THE DATABASES ON PERIODICITY IN NUCLEOTIDE (MRFGS) AND AMINO ACID (MRFPS) SEQUENCES

Chaley M.B., Frenkel F.E., Korotkov E.V.[1]

Center “Bioengineering” RAS
Russian Federation, Moscow 117312, Prospekt 60-letiya Oktyabrya, 7/1

We have built the first versions of the databases on periodicity in nucleotide and protein sequences named as MRFGS and MRFPS (Miscellaneous Repetitive Features in Genetic (Protein) Sequences). The MRFGS database contains periodic DNA/RNA sequences having period length multiple to 3: 6, 9, 12,…,120 bases. The MRFPS database includes protein sequences with the periodicity ranged from 2 to 1046 amino acid residues, the sequences having period length greater than 100 residues make up only 14% of the whole volume of this database. The total numbers of the records in the MRFGS and MRFPS databases are appropriately equal to 12067 and 1072. The MRFGS and MRFPS databases are currently accessible by URL: http://genomics.narod.ru.

Key words: DNA/RNA periodicity, protein sequence periodicity, databases.

Yet 10 years ago scientific community did not conceive in full measure what hard efforts will be needed for a systematization of our knowledge about the genome and protein sequences, to serve as a reliable basis for theoretical prediction of structure and interaction machineries of cell molecules. We entered in epoch of genomics and proteomics, the number of various biological databases increases almost exponentially, new methods of DNA/RNA and protein sequences’ analysis appeared – but now as before we are only at a foundation of a huge pyramid of knowledge about how molecular and genetic machineries work.

For the last five years the main efforts of our research group were concentrated at study of periodicity in nucleotide and amino acid sequences and a possible relationship between periodicity and a spatial structure and functional activity of nucleotide and protein sequences [1,2,3,4,5,6]. Latent periodicity of 21 bases has been earlier revealed in many chemoreceptor genes of different bacteria [4] just in cytoplasmic region of MCP proteins, which is believed about has been formed of a few α-helixes [7]. It has been also noted for the latent periodicity of 19 amino acid residues to be a characteristic feature for NAD^+-binding sites of various proteins; periodicity of 2 amino acids is typical for transmembrane domain of various receptors [5]. A reason for the latent periodicity to arise in nucleotide sequences might be either multifold duplications of some DNA/RNA fragment accompanied by bases’ divergency or internal convergency process keeping up by adaptive structural evolution of DNAs or their encoded proteins. Probable biological sense of the latent periodicity is to serve as a scale-rule of the sequences in forming their complexes with other proteins or nucleotide sequences.

It stands no reason, that more complete study of how frequently the latent periodicity may be encountered in the proteins and nucleotide sequences, and also the study of its meaning as a characteristic feature of definite functional regions will become possible only after creation of special databases accessible for analysis for a wide circle of the molecular biologists. We have built the first version of databases on periodicity in nucleotide and amino acid sequences named as Miscellaneous Repetitive Features in Genetic (MRFGS) and Protein Sequences (MRFPS). Data constituted a content of these databases have been revealed in the result of sequences’ analysis through the GenBank database release 116 and the SwissProt database release 38. Periodic sequences in the proteins and DNAs have been revealed by an original method for searching the latent periodicity, which as it has been shown earlier reveals a wide spectrum of periodicity ranged from perfect periodicity and strongly eroded, although staying visible, to the latent periodicity [5,6]. Periodicity, which is named as latent, may be identified only on the basis of not casual bases’ (or amino acid residues’) statistics in separate sites of a period. It is not quite necessarily that only one kind of base (residue) must dominate in such sites. In a case of nucleotide sequences it may also be

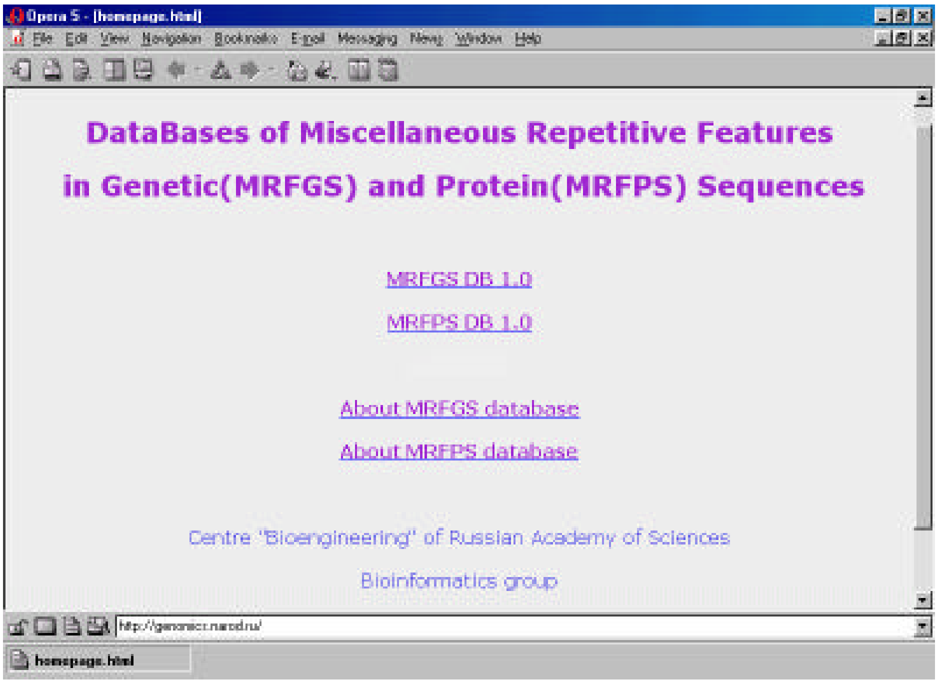
[1] Corresponding author
E-mail address: katrin22@mtu-net.ru; bioinf@gagarinclub.ru

