The third nucleotide of the Gly coding triplet remembers the periodicity of the collagen chain

V.Ju. Makeev*, V.G. Tumanyan, N.G. Esipova

V.A. Engelgardt Molecular Biology Institute RAS, Moscow, Russian Federation

Received 20 April 1995

Abstract Collagen is a fibrous protein with a primary structure with complex periodical features. We show using symbolic Fourier transform of the collagen cDNA sequence that basic periodical patterns appear there also. Strikingly they are present in the third position of triplets encoding Gly, which occupies each third position in the sequence of the protein, and to which selection on the protein level does not applied. Thus, the gene of collagen seems to appear due to pra-gene multiplication.

Key words: Collagen; Gene duplication; Sequence analysis; Sequence Fourier transform

1. Introduction

The fibrillar protein collagen plays a vital part in body structure of animals [1]. In addition to structural and bearing functions collagen is e.g. XIII platelet factor [2].

The so-called large periods in collagen first discovered on electron microscopic and low angle X-ray photographs and later identified in the primary structure of the protein are considered to be classical results. The length of basic period is 234 amino acid residues [3].

This largest period reflects the fundamental periodicity of the molecule and predetermines the molecules aggregation. Many periods – D divisors such as D/13, D/6, D/5 and other were found by different researchers (see e.g. [4,5]).

As primary structures of collagen were determined Fourier and correlation analysis were used for studying periodical patterns being limited to amino acid sequences (see e.g. [5,6]). One of the interesting attempts to explain the origin of collagen periodical structure was the reduplication of a gene precursor.

Collagen may serve as a unique object for a study of gene multiplication processes in coding sequences. Indeed, since every third amino acid in collagen protein is Gly, the Gly-coding triplets of the collagen gene does not undergo any selection induced from the protein level. Thus, every third nucleotide in Gly codons depends only on the factors reflecting DNA environment and gene history.

It should be noted, that despite long and broad study of periodical patterns in collagen, very little is known about the function of these patterns. Most researchers regard them as a requirement for constructing a specific collagen structure, which until now is not known in detail. However, sequences of collagen genes, which appeared lately, provide us with the opportunity to study the other possible reason of the periodical structure of the collagen, namely the history of gene evolution. We undertook this study in order to draw the attention to the fact, that periodical patterns may appear due to both reasons, and it seems, that the basic period of 234 amino acid arises from genetic factors.

*Corresponding author.

2. Materials and methods

Recently, Silverman et al. [7,8] suggested a reasonable way of solving the problem of transforming the symbolic sequence into the set of digital sequences to which the Fourier transform may be implemented. Thus, the powerful Fourier method of the periodical patterns search now may be used to investigate the structure of native DNA sequences. In our recent paper [9] we show, how the Fourier transform may be used to study structure of proteins.

For practical usage we developed the following algorithm. For each possible period the sequence was cut to the largest possible length multiple to the integer number of periods and closed to the ring. After that the auto-correlation function was calculated on the ring, i.e. the number of identical letters at the given distance in dependence on that distance:

$$R(k) = \frac{1}{\sqrt{N}} \sum_{i=0}^{N} \Delta a_i a_{i+k} \bmod N$$

The Fourier transform of this function serves as a characteristic of periodicity's power. The closing of the sequence into the ring of different length allows to put all the periods into a resonance with the length of the sequence. This serves two purposes. On the one hand, the periods with the length comparable to the length of the whole sequence can be resolved, e.g. the period D in the collagen, which is repeated only four times on the length of the molecule; on the other hand the powerful periods such as periods of 3 and 9 bp in the nucleic acid sequence of the collagen, related to the strict Gly repeat, do not contribute to the power of other periods and do not interfere with the final picture.

Thus the method developed provides the possibility of studying sequences containing both strict repeats and disperse periodical patterns. This allowed us to study periodical distribution of nucleotides in Gly-coding triplets. Note, that till now researchers who used similar approaches were urged to throw Gly's out of the sequence (see e.g. [5]).

Naturally our approach allows to reproduced results obtained for the amino acid sequences; to check this we reproduced in several series of calculations results published in [4–6].

The main purpose of our work was to study periodicities in nucleic acid sequence of the collagen cDNA. We have studied the most representative among sequences deposited in EMBL bank (HSCOL3A1R). It is α_1 chain of human collagen (III).

Chain α_1 is of the greatest interest for studying because this chain is represented in the collagen molecule either two (I) or three (III) times.

We studied sequences composed from the first, second and third nucleotides in each codon (Figs. 1–3). The sum gives evidently a full-power spectrum of the nucleic acid sequence. One can see that in spite of a not very clear picture a number of main periods characteristic for the amino acid sequence are present. At the same time periods characteristic only for the DNA sequence are also represented.

