DNA RESEARCH 6, 153–163 (1999)

Method Revealing Latent Periodicity of the Nucleotide Sequences Modified for a Case of Small Samples

Maria B. CHALEY,* Eugene V. KOROTKOV, and Konstantin G. SKRYABIN

Centre Bioengineering of Russian Academy of Sciences, Prospect 60-letiya Oktyabrya, 7/1, 117312, Moscow, Russia

(Received 25 May 1998; revised 12 April 1999)

Abstract

An earlier reported method for revealing latent periodicity of the nucleotide sequences has been considerably modified in a case of small samples, by applying a Monte Carlo method. This improved method has been used to search for the latent periodicity of some nucleotide sequences of the EMBL data bank. The existence of the nucleotide sequences' latent periodicity has been shown for some genes. The results obtained have implied that periodicity of gene structure is projected onto the periodicity of primary amino acid sequences and, further, onto spatial protein conformation. Even though the periodic structure of gene sequences has been eroded, it is still retained in primary and/or spatial structures of corresponding proteins. Furthermore, in a few cases the study of genes' periodicity has suggested their possible evolutionary origin by multifold duplications of some gene's fragments.

Key words: computer sequence analysis; latent periodicity; gene structure; protein structure; relationship between gene and protein structures

1. Introduction

A study of nucleotide and amino acid sequences' periodicities is one of the ways to explore the structure of various genes and their encoded proteins.[1–6] One can suppose that periodicity of spatial organisation of proteins leads to certain structural arrangements of nucleotide and/or amino acid sequences and vice versa.[7] Furthermore, analysis of periodicity in gene sequences suggests that particular genes have probably arisen by a consecutive chain of duplications of some DNA fragment.[8,9]

In general, periodicity may be classified as homologous (perfect), eroded and latent. Homologous periodicity implies, for instance, periodicity of such a kind $(ATCGT)_n$, where n may range from a few units to the figures which are large enough. Eroded (imperfect) periodicity may be represented as a sequence of repeated similar units in which some changes of nucleotides have occurred. So, different repeated units of eroded periodicity have no perfect homology between themselves. In the case of latent periodicity, one can say only about statistical sufficiency for particular nucleotides to be present in each site of the periodic unit. For instance, the latent periodicity may be: $\{(A/G)N(C/T)(G/C/T)(T/A)\}_n$. Here, two bases, A and G, are encountered at the first position of the latent periodic units five nucleotides long in the majority of cases, at the second position any base may occur, at the third — C and T are the most probable and so on. If a sequence having the latent periodicity is long enough then such a period may be considered as statistically significant, i.e., the probability that a latent periodicity will arise by chance is quite small.[8,9]

A search of periodic structure of the genes implies two tasks to be solved. The first task is to elaborate mathematical methods revealing so far eroded (latent) periodicity. If some genes have really arisen by duplication of a short sequence then now these fragments may have diverged so far that in the majority of cases it is impossible to see any homology between them directly. The same could be said about the latent periodicity resulting from periodic organisation of a spatial protein structure.

The known algorithms mainly reveal homologous periodicity including imperfect periodicity.[1–5] Other approaches are focused on searching latent periodicity which is known beforehand.[10] However, to reveal latent periodicity which is not known in advance and to value a statistical significance of the found periodicity is very difficult using these approaches.

The second task which is very actual in searching for periodic gene structures is to reveal a latent periodicity pattern distorted by nucleotide insertions (deletions). This much more complicated task is hard enough even for the most modern computing machines. An analyt-

Communicated by Masahiro Sugiura

* To whom correspondence should be addressed. Tel. +7-095-135-2161, Fax. +7-095-135-0571, E-mail: mariam@biengi.ac.ru

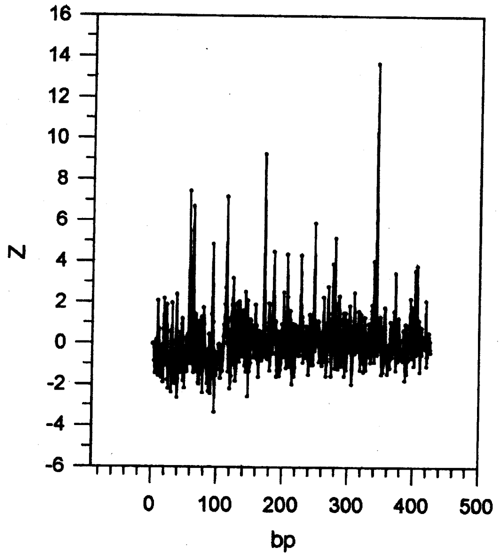

Figure 2. A set of Figs. 2–7 presents the spectra of Z values for the revealed latent periodicity regions of six various genes. Z values have been determined for the latent periodicity of $2, 3, 4, \cdots, n$ bases. The maximum length of the analysed period was equal to one half of the full length of a latent periodicity region. Z values equal to not less than 5 corresponded to a probability of less than 10^{-6} that the latent periodicity would arise by chance. The greater the value of Z, the lower the probability of an accident. Here, the figure shows the spectrum of Z values for the revealed latent periodicity region of toxin A gene of *C. difficile* (Accession number X51797). The large Z value equal to 13.68 for period's length of 342 bases suggests a very probable duplication of DNA fragment of such length on the gene sequence.

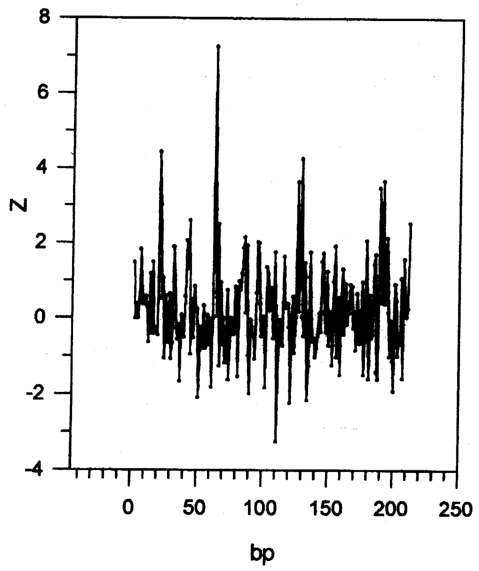

Figure 3. The spectrum of Z values for the latent periodicity region of *E. coli* RhsD gene (Accession number X60999). The latent periodicity of 63 bases corresponds to a known amino acid motif GxxxRYxYDxxGRL(I/T) which has a tendency to be present at each 21-amino-acids periodicity.[17]

mation of existing periodicity in the nucleotide sequences. Table 1 includes data about some genes in which the latent periodicity was revealed, with the EMBL Accession numbers indicated. The co-ordinates, the period length L and its corresponding Z value are also shown for each latent periodicity region. The probability that the latent periodicity arose by accident was no more than 10^{-6} for all examples in Table 1.

