studia biophysica, Berlin, Band 71, 1978, Heft 3, S. 173 - 180
Printed in the German Democratic Republic

ON THE COMPLEXITY OF CYTOCHROME C AND THE INFLUENCE OF THE GENETIC CODE

W. EBELING, R. MAHNKE
Sektion Physik, Wilhelm-Pieck-Universität, 25 Rostock

The complexity of 22 cytochrome c sequences is analyzed. The calculations are based on the THIELE-HEINZ subword complexity. This mathematically well-defined quantity may serve as a measure to examine the information content of the proteins. The calculations show that the complexity values of real proteins are considerably smaller than that of randomized protein sequences. It is shown that the complexity distribution of real proteins is very near to that of artificial protein sequences which were generated from random DNA-sequences by the genetic code.

1. Introduction

Since 1838 (the year BERZELIUS made the suggestion to designate proteins by that terminus) natural scientists have focused a great deal of research work on proteins as an essential part of all living organism and thus partaking in all cell processes /12/. Because proteins are synthesized under direct control of the information stored in the DNA they play an important role in information transfer. Note that the information stored in a protein is given by the number and arrangements of the 20 amino acids only. If we know the sequence of amino acids in a given protein we are able to calculate the information content of that protein. Following the general ideas of KOLMOGOROV /8/ we introduce now the term "complexity of a sequence" as a measure of the quantity of information stored in the sequence. In earlier papers was proposed to apply the concept of complexity to the analysis of biopolymers /3,4/. Among the different possibilities to measure the complexity of a sequence we choose here the subword complexity proposed by THIELE and HEINZ /6,10/ This complexity is a mathematically precise measure to examine the randomness or non-randomness of sequences, the algorithm given in the following paragraph is very simple.

It is the purpose of the present paper to apply this theory to the cytochrome c family to give values of the complexity for

different protein sequences. We want to point out that the calculations of the complexity require only the knowledge of the biopolymer sequence data. We believe that for the quantitative analysis of sequence data this quantity could play a similar role as the genetic distance, a concept which was proved to be very useful for the construction of phylogenetic trees /1/.

After giving the algorithm for calculating the complexity of sequences based on the subword property (chapter 2) we present the complexity values for 22 cytochrome c sequences of 104 sites and discuss their distribution in chapter 3. We investigate the result that the complexity of arbitrarily randomized sequences of length 104 is greater than that of real proteins. This existing structure in the cytochrome c proteins is connected with the reality of the genetic code. In chapter 4 it is shown that randomized DNA-sequences of length 3×104 translated by the genetic code to protein sequences of length 104 yield nearly the same complexity as that of natural cytochrome c proteins.

2. Subword complexity of sequences

Various complexity measures have been suggested for different areas /10/. We use here a very simple but sensitive measure for the calculation of the complexity of sequences (e.g. proteins, DNA). This algorithm, described in detail elsewhere /6/, /9/, is based on the subword property of sequences. In our notation n is the number of elements or building stones (in our case n = 20 because there are 20 different amino acids) and l(q) is the length of a sequence q (in our case l(q) = 104). The complexity of a sequence q is defined as follows

$$K(q) = \left(2 \cdot \sum_{i=1}^{\ell(q)} A_i(q, \ell(q), n) - \ell(q)\right) / \ell(q) \qquad (1)$$

with

$$A_i(q, \ell(q), n) = \{\, p \mid p \text{ in } q \text{ and } \ell(p) = i \,\} \geq 1$$

A_i $(i = 1, 2, \ldots, \ell(q))$ is the number of subwords or parts of q with length i. For example: q = CCD, l(q) = 3; $A_1 = \{C, D\}$, $A_2 = \{CC, CD\}$, $A_3 = \{CCD\}$ and $K(q) = \frac{2}{3} \cdot 5 - 1 = 7/3$. To give values for the upper and lower bound of K(q) we use an estimation for $A_i(q, \ell(q), n)$. After short calculations we get the following inequality

$$1 \leq K(q) \leq \frac{2}{\ell(q)} \sum_{i=1}^{\ell(q)} \mathrm{Min}\{\ell(q) + 1 - i, n^i\} - 1 \qquad (2)$$