3. Results and discussion

It can be seen that power spectra of the sequences corresponding to the first and second positions (Figs. 1,2) are alike, and similar to the spectrum of amino acid sequence of the protein [4]. Whereas the spectrum of the sequence composed

SSDI 0014-5793(95)00490-4

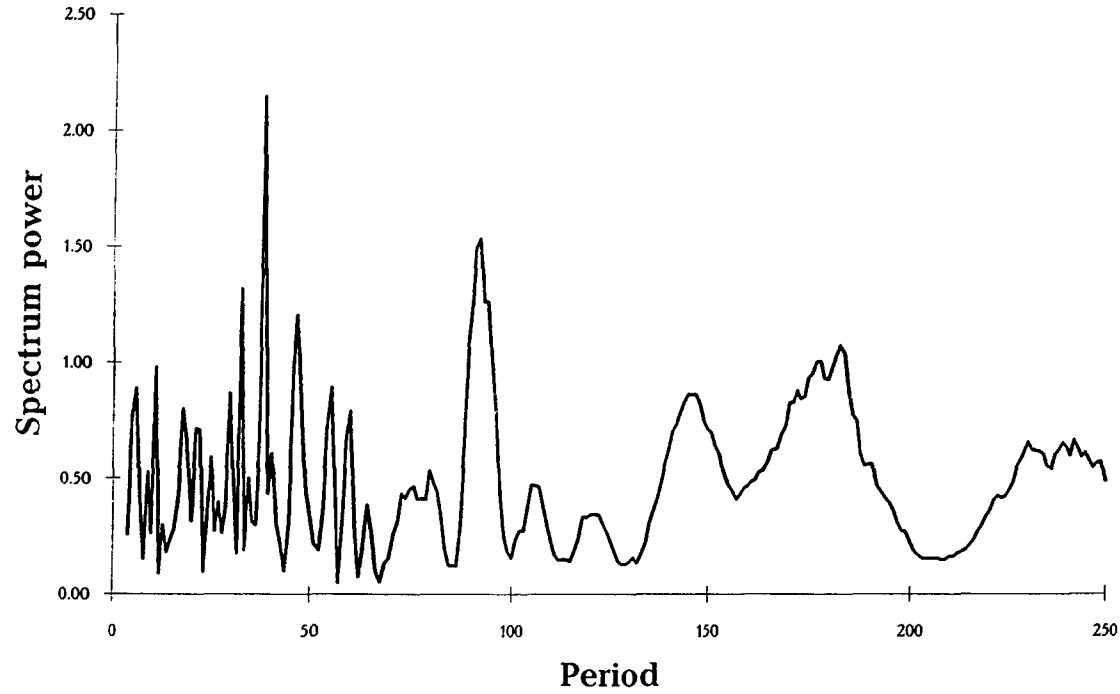

Fig. 1. The power spectrum of the sequence composed from nucleotides occupying first positions of triplets in collagen DNA of α_1(III) chain. Maximum corresponds to the period of 39 triplets equal to D/6. D-period not revealed.

from the nucleotides occupying the third position differs essentially from the previous two. Note, that period D can be seen only on the spectrum of the sequence composed from the nucleotides occupying the second position in codon. This agrees with a point of view that the second position of the code influence most the chemical properties of coded amino acid. For example, T in second position determines hydrophobic amino acid (with an exception of the code for Trp, but Trp lacks in collagen).

The spectrum of the sequence composed from the nucleotides located at the third position is distinctively remarkable. It is clearly seen that it incorporates two different regions, i.e. long and short periods (Fig. 3). The most prominent peak corresponds to D period in collagen.