3.1. *Arguments in the genes' latent periodicity is full of sense*

3.1.1. *Latent gene periodicity overlaps with known repeats*

A few examples of genes where a search of the latent periodicity has led to known periodic structures are discussed here to prove the reliability of our method.

Thus, in clone including toxin A gene of *Clostridium difficile.* (Accession number X51797), a region of the latent periodicity 853 bp long with a repeated unit of 342 bp was revealed in the frame of known repeated region of toxin A.[14,15] The latent periodicity region overlaps with the central part of a known periodic region and spans one-third of its total length (2499 bp). The structure of the periodic region at the 3′- end of toxin A gene of *C. difficile* has been described earlier as consisting of alternating subsequences of classes I and II. The subsequences of class II were also divided into groups A, B, and C. On average, from three to five subsequences of class II were located between each pair of class I subsequences. Amino acid sequences 27 amino

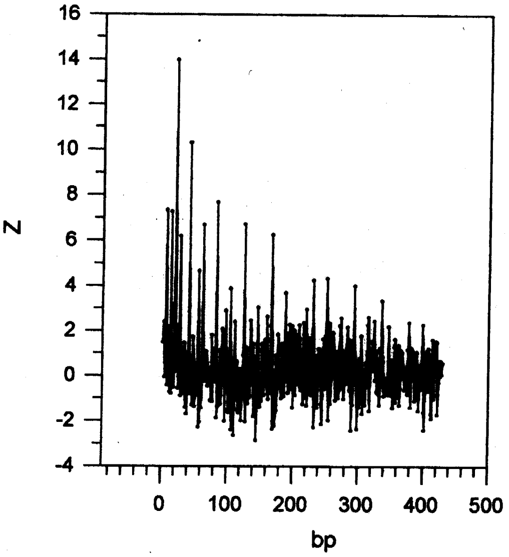

Figure 4. A common periodicity of 21 bases has been demonstrated in the genes of different isoforms of tropomyosin. The spectrum of Z values on this figure is one for the gene with Accession number M19713. Such gene periodicity corresponds to the known period of 7 amino acids in tropomyosin.

spond to a period of 735 bases ($Z = 9$). In clone U35636, a main period in the gene of 732 bases ($Z = 52$) was first determined, which corresponded to the 7-module superrepeat in the protein. Upon further analysis, all periodic units of 732 bases showed periodicity of 105 bases ($Z > 6$). The reason why the periodicity of 105 bases was not obvious in the long sequence was probably due to deletions and insertions distorting the original fine periodicity of the gene. In pairwise comparison the sequences of hidden repeats of 105 and 732 bases showed no more than 40% homology within each group.

Since a clear relationship between gene and protein periodicity have been shown above in terms of its functional significance for protein complex assembly, we consider a few examples where the related latent periodicity of a gene seems to be a single pointer to a mechanism of polypeptide outer interaction based on structural peculiarities of its amino acid sequence. In all next examples no peculiarities in genes' sequences and in their protein products besides the revealed latent periodicity have been shown before.

The latent periodicity of 48 bases over the length of 313 bases was found in the gene encoding a precursor of apocytochrome f of cyanobacterium *Agmenellum quadruplicatum* (Accession number M74514). The region of the latent periodicity included an area encoding a gem-binding domain CANCH. According to Widger's data,[23] gem-binding domain of mature peptide of apocytochrome f is located at amino acids 62–66. A corresponding DNA region ranging from 851 to 865 bases was inside the region of revealed latent periodicity whose coordinates in the clone of the EMBL data bank were the bases ranging from 829 to 1141 (see also Table 1).

In clones including the *tolA* and *tolB* genes of *E. coli* (Accession number M28232) the latent periodicity of the *tolB* gene of 33 bases, over the length of 659 bases, was revealed. Products of genes *tolA* and *tolB* are components of a multistep translocation system of *E. coli* which provides transport of such large molecules as colicins and filamentous DNAs through the cell membrane. It has been earlier determined by Levengood and Webster[24] that an ORF including the *tolB* gene seems to encode two proteins. The larger protein (47.5 kDa) 431 amino acids long and the smaller one (43 kDa), which is presumably the result of processing or in-frame restart within this ORF. The smaller protein (43 kDa) was exclusively found in periplasm, the larger one (47.5 kDa) was mainly associated with bacterial cytoplasmic membrane. However, a sufficient amount of 47.5-kDa protein was always detected out in the membrane fraction. There were no data which would point to a mechanism of protein binding with membrane fraction. According to our data nearly one-half of the *tolB* gene (659 of 1296 bases) has the latent periodicity. The region of the gene's latent periodicity which translated into the amino acid sequence corresponds to a region of amino acids ranging from 142 to 361 while the whole protein (47.5 kDa) is 431 amino acids long. One can suppose the latent periodicity may be related to a mechanism of 47.5-kDa protein binding with the membrane fraction.

The latent periodicity region of phosphofructokinase A gene (*pfkA*) of the length of 897 bases (close to the full length of the gene equal to 963 bases) was revealed in *E. coli* clones (Accession numbers X02519 and L19201). Phosphofructokinase catalyses the phosphorylation of fructose 6-phosphate into fructose 1,6-biphosphate. Two phosphofructokinases were identified in *E. coli*: phosphofructokinase 1 (or phosphofructokinase A) and phosphofructokinase 2 (or phosphofructokinase B).[25] Ninety percent of phosphofructokinase's activity is attributable to phosphofructokinase 1. A comparative analysis of amino acid sequences showed similarity between *E. coli* phosphofructokinase-1 and analogous ferments of *Bacillus stearothermophilus* and of rabbit muscles, but no similarity was detected with *E. coli* phosphofructokinase-2.[26] The ferment is a tetrameter consisting of four identical units of 320 amino acids long.[26,27] According to our data, the phosphofructokinase 1 primary sequence might be formed by repetition of a fragment of 40 amino acids which corresponds to the gene's latent periodicity of 120 bases.

