Molecular Biology, Vol. 33, No. 4, 1999, pp. 611–617. Translated from Molekulyarnaya Biologiya, Vol. 33, No. 4, 1999, pp. 688–695.

MISCELLANEOUS

UDC 577.212.2

Latent Periodicity Regions in Amino Acid Sequences

E. V. Korotkov[1], M. A. Korotkova[2], V. M. Rudenko[1], and K. G. Skryabin[1]

[1] *Bioengineering Center, Russian Academy Sciences, Moscow, 117312 Russia;*
E-mail: korotkov@biengi.msk.su
[2] *Moscow Engineering Physics Institute, Moscow, 115409 Russia*

Received June 11, 1998

Abstract—We a propose mathematical approach to searching for latent periodicity in amino acid sequences. The statistical significance of the latent periodicity was evaluated by the Monte Carlo method. Search for latent periodicity in the SWISS-PROT bank sequences reveals over 10% proteins with latent periodicity including enzymes. We present amino acid sequences with latent periodicity revealed in lipoamide dehydrogenase (DLDH_AZOVI), asparagine synthase (ASN1_YEAST), endochitinase 2 (CHI2_COCIM), and Rho/Rac effector (CTRO_MOUSE). These proteins feature latent periodicity regions 19, 6, 11, and 7 amino acid long, respectively. The possible functional and evolutionary significance of the latent periodicity regions is discussed.

Key words: latent periodicity; reciprocal information; amino acid sequence

INTRODUCTION

At present the life science enters a new, postgenomic era, owing to sequencing of the full genomes of certain bacteria, yeast (*Saccharomyces cerevisiae*), and nematode (*Caenorhabditis elegans*) as well as ongoing genome sequencing of more complex organisms (plants, animals, and man). Availability of these huge volumes of genetic information will inevitably exert significant influence on our understanding of life, since many problems related to emergence and evolution of single traits now can be considered at the level of complete genomes. This confers especial significance on studying the organization of available nucleotide and amino acid sequences and developing mathematical algorithms for analysis of these sequences, aimed at fuller extraction of biological information from huge genetic texts.

Studies on amino acid sequence periodicity may bring light to structural properties of the protein sequences and their relationship to protein spatial structure and evolution. At present two mathematical approaches are widely used in such studies.

The first approach is algorithmic studies of a sequence character code revealing character code periodicity in the presence of deletions, inserts, and a relatively high number of substituted characters. However, the number of substitutions should not disturb the visible similarity between the periods [1–7]. The available algorithms for character code periodicity rely on known dynamic programming procedures for comparing two character code sequences [8]. Algorithmic approaches also include tests for complexity of the genetic texts revealing duplications in nucleotide sequences of various genes [9–11]. These studies are based on Kolmogorov's text complexity [12].

The second approach applies Fourier transform to analysis of character code sequences [13–20]. The main difficulty of using the Fourier transform mathematics is selecting the mode of character-to-numerical transformation of the sequence and storing all statistical properties of the character code sequence; in addition, sequence periodicity cannot be revealed in the presence of inserts and deletions.

The mathematical methods used reveal chiefly the "homologous periodicity" of the amino acids sequences, implying high similarity between the periods in the studied sequence. The kind of periodicity detectable by Fourier transform is defined by the mode of character-to-numerical transformation or by the similarity matrix (such as PAM250) generated on the basis of evolutionary comparison of closely related proteins in the case of algorithmic approaches.

Let us consider the situation of a character code sequence with periodicity lacking statistically significant similarity between the periods (hereafter referred to as latent periodicity). In this case such periodicity can only be revealed by analysis of all periods at the same time. Let us consider a three-character amino acid periodicity {(Lys\Asn\Ile\Met\Thr)(Arg\Ser\Pro\Hys\Glu)(Gln\Val\Ala\Asp\Gly)}. The amino acid residues allowed in the first, second, and third positions of the period are enclosed in parentheses. The ambiguity of each position is clearly different; different amino acid sets are allowed in the first, second, etc. positions. Sequences with this kind of periodicity may acquire the form: [LysArgGln][AsnSerVal][IleProAlaMet]

0026-8933/99/3304-0611$22.00 © 1999 МАИК "Наука/Interperiodica"

[HysAspThr][GluGlyLys][HysAspGly]... (the periods are enclosed in brackets). Similarity between the periods of this sequence is clearly statistically insignificant or absent; therefore, such periodicity can be missed during the search based on Fourier transform or dynamic programming. There is a lot of amino acid sequences with this kind of periodicity. Statistical significance of periodicity requires sufficient length of the corresponding sequence. The mathematical approach revealing latent periodicity in amino acid sequences is described below.

