Biophysical Chemistry 42 (1992) 7–11
Elsevier Science Publishers B.V., Amsterdam

BIOCHE 01610

Entropies of coding and noncoding sequences of DNA and proteins

Gordan Lauc [a,*], Igor Ilić [b], Marija Heffer-Lauc [c]

[a] *Laboratory of Physical Chemistry, Faculty of Science, University of Zagreb, Marulićev trg 19, 41000 Zagreb (Croatia)*
[b] *Department of Electrical Engineering, University of Osijek, Istarska 3, 54000 Osijek (Croatia)*
[c] *Department of Chemistry and Biochemistry, Faculty of Medicine, University of Zagreb, Šalata 3, 41000 Zagreb (Croatia)*

(Received 6 December 1990; accepted in revised form 7 May 1991)

Abstract

The entropies of protein coding genes from *Escherichia coli* were calculated according to Boltzmann's formula. Entropies of the coding regions were compared to the entropies of noncoding or miscoding ones. With nucleotides as code units, the entropies of the coding regions, when compared to the entropies of complete sequences (leader and coding region as well as trailer), were seen to be lower but with a marginal statistical significance. With triplets of nucleotides as code units, the entropies of correct reading frames were significantly lower than the entropies of frameshifts +1 and −1. With amino acids as code units, the results were opposite: Biologically functional proteins had significantly higher entropies than proteins translated from the frameshifted sequences. We attempt to explain this paradox with the hypothesis that the genetic code may have the ability of lowering information content (increasing entropy) of proteins while translating them from DNA. This ability might be beneficial to bacteria because it would make the functional proteins more probable (having a higher entropy) than nonfunctional proteins translated from frameshifted sequences.

Keywords: Entropy; Information content; Genetic code; Frameshift mutation

1. Introduction

The thermodynamic quantity entropy (S) has been, since Boltzmann has defined it as

$$S = -k_B \sum_{i=1}^{n} P_i \ln P_i \qquad (1)$$

(k_B denoting the Boltzmann constant), closely related with disorderliness, randomness and sometimes even used as a measure for the lack of information. While the first two relations are universal, the identification of entropy with missing information has proved to be controversial [1–5], and of limited use, being applicable only to those systems where the information could be clearly defined, for example in the genetic code [6].

In 1949 C.E. Shannon proposed a similar function,

$$H = -\sum_{i=1}^{n} P_i \log_2 P_i \qquad (2)$$

as a measure for information per digit in a coded message [7]. The applicability of his approach to biological systems has been widely discussed ever

* To whom correspondence should be addressed.

since [8–13]. Most controversies arose from disputes concerning actual signs (+ or −) in these relations [13,14]. However, this debate seems to be settled because it is now widely accepted that, beside average information per digit of coded message, Shannon's H function (eq. 2) also measures the missing information about the system [6,15].

The initial idea of this paper was to apply Boltzmann's formula (eq. 1) in calculating entropies of functional DNA and protein sequences and to compare the obtained values to the entropies of noncoding or miscoding sequences in order to find out whether this approach could discriminate between coding and noncoding DNA sequences, and whether it could be used as a possible method to verify the open reading frame (one of the three possible frames that gives the biologically functional protein).

Although thus calculated entropies did not prove to be accurate enough to distinguish between coding and noncoding DNA sequences, they have shown to be very specific in discriminating reading frames from frameshifts. The entropies of the reading frames were significantly lower than those of the frameshifted sequences, which is in accordance with the intuitive expectation that the sequences containing a definite biological information should have a lower entropy than the frameshifted sequences which lack that information.

Unexpected results emerged upon the translation of these DNA sequences to proteins. Protein sequences originating from the reading frame, i.e. biologically functional proteins, showed significantly *higher entropies* than the protein sequences which were translated from frameshifted sequences (sequences unable to yield the functional proteins).

