ESAIM: Probability and Statistics December 2001, Vol. 5, 171–181
URL: http://www.emath.fr/ps/

STATISTICAL TOOLS FOR DISCOVERING PSEUDO-PERIODICITIES IN BIOLOGICAL SEQUENCES

Bernard Prum[1], Élisabeth de Turckheim[2] and Martin Vingron[3]

Abstract. Many protein sequences present non trivial periodicities, such as cysteine signatures and leucine heptads. These known periodicities probably represent a small percentage of the total number of sequences periodic structures, and it is useful to have general tools to detect such sequences and their period in large databases of sequences. We compare three statistics adapted from those used in time series analysis: a generalisation of the simple autocovariance based on a similarity score and two statistics intending to increase the power of the method. Theoretical behaviour of these statistics are derived, and the corresponding tests are then described. In this paper we also present an application of these tests to a protein known to have sequence periodicity.

Mathematics Subject Classification. 62G10, 62P10.

Received February 5, 2001. Revised July 31 and October 12, 2001.

1. Introduction

Molecular biology deals with sequences of letters and mathematical tools are requested for automatic analysis of large sequence or set of sequences now available. Proteins are such sequences written in a 20 letter alphabet – the amino-acid (a.a.) alphabet. It is well known that the spatial folding of these molecules strongly depends on the sequence of a.a. For instance membrane proteins present a periodic structure of hydrophobic a.a. (for the segments inside the membrane) and hydrophilic a.a. (for the segments outside). Other examples are the leucine heptads in coiled-coil proteins, repeated cysteines and histidines signatures in zinc finger knots, whose analysis has been a long standing problem in computational molecular biology. The difficulty lies in the fact that repeats often show weak similarities due to evolutionary divergence, such that they are only recognised from a sensitive self comparison. In this paper we propose a procedure to identify the existence of a periodic structure with a random variation. If repeated motifs are present, even with some variation, they could also be detected by such tools.

Two methods have dominated the effort to identify general periodic structures. The first one is based on the dot-plot method which consists in reporting in a 2-way picture a point in (i, j) if the a.a. in positions i and j are similar according to a given criterion. Such a dot-plot often shows a characteristic pattern that allows to deduce the existence of repeats by visual inspection [4, 6]. Boguski [3] used sophisticated sequence comparison methods to refine the dot-plot of the self-comparison and Heringa and Argos [14] applied clustering techniques

Keywords and phrases: Biological sequences, proteins, periodicity, autocovariance funtion.

[1] Laboratoire Statistique et Génome, URA 8071 du CNRS, La Génopole, Université d'Evry, France; e-mail: prum@genopole.cnrs.fr

[2] Institut National de la Recherche Agronomique, BIA, 78352 Jouy-en-Josas, France; e-mail: et@jouy.inra.fr

[3] Max-Planck-Institut für Molekulare Genetik, Ihnestr. 73, 14195 Berlin, Germany; e-mail: vingron@molgem.mpg.de

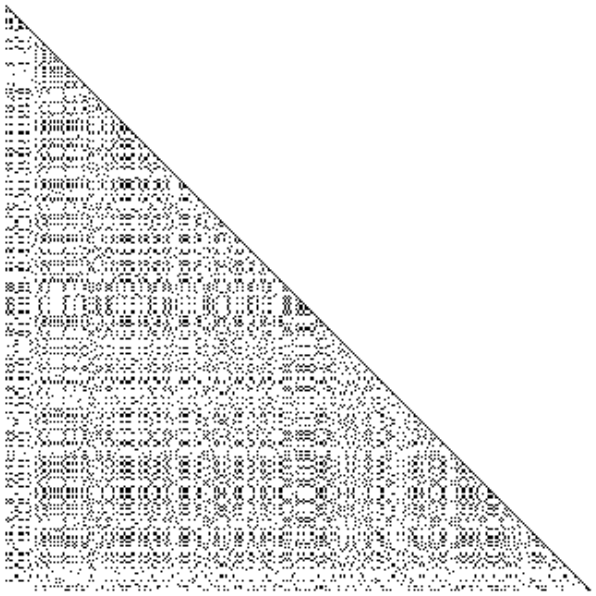

FIGURE 1. Human Apolipoprotein E. Dot-plot of points (i, j) such as the entry in the matrix PAM250 corresponding to $X(i)$ and $X(j)$ is greater than or equal to 2.