One can give different explanations of latent periodicity. It can correlate with periodicity in the protein spatial structure. Ancient cases of multiple duplications in the amino acid sequences that accumulated lot of substitutions can also give rise to latent periodicity.

Successful search for latent periodicity in amino acids sequences requires a novel mathematical approach, since application of dynamic programming and Fourier transform cannot reveal latent periodicity in many cases. To solve this problem, we modified the method of searching for latent periodicity in nucleotide sequences [21–23] based on revealing extended similarity between an artificial periodic sequence and a character code sequence [24]. Here we used the Monte Carlo method to evaluate the statistical significance of the observed periodicity. Analysis of the SWISS-PROT protein bank reveals at least 10% proteins with latent periodicity.

EXPERIMENTAL

The search for latent periodicity regions in amino acid sequences was realized through comparison of an artificial periodic and amino acid sequences. The alphabet of the artificial sequences has $S(i)$ characters. In order to reveal a 2-aa periodicity, we generate artificial sequence $S(1)S(2)S(1)S(2)S(1)S(2)\ldots$ An artificial sequence $S(1)S(2)S(3)S(1)S(2)S(3)S(1)S(2)S(3)$ is generated to reveal a 3-aa period, while sequence $S(1)S(2)\ldots S(n)S(1)S(2)\ldots S(n)S(1)S(2)\ldots S(n)$ is generated to reveal a period of n amino acids. The length of the artificial sequence equals the length of the analyzed one. We shall study amino acid sequence periodicity with a period ranging from 2 to $L/2$, where L is the length of the analyzed sequence. After comparison of the artificial and query sequences, we have a matrix $M(20,n)$, where n is period length. Element $M(i, j)$ of this matrix stores the number of i-type amino acid (i = 1,20) that match the character $S(j)$ in the artificial periodic sequence. We use reciprocal information calculated from the matrix $M(20, 2)$ as a measure of similarity between the artificial sequence with period n and an amino acid sequence. We propose that the characters in the artificial and amino acid sequences are independent; hence, according to [25], the doubled reciprocal information will be distributed as X^2 with $19(n - 1)$ degrees of freedom. The more there is reciprocal information, the less is the probability of random relationship between the characters in the artificial periodic and amino acid sequences [25]. If periodicity is not observed along the whole amino acid sequence, reciprocal information between the artificial periodic sequence and any fragment of the initial sequence should be defined. This procedure, described in detail for nucleotide sequences, was applied to amino acid ones without modification [23]. However, in the case of a short amino acid sequence fragment the doubled reciprocal information will deviate from the X^2 distribution with $19(n - 1)$ degrees of freedom. We accept that for short amino acid sequence the mean value of each element in the $M(20, n)$ matrix should be less than 5. Hence, for an amino acid sequence shorter than $100n$ the X^2 distribution cannot be used to evaluate the statistical significance of match between the artificial periodic and the amino acid sequences. In this case the Monte Carlo method can be used and random matrices M' should be generated with the same $X(i)$, $i = 1, 2, \ldots, 20$ and $Y(j)$, $j = 1, 2, \ldots, n$ (22) as in the initial matrix resulting from comparison between the artificial periodic and the amino acid sequences, where:

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

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

$X(i)$ shows the number of different amino acid residues in the analyzed sequence. $Y(j)$ equals the number of $S(j)$ characters in the artificial periodic sequence. Let us assume that the reciprocal information $I(1)$ and matrix M(20, n) resulted from comparison of the artificial periodic and the amino acid sequences, where 20 is in number of amino acids and n is the period length. In order to evaluate the statistical significance of the observed relationship between the sequences using the Monte Carlo method, random matrices $M'(20, n)$ should be generated with the same rows and columns totals as the $M(20, n)$. The matrices were generated according to [26]. Let the number of M' matrices be N. The number $N1$ of M' matrices with reciprocal information $I' > I(1)$ can be determined, which makes it possible to evaluate the probability F of random relationship between the artificial periodic and the amino acid sequences. The probability F can be approximated as $N1/N$. If the amino acid sequence has pronounced periodicity, the $I(1)$ has a high value and it is hard or even impossible to generate a significant number of M' matrices in order to have $N1 > 0$ and evaluate the probability F. In this case it is convenient to introduce $Z = (I(1) - I'(\mathrm{av}))/\sqrt{D}$ as a measure of statistical significance of similarity between the artificial peri-