Since translation according to the genetic code is the only operation that has been performed on those sequences, we assume that the genetic code may have the ability of increasing entropy (decreasing information content) of proteins while translating them from the DNA sequence. This ability might be beneficial to organisms because it would make the functional proteins more probable (having a higher entropy) than nonfunctional proteins produced by frameshifted translation of the genetic information.

2. Methods

Basic statistical unit in this study was one DNA sequence. Calculations were performed in three ways, i.e. assuming nucleotides, triplets of nucleotides and amino acids as code units. Entropy (S) was calculated according to the modified Boltzmann formula, with the proportionality constant k chosen to fit the binary code, i.e. $k = 1/\ln 2$.

$$S = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{3}$$

where P_i denotes the probability of a particular (ith) code unit (nucleotide, triplet, or amino acid); for instance, the probability of the amino acid glycine in the enzyme serine-tRNA synthetase from *E. coli* is 0.0696 (number of glycines in the sequence divided by the total number of amino acids in the sequence).

With nucleotides as the code units, the entropies of the coding region were compared to the entropies of the complete sequence (leader, coding region and trailer). With triplets as the code units, the entropies of the reading frames were compared to the entropies of frameshifted sequences (sequences derived from the original sequence reading with frame shifted for one (+1) or two nucleotides (−1)). With amino acids as the code units entropies of the biologically functional proteins were compared to the entropies of nonfunctional polypeptides which would be produced if frameshifted sequences were translated.

The significance of entropy differences was tested by using Student's t test (p_t) and the non-parametric Mann–Whitney test (p_{MW}).

DNA sequences were collected from the computer database 'Micro genius' and checked against the original papers [16–37] for starting and ending points of the coding regions. Frequencies, probabilities and entropies were calculated by means of an original computer program. Further computations and testing of the significance were

performed by using commercially available software ('QuattroPro' and 'StatGraphics').

3. Results

The entropies of protein coding genes have been calculated, compared and related to the information content and biological significance. The analysis covered the total of 42,353 base pairs (bp) organized in 20 sequences. The average length of a sequence was 2,117 bp ranging from 595 bp to 5,227 bp. All sequences were protein coding genes from Escherichia coli and together they represented approximately 1% of its genome.

The average entropy of a complete gene expressed in binary units, calculated with a nucleotide as a code unit was 1.99225 bits per nucleotide, ranging from 1.97364 bits to 1.99992 bits. The entropy of the coding region was only slightly lower (Table 1), the difference being statistically insignificant ($p_t = 0.20$, $p_{MW} = 0.13$).

With triplets as code units, the entropies of biologically functional sequences were compared to the entropies of frameshifted sequences. The obtained results (Table 1, Fig. 1) were in agreement with the intuitive expectation that the sequences containing biological information should have a lower entropy than the frameshifted sequences which lack that information. Observed differences (average of 0.21474 bits and 0.28872 bits between entropy of the reading frame and frameshifts +1 and −1, respectively) were found to be statistically significant. Both Student's test and Mann–Whitney's test estimated the probability of null (i.e. no difference) hypothesis fairly below one percent ($p_t = 1.5\ 10^{-4}$ and $p_{MW} = 3\ 10^{-4}$ for the difference between the reading frame and the frameshift +1; $p_t = 4.5\ 10^{-6}$ and

Table 1

Entropy values (mean, standard deviation, minimum and maximum value of S) in bits per code unit calculated for protein coding genes from *E. coli* with nucleotides, triplets and amino acids as code units