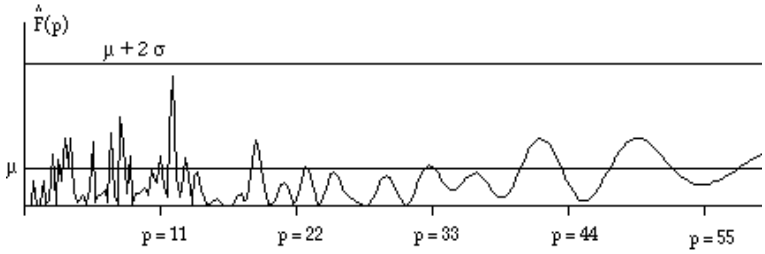

FIGURE 2. Fourier transform for the hydrophobicity of the human Apolipoprotein E. It is plotted as a function of the period p. The lower horizontal line indicates the expectation μ of the Fourier transform; the higher one indicates the level μ plus twice its standard deviation. No point appears to be significative. Note that a peak is almost significative for $p = 12$.

to the result of a sequence comparison. In Figure 1, we present the dot-plot corresponding to the protein that we have chosen to illustrate our method, namely the human Apolipoprotein E (for more details see Sect. 4). It appears rather difficult to "see" an obvious periodicity in this picture.

Alternatively, Fourier analysis can be applied to identify periodic features in a numerical vector chosen to represent the amino acid sequence, for example the hydrophobicity values. Both methods have been applied by McLachlan *et al.* [18–23]. The Fourier method was also applied to study the helical amphipaticity [7]. Figure 2 displays the Fourier transform of the same Apolipoprotein E when it is encoded using Ketty–Doolittle scale of hydrophobicity, which is the most commonly used one.

Each of the two major approaches has its deficiencies and no method has resulted in a general, automatic procedure to identify periodicities in sequences. Calculating a dot-plot still requires human interaction to recognise the characteristic pattern associated with repeats. Furthermore, when the period is small (*e.g.* only 7 amino acids long as in the heptad repeat of coiled-coil proteins) the pattern of the dot-plot is hardly discernible.

Converting the amino acid sequence into a numerical sequence in order to apply spectral analysis requires prior assumptions about the physical significance of the repeated motifs in order to chose an encoding and even given an encoding may produce only a weak signal.

Let us note that a third approach has been proposed by Coward [8]; for each possible period k, he searches the k-periodic sequence which is the nearest of the considered one according to a given similarity between amino acids.

In this paper, we propose an approach which adapts the autocovariance function to the case of protein sequences [2, 5]. In time series analysis, a very simple way to evaluate dependence between a sequence Y_t and its shifted copy with lag k uses $r(k) = E(Y_t Y_{t+k}) - E(Y_t)E(Y_{t+k})$. We propose to replace in this definition the product of real numbers YY' by a the score $\varphi(X, X')$ measuring the similarity of the two amino-acids X and X'. Hence if a word W appears in a repetitive way – or appears with minor changes –, the original sequence and its shifted version will give rise to a large value of the sum of scores of aligned amino-acids.

In the next section, we introduce this method more precisely and propose various statistics that are easy to compute, which one can use to detect periodicities in protein sequences. The mathematical properties of the proposed statistic are established Section 3. In Section 4, we first present an example of treatment of a protein which is known to present a periodicity and then another one, treating a protein without any periodicity.

Even though derived from a simple model, our approach works generally very well on proteins. Several violations of our assumptions are easily accommodated: neither small gaps on the repeated unit nor leading or trailing sequences do much harm to the results. If the method is to be used to find repeated motifs, it will not be able to detect repeats if there are only a few copies of the repeated unit, if their spacing is not uniform or if the alignment of different copies of the repeated unit requires many insertions and deletions.

The complexity of our algorithm is of order of the square of the sequence length, comparable *e.g.* to that of alignment algorithms. It is thus fast enough to analyse proteins of several thousand amino acids length without frustrating waiting time.

2. Method

Time-series analysis deals with sequences of random variables Y_t, where t takes integer values (say from 1 to n), and Y_t is real valued. Assuming stationarity, the statistical analysis is then founded on the empirical autocovariance function, which is defined as

$$c(k) = \frac{1}{n} \sum_i [Y_i Y_{i+k} - \bar{Y}^2]. \tag{1}$$

If there is a "hidden periodicity" with period equal to some k_0, then $c(k_0)$ will take a large value.

If we denote $S = (X_1, X_2, \dots, X_n)$ the amino-acid sequence, the main difference between the classical setup and our data setting lies in the fact that for each t the observation X_t is no more a real, but takes its value within a finite alphabet $\mathbf{A}$, the set of the 20 amino-acids. A possibility is to transform these values into real numbers, as it is done in Fourier analysis (*i.e.* change an amino-acid into its hydrophobicity, for example).

It is much more convenient to avoid this step and to understand the fact that some $c(k_0)$ takes a large value in (1) when X_t and X_{t+k} are similar. Actually, the comparison of amino-acids is very much used alignments of protein sequences, and there are different possible scores quantifying "how similar" two amino-acids are: all the PAM matrices introduced by Dayhoff [9], as well as BLOSUM matrices precisely describe this similarity according to biological knowledge [13].

