J. Mol. Model. **1999**, *5*, 103 – 115

Journal of Molecular Modeling

FULL PAPER

Latent Periodicity of Protein Sequences

Maria A. Korotkova[1], Eugene V. Korotkov[2], and Valentina M. Rudenko[2]

[1]Department of Cybernetics, Moscow Physical Engineering Institute, Kashirskoe shosse 31, 115409, Moscow, Russia.

[2]Center of Bioengineering, Russian Academy of Sciences, Prospect 60-tya Oktyabrya, 7/1, 117312, Moscow, Russia.
Tel: +7-095-430-32-70; Fax: +7-095-135-05-71. E-mail: katrin2@biengi.ac.ru

Received: 18 February 1999/ Accepted: 8 April 1999/ Revised: 12 May 1999/ Published: 11 June 1999

Abstract This article is in the area of protein sequence investigation. It studies protein sequence periodicity. The notion of latent periodicity is introduced. A mathematical method for searching for latent periodicity in protein sequences is developed. Implementation of the method developed for known cases of perfect and imperfect periodicity is demonstrated. Latent periodicity of many protein sequences from the SWISS-PROT data bank is revealed by the method and examples of latent periodicity of amino acid sequences are demonstrated for: the translation initiation factor EIF-2B (epsilon subunit) of Saccharomyces cerevisiae from the E2BE_YEAST sequence; the E.coli ferrienterochelin receptor from the FEPA_ECOLI sequence; the lysozyme of Bacteriophage SF6 from the LY_BPSF6 sequence; lipoamide dehydrogenase of Azotobacter vinelandii from the DLDH_AZOVI sequence. These protein sequences have latent periods equal to six, two, seven and 19 amino acids, respectively. We propose that a possible purpose of the amino acid sequence latent periodicity is to determine certain protein structures.

Keywords Latent periodicity, Protein sequences, Mutual information

Introduction

The study of protein sequence periodicity is one of many approaches to protein sequence investigation. The purpose of these investigations is to find structural peculiarities of amino acid sequences and their relations with the spatial organisation and function of proteins. Two main mathematical approaches are widely used for this purpose. The first method is classical Fourier transformation of a symbolic sequence [1-9]. The main fundamental difficulty of Fourier analyses or autocorrelation techniques for symbolic sequences is the transformation of symbolic to numerical sequences saving all statistical properties of an initial sequence. Typical mapping of amino acids to number (or numbers) is taken from the physico-chemical properties of amino acids [8,9]. Considering the amino acid sequence as a symbolic sequence, we can propose a very large number of mappings of the amino acids to number (or numbers) and each type of mapping may find some set of amino acid periodicities. For example, we can transform amino acid sequence to numerical sequence as follows: {Ala≡10, Val≡1}, {Rest amino acid≡0}. In this case autocorrelation techniques and Fourier transformation can find the periodicity of Ala completely, the periodicity of Val is less expressed and the periodicities

Correspondence to: E. V. Korotkov

of the remaining amino acids are missed. To find all types of periodicity in a symbolic sequence, we should test all types of mapping, but this requires testing of an almost infinite number of mappings. Thus, Fourier transformation in its present form may miss some periodicities of amino acid sequences and we should develop a new method of autocorrelation function calculation (or similar function) that allows us to find all periodicities of the amino acid sequence.

The second method is algorithmic investigations of symbolic sequences. Methods of algorithmic search of tandem repeats were developed earlier [10-17]. These algorithmic methods are usually based on the dynamic programming of the alignment of pair sequences [18]. For searching amino acid periodicity by the dynamic programming method, we should determine the weights (w_{ij}) for the amino acid coincidences a_ia_j, where a_i, a_j are amino acids, i=1,...,20, j=1,...,20. Weights determine a set of periodicities that can be found in a sequence. Usually, these weights are taken from the PAM-250 (or similar matrices) where w_{ij} were calculated by comparison of closely related proteins. Weights w_{ii} are greater than w_{ij} for i≠j (i,j=1,...,20) in most cases. This means that we may find the periodicity of an amino acid sequence by the dynamic programming method if the level of similarity between the periods is high and we may miss a periodicity if the level of similarity is low or zero. For example, the dynamic programming method in its present form misses the latent periodicity of the amino acid sequence:

[LysArgGln][AsnSerVal][IleProAla][MetHysAla][AsnGluGln][LysHysVal]...

The brackets show the periods. This example sequence has the following amino acid latent periodicity: {(Lys\Asn\Ile\Met)(Arg\Ser\Pro\Hys\Glu)(Gln\Val\Ala)}. The first position of the period has one set of amino acids, the second position of the period has another. The third position has a set of amino acids that differs from the sets for the first and second positions. If the sequence is long enough, then latent periodicity can be found to be statistically important.

Latent periodicity is an integral statistically important property of an amino acid sequence having a generalised period of some length and a set of used amino acids for every position of the period. Statistically important similarity between any two periods is low or absent if latent periodicity occurs in an amino acid sequence. An example of an amino acid sequence with latent periodicity is shown above.

Many types of latent periodicity may exist for a given length of a period. To reveal these types of latent periodicity by the dynamic programming method, we should test a large number of weight matrices for a_ia_j (i=1,...,20; j=1,...,20) coincidences. A full test of all possible matrices requires an astronomic time because the number of the latent periodicity types increases rapidly with increasing period length.

The first purpose of this report is to develop a new method for calculation of a function that is, on the one hand, similar to an autocorrelation function and does not require the mapping of the amino acids to numbers and, on the other hand, may reveal any statistically important periodicity of an amino acid sequence, including latent periodicity. For these purposes we modify the mathematical method that we elaborated earlier for DNA sequences [19-21]. The method uses the principle of the enlarged similarity between symbolic sequences [22]. In the present study this method has been changed considerably for the analysis of the amino acid alphabet (20 letters) by the use of Monte-Carlo calculations.

The second purpose of this report is to show how the developed autocorrelation function for symbolic sequences works on real amino acid sequences. For this purpose we analyse the protein sequences from the SWISS-PROT data bank and show some examples of the amino acid sequences with latent periodicity from several proteins. We use a weight profile analysis for the determination of a biological sense of amino acid sequences found with periodicity. The possible meaning of the amino acid sequence latent periodicity is discussed.

Materials and methods

We use the comparison of a protein sequence with artificial symbolic sequences for finding a latent periodicity in an amino acid sequence, as done earlier for DNA sequences [19-21]. The alphabet of artificial sequence contains S_i letters. If we search a period equal to two amino acids, we generate the artificial sequence $S_1S_2S_1S_2S_1S_2$..... The artificial sequence $S_1S_2S_3S_1S_2S_3S_1S_2S_3$... is generated if we search a period equal to three amino acids. The sequence S_1S_2...$S_nS_1S_2$...$S_nS_1S_2$...S_n... is created for finding period of length n. The length of the artificial sequence is equal to the length L of the protein sequence to be analysed. The artificial sequences with periods from two to L/2 are compared with the protein sequence in question one after another. For example, if we search a period of length two in the amino acid sequence, then we compare two sequences:

```
A V L I P W D E ...
S1S2S1S2S1S2S1S2...
```

Matrix M(20,n) is filled for each comparison. Element M(i,j) shows the number of i type amino acids (i=1,...,20) that are situated opposite the letter S_j (j=1,...,n) of the artificial sequence.