Sample regions with latent periodicity from SWISS-PROT amino acid sequences

SWISS-PROT sequence identifier	Protein	Period length (number of residues)	References
ASN1_YEAST	Glutamine-hydrolyzing asparagine synthase	6	[40, 41]
CHI2_COCIM	Endochitinase	11	[42]
CTRO_MOUSE	Rho/Rac-effector	7	[34]
DLDH_AZOVI	Lipoamide dehydrogenase	19	[35]

odic and the amino acid sequences; $I(m)$ is the mean I' for the set of random matrices M' and D is dispersion of I' for the set of matrices M'. If Z equals 0, there is no similarity between the artificial periodic and the amino acid sequences. If $Z > 3.0$, the distribution of Z fits well into normal distribution.

Generation of matrices according to [26] is equivalent to generation of Bernoulli sequences with the same amino acid composition as the analyzed sequence. In this case, the relationship between amino acids need not be accounted for, since this would distort the relationship between Z and the period length. For instance, accounting for paired correlation between the amino acids at generation of M' matrices usually results in low $Z(2)$ values so that 2-aa periods cannot be revealed.

After the calculations we have the relationship $Z(n)$, where n ranges from 2 to $L/2$. This also allows us to reveal duplications in the analyzed sequence. Such spectrum shows periodicity of various length in the protein. The most significant periodicity has the top Z value. At the same time, all multiple periods will also have high Z values (however, not as high as for the basic period) as was already noted for nucleotide sequences [23].

Latent periodicity may be present in only a part of the analyzed amino acid sequence. In this case all subsequences of the initial amino acid sequence should be tested and the $Z(n)$ spectrum should be determined for each of them [23]. These calculations decelerate the search for latent periodicity. For instance, analysis of a 500-aa sequence takes less than 3 min on a Pentium-200 IBM PC.

We studied the $Z(n)$ behavior for randomly generated amino acid sequences 10^8 characters long. In this case $Z(n)$ never exceeded 5.6. This value can be considered as the threshold for latent periodicity search in amino acid sequences. Let us accept periodicity in an amino acid sequence if $Z(n) > 5.6$.

RESULTS AND DISCUSSION

The above algorithm was realized in a C^{++} program complex and was applied to periodicity search in the SWISS-PROT amino acid sequences. The result of this investigation demonstrated that over 10% of the amino acid sequences analyzed have regions with latent periodicity. It is probable that such sequences include protein regions with lowered Kolmogorov complexity [9–11]. The proteins with latent periodicity regions include enzymes, DNA-binding proteins, oncoproteins, etc. In addition to proteins with latent periodicity, we have revealed proteins with homologous periodicity (sometimes imperfect) as well as numerous duplications, which agrees with the published data [3, 6, 7, 27, 28]. The revealed regions feature high Z values. We assume that $Z > 18.0$ indicates homologous or almost homologous periodicity (similarity level between single periods will exceed 90%).

We can exemplify the functioning of this algorithm by revealing a perfect homologous periodicity (Fig. 1). The first SWISS-PROT sequence with homologous periodicity is 110K_PLAKN. This sequence has 12 repeats ETQNTVEPEQTE [29]. Figure 1 demonstrates maximum $Z = 44$ for a period of 12 aa. Probability of random appearance (P) of such a sequence can be evaluated from formula $P = (e^{Z1})/(Z\sqrt{2\pi})$, where $Z1 = Z^2$, and equals $P < 10^{-100}$ [30]. For periods of 24, 36, and 48 aa (multiples of 12), Z equals 24.5, 26.6, and 17.1, respectively. Interference of the 12-aa period also increases the Z for the periods multiple of 2, 3, and 4 aa [23]. However, Z values for these periods are much lower than that for the 12-aa period. Another example of this algorithm at work is revealing latent periodicity of the SWISS-PROT sequence ABA1_ASCSU fragment [31–33]. This fragment has three tandem repeats

AKILHYYDELEGDAKKEATEHLKGGCREILKHV
VGEKAAELKNLKDSGASKEELKAKVEEALHAV
TDEEKKQYIADFGPACKKIYGVHTSRRRRHHFT
LESSLDTHLKWLSQEQKDELLKMKKDGKAKKELE

133 aa long, which gives us $Z(133) = 50.6$. This period increases the Z for the fractional periods of 7, 19, and 38 aa ($Z(76) = 11.3$, $Z(19) = 18.5$, and $Z(38) = 8.0$). The above examples demonstrate extremely high Z values in the case of homologous periodicity.