two and even three kinds of bases, but in a case of amino acid sequences may be even more numerous groups out of 20 possible amino acids in the sites.

A statistical significance criterion for revealed latent periodicity is Z value: $Z=(I(1)-I'_m)/\sigma$, which is named as Score and has been earlier described in detail elsewhere [4,5,6]. The value Z greater than 5 for both nucleotide and amino acid sequences is correspond to accidental probability to reveal the latent periodicity of less than 10^{-6} [4,5].

In creating the first version of the databases on periodicity of protein and genetic sequences, we have chosen a critical level of Z equal to a little more than minimum of significant Z value (Z=5) for revealed periodicity: Z is equal to 13 for both nucleotide and amino acid sequences. Thereby in the first version of the MRFGS and MRFPS databases we have restricted data on periodicity by perfect periodicity, nearly perfect and strongly eroded periodicity. So, the latent periodicity has made up a small part of both databases. This restriction has been done meaningly, because we believe that one can approach to insight into biological sense of the latent periodicity only after exhaustive research will be done on relationship between strongly eroded periodicity and its functional and structure forming role in the genetic sequences and proteins.

The first version of the MRFGS and MRFPS databases is accessible via Internet by the URL: http://genomics.narod.ru. A home page of these databases is shown in figure 1. One can choose a database of interest or preliminary study the database description on corresponding page "About".


Opera 5 - [homepage.html]
File Edit View Navigation Bookmarks E-mail Messaging News Window Help
http://genomics.narod.ru/
homepage.html

DataBases of Miscellaneous Repetitive Features

in Genetic(MRFGS) and Protein(MRFPS) Sequences

MRFGS DB 1.0

MRFPS DB 1.0

About MRFGS database

About MRFPS database

Centre 'Bioengineering' of Russian Academy of Sciences

Bioinformatics group

Figure 1. A home page of the MRFGS and MRFPS databases is shown.

The MRFGS and MRFPS databases have a similar structure of records. Every periodic sequence founded in the GenBank or SwissProt databases is registered with a unique accession code "Periodicity accession number" in appropriate MRFGS or MRFPS database, at that accession number of a primary sequence from the GenBank or SwissProt where periodicity has been found is also written. Coordinates of periodicity region, its a whole sequence, a significance value Z (Score) of revealed periodicity, and a period length are also written.

At present time DNA/RNA periodic sequences with period length multiple to 3: 6, 9, 12,…,120 bases have been collected in the MRFGS database. Totally 12067 sequences having periodicity are kept in the first version of the MRFGS. A graphics of a numerical distribution of periodic sequences according to period length is shown in figure 2. As one can see, the sequences of DNA/RNA having periodicity of 6 bases (2555 sequences) represent the most numerous group. The numbers of the other periodic sequences are on average distributed uniformly with some preference for the periods of 9 bases (516 sequences), 12 bases (690 sequences), 15 bases (528 sequences), 21 bases (569 sequences) and 84 (616 sequences). A whole volume of the first version of the MRFGS database occupies a little less than 31 Mb of hard disk space.