The measure of a similarity between two compared sequences is selected as mutual information, calculated from matrix M(20,n). Double mutual information is distributed as X^2 with 19(n-1) degrees of freedom if a random amino acid sequence and an artificial sequence are compared; the average value of the mutual information is equal to 19(n-1). If the mutual information value is increased, then the probability of a random relation between the amino acid sequence and the artificial periodical sequence is decreased [23].

However, if the length of compared sequences is insufficient, then the double mutual information 2I is not distributed as X^2 with 19(n-1) degrees of freedom. We believe that an amino acid sequence has insufficient length if the average value of matrix element M(20,n) is less than five. This is true for compared sequences of length less than 100n, where n is the length of the period to be analysed. In this case, it is impossible to calculate the correct estimate of the probability of random creation of latent periodicity using the 2I value. The Monte-Carlo method is used for this estimate for the case of insufficient length of the compared sequences. We generate random matrices that have the same values of X(i), i=1,...,20 and Y(j), j=1,...,n [24].

$$X(i)= \sum_{j=1}^{n} M(i,j) \qquad (1)$$

$$Y(j)= \sum_{i=1}^{20} M(i,j) \qquad (2)$$

The X(i) value is equal to the number of amino acids of i type in the protein sequence. The Y(j) value is equal to the number of S(j) letters in the artificial sequence.

Suppose that we calculate the probability of random similarity between two sequences. Let the mutual information between the compared sequences be I(1) and the matrix M(20,n) where 20 is the size of alphabet used for the amino acid sequence and n is the size of alphabet of the artificial sequence. Let the sum of all elements of matrix M be L. First of all we generate a P1 sequence of length L. The first X(1) elements of the P1 sequence are equal to one, the elements from X(1)+1 to X(1)+X(2) are equal to two and so on. The last elements from L-X(20)+1 to L are equal to 20. After this we generate a random sequence P2 of length L. Then the sequence P2 is sorted into ascending rank, moving the value in sequence P1 along with the number in sequence P2. Let us introduce the new matrix M'(20,n). The randomised value of each M'(i,1) is obtained by computing the number of ones, twos, threes, etc. in the first Y(1) elements of the P1 sequence. Similarly, the number of each M'(i,2) is obtained by considering the number of ones, twos, threes, etc. in the elements of the P1 sequence from Y(1)+1 to Y(1)+Y(2). The same procedure is repeated for all sites, resulting in a new data matrix M'(20,n) that is randomised but in which the column totals and row totals are equal to those of the original M matrix [24].

We generate 500 matrices M'(20,n) for each period length from two to L/2. Than we calculate the mutual information I' for each M' matrix and determine the distribution F(I'). Distribution F(I') shows the number M'(20,n) matrices that lie in any interval (I'-0.5; I'+0.5). It is suitable to take a value $Z=(I(1)-I'(average))/\sqrt{D}$ as a quantitative measure of periodicity in an amino acid sequence. Here I'(average) is an average value of I' for many M' matrices, D is the variance of I' calculated with help of the F(I') distribution. If Z is low or equal to zero, then the similarity between artificial and protein sequences is absent and the protein sequence does not have any periodicity. Normal distribution has been used for the estimation of the probability P(Z>X) for X more 5.0 [25]. We selected protein sequences with a high level of Z. If Z is equal to 5.8 and more, the probability of revealing of a random periodicity in an amino acid sequence is less than 10^{-8}. We tested our method using random amino acid sequences generated artificially. We analysed $5x10^4$ random sequences of length 2000. Z <5.8 was obtained for any case of amino acid periodicity in random sequences. For the analysis of the SWISS-PROT data bank, we can consider periodicity of amino acid sequences with $Z \geq 5.8$ as statistically important.

We calculated the spectrum of Z for different period length n for the analysed protein sequence. This spectrum shows the presence of different amino acid periods in the analysed protein sequences. The Z(n) spectrum is calculated for period lengths up to L/2. This gives us the chance not only to find the latent protein periodicity, but also all duplications in a protein sequence. The Z(n) spectrum shows the fundamental frequency and its harmonics as would be shown by an autocorrelation function. We can consider the Z(n) spectrum as a function similar to an autocorrelation function in the method of Fourier transformation [8,9], but with four essential advantages.

1. The Z(n) spectrum does not require any type of mapping of the amino acids to numbers.

2. The Z(n) spectrum shows any type of periodicity of a protein sequence including periodicity with similarity between periods and latent periodicity.

3. The Z(n) spectrum allows the simple estimation of the statistical importance of the periodicity found.

4. The Z(n) spectrum is consistent for extremely low lengths of the analysed amino acid sequences.

The set of amino acids used for each position in a period can be different for different sequences and for different period lengths. The specific type of amino acid presented in each position of a period may be determined by analysing the M(20,n) matrix. We can consider the M(20,n) matrix as the type of periodicity found.

We use a sliding window with length equal to 1000 amino acids for the analysis of amino acid sequences from the SWISS-PROT data bank. If the length of an analysed sequence is less than 1000 amino acids, we take the window length equal to the length of a sequence. Periodicity can occupy a part of an amino acid sequence only. To find the amino acid region with the most expressed periodicity, we should test all positions of the left and right borders in the window and we should calculate Z for each position of the borders. We take a step for a rough shift of the borders equal to 20 amino acids and select the best positions of the left and right borders. The best positions produce the maximum value of Z(n). After this we define the exact border positions by variation of the borders with a shift of one amino acid. Exact determination requires about 40x40=1600 calculations of Z for each period length. These calculations allow us to find the exact position of the amino acid region with periodicity in a window, but increase the time of analysis. After analysis of one window position, we move a window on 500 amino acids and repeat the calculations. We selected all amino acid sequences from the SWISS-PROT data bank with Z(n)>5.8.

Time for analysis of amino acid sequence is proportional to L_1L_2N, for $L_1 > L_2$, where L_1 is the length of the amino acid sequence to be analysed, L_2 is the length of the window and N is the number of periods tested. We analysed periods from two to $L_2/2$ amino acids for $L_1 > L_2$. For example, the time for analysis of the periodicity of one sequence of length 1000 amino acids was about 2 minutes for a Pentium 200 processor.