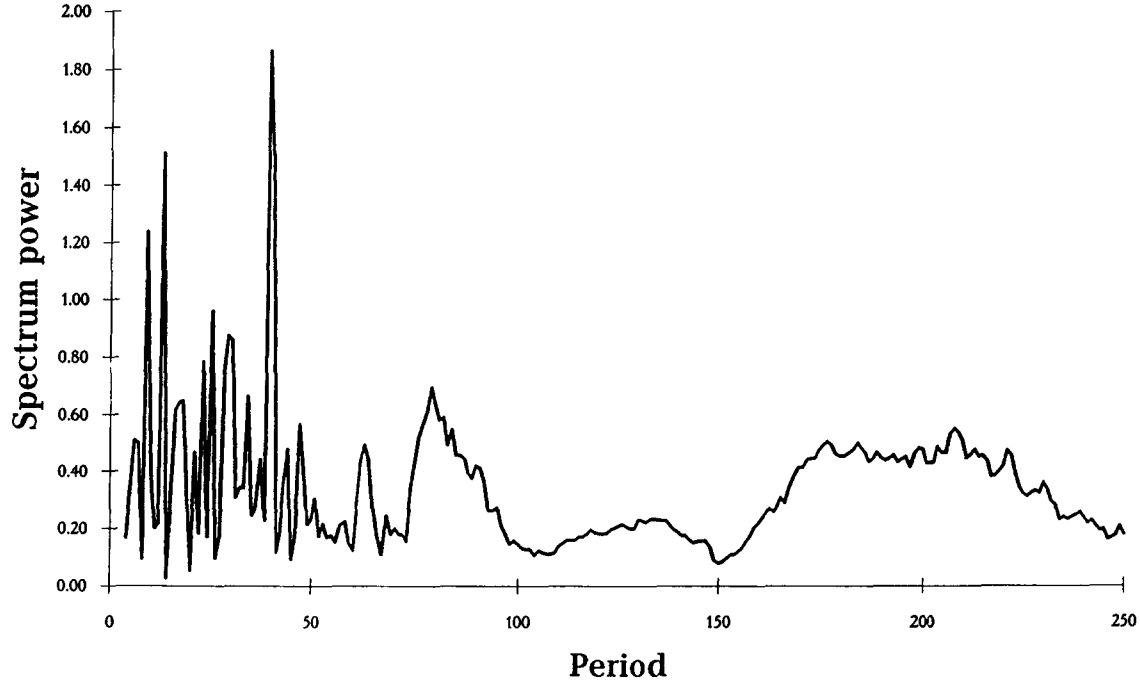

Fig. 2. The corresponding power spectrum for second positions of the triplets. Maximum is the same (D/6). D-period is rather weak.

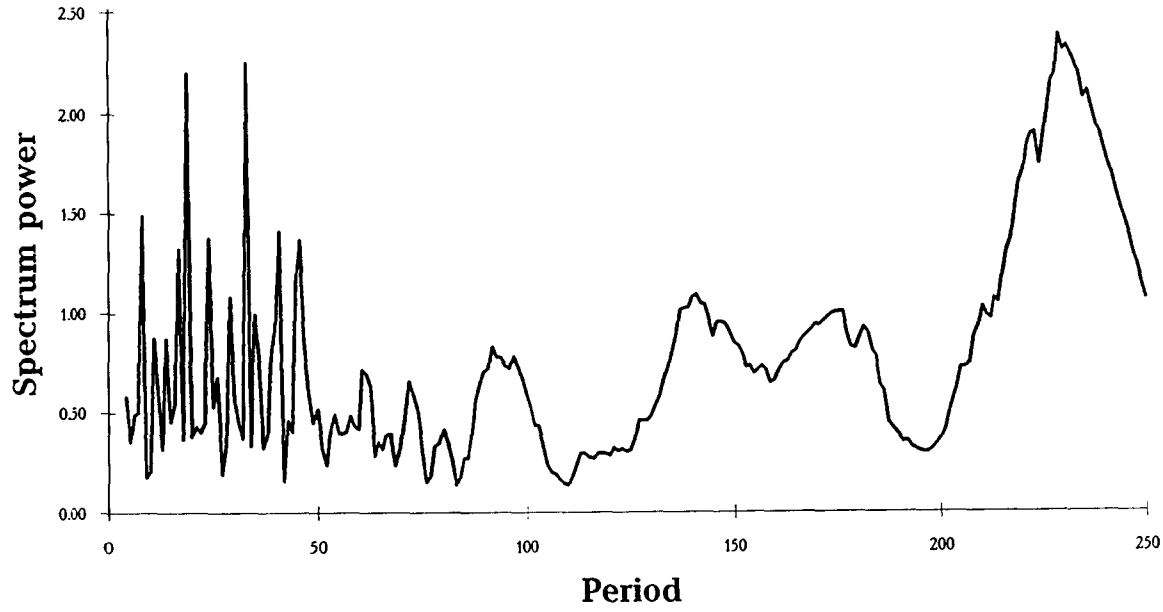

Fig. 3. The same for the third positions in triplets. Maximum in the long periods region (232) is close to D (234). In a short period region one can observe a regular system of the periods which has no analogy to the amino acid sequence (8, 19 and 33 (three) nucleotides). Note that prevailing nucleotides in the third position are A and T, which in the third position interchange either without changing the amino acid at all, or without essentially changing its properties (I–L, etc.).

Thus, it is not only that period D has been found for the first time on the nucleic acid sequence but it has been found on the low-significant or non-significant code position, being visible there clearer then on the sequences composed from the bases occupying significant positions.