In a clone of *E. coli* (Accession number J01687) a latent periodicity of 90 bases was revealed on almost one-half of the full length of DNA primase gene *dnaG* (811 of 1746 bases). The product of the *dnaG* gene is a kind of RNA polymerase called primase which interacts with DNA to synthesise RNA olygonucleotides directing DNA synthesis. The *in vitro* system of DNA synthesis for G4 phage DNA-primase recognises a specific region of single-strand viral template and in the presence of single-strand binding protein (SSB) it synthesises complementary RNAs 14–29 nucleotides long (RNA primers) which then are continued by DNA polymerase III synthesising the full complementary strand of DNA. Periodic structure of DNA primase is not known.[28]

A region of the latent periodicity of 102 bases was revealed in a clone containing a menaquinone (MQ) operon of *Bacillus subtilis* (Accession number M74538), nearly from the very beginning of the *menB* gene along the length of 472 bases. MQ plays a central role in an oxidative respiration chain of *B. subtilus.* Biosynthesis of MQ requires the formation of naphthoquinone-ring through a series of specific reactions issued from the shikimate-pathway. Earlier MQ-specific reactions catalyse the formation of *o*-succinylbenzoate (OSB) from isochorismate; later reactions convert OSB into dihydroxynaphthoate (DHNA), using OSB-CoA (co-ferment A) as a medium. The sequences of the *menE* and *menB* genes have been cloned, and they encode OSB-CoA synthase and DHNA synthase, respectively. The *menB* protein product converts OSB-CoA complex formed from OSB with the help of *menB* product into DHNA. No structural peculiarities at the 5'-end of the *menB* gene or in its product were noticed before.[29]

In a clone of *Haemophilus influenzae* (Accession number Z33502), which includes a fimbrial gene cluster, in the HifC gene over the length of 1005 bases a latent periodicity of 117 bases was revealed (Fig. 6). A region of latent periodicity was localised in the central part of the gene and makes up almost one half of its full length. The bacteria *H. influenzae* causes chronic bronchitis, inflammation of the middle ear and meningitis. A structural complex fimbriae is formed at the surfaces of bacteria when they colonise mucous membranes of human cells. This complex is encoded by a cluster consisting of five genes: HifA (a main fimbrial subunit), HifB (a chaperone), HifC (an outer membrane usher), HifD and HifE (two minor subunits). Mutation analysis of individual fimbrial genes has demonstrated that HifB and HifC are the essential genes for fimbrial assembly.[30] One can suppose the revealed periodicity in HifC gene is related with the amino acid sequence periodicity of its product, and furthermore it is required in forming the fimbriae complex.

In a clone of *Agrobacterium rhizogenes* (Accession number X51418) the latent periodicity of the *virA* gene was revealed over the length of 553 bases, what was equal to one-fifth of the full gene's length. A repeated unit of the latent periodicity is 105 bases long. *virA* is one of the genes of a complicated mechanism which promotes the growth of malignant tumours of plants infected by *A. rhizogenes* or *A. tumeraciens* bacteria. Genes of the *vir* locus (*virA*, *virB*, *virG*, *virC*, *virD*, *virE*) are exactly induced by phenolic components of plants such as acetosyringone. It is supposed that the product of *virG* gene is a positive regulator of expression of *vir* genes' set, and product of *virA* gene seems to play a role in activation of *virG* gene.[31]

The latent periodicity of 117 bases over the length of 983 bases was revealed in a clone including alpha-glucan phosphorylase gene (E.C. 2.4.1.1) of *E. coli* (Accession number J03966).[32] A region of the latent periodicity spans slightly less than one-half of the whole gene.

3.3. Uniform latent periodicity of some bacterial chemoreceptor genes

Findings of special interest strengthen the supposition about a relationship between the latent periodicity of genes and the structural peculiarities of its encoded protein and, furthermore, the protein's function are the latent periodicity of 21 bases in genes of various bacterial chemoreceptors (see also Fig. 7).

In a clone including the chemoreceptor gene *mcpA* of *Caulobacter crescentus* (Accession number X66502) a latent periodicity of 21 bases was revealed over the length of 1172 bases, i.e., on more than one-half of the full gene length. When being translated into the amino acid sequence, a region of the latent periodicity corresponded to amino acids 189 to 578. The bacterial signal of transduction *mcpA* is an integral protein receptor. There are two transmembrane domains on its amino acid sequence: TM1 consisting of the first 40 amino acids, and TM2 which is located at amino acids 210 to 235. As it was earlier supposed, the regions of metilation are peptides: K1, from amino acids 400 to 420; and R1, from amino acids 585 to 600.[33] As one can see, the amino acid sequence corresponding to the region of latent periodicity includes transmembrane domain TM2 and peptide K1.

The latent periodicity of 21 bases over the length of 828 bases was also revealed in the *mcpA* gene of *B. subtilis* (Accession number L29189; Fig. 7). The corresponding region of amino acids ranges from 333 to 608, and the full length amino acid sequence corresponding to the whole gene is equal to 662 amino acids.[34]

In a clone of *Enterobacter aerogenes* (Accession number M26411) the latent periodicity of *tse* chemoreceptor gene (tse - taxis to serine) of 21 bases was revealed over the length of 792 bases. The corresponding region of the amino acid sequence ranges from amino acids 199 to 462. The full amino acid sequence of receptor is 558 amino acids in length.[35]

In clones of *E. coli* (Accession numbers J01705, U14003) regions of latent periodicity of 21 bases were revealed over the lengths of 644 bases and of 670 bases in the *tar* and *tsr* chemoreceptor genes, respectively. The full length of *tar* chemoreceptors is 554 amino acids,[36] and for *tsr* chemoreceptor it is 551 amino acids.[37] The amino acids of *tar* receptor range from amino acids 245 to 458 and those of the *tsr* receptor range from amino acids 242 to 465, corresponding to the regions of latent periodicity of respective genes. It was noted before the *tar* and *tsr* genes seemed to be the members of the same family which evolved from a common precursor.[38] A region of maximum identity of *tar* and *tsr* proteins covers the amino acids from 360 to 407. It has been supposed that the function of this region is to carry out interactions between transducers or to form stable multimers or reversible interactions of transducers when they are binding with ligands.[38] The metilation region K1 for *tar* and *tsr* proteins is from amino acids 295 to 317. Amino acids of *tar* and *tsr* proteins corresponding to the latent periodicity regions of the genes include peptide K1 and the region of maximum identity of the proteins which is important for transfer of the chemoreception signal into the cytoplasm.

For all examples of chemoreceptors given above it may be noted that the latent periodicity of 21 bases, as a rule, occurred in that region of the gene which corresponded to the second transmembrane and cytoplasmic domains. It is very unlikely that such identical periodic structures of chemoreceptor genes is casual, and it is conditioned by the same mechanism of transduction of the signal of binding with ligand at a membrane surface into cytoplasm.

