Biochemistry (Moscow), Vol. 71, No. 1, 2006, pp. 18-31. Translated from Biokhimiya, Vol. 71, No. 1, 2006, pp. 26-41.

Originally published in Biochemistry (Moscow) On-Line Papers in Press, as Manuscript BM05-052, October 16, 2005.

Identification of Latent Periodicity in Amino Acid Sequences of Protein Families

V. P. Turutina[1], A. A. Laskin[2], N. A. Kudryashov[2], K. G. Skryabin[1], and E. V. Korotkov[1,2]*

[1]Bioengineering Center, Russian Academy of Sciences, pr. 60-letiya Oktyabrya 7/1, 117312 Moscow, Russia; fax: (7-095) 135-0571; E-mail: katrin22@mtu-net.ru
[2]Moscow Physical Engineering Institute (Technical University), Kashirskoe Shosse 31, 115409 Moscow, Russia

Received February 24, 2005
Revision received May 16, 2005

Abstract—For detection of the latent periodicity of the protein families responsible for various biological functions, methods of information decomposition, cyclic profile alignment, and the method of noise decomposition have been used. The latent periodicity, being specific to a particular family, is recognized in 94 of 110 analyzed protein families. Family specific periodicity was found for more than 70% of amino acid sequences in each of these families. Based on such sequences the characteristic profile of the latent periodicity has been deduced for each family. Possible relationship between the recognized latent periodicity, evolution of proteins, and their structural organization is discussed.

Key words: latent periodicity, alignment, information decomposition

Study of periodicity of amino acid sequences may help to better understand structural organization of proteins and their evolutionary relations. In 1970, S. Ohno suggested that the main evolutionary mechanism involves duplication and divergence of DNA nucleotide sequences [1]. At the level of genes, such mechanism may create new encoding sequences by duplication of relatively short DNA sequences (of several dozens or hundreds of nucleotides in length) followed by their evolutionary divergence [2, 3]. If the process of new gene formation by multiple tandem duplications was rather massive, the encoding nucleotide sequences and their corresponding amino acid sequences would preserve signs of tandem duplications in the form of certain periodicity.

It is reasonable to suggest that such periodicity would be hard to recognize in modern genes because of significant divergence of an initial sequence due to accumulation of base substitutions, deletions, and inserts. We may also suggest that nucleotide sequences of modern genes derived from a common predecessor resulting from multiple tandem duplications will be characterized by such dissipated periodicity, which will influence amino acid sequences. This hypothesis implies that amino acid sequences exhibiting the same biological functions should possess similar (although weakly expressed) periodicity. For example, such amino acid sequences would represent the sequences of an NAD^+-binding domain, sequences of active sites of various protein kinases, and many other proteins.

However, mathematical methods and approaches have revealed periodicity only in a relatively small class of proteins [4-17]. All repeats found in amino acid sequences can be subdivided into three groups [18]. The first group includes sequentially duplicated structurally and functionally independent units, which may operate independently. Zinc finger domains represent a good example of such repeats. The second group includes well recognizable repeats constituting a single functional subunit; these repeats do not function independently. A well-known β-barrel pattern six amino acid residues in length is an example of these repeats. The β-barrel patterns form a structure that can be defined as a left-handed β-helix or a three-edged prism, and each repeat forms the β-sheet constituting part of the edge of this prism. There are many examples illustrating formation of such structures by repeats of longer periodical length [19]. The third less studied group of amino acid repeats contains sequences of proteins lacking significant intrinsic homology. This class of repeats can be defines as the class of repeating motifs because it is only possible to elucidate some order in the position of amino acid residues. This phenomenon is observed in the case of the leucine

Abbreviations: ID) information decomposition; ND) noise decomposition; PWM) position weight matrix.

* To whom correspondence should be addressed.

0006-2979/06/7101-0018

residue of the leucine zipper [20] or in distribution of amino acid residues exhibiting some common properties (e.g., hydrophilicity or hydrophobicity) on periodic positions.

This classification shows that the level of repeat homology inside protein domains probably depends on the evolutionary time of their existence. Similarity of recently formed repeats is evident; however, it seems likely that long-term life of basic types of enzymes hid repeats in their amino acid sequences from traditional methods of search. The existence of hidden repeats in protein sequences can also be attributed for convergent processes determined by structure–functional features of protein folds. Unfortunately, consequences of functional divergence of repeats and consequences of convergent processes cannot be distinguished without the use of additional information, which may not be available in a particular case.

A reasonable question arises: is hidden periodicity related to gene formation by multiple tandem duplication actually present in amino acid sequences? Since the rate of tandem duplications in genes is higher than the rate of convergent processes, any evidence for the existence of hidden periodicity may be indirect and related to discovery of massive cases of hidden periodicity in proteins. Lack of information on primary structure of various protein families can be explained by the fact that mathematical methods used for such search in symbolic sequences are not entirely adequate and therefore they cannot find weakly defined periods. We have previously shown [21, 22] that methods for search of periodicity in symbolic sequences based on Fourier transformation and dynamic programming have some important limitations. These limitations do not allow finding weakly defined periodicity in symbolic sequences. For example, dynamic programming can mainly detect periodicity with inserts and deletions of symbols; however, this requires high level of homology between repeats, which is set by use of certain weight matrices for symbol coincidence (like PAM or BLOSUM). Methods based on Fourier transformation cannot reveal periodicity with deletions and inserts of symbols as well as weakly defined periodicity with the length of period comparable or even longer than the alphabet size of the analyzed symbolic sequence [21, 22].

For the search of weakly defined periodicity or hidden periodicity in symbolic sequences, we have previously developed the method of Information Decomposition (ID), which lacks many shortcomings of methods based on Fourier transformation and dynamic programming [21, 22]. Our method is based on the use of a new mathematic measure for determination of similarity between compared symbolic sequences. We have used the mutual information measure as such a measure [23]. Such mode for calculation of similarity degree can ignore preexisting weights of coinciding symbols (determined by the matrices like PAM or BLOSUM). This allows not only homologous sequences but also sequence similarity to be detected by simple and complex recoding [23]. Use of the ID method revealed amino acid sequences with hidden periodicity of various lengths among proteins of the Swiss-prot database. The periodicity matrix showing number of amino acids found in various positions of the period is an important characteristic of recognized periodicity (see the "Methods and Algorithms" section).