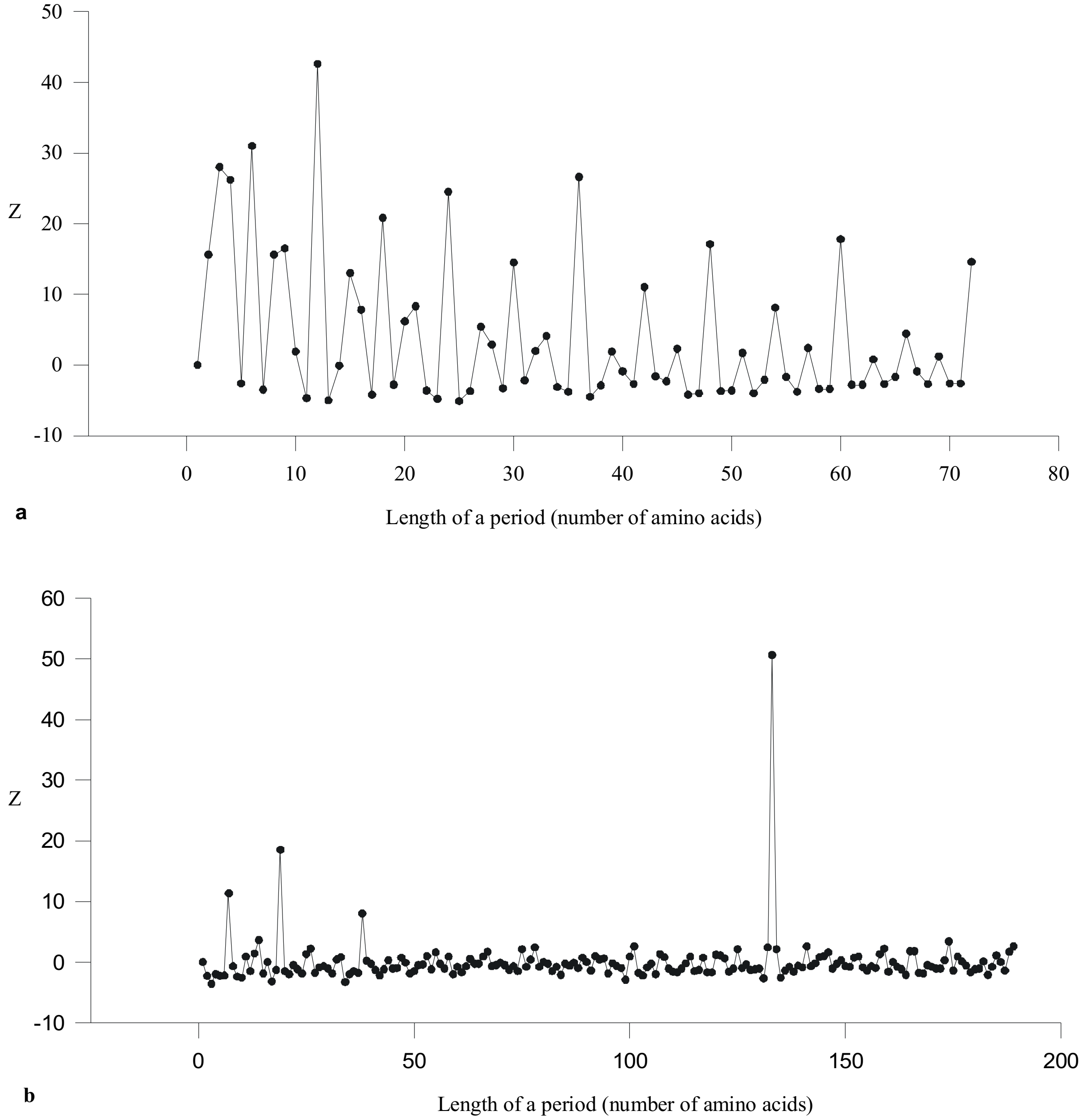

Figure 1 *The Z values for different period lengths of artificial sequences for the sequences with perfect periodicity from the 110K_PLAKN sequence (a) [30] and the ABA1_ASCSU sequence (b) [31-32] of the SWISS-PROT data bank. The ordinate axis shows the Z value. The abscissa axis shows the period length in the amino acids*

For determination of the biological sense of the revealed latent periodicity, we use a weight method for searching the periodicity determined by the M(20,n) matrix in the amino acid sequences with known function in SWISS-PROT data bank. We used the M(20,n) matrix for calculation of the p(i,j) probabilities for the occurrence of j type of amino acids at i period position. t(j) is defined as an estimate of the probability that amino acid j occurs within the sequence with periodicity (0≤t(j)≤1):

$$t(j)=n(j)/N \qquad (3)$$

Here n(j) is the number of occurrences of amino acids in the sequence, and N is the total length of the sequence. A weighting w(i,j) for each amino acid at position i is introduced as: w(i,j)=p(i,j)ln(p(i,j)/t(j)). After determining the weight matrix containing elements w(i,j), it is possible to search for periodicity in amino acid sequences from the data bank. The

Table 1 *The characterising periods for regions found with latent periodicity are shown. N— is an arbitrary amino acid. The sequence identifier, the period length, the positions of the amino acids sequences with latent periodicity in the corresponding protein, the Z value and the characterising period are shown*

Sequence identifier	Period length	Coordinates of the sequence with latent periodicity	Z	Characterising period
E2BE_YEAST	6	382-448	5.8	1 5 GlyAspAsnN—-N—Ile Leu
DLDH_AZOVI	19	151-301	5.6	1 5 10 N—-N—-AlaLeuAspLysGlnN—N—-N—- 11 15 ThrGlyN—-N—LeuGlyAlaIleN—-
LY_BPSF6	7	31-151	7.4	1 5 GlnAlaArgValSerAspLeu ThrLysSerIleAsp
FEPA_ECOLI	2	51-601	8.7	1 LeuAsn Phe

first n amino acids are a(1)a(2)...a(n) observed and the weight, W, of these sequences is determined as:

$$W = \sum_{i=1}^{n}\sum_{j=1}^{20} w(i,j)d(i,j) \tag{4}$$

This is simply the sum of the weightings for the amino acid j present at each position i of the sequence; d(i,j)=1 for j=a(i), and d(i,j)=0 for all another j. The full sum therefore contains n weightings. We select Z as the measure of the statistical importance of W. $Z=(W-W'))/\sqrt{D}$, where W' and D is the middle value and variance of W for a set of random amino acid sequences with the same amino acid frequencies as the amino acid sequence tested. If the sequence has Z>6.0 then the amino acid sequence is similar to the periodic sequences determined by the M(20,n) matrix.

We determine the "characterising period" showing the amino acids forming a latent periodicity. The characterising period is generated as follows. We select a characterising acid type for each position j in a period. The $X^2(i,j)$ value is calculated for all positions of a period using the matrix M(20,n). Let M(i,j) be the element of M(20,n) matrix, n the period length and X(i) the number of i type amino acid in the protein sequence. The $X^2(i,j)$ value is calculated for positions with $M(i,j)\geq 2$ and $M(i,j)>X(i)/n$ as:

$$X^2(i,j)=(M(i,j)-X(i)/n)^2/(X(i)/n) \tag{5}$$

For other positions we take $X^2(i,j)=0$. We show all amino acids with $X^2(l,k) \geq 4.0$ in descending order of $X^2(l,k)$ for all positions k with some $X^2(l,k) \geq 4.0$. If there is no element $X^2(i,k)>4.0$ for some k, then this position is marked as arbitrary by N—. Each position of the characterising period shows amino acids whose enrichment exceeds the average frequency of this amino acid in the region. The characterising periods for some regions with the latent periodicity are shown below.

Results and discussion

The method developed was applied to reveal the latent periodicity in protein sequences from the SWISS-PROT data bank. It was found that not less than 10% of all amino acid sequences from SWISS-PROT have regions with latent periodicity. Also, we found homological protein repeats and many cases of duplications inside proteins, as found earlier [12,26,27]. Firstly, we show how our method can reveal perfect periodicity in amino acid sequences. It illustrates the power of the mathematical method and the level of possible Z(n) (Figure 1a and b). The first sequence with perfect periodicity is part of the 110K_PLAKN sequence from SWISS-PROT; the sequence contains twelve direct tandem repeats ETQNTVEPEQTE [28]. It can be seen from Figure 1a that Z has a maximum for n=12 and Z(12)=44. The probability P for random occurrence of such periodicity can be estimated to be less that 10^{-100} using formula [29]: $P=(e^{-Z1})/(Z\sqrt{2\pi})$, where $Z1=Z^2$. For n=24, 36 and 48 Z is equal to 24.5, 26.6 and 17.1 correspondingly and for n=3, 4, 6, 8 and 9 we have Z equals to 28, 26.5, 31, 15 and 15.8. The period equal to twelve amino acids gives rise to these periods and makes large values of Z. However, Z(12) is the maximum value of Z for all periods.