Table 2. Results of the latent period analysis for five bacterial chemoreceptors (see Table 1 and/or text for detail). The numbered sites of the latent periodicity unit correspond to the columns. The frequencies of bases in each site are shown along the rows. The last row presents a "quasi-consensus" of the latent period of 21 bases long, consisting of the separated bases whose frequencies are greater than 0.2500.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	.1179	.2308	.2769	.2667	.4359	.2308	.2821	.3282	.2205	.1282	.1846	.2154	.1590	.2769
T	.1538	.1949	.1538	.0872	.1282	.2564	.0923	.1590	.2205	.2667	.1897	.1744	.1026	.1744
C	.3949	.2513	.2462	.2974	.1897	.2923	.3077	.1744	.3128	.3128	.2615	.4205	.3026	.2718
G	.3333	.3231	.3231	.3487	.2462	.2205	.3179	.3385	.2462	.2923	.3641	.1897	.4359	.2769
	C/G	G	G/A	G/C	A	C	G/C/A	G/A	C	C/G	G	C	G/C	A/G/C

	15	16	17	18	19	20	21
A	.2410	.2821	.2821	.2564	.2103	.2103	.2051
T	.1795	.0769	.1590	.2513	.1385	.2051	.2103
C	.3179	.2821	.2359	.2667	.3333	.2718	.2513
G	.2615	.3590	.3231	.2256	.3179	.3128	3333
	C	G/A/C	G/A	N	C/G	G	G

However, it is natural that binding of the membrane-spanning domain to a ligand would show more structural variability.

We analysed a set of all untruncated units of the latent periodicity of 21 bases for the five chemoreceptor genes described above. The frequencies of nucleotides A, T, C, and G were determinated at each site of the latent period. Table 2 shows the results. The last row of Table 2 presents a quasiconsensus of the chemoreceptor's latent period, that is a sequence of the most frequent bases at each site of the period. The most probable sequence of the corresponding amino acid codons according to the quasiconsensus is CGG GAC GGC CGC GA(G)C GGN CGG. Its corresponding amino acid sequence is ARG ASP GLY ARG ASP(GLY) GLY ARG.

3.4. Conclusions

Our research on the latent periodicity in separate genes implies that these genes have probably arisen as a result of numerous duplications of some DNA fragment. At present, however, copies of such fragment are so far eroded that no homology may be revealed between them. However, a periodicity existed at the level of nucleotide sequence is still retained on primary amino acid sequence or is traced in spatial protein organisation. The periodic spatial organisation of protein might also influence a formation of gene's latent periodicity. We believe that gene structure being interrelated with protein structure may provide a key both to the conformational interaction of proteins among themselves and to the formation of their structural complexes. Proteins having the same kind of gene periodicity probably function in a similar fashion. Our research is only the first step in this new and interesting area of research.

References

1. Cheever, E. A., Overton, G. C., and Searls, B. B. 1991, Fast fourier transform-based correlations of DNA sequences using complex plane encoding, *CABIOS*, **7**, 143–154.
2. Heringa, J. and Argos, P. 1993, A method to recognize distant repeats in protein sequences, *Proteins*, **17**, 391–411.
3. McLachlan, A. D. 1993, Multichannel fourier analysis of patterns in protein sequences, *J. Phys. Chem.*, **97**, 3000–3009.
4. Benson, G. and Waterman, M. S. 1994, A method for fast database search for all k-nucleotide repeats, *Nucleic Acids Res.*, **22**, 4228–4836.
5. Chechetkin, V. R., Knizhnikova, L. A., and Turygin, A. Yu. 1994, Three-quasiperiodicity, mutual correlations, ordering and long modulations in genomic nucleotide sequences of viruses, *J. Biomol. Str. and Dyn.*, **12**, 271–299.
6. Benson, G. 1997, Sequence alignment with tandem duplication, *J. Comput. Biol.*, **4**, 351–367.
7. Korotkov, E. V. and Korotkova, M. A. 1995, Latent periodicity of some human gene DNA sequences, *DNA Seq.*, **5**, 353–358.
8. Korotkov, E. V., Korotkova, M. A., and Tulko, J. S. 1997, Latent sequence periodicity of some oncogenes and DNA-binding protein genes, *CABIOS*, **13**, 37–44.
9. Korotkov, E. V. and Phoenix, D. A. 1997, In: Altman, R. B., Dunker, A. K., Hunter, L., Klein, T. (eds) Proceedings of Pacific Symposium on Biocomputing 97, Word Scientific Press, Singapore-New-Jersey-London, pp. 222–231.
10. Chechetkin, V. R. and Turygin, A. Y. 1996, Study of correlations in DNA sequences, *J. Theor. Biol.*, **178**, 205–217.
11. Korotkov, E. V. and Korotkova, M. A. 1996, Enlarged similarity of nucleic acid sequences, *DNA Res.*, **3**, 157–164.
12. Roff, D. A. and Bentzen, P. 1989, The statistical analysis of Mitochndrial DNA polymorphisms: χ^2 and the

problem of small samples, *Mol. Biol. Evol.*, **6**, 539–545.