The ID method recognizes weakly defined periodicity in various symbolic sequences [21-23]. However, this method in its present state cannot detect weakly defined periodicity in the presence of symbol inserts or deletions. Consequently, we can find hidden periodicity only in a small proportion of amino acid sequences of a protein family. So we have developed the method of Noise Decomposition (ND) for detection of weakly defined periodicity of a given type by the method of dynamic programming (the type is set by a periodicity matrix) [24]. The ND method uses the periodicity matrix (initiating matrix) for hidden periodicity found by the ID in an amino acid sequences (initiating sequence) of some protein family as input data. Based on this matrix the position weight matrix (PWM) is calculated; this PWM is then used to search for cyclic alignments in Swiss-prot database versus PWM by means of modified profile analysis [25, 26]. The cyclic alignments obtained versus PWM are further separated into false and true alignments on the basis of relationship of proteins where they have been found to the protein family of the initiating sequence. False alignments are then considered as noise. On the basis of such separation a new PWM is then constructed. The goal for a new PWM construction consists in the most exhaustive discovery of amino acid sequences related to the protein family of the initiating sequence among many true alignments and maximally possible reduction of number of statistically significant false alignments (see the "Methods and Algorithms" section for details). Such effect is usually achieved after 3-10 rounds of sequential scanning of the Swiss-prot database versus always-new PWM [24].

Combined use of ID and ND detected weakly defined or hidden periodicity in more than 80% of NAD^+-binding domains [24] and also in 16 various protein families [27]. Analysis of our previous results raises the following question: does the existence of hidden or weakly defined periodicity of amino acid sequences containing active sites of various proteins reflect a common feature? So the major goal of this study was to demonstrate that the hidden periodicity is typical not only to separate proteins, but to 70-100% of amino acid sequences belonging to one protein family. In the present study, we demonstrate data on specific hidden periodicity for 94 various protein families differing in their biological functions. These results support the hypothesis of evolutionary origin of genes by multiple tandem duplication.

METHODS AND ALGORITHMS

Hidden periodicity. Suppose that we have a sequence S, which consists of N subsequences S^i ($i = 1, 2, ..., N$) of equal length L:

$$S \equiv S^1S^2 \ldots S^N \equiv \{s_1^1 s_2^1 \ldots s_L^1 s_1^2 s_2^2 \ldots s_L^2 \ldots s_1^N s_2^N \ldots s_L^N\},$$

where s_i^j is an amino acid residue. Suppose we are looking for a period of the length L in the sequence S ignoring possibilities of inserts and deletions of symbols. We should evaluate global homology between subsequences S^i ($i = 1, 2, ..., N$) to find such period. If global homology between subsequences S^i is statistically significant, we can conclude that symbol periodicity with period L exists in the sequence S. Possibilities for finding periodicity in the sequence S will depend on the mode of insertion of the quantitative measure determining similarity of subsequences S^i. At the present time, the method of dynamic programming and Fourier transformation are often used for these purposes. For introduction of quantitative measure for global homology of subsequences S^i these methods used search for homology between these subsequences S^i. Using the method of dynamic programming the search for homology is set by PAM or BLOSUM matrices [28, 29]; in these matrices, weight of coinciding amino acids is always higher than the weight of non-coinciding amino acids, and during use of Fourier transformation the search for homology is set by laws of autocorrelation function construction [30]. Earlier we showed [21] that quantitative measures using homology search between subsequences S^i can miss hidden periodicity of the length L in the sequence S because of lack of statistically significant homology between subsequences S^i, and periodicity can be recognized only at a sufficient number of S^i periods.

Consider a notion of hidden periodicity in more detail. For introduction of the quantitative measure of subsequences S^i homology, we need to construct multiple alignment without inserts and deletions of symbols and put them in sequential order. The total weight of this multiple alignment, which can be considered as the quantitative measure of similarity of subsequences S^i, can be introduced as the sum of weights of all positions:

$$W = \sum_{i=1}^{L} W_i . \qquad (1)$$

The traditional approach calculates weight of position as the sum of all possible pairs of amino acids that could be found between compared sequences:

$$W_i = \sum_{\alpha} \sum_{\beta>\alpha} P(S_i^\alpha, s_i^\beta), \qquad (2)$$

where α and β show numbers of subsequences S^i; P is some weight matrix such as PAM or BLOSUM. This expression can be also introduced as the following sum:

$$W_i = 0{,}5 \sum_{l,k} m(i, l)[m(i, k) - \delta_l^k]P(l, k) , \qquad (3)$$

where $\delta_l^k = 1$ (at $l = k$) and 0 (in all other cases). The function δ_l^k is introduced to exclude similarity of the subsequence S^i to itself from consideration. Variable values l and k show a type of amino acid; $m(i, l)$ represents amount of amino acid type l in the position i at multiple alignment. Earlier we proposed another measure of similarity [31-33], which may be defined as the "information content" [34]:

$$W_i' = \sum_{l=1}^{20} m(i,l) \ln \frac{Km(i,l)}{x(i)y(l)} , \qquad (4)$$

where $K = NL$, $x_i = \sum_{l=1}^{20} m(i,l)$, and $y_1 = \sum_{i=l}^{L} m(i,l)$.

It is clear that the measures of homology of subsequences S^i determined by formulas (3) and (4) are different and so alignment may have higher weight using formulas (1) and (4) and lower weight using formulas (1) and (3) and *vice versa*. However the term "high weight" is lowly informative especially during comparison of weights determined by means of different mathematical measures. For each of the introduced measures we should determine probability p of that weight (higher or equal to W) would be found during alignment of purely casual sequences. In the case of periodicity search of length L in R independent sequences S (e.g., analysis of R sequences from Swiss-prot database) the probability f should be considered as p probability:

$$f = 1 - (1 - p)^R. \qquad (5)$$

For evaluation of probability p during search for period of length L in the sequence S using the measures determined by formulas (3) and (4), we may accidentally mix initial sequence S and create many random sequences Q_i ($i = 1, 2, ...$) with length equal to sequence S. Using many sequences Q_i we can determine the value Z as:

$$Z = \frac{W - E(W)}{\sqrt{D(W)}} , \qquad (6)$$

where $E(W)$ and $D(W)$ are the mean and dispersion of the weight W, respectively; they are calculated for many random sequences Q_i. High values of Z correspond to all lower values of probability p; they suggest the existence of significant similarity between subsequences S^1, S^2, ..., S^N. For evaluation of only non-accidental similarity some threshold Z value with probability $p < 0.05$ is usually determined. Homologies with Z value exceeding threshold level are considered as non-accidental.

As mentioned above, different weights and different Z values may give differences in homology degree. In

some cases, periodicity may be evident using the information measure (formula (4)) and may be missed by the methods of homology search (formula (3)). Let us define probability p, which corresponds to Z value calculated by formulas (3), (1), and (6) as α and define probability p, which corresponds to Z value calculated by formulas (4), (1), and (6) as β. Let us also assume that the sequence S contains hidden periodicity of length L provided that probability value $\alpha > 0.05$ and probability value $\beta < 0.05$. Suppose that the sequence S contains periodicity related to homology between subsequences S^i at $\alpha < 0.05$ irrespectively to probability values β. Let us also suppose that the sequence S lacks periodicity of the length L at α and $\beta > 0.05$.

