Biophysical Chemistry, 44 (1992) 77–78
Elsevier Science Publishers B.V., Amsterdam

BIOCHE 01676

Letter to the Editor

Comments on the entropies of genetic codes of DNA and proteins

Akiyoshi Wada

Sagami Chemical Research Center, 4-1, Nishi-Ohnuma 4, Sagamihara, Kanagawa 229 (Japan)

(Received 13 March 1992)

In a recent publication entitled "Entropies of coding and noncoding sequences of DNA and proteins", Lauc, Ilić and Heffer-Lauc [1] made a comparison between entropies of the genetic code of DNAs and those of the primary sequence of proteins, where the entropy was calculated by Boltzmann's formula,

$$S = -\sum_i P_i \log_2 P_i, \qquad (1)$$

with the probability of ith coding unit being, P_i. They discovered a very interesting distinction between DNA and proteins: When triplets of nucleotides are taken as coding units, the entropy of the sequence based on the correct reading frame is significantly lower than the entropy of either of the frames shifted $+1$ or -1. When amino acids are adopted as coding units, in contrast, amino acid sequences based on the correct reading frame have higher entropies than the sequences which are produced by the frameshifts.

They found this a paradox, and hypothesized that the genetic code may have the ability to lower information content, i.e. to increase entropy, of proteins while translating them from DNA. I concur that the evidence may reveal something about the basic strategy of living organisms learned during their long evolutionary history; for instance, as the authors claimed, it would make the functional proteins more probable (having a higher entropy) than nonfunctional proteins translated from wrongly shifted frames. At the same time, however, I wish to point out that the findings can simply and directly be elucidated in terms of the nature of the universal codon table and the specific codon-usage strategy in the translation process as follows.

A codon consists of three nucleotides and a DNA sequence can be written as—1 2 3 1′2′3′1″2″3″—where 1, 2, and 3 denote the first, second, and third letter of a codon, respectively, and the single and double primes are used to distinguish sites belonging to different codons. In the correct frame, the sequence is divided into codon units as $\boxed{1\,2\,3}\boxed{1'2'3'}\boxed{1''2''3''}$, while frame shifted ones are $\overline{1\,|\,2\,3\,1'\,|\,2'3'1''|2''3''}$ and $\overline{1\,2|3\,1'2'\,|3'1''2''|3}$, for $+1$ and -1 shifts, respectively.

There are two distinctive conditions in the universal codon table itself and in its use in the translation event:

Condition A: A coding redundancy exists caused by the "wobbling" in the codon's "third letter", thus, there appear a number of synonymous codons which code for one amino acid but with a different third nucleotide. There is no such large redundancy at the second site, which uniquely determines the corresponding amino acid (except serine), while the first site has little more freedom than the second for synonymous base substitution.

Condition B: Codon usage within a group of synonymous codons is not arbitrary but is restricted to only a few members (distinctive ones are called optimal codons) in the group [2–5]. According to a statistical study on the base sequence of *Escherichia coli* DNA made by Blake et al. [6], the average codon probability of triplets in reading frame segments appears to be bimodal in distribution with about 40% of sequences having an average probability of 0.028 (36 codons are used) and the remaining 60% with values near 0.024 (42 codons are used). The average codon probability of triplets in nonreading frame segments has a normal distribution with a mode at 0.016, which is nearly 1/64, i.e. the random occurrence of all triplets.

From these two conditions, it is evident that the codons' third letter is bound by the type of doublet of the two letters that precede it, 1-2; a joint probability $P_{1\text{-}2,3}$ is defined [7,8]. The existence of the probability of the joint event implies a constraint which has the effect of reducing the entropy [9].

In both +1 and −1 frameshifts, however, the constraint caused by the 1-2-3 joint probability within a codon is partly removed [7], resulting in an increase in the entropy. This scheme based on the joint probability explains further details of the evidence presented by Lauc et al. that the increase in entropy is highest in the −1 frameshift, because the −1 shift cuts off the joint between the third letter and the preceding doublet 1-2, while the +1 shift leaves the 2-3 joint intact.

The reverse result reported for the amino acid sequence can be explained as follows. In the frame shifts, spurious codons, 2 3 1′ and 3 1′2′, obtain a true codon's third letter at the second and the first sites, respectively, both of which sites almost unequivocally determine the amino acid. Several amino acid species, therefore, have no (or very little) chance to appear in the translated sequence because the usage of the third letter is restricted (condition B), and the entropy of the amino acid sequence is reduced accordingly. This scheme is reasonably consistent with the details of the finding that the entropy is lowest in the +1 frameshift, and next to lowest in the −1 frameshift, for the former has the letter 3 at the second site where the strongest constraint is placed [10,11].

In conclusion, the intriguing nature of the entropy of DNA and protein coding sequences that has been revealed by Lauc et al. can be elucidated in terms of the coding redundancy, which almost solely originates in the third letter of the universal codon-table, and in the highly biased usage of codons.

References

1 G. Lauc, I. Ilić and M. Heffer-Lauc, Biophys. Chem. 42 (1992) 7.
2 R. Grantham, C. Guiter, M. Gouy, R. Mercier and A. Pave, Nucleic Acids Res. 8 (1980) r49.
3 T. Ikemura and H. Ozeki, Cold Spring Harbor Symp. Quant. Biol. 47 (1983) 1087.
4 H. Grosjean and W. Fiers, Gene 18 (1982) 199.
5 P.M. Sharp, E. Cowe, D.G. Higgins, D.C. Shields, K.H. Wolfe and F. Wright, Nucleic Acids Res. 16 (1988) 8207.
6 R.D. Blake, P.W. Hinds, S. Earley, A.L. Hillyard and G.R. Day, in: Biomolecular Stereodynamics 4, eds. R.H. Sarma and M.H. Sarma (Adenine Press, New York, NY, 1986) 271 pp.
7 R. Hanai and A. Wada, J. Mol. Biol. 207 (1989) 655.
8 R. Hanai and A. Wada, J. Mol. Evol. 32 (1990) 109.
9 L. Brillouin, Science and information theory (Academic Press, New York, NY, 1962) p. 17.
10 A. Wada, A. Suyama and R. Hanai, J. Mol. Evol. 30 (1991) 374.
11 A. Wada, Adv. Biophys. 28 (1992) in press.