13. Pagano, M. and Halvorsen, K. T. 1981, An algorithm for finding the exact significance levels of r × c contingency tables, *J. Am. Stat. Assoc.*, **76**, 931–934.
14. Jonson, J. L., Dove, C. H., Price, S. B., Sickles, T. W., Phelbs, C. J., and Wilkins, T. D. 1988, In: Hardie, J. M. and Borriello, S. P. (eds) Anaerobes Today, Wiley and Sons Ltd., pp. 115–123.
15. Sauerborn, M. and von Eichel-Streiber, C. 1990, Nucleotide sequence of *Clostridium difficile* Toxin A, *Nucleic Acids Res.*, **18**, 1629–1630.
16. Fujino, I., Beguin, P., and Aubert, J. P. 1993, Organization of a Clostridium Thermocellum gene cluster encoding the cellulosomal scaffolding protein Cip A and a protein possibly involved in attachment of the cellulosome to the cell surface, *J. Bacteriol.*, **175**, 1891–1899.
17. Feulner, G., Gray, J. A., Kirschman, J. A., et al. 1990, Structure of the rhsA locus from Escherichia coli K-12 and comparison of rhsA with other members of the rhs multigene family, *J. Bacteriol.*, **172**, 446–456.
18. Phillips Jr., G. N., Fillers, J. P., and Cohen, C. 1986, Tropomyosin crystal structure and muscle regulation, *J. Mol. Biol.*, **192**, 111–131.
19. McLachlan, A. D. and Stewart, M. 1976, The 14-fold periodicity in a-tropomyosin and the interaction with actin, *J. Mol. Biol.*, **103**, 271–298.
20. McLachlan, A. D., Stewart, M., and Smillie, L. B. 1975, Sequence repeats in a-tropomyosin, *J. Mol. Biol.*, **98**, 281–291
21. Stone, D., Sodek, J., Jonson, P., and Smillie, L. B. 1975, Proc. 9th FEBS Meeting (Budapest), **31**, 125–136.
22. Wang, K., Knipfer, M., and Huang., Qi-Q. at al. 1996, Human skeletal muscle nebuline sequence encodes a bluprint for thin filament architecture, *J. Biol. Chem.*, **271**, 4304–4314.
23. Widger, W. R. 1991, The cloning and sequencing of the Synechoccus sp. pcc 7002 petCA operon: Implications for the cytochrome c-553 binding domain of cytochrome f, *Photosyn. Res.*, **30**, 71–84.
24. Levengood, S. K. and Webster, R. E. 1989, Nucleotide sequence of the tolA and tolB genes and localization of their products: Components of a multistep translocation system in Escherichia coli, *J. Bacteriol.*, **171**, 6600–6609.
25. Fraenkel, D. G., Kotlarz, D., and Buc H. 1973, 2-Fructose 6-phosphate kinase activities in Escherichia coli, *J. Biol. Chem.*, **248**, 4865–4866.
26. Hellinga, H. W. and Evans, P. R. 1985, Nucleotide sequence and high-level expression of the major Escherichia coli phosphofructokinase, *Eur. J. Biochem.*, **149**, 363–373.
27. Plunkett, G., Burland, V., Daniels, D. L., and Blattner, F. R. 1993, Analysis of the Escherichia coli genome. III. DNA sequence of the region from 87.2 to 89.2 minutes, *Nucleic Acids Res.*, **21**, 3391–3398.
28. Smiley, B. L., Lupski, J. R., Svec, P. S., McMacken, R., and Godson, G. N. 1982, Sequences of the Escherichia coli dnaG primase gene and regulation of its expression, *Proc. Natl. Acad. Sci. USA*, **79**, 4550–4554.
29. Drissol, J. R. and Taber, H. W. 1992, Sequence organization and regulation of the Bacillus subtilis men BE operon, *J. Bacteriol.*, **174**, 5063–5071.
30. van Ham, M. S., van Alphen, L., Mooi, F. R., and van Putten, J. P. 1994, The fimbrial gene cluster of Haemophilus influenzae type b, *Mol. Microbiol.*, **13**, 673–684.
31. Endoh, H., Aoyama, T., Hirayama, T., and Oka, A. 1990, Characterization of the virA gene of the agropine-type plasmid pRiA4 of Agrobacterium rhizogenes, *FEBS Lett.*, **271**, 28–32.
32. Yu, F., Jen, Y., Takeuchi, E. et al. 1988, Alpha-glucan phosphorylase from Escherichia coli: Cloning of the gene and purification and characterization of the protein, *J. Biol. Chem.*, **263**, 13706–13711.
33. Alley, M. R. R., Maddock, J. R., and Shapiro, L. 1992, Polar localization of a bacterial chemoreceptor, *Genes Dev.*, **6**, 825–236.
34. Hanlon, D. W. and Ordal, G. W. 1994, Cloning and characterization of genes encoding methyl-accepting chemotaxis proteins in Bacillus subtilis, *J. Biol. Chem.*, **269**, 14038–14046.
35. Dahl, M. K., Boos, W., and Manson, M. D. 1989, Evolution of chemotactic — signal transducers in enteric bacteria, *J. Bacteriol.*, **171**, 2361–2371.
36. Krikos, A., Mutoh, N., Boyd, A. W., and Simon, M. I. 1983, Sensory transducers of *E. coli* are composed of discrete structural and functional domains, *Cell*, **33**, 615–622.
37. Burland, V., Plunkett, III G., Sofia, H. J., Daniels, D. L., and Blattner, F. R. 1995, Analysis of the Escherichia coli genome VI: DNA sequence of the region from 92.8 through 100 minutes, *Nucleic Acids Res.*, **23**, 2105–2119.
38. Boyd, A., Krikos, A., and Simon, M. I. 1981, Sensory transducers of Escherichia coli are encoded by homologous genes, *Cell*, **26**, 333–343.

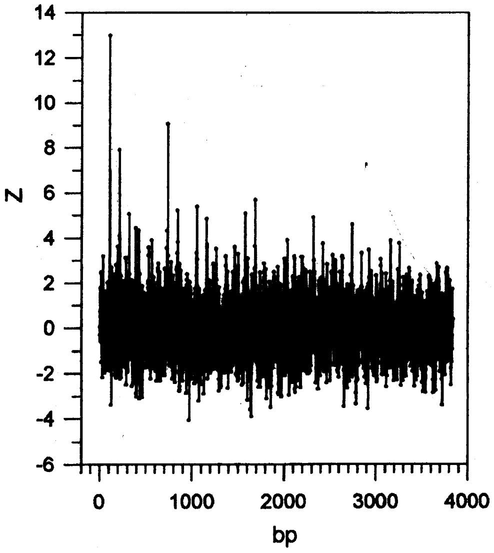

Figure 5. The spectrum of Z values for the latent periodicity of human nebulin gene (Accession number U35637). The biggest z value equal to 13.0 corresponds to a period of 105 bases, which is in agreement with the protein modular structure of ~ 35 residues.

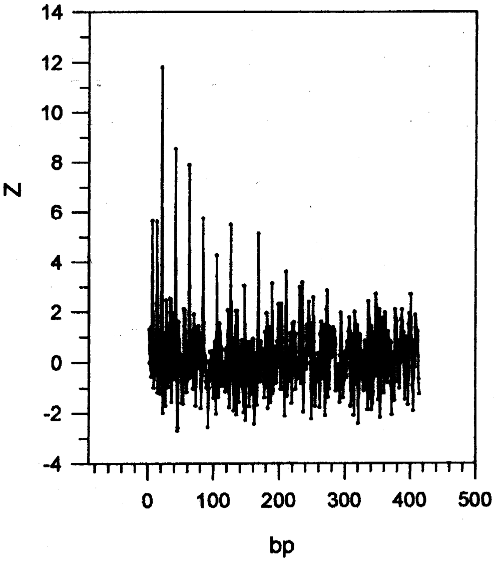

Figure 7. The spectrum of Z values of the latent periodicity region of *B. subtilis* mcpA gene (Accession number L29189). According to our data, the latent periodicity of 21 bases seems to be common for bacterial chemoreceptors (see text for explanation).