In reality such difference in probabilities α and β can be realized quite often if we analyze rather long sequence and we meet some multiplicity smaller or equal to the size of the alphabet used rather than a single symbol in each position. In this case, the number of homologous coincidences can be relatively small for each position. Since weights of homologous coincidences in BLOSUM or PAM type matrices are significantly higher than weights of non-homologous coincidences, final W value can be relatively small at small number of homologous coincidences; this will provide sufficiently high value of α. At the same time, β is determined on the basis of deviations of symbol frequencies in each position i from the symbol frequencies determined for all the sequence S analyzed. Such deviations can be significant which will result in small and statistically significant value of β at high and statistically insignificant value of α.

As an example let us consider two amino acid sequences; one of them possesses perfect periodicity whereas the other one has hidden periodicity. Determine α and β probabilities using the method of Monte-Carlo. For each amino acid sequence, we generate 500 random sequences with the same amino acid composition by random transposition of amino acid residues over the whole amino acid sequence. Using formulas (3) and (1) we calculate weight means W for each of 500 randomly generated sequences and then determine $E(W)$ and $D(W)$ and finally calculate Z value by formula (6). The same calculations we make for the weight determined by formulas (4) and (1). In the result we will have two Z values for each amino acid sequence: one value has been calculated by formulas (3) and (1) and the other by formulas (4) and (1).

Let us take an amino acid sequence containing 20 (N = 20) tandem repeats ANDKVHG as the first sequence. In subsequent consideration, this sequence of 140 amino acid residues in length will play a role of the sequence S and the length of subsequence S^i consists of seven amino acid residues; this means that we are looking for the period of seven amino acids. For the sequence S, we generate multiple alignment of periods; here they play the role of the subsequence S^i. These periods are positioned one under the other. For this multiple alignment of periods Z value calculated by formulas (3), (1), and (6) using BLOSUM50 matrix is 54.5 ± 6.6, whereas Z value calculated by formulas (4), (1), and (6) is 52.2 ± 4.5. These calculation show that in the case perfect periodicity both methods of calculation of the weight function give very similar results, which are indistinguishable within the error of calculation by the Monte-Carlo method.

Table 1. Matrix used for generation of artificial sequence with hidden periodicity seven amino acid residues in length. In each position of the period probabilities of selection of any amino acid are equal to each other

Position of period	Multiplicity of amino acids used in a given position of the period
1	ARMWVDLGEPK
2	TKLWPVEHFY
3	DWEYHQLGSFK
4	RMHKTISDC
5	MKPQHYVDFI
6	RNDHALIKFTG
7	DVNWMFKES

In the second example we consider an amino acid sequence given below which is characterized by the existence of hidden period of seven symbols in length (L = 7, N = 20). Assume that amino acids listed in Table 1 can be in each position of the period with equal probability.

One of possible sequences where such periodicity is actually observed is the following:

AKLRIDM**DWSMVKD**GVLDDGV**DYGRIDE**KWKIHIW**RHLRKKW**GFQSMDV**RKYDVRN**DPLDVDK**VPDMPLD**LVLDKRD**DLYDINE**EHLRKHD**GYEHQLD**EHWMKGF**AESSYRV**MEWRPRV**ETYTDLD**GWSTFID**REFDQRD.**

Period in the sequence is marked either in bold or using "ordinary" letters; however, since it is a hidden period it is impossible to find it "visually" by homology between separate periods. Let us evaluate probability values α and β for this sequence. For this sequence, Z values determined by the Monte-Carlo method are 2.5 ± 0.4 and 7.3 ± 0.7. Using normal distribution for evaluation of distribution of Z value probability $\alpha > 0.05$ and probability $\beta < 10^{-9}$, respectively. Thus, is it clear that such periodicity will be statistically insignificant if we use the weight introduced on the basis of PAM or BLOSUM matrices (formulas (3) and (1)) and this is recognized at statistically significant level using the information measure introduced by formulas (4) and (1).

Information decomposition. To find initiating sequences with hidden periodicity required for calcula-

tion of initiating matrices of decomposition we have analyzed the whole Swiss-prot database by the method of information decomposition. Details of this method are given in [21, 22]. For construction the ID spectrum for amino acid sequence S (s_1, s_2, s_3, ..., s_L) we have generated a multiplicity A (with volume of $L/2$) of artificial periodic sequences (each of L in length) with the lengths of period from 2 to $L/2$. For example, the sequence A_2 with the period of two symbols was generated as 12121212121212..., the sequence A_3 of three symbols in length was generated as 123123123..., the sequence A_k of k symbols in length was generated as 12...k12...k12...k... (numbers were considered as symbols). Afterwards we calculated mutual information between the sequence S and the sequence A_n, n = 2, 3, ..., $L/2$. For calculation of mutual information coincidence matrix $M(n)$ of $n \times 20$ in size was filled (n is the length of the period in the sequence A_n). The matrix $M(n)$ shows number of coincidences between symbols of artificial sequence and amino acids. Mutual information was calculated by the following equation:

$$I(n) = \sum_{1}^{n}\sum_{1}^{k} m(i,j)\ln m(i,j) - \sum_{1}^{n} x(i)\ln x(i) - \sum_{1}^{k} y(j)\ln y(j) + \\ + L\ln L\,, \qquad (7)$$

where $m(i, j)$ is matrix element, $x(i)$ (i = 1, 2, ..., n) is number of symbols in the sequence A_n, $y(j)$ (j = 1, 2, ..., k) is number of amino acids in sequence S. For evaluation of statistical significance of the calculated mutual information, we used the Monte-Carlo method. For each sequence A_n we generated multiplicity of random sequences Q^n by random mixing of the sequence A_n. For each period n on multiplicity Q^n we calculated the mean of mutual information $J(n)$, its dispersion $D(J(n))$, and then $Z(n)$ value by the following formula:

$$Z(n) = \{J(n) - \overline{J(n)}\} / \sqrt{D(J(n))}\,. \qquad (8)$$

Information decomposition of a symbolic sequence is thus represented as a $Z(n)$ spectrum. Searching for hidden periodicity in the Swiss-prot database was limited by the period lengths from 2 to 100 amino acids. Statistical tests revealed that the threshold for Z value during selection of statistically significant periodicity for period lengths from 2 to 100 using analysis of the whole Swiss-prot database by the ID method was 6.0 [27, 31]. Using the ID method we found more than 5000 amino acid sequences in Swiss-prot-41 database; these sequences demonstrated periodicity within the interval of period length n from 2 to 100 amino acid residues with $Z(n)$ values exceeding 6.0. These sequences were considered as initiating amino acid sequences. For each initiating sequence we determined periodicity matrix $M(n)$ and $Z(n)$, where n shows length of hidden period found. Using this approach, we generated more that 5000 initiating matrices. The matrices $M(n)$ were then used as the initiating matrix in the ND method. We ranged the $M(n)$ matrices by the increasing of order of Z value. In this study, we used the first 110 initiating matrices $M(n)$ found by the ID method (increase in Z value from 6.0 and above). Taking into consideration R tests made during Swiss-prot database search for matrices (formula (5)) the threshold for $Z(n)$ of 6.0 provided probability $\alpha > 0.05$ and $\beta < 0.05$. These matrices determined selection of 110 protein families, which we investigated in the present work by iteration analysis for presence of hidden periodicity specific for each protein family.