If the number of elements n is greater than the length of the sequence $l(q) < n$ we obtain from eq.(2) the inequality

$$1 \leq K(q) \leq l(q) \qquad (3)$$

This means that a totally regular sequence has a complexity $K(q) = 1$ and on the other hand a "perfectly random string" a complexity of $K(q) = l(q)$. For every real sequence we get a value of $K(q)$ which fulfils eq.(3).

In our case n = 20, $l(q) = 104$, and therefore $l(q) > n$, after using eq.(2) we obtain a similar inequality

$$1 \leq K(q) \leq l(q) - 2\,\frac{l(q) - n}{l(q)} \qquad (4)$$

After inserting the values of n and $l(q)$ in eq.(4) we get

$$K_{max} = 102.3846 \qquad (5)$$

as maximum (upper limit) of the subword complexity of sequences with 104 sites, or in other words $K_{max} = 98.4467\%$ if $l(q) = 104$ equals 100%.

Distinguishing different sequences it is obvious that only protein sequences with complexity values nearby K_{max} have the chance to act as stable sources of all the information which is necessary to guarantee the structure and function of living organism. Note that for simplicity we represent $l(q) \cdot K(q)$, instead of $K(q)$ in the following figures, because we get then natural numbers, e.g. $l(q) \cdot K_{max} = 10648$.

3. Complexity values of 22 cytochrome c sequences

Using the basic equation (1) which defines the complexity we are able to calculate the values for given protein sequences. The calculations were performed using a computer program (details of the program are given in /9/). It takes about 30 sec. (on ESER 1040) to compute one complexity value of a sequence of 104 sites.

TABLE 1 shows the results for a large variety of organism. Several things should be noticed. All cytochrome c sequences which are investigated have a high complexity value or information content which is quite near to the theoretically possible maximum 10648. There are very small differences in the values for the different organism. Because of these small differences there exists evidently no order of rank. We point out that all living organism investigated here have about the same complexity of cytochrome c proteins. Taking the mean value $\bar{K}$ (arithmetical ave-

rage) we get

$$\bar{K} = 101.9960 \quad or \quad 98.073\% \qquad (6)$$

TAB. 1 Complexity values of 22 cytochrome c sequences

Organism	Complexity (eq.(1)) absolute values	per cent	Complexity Length K(q) * l(q)
1. Human	102.0000	98.077	10608
2. Chimpanzee	102.0000	98.077	10608
3. Rhesus monkey	102.0000	98.077	10608
4. Horse	101.9615	98.040	10604
5. Donkey	101.9808	98.058	10606
6. Bovine	101.9615	98.040	10604
7. Pig	101.9615	98.040	10604
8. Sheep	101.9615	98.040	10604
9. Dog	101.9231	98.003	10600
10. Rabbit	102.0192	98.095	10610
11. California graywhale	101.9808	98.058	10606
12. Kangaroo	101.9038	97.984	10598
13. King penguin	102.0385	98.114	10612
14. Chicken	102.0769	98.151	10616
15. Turkey	102.0769	98.151	10616
16. Pigeon	102.0577	98.132	10614
17. Peking duck	102.0192	98.095	10610
18. Snapping turtle	102.0000	98.077	10608
19. Rattle snake	101.8654	97.947	10594
20. Bullfrog	102.0769	98.151	10616
21. Puget sound dogfish	102.0385	98.114	10612
22. Pacific lamprey	102.0192	98.095	10610

The protein sequence data are taken from /2/.