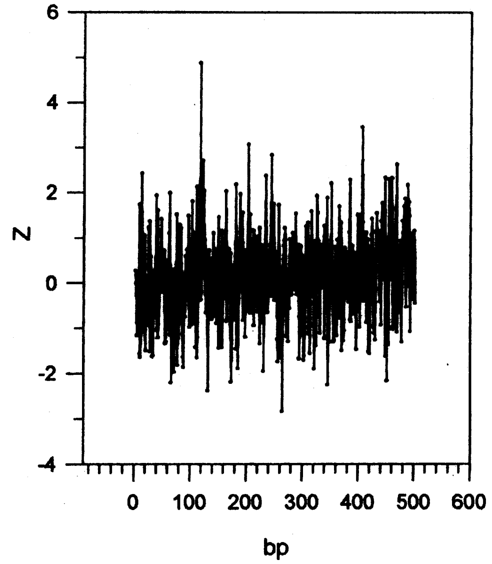

Figure 6. The spectrum of Z values for the latent periodicity region of *H. influenzae* HifC gene (Accession number Z33502). A periodicity of the gene of 117 bases is shown for the first time.

acids long corresponded to the nucleotide subsequences of class I; sequences of 21 amino acids corresponded to class II nucleotide subsequences.[14] The value Z equal to 13.68 points at a very probable duplication of a 342-bp DNA fragment in the region of revealed latent periodicity (Fig. 2). The length of the latent periodicity region accounts for about 2.5 copies of repeated DNA fragment of 342 bp. Each copy (including incomplete copies) was analysed for the presence of periodicities. The first copy showed a latent periodicity of 63 bp (corresponding to 21 a.a.) with a very large Z value of 13.09. In the second copy the latent periodicity was not revealed at a significant level. In the third incomplete copy, the latent periodicity of 63 bp with Z equal to 10.3 was present. So, our data are in a good agreement with earlier known data about periodic organisation of studied DNA region and besides point at a probable duplication of DNA fragment of 342 bp in this region, what was not observed before.

In clone of *Clostridium thermocellum* (Accession number X67506) a region of the latent periodicity of 297 bp over the length of 1138 bp coincides with unknown repeats revealed in ORF1 (open reading frame) of the clone which encodes an S-layer protein. Two of four unknown repeats are almost totally overlapped with the region of latent periodicity (see corresponded co-ordinates in Table 1). The large Z value equal to 12.99 for the latent period of 297 bp assumes that the latent periodicity of this region has arisen by multiple duplications of DNA fragment 297 bp long. Let us note that *C. thermocellum* is an anaerobic bacterium producing a highly active thermostable cellulose system whose different cellulolytic components are bound in a whole high molecular complex. A peptide encoded by ORF1 is one of the components of this complex and can be disposed at a cell's surface thanks to S-layer repeats.[16] For four repeated regions in ORF1, two of which were also found by the method searching the latent periodicity, no data about their internal structure and functions were adduced earlier.

Table 1. Data about 17 regions of the latent periodicity revealed in various genes, listing in order of their discussion in the text, are demonstrated. Accession codes of the EMBL data bank, co-ordinates of the genes and some of their known peculiarities, length of the latent period L and its corresponding value Z for each latent periodicity region are also shown. Z value equal to 5 corresponds to a probability of the latent periodicity to arise by accident of less than 10^{-6}. The more increased Z values correspond to the much less probabilities.

species	gene/product	Accession numbers	co-ordinates of gene and its peculiarities	co-ordinates of the latent periodicity region	L (bp)	Z
C. difficile	toxA/toxin A	X51797	1282-9414, repetitive region: 6826-9324	7533-8385	342	13.68
C. thermocellum	pORF1/S-layer protein	X67506, S43430	259-5253, repeate regions: 364-831, 877-1344, 1483-1950, 2077-2547. 2569-4389 - TP SDEP repeats, 4615-4803 - S-layer-like repeat, 4804-5013 - S-layer-like repeat, 5014-5196 - S-layer-like repeat	1360-2497	297	12.99
E. coli	RhsD/product of RhsD ORF	X60999	460-4740, RhsD core: 460-4206	1838-2261	63	7.25
A. quadruplicatum	petA/apocyto-chrome f precursor	M74514	788-1642	829-1141	48	6.64
E. coli	tolB/component of a translocation system of colicins and filamentous DNAs	M28232	1474-2769	1898-2556	33	7.80
E. coli	pfkA/ phosphofructo-kinase 1	X02519	115-1077	160-1056	120	5.54
E. coli	pfkA/ phosphofructo-kinase 1	L19201	69185-70147	69229-70125	120	5.88
E. coli	dnaG/ DNA primase	J01687	1041-2786	1265-2075	90	5.16
B. subtilis	menB/DHNA - synthase	M74538, M74182, M74183	4157-4942	4167-4638	102	6.18
H. influenzae	Hif C	Z33502	2937-5450	3239-4243	117	4.88
A. rhizogenes	virA	X51418	463-2952	696-1228	105	5.35
E. coli	alpha-glucan phosphorylase (EC 2.4.1.1)	J03966	61-2490	403-1385	117	5.19
C. crescentus	mcpA/ chemoreceptor	X66502	927-2900	1490-2661	21	10.75
B. subtilis	mcpA/ chemoreceptor	L29189	4505-6490	5502-6329	21	11.81
E. aerogenes	tse/chemoreceptor	M26411	1332-3005	1926-2717	21	9.67
E. coli	tar/chemoreceptor	J01705, V01504	151-1812	883-1526	21	10.63
E. coli	tsr/chemoreceptor	U14003	282491-284146	283216-283885	21	10.63

3.1.2. *Relations between the latent periodicity of gene and the structure of its encoded protein*

An example of the relationship between the latent DNA periodicity and structural peculiarities of amino acid sequence one could see in clone of *Escherichia coli* (Accession number X60999) which contains the RhsD gene. The Rhs family was so named according to its function as rearrangement hot spots in intrachromosomal recombination. Members of the family were identified by their activity as sites of crossover. Their distinctive feature is the availability of a highly conserved core of 3.7 kb long. The cores are usually flanked by sequences which have no similarity. The similarity of cores usually begins with start codon of ORF (open reading frame). Cores of RhsA, RhsB, RhsC diverge at 1 or 2 per cent, and only RhsD core is different from others at 18%. A product of ORF has repeated amino acid motif GxxxRYxY-DxxGRL(I/T) which is present 28 times. The motif has a tendency to encounter at each 20 or 21 amino acids especially in a region of amino acids ranging from 418 to 711.[17] The DNA bases corresponding to this amino acid region have co-ordinates in the clone from 1693 to 2572. A region of the revealed latent periodicity of 63 bases was placed from 1838 to 2261 bases and was localised in the frame of the bases corresponding to the repeated amino acids' motif (Fig. 3).

Let us note that only partial overlapping of the revealed latent periodicity with known periodic regions was a consequence of deletions (or insertions) occurring in nucleotide sequence, because the method which we have used previously can not reveal periodicity under such circumstances. According to the examples given above, one can suppose the existence of a relationship between the latent periodicity of a gene and the structural peculiarities of the amino acid sequence of its encoded protein.