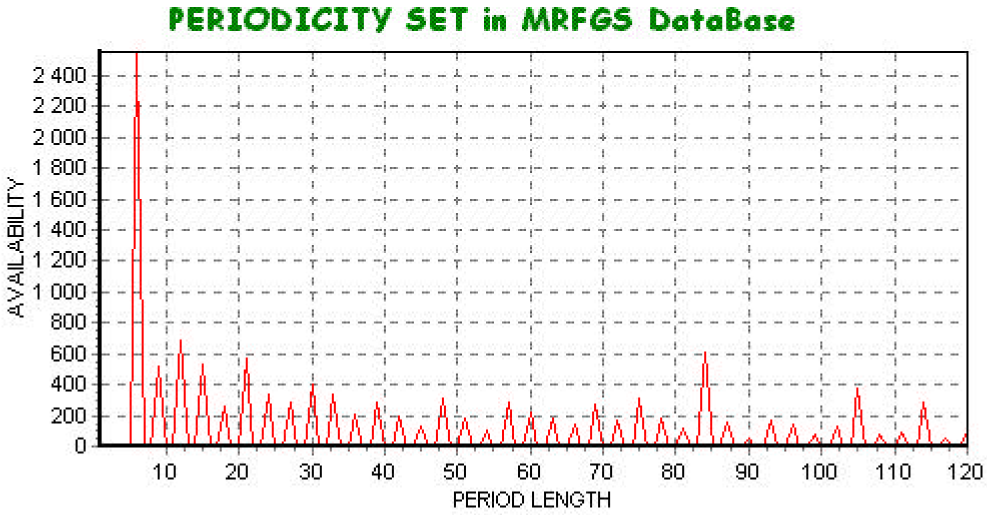


Figure 2. A distribution of the numbers of periodic sequences in the MRFGS database is shown according to the period lengths.

Totally 1072 periodic protein sequences have been collected in the first version of the MRFPS database that amounts to about 4.5% of the total protein number in the SwissProt release 38. A periodicity in proteins is presented by period lengths equal to 2, 3,…,1046 amino acid residues. In figure 3 one can see a whole spectrum of the MRFPS sequences' numbers according to period length. The most numerous protein sequences are those with period lengths of no more than 100 residues. There are only 6 cases with period lengths of more than 1000 amino acid residues. It is the most likely that periodicity of more than 100 amino acid residues is conditioned by duplications in the gene sequences of corresponding proteins.

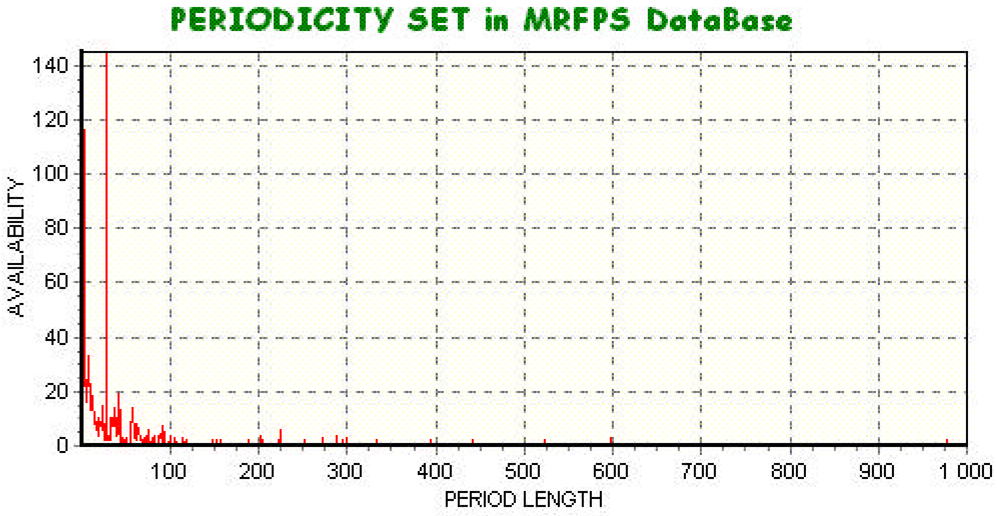


Figure 3. A distribution of the numbers of periodic sequences in the MRFPS database is shown according to the period lengths.

Figure 4 discloses the most dense interval of the common periodicity spectrum in the MRFPS database (Fig. 3). As one can see, the periodicities of 28 residues (145 sequences), 3 residues (116 sequences) and 2 amino acid residues (79 sequences) are the most numerous. A whole volume of the first version of the MRFPS database occupies about 6 Mb of hard disk space.