Figure 1 presents examples of the regions with latent periodicity. The corresponding amino acid sequences and matrices are given in Figs. 2 and 3, respectively. These matrices demonstrate that most positions in the periods are degenerate; however, certain positions have a nearly homogeneous amino acid composition. For instance, this is true for positions 3

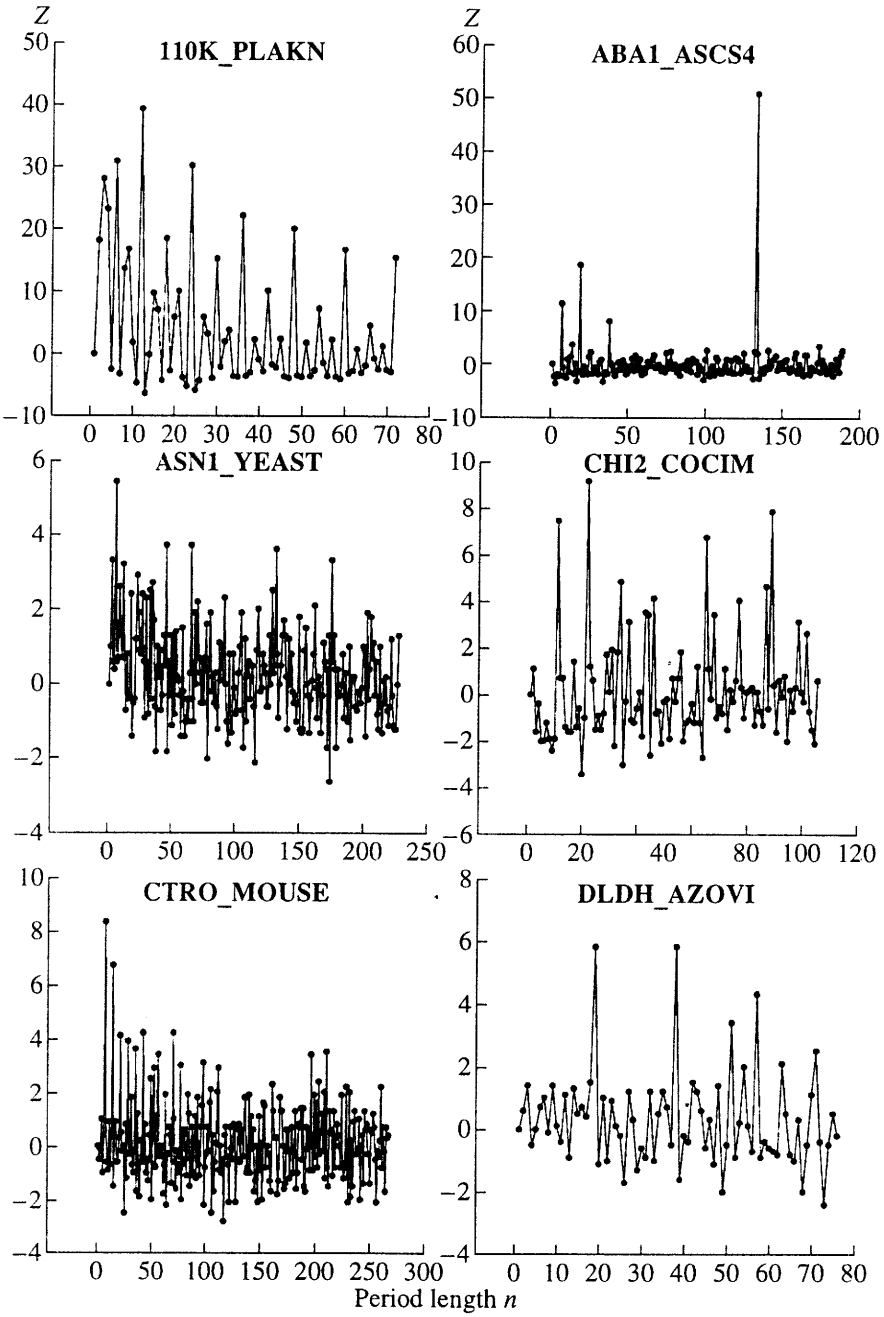

Fig. 1. *Z* values as a function of period length *n* for amino acid regions with perfect latent periodicity from sequences 110K_PLAKN and ABA1_ASCSU, and regions with latent periodicity from sequences ASN1_YEAST, CHI2_COCIM, CTRO-MOUSE, and DLDH _AZOVI. The corresponding amino acid sequences are presented in Fig. 2.

and 5 in the 19-aa periodicity of the DLDH_AZOVI sequence.