Iteration analysis. We used initiating matrices $M(n)$ for generation of a position weight matrix for analysis of hidden amino acid periodicity; at this stage it took into consideration possible insertions and deletions of amino acids (Scheme). Elements of the corresponding PWM U for subsequent cycling profile analysis were determined by the periodicity matrix using the following formula [22]:

$$U_{i,j} = C\ln\frac{p_{i,j} + \varepsilon f_i}{f_i + \varepsilon}\,, \qquad (9)$$

where $U_{i,j}$ is an element of the PWM for the symbol i in position j; $p_{i,j} = m(i, j)/y(j)$, $m(i, j)$ are the elements of the periodicity matrix $M(n)$, $y(j) = \sum_{i=1}^{20} m(i, j)$, and f_i is frequency of symbols type i in the amino acid sequence where hidden periodicity was found. A small number ε was included into Eq. (9) for exclusion of zero values from consideration; in all calculations $\varepsilon = 10^{-5}$. The scale coefficient C in Eq. (9) may be of any value; its concrete mean does influence either alignment pathway or its statistical significance. Usually C value is selected to be rather high, for approximation of weight values to round numbers for acceleration of calculations without loss of accuracy.

For generation of a new position weight periodicity matrix U we have used the ND method; such U matrix should reveal hidden periodicity at statistically significant level in all amino acid sequences of a protein family but miss statistically significant homology in all other amino acid sequences (Scheme). The principle of the iterative ND method consists in the following.

At the first stage, we are looking for amino acid sequences in Swiss-prot database; these sequences should have statistically significant cyclic alignment ($Z_1 > 6.0$) versus PWM U. For this purpose, we used the method of cyclic alignment and modified profile analysis described earlier [24-26]. The Z_1 value for weight alignment was calculated by the Monte-Carlo method using cyclic alignment versus multiplicity of random sequences [24-26]. Such scanning resulted in selection of statistically significant amino acid sequences, which were optimally

aligned during cyclic alignment versus PWM U. Subsequently the term "alignment" will mean cyclic alignment versus PWM.

At the second stage, we subdivided resultant alignments into two multiplicities. The first multiplicity contained amino acid sequences with Z_1 values >6.0; these sequences were characterized by the same functional meaning as the initiating amino acid sequence. This multiplicity can be defined as true alignments. Inclusion of proteins into this category was based on their description (DE field), key words (KW field), and by table of locus features (FT field) of the Swiss-prot database; these characteristics would be identical to corresponding fields of the initiating amino acid sequence. The second multiplicity included all other amino acid sequences; we defined this multiplicity as the multiplicity of false alignments.

At the third stage, we modified PWM U for two purposes. First, we wanted to change it for detection of maximally possible homologs related to multiplicity of true alignments with Z_1 level >6.0 during modified profile analysis. We also wanted to find all sequences listed in Swiss-prot database that are characterized by the same functional meanings as the initiating amino acid sequence. Second, we wanted to reduce maximally (or totally abolish) the number of homologs in the multiplicity of false alignments with Z_1 value >6.0. For these purposes, we modified Eq. (9) as follows:

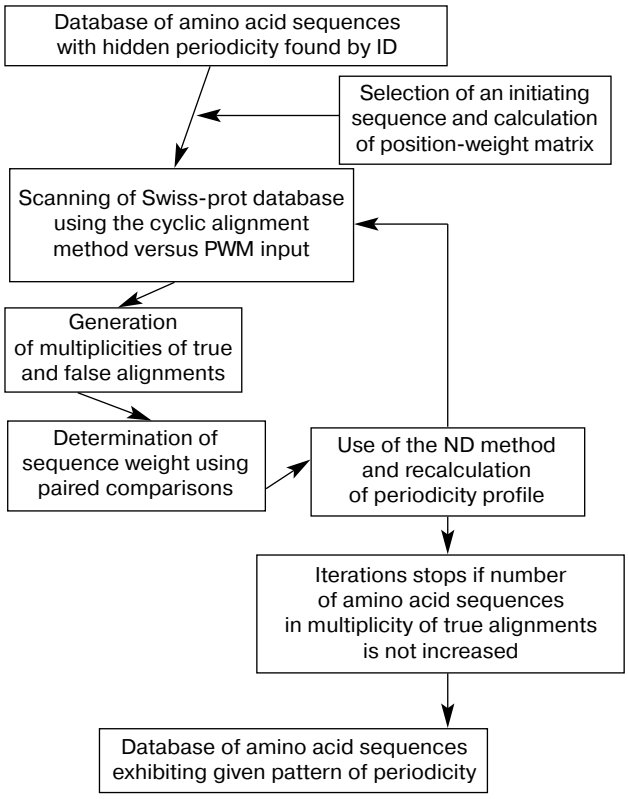
$$\overline{U}_{i,j} = C \ln \frac{r_{i,j}}{\pi_{i,j}} , \qquad (10)$$

where $r_{i,j}$ is a weighted $p_{i,j}$ value.

Multiplicity of true alignments can contain both homologous sequences and globally homologous sequences. This will result in an excessive amount of any amino acid in certain positions of the period. Taking into consideration this effect it is necessary to compare all selected amino acid sequences and to calculate weight for each sequence of multiplicity of true alignments. This weight should reflect representation of the sequence in multiplicity of true alignments. Let us define weight of alignment between sequences k and l as $S(k, l)$. After that we can calculate $T(k)$ value, which shows representation of a given sequence in multiplicity of true alignments as:

$$T(k) = \sum_l \max (0, S(k, l)/\{\max(S(k, k), S(l, l))\}). \quad (11)$$

The index l runs over the whole multiplicity of true alignments. In the sum of Eq. (11) the member for $l = k$ is always equal to one (any sequences is self-similar); members of the sum for unrelated sequences are equal to zero and members of the sum for homolog sequences may have values from 0 to 1. Consequently, we get $T(k) = 1$ for the case lacking homolog sequences with the index k; $T(k) = N$ when all N sequences in the multiplicity are identical, and $T(k)$ is equal to some number from 1 to N when there are some homologs in the multiplicity.