We then propose to define a statistic similar to the sample autocovariance given in (1):

$$Y_n(k) = \frac{1}{\sqrt{n}} \sum_i [\varphi(X_i, X_{i+k}) - \hat{\varphi}_0] \tag{2}$$

where φ is a similarity score for amino-acids and $\hat{\varphi}_0$ is an estimation of $\varphi_0 = E(\varphi(X, X'))$ where X and X' are letters of $\mathbf{A}$ chosen independently with a common distribution. The $n^{-1/2}$ normalisation factor is the appropriated rate for the central limit convergence.

Based on such statistics, we wish to test the null hypothesis H_0: the letters of the observed sequence $\mathcal{S}$ are chosen independent according to a same distribution on $\mathbf{A}$. In particular, under H_0 there will be no periodicity in $\mathcal{S}$.

To carry out such tests, one must determine the distribution of the finite sequence of $Y_n(k)$ for $k = 1, \ldots, K$ – at least asymptotically, as n tends to infinity –, in order to determine the cut-off value u such that the test

$$H_0 \text{ is rejected} \iff \exists\, k_0\ ,\ Y_n(k_0) \geq u$$

has a fixed error value α.

If for a given k_0, $Y_n(k_0)$ is large and if the other $Y_n(k)$ are small, there is an evidence that there exists a dependence between X_i and X_j with a lag k between i and j producing similar a.a. at k distant sites. In such a case, $Y_n(k_0)$ will be large as well as $Y_n(2k_0), Y_n(3k_0) \ldots$

Large values of $Y_n(k)$ for many k would show another type of dependency between the X_i like that of autoregressive time series.

It turns out that this asymptotic distribution is much easier to obtain if we consider the sequence of amino-acids written on a circle; in other words the sum $i+k$ is to be understood "modulo" n. For a rather long sequence (typically in our examples n is between 300 and 2000 amino-acids), if we search rather short periodicities (say between some units to 20), this convention does not change much the value of $Y_n(k)$. In other cases, the protein (or a part of the protein) is constituted by 2 or 3 repetitions of a same long sequence, because a duplication of part of its gene took place during the historical evolution. In such cases, "closing" the sequence is even useful, taking into account the alignment of the last period with the first one.

The asymptotic behaviour of $Y_n(k)$ is given in the following theorem:

Theorem 2.1. *If we denote by* $\sigma^2 = \mathrm{Var}(\varphi(X, X'))$ *and by* $\rho^2 = \mathrm{Var}(\alpha(X))$ *where* $\alpha(x) = \mathrm{E}(\varphi(x, X'))$*, then, under* H_0*, when* n *tends to infinity, the finite sequence of the* $Y_n(k)$ *(for* $k = 1, \cdots, K$*) converges in distribution to a sequence of independent centred Gaussian variables with same variance* $(\sigma^2 - 2\rho^2)^2$*.*

We can then consider three statistics for testing the presence of periodicities in a protein sequence:

1) we can compare $Y_n(k)$ to z times the computed standard deviation, where z is the suitable quantile of the standard Gaussian $\mathcal{N}(0; 1)$;

2) if k_0 is a period, then $2k_0, 3k_0, cdots$ are periods also. Hence we can introduce, for a fixed number of terms J such as $Jk \leq n$

$$Z_n(k) = \sum_{j=1}^{J} Y_n(jk).$$

We call $Z_n(k)$ the cumulated statistic. According to the asymptotic zero-correlation between the $Y_n(k)$, $Z_n(k)$ will converge in distribution to a $\mathcal{N}(0; J\ (\sigma^2 - 2\rho^2)^2)$;

3) another possible statistic is

$$C_n(h) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n-h} Y_n(k)\ Y_n(k+h)\ .$$

We shall call this statistic the quadratic one. In the next section we prove that $C_n(h)$ converges to a Gaussian limit and we give its asymptotic variance.

Remark. As we do not consider applications where k would tend to infinity, we do not need to prove more than finite dimensional convergence.

3. Mathematical proofs

Let $X_1, \dots, X_n$ be iid variables on a finite set $\mathbf{A}$. We denote by μ their common distribution. We consider a symmetrical score function

$$\varphi : \mathbf{A} \times \mathbf{A} \to \mathbf{R}, \qquad \varphi(a,b) = \varphi(b,a)$$

and we let

$$\varphi_0 = \mathrm{E}(\varphi(X, X')) \qquad \sigma^2 = \mathrm{Var}(\varphi(X, X'))$$

$$\alpha(x) = \mathrm{E}(\varphi(x, X')) \qquad \rho^2 = \mathrm{Var}(\alpha(X))$$

where X and X' are independent with distribution μ. Obviously, $E(\alpha(X)) = \varphi_0$. To have simpler results, we "close" the sequence $X_1, \dots, X_i, \dots, X_n$ and consider that the index i lives on a torus, so that $i = i + n$.