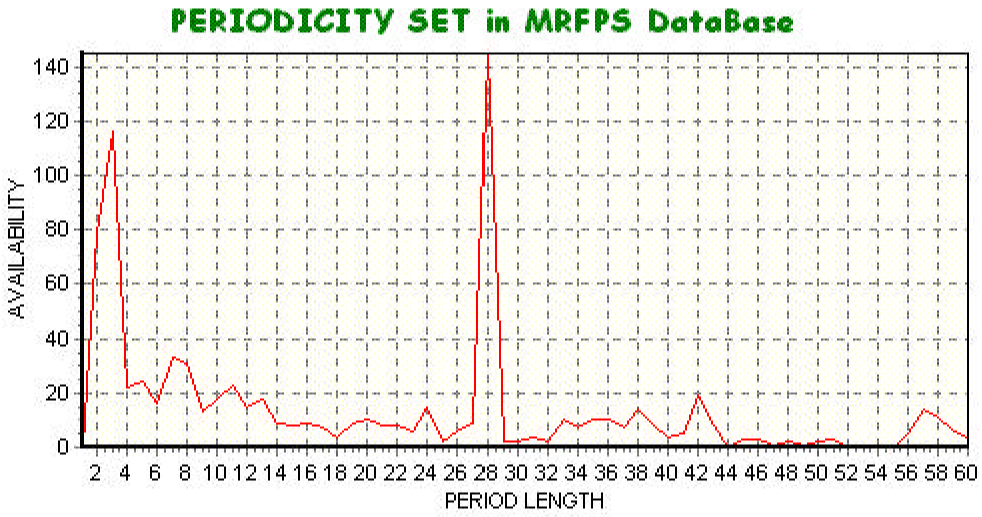
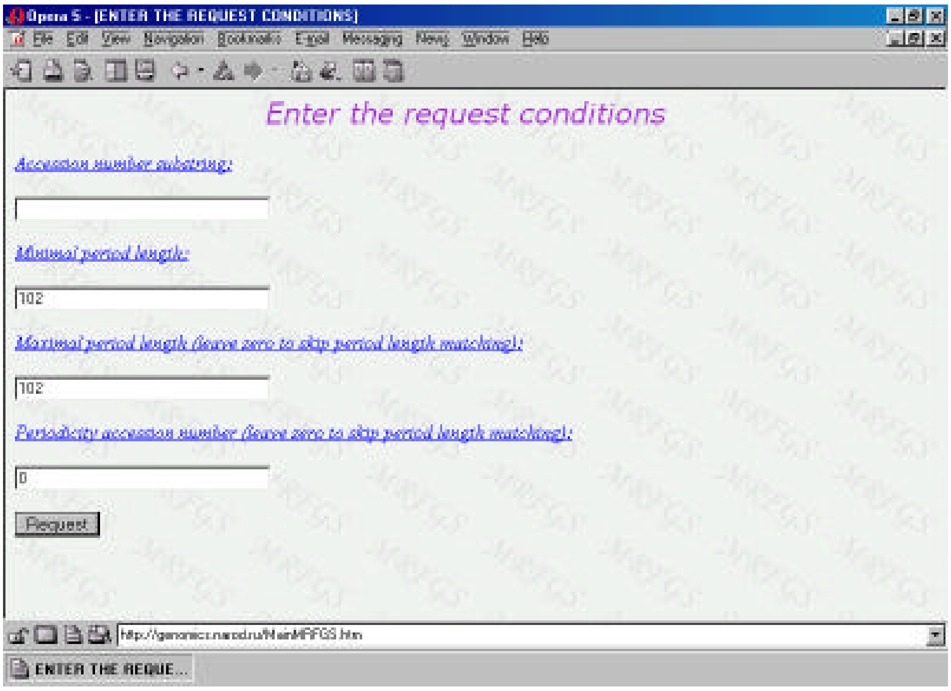


Figure 4. A distribution of the numbers of periodic sequences in the MRFPS database is shown over the limited interval of period lengths ranged from 2 to 60 amino acids.

Both the MRFGS and MRFPS databases support the same user interface for information request. A page shown in figure 5 allows shape the request by filling up the input fields. It is enough to enter an appropriate GenBank (SwissProt) accession number in the field “Accession number substring” to know if there are the periodicities in a primary sequence from the GenBank (or SwissProt), at that other fields may stay clear. One can send the request on availability of the periodic sequences with a definite period length or definite length interval. If user has already worked with the MRFGS and MRFPS databases, he (she) can send the request specifying only “Periodicity accession number” of a sequence in the appropriate database. Figure 5 shows the request on periodic sequences in the MRFGS with period length equal to 102 bases. Figure 6 demonstrates a list of the sequences received on the request.



Opera 5 - [ENTER THE REQUEST CONDITIONS]
File Edit View Navigation Bookmarks E-mail Messaging News Window Help
Request
http://genomics.narod.ru/MainMRFGS.htm
ENTER THE REQUE...

Enter the request conditions

Accession number substring:

Minimal period length:

102

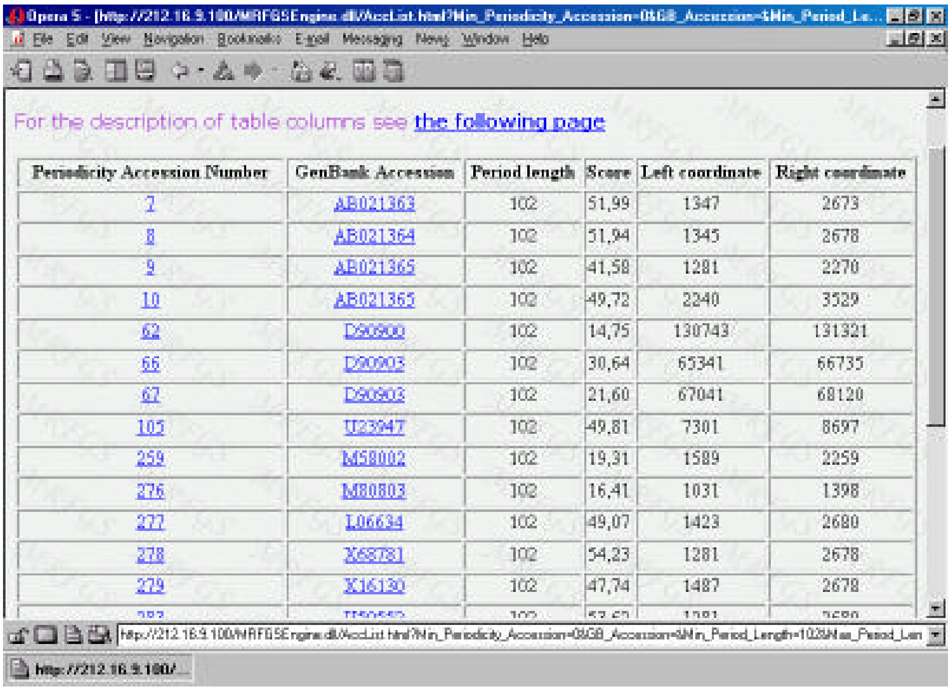
Maximal period length (leave zero to skip period length matching):

102

Periodicity accession number (leave zero to skip period length matching):

0

Figure 5. An interface of user request in the MRFGS database is the same as in the MRFPS.

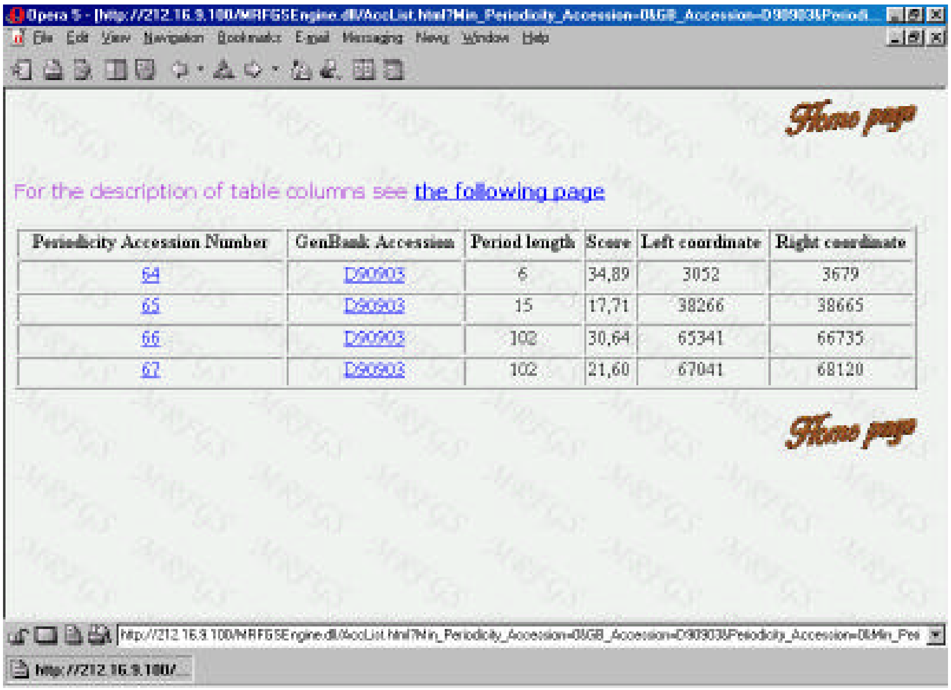

Opera 5 - [http://212.16.9.100/MRFGSEngine.dll/AccList.html?Min_Periodicity_Accession=0&GB_Accession=&Min_Period_Le...
File Edit View Navigation Bookmarks E-mail Messaging News Window Help
http://212.16.9.100/MRFGSEngine.dll/AccList.html?Min_Periodicity_Accession=0&GB_Accession=&Min_Period_Length=102&Max_Period_Len
http://212.16.9.100/...

For the description of table columns see the following page

Periodicity Accession Number	GenBank Accession	Period length	Score	Left coordinate	Right coordinate
7	AB021363	102	51,99	1347	2673
8	AB021364	102	51,94	1345	2678
9	AB021365	102	41,58	1281	2270
10	AB021365	102	49,72	2240	3529
62	D90900	102	14,75	130743	131321
66	D90903	102	30,64	65341	66735
67	D90903	102	21,60	67041	68120
105	U23947	102	49,81	7301	8697
259	M58002	102	19,31	1589	2259
276	M80803	102	16,41	1031	1398
277	L06634	102	49,07	1423	2680
278	X68781	102	54,23	1281	2678
279	X16130	102	47,74	1487	2678

Figure 6. A list of all records in the MRFGS database received on the request of periodicity of 102 bases.