3.2. *Latent gene periodicity may provide a key to the mechanism of polypeptide interactions*

What may the found hidden periodicity of a gene mean for its encoded protein? It would be reasonable to expect the hidden gene periodicity, at best, to be translated into eroded but still observable periodicity of a protein or into hidden periodicity of polypeptide sequence. Even weak periodicity of polar and hydrophobic residues in primary structure of protein can determine both spatial conformation of the protein and its various outer interactions in different protein-protein or protein-ribonucleic or protein-substrate complexes. Thus, the periodicity of proteins may be considered as a kind of sectoring of the areas of interactions. Regulatory complex of human skeletal muscle proteins consisting of nebulin, tropomyosin and troponin is a clear example of such scaleable interactions. Further, we discuss how the imperfect periodicity of nebulin and tropomyosin corresponds to the latent periodicity of their respective genes. This permits us to speculate that latent periodicity found in many genes listed in Table 1 trends at least to a weak periodicity of the proteins needed for outer interactions.

The thin filaments of skeletal muscles of vertebrates are formed due to interaction between actin α-helix and the coiled coil structure of tropomyosin, so that one molecule of tropomyosin binds seven monomers of actin.[18] Such interaction is supported by charged residues of a 42 a.a. repeat which occurs 7 times in the tropomyosin sequence.[18–20] This repeat preserves six repeating fragments of 7 amino acids.[21] Imperfect tropomyosin periodicity of 7 amino acids was revealed before it was shown that the tropomyosin evolved by multiplicit duplication of a set of 42 amino acids.[21] The search for hidden periodicity in tropomyosin genes of different isoforms (Accession numbers M12125, M19713, M74817, X04201, X06825) revealed periodicity of 21 bases (see Fig. 4) in full agreement with the known period of 7 amino acids. Homology between the pairs of 21 different base periods are ranging from 30% to 50%.

Not so long ago a complex periodic organization of human nebulin has been shown.[22] The main part of the nebuline sequence is built from $\sim$ 150 tandem copies of $\sim$ 35 residue modules which were classified into seven types based on their homology. These modules are combined in super-repeats of a 7-module set (one of each type in the same order). One supposes that the full number of duplicated nebulin modules is determined by the length of thin filaments and the number of actin monomers per helical strand. The 7-module super-repeats arose in nebulin evolution to provide appropriately spaced sites for tropomyosin and troponin binding. In a hypothetical model each super-repeat binds to 7 actin monomers, one tropomyosin and one troponin complex.[22]

Consensus sequences for each type of nebuline modules have been written according to 50% and more homology for each site.[22] On the average, from one-third to one-half positions of module consensus sequence are occupied by conservative amino acids. That is one may say the nebulin amino acid sequence shows vastly eroded periodicity. We analysed a sequence of nebulin gene (Accession numbers U35636 and U35637) to reveal hidden base periodicity, corresponding to earlier determined modules and super-repeats in the protein sequence. In clones U35636 and U35637 the gene parts corresponding to amino acids from 29 to 2468 and from 23 to 2580, respectively (here amino acid numbers are the same as in [5]), were chosen for the test. This was done to start the latent periodicity search from the gene point which corresponds to the beginning of the whole module in protein. The most appropriate pattern to earlier known data about modules of $\sim$ 35 amino acids in the protein was found in clone U35637 (see Fig. 5). A hidden period of 105 bases ($Z = 13$) was apparent there. The existence of super-repeats in the protein was also traced in the gene sequence, because the next large peak of Z value correical solution of the task has not yet been found, and direct sorting of all possible variants of insertions (deletions) together with all kinds of latent periodicity requires a computer of super high performance. It seems likely that such a way will continue to be impossible for a long time. Search of the analytical solution is also complicated by the fact that latent periodicity of symbolic sequences (DNA bases' sequence is considered as a chain of symbols of four letter alphabet) is neither a property of one separated period nor of periods' pairs, it is rather the property of a whole complex of repeated units.

One of the methods revealing latent periodicity has been reported earlier.[7–9] It is based on a principle of enlarged similarity of the symbolic sequences[11] and consists in comparing the artificial periodic sequences which have periods of different length with the symbolic sequences. With this method it has been shown that a valued part of all known genes (more than 20%) has regions of latent periodicity. However, the method has some shortcomings which are conditioned by a variation of the double mutual information value 2I from χ^2 distribution in the case of small samples.

Here, by applying a Monte Carlo method,[12] the earlier proposed approach has been considerably modified so that it is also reliable in the case of small samples. This improved mathematical approach to the search for latent periodicity was applied to nucleotide sequences selected from the EMBL data bank. The latent periodicity of some genes was shown and the possible sense of the revealed latent periodicities of the sequences was discussed.

2. Methods

2.1. Search of latent periodicity using artificial periodic sequences

A comparison of the artificial periodic sequences with a nucleotide sequence was used to reveal the latent periodicity of the latter, as it has been reported earlier.[7–9] An alphabet of the artificial sequences consisted of S(i) letters, here $i = 1, \cdots, n$. If a period of two bases was searched then a sequence $S(1)S(2)S(1)S(2)S(1)S(2)\cdots$ was generated. To seek a period of 3 bases, a sequence $S(1)S(2)S(3)S(1)S(2)S(3)S(1)S(2)S(3)\cdots$ was created. In general, to reveal a period of n bases long, a sequence $S(1)S(2)\cdots S(n)S(1)S(2)\cdots S(n)S(1)S(2)\cdots S(n)\cdots$ was used. The length of the artificial sequence was chosen to be equal to the length of the analysed nucleotide sequence. The artificial sequences having periods of $2, 3, \cdots, n$ letters were compared in turn with the analysed sequence. Each comparison resulted in filling up a matrix $M(4,n)$. Here, an element $M(i,j)$ of the matrix showed a quantity of nucleotides of i-kind ($i = A, T, C, G$) which stood opposite a letter $S(j)$ on the artificial sequence. The double mutual information value 2I was chosen as a measure of similarity, and it was counted proceeding from an $M(4,n)$ matrix.[7,8] An independent varying of both left and right borders of the artificial sequence together with DNA sequence was used to search for a periodicity region whose total length was unknown beforehand. This method has been earlier described in detail by Korotkov and Korotkova,[7] and by Korotkov et al.[8] However, a certain minimum length was required to apply the method.

It may be considered that a nucleotide sequence is not sufficiently long, if a value of any $M(i,j)$ element is less than 5. In this case, the double mutual information value 2I varies from χ^2 distribution, and it becomes impossible to estimate precisely the probability that the latent periodicity arose by accident. Such a situation is called a case of small samples. It usually occurs if an analysed sequence is shorter than 20n bases, where n is the period's length.