3.1. Asymptotic behaviour of $Y_n(k)$

For $k \geq 1$, we define $Y_n(k)$ as

$$Y_n(k) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\varphi(X_i, X_{i+k}) - \hat{\varphi}_0) \ ,$$

where

$$\hat{\varphi}_0 = \frac{2}{n(n-1)} \sum_{i<j} \varphi(X_i, X_j).$$

We recall the following theorem ([24], pp. 192-194) of convergence in distribution for degenerate U-statistics, which shows that the rate of convergence is n^{-1} (instead of $n^{-1/2}$ for non degenerate statistics).

Theorem 3.1. *Let X_i be iid for $i = 1, \dots, n$ and let $h(x, x')$ be a symmetrical function such that*

$$i \neq j \quad \Rightarrow \quad \mathrm{E}(h(X_i, X_j)) = 0 \quad \text{and} \qquad \mathrm{Var}(h(X_i, X_j)) < \infty.$$

Consider the U-statistic

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} h(X_i, X_j).$$

If the "projection" of $h(X_i, X_j)$ on X_i, which is defined as the conditional expectation

$$\tilde{h}(X_i) = \mathrm{E}[h(X_i, X_j) \mid X_i],$$

is a constant (i.e. $\mathrm{Var}(\tilde{h}(X_i)) = 0$), then $n\,U_n$ converges in distribution towards a (finite) random variable.

We consider the degenerate U-statistics related to $\hat{\varphi}_0$ where $\varphi(X_i, X_j)$ is replaced by $\psi(X_i, X_j)$ with

$$\psi(X_i, X_j) = \varphi(X_i, X_j) - \alpha(X_i) - \alpha(X_j) + \varphi_0.$$

Lemma 3.2. *If $i \neq j$, then* $\mathrm{Var}(\psi(X_i, X_j)) = \sigma^2 - 2\rho^2$.
If $\{i, j\} \neq \{i', j'\}$, then $\mathrm{Cov}(\psi(X_i, X_j), \psi(X_{i'}, X_{j'})) = 0$.

This lemma immediately follows from the following identities:
$\mathrm{Cov}(\varphi(X_i, X_j), \varphi(X_{i'}, X_{j'})) = 0$ if the four values (i, j, i', j') are all different,
$\mathrm{Cov}(\varphi(X_i, X_j), \varphi(X_{i'}, X_{j'})) = \rho^2$ when $i = i'$ and (i, j, j') are different,
$\mathrm{Cov}(\varphi(X_i, X_j), \varphi(X_{i'}, X_{j'})) = \sigma^2$ when $i = i'$ and $j = j'$ (with $i \neq j$),
$\mathrm{Cov}(\alpha(X_i), \varphi(X_{i'}, X_{j'})) = 0$ if (i, i', j') are different and $\mathrm{Cov}(\alpha(X_i), \varphi(X_{i'}, X_{j'})) = \rho^2$ when $i = i'$ or $i = j'$.
By elementary calculus, we can write $Y_n(k)$ using the $\psi(X_i, X_j)$ as:

$$Y_n(k) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, X_{i+k}) - \frac{2}{(n-1)\sqrt{n}} \sum_{i<j} \psi(X_i, X_j) \ .$$

In other words,

$$Y_n(k) = \frac{1}{\sqrt{n}} \ S_n(k) + \sqrt{n} \ U_n \ ,$$

where

$$S_n(k) = \sum_{i=1}^{n} \psi(X_i, X_{i+k})$$

and

$$U_n = \frac{-2}{n(n-1)} \sum_{i<j} \psi(X_i, X_j) \ .$$

The projections of the terms involved in this U-statistic are zero:

$$\tilde{\psi}(X_i) = E(\psi(X_i, X_j)|X_i).$$

As $\mathrm{Var}(\psi(X_i, X_j)) = \sigma^2 - 2\rho^2$ is finite, as a result of Theorem 3.1, nU_n converges to a finite random variable. Hence we have:

Lemma 3.3. *When n tends to infinity, $U_n(k) = O_P(n^{-1})$.*

As a consequence, $\lim \ Y_n(k) = \lim \ \frac{1}{\sqrt{n}} \ S_n(k)$. For a fixed k, we can applied the following theorem to the sequence $\xi_i = \psi(X_i, X_{i+k})$ ([15], [10]):

Theorem 3.4. *For a sequence of identically distributed variables ξ_i such that* $\mathrm{E}(\xi_i) = 0$, *define* $S_n = \xi_1 + \xi_2 + ... + \xi_n$. *If* $\mathrm{Var}(S_n) = n\ \sigma^2 + O(1)$, *with* $\sigma^2 > 0$ *and* $\mathrm{E}(|\xi_i|^{2+\delta}) < \infty$ *for some* $\delta > 0$, *then* $\frac{1}{\sqrt{n}} \ S_n$ *converges in distribution to a Gaussian variable* $\mathcal{N}(0; \sigma^2)$.