Scheme illustrating search for hidden periodicity in protein families

We then calculated periodicity matrix M^k (which is similar to the M matrix) for each amino acid sequence k of multiplicity of true alignments and then these matrices were summed with weights equal to $1/T(k)$. In the result, we calculated the weighted periodicity matrix M for the whole multiplicity of true alignments:

$$m(i, j) = \sum_k m^k (i, j) / T (k) , \qquad (12)$$

where k runs over all sequences of multiplicity of true alignments, $m^k(i,j)$ represents the element of the matrix M^k, and $m(i,j)$ is the element of the matrix M. The values $r_{i,j}$ were calculated as using the following equation:

$$r_{i,j} = \frac{m(i, j)}{\sum_i m(i, j)} . \qquad (13)$$

Afterwards we calculated $\pi_{i,j}$:

$$\pi_{i,j} = c_0 f_i + c_1 q_{i,j} , \qquad (14)$$

where:

$$q_{i,j} = \frac{\sum_k q^k{}_{i,j}}{\sum_i \sum_k q^k{}_{i,j}}, \quad (15)$$

where k runs over all sequences of multiplicity of false alignments. Frequencies $q^k_{i,j}$ are similar to frequencies $m^k_{i,j}$, but they are determined for k sequence of multiplicity of false alignments; f_i is probability of amino acid type i determined by all amino acid sequences of the Swiss-prot database.

Coefficients c_0 and c_1 (their sum should be equal to 1) were experimentally chosen for better selectivity of profile analysis (at constant sensitivity of this analysis). We chose the best values for c_0 and c_1 using amino acid sequences of the homeodomain and cytochrome P450 (No. 1 and No. 4 of Table 2). Coefficient c_0 varied from 0.99 to 0.01 and c_1 varied from 0.01 to 0.99, respectively. We determined c_0 and c_1 values creating multiplicity of false alignment of the smallest volume. Reduction of the volume of multiplicity of false alignments was insignificant at $c_0 > 0.95$, whereas at $c_0 < 0.5$ significant change in PWM and significant reduction of volume of multiplicity of true alignments occurred. Similar results for choosing c_0 and c_1 coefficients were also obtained for the family of serine/threonine protein kinases (No. 94, Table 2). Our calculations demonstrated that optimal values for c_0 and c_1 coefficients are within 0.8 and 0.2, respectively. These coefficient values were used in this study.

After calculation of a new PWM $\overline{U}_{i,j}$ we started the first stage of calculation, and the calculation cycle was repeated again. Such iteration procedure was usually repeated from 3 to 10 times. The goals of these iterations were: 1) increase of number of alignments for proteins from the protein family typical for an initiating amino acid sequence in the multiplicity of true alignments ($Z_1 > 6.0$); 2) reduction to minimal quantity of alignments versus $\overline{U}_{i,j}$ with $Z_1 > 6.0$ in multiplicity of false alignments. Iteration stopped when a volume of multiplicity of true alignments did not exhibit any trend to increase.

RESULTS AND DISCUSSION

Using iteration analysis, we revealed hidden periodicity in 94 protein families. With a few exceptions from 70 to 100% of proteins of each family have hidden periodicity. For most of these proteins the number of false alignments was either nil or did not exceed 10% of the total number of true alignments. However, in some families the number of false alignments reached 20%.

In 15% of protein families, such iteration procedure did not provide the requested specificity for recognition of hidden periodicity of the whole protein family. This means that for the 15% of protein families, hidden periodicity has been found in less than 70% of proteins belonging to these particular families (i.e., the number of true alignments was not representative). We believe that this phenomenon may be attributed to strong evolutionary divergence of periodicity in a protein family and also to the presence of large sized inserts or deletions of amino acids of amino acid sequences of the protein family.

Only in several of these 15% of families we have detected many amino acid sequences with $Z > 6.0$ among multiplicity of false alignments; this has been detected after several iterations. It is possible that such phenomenon is related to the presence of identical structural motifs in various proteins, for example, α-helices and/or certain α/β structures. This phenomenon may suggest the existence of convergent processes in amino acid sequences of protein families.

Table 2 lists protein families, length of the periodicity found, number of the same name proteins in the Swiss-prot database, and also number of the same name proteins where hidden periodicity has been found. Table 3 shows examples of amino acid alignments for amino acid sequences of cytochrome P450, homeodomain, MADs domain, T domain, and halcon synthetase. Alignment examples for each protein family are shown at the Website: http://bioinf.narod.ru/new1. Table 2 shows that the number of deletions and inserts may be relatively small (alignment for homeodomain) and quite large (MADs domain). These examples demonstrate that hidden periodicity of amino acid sequences listed in Table 2 is impossible to detect by the ID method in its present state [21] due to the presence of amino acid inserts and deletions versus the period observed.

For control experiment we generated 100 random matrices with $Z(n) > 6.0$ with period length varied from 10 to 50 amino acids. These matrices were considered as the initiating matrices, and we treated them using the iteration analysis procedure described in the corresponding section of the "Methods and Algorithms" section. In the result we did not reveal any protein family of the Swiss-prot database in which at least one of 100 randomly generated matrices was found by cyclic alignment even in a small proportion of sequences (>10%) of this particular family. For each periodicity listed in Table 2, we were looking for false alignments in 94 random massifs with number of sequences, amino acid frequencies, and distribution of sequence lengths (122,564 sequences) as in the Swiss-prot database. At such comparison, we found from 0.6 to 1.2 cases of false cyclic alignments of PWM (for $Z_1 > 6.0$) per volume of the Swiss-prot database.

This result shows that the detection of amino acid sequences of protein families characterized by hidden periodicity is rather specific. This result also shows three possible reasons underlying the origin of false alignments in the Swiss-prot database.

First, some proteins may have earlier unknown biological functions, which have not been annotated in the Swiss-prot database, and these functions are identical to biological functions of the initiating sequence. Under our

Table 2. List of protein families where hidden periodicity was found by means of iteration profile analysis

No.	Protein family name	Length of period	Number of proteins of a family present in Swiss-prot database (release 41)	Number of proteins of a family, which were included into multiplicity of true alignments	Number of proteins of a family, which were included into multiplicity of false alignments
1	2	3	4	5	6
1	Homeodomain containing proteins (homeodomain)	14	725	572	42
2	Proteins containing MADS domain (MADS domain)	13	73	70	1
3	T domain containing proteins (T domain)	14	65	59	9
4	Cytochromes P450 (active site)	14	665	577	31
5	Pyruvate kinases (ADP binding site)	11	67	59	6
6	Cpn10 proteins	14	133	121	8
7	Lysozyme C (active site)	6	64	60	8
8	Hexon proteins	34	21	20	3
9	Glycosyl transferases	33	57	47	4
10	cGMP-dependent protein kinases (nucleotide phosphate binding site of cGMP1, cGMP2e)	61	15	12	2
11	120 kD antigens	45	20	17	0
12	Vitamin B12 receptors	6	2	2	0
13	Carboxyl esterases type B/lipase family (active site)	18	80	73	9
14	SNAP-23, SNAP-25 family (T domain)	7	6	6	0
15	AB family of cyclins	4	74	67	11
16	Heat shock protein-90 family (A domain, substrate binding site, B-domain)	14	94	84	6
17	Immunoglobulin constant region domain	11	8	7	2
18	α-Subunit of methyl coenzyme M reductase	14	10	10	1
19	Phosphoenolpyruvate carboxylases (active site)	50	110	108	9
20	EF hand calcium binding sites	36	545	431	26
21	Catalases (active site)	10	118	101	4
22	Major histocompatibility complex class II antigens, β-chain	20	48	45	0
23	DNA binding motif (leucine zipper domain)	7	180	129	16
24	Interleukin 12, α-chain (IL-12A)	14	13	13	0
25	Halcon synthetases	17	119	118	5
26	Chitin synthetases (possibly transmembrane region, for many proteins domain was not clearly determined)	33	56	56	2

Table 2. (Contd.)