2.2. A Monte Carlo method for small samples

The problem of statistical analysis when a sample size is small has been discussed in detail by Roff and Bentzen.[12] The calculated χ^2 may not be reliable to assess the significance of the observed value in this case as χ^2 value may be inflated upward. Some alternate approach is needed to estimate the significance of a data matrix when the values within the cells are very small. For Fisher's exact permutation test and its modified algorithm introduced by Pagano and Halvorsen,[13] the number of required permutations is increased with the number of rows, columns, and total sample size in a nonlinear fashion. Searching the latent gene periodicity implies the analysis of matrix $4 \times n$, where n starting from 3 may in fact be large enough (for example, 30, 90 or 300) because the maximum n is equal to one half of the length of latent periodicity region. So, Fisher's test and Pagano and Halvorsen's algorithm[13] are not computationally feasible in such a case. A Monte Carlo method may be used to generate the distribution of χ^2 expected values by considering all possible arrangements of the data set subjected to the constraint of constancy of the row and column totals. According to the method, one generates casual matrixes $M'(4,n)$ which have the same sum over columns ($X(i) = \sum_j M(i,j), j = 1, \cdots, n$) and over rows ($Y(j) = \sum_j M(i,j), i = 1, \cdots, 4$) as matrix $M(4,n)$. The $X(i)$ value is equal to number of nucleotides of i-kind on an analysed sequence. The $Y(j)$ value is equal to number of $S(j)$ letters in an artificial sequence. Let the mutual information between nucleotide and artificial periodic sequences, corresponding to M matrix, be equal to $I(1)$ ($2I$ is assumed to follow χ^2 distribution). Let us denote mutual information corresponding to M' as I'. Further, a set of N matrixes M', and the number of events $N1$, where $I' > I(1)$, is revealed. Then a probability F of the latter event ($I' > I(1)$) may be appreciated as $N1/N$. If

M(4,n), n=3

	1	2	3	X(i)
A	1	0	1	2
T	2	0	0	2
C	0	2	0	2
G	0	1	2	3
Y(j)	3	3	3	

I

275487	1
20289	1
374169	2
519336	2
845347	3
724637	3
970126	4
943592	4
152687	4

II

970126	4
943592	4
845347	3
724637	3
519336	2
374169	2
275487	1
152687	4
20289	1

M'(4,n), n=3

	1	2	3	X(i)
A	0	0	2	2
T	0	2	0	2
C	1	1	0	2
G	2	0	1	3
Y(j)	3	3	3	

Figure 1. An application of data set randomization algorithm[12] is shown for a short nucleotide sequence of 9 bases being tested on the periodicity of 3 bases. Matrix M represents the original observation. The first column of matrix I contains $L = \Sigma X(i) = \Sigma Y(j)$ random numbers. The second column contains $X(1)$ ones, $X(2)$ twos, $X(3)$ threes, $X(4)$ fours according to matrix M, which in general may be placed arbitrarily. Then, the random numbers are sorted into descending rank, moving along with the numbers in the second column of matrix I. Matrix II shows the result of the process. The next step is filling up of matrix M' on the base of randomized numbers in the second column of matrix II, as described in section 2.2 of the text.

the F value is less than 0.05 then one considers the latent periodicity of the analysed nucleotide sequence has not arisen by accident.

An algorithm of matrix randomizatiom has been proposed by Roff and Bentzen.[12] The matrix $M(4, n)$ shows the numbers of coincidences between the sequences' elements. Here, 4 and n are the sizes of nucleotide and artificial sequence alphabets, respectively. Let L be a total sum of all the matrix's elements. To produce randomization of data matrix $M(4, n)$, one constructs a matrix of two columns and L rows. The first column consists of L random numbers. The second column contains $X(1)$ ones, $X(2)$ twos, $X(3)$ threes and $X(4)$ fours. Then the elements of the first column are sorted according to their ascending (or descending) rank, moving along with corresponding elements of the second column. The last matrix is used to fill up a randomized matrix $M'(4, n)$ having the same values $X(i)$ and $Y(j)$ as the matrix $M(4, n)$. The elements $M'(1,1)$, $M'(2,1)$, $M'(3,1)$, $M'(4,1)$ are obtained by computing the number of ones, twos, threes and fours in the first $Y(1)$ rows of the second column. The elements $M'(1,2)$, $M'(2,2)$, $M'(3,2)$, $M'(4,2)$ are obtained by the same computation in the rows from $Y(1)+1$ to $Y(1)+Y(2)$ of the second column. The same procedure is repeated for the rows from $Y(2)+1$ to $Y(2)+Y(3)$, and from $Y(3)+1$ to $Y(4)$. Figure 1 shows a simple example of matrix $M(4,3)$ randomization following the described steps of the algorithm. After the elements of $M'(4, n)$ have been determined, the mutual information value I' is counted as it was reported earlier.[7,8]

2.3. Z value as a measure of accident

In practice, to be more sure the value F was precisely appreciated, one has to generate so many M' matrixes in order to get $N1$ equal to not less than 10. Even though a nucleotide sequence has noticeable periodicity, the number of M' matrixes is indeed an astronomical value! For instance, one analysed sequence of 200 nucleotides long which consisted of repeated units of 9 bases had the value of $I(1)$ equal to 54.8. More than 10^9 accidental matrixes are needed to appreciate the value F. So, it was more convenient to use of a value $Z = (I(1)\text{-}I'_m)/\sigma$ as a measure of accident that led to the appearance of the periodic structure in the nucleotide sequence. I'_m was a mean value of the mutual information I' over the set of M' matrixes, and σ was equal to the square root of dispersion of the I' value over the set. According to conducted estimations the value Z equal to 5 corresponded to the accidental probability F of no more than 10^{-6}. The values of Z greater than 5 corresponded to the much less probability F. A spectrum of Z values for all possible period lengths n was determined in the result of the calculations. The maximum period length was equal to one half of the full length of an analysed sequence. This made it possible to reveal not only the latent periodicity but also all possible duplications inside a nucleotide sequence. The obtained spectrum of Z values showed the presence of various periodicities in an analysed nucleotide sequence without concrete knowledge of a kind of erosion in each period's site. The examples of Z spectra for some genes are shown in Figs. 2–7.

3. Results and Discussion

The approach described above was applied to a search for the latent periods in nucleotide sequences of the EMBL data bank. The analysis done showed that about 30% of all known genes have regions of latent periodicity which are difficult to reveal by other methods. It should be noted that the elaborated approach does not take into account deletions or insertions of nucleotides. So, periodicity of nucleotide sequence was not revealed if insertions or deletions of the nucleotides took place. For this reason, the value of 30% shoufd be considered as a minimum esti-