As k is fixed, ξ_i and ξ_j are independent except when $|i - j| = k$, but even in this case

$$\mathrm{Cov}(\xi_i, \xi_j) = 0.$$

Hence $\mathrm{Var}(S_n(k)) = n\mathrm{Var}(\psi(X_i, X_{i+k})) = n(\sigma^2 - 2\rho^2)$; as ξ_i is bounded, Theorem 3.4 implies:

Theorem 3.5. $Y_n(k) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\varphi(X_i, X_{i+k}) - \hat{\varphi}_0)$ *converges in distribution to a Gaussian variable* $\mathcal{N}(0; \sigma^2 - 2\rho^2)$.

The same theorem applied on finite linear combinations $\sum\ a_{k_r} Y_n(k_r)$ shows the convergence of any finite sequence $(Y_n(k_1), \dots, Y_n(k_r))$ to a vector of i.i.d. Gaussian variables. If we set $Y_n(k) = 0$ for $k > n$, then the process $Y_n(.)$ converges in distribution to a Gaussian white noise $W(.)$ with variance equal to $\sigma^2 - 2\rho^2$.

It is important to note that this weak convergence of $Y_n(.)$ to $W(.)$ does not imply the convergence of the moments. In particular we will show in Theorem 3.8 that the "autocovariance" $C_n(h) = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{k=1}^{n-h} Y_n(k)\ Y_n(k+h)$ does not converge to the "autocovariance" of $W(.)$, which is $\mathcal{N}(0; (\sigma^2 - 2\rho^2)^2)$.

3.2. **Asymptotic behaviour of** $C_n(h)$

For $h \geq 1$, we consider

$$C_n(h) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n-h} Y_n(k)\ Y_n(k+h).$$

Define

$$\tilde{C}_n(h) = \frac{1}{n^{3/2}} \sum_{k=1}^{n-h} S_n(k)\ S_n(k+h).$$

Lemma 3.6. *If $\tilde{C}_n(h)$ converges in distribution to a variable Z when n tends to infinity, then $C_n(h)$ converges to the same limit.*

From $Y_n(k) = \dfrac{1}{\sqrt{n}}\ S_n(k) + \sqrt{n}\ U_n$, we have

$$C_n(h) = \tilde{C}_n(h) + U_n\ \frac{1}{\sqrt{n}} \sum_{k=1}^{n-h} [S_n(k) + S_n(k+h)] + n^{3/2}\ U_n^2.$$

From Lemma 3.3, we know that $n^{3/2}\ U_n^2$ is $O_P(n^{-1/2})$; this lemma and the convergence of $n^{-1/2} \sum S_n(k)$ implies, using Cauchy–Schwartz inequality, that the middle term of the right hand side also converges to zero.

Note that $\tilde{C}_n(h)$ can be written as

$$\tilde{C}_n(h) = \frac{1}{n^{3/2}} \sum_{k=1}^{n-h} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi(X_i, X_{i+k})\ \psi(X_j, X_{j+k+h}).$$

We compute $V = \mathrm{Var}(\tilde{C}_n(h))$.

Let us consider the sequence $(\xi_{i,j,k})$ with indices in the cube $[1, \dots, n]^3$

$$\xi_{i,j,k} = \psi(X_i, X_{i+k})\ \psi(X_j, X_{j+k+h}).$$

To compute the variance of $\tilde{C}_n(h)$, we only have to consider terms of order n^3 since terms of order less than n^3 will disappear.

To each index, (i, j, k) corresponds the set $A(i, j, k)$ of the points $\{i, i+k, j, j+k+h\}$ involved in the calculus of $\xi_{i,j,k}$. A systematic screening of all possible cases leads to the lemma:

Lemma 3.7. $\mathrm{Cov}(\xi_{i,j,k}, \xi_{i',j',k'}) = 0$ *except in only two cases:*

1) *$A(i, j, k)$ contains 4 distinct points (i.e. $\#A(i, j, k) = 4$), and $A(i', j', k') = A(i, j, k)$;*

2) $\#A(i,j,k) = 3$, $\#A(i',j',k') = 3$, *and* $\#(A(i,j,k) \cup A(i',j',k')) = 4$.

Case 1. First note that when $A(i,j,k)$ contains 4 distinct points, say u, v, s and t

$$\mathrm{Var}(\xi_{i,j,k}) = \mathrm{Var}(\psi(X_u, X_v)\ \psi(X_s, X_t)) = \mathrm{Var}(\psi(X_u, X_v))^2 = (\sigma^2 - 2\rho^2)^2.$$

But it turns out that in this case, there are exactly 3 other choices (i',j',k') such that $A(i,j,k) = A(i',j',k')$ described in the rows C2, C3 and C4 in the following table:

C1	i	j	k
C2	$i' = i$	$j' = i + k$	$k' = j - i$
C3	$i' = j + k + h$	$j' = j$	$k' = n - i - j - h$
C4	$i' = j + k + h$	$j' = i + k$	$k' = n - i - j - k - h$

This defines an equivalence between indices (i,j,k), each class containing exactly 4 indices; there are then $\dfrac{n^3}{4}$ classes.