1	2	3	4	5	6
27	CF0 subunits (membrane channel for proton) of ATP synthase, chain A	5	155	120	11
28	CF0 subunits (membrane channel for proton) of ATP synthase, chain B	11	74	58	9
29	CF1 subunit (catalytic core) of ATP synthase, chain B	7	135	134	0
30	CF1 (catalytic core) of ATP synthase, chain A	9	92	89	4
31	Chemotaxis proteins (active site, regulatory domain, transmembrane domain)	7	291	198	24
32	S3 family of ribosomal proteins (KH domain type 2)	4	191	162	17
33	Hemoglobins β, Δ, ε, γ, and ρ chain	5	243	238	7
34	Eukaryotic initiation factor 4A	3	28	26	3
35	Acetyl-CoA carboxylase (coenzyme A binding site, biotin binding site, active site)	21	8	8	1
36	54 kD proteins recognizing signal sequence of secretory proteins (SRP54) (G-domain, M-domain)	9	55	52	7
37	Arginases (manganese binding site)	12	24	21	2
38	Phorbol ester and diacyl glycerol binding sites	36	108	88	0
39	β-Galactosidases (nucleophilic site)	14	41	31	5
40	Serpins	10	151	134	30
41	Adenine phosphoribosyl transferases	13	68	66	7
42	AMP deaminases (active site)	9	8	8	0
43	MCM protein family	13	43	39	2
44	Endorphins (NPP binding site)	41	40	39	0
45	Bacterial surface proteins (S-layer, SLH domain)	8	38	34	5
46	Citrate synthases	6	71	50	13
47	Exoribonucleases II	15	11	11	0
48	3-Phosphoshikimate-1-carboxyvinyl transferases (EPSP-domain)	12	81	77	4
49	Envelope proteins type A	5	5	4	0
50	Protein family of substrates for CRK (CRK-associated substrate)	42	5	5	0
51	Translation factors COE1, COE2, COE3, COE4 (domain for ser/thr/pro-rich)	22	12	12	0
52	Bcl-2 apoptosis regulators (domains BH4, BH3, BH2, BH1, transmembrane site)	20	9	8	1

Table 2. (Contd.)

1	2	3	4	5	6
53	Aspartate aminotransferases (pyridoxal phosphate binding site)	16	54	52	6
54	1-Aminocyclopropane-1-carboxylate oxidases (ACC oxidase)	32	25	24	0
55	Acetate kinases	11	49	48	6
56	Acetyl-coenzyme A synthetases	13	23	22	2
57	Antigen 85B precursors (active site)	4	7	7	2
58	Fructose *bis*-phosphate aldolase (C1-phosphate group substrate binding site)	3	35	34	3
59	Muscarinic acetylcholine receptors M1, DM1, M2, M3, M4, M5 (extracellular and transmembrane domains)	15	29	28	0
60	Cellulose synthases (catalytic subunit, active site)	32	13	13	0
61	Heat shock proteins	13	68	66	8
62	M10A peptidase family (active site, collagen binding site)	48	74	67	4
63	Acetylcholine receptors (transmembrane and extracellular domain)	7	132	89	7
64	Cytochrome *c* oxidases (B copper ion binding site)	7	148	132	11
65	Acetolactate synthase (active site)	25	51	37	4
66	Serine carboxypeptidases (active site, substrate binding site)	4	14	12	3
67	Proopiomelanocortin protein family (POMC)	33	22	22	0
68	Arginine kinases (active site)	18	18	18	2
69	Ribosomal acidic proteins P0 (L10E)	26	49	47	5
70	Arrestins	3	34	32	6
71	Cytochrome *c* oxidases, polypeptide II (intramembrane mitochondrial domain, A copper ion binding site)	5	220	216	14
72	4-Coumarate-CoA ligases	30	18	16	0
73	Triose phosphate isomerases (TIM) (active site)	6	129	118	7
74	Histidine and phenylalanine ammonia lyases (active site of 4-methylene-imidazole-5-histidine ammonia lyases)	28	80	80	0
75	Asparagine synthetases	7	25	23	2
76	Argininosuccinate synthetases (NP-ATP binding site)	40	64	62	1

Table 2. (Contd.)

1	2	3	4	5	6
77	Serine proteases (active site)	25	860	550	33
78	β-Glucuronidase precursors (proton donor domain)	18	7	7	0
79	β-Fructofuranosidase (active site)	29	32	30	2
80	α-Glycosidase (active site)	17	32	28	2
81	Cytochrome *c* oxidase, polypeptide III	11	137	130	10
82	RNA polymerase σ factors proD (the polymerase binding site)	13	47	46	5
83	Bacteriorhodopsins and archeorhodopsins (extracellular domain, transmembrane helix A-F)	12	21	21	3
84	Major histocompatibility complex type II (MHC II), precursors of α and β chains (extracellular domain)	41	75	72	5
85	Histone deacetylase family, subfamily 1	8	12	12	1
86	Coproporphyrinogen III oxidase	37	28	28	4
87	cAMP receptor protein (CAP) family (polypro domain)	7	10	10	1
88	Ligand-gated ion channel protein family	7	222	175	11
89	Crystalline entomocidal protoxin	9	79	71	4
90	CPCE and CPCF proteins	32	24	24	3
91	CAS protein family (kinase substrate, SH3 binding site)	42	7	7	0
92	IGF-binding proteins 1,2,3,4 (thyroglobulin type I domain)	21	31	27	1
93	DAHP synthetases	10	28	28	2
94	Serine/threonine protein kinases (active site)	18	1116	903	37

iteration analysis, such proteins will be referred to the category of false alignments.

Second, proteins sharing common evolutionary origin but exhibiting various biological functions now may have related periodicity. This may result in interpretation of evolutionary related proteins as false alignments during our iteration analysis.

