al.* (Zuckerkandl *et al.*, 1971; Vogel and Zuckerkandl, 1971) have interpreted the entropy increase as a randomization process. Our interpretation, however, is different from theirs. In so far as the central dogma is correct, the random process will occur only on DNA sequence. Then, because of the biased nature of genetic

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Asx	Glx	gap
Ala	9224	12	64	55	0	13	59	92	0	27	10	24	0	11	89	110	89	0	25	82	0	0	47
Arg	5	9629	21	0	0	51	8	23	48	13	14	59	49	0	10	19	15	33	0	0	0	0	16
Asn	21	0	9258	95	0	0	8	23	79	0	0	41	0	10	10	84	37	0	0	0	0	0	8
Asp	16	0	106	9422	0	0	269	41	16	0	0	12	0	0	30	45	7	0	25	0	0	0	55
Cys	0	0	0	0	9984	0	0	0	0	27	0	0	0	0	0	26	15	0	0	14	0	0	0
Gln	21	48	11	8	0	9517	118	9	79	27	5	47	0	0	20	13	7	0	0	21	0	0	8
Glu	53	12	32	206	0	89	9286	32	0	0	0	47	0	0	30	26	0	0	0	21	357	0	16
Gly	100	24	64	32	0	64	50	9592	0	0	10	24	0	0	20	58	22	0	25	27	0	0	55
His	5	24	42	24	0	64	0	5	9666	0	10	18	0	72	10	26	0	33	37	0	0	385	16
Ile	5	36	21	8	0	0	8	5	0	9295	87	6	0	51	20	26	37	33	0	178	0	0	16
Leu	21	12	11	0	0	25	25	5	0	186	9596	6	146	62	69	6	7	0	37	69	0	0	32
Lys	16	108	95	24	0	25	17	18	16	40	5	9556	0	0	10	0	22	0	0	7	0	0	16
Met	0	12	0	0	0	25	0	0	0	40	63	18	9563	10	0	0	7	33	0	55	0	0	0
Phe	11	0	0	0	0	0	0	0	0	13	58	6	49	9547	10	19	7	100	258	14	0	0	16
Pro	58	24	11	0	0	13	0	0	16	27	10	6	0	0	9524	39	7	0	0	0	0	0	16
Ser	211	24	170	32	16	25	0	73	16	0	14	36	49	21	69	9245	186	0	25	27	0	0	71
Thr	137	0	32	16	0	38	42	23	48	80	5	41	97	21	30	200	9441	0	0	34	0	0	16
Trp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9601	12	0	0	0	0
Tyr	11	12	21	8	0	0	8	0	16	13	5	6	0	154	0	19	7	133	9533	14	0	0	8
Val	53	0	32	16	0	13	50	14	0	213	91	24	49	31	30	19	52	33	12	9425	0	0	31
Asx	11	0	0	0	0	13	34	9	0	0	0	18	0	0	0	6	22	0	0	7	9643	0	24
Glx	5	12	0	0	0	0	0	5	0	0	0	6	0	0	0	6	7	0	0	0	0	9231	16
gap	16	12	11	55	0	25	17	32	0	0	19	0	0	11	20	6	0	0	12	7	0	385	9520

Fig. 1. Transition probability matrix for protein as a whole. The elements are shown multiplied by 10000.