It is easy to see that $\xi_{\mathrm{C1}} = \xi_{\mathrm{C4}}$ and $\xi_{\mathrm{C2}} = \xi_{\mathrm{C3}}$. The contribution of the four variables ξ associated to a given $A = \{u, v, s, t\}$ to $\tilde{C}_n(h)$ is $2(\xi_{\mathrm{C1}} + \xi_{\mathrm{C2}})$. Therefore, its contribution to V is

$$8\ \mathrm{Var}(\xi_i) + 8\ \mathrm{Cov}(\xi_{\mathrm{C1}}, \xi_{\mathrm{C2}}) = 8(\sigma^2 - 2\rho^2)^2 + 8c,$$

where

$$c = \mathrm{Cov}(\psi(X_u, X_s)\ \psi(X_u, X_t), \psi(X_v, X_s)\ \psi(X_v, X_t)),$$

or equivalently

$$c = \mathrm{E}(\psi(X_u, X_s)\ \psi(X_u, X_t)\ \psi(X_v, X_s)\ \psi(X_v, X_t)).$$

As there are $\dfrac{n^3}{4}$ classes, the contribution to V of the terms described in Case 1 is $V_1 = 2(\sigma^2 - 2\rho^2)^2 + 2c$.

Case 2. We have now to consider the situation described in point 2 of Lemma 4. A systematic study of the possible cases shows that there are only 4 kinds of pairs $A(i,j,k)$, $A(i',j',k')$ giving terms of order n^3:

$i = j$	$i' = j'$	$i + k = i' + k'$	$\Rightarrow j + k + h = j' + k' + h$
$i = i'$	$j = j'$	$i + k = j + k + h$	$\Rightarrow i' + k' = j' + k' + h$
$i = j$	$j' = i + k$	$i' = j + k + h$	$\Rightarrow i' + k' = j' + k' + h$
$i' = j'$	$j = i' + k'$	$i = j' + k' + h$	$\Rightarrow i + k = j + k + h$

Up to order n^2, there are n^3 possible choices of $(i,j,k), (i',j',k')$ associated to each kind of pair of sets A. Hence, the contribution on the terms described in Case 2 to V is $V_2 = 4c$.

Hence $V = V_1 + V_2 = 2(\sigma^2 - 2\rho^2)^2 + 6c$, and we can conclude:

Theorem 3.8. *$C_n(h)$ converges in distribution to a Gaussian $\mathcal{N}(0; 2(\sigma^2 - 2\rho^2)^2 + 6c)$, where*

$$c = \mathrm{E}(\psi(X_u, X_s)\ \psi(X_u, X_t)\ \psi(X_v, X_s)\ \psi(X_v, X_t)).$$

Remark. In practice, c has to be estimated. Theoretically, it can be expressed using 4^{th}-order moments of $\varphi(X, X')$ and $\alpha(X)$, but it is shorter to estimate the expectations of the probabilities of each of the 20 a.a., then by "plug-in" get an estimation of $\alpha(x)$ and then of c. This gives a consistent estimate $\hat{V}$ of V and $\hat{V}^{-1/2}\ C_n(h)$ converges to a Gaussian variable $\mathcal{N}(0;1)$.

```
MKVLWAALLV  TFLAGCQAKV  EQAVETEPEP  ELRQQTEWQS  GQRWELALGR
FWDYLRWVQT  LSEQVQEELL  SSQVTQELRA  LMDETMKELK  AYKSELEEQL
TPVAEETRAR  LSKELQAAQA  RLGADMEDVC  GRLVQYRGEV  QAMLGQSTEE
LRVRLASHLR  KLRKRLLRDA  DDLQKRLAVY  QAGAREGAER  GLSAIRERLG
PLVEQGRVRA  ATVGSLAGQP  LQERAQAWGE  RLRARMEEMG  SRTRDRLDEV
KEQVAEVRAK  LEEQAQQIRL  QAEAFQARLK  SWFEPLVEDM  QRQWAGLVEK
VQAAVGTSAA  PVPSDNH
```

FIGURE 3. The sequence of Human Apolipoprotein E (P02649).

