Search of periodicities in primary structure of biopolymers: a general Fourier approach

V.Ju. Makeev and V.G. Tumanyan

Abstract

We discuss a new convenient way to study periodical patterns in primary structures of biopolymers which appeared recently. For the sequence of a biopolymer the symbolic correlation function is constructed, which is used as a digital sequence thus allowing us to perform a Fourier transform. Another fruitful technical improvement is the closing of the sequence in the ring with further scanning of the ring length, which allows the study of periods of the order of the sequence length. This approach makes it possible to take into account any scores describing similarity between symbols and to compare results obtained using different Fourier-like and correlation matrix techniques. An algorithm to compute Fourier spectrum power allows detection of vague periods in sequences containing strong repeats. A PASCAL program, SYMFOUR, has been written and tested on both sequences with periodical patterns, already reported, and sequences and other sites interesting from a biological point of view.

Introduction

Periodical organisation of primary structures is an inherent property of many biopolymers. The study of periodical patterns in primary structures is fruitful in many ways. For example, it allows detection of repetitive sequences (Bell and Torney, 1993; Kobe and Deisenhofer, 1994) and characteristic sites with periodical structure, such as the leucine zipper (Landschulz et al., 1988). In DNA a periodical structure may indicate protein coding regions. If this is the case the period 3 is very characteristic, associated with triplet organisation of the genetic code (Shepherd, 1981).

The most promising area of periodicity searches appears to be the study of fibrous proteins. Here a very sophisticated system of periods is characteristic (McLachlan and Stewart, 1976; Hofmann et al., 1980). The study of regularities in primary structures of fibrous proteins facilitates understanding of both the molecule itself and the molecular aggregate structure formation.

A variety of approaches for searching for periodical patterns in biological sequences can be reduced to two basic methods, namely Fourier analysis and the study of internal homologies.

However, it seems that an adequate technique for periodical pattern detection so far has not been developed. Indeed, to use an arbitrary physical feature, such as ionic potential (Veljkovich et al., 1985), partial charge of amino acid (McLachlan and Stewart, 1976), hydrophobicity (Hofmann et al., 1980) or any other similar parameter, as a digital data for Fourier transform means that the sequence is examined from a rather arbitrary point of view. These properties seem to be unrelated directly to the sequence of symbols. An attempt to describe a biological sequence with a single digital parameter can give only limited results. Anyway, the comparison of results obtained using various digitizing techniques always presents a problem for this kind of investigation. Hardly any digitizing may be taken for an optimal one.

On the other hand, methods of homology search, developed as applications for alignment, have reached a significant level of sophistication (Gelfand, 1992). Various homology matrices have been constructed and studied comparatively by many researchers. As a result, attempts have been undertaken to seek periodical patterns as internal homologies in a sequence. Despite doubtless success (Hulms et al., 1973), such techniques rank below Fourier analysis in respect of consistency and ease of interpretation. If the system of periods is complex and includes periods multiple to each other, which is the case in fibrous proteins such as collagen (Hofmann et al., 1980), homology matrices become filled rather homogeneously, without clear diagonals. Moreover, little can be done to distinguish multiple periods.

Evidently, the appropriate approach should unify the advantages of both methods currently in use. In a previous short communication (Makeev and Tumanyan, 1994) we suggested a technique that conveniently combines the advantages of Fourier analysis and the internal homology method, which partially relates to the multichannel Fourier analysis method proposed by McLachlan (1993). In this paper we describe a practical study of various primary structures using our approach, together with the method’s further developments.
System and methods

The PASCAL Turbo (version 6.0) program SYMFOUR, with a Turbo Vision interface, has been written to run on IBM-PC compatible microcomputers under the MS-DOS operating system. We used several comparison matrices, such as the Dayhoff matrix, genetic code similarity matrix and several matrices combining amino acids with similar chemical properties. Also, several matrices of a special type were composed especially for reasons of analysing the periodical structure of collagen, such as the matrix of interchange between collagen basic units, described by Holmann et al. (1980). All sequences used were taken from the SwissProt Sequence Database; the cDNA of collagen was taken from the EMBL bank. Pictures were plotted with MS EXCEL 5.0.

Algorithm

The basic idea of the algorithm is clear. While it is difficult to put a digital sequence allowing Fourier transform into correspondence with a given symbolic sequence, it is much easier to construct an autocorrelation function. For this function one or other homology matrix may be readily used. This allows substitution of an arbitrary digitizing procedure by a plainly expressed similarity of symbols.