It should be interesting to consider the functional significance of the amino acid sequences with latent periodicity. A latent periodicity region of the ASN1_YEAST includes the most part of the glutamine-hydrolyzing asparagine synthase. One can suppose that the latent periodicity is essential for the spatial structure of this protein and its interaction with the ligands.

A latent periodicity region of 7 aa from the Rho/Rac effector (CTRO_MOUSE) includes a leucine zipper region ($L–X_6–L–X_6–L–X_6–L$) with 7-aa periodicity [34]. One can suppose that this region of the Rho/Rac effector resulted from multiple duplica-

```
ASN1_YEAST
LDGMFAWTLYDAKQDRIVAARDPIGITTLYMGRSSASPKTVYFASELKCLTDDCDTITAFPPGHVYDSKTDK
ITRYFTPDWLDEKRIPSTPIDYMAIRHSLEKAVRKRLMAEVPYGVLLSGGLDSSLIASIAARETAKATNDVE
PSTYDSKARHLAGIDDDGKLHTAGWTSLHSFAIGLPNAPDLQAARKVAKFIGSIHHEHTFTLQEGLDALDDV
IYHLETYDVTTIRASTPMFLLSRKIKAQGVKMVLSGEGSDEIFGGYLYFAQAPSAAEFHTESVQRVKNLHLA
DCLRANKSTMAWGLEARVPFLDREFLQLCMNIDPNEKMIKPKEGRIEKYILRKAFDTTGEPDAKPYLPEEIL
WRQKEQFSDGVGYSWIDGLKDTAEAVISDEMFASPKAEWGSDIPTTKEAFWYRLKFDALFPQKTVADTVMRW
IPKADWGCAEDPSGRYAQIHEKHIE

CHI2_COCIM
TLSASTTPSSPSTVSPSSTMQTTSTGSTSIETVTTRSQEPPSTTISTRSASTEPVTTRSQEPPSTTISTRSA
STETVTTRSQEPPSTTISTWSASTETSTSSQDSPSTTISTKSAPTGTVTTRSQDLPSTTISTRSPETETETA
TTKSQGSPSITLSTRSSSAETVSTRSQHSSSTTISTKSAPTETGTTSEHSTSMPVSTRSASTETVITRS

CTRO_MOUSE
IQELQEKLEKAVKASTEATELLQNIRQAKERAERELEKLHNREDSSEGIKKKLVEAEERRHSLENKVKRLET
MERRENRLKDDIQTKSEQIQQMADKILELEEKHREAQVSAQHLEVHLKQKEQHYEEKIKVLDNQIKKDLADK
ESLENMMQRHEEEAHEKGKILSEQKAMINAMDSKIRSLEQRIVELSEANKLAANSSLFTQRNMKAQEEMISE
LRQQKFYLETQAGKLEAQNRKLEEQLEKISHQDHSDKSRLLELETRLREVSLEHEEQKLELKRQLTELQLSL
QERESQLTALQAARAALESQLRQAKTELEETTAEAEEEIQALTAHRDEIQRKFDALRNSCTVITDLEEQLNQ
LTEDNAELNNQNFYLSKQLDEASGANDEIVQLRSEVDHLRREITEREMQLTSQKQTMEALKTTCTMLEEQVL
DLEALNDELLEKERQWEAWRSVLGDEKSQFECRVRELQRMLDTEKQSRARADQRITESRQVVELAVKEHKAE
ILALQQALKEQKLKAESLSDKLNDLEKKHAM

DLDH_AZOVI
SGSKPVEIPPAPVDQDVIVDSTGALDFQNVPGKLGVIGAGVIGLELGSVWARLGAEVTVLEAMDKFLPAVDE
QVAKEAQKILTKQGLKILLGARVTGTEVKNKQVTVKFVDAEGEKSQAFDKLIVAVGRRPVTTDLLAADSGVT
LDERGFI
```

Fig. 2. Amino acid sequences with latent periodicity; SWISS-PROT sequence identifier is given prior to the sequence.