The third reason for appearance of false alignments during the Swiss-prot searching can be attributed to the presence of many tandem copies of a tandem repeat, which is characterized by a large positive value obtained on the treatment with position specific matrix $\overline{U}_{i,j}$ using the modified method of profile analysis [25, 26].

Such cases of tandem homologous repeats have low statistical probability in a random text and therefore they are absent in our randomly generated amino acid sequences. Total weight of an amino acid sequence containing many such repeats may be quite high, and this provides Z_1 value exceeding 6.0. However, we observed this phenomenon only in the case of perfect repeats lacking amino acid deletions or inserts. Such homologous perfect periodicity was easily recognized in our results and excluded from subsequent consideration by constructing the ID spectrum for the initial amino acid sequence which gave $Z_1 > 6.0$ after alignment with the corresponding hidden periodicity matrix. After using this

Table 3. Examples of cyclic alignments for cytochrome P450, homeodomain, MADs domain, and halcon synthetase

Code accession number	Description of the analyzed sequences in the literature	Z	Coordinates of the region with hidden periodicity	Alignment
P10633	Cytochrome P450 2D1	8.3	223–391	23456789abcde123456789abcde123456789abcde1-----2345--6789abc PEVLNTFPA-LLRIPGLADKVFQGQKTFMALLDNLLAENRTTWDPAQPPRNLTDAFLAEV de---123456789abcde123456789abcde123456789abcde123456789abcd EKAKGNPE--SSFNDENLR-MVVVDLFTAGMVTTATTLTWALLLMILYPDVQRRVQQEID e123456789abcde-123456789abcde123456-789abcde123456789 EVIG-QVRCPEMTDQAHMPYTNAVIHEVQRFGDIAPLNLPRFTSCDIEVQDFVI
P42585	Homeodomain EGHBX3	8.1	78–142	789abcde123456789abcde123456789abcde123456789abcde123456789a QSQSKRRVL-FNKFQISQLEKRLKQ-RYLTAQERQELAHTIGLTPTQVKIWFQNH---AY bcde123456 KMKRLFHDDH
O64645	MADs domain AGL20	11.0	8–108	789abcd123456789abc-d123456789abcd123456789abcd1-23456789abc KRIENATSRQVTFSKRRNGLLKKAFELSVLCDA-----EVSLIIFSPKGKLYEFA----S d123456------789abcd123456789abcd123-456789abcd1234--56789 SNMQDTIDRYLRHTKDRVSTK---PVS--EENMQHLKYEAA---NMMKKIEQLEASKR
P24824	Halcon synthetase WHP1	18.0	220–350	789abcdefgh1--234--5-6789abcdefgh123456789abcdefgh12-3456789 GAAAGRGGADPDGRVERPLFQLVSAAQ--TILPDSE---GAID---GHLREVGLAFHLLK abcdefgh123456789a bcdefgh123456789abcdefgh12345678-9abcdefg DVPGLISKNIER ALEDAFEPLGISDWNSIFWVAHPGGPAILDQVEAKVGLDKAR-MRAT h123456789--abcdefgh1 RHVLSEYGNMSSACVLFILDE

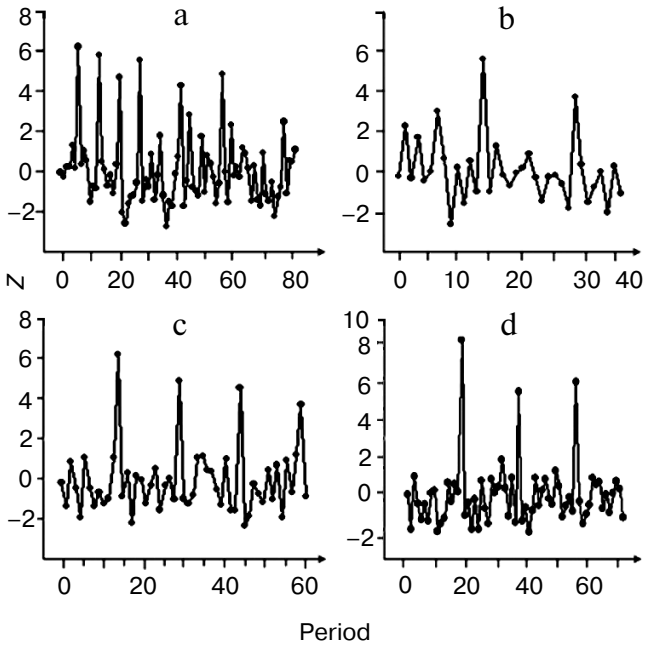
approach, the mean of false alignments for all protein families listed in Table 2 varied from 0 to 12. This value better corresponds to calculations made by the Monte-Carlo method shown above and the presence of these false alignments can be explained by the first two reasons. We also analyzed all true alignments for the presence of homologous tandem repeats, and we did not find them in the families listed in Table 2. This means that high value of Z_1 for true alignments is not related to the presence of highly homologous tandem repeats in the amino acid sequences.

We believe that taking into consideration the total number of amino acid sequences in the Swiss-prot database (exceeding 120,000 sequences) the selectivity and specificity of our approach are reasonably good. We recognize from 70 to 100% of proteins of one family and in most cases the proportion of false alignments varies from 0 to 10% of true alignments, and these values may be further reduced by filtration of perfect tandem repeats from false alignments. From this viewpoint, our results may be useful for prediction of biological functions of unknown amino acid sequences.

We have also used the ID method for searching for hidden periodicity in amino acid sequences aligned versus a PWM. Examples given in the figure show that hidden periodicity is definitely recognized by the ID method at high statistical significance. This result shows that in spite of the multiple iteration process and noise decomposition the hidden periodicity is found in resultant amino acid sequences at statistically significant level but only with certain numbers of inserts and deletions.

For comparative analysis, we tried to find the hidden periodicity listed in Table 2 using the RADAR [4] and REPRO [35] programs, but these programs failed to recognize any case of periodicity shown in Table 2.

It is possible that the number of protein families with hidden periodicity significantly exceeds 94, but this study has not been designed for elucidation of hidden periodicity in all known protein families. The Swiss-prot database contains information about 6000 protein families [36]. These protein families cover 75% of the amino acid sequences annotated in the Swiss-prot database, and these sequences may have several thousand spatial conformations [37]. However, we believe that the bulk of information on the presence of hidden periodicity in diverse protein families (exhibiting various functional and biological importance) together with previously obtained results [24, 27] clearly demonstrates that we are dealing with a common biological phenomenon. Subsequent studies will clarify whether this is a common phenome-



Information decomposition of some amino acid sequences belonging to protein families listed in Table 2. The ordinate axis shows statistical significance for each period showing an argument of normal distribution, the abscissa axis shows length of the period expressed as number of amino acid residues. Statistically significant periodicity has $Z(n) > 6.0$. The plot shows the main period and periods divisible into the main period. a) Cytochrome P450 (accession number P10633), period length of 7 amino acid residues. b) Homeodomain (accession number P42585), period length of 14 amino acid residues. c) MADs domain (accession number O64645), period length of 13 amino acid residues. d) Halcon synthetase (accession number P24824), period length of 17 amino acid residues

non for various protein families; if so, they will help to recognize hidden periodicity in various protein families.