Comparing eq.(5) and eq.(6) we state the divergence of the complexity values of Table 1 from the maximum of complexity. In other words the proteins are not totally random, a small degree of non-randomness (regularity) exists. This fact supposed by several authors /5/ involves especially the influence of the genetic code. The present-day genetic code generates a certain degree of structure in the protein sequences. In the last paragraph this subject is investigated.

4. Influence of the genetic code

A glance at the genetic code shows that the 64 triplets are not distributed at random among the 20 amino acids and that there are three nonsense or chain terminating codons /5/. The frequency of a protein j is given by the following formula

$$P_j = r_j \cdot s_i \qquad (7)$$

with r_j = number of codons coding for amino acid j
(r_j = 1,2,3,4 or 6)
s_i = frequency of codon i.

Assuming the equality of base frequency in the DNA s = 1/4 (that means a base composition of 50% C + G), we fix s_i = 1/61 for every codon. The calculated frequencies are presented in FIG. 1 (dash-dot line).

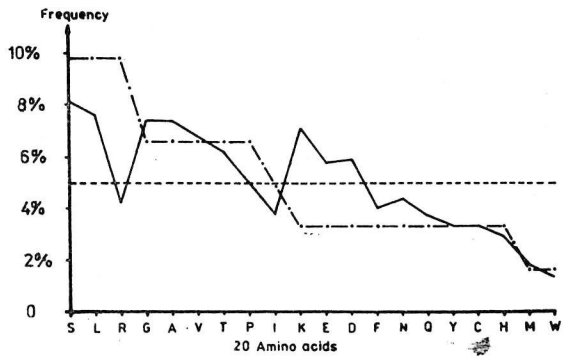

FIG.1 Frequencies of amino acids (abbreviations are taken from DAYHOFF /6/) under the influence of the genetic code
- · —— · —— theoretical line after eq.(7)
———— experimental line after KING, JUKES /9/
The straight line (— — —) shows equal frequency for every amino acid.

Compared with it the solid line of FIG.1 shows the experimental results after KING, JUKES /7/. They give the number of occurrences of the amino acids among 1492 amino acid residues in 53 vertebrate polypeptides. These curves reflect the influence of the genetic code on the frequency of amino acids.

In contrast, we consider now the random protein model with equal frequencies p_j = 1/20 for every amino acid (see dashed line in FIG.1). Using a random number generator which generates the 20 amino acids with equal probability we construct 100 amino acid sequences of length 104. After calculating the complexity of each random protein we get a frequency distribution of proteins via complexity. The result is shown in FIG.2 (dash line — —x— —x—).

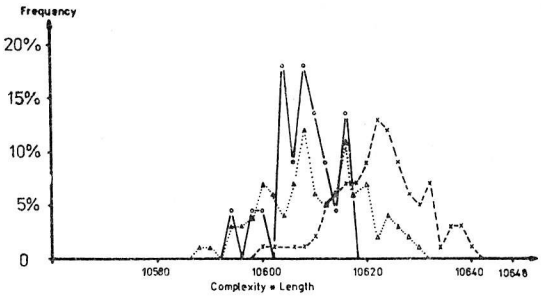

FIG.2 Comparison of frequency distributions via complexity for
a) 22 cytochrome c proteins (after TABLE 1) ——o——
b) 100 random proteins with equal frequencies for every amino acid ——x——
c) 100 proteins with frequencies generated by the genetic code ...Δ...

The complexity values of arbitrarily randomized protein sequences are significantly higher than the values of real cytochrome c sequences, their frequency distribution is also given in FIG.2 (solid line ——o——). For the mean value (arithmetical average) of the random protein model (RPM) we get

$$\bar{K}_{RPM} = 102.1388 \quad or \quad 98.210\,\% \tag{8}$$

Compared with the mean complecity of cytochrome c (see eq.(6)) the divergence of the values is remarkable,