	ASN1_YEAST						CHI2_COCIM			CTRO_MOUSE							DLDH_AZOVI			
	1					5	1	5	10	1					5		1	5	10	15
Lys	11	2	4	5	3	7	11000000001			9	9	5	9	4	9	5	02010400011100001 01			
Asn	2	1	1	0	1	1	00000000000			2	2	0	4	6	3	3	00001001000000000000			
Ile	7	6	6	2	3	5	00001003410			1	0	5	4	0	1	10	01100101001000030000			
Met	2	3	0	3	1	2	00000110000			0	1	1	0	1	7	3	00010000000000000000			
Thr	1	6	6	12	2	4	20236599998			7	3	1	0	1	0	2	1	20001000200000001 30		
Arg	1	4	7	4	6	1	41000000015			8	6	2	6	6	8	4	00000000000121010 00			
Ser	2	6	4	6	6	5	88485631375			7	8	2	4	3	4	4	20000101000000000 11			
Leu	5	6	8	6	10	3	00100000020			3	42	11	7	3	42	0	21030111110130000 00			
Tyr	4	1	0	7	0	5	00000000000			0	1	1	0	0	1	0	00000000000000000000			
Phe	8	5	0	1	2	3	00000000000			2	1	1	0	1	0	0	00010110000100100000			
Cys	1	2	0	0	1	1	00000000000			1	0	0	0	1	0	1	00000000000000000000			
Trp	3	0	4	0	0	3	10000000000			0	0	1	0	0	1	0	00000000010000000000			
Pro	7	5	1	2	6	3	02155112000			0	0	0	0	0	0	0	00100011120000100 00			
Hys	1	2	3	4	3	0	00110000000			4	3	2	1	2	0	3	00000000000000000000			
Gln	0	2	4	2	0	3	23200010000			6	7	10	5	12	10	2	11000030000011000000			
Val	0	1	6	0	7	5	00000014201			6	2	2	1	0	1	5	10110002313010313 01			
Ala	8	11	6	4	7	10	11410100100			1	6	12	3	8	6	8	00410000121000310 11			
Asp	5	5	7	9	11	1	02000000000			5	5	3	3	4	3	1	00004000001202000 11			
Glu	4	1	6	5	4	9	01401530000			13	16	7	18	16	17	6	02002000000110020 20			
Gly	5	7	3	4	3	5	10011000100			1	2	0	2	0	0	0	01100021001104002 03			

Fig. 3. Matrices $M(20,n)$ for amino acid sequences with latent periodicity; n is length of the latent period; n equals 6, 11, 7, and 19 aa for sequences ASN1_YEAST, CHI2_COCIM, CTRO-MOUSE, and DLDH _AZOVI, respectively; rows and columns correspond to amino acid residues and the latent period positions.

tions of the 7-aa fragment, and that the periodicity of the leucine zipper region reflects its origin.

A 19-aa periodicity region of the DLDH_AZOVI sequence includes the NAD^+-binding site of lipoamide dehydrogenase [35]. The conformation of the NAD^+-binding site is highly conservative and is present in many proteins. The region of the NAD^+-binding site features four α-helices and six β-sheets; however, the number of α-helices and β-sheets can vary among proteins [36, 37]. The region of NAD^+ binding is 32 aa long and has the βαβ conformation [38, 39]. These amino acid residues were studied in detail in various proteins, and 11 conserved positions with definite amino acid composition were revealed [38]. However, the latent periodicity specific for this amino acid sequences has not been noticed. The period of 19 aa includes a single β-sheet and a single α-helix. Multiple duplication of this sequence could

give rise to the structure of the NAD^+-binding site. We observed a similar periodicity in many other proteins with NAD^+-binding sites.

The mathematical approach developed allows revealing much less pronounced periodicity in the amino acid sequences as compared with Fourier transform or algorithmic approaches, as well as the types of periodicity undetectable by these methods. However, the approach has a set of limitations. As with Fourier transform, it does not account for inserts/deletions in the analyzed sequence. Since the number of indels can be relatively high, a considerable amount of the amino acid sequences with the latent periodicity are missed by this method, and 10% is the lower bound proportion of proteins with latent periodicity. At present we are developing a modified approach revealing latent periodicity in the amino acid sequences with a relatively low number of indels. Development of a more general approach based on dynamic programming is a labor-consuming task, since the type of latent periodicity in the amino acid sequence is not known in advance, and exhaustion of all possible latent periodicity types is hardly possible even with supercomputers. Nevertheless, the proposed approach even in its present state can be a powerful tool of searching for evolutionary distant multiple duplications in various proteins.