The second example of perfect periodicity is shown in Figure 1b. The amino acid sequence from the ABA1_ASCSU

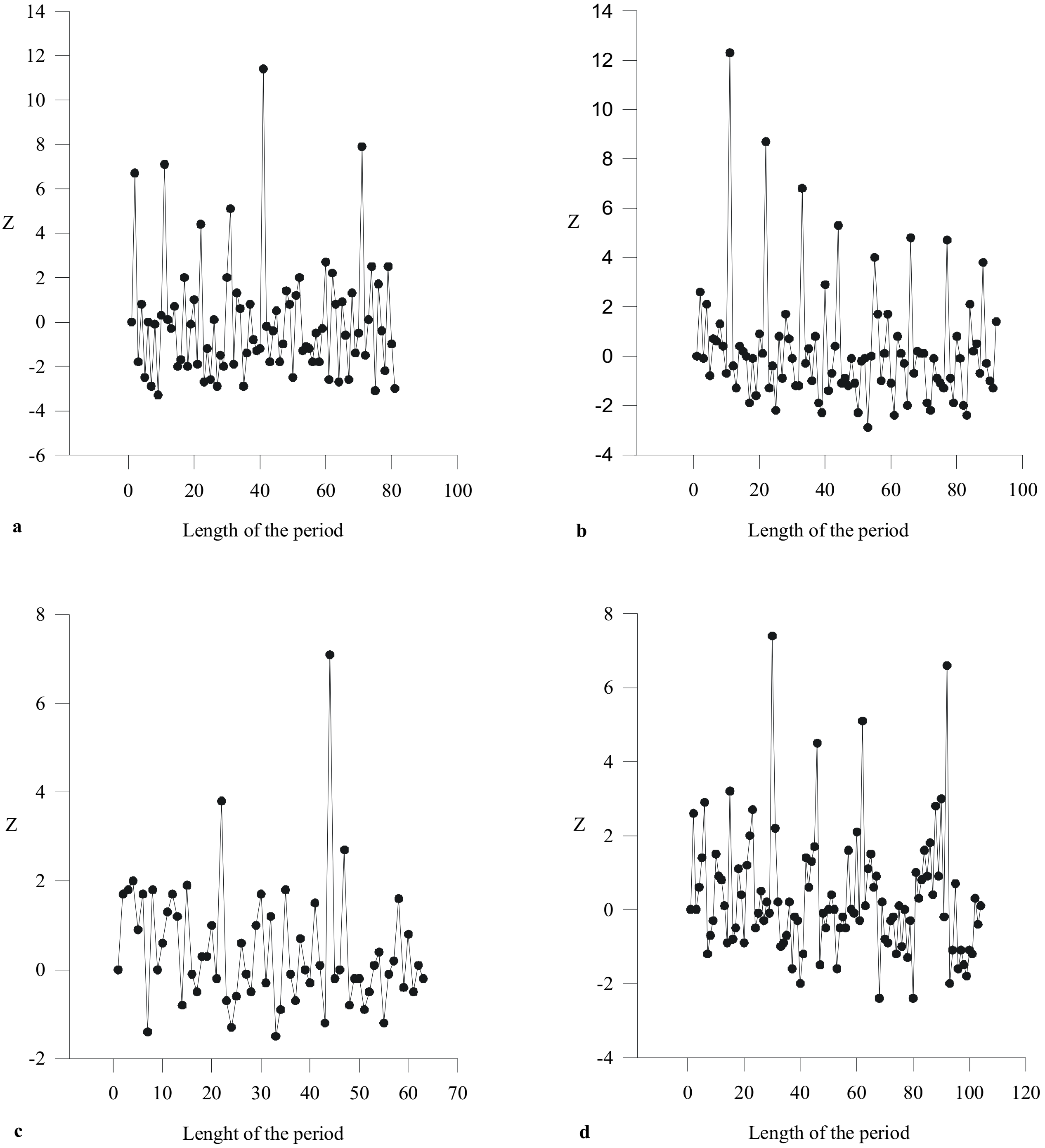

Figure 2 *Periodicity of amino acid sequences from the SWISS-PROT sequences: (a) BAR_CHITE; (b) APA1_HUMAN; (c) CGRA_MOUSE; (d) TF3A_RANPI*

sequence of the SWISS-PROT data bank [30-32] has three tandem repeats. The length of period is equal to 133 amino acids and Z(133)=50.6. The period of length 133 amino acids gives rise to Z(7)=11,3, Z(19)=18.5 and Z(38)=8.0. But it is obvious that the period equal to 133 amino acids has the maximum value of Z. Two examples show that both tandem amino acid repeats and perfect periodicity can be revealed by our method with extremely high values of Z.

We tested the periodicity of amino acid sequences from the SWISS-PROT sequences BAR_CHITE, APA1_HUMAN, CGRA_MOUSE and TF3A_RANPI (Figure 2). These sequences contain the amino sequences for the giant secretory