Let there be a sequence of biopolymer $a(k)$ of length $N$ written in the alphabet $A$. In the case of a nucleic acid $A = \{a, c, t, g\}$, while in the case of proteins $A$ is a set of amino acids. Let $a_1, a_2$ be a pair of letters of such an alphabet. Let the correlation matrix be set, matching with any pair of letters of the alphabet $a$ certain number $L(a, a')$ (for example the Dayhoff matrix for amino acids). In this case the correlation function of general form is determined by the expression

$$S(d) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} L[a(k), a(k+d)]$$

(1)

When $k + d > N$ it should be taken on mod $N$, which corresponds to the closing of the sequence in the ring in Benson style (Benson, 1990).

In the spirit of (McLachlan, 1993) we are going to use as the measure of periodicity the Fourier spectrum of the correlation function (1)

$$\tilde{S}(\omega_q) = \frac{1}{\sqrt{N}} \sum_{k=0}^{n-1} S(k) e^{-i\omega_q k}.$$  

(2)

In the case of sequences of real numbers $x_k$ the Fourier transform

$$\tilde{x}(\omega_q) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} e^{-i\omega_k k} x(k), \quad \omega_q = \frac{2\pi q}{N}, \quad q = 0, \ldots, N - 1.$$  

(3)

links with the Fourier transform of its correlation function

$$J(d) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_k x_{(k+d) \mod N}$$  

(4)

with a relation

$$\tilde{J}(\omega_q) = \sum_{d=0}^{N-1} e^{-i\omega_q d} J(d) = \tilde{x}(\omega_q) \tilde{x}^*(\omega_q).$$  

(5)

It is easy to show (McLachlan, 1993; Makeev and Tumanyan, 1994), that the symbolic correlation function is linked with the ‘Fourier transform of the sequence’ by a similar relation. Construct for each letter $\alpha$ from $A$ a characteristic sequence $I^\alpha(k)$, in which 1 represents numbers occupied in the studied sequence by the given letter and 0 all other sites. Then the Fourier transform of our symbolic correlation function (2) links with the Fourier transforms of characteristic sequences

$$\tilde{I}(\omega_q) = \frac{1}{\sqrt{N}} \sum_{k=0}^{n-1} I_k e^{i\omega_q k}$$  

(6)

with relation

$$\tilde{S}(\omega_q) = \sum_{\alpha, \beta \in A} L(\alpha, \beta) \tilde{I}^\alpha(\omega_q) \tilde{I}^\beta(\omega_q).$$  

(7)

In the case of a single correlation matrix the formula is much simpler

$$\tilde{S}(\omega_q) = \sum_{\alpha \in A} \tilde{I}^\alpha(\omega_q) \tilde{I}^{\alpha^*}(\omega_q)$$  

(8)

(for details see Makeev and Tumanyan 1994). This relation was used by several authors to study the periodicities in DNA (Silverman and Linsker, 1986; Tavare and Giddings, 1989). Thus our approach naturally includes the well-known Silverman and Linsker technique.

If $L$ is a symmetrical matrix, then the spectrum $\tilde{S}(\omega_q)$ is real [it is easy to demonstrate by changing the order of summation in (7)].

Note that in contrast to the usual correlation function spectrum, spectrum (7) in the general case is not positive. Indeed, if non-diagonal elements in $L$ are much greater than diagonal ones, then $\tilde{S}(\omega_q)$ may have negative values. However, this case seems to be non-realistic, because in this case symbols of the alphabet look more like each other than themselves.

In simple cases the generalized spectrum may be calculated analytically. One useful example is a random sequence of Bernoulli type.

Let us calculate the average spectral power of the Fourier transform of a random Bernoulli type sequence of length $N$, calculated when the arbitrary correlation matrix
Search of periodicities in primary structure of biopolymers: a general Fourier approach

L is used. We use the correlator's technique taken from McLachlan and Stewart (1976).

To calculate the average of the correlation function \( \langle R(k) \rangle \) one needs the pair distribution function \( \rho_{\alpha\beta}(n,k) \), i.e. the probability that amino acids \( \alpha \) and \( \beta \) are at \( n \) and \( k \). According to McLachlan and Stewart (1976) this probability is

\[
\rho_{\alpha\beta}(n,k) = \frac{1}{N(N-1)} (1 - \Delta_{n,\alpha})(1 - \Delta_{k,\beta}) + \frac{1}{N} \Delta_{\alpha,\beta} \Delta_{k,n}
\]

(9)

\( \Delta_{pq} = 1 \) if \( p = q \) and 0 if \( p \neq q \).