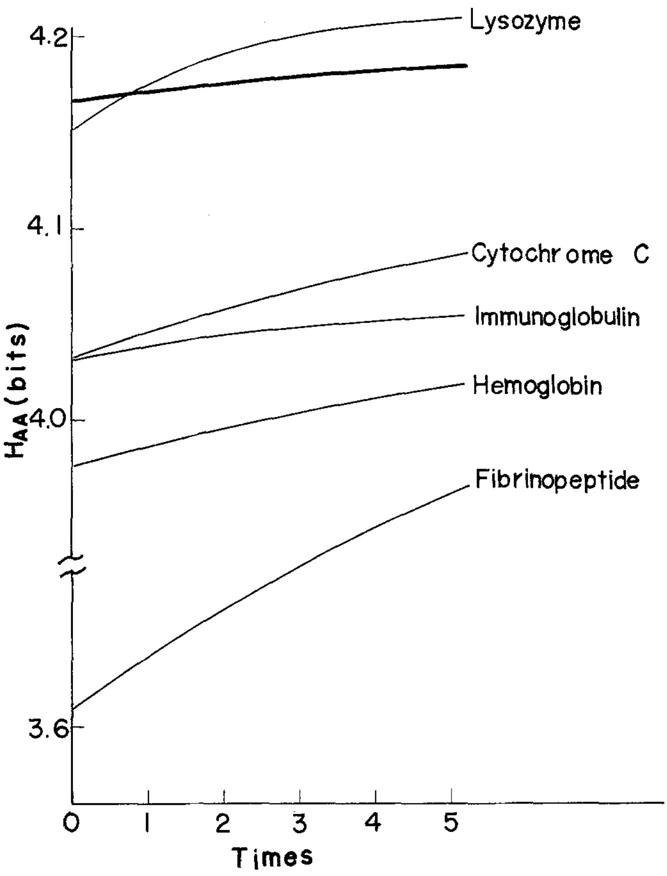

Fig. 2. Entropy H_{AA} as a function of times of operation of transition matrix on the contemporary amino acid frequencies. Bold line shows protein as a whole. The time scale of the abscissa is arbitrary and it differs for different protein families.

code, the entropy H_{AA} will not necessarily increase, but rather decrease. We regard the increasing trend of H_{AA} as a tendency against randomization. The H_{AA} value of the amino acid sequence coded by completely random DNA sequence, in which case amino acid frequency is proportional to the degree of codon degeneracy, is 4.14, while the observed H_{AA} values of the contemporary total protein are higher than this value, and now increasing. This suggests the possibility that, at the days of the origin and establishment of the genetic code, the amino acid frequencies of protein were highly biased owing to the biased nature of the code (Watson, 1970), and in the course of evolution the nucleotide base compositions and base doublet frequencies have become biased so as to decrease the bias of the amino acid frequencies and to increase the entropy H_{AA}. We will discuss on this subject in the following section.

2. Genetic Code Table and Entropy

Smith (1969) has shown that the maximum in the entropy function of protein provides an insight into the question as to why all vertebrates have a base C+G composition of about 42% (Sueoka, 1965). Since then, several authors have discussed this problem (King, 1972; Gatlin, 1972a, b; Miyazawa and Miyata, personal communication), and we have further developed as the following manner (Hasegawa and Yano, 1972a, b).

Now, DNA/mRNA is assumed to be a stationary second-order Markov chain of four kinds of bases; thymine (T, uracil U in mRNA), cytosine (C), adenine (A), and guanine (G). As this base sequence is assumed to be stationary, probabilities of four kinds of bases, $D(i)$, doublet frequencies, $D(i, j)$, and triplet frequencies, $D(i, j, k)$, in which $i, j, k \in \{\mathrm{U, C, A, G}\}$, are constant at any part of the strand. Then,

$$D(i, j) = \sum_{k \in \{\mathrm{U, C, A, G}\}} D(k, i, j) = \sum_{k} D(i, j, k), \tag{3}$$

$$D(i) = \sum_{j} D(i, j) = \sum_{j} D(j, i), \tag{4}$$

and the normalization condition

$$\sum_{i, j, k} D(i, j, k) = 1. \tag{5}$$

There exist $64 D(i, j, k)$'s, but the conditions (3) and (5) reduce the number of independently variable triplet frequencies to 48. Here, we are concerned with mRNA and are free from Watson-Crick constraint.