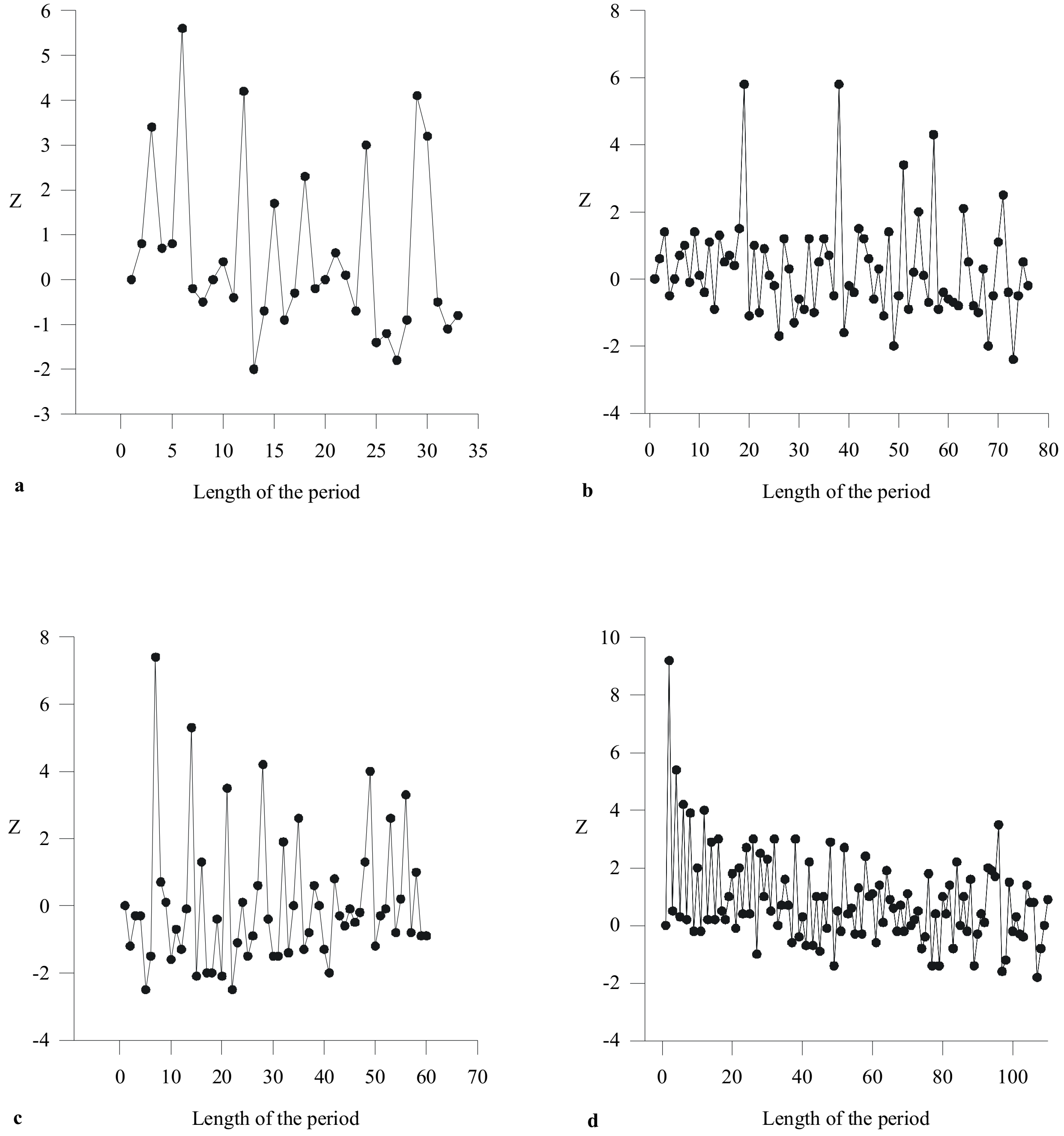

Figure 3 *Z(n) spectrum for different amino acid sequences of the SWISS-PROT data bank: (a) E2BE_AEAST; (b) DLDH_AZOVI; (c) LY_BPSF6; (d) FEPA_ECOLI. The ordinate axis shows the Z value. The abscissa axis shows the period length in the amino acids.*

protein, apolipoprotein, γ-crystallin and transcriptional factor IIIA (TF3A). The positions of sequences found with periodicity in these proteins are 1-163, 70-254, 37-163 and 1-209. The sequences found contain periods equal to eleven amino acids for BAR_CHITE and APA1_HUMAN and 44 and 30 amino acids for the CGRA_MOUSE and TF3A_RANPI SWISS-PROT sequences, respectively. These periods for amino acid sequences have been found earlier [12], but for CGRA_MOUSE and TF3A_RANPI the sequence periods are slightly different (44 and 30 in our method com-

Table 2 *M(20,n) matrices for the amino acid sequences with latent periodicity*

Type of amino acid	Sequence identifiers from the SWISS-PROT data bank			
	E2BE_YEAST	DLDH_AZOVI	LY_BPSF6	FEPA_ECOLI
	Position in period 1 5	Position in period 1 5 10 15	Position in period 1 5	Position in period 1
Lys	000021	0201040001110000101	0 2 0 0 0 0 0	11 11
Asn	207100	0000100100000000000	1 0 1 0 1 1 0	14 30
Ile	200016	0110010100100003000	0 0 0 5 1 1 1	14 14
Met	000101	0001000000000000000	0 0 0 0 0 0 0	3 5
Thr	000110	2000100020000000130	3 3 2 0 1 2 0	17 26
Arg	010020	0000000000012101000	1 1 5 1 0 2 0	13 20
Ser	121210	2000010100000000011	1 2 4 1 4 1 0	16 21
Leu	000023	2103011111013000000	0 2 0 3 1 011	34 5
Tyr	100000	0000000000000000000	0 0 0 0 0 0 0	14 7
Phe	010000	0001011000010010000	0 0 0 0 0 0 0	13 1
Cys	000110	0000000000000000000	0 0 0 0 0 0 0	1 1
Trp	000000	0000000001000000000	0 0 0 0 0 0 0	11 4
Pro	000100	0010001112000010000	0 0 0 0 0 0 0	7 13
Hys	001000	0000000000000000000	0 0 0 1 0 0 1	3 4
Gln	000000	1100003000001100000	2 2 0 0 1 2 0	14 8
Val	000200	1011000231301031301	2 1 1 7 0 0 1	21 10
Ala	120100	0041000012100031011	0 3 1 0 1 3 3	17 20
Asp	240110	0000400000120200011	1 1 1 0 4 4 0	15 21
Glu	000100	0200200000011002020	5 0 1 0 0 0 0	10 21
Gly	212000	0110002100110400203	1 0 2 0 3 1 0	28 33

pared with 43 and 36 in the Heringa and Argos method). The origin of the differences lies in the method of creation of the deletions and insertions by the Heringa and Argos method [12]. Results obtained by the proposed method agreed well with earlier results. The Z value is more than 7.0 and shows that the periodicity found is caused by similarity between periods. In addition, we found periodicity in the BAR1_CHITE sequence with periods equal to two and 41 amino acids. We can thus see that the proposed method is capable of finding all periods obtained earlier and some others.

In Figure 3 we show several examples of protein regions with latent periodicity. The positions of regions found with latent periodicity in the corresponding proteins are shown in Table 1. Regions with latent periodicity were found in the following genes:

1. translation initiation factor EIF-2B (epsilon subunit) of Saccharomyces cerevisiae from the E2BE_YEAST sequence [33];

2. E.coli ferrienterochelin receptor from the FEPA_ECOLI sequence [34];

3. lysozyme of Bacteriophage SF6 from the LY_BPSF6 sequence [35];

4. lipoamide dehydrogenase of Azotobacter vinelandii from the DLDH_AZOVI sequence [36];.

These protein sequences have latent periods of length six, two, seven and 19 amino acids respectively. M matrices are shown in Table 2. As is clear from these matrices, the set of amino acids used in each position of the latent period found contains many types of amino acids. However, there are positions with uniform amino acid composition. This is almost the case for positions three and seven of the latent periodicity from the E2BE_YEAST and LY_BPSF6 sequences, respectively.

The ”characterising periods” for all regions found with latent periodicity are shown in Table 1. Each site of the characterising sequence shows the amino acids forming the latent periodicity. For example, the third amino acid of the integrated sequence of the DLDH_AZOVI sequence isAla. This means that 3+19k (k=1,2,3,…) positions of the region with latent periodicity of the DLDH_AZOVI sequence are enriched by Ala. This enrichment exceeds the average frequency of Ala in the region. We can assume that the region with latent protein periodicity was created by duplications of the elementary amino acid sequence.