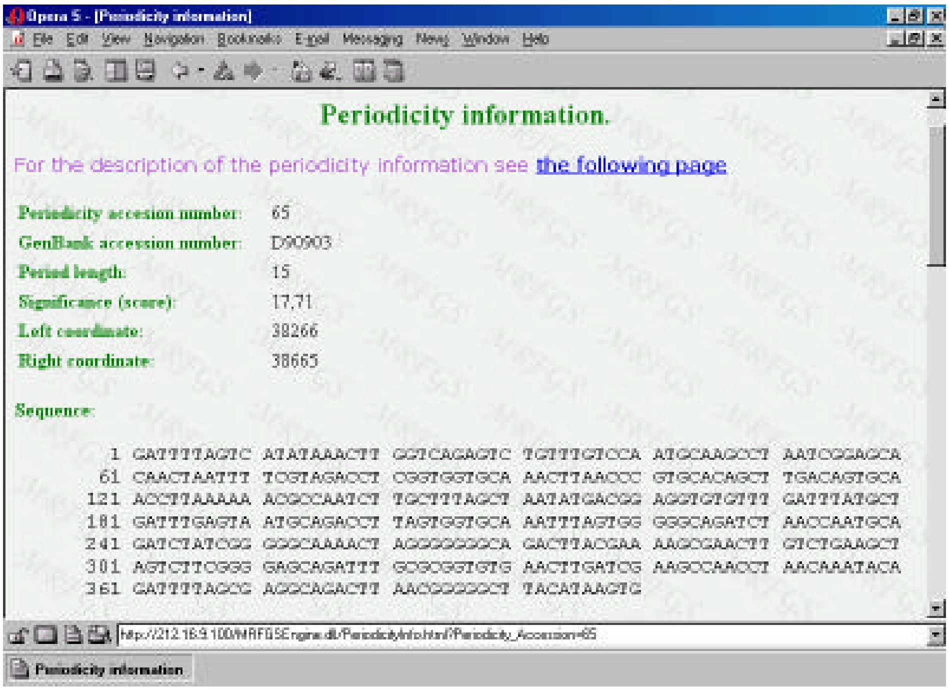
A hyperlink to the GenBank (or SwissProt) accession number leads to a list of all periodicities revealed in a primary sequence from the GenBank (SwissProt) databank. Figure 7 shows such a list for the GenBank accession number D90903. A hyperlink to the Periodicity accession number of a unique sequence in the MRFGS (MRFPS) database leads to a page "Periodicity information" (see fig. 8) which contains a whole information about the sequence. The following features are sequentially enumerated: the Periodicity accession number of the sequence; the GenBank accession number of the primary sequence in which periodicity has been found, the length of period, the value Z (Score) of significance of the revealed periodicity, the left and right coordinates of periodicity region in the primary sequence, and the whole sequence of this region.


Opera 5 - [http://212.16.9.100/MRFGSEngine.dll/AccList.html?Min_Periodicity_Accession=0&GB_Accession=D90903&Periodi...
File Edit View Navigation Bookmarks E-mail Messaging News Window Help
Home page
Home page
http://212.16.9.100/MRFGSEngine.dll/AccList.html?Min_Periodicity_Accession=0&GB_Accession=D90903&Periodicity_Accession=0&Min_Per
http://212.16.9.100/...

For the description of table columns see the following page

Periodicity Accession Number	GenBank Accession	Period length	Score	Left coordinate	Right coordinate
64	D90903	6	34,89	3052	3679
65	D90903	15	17,71	38266	38665
66	D90903	102	30,64	65341	66735
67	D90903	102	21,60	67041	68120

Figure 7. A list of all periodic regions in the primary GenBank sequence having accession number D90903.


Opera 5 - [Periodicity information]
File Edit View Navigation Bookmarks E-mail Messaging News Window Help
Periodicity information

Periodicity information.

For the description of the periodicity information see the following page

Periodicity accesion number:	65
GenBank accession number:	D90903
Period length:	15
Significance (score):	17,71
Left coordinate:	38266
Right coordinate:	38665

Sequence:

```
  1 GATTTTAGTC ATATAAACTT GGTCAGAGTC TGTTTGTCCA ATGCAAGCCT AATCGGAGCA
 61 CAACTAATTT TCGTAGACCT CGGTGGTGCA AACTTAACCC GTGCACAGCT TGACAGTGCA
121 ACCTTAAAAA ACGCCAATCT TGCTTTAGCT AATATGACGG AGGTGTGTTT GATTTATGCT
181 GATTTGAGTA ATGCAGACCT TAGTGGTGCA AATTTAGTGG GGGCAGATCT AACCAATGCA
241 GATCTATCGG GGGCAAAACT AGGGGGGGCA GACTTACGAA AAGCGAACTT GTCTGAAGCT
301 AGTCTTCGGG GAGCAGATTT GCGCGGTGTG AACTTGATCG AAGCCAACCT AACAAATACA
361 GATTTTAGCG AGGCAGACTT AACGGGGGCT TACATAAGTG
```

http://212.16.9.100/MRFGSEngine.dll/PeriodicityInfo.html?Periodicity_Accession=65

Figure 8. A beginning of page "Periodicity information" holding a whole information about periodic sequence in the MRFGS database.