If this is the case, then the result obtained unambiguously indicates that the long periodical structure of the collagen originates from a duplication of the main D-period, and so the corresponding periodicity is not a result of the selection on the amino acids level, caused by the necessity to provide precise packing of the molecules into the structures of higher order [10].

In order to finally test the conclusion made, we undertook an *experimentum crucis*. We composed a sequence formed by nucleotides located at the third position of the triplets coding Gly, which is every third amino acid in a collagen. This nucleotide is obviously insignificant one, but the most important thing here is that the selection at the amino acid level surely does not take place here. As is shown in Fig. 4 in the sequence

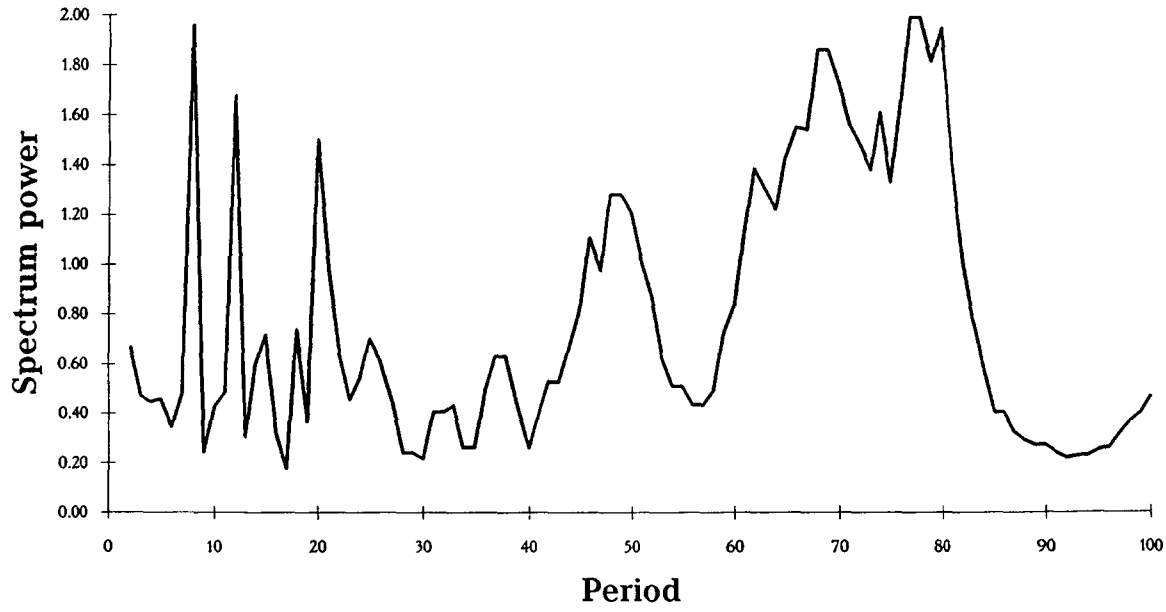

Fig. 4. The power spectrum of the sequence composed from nucleotides occupying third positions of triplets coding Gly. Maximum (78 nona-nucleotides) exactly coincides with D. Three prominent peaks in the short range region are originated from correspondent peaks on the Fig. 3.

of the nucleotides located in the third position of the Gly coding codons, one can clearly see the periodicity of 78 glycines, exactly corresponding to D ($234 = 78 \times 3$). Thus a strong argument in favour of the origin of the basic periodicity in α_1 chain in collagen by gene duplication is obtained.

Acknowledgements: This study was partially supported by Russian Human Genom Project, INTAS-93-36-15 and ISF Grant N21000.

References

[1] Ramachandran, G.N. and Reddi, A.H. (1976) Biochemistry of Collagen, Plenum Press, New York.
[2] Saito, Y., Imada, T., Takagi, J., Kikuchi, T. and Imada, Y. (1986) J. Biol. Chem. 291, 1355–1358.
[3] Hulmes, D.J.S., Miller, A., Parry, D.A.D., Piez, K.A. and Woodhead-Galloway, J., (1973) J. Mol. Biol. 79, 137–148.
[4] Hoffmann, H., Fietzek, P.P. and Kuhn, K. (1980). J. Mol. Biol. 141, 293–314.
[5] McLachlan, A.D. (1977) Biopolymers 16, 1271–1297.
[6] Bear, S.R., Adams, J.B. and Poulton, J.V., J. Mol. Biol. 118, 123–126.
[7] Silverman, B.D. and Linsker, R. (1986) J. Theor. Biol. 118, 295–300.
[8] Voss, R., (1992) Phys. Rev. Lett. 68, 3805–3808.
[9] Makeev, V.Ju., Tumanyan, V.G. (1994) Biophysics 39, no. 2, pp. 263–266.
[10] Miller, A., (1982) Trends Biochem. Sci. 7, 13–18.