$$\bar{K} < \bar{K}_{RPM} \tag{9}$$

Finally we consider the random nucleotide model (RNM). We assume for simplicity equal occurrence of the nucleotide bases in 100 random generated DNA sequences of length $3 * 104$. After translating these random sequences into proteins by means of the genetic code and calculating their complexity we get the result presented in FIG.2 (dot line ...Δ...). The remarkable result is that this frequency distribution and its mean value

$$\bar{K}_{RNM} = 102.0215 \quad or \quad 98.079\,\% \tag{10}$$

concide quite well with the result for natural cytochrome c proteins. Note that modelling the translation process of DNA into protein by the genetic code is identical with the generation of a protein with the unequal amino acid distribution of FIG.1 (dash-dot line). Because of the degeneracy of the genetic code, that is that amino acid occurrences are not equal, the proteins have a

certain degree of structure /11/, the complexity has not its maximum value.

In conclusion we summarize the main results. The subword complexity appears to be a new quantitative characteristics of the complexity or information content of real biopolymer sequences which may serve for quantitative comparisons of real proteins or nucleotid sequences. In order to give an example we presented here an analysis of 22 cytochrome c sequences. The results show that the complexity distribution is very near to that of artificial protein sequences which were generated from random DNA sequences by the genetic code. Of course this must not be interpreted as the result of a random process originating the real DNA sequences which generated the real cytochrome c sequences. On the contrary, the real DNA sequences are clearly the result of a long evolution process and store the whole information necessary to maintain the structure and function of living organism. But on the other hand, from the point of view of the mathematical analysis the DNA sequences which produced the present day cytochrome c sequences have evidently random structures, i.e. maximal complexity in the sense given above. They seem to be very near to the aperiodic crystals which were predicted by SCHROEDINGER. Evidently nature uses sequences of high complexity in order to store a maximum of information. The subword complexity discussed here reflects the quantity of information stored in a sequence but it is of course unable to describe the quality of the information stored in a biopolymer.

References

/1/ BEYER, W.A. et al., Math. Biosci. 19 (1974) 9
/2/ DAYHOFF, M.O., "Atlas of Protein Sequences and Structure", Silver Spring (1969)
GEISSLER, E. (Ed.), "Kleine Enzyklopädie Leben", Leipzig (1976)
/3/ EBELING, W., FEISTEL, R., JIMENEZ MONTANO, M.A., Rostocker Physikalische Manuskripte 2 (1977) 105
/4/ EBELING, W., MAHNKE, R., to be published in "Problems of Contemporary Biophysics"
/5/ GATLIN, L.L., J. Mol. Evol. 3 (1974) 189
YOCKEY, H.P., J. theoret. Biol. 67 (1977) 345, 377
/6/ HEINZ, M., Elektr. Inform. u. Kybern. 13 (1977) 27
/7/ KING, J.L., JUKES, T.H., Science 164 (1969) 788
/8/ KOLMOGOROV, A.N., Probl. Peredachi Inform. 1 (1965) 3
/9/ MAHNKE, R., Wiss. Z. WPU Rostock (in press)

/10/ THIELE, H., Nova acta Leopold. NF 27/1 (1972) Nr. 206
THIELE, H., in "Organismische Informationsverarbeitung" (Hrsg.: F. Klix), Berlin (1974)
/11/ YANO, T., HASEGAWA, M., J. Mol. Evol. 4 (1974) 179
/12/ ZWILLING, R., Umschau 78 (1978) 170

Сложность 22-х последовательностей белка цитохрома *C* анализируется. Вычисления базируются на определении сложности на основе частных слов по Тиле-Хейнцу. Эта математическая величина служит мерой количества информации в белках. Вычисления показывают, что сложность реальных белков меньше случайных последовательностей. Показывается, что распределение сложностей реальных белков близко к искусственным последовательностям, которые были генерированы из случайных ДНК-последовательностей с помощью генетического кода.

Eingegangen am 17. 8. 1978

Prof. W. EBELING, Wilhelm-Pieck-Universität Rostock, Sektion Physik, Universitätsplatz 3, DDR 25 Rostock