The entropy of stationary second-order Markov chain of mRNA is

$$H^{2\mathrm{M}}_{\mathrm{mRNA}} = -\sum_{i, j, k} D(i, j, k) \log_2 \{D(i, j, k)/D(i, j)\}. \quad \text{(bits)} \tag{6}$$

Still more, we define

$$H^{\mathrm{M}}_{\mathrm{mRNA}} = -\sum_{i, j} D(i, j) \log_2 \{D(i, j)/D(i)\} \quad \text{(bits)}, \tag{7}$$

$$H^{\mathrm{R}}_{\mathrm{mRNA}} = -\sum_{i} D(i) \log_2 D(i) \quad \text{(bits)}. \tag{8}$$

Base sequence of mRNA is translated into amino acid sequence according to the genetic code. The probability of the I-th amino acid, $P(I)$, and probability of amino acid doublet I-J, $P(I, J)$, of the translated amino acid sequence are

$$P(I) = \sum_{(i, j, k)} D(i, j, k), \tag{9}$$

$$P(I, J) = \sum_{(i, j, k)} \sum_{(l, m, n)} D(i, j, k, l, m, n) \tag{10}$$

where summation is over the degenerate codons of each amino acid, and besides 20 kinds of amino acid, chain terminating word is included. The entropy of amino acid sequence considered as a simple Markov chain is written as

$$H_{\mathrm{AA}}^{\mathrm{M}} = -\sum_{I=1}^{21} P(I, J) \log_2 \{P(I, J)/P(I)\} \quad \text{(bits)} \tag{11}$$

and that considered as a random chain is

$$H_{\mathrm{AA}}^{\mathrm{R}} = -\sum_{I=1}^{21} P(I) \log_2 P(I) \quad \text{(bits)}. \tag{12}$$

Let's maximize the entropy function of amino acid sequence $H_{\mathrm{AA}}^{\mathrm{M}}$ on the phase space spanned by 48 independent variables of $\{D(i, j, k)\}$ by the Monte Carlo Method. The results are

$D(\mathrm{U})=0.2630$, $D(\mathrm{C})=0.1757$,
$D(\mathrm{A})=0.3171$, $D(\mathrm{G})=0.2442$,
$H_{\mathrm{AA}}^{\mathrm{R}}=4.376$ bits, $H_{\mathrm{AA}}^{\mathrm{M}}=4.367$ bits,
$H_{\mathrm{mRNA}}^{\mathrm{R}}=1.970$ bits, $H_{\mathrm{mRNA}}^{\mathrm{M}}=1.865$ bits, $H_{\mathrm{mRNA}}^{2\mathrm{M}}=1.818$ bits.

C+G content of $H_{\mathrm{AA}}^{\mathrm{M}}$ maximum DNA is 42%, which coincides with Smith's result, and around this value the base composition of higher organisms converge. In Figure 3 the deviations of dinucleotide frequencies from random expectation, dfr, defined by

$$\mathrm{dfr} = \frac{D(i, j)}{D(i)\, D(j)} - 1 \tag{13}$$

are shown for human spleen DNA, *E. coli* DNA (Josse *et al.*, 1961; Swartz *et al.*, 1962), and DNA with the maximum $H_{\mathrm{AA}}^{\mathrm{M}}$ under Watson-Crick constraint. It is clearly seen that the doublet pattern of human DNA (any vertebrate DNA resembles closely) is more close to that of the $H_{\mathrm{AA}}^{\mathrm{M}}$ maximum DNA than that of *E. coli*. An especially distinctive feature is that doublet $\mathrm{C}_p\mathrm{G}$ frequencies of human DNA and the $H_{\mathrm{AA}}^{\mathrm{M}}$ maximum DNA are both scarce. The anomalous rarity of doublet $\mathrm{C}_p\mathrm{G}$ in vertebrate DNA are noted by Josse *et al.* (1961), Swartz *et al.* (1962), and King and Jukes (1969). Our result gives a reasonable explanation from the standpoint of informational theory. We take the entropy $H_{\mathrm{AA}}^{\mathrm{M}}$ as a measure of a variety of protein per one amino acid. Evolutionary process is considered to be accompanied with diversification of proteins so as to carry out various kinds of functions. As a consequence, the base compositions and the base doublet frequencies of DNA would have tended to the direction to increase the entropy $H_{\mathrm{AA}}^{\mathrm{M}}$.

In Figure 4, the entropies of protein H_{AA}, calculated from the experimental data of 20 amino acid frequencies for various organisms, are plotted as a function of C+G content. It is clearly seen that the C+G content of DNA is related to the entropy of amino acid sequence, and higher organisms with complex structure have higher H_{AA} values than those of lower organisms.