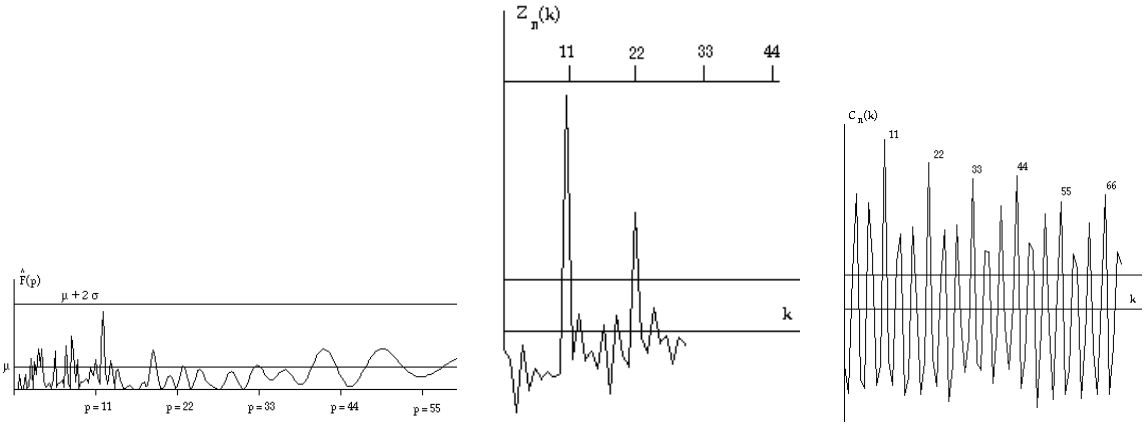

FIGURE 4. Apolipoprotein E ($n = 317$). a) The autocovariance function $Y_n(k)$. Horizontal lines indicate $+2.33\sigma$ and -2.33σ, where σ is the standard deviation of $Y_n(k)$. A significant periodicity is 11. Note also a lower significant periodicity $p = 11/3$ which probably corresponds to the presence of α-helices (11/3 is the number of amino acids in each turn). b) The cumulant statistics $Z_n(k)$ for $k = 1, \cdots, 30$. The scale is 2.5 greater than for $Y_n(k)$. The horizontal line indicates $+2.33\sigma$, where σ is the standard deviation of $Z_n(k)$. Significativity of $k = 11$ and $k = 22$ is much higher than in Figure 4a. c) The quadratic statistic $C_n(k)$. The upper line indicates $+2.33\sigma$, where σ is the standard deviation of $C_n(k)$. Very significative peaks appear for all periods $p = 11q$. Again the periodicity 11/3 appears.

4. Application

We now show on an example, how the method can be applied to a protein. We chose the Apolipoprotein E, which is a membrane protein. In all the examples, we use the PAM250 matrix as $\varphi(X, X')$.

Apolipoprotein E (SwissProt Access Number: P02649) is an Human protein which mediates binding internalisation and catabolism of lipoprotein particles, especially in the hepatic tissues. It has been shown to contain a period of length 11 [18, 21] and the copies of the repeated unit are rather well conserved. Figure 3 gives the complete sequence of this protein.

Applying our method to this case results in a significative peak of the autocovariance function $Y_n(k)$ at the period 11 (Fig. 4a); the peak is much more significant (approximately 12 standard deviation above the mean) when considering the cumulated statistic $Z_n(k)$ (Fig. 4b); it also appears for the quadratic statistic $C_n(k)$ (Fig. 4c). Hence all these statistics prove a very significant periodicity with $k = 11$.

We applied similar treatment to a number of other proteins known for presenting pseudo-periodicity, as for example the myosin of the rod of *C. elegans* (SwissProt P 12844), rat α-farnesyl transferase (Q02769),

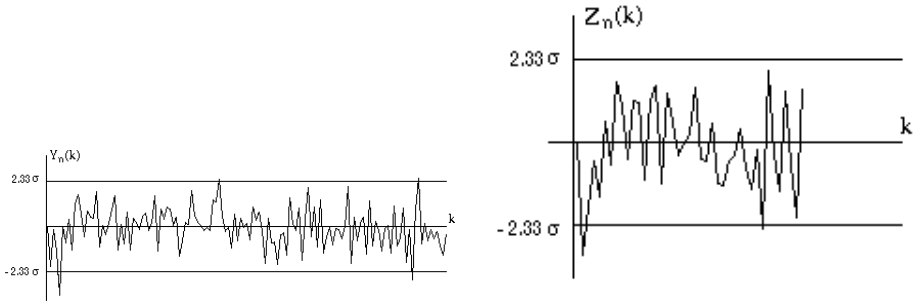

FIGURE 5. Plots of $Y_n(k)$ and $Z_n(k)$ for the protein AAHA of *methylophylus*. No period appears as significative.

Salmonella typhimurium flagellin (P03001) or *Xenopus laevis* transcription factor III (P03001). In all cases, our statistics clearly detect periodicities (for example $p = 7$ for the myosin or the farnesyl transferase, $p = 6$ for flagellin and $p = 30$ for transcription factor).

Finally, we tested our method on a protein without any periodicity. We chose a globular protein, AAHA which is the A-chain of methanol dehydrogenase from the bacteria *methylophylus* (PDB reference: 4aah). The graphs of $Y_n(k)$ and $Z_n(k)$ do not show any significant period.