Code unit	$\bar{S}$	SD(S)	Min(S)	Max(S)
Nucleotides				2.00000
(a) All	1.99225	0.00922	1.97364	1.99992
(b) Coding	1.98967	0.00982	1.95919	1.99921
(a)−(b)	0.00292			
Triplets				6.00000
(a) Correct frame	5.30943	0.20737	4.78985	5.62513
(b) Frameshift +1	5.52416	0.12720	5.20989	5.69535
(c) Frameshift −1	5.59815	0.14149	5.19701	5.72021
(b)−(a)	0.21474			
(c)−(a)	0.28872			
Amino acids				4.39232
(a) Correct frame	4.09358	0.10029	3.84000	4.21102
(b) Frameshift +1	3.95859	0.08588	3.73490	4.11519
(c) Frameshift −1	4.01419	0.06463	3.88908	4.15887
(b)−(a)	−0.13499			
(c)−(a)	−0.07939			

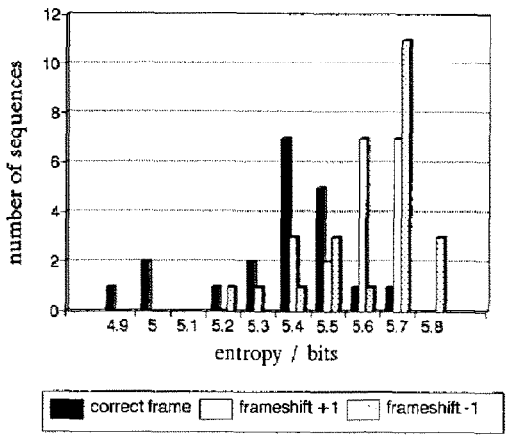

Fig. 1. Distribution of entropies calculated with triplets as code units, assuming correct and frameshifted reading of *E. coli*'s DNA sequences.

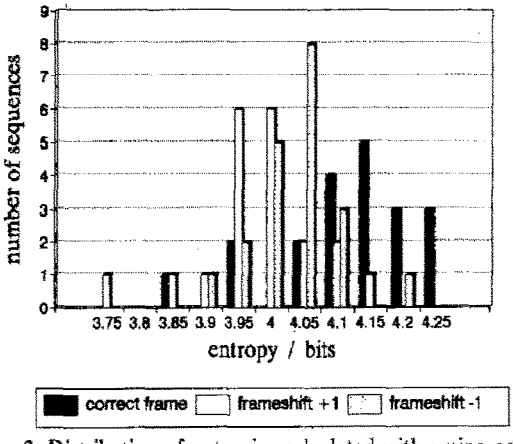

Fig. 2. Distribution of entropies calculated with amino acids as code units, assuming correct and frameshifted translation of *E. coli*'s DNA sequences.

$p_{MW} = 2\ 10^{-5}$ for the difference between the reading frame and the frameshift -1).

The analogous analysis with amino acids as code units (Table 1, Fig. 2) gave opposite results: Protein sequences which are biologically functional and therefore contain some evolutionary information have shown higher entropies than protein sequences which are products of the frameshift and hence lack any biological information and function. The calculated differences (Table 1) are also highly significant: $p_t = 2.5\ 10^{-5}$ and $p_{MW} = 2\ 10^{-4}$ for the difference between the reading frame and the frameshift $+1$; $p_t = 2.5\ 10^{-3}$ and $p_{MW} = 2\ 10^{-3}$ for the difference between the reading frame and the frameshift -1.

4. Discussion

Results of this analysis support the idea that entropy calculated according to Boltzmann's formula could be used as an indicator for the functional information in DNA. Entropy difference between coding and noncoding regions, calculated with a nucleotide as code unit, was only of marginal significance ($p_t = 0.2$, $p_{MW} = 0.13$). Since scarcity of extensive noncoding regions in *E. coli*'s genome has made it impossible to calculate the exact entropy of noncoding DNA, entropies of coding regions could have been compared only to the entropies of complete mRNA sequences (leader, coding region and trailer), and such an inadequate comparison (coding region was partially compared to itself) might explain the lack of significance.

The results obtained with triplets as code units seem to be unequivocal in stating the statistical significance of the difference between entropies of the correct reading frames and entropies of the frameshifted sequences leaving no dispute about the adequacy of entropy as an indicator for the biologically functional information of DNA. It is quite essential that these results were obtained with triplets as the code units because reading in triplets is the way the nature itself decodes the information from DNA. This resemblance reinforces entropy as an indicator for the functional information in DNA.