There is another mathematical approach revealing "fuzzy" amino acid periodicity [16]. However, this method reveals latent periodicity only if all period positions have the same type of "fuzziness." Moreover, the type of such "fuzziness" should be known in advance for each period position, which requires a lot of computing and complicates realization of this approach.

The revealed latent periodicity in the amino acid sequences from EMBL {Sic!} cannot be reduced to more homologous periodicity of a limited set of local alphabets (resulting from unification of certain amino acids). First, such alphabets will be different for every region of the revealed latent periods and will depend on the period length. Second, such alphabets will significantly differ with proteins even provided the same period length. Apparently, this indicates the relationship between latent periodicity and the evolutionary origin of proteins through multiple duplications of various short sequences.

Thus, here we demonstrate latent periodicity in the primary structure of proteins. Earlier we have demonstrated this in the nucleotides sequences [23]. It is quite probable that a sizeable portion of the coding sequences feature latent periodicity. At the same time, the functional significance of the latent periodicity can be different. One can suppose involvement of latent periodicity in the formation of protein spatial structure and realization of their biological function. Latent periodicity in an amino acid sequence can provide for specific spatial organization of the protein globules and alternation of β-sheets and α-helices. Stated differently, latent periodicity is a code defining the spatial organization of the corresponding amino acid sequences.

On the other hand, latent periodicity can reflect the origin and evolution of a protein globule. In this case, the latent periods can be considered as elementary units in protein structure, and protein evolution is related to relatively simple multiple duplications of the primary short sequences.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 97-01-00205).

REFERENCES

1. McLachlan, A.D., *J. Mol. Biol.*, 1983, vol. 169, pp. 15–30.
2. Boswell, D.R. and McLachlan, A.D., *Nucleic Acids Res.*, 1984, vol. 12, pp. 457–464.
3. Heringa, J. and Argos, P., *Proteins*, 1993, vol. 17, pp. 341–391.
4. Huang, X., Hardison, R.C., and Miller, W., *Comput. Applic. Biosci.*, 1990, vol. 6, pp. 373–381.
5. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, F., and Wootton, J.C., *Science*, 1993, vol. 262, pp. 208–214.
6. Benson, G. and Waterman, S., *Nucleic Acids Res.*, 1994, vol. 22, pp. 4828–4836.
7. Benson, G., *J. Comput. Biol.*, 1997, vol. 4, pp. 351–367.
8. Needleman, S.B., Wunsch, C.D., *J. Mol. Biol.*, 1970, vol. 48, pp. 443–453.
9. Gusev, V.D., Kulichkov, V.A., and Chupakhina, O.M., *Mol. Biol.*, 1991, vol. 25, pp. 825–834.
10. Gusev, V.D., Kulichkov, V.A., and Chupakhina, O.M., *Mol. Biol.*, 1991, vol. 24, pp. 1080–1089.
11. Gusev, V.D., Kulichkov, V.A., and Chupachina, O.M., *Biosystems*, 1993, vol. 30, pp. 201–214.
12. Kolmogorov, A.N., *Teoriya informatsii i teoriya algoritmov* (Theory of Information and Theory of Algorithms), Moscow: Nauka, 1987, pp. 212–223.
13. McLachlan, A.D., *Biopolymers*, 1977, vol. 16, pp. 1271–1297.
14. Cornette, J.L., Caese, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., and Delisi, C., *J. Mol. Biol.*, 1987, vol. 195, pp. 659–685.
15. Kolaskar, A.S. and Kulkarni-Kale, F., *J. Mol. Biol.*, 1992, vol. 223, pp. 1053–1061.
16. McLachlan, A.D., *J. Phys. Chem.*, 1993, vol. 97, pp. 3000–3009.
17. McLachlan, A.D. and Stewart, M., *J. Mol. Biol.*, 1994, vol. 235, pp. 1278–1290.
18. Chechetkin, V.R., Knizhnikova, L.A., and Turygin, A.Yu., *J. Biomol. Struct. Dyn.*, 1994, vol. 12, pp. 271–299.
19. Makeev, V.Ju. and Tumanyan, V.G., *Biofisika*, 1994, vol. 39, pp. 294–297.