It is very important to consider the functional meaning of the protein regions where latent periodicity has been found. We attempted to reveal similar periodicity in the amino acid sequences with known biological sense. We used M(20,n) matrices from regions found with latent periodicity as the base for determination of the weight profiles. The latent pe-

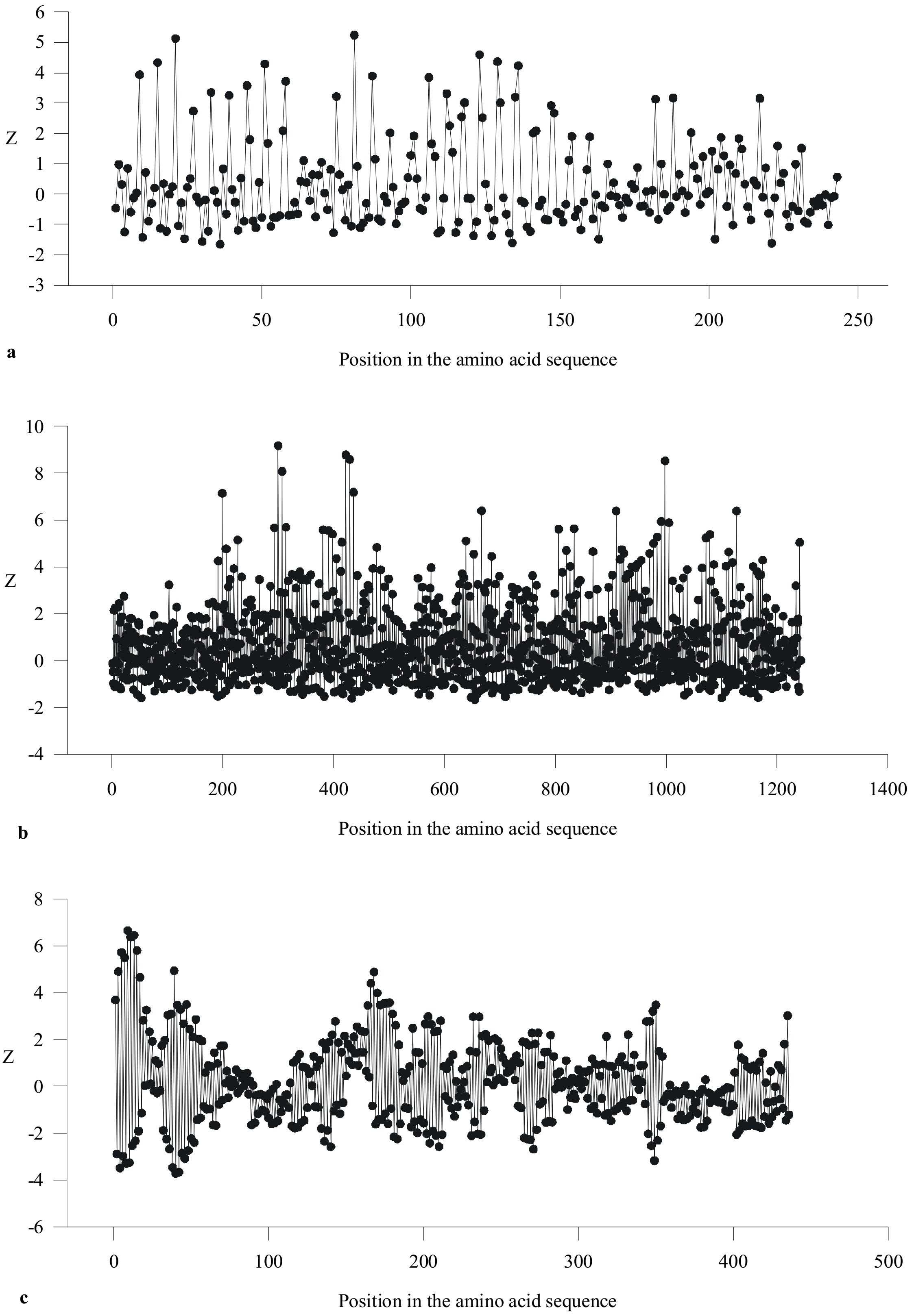

Figure 4 *Weight profiles for the sequences from the SWISS-PROT data bank are shown. (a) LPXA_ECOLI; (b) DYNA_HUMAN; (c) CPXA_ECOLI. The matrices M(20,6), M(20,7) and M(20,2) for the periodicity of the regions from E2BE_YEAST, LY_BPSF6 and FEPA_ECOLI were used for the creation of these profiles*

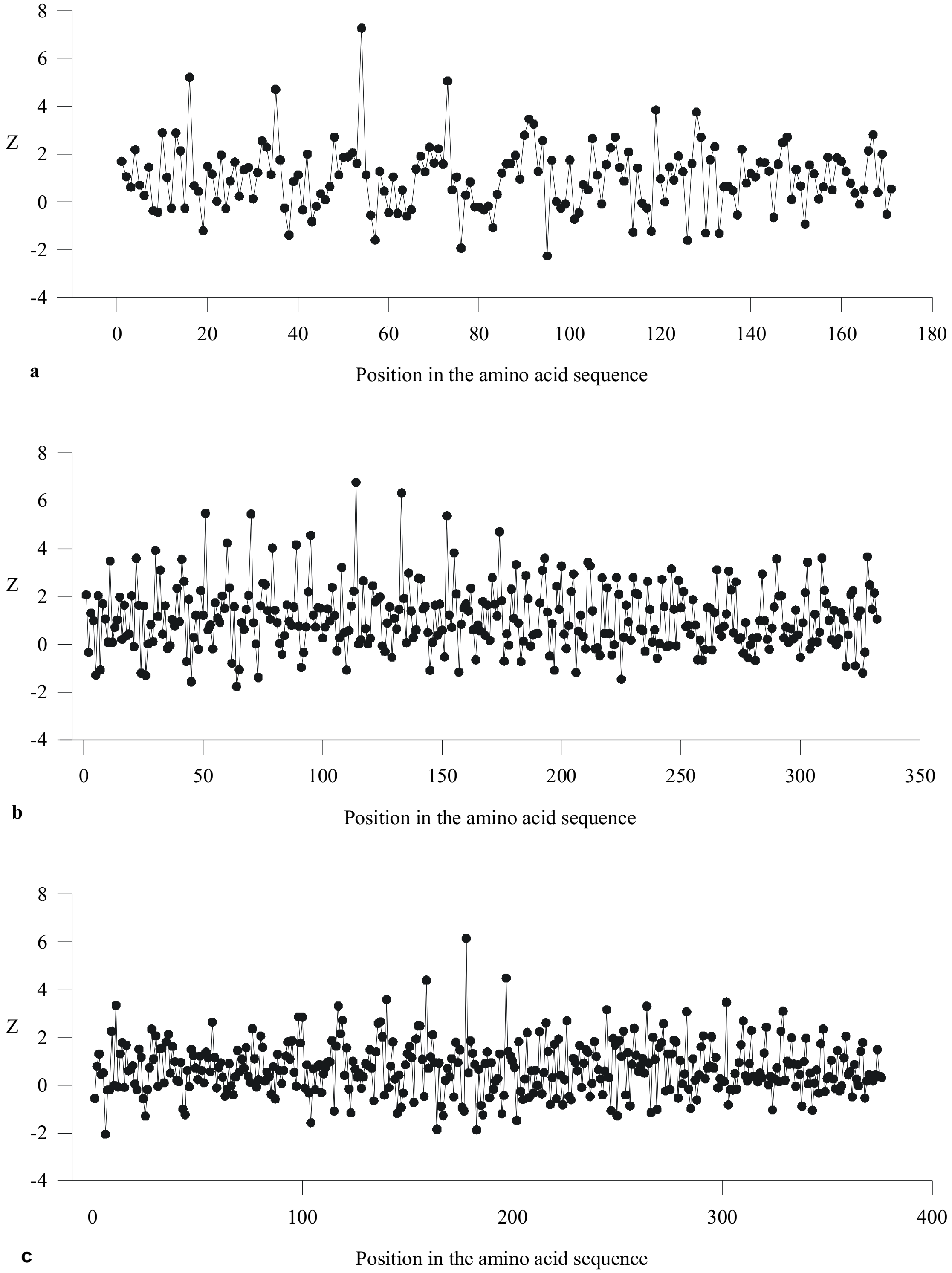

Figure 5 *The weight profiles for the sequences from the SWISS-PROT data bank are shown. (a) ENTA_ECOLI; (b) TODA_PSEPU; (c) DHAP_HUMAN. The matrix M(20,19) for the periodicity region for DLDH_AZOVI was used for the creation of these profiles*