Interesting results were obtained when entropies were calculated with amino acids as code units, i.e. *after* translation from DNA to proteins. The translation according to the genetic code results in an inversion of the relationship between entropies of biological sense and nonsense. The protein sequence originating from the correct reading frame, thus representing the biologically functional polypeptide, shows a higher entropy than the polypeptides generated through frameshifted translations which are a biological nonsense. This finding seems paradoxical: How can a molecule which obviously contains definite information (a biologically functional protein) have a higher entropy than sequences which lack any biological sense (sequences produced by frameshift)? We suggest a hypothesis which may help to explain the paradox.

Protein sequences, functional and frameshifted, are translated from DNA sequences according to the genetic code. The present genetic code is a product of billion years of evolution and its redundancy provides many favorable qualities. Some of them are well known, e.g. suppressing the nonsense mutations (mutations to stop codon) and propagation of the same sense and neutral mutations, and some are still to be discovered, maybe the tendency to retain information on both strands of DNA [38]. The ability of lowering information content during translation is perhaps another intrinsic property of the genetic code. If this were true, it would enable bacteria to suppress frameshift mutations by lowering the entropy of proteins produced by frameshift. This would make the correct reading frame the most probable and thus propagate its translation to the corresponding protein.

Our hypothesis that genetic code may have the ability for some kind of 'hiding' information during translation might seem, unexpectedly strange, but if we consider other, equally strange though already established facts (e.g. the correlation of the expression of most proteins with the fitness of the codon anticodon pairing and its consecutive determination by the codon usage [39–41]; claims on structural complementarity between proteins

coded on complementary strands of DNA [42–45]; suggestions that genetic code redundancy is in fact 'optimal utilization of 64 codons for the purpose of retaining information in two complementary strands of DNA' [38], etc.), we cannot reject the possibility that lowering the information content during the translation might be another remarkable property of the genetic code.

Acknowledgments

The authors are grateful to Prof. Vl. Simeon for hours of stimulating discussions and helpful suggestions on the manuscript. We also thank Prof. M. Flögel-Mršić for extensive help in preparing the revised version of the manuscript.