What does the increase of entropy of amino acid sequence mean? The theory of non-Darwinian evolution developed by Kimura (1968, 1969) and King and Jukes

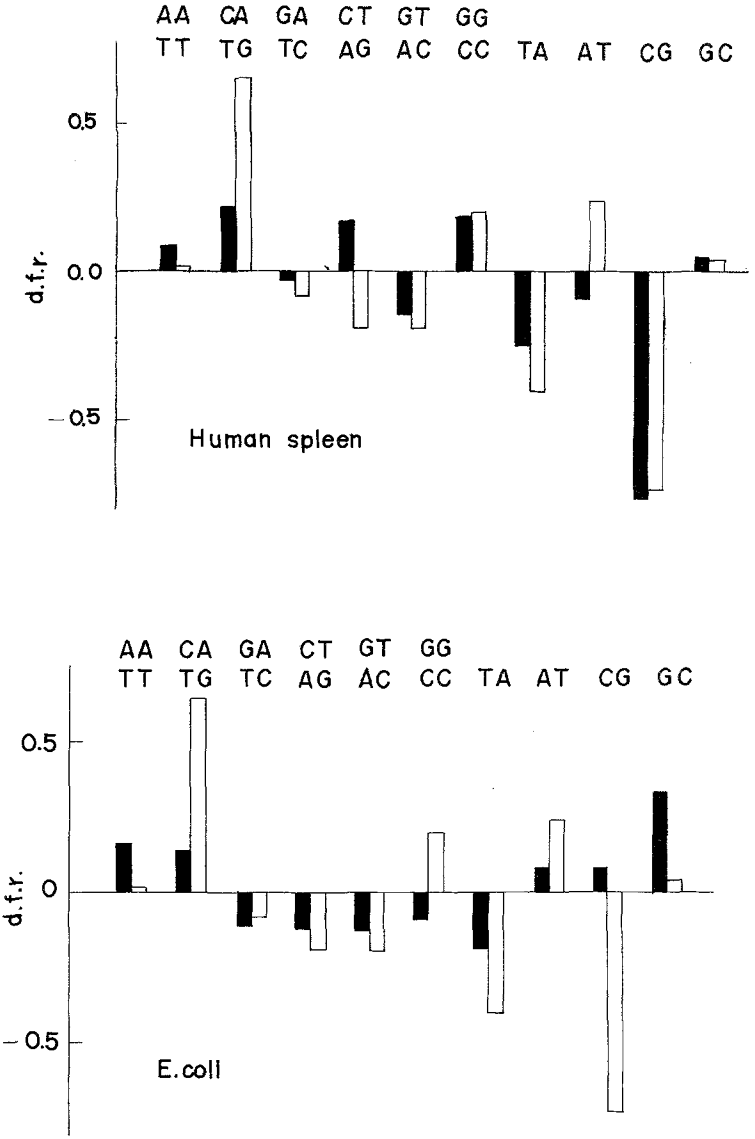

Fig. 3. Deviation of dinucleotide frequencies of DNA from random expectation. ☐ DNA with the maximum H^{M}_{AA}, ■ (a) human spleen DNA (C+G 40.4%), (b) E. coli DNA (C+G 49.8%).

(1969) is well established in the field of molecular evolution. Yet, if neutral mutations are predominant and randomization takes place on the level of nucleotide sequence in DNA, the frequency of each amino acid will tend to be proportional to the degree of its codon degeneracy, and the entropy will decrease. However, this is not the case. Although we cannot yet give a complete and satisfactory answer to this question,