20. Makeev, V.Ju. and Tumanyan, V.G., *Comput. Appl. Biosci.*, 1995, vol. 12, pp. 49–54.
21. Korotkov, E.V. and Korotkova, M.A., *DNA Sequence*, 1995, vol. 5, pp. 353–358.
22. Korotkov, E.V. and Phoenix, D.A., *Proc. Pacific Symp. Biocomputing 97. Maui. Hawaii, USA*, Singapore: Word Scientific Press, 1996, pp. 222–231.
23. Korotkova, M.A., Korotkov, E.V., and Tulko, J.S., *Comput. Appl. Biosci.*, 1997, vol. 17, pp. 37–44.
24. Korotkov, E.V. and Korotkova, M.A., *DNA Res.*, 1996, vol. 3, pp. 157–164.
25. Kullback, S., *Information Theory and Statistics*, London: John Wiley & Sons Inc., 1959.
26. Roff, D.A. and Bentzen, P., *Mol. Biol. Evol.*, 1989, vol. 6, pp. 539–545.
27. Heringa, J., *Comput. Chem.*, 1994, vol. 18, pp. 233–243.
28. Heringa, J. and Taylor, W.R., *Current Opin. Struct. Biol.*, 1997, vol. 7, pp. 416–421.
29. Perler, F.B., Moon, A.M., Qiang, B.Q., Meda, M., Dalton, M., Card, C., Schmidt-Ullrich, R., Wallach, D., Lynch, J., and Donelson, J.E., *Mol. Biochem. Parasitol.*, 1987, vol. 25, pp. 185–193.
30. Feller, W., *An Introduction to Probability Theory and Its Applications*, N.Y.: John Wiley & Sons, 1970.
31. Christie, J.F., Dunbar, B., Davidson, I., and Kennedy, M.W., *Immunology.*, 1990, vol. 69, pp. 596–602.
32. Spence, H.J., Moore, J., Brass, A., and Kennedy, M.W., *Mol. Biochem. Parasitol.*, 1993, vol. 57, pp. 339–343.
33. Kennedy, M.W., Brass, A., Mccruden, A.B., Price, N.C., Kelly, S.M., and Cooper, A., *Biochemistry*, 1995, vol. 34, pp. 6700–6710.
34. Madaule, P., Furuyashiki, T., Reid, T., Ishizaki, T., Watanabe, G., Morii, N., and Narumiya, M., *FEBS Lett.*, 1995, vol. 377, pp. 243–248.
35. Westphal, A.H. and de Kok, A., *Eur. J. Biochem.*, 1988, vol. 72, pp. 299–305.
36. Lilias, M.G., Branden, C.I., and Banaszak, L.J., *The Enzymes*, London: Acad. Press, 1975, vol. 10, p. 68.
37. Thekkumkara, T.J., Pous, G., Sumair, M., Jentoft, J.E., and Patel M.S., *Alpha-keto Acid Dehydrogenase Complexes. Organization, Regulation and Biomedical Ramifications*, Roche, T.E. and Patel, M.S., Eds., N.Y.: Annals New York Acad. Sci., 1989, vol. 573, pp. 113–129.
38. Wierenga, R.K., Terpstra, P., and Hol, W.G.J., *J. Mol. Biol.*, 1986, vol. 187, pp. 101–108.
39. Rice, D.W., Schulz, G.E., and Guest, J.R., *J. Mol. Biol.*, 1984, vol. 174, pp. 483–496.
40. Dang, V.D., Bolotin-Fukuhara, M., Daignan-Fornier, B., *EMBL/GENBANK/DDBJ Data Banks, Acc. Number P49089*, 1995.
41. Johnston, M., Andrews, S., Brinkman, R., Cooper, J., Ding, H., Du, Z., Favello, A., Fulton, L., Gattung, S., Greco, T., Kirstein, J., Kucaba, T., Hallswoth, K., Hawkins, J., Hiller, L., Jier, M., Johnson, D., Johnston, L., Langston, Y., Latreille, P., Le, T., Mardis, E., Menezes, S., Miller, N., Nhan, M., Pauley, A., Peluso, D., Rifken, L., Riles, L., Taich, A., Trevaskis, E., Vignati, D., Wilcvox, L., Wohldman, P., Vaudin, M., Wilson, R., and Waterston, R., *EMBL/GENBANK/DDBJ Data Banks, Acc. Number P49089*, 1995.
42. Pishko, E.J., Kirkland, T.N., and Cole, G.T., *Gene*, 1995, vol. 167, pp. 173–177.