Calculate the average correlation function

\[
\langle R(k) \rangle = \frac{1}{\sqrt{N}} \sum_{m} \langle L(s_{m}, s_{m+k}) \rangle
\]

\[
= \frac{1}{\sqrt{N}} \sum_{m,\alpha,\beta} \rho_{\alpha\beta}(m, m+k)L(\alpha, \beta)
\]

\[
= \frac{(1 - \Delta_{k,0})}{\sqrt{N(N-1)}} \sum_{\alpha} L + \frac{(N\Delta_{k,0} - 1)}{\sqrt{N(N-1)}} \text{tr} L
\]

(10)

Here \( \sum L \) is the sum of all elements from \( L \) and \( \text{tr} L \) is the sum of its diagonal elements.

The average spectral power \( \langle Z(\omega_q) \rangle \) is calculated separately for the cases of \( \omega_q = 0 \) and \( \omega_q \neq 0 \). In the first case we have:

\[
\langle Z(\omega_q = 0) \rangle = \frac{1}{\sqrt{N}} \sum_{k} \langle R(k) \rangle = \frac{1}{N} \sum L
\]

(11)

In the second case, the calculations

\[
\langle Z(\omega_q = 0) \rangle = \frac{1}{\sqrt{N}} \sum_{k} e^{-i\omega_q k} \langle R(k) \rangle
\]

\[
= \frac{1}{\sqrt{N}} \sum_{k \neq 0} e^{-i\omega_q k} \langle R(k) \neq 0 \rangle + \frac{1}{\sqrt{N}} R(0)
\]

(12)

are convenient to do, noting that \( \langle R(k \neq 0) \rangle \) does not depend on \( k \). Thus

\[
\sum_{k = 0}^{N} e^{-i\omega_q k} \langle R(k \neq 0) \rangle = 0
\]

\[
\sum_{k \neq 0} e^{-i\omega_q k} \langle R(k \neq 0) \rangle = -\langle R(k \neq 0) \rangle
\]

(13)

is valid, which leads to

\[
\langle Z(\omega_q \neq 0) \rangle = \frac{1}{\sqrt{N}} ([R(0)] - \langle R(k \neq 0) \rangle)
\]

(14)

Or, after all substitutions are made, using (10)

\[
\langle Z(\omega_q \neq 0) \rangle = \frac{N \text{tr} L - \sum L}{N(N-1)}
\]

(15)

This is the resulting average spectral power. Finally

\[
\langle Z(\omega_q \neq 0) \rangle = \frac{N \text{tr} L - \sum L}{N(N-1)},
\]

\[
\langle Z(\omega_q = 0) \rangle = \frac{1}{\sqrt{N}} \sum_{k} \langle R(k) \rangle = \frac{1}{N} \sum L.
\]

(16)

We used these formulae to normalize results obtained with different matrices in order to compare them.

Thus we used as a measure of periodicity of a symbolic sequence, the 'symbolic Fourier transform' (7), related to the Fourier transform of the random sequence.

The main problem to overcome in Fourier analysis of sequences of finite length is a problem of detecting periods of a length comparable with the length of the sequence, for example \( \sim N/4 \). Another problem arises when a sequence contains strict repeats, such as the sequence of collagen, where each third amino acid is Gly. In this case harmonics of the strict repeat interfere with the final picture. In order to tackle this problem we used the following algorithm. For each period the sequence was cut to the longest possible length multiple of the integer number of periods and closed to a ring. The autocorrelation function is naturally calculated on a ring [see definition (1)]. This closure of the sequence into a ring of different length allows inclusion of all the periods in the resonance with the whole length of the sequence during calculation of the Fourier transform. Note that we have not used from the Fast Fourier Transform (FFT) algorithm. In FFT the length of the sequence should be to the power of 2 and in our case it varies during calculation. This serves two purposes. On the one hand, periods with a length comparable to the length of the whole sequence may be resolved and, on the other hand, strict repeats do not contribute to the power of other periods and distort the final picture.

Implementation

In this part some examples of application of SYMFOUR are given. A sequence with a length of 1000 amino acids is processed in \( \sim 20 \) min on an IBM-compatible AT386/87 DX40. The user interface is shown in Figure 1.

H4 histone

There is a clear periodicity of four amino acids along all the sequence of H4 histone, linked probably to its quasi-spiral structure. Figure 2 represents the data obtained using a single similarity matrix (in this case all the amino acids are absolutely unique) and using a Dayhoff matrix.