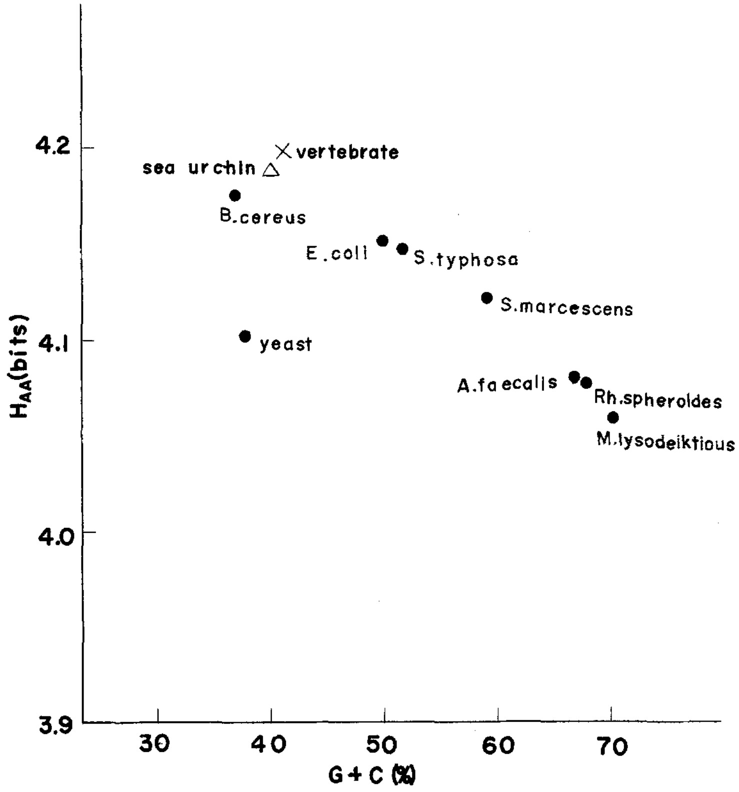

Fig. 4. Entropy H_{AA}, which is calculated from the experimental data of the frequencies of amino acids, as a function of C+G content.

this phenomenon may contain very important and prospective problems in the field of molecular evolution.

We wish to acknowledge Prof. Noda for continuous encouragement.

References

Dayhoff, M. O.: 1972, *Atlas of Protein Sequence and Structure.*

Dayhoff, M. O., Eck, R. V., and Park, C. M.: 1969, in M. O. Dayhoff (ed.), *Atlas of Protein Sequence and Structure*, National Research Foundation, Silver Spring, Md., p. 75.

Gatlin, L. L.: 1972a, *Math. Biosci.* **13**, 213.

Gatlin, L. L.: 1972b, *Information Theory and the Living System*, Colombia University Press, New York.

Hasegawa, M. and Yano, T.: 1972a, *Origin of Life and Evolution*, Kyoritsu Publ. Co., Tokyo (in Japanese), p. 134.

Hasegawa, M. and Yano, T.: 1972b, submitted to *Math. Biosci.*

Josse, J., Kaiser, A. D., and Kornberg, A.: 1961, *J. Biol. Chem.* **236**, 864.

Kimura, M.: 1968, *Nature* **217**, 624.

Kimura, M.: 1969, *Proc. Nat. Acad. Sci. U.S.* **63**, 1181.

King, J. L.: 1972, *Proc. 6th Berkeley Symp. Math. Stat. Prob.* University of California Press, Berkeley.

King, J. L. and Jukes, T. H.: 1969, *Science* **164**, 788.

Miyazawa, S. and Miyata, T.: Personal communication.

Ohta, T. and Kimura, M.: 1971, *Science* **174**, 150.

Smith, T. F.: 1969, *Math. Biosci.* **4**, 179.

Sueoka, N.: 1965, in V. Bryson and H. J. Vogel (eds.), *Evolving Genes and Proteins*. Academic Press, New York, p. 480.

Swartz, M. N., Trautner, T. A., and Kornberg, A.: 1962, *J. Biol. Chem.* **237**, 1961.

Vogel, H. and Zuckerkandl, E.: 1971, in E. Schoffeniels (ed.), *Molecular Evolution II, Biochemical Evolution and the Origin of Life*, North-Holland Publ. Co., Amsterdam, p. 352.

Watson, J. D.: 1970, *Molecular Biology of the Gene*, W. A. Benjamin, Inc., New York, p. 409.

Yano, T. and Hasegawa, M.: 1973, submitted to *J. Mol. Evol.*

Zuckerkandl, E., Derancourt, J., and Vogel, H.: 1971, *J. Mol. Biol.* **59**, 473.