References

1 K.R. Popper, Nature 207 (1965) 233.
2 W. Büchel, Nature 213 (1967) 319.
3 K.R. Popper, Nature 213 (1967) 320.
4 H.W. Woolhouse, Nature 213 (1967) 952.
5 H.W. Woolhouse, Nature 216 (1967) 200.
6 A.R. Peacocke, The physical theory of biological organization (Oxford University Press, Oxford, 1989).
7 C.E. Shannon and W. Weaver, The mathematical theory of communication (The University of Illinois Press, Urbana, IL, 1949).
8 L. Brillouin, Science and Information Theory (Academic Press, New York, NY, 1962).
9 G.C. Theoridis and L. Stark, Nature 224 (1969) 860.
10 L. Gatlin, Information theory and the living system (Columbia University Press, New York, NY, 1972).
11 J.S. Wicken, J. Theor. Biol. 72 (1978) 191.
12 J.S. Wicken, J. Theor. Biol. 77 (1979) 349.
13 M. Eigen, Naturwissenschaften 58 (1971) 465.
14 J.A. Wilson, Nature 219 (1968) 534.
15 H. Atlan, in: Autopoiesis, a theory of living organisation, ed. M. Zeleny, Hierarchical self-organization in living systems: noise and meaning (North Holland, Amsterdam, 1981).
16 M. Hartlein, Nucl. Acid. Res. 15 (1987) 1005.
17 R. Freedman, B. Gibson, D. Donovan, K. Biemann, S. Eisenbeis and J. Parker, J. Biol. Chem. 260 (1985) 10063.
18 H. Cudny, J.R. Lupski, G.N. Godson and M.P. Deutcher, J. Biol. Chem. 261 (1986) 6444.
19 T. Keng, T.A. Webster, R.T. Sauer and P. Schimmel, J. Biol. Chem 257 (1982) 12503.
20 T.A. Webster, B.W. Gibson, T. Keng, K. Biemann and P. Schimmel, J. Biol. Chem. 258 (1983) 10637.
21 J. Bouvier, J.C. Patte and P. Straiger, Proc. Natl. Acad. Sci. U.S.A. 81 (1984) 4139.
22 G. Branlant and C. Branlant, Eur. J. Biochem. 150 (1985) 61.
23 W.C. Herlihy, N.J. Royal, K. Biemann, S.D. Putney and P.R. Schimmel, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 6531.
24 R. Breton, H. Sanfacon, I. Papayannopoulos, K. Biemann and J. Lapointe, J. Biol. Chem 261 (1986) 10610.
25 F. Yamao, H. Inokuchi, A. Cheung, H. Ozeki and D. Söll, J. Biol. Chem 257 (1982) 11639.
26 C.V. Hall and C. Yanofsky, J. Bacteriol. 148 (1981) 941.
27 C.V. Hall, M. vanCleemput, K.H. Muench and C. Yanofsky, J. Biol. Chem. 257 (1982) 6132.
28 F. Dardel, G. Fayat and S. Blanquet, J. Bacteriol. 160 (1984) 1115.
29 P.K. Chanda, M. Ono, M. Kuwano and H. Kung, J. Bacteriol. 161 (1985) 446.
30 F.R. Mooi, M. van Buuren, G. Koopman, B. Roosendaal and F.K. de Graaf, J. Bacteriol. 159 (1984) 482.
31 J. Piette, H. Nyunoya, C.J. Lusty, R. Cunin, G. Weyens, M. Crabeel, D. Charlier, N. Glansdorff and A. Pierard, Proc. Natl. Acad. Sci. U.S.A. 81 (1984) 4134.
32 H. Nyunoya, C.J. Lusty, Proc. Natl. Acad. Sci. U.S.A. 80 (1983) 4629.
33 C. Debouck, A. Riccio, D. Schumperli, K. McKenney, J. Jeffers, C. Huges and M. Rosenberg, Nucl. Acid. Res. 13 (1985) 1841.
34 M. Brune, R. Schumann and F. Wittinghofer, Nucl. Acid. Res. 13 (1985) 7139.
35 P. Stragier, O. Danos and J.C. Patte, J. Mol. Biol. 168 (1983) 321.
36 P. Stragier and J.C. Patte, J. Mol. Biol. 168 (1983) 333.
37 F. Valle, E. Sanvicente, P. Seeburg, A. Coyarrubias, R.L. Rodrigez and F. Boliva, Gene 23 (1983) 199.
38 J.E. Zull and S.K. Smith, Trends Biochem. Sci. 15 (1990) 257.
39 P.W. Finch, A. Storey, K. Brown, I.D. Hickson and P.T. Emmerson, Nucl. Acid. Res 14 (1986) 8583.
40 H. Grosjean and W. Fiers, Gene 18 (1982) 199.
41 M. Gouy and C. Gautier, Nucl. Acid. Res. 10 (1982) 7055.
42 J.E. Blalock and E.M. Smith, Biochem. Biophys. Res. Commun. 121 (1984) 203.
43 K.L. Bost, E.M. Smith and J.E. Blalock, Proc. Natl. Acad. Sci. U.S.A. 82 (1985) 1372.
44 J.J. Mulchahey, J.D. Neill, L.D. Dion, K.L. Bost and J.E. Blalock, Proc. Natl. Acad. Sci. U.S.A. 83 (1986) 9714.
45 V.P. Knutson, J. Biol. Chem 263 (1988) 14146.