Interestingly, the period lengths are quite variable for various families (see Table 2). However, the presence of the same period length in various protein families does not necessarily mean similar periodicity. The form of hidden periodicity is determined by the matrix $M(i,j)$, which can have different form even for the same period length.

Thus, in the present study we have demonstrated that many functionally important protein families are characterized by the existence of hidden periodicity typical for all (or at the least majority of) proteins of this particular family. Our results confirm an earlier proposed hypothesis [1-3] that evolution of genes may involve multiple tandem duplications. Earlier hidden periodicity was also demonstrated at the nucleotide level [24, 32, 33]. We suggest that a protein with certain biological function is characterized by a corresponding periodicity. It is possible that certain amino acid periodicity may facilitate formation of a certain protein globule and thus provide interaction of adjacent amino acid residues in it. If this suggestion is correct the evolutionary process might shuffle a relatively small number of amino acid periods, which are then repeated in a tandem manner for creation of a protein of certain functional meaning. Subsequent evolutionary process could just employ evolutionary selection of preexisting genes for augmentation of a certain functions of their protein products by means of certain amino acid substitutions (inserts or deletions). Finally, we can get modern proteins where hidden periodicity can be hardly recognized.

The results obtained in the present study suggest the existence of convergent origin of the hidden periodicity found in each protein family. We can also assume that the hidden amino acid periodicity was formed as a consequence of realization of the same biological function (and consequently the same spatial conformation) attributed to amino acid sequences of the same family. Although we do not personally believe in such a possibility, the results of the present study cannot rule out it. The results of the present study demonstrate the existence of some correspondence between certain classes of amino acid periods and functional significances of proteins where these periods are observed as the hidden periodicity. If such correspondence actually exists in all protein families, the presence of hidden periodicity in various protein families may be found during subsequent accumulation of data.

REFERENCES

1. Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag, Berlin.
2. Ohno, S., and Epplen, J. T. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3391-3395.
3. Ohno, S. (1984) *J. Mol. Evol.*, **20**, 313-321.
4. Heger, A., and Holm, L. (2000) *Proteins*, **41**, 224-237.
5. Landau, G. M., Schmidt, J. P., and Sokol, D. (2001) *J. Comp. Biol.*, **8**, 1-18.
6. Neuwald, A. F., and Poleksic, A. (2000) *Nucleic Acids Res.*, **28**, 3570-3580.
7. Coward, E., and Drablos, F. (1998) *Bioinformatics*, **14**, 498-507.
8. Katti, M. V., Sami-Subbu, R., Ranjekar, P. K., and Gupta, V. S. (2000) *Protein Sci.*, **9**, 1203-1209.
9. Andrade, M. A., Perez-Iratxeta, C., and Ponting, C. P. (2001) *J. Struct. Biol.*, **134**, 117-131.
10. Benson, G. (1999) *Nucleic Acids Res.*, **27**, 573-580.
11. Heringa, J. (1998) *Curr. Opin. Struct. Biol.*, **8**, 338-345.
12. Heringa, J., and Argos, P. (1993) *Proteins*, **17**, 391-341.
13. Murray, K. B., Taylor, W. R., and Thornton, J. M. (2004) *Proteins*, **57**, 365-380.
14. Rackovsky, S. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 8580-8584.
15. McLachlan, A. D. (1993) *J. Phys. Chem.*, **97**, 3000-3006.
16. Jackson, J. H., George, R., and Herring, P. A. (2000) *Biochem. Biophys. Res. Commun.*, **268**, 289-292.
17. Makeev, V. Ju., and Tumanyan, V. G. (1996) *Comput. Appl. Biosci.*, **12**, 49-54.
18. Marcotte, E. M., Pellegrini, M., Yeates, T. O., and Eisenberg, D. (1999) *J. Mol. Biol.*, **293**, 151-160.

19. Andrade, M. A., Ponting, C. P., Gibson, T. J., and Bork, P. (2000) *J. Mol. Biol.*, **298**, 521-537.
20. Landschulz, W. H., Johnson, P. F., and McKnight, S. L. (1988) *Science*, **240**, 1759-1764.
21. Korotkov, E. V., Korotkova, M. A., and Kudryashov, N. A. (2003) *Phys. Lett. A*, **312**, 198-210.
22. Korotkov, E. V., Korotkova, M. A., and Kudryashov, N. A. (2003) *Mol. Biol. (Moscow)*, **37**, 436-451.
23. Korotkov, E. V., and Korotkova, M. A. (1996) *DNA Res.*, **3**, 157-167.
24. Laskin, A. A., Korotkov, E. V., Chaley, M. B., and Kudryashov, N. A. (2003) *Mol. Biol. (Moscow)*, **37**, 561-570.
25. Chaley, M. B., Korotkov, E. V., and Kudryashov, N. A. (2003) *DNA Sequence*, **14**, 37-52.
26. Korotkov, E. V., and Korotkova, M. A. (2000) *Mol. Biol. (Moscow)*, **34**, 348-353.
27. Laskin, A. A., Korotkov, E. V., and Kudryashov, N. A. (2004) in *Bioinformatics of Genome Regulation and Structure* (Kolchanov and Hofestaedt, R., eds.) Kluwer Press, New-York, pp. 135-144.
28. Holmes, I., and Durbin, R. (1998) *J. Comput. Biol.*, **5**, 493-504.
29. Henikoff, S., and Henikoff, J. G. (1993) *Proteins*, **17**, 49-61.
30. Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C., and Marcourt, L. (2000) *J. Theor. Biol.*, **206**, 323-326.
31. Korotkova, M. A., Korotkov, E. V., and Rudenko, V. M. (1999) *J. Mol. Model.*, **5**, 103-115.
32. Chaley, M. B., Korotkov, E. V., and Skryabin, K. G. (1999) *DNA Res.*, **6**, 153-163.
33. Korotkov, E. V., Korotkova, M. A., and Tulko, J. S. (1997) *CABIOS*, **13**, 37-44.
34. Kullback, S. (1959) *Information Theory and Statistics*, John Wiley & Sons, New-York.
35. George, R. A., and Heringa, J. (2000) *Trends Biochem. Sci.*, **25**, 515-517.
36. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) *Nucleic Acids Res.*, **32**, Database issue D138-D141.
37. Wolf, Y. I., Grishin, N. V., and Koonin, E. V. (2000) *J. Mol. Biol.*, **299**, 897-905.