5. Conclusion

We applied simple statistics mimicking those used in time series analysis to catch periodicity in sequences of letters. When the simple alignment statistic $Y_n(k)$ has the behavior of a periodic time series, the set of examples of proteins shows that the statistic $Z_n(k)$ cumulating the $Y_n(h)$ for h chosen as multiples of k clearly isolates the main period. On a set of other examples not shown here, it appeared that the quadratic statistic $C_n(h)$ defined like an autocorrelation of the centered sequence $Y_n(k)$ does not select the main period but shows a much more periodic behavior with was not at all the expected one. On the set of the tested proteins, the behaviour of this statistic was very disappointing.

Therefore, the cumulative statistic is therefore a suitable tool to find periodicities in sequences of letters, and in particular in protein sequences. Test based on this statistic seem to be more powerful than the one based on the simple alignment statistic.

References

[1] P. Argos, Evidence for a repeating domain in type I restriction enzyme. *European Molecular Biology Organization J.* **4** (1985) 1351-1355.

[2] G. Benson and M.S. Waterman, A method for fast data search for all k-nucleotide repeats. *Nucleic Acids Res.* **20** (1994) 2019-2022.

[3] M.S.M. Boguski, R.C. Hardison, S. Schwart and W. Miller, Analysis of conserved domains and sequence motifs in cellular regulatory proteins and locus control using new software tools for multiple alignments and visualization. *The New Biologist* **4** (1992) 247-260.

[4] G.M. Bressan, P. Argos and K.K. Stanley, Repeating structure of chick tropoelastin revealed by complementary DNA cloning. *Biochemistry* **26** (1987) 1497-1503.

[5] P.J. Brockwell and R.A. Davis, *Time Series: Theory and Methods*. Springer-Verlag (1987).

[6] R.S. Brown, C. Sander and P. Argos, The primary structure of transcription factor TF III A has 12 consecutive repeats. *Federation of European Biochemical Society Letter* **186** (1985) 271-274.

[7] J.L. Cornette, K.B. Cease, H. Margalit, J.L. Sponge, J.A. Berzofsky and Ch. DeLisi, Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Molecular Biology* **195** (1987) 659-685.

[8] E. Coward, Detecting periodicity pattern in biological sequences. *Bioinformatics* **14-6** (1998) 498-507.

[9] M.O. Dayhoff, R. Schwartz and B.C. Orcutt, A model of evolutionary change in protein, edited by M.O. Dayhoff. National Biomedical Research Foundation, Washington D.C., *Atlas of Protein Sequences and Structure* **5-3** (1978) 345-352.

[10] P. Doukhan, Mixing, properties and examples. Springer Verlag, *Lecture Notes in Statist.* **85** (1985).
[11] V.A. Fischetti, G.M. Landau and P.H. Seller, Identifying period occurences of a template with application to protein structure. *Inform. Process. Lett.* **45-1** (1993) 11-18.
[12] W. Fitch, Phylogenies constrained by cross-over process as illustrated by human hemoglobins an a thirteen-cycle, eleven amino-acid repeat in human apolipoprotein AI. *Genetics* **86** (1977) 623-644.
[13] S. Hennikoff and J.G. Henikoff, Amino acid substitution matrices from protein blocks for database research. *Nucleid Acid Res.* **19** (1992) 6565-6572.
[14] J. Heringa and P.Argos, A method to recognize distant repeats in protein sequences. *Proteins* **17-4** (1993) 391-441.
[15] I.A. Ibragimov, On a central limit theorem for dependent random variables. *Theory Probab. Appl.***15** (1975).
[16] S. Labeit, M. Gautel, A. Lakey and J. Trinick, Towards a molecular understanding of titin. *European Molecular Biology Organization J.* **11** (1992) 1711-1716.
[17] A. Lupas, M. van Dyke and J. Stock, Predicting coiled coils from protein sequences. *Science* **252** (1991) 1162-1164.
[18] A.D. McLachlan, Analysis of periodic patterns in amino-acid sequences: Collagen. *Biopolymers* **16** (1977) 1271-1297.
[19] A.D. McLachlan, Repeated helical patterns in apolipoprotein AI. *Nature* **267** (1977) 465-466.
[20] A.D. McLachlan and J. Karn, Periodic features in the amino-acid sequence of nematod myosin rod. *J. Molecular Biology* **220** (1983) 79-88.
[21] A.D. McLachlan and M. Stewart, The 14-fold periodicity in alpha-tropomyosin and the interaction with actin. *J. Molecular Biology* **103** (1976) 271-298.
[22] A.D. McLachlan, M. Stewart, R.O. Hynes and D.J. Rees, Analysis of repeated motifs in talin rod. *J. Molecular Biology* **235-4** (1994) 1278-1290.
[23] J. Miller, A.D. McLachlan and A. Klug, Repetitive zinc-binding domains in the transcription factor IIIA from Xenopus oocytes. *European Molecular Biology Organization J.* **4** (1985) 1609-1614.
[24] R.J. Serfling, *Approximation Theorems of mathematical statistics*. Wiley (1980).