Besides the information shown in figure 8, page "Periodicity information" contains a matrix of A, T, C, G base quantities for every period site of a revealed periodic sequence. On the base of this matrix "Chi-squared graphics" is built – a graphics showing how a set of bases' quantities for every period site varies from a casual one expected for this periodic sequence according to χ^2 criterion with the number of freedom degrees equal to the period length. Down to this page a spectrum of all possible periodicities in a revealed region is built (see fig.9 for details). The

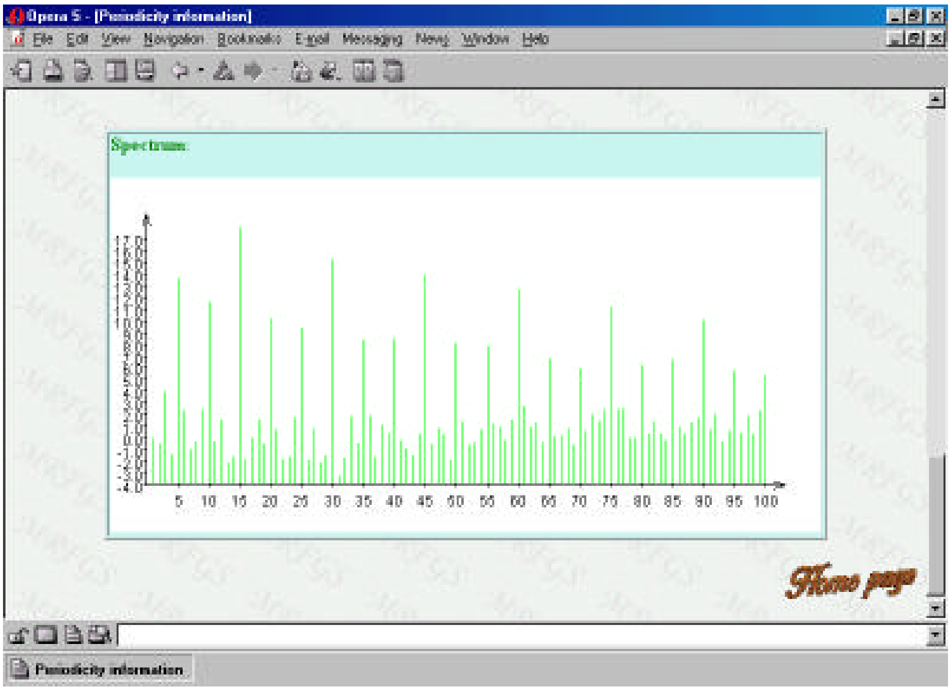
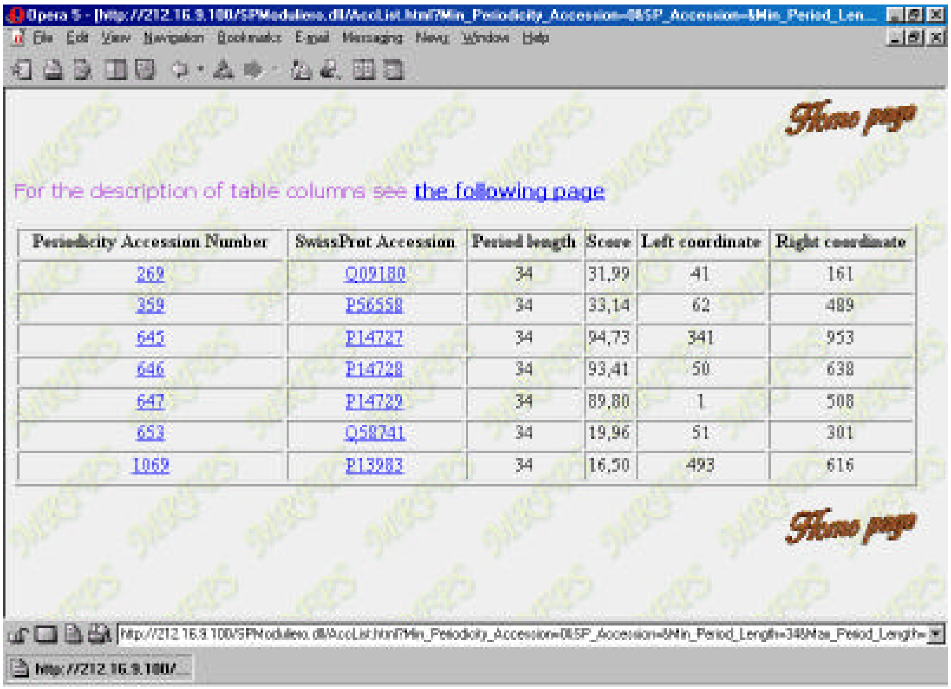


Figure 9. A spectrum of Z values (statistical significance measure) for all tested period lengths in a periodic sequence with accession code equal to 65 in the MRFGS database.

spectrum shows the value Z according to analyzed period length. The greatest Z value and its corresponding period length in the spectrum may not be the same with a Score and a period length pointed at the beginning of “Periodicity information” page. This disagreement arises because the spectrum allows only select a few large Z values and their corresponded period lengths for further analysis. Final conclusion about a certain period length and a statistical significance value of revealed periodicity (Score) is accepted after testing the every selected period length by a Monte Carlo method.