HIV reverse transcriptase

In HIV reverse transcriptase (fragment of Pol protein
Fig. 1. User interface.

Fig. 2. The power spectra of the H4 histone sequence (H4_human) obtained with different similarity matrices: Dayhoff and single. Note the periodicity of four, which may clearly be seen along all the histone molecule, especially in the distribution of charged amino acids. Note that usage of the Dayhoff matrix provides a clearer picture.

POL_HV1BR) Fourier transform shows a strong 6-fold periodicity (Figure 3). A closer inspection by subsequently cutting the sequence into two shows that this periodicity originates from a rather short 42 amino acid site located at position 438 from the beginning of the Pol protein.

\[
\begin{align*}
P &\ C &\ I &\ K &\ V &\ R \\
Q &\ L &\ C &\ K &\ L &\ L \\
P &\ G &\ T &\ K &\ A &\ L \\
I &\ E &\ V &\ I &\ P &\ L \\
T &\ E &\ E &\ A &\ E &\ L \\
E &\ L &\ A &\ E &\ N &\ R \\
F &\ I &\ L &\ K &\ E &\ P
\end{align*}
\]

Fig. 3. The power spectrum of HIV reverse transcriptase (fragment of Pol protein). A strong 6-fold period is observed.
Search of periodicities in primary structure of biopolymers: a general Fourier approach

Human collagen; Unpolar amino acids

Fig. 4. Periodical patterns in the sequence of collagen (Ca11 hum). The basic periods of 39 and 47 amino acids are characteristic of all sets of amino acids, while there exist many periods characteristic only of amino acids with specific features. Thus the basic periods create the most prominent peaks in the spectrum obtained when the Dayhoff matrix is used.

This Leu-rich sequence may be relevant to Leu-rich repeats (Kobe and Deisenhoffer, 1994) and thus important for reverse transcriptase function.

Collagen

Figure 4 shows complex periodical structure of the collagen sequence. One can clearly see that different types of amino acids in collagen have their own periodical patterns, whereas periods of 39 and 234 are characteristic of every amino acid family.

In Figure 5 the Fourier transform of collagen DNA is presented. The analysis of periodical patterns in collagen DNA was for the first time performed in our previous paper (Makeev et al., 1995). It can be seen that although the picture is not clear, the basic period of 117 (39 x 3) may be seen here just as in the spectrum of the protein. However, there are also some periods characteristic only of the DNA sequence. Note that in DNA analysis it is of no use to introduce similarity matrices, because there are only four letters in the DNA alphabet. However, sometimes it is useful to use words of certain length as letters; the similarity score for such words may be introduced in common way.

Discussion

It seems that matrix Fourier analysis provides a powerful tool to study the organization of primary structure, as well as patterns in the distribution of amino acids of specific type. Also, it allows simultaneous investigation of periods characteristic of DNA and proteins and comparison of them. This allows separation of the periods in DNA that are not related to the protein, which might be useful for the study of DNA properties. Moreover, Veljovic and colleagues (Veljovic et al., 1985) claim that Fourier transform of the protein sequence presents information about function which is difficult to derive from the primary structure alone, using for example, analysis of homologies. Thus we think the Fourier analysis may be used to classify proteins into families according to global dependencies in their primary structures, which may be of use for protein folding investigations, especially of fibrous proteins.

Another seemingly profitable field of Fourier implementation is DNA long correlation analysis and fractal structures searches (Peng et al., 1992), where Fourier transform allows creation of unique algorithms to analyse long DNA strands.
Conclusion

We suggest the use of Fourier transform to analyse periodical structures in sequences of symbols from arbitrary alphabets for which a comparison matrix is set. Our approach is based on the general relation between Fourier transform and a correlation function of a Wiener-Hinchin type. This technique allows comparison of periodical patterns of sequences written in different alphabets, such as DNA and proteins, and comparison of the results obtained through various digitizing procedures. A closing of the sequence in a ring with scanning of the ring length allows resolution of periods with a length comparable with the length of the whole sequence and vague periods multiple to strong repeats. The program SYMFOUR is written in PASCAL. The program is available on EMAIL makeev@imb.imb.ac.ru.

Acknowledgements

The authors are pleased to thank A.P.Lifanov, M.S.Gelfand and N.G.Esipova for helpful discussions and attention to their work. This study was partially supported by ISF grant JIGIOO and INTAS grant 93-3615.

References