riodicity of the amino acid sequence from translation initiation factor EIF-2B (epsilon subunit) of Saccharomyces cerevisiae (E2BE_YEAST sequence) is similar to the periodicity of the LPXA_ECOLI sequence (Figure 4). The periodicity of the LPXA_ECOLI sequence is important for the creation of the left-handed parallel β helix in the structure of UDP-N-acetylglucosamine acyltransferase [37]. The periodicity of the LPXA_ECOLI sequence contains six and four coils interrupted at two corners by external loops. We can see in Figure 5 two regions where periodicity similar to M(20,6) matrix is discovered. The first region corresponds to six coils (1-90 amino acids) and second corresponds to four coils (100-150 amino acids). Functional sense of the hexapeptide repeats found in the E2BE_YEAST sequence may be analogous to those of the hexapeptide repeats in the LPXA_ECOLI sequence and may be conditioned by the interaction of β sheets in the β helices or in the similar structures [37].

The periodicity of the lysozyme of Bacteriophage SF6 from the LY_BPSF6 sequence [35] is similar to the periodicity from the DYNA_HUMAN sequence [38,39] and corresponds to the coiled coil regions of the dynactin (206-505, 920-1029 amino acids in Figure 4). Coiled coil structures typically show the heptad motif abcdefg where the a and d positions are chiefly occupied by apolar residues, the e and g positions contain charged amino acids [40-43]. The M(20,7) matrix for periodicity in the lysozyme from the LY_BPSF6 sequence is typical for the heptad periodicity in coiled coil [40,41]. We propose that heptad periodicity may be important for creation of the protein structures by the interaction of α-helices. Heptad periodicity in protein sequences was found as 3,5 amino acid periodicity by Fourier transformation methods and significant 3,5 amino acid periodicity for the α-helix formation was also revealed earlier [4,8].

Two amino acid periodicity of the E.coli ferrienterochelin receptor from the FEPA_ECOLI sequence is found by weight profile in the CPXA_ECOLI sequence that contains the sensor protein cpxA of E.coli [44,45]. The region from the eighth to the 30th amino acid is a transmembrane domain of cpxA protein and the periodicity of this sequence is similar to the periodicity of the ferrienterochelin receptor (Figure 4). We tested the entire SWISS-PROT data bank and found more than 400 transmembrane regions from different proteins that have two amino acid periodicity of this type. We did not find this periodicity in the β-structures outside the transmembrane regions. We propose that this type of two amino acid periodicity may be a common property of some transmembrane domains.

The region with latent periodicity in the DLDH_AZOVI sequence includes the NAD$^+$-binding site of the lipoamide dehydrogenase [36]. The NAD$^+$-binding site is a highly conserved protein structure that occurs in many proteins. The region containing the NAD$^+$-binding site includes four α-helices and six β-structures, but variations in the number of structure elements are possible [46,47]. The precise structure of the NAD$^+$-binding site includes 32 amino acids and contains a βαβ-fold [48,49]. Eleven rules describe the types of amino acids that should occur at the specific positions in this peptide fragment [48]. The rules differ from that of the matrix shown in Table 1 for the DLDH_AZOVI sequence. In this study 19 amino acid repeats were found including the βα-fold of NAD$^+$-binding site. The regular alternation of β-structures and α-helices can create a latent periodicity of amino acid sequences. Duplications of 19 amino acid primary sequence could create the region containing the NAD$^+$-binding site.

We tested for the presence of 19 amino acid periodicity of NAD$^+$ regions in amino acid sequences of the SWISS-PROT data bank. More than one thousand proteins have similar periodicity. Most of these regions are observed in NAD$^+$ sites, ATP and GTP-binding sites. Three examples of the weight profiles of 19 amino acid periodicity for ENTA_ECOLI, TODA_PSEPU and DHAP_HUMAN sequences from the SWISS-PROT data bank are shown in Figure 5. These entries contain dihydro-2,3-dihydrobenzoate dehydrogenase [50], ferredoxin NAD$^+$ reductase [51] and aldehyde dehydrogenase [52]. The high level of score (Z) present in the regions that contain NAD$^+$ sites of these proteins (1-100, 120-180 and 100-200 amino acids for the ENTA_ECOLI, TODA_PSEPU and DHAP_HUMAN sequences, respectively). It should be noted that similarity between the NAD$^+$ sites from the DLDH_AZOVI and from the ENTA_ECOLI, TODA_PSEPU and DHAP_HUMAN sequences is absent. We propose that 19 amino acid periodicity of the NAD$^+$ site is important for creating a specific conformation of this site. It is possible that 19 amino acid periodicity is a specific amino acid code of the NAD$^+$ binding sites of many known proteins.

The mathematical method developed can find all cases of a perfect periodicity (without deletions and insertions) and some more diverged periodicity than can be found by the methods using Fourier transformation or algorithmic approaches. If the Z value lies between 5.8 and 7.0, we usually find latent periodicity. If the Z value lies between 7.0 and 15.0, we usually find imperfect homology periodicity. We find perfect homology periodicity for Z>15.0. The application of this method to analysis of amino acid sequences can reveal cases of a periodicity that could be missed by earlier methods (5.8<Z<7.0).

An insufficiency of the developed method in its present form is the impossibility to find periodicity with deletions and insertions. Because insertions and deletions occur very often, the majority of regions with latent periodicity will not be revealed by the present method. We consider 10% as the lowest quantity of protein sequences with latent periodicity. It is possible to modify the method of latent periodicity search in cases of low numbers of deletions and insertions (no more than one insertion or deletion per 50 amino acids) and we are developing the improved method now. In its present form our method could become a good addition to existing methods for the search of very ancient tandem diverged repeats in amino acid sequences.

A method for discovering latent amino acid periodicity had also been developed earlier [4]. However, it can find latent periodicity only when the sequence has the same set of used amino acids for all positions of the period. Moreover, it is necessary to know the set of used amino acids before a protein sequence can be analysed by this method. One should make very complicated calculations in this case, because the number of possible sets of used amino acids in all positions of periods is very high. Analysis of a large number of sets of used amino acids is required and it is hardly attainable, even for modern powerful computers. Our autocorrelation function calculation allows us to find any type of protein periodicity without assumptions about sets of used amino acids before calculations and obtained sets of used amino acids for different positions of a period may be different. These advantages are obtained by using a matrix M(20,n) and the mutual information as the measure of correlation of the amino acids and the letters of the artificial periodical sequences.

Thus, the results obtained show that latent periodicity often exists in protein sequences. Latent periodicity has been found in DNA sequences of different genes [19-21]. It is possible that significant part of coding regions has latent periodicity. The meaning of a found periodicity may be different, in some cases latent periodicity can be related with the laws of a protein spatial organisation. We can assume that several alternating of βα-folds and some other protein structures require latent periodicity of amino acids sequences. Periods can be considered as elementary ”bricks” of certain protein structures. We suggest the existence of different codes that can be written in protein sequences and can determine amino acid structures and biological functions. This idea was proposed earlier [8]. In our cases pair amino acid periodicity may code some class of transmembrane regions of proteins, 19 amino acid periodicity may code a wide class of NAD^+ binding sites, heptad periodicity code the α helices and hexapeptide periodicity determine the protein structures where the specific interactions between β-sheets determine the protein structure.