As it has been already mentioned above the structure of records (entries) and also the request procedure in the MRFGS and MRFPS databases are nearly the same, so we will be further restricted ourselves by only short notes on the MRFPS database. Figure 10 shows the information received upon the request of periodicity of 34 amino acid residues in the MRFPS database which may be considered as mating to the MRFGS request on periodicity of 102 bases (see fig.6).


Opera 5 - [http://212.16.9.100/SPModuleeo.dll/AccList.html?Min_Periodicity_Accession=0&SP_Accession=&Min_Period_Len...
File Edit View Navigation Bookmarks E-mail Messaging News Window Help
Home page

For the description of table columns see the following page

Periodicity Accession Number	SwissProt Accession	Period length	Score	Left coordinate	Right coordinate
269	Q09180	34	31,99	41	161
359	P56558	34	33,14	62	489
645	P14727	34	94,73	341	953
646	P14728	34	93,41	50	638
647	P14729	34	89,80	1	508
653	Q58741	34	19,96	51	301
1069	P13983	34	16,50	493	616

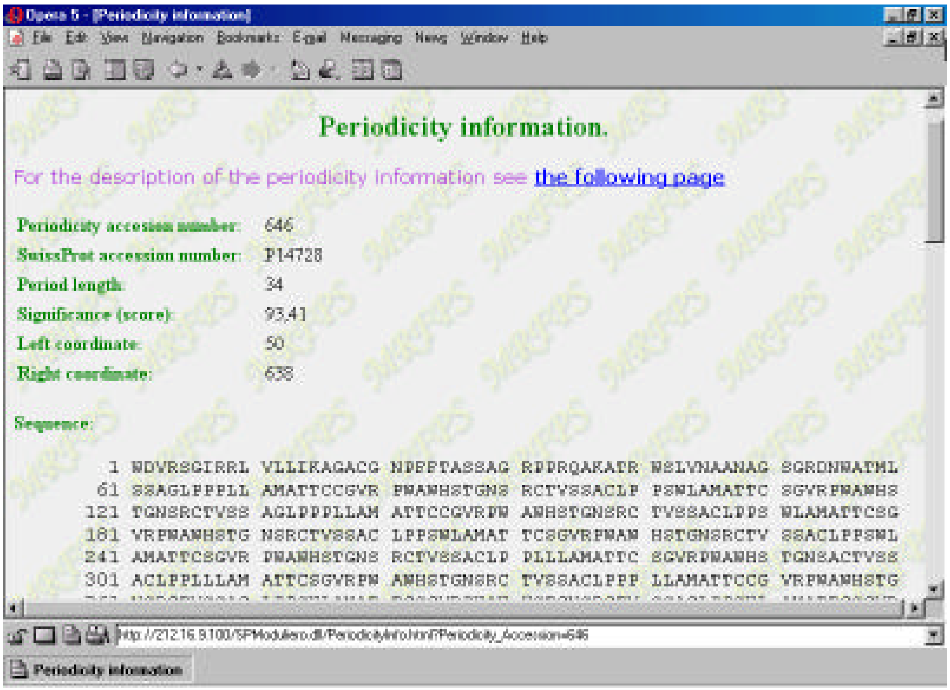
Home page
http://212.16.9.100/SPModuleeo.dll/AccList.html?Min_Periodicity_Accession=0&SP_Accession=&Min_Period_Length=34&Max_Period_Length=
http://212.16.9.100/...

Figure. 10. A list of periodic sequences received on the request of 34 amino acid residues’ periodicity in the MRFPS database.

In figure 11 the “Periodicity information” page is shown for a sequence having Periodicity accession number in the MRFPS database equal to 646 – for one of the sequences with periodicity of 34 amino acid residues. Amino acid sequence of a periodicity region is drawn in one letter code. Down to the page “Periodicity information” a matrix of quantitive distribution of amino acid residues over a period sites is situated (not seen in figure 11), that helps to present what kind of periodicity exists in a protein sequence: perfect, eroded, strongly eroded or latent. A spectrum of Z values for all probable period lengths in the sequence follows the matrix.

For all protein sequences having periodicity of 34 residues in the MRFPS database, excepting one (SwissProt accession P13983), there are corresponding gene sequences in the MRFGS database. Even for such a little