Acknowledgements This work was supported by Russian Fund of Fundamental Investigation, Grant 98-01-00205.

References

1. McLachlan, A.D. *Biopolymers*. **1977**, *16*, 1271.
2. Cornette, J.L.; Caese, K.B.; Margalit, H.; Spouge, J.L.; Berzofsky, J.A.; Delisi, C. *J.Mol.Biol*. **1987**, *195*, 659.
3. Kolaskar, A.S.; Kulkarni-Kale, U. *J.Mol.Biol*. **1992**, *223*, 1053.
4. McLachlan, A.D. *J.Phys.Chem*. **1993**, *97*, 3000.
5. McLachlan, A.D.; Stewart, M. *J.Mol.Biol*. **1994**, *235*, 1278.
6. Makeev, V.Ju.; Tumanyan, V.G. *Comput. Appl. Biosci*. **1995**, *12*, 49.
7. Chechetkin, V.R.; Lobzin, V.V. *J.Theor.Biol*. **1998**, *190*, 69.
8. Weiss, O.; Herzel, H. *J.Theor.Biol*. **1998**, *190*, 341.
9. Rackovsky, S. *PNAS*. **1998**, *95*, 8580.
10. McLachlan, A.D. *J.Mol.Biol*. **1983**, *169*, 15.
11. Boswell, D.R.; McLachlan, A.D. *Nucl. Acid. Res*. **1984**, *12*, 457.
12. Heringa, J.; Argos, P. *Proteins* **1993**, *17*, 341.
13. Huang, X.; Hardison, R.C.; Miller, W. *Comput. Applic. Biosci*. **1990,** *6*, 373.
14. Lawrence, C.E.; Altschul, S.F.; Boguski, M.S.; Liu, J.S.; Neuwald, F.; Whotton, J.C. *Science* **1993**, *262*, 208.
15. Benson, G.; Waterman, S. *Nucl.Acid Res*. **1994**, *22*, 4828.
16. Benson, G. *J.Comput. Biol*. **1997**, *4*, 351.
17. Herzel, H.; Grobe, I. *Physica A*. **1995**, *216*, 518.
18. Needleman, S.B.; Wunsch, C.D. *J.Mol.Biol*. **1970**, *48*, 443.
19. Korotkov, E.V.; Korotkova, M.A. *DNA Sequence* **1995**, *5*, 353.
20. Korotkov, E.V.; Phoenix, D.A. *Proceedings of Pacific Symposium on Biocomputing 97*; Word Scientific: Maui, Hawaii, USA, 1997; p 222.
21. Korotkov, E.V.; Korotkova, M.A.; Tulko, J.S. *Comput. Appl. Biosci*. **1997**, *13*, 37.
22. Korotkov, E.V.; Korotkova, M.A. *DNA Research*. **1996**, *3*, 157.
23. Kullback S. *Information theory and statistics*; Wiley: London, UK, 1959.
24. Roff, D.A.; Bentzen, P. *Mol.Biol.Evol*. **1989**, *6*, 539.
25. Hudson, J.D. *Statistics for physics*; MIR: Moscow, 1967.
26. Heringa, J. *J. Comput.Chem*. **1994**, *18*, 233.
27. Heringa, J.; Taylor, W.R. *Current Opin. in Struc.Biol*., **1997**, *7*, 416.
28. Perler, F.B.; Moon, A.M.; Qiang, B.Q.; Meda, M. Dalton, M.; Card, C.; Schmidt-Ullrich, R.; Wallach, D.; Lynch, J.; Donelson, J.E.; *Mol. Biochem. Parasitol*. **1987**, *25*, 185.
29. Feller, W. *An introduction to probability theory and its applications*; Wiley: New-York, USA, 1970.
30. Christie, J.F.; Dunbar, B.; Davidson, I.; Kennedy, M.W. *Immunology*. **1990**, *69*, 596.
31. Spence, H.J.; Moore, J.; Brass, A.; Kennedy, M.W. *Mol. Biochem. Parasitol*. **1993**, *57*, 339.
32. Kennedy, M.W.; Brass, A.; Mccruden, A.B.; Price, N.C.; Kelly, S.M.; Cooper, A. Biochemistry. **1995**, 34, 6700.
33. Bushman, J.L.; Asuru, A.I.; Matts, R.L.; Hinnebusch, A.G. *Mol. Cell. Biol*. **1993,** *13*, 1920.
34. Lundrigan, M.D.; Kadner, R.J. *J. Biol. Chem*. **1986**, *261*, 10797.
35. Verma, M. *Curr. Microbiol*. **1986**, *13*, 299.
36. Westphal, A.H.; de Kok, A. *Eur. J. Biochem*. **1986**, *72*, 299.
37. Raetz, C.R.H.; Roderick, S.L. *Science* **1995**, *270*, 997.
38. Tokito, M.K.; Howland, D.S.; Lee, V.M.; Holzbaur, E.L. *Mol. Biol. Cell* **1996**, *7*, 1167.
39. Holzbaur, E.L.; Tokito, M.K. *Genomics* **1996**, *31*, 398.
40. Conway, J.F.; Parry, D.A.D. *Int.J. Biol. Macromol*. **1990**, *12*, 328.
41. Conway, J.F.; Parry, D.A.D. *Int.J. Biol. Macromol*. **1991**, *13*, 14.
42. Cohen, C.; Perry, D.A. *Science* **1994**, *263*, 488.
43. Brown, J.H.; Cohen, C.; Parry, D.A.D. *Proteins* **1996**, *26*, 134.
44. Weber, R.F.; Silverman, P.M. *J.Mol.Biol*. **1988**, *203*, 467.
45. Rainwater, S.; Silverman, P.M. *J.Bacteriol*. **1990**, *172*, 2456.

46. Lilias, M.G.; Branden, C.I.; Banaszak, L.J. In: *The enzymes.*, 3rd ed.; Academic Press: London, UK, 1975; Vol. 10, p 68.
47. Thekkumkara, T.J.; Pous, G.; Sumair, M.; Jentoft, J.E.; Patel, M.S. In *Alpha-keto acid dehydrogenase complexes. Organization, regulation and biomedical ramifications*; Roche, T.E,; Patel, M.S., Eds.; Annals of the New-York Academy of Sciences: New-York, USA, 1989; Vol. 573, p 113.
48. Wierenga, R.K.; Terpstra, P.; Hol, W.G. *J.Mol.Biol.* **1986**, *187*, 101.
49. Rice, D.W.; Schulz, G.E.; Guest, J.R. *J.Mol.Biol.* **1984**, *174*, 483.
50. Nahlik, M.S.; Brickman, T.J.; Ozenberger, B.A.; McIntosh, M.A. *J.Bacteriol.* **1989**, *171*, 784.
51. Zylstra, G.J.; Ginson, D.T. *J.Biol.Chem.* **1989**, *264*, 14940.
52. Hsu, L.C.; Chang, W.C.; Shibuya, A.; Yoshida, A. *J.Biol.Chem.* **1992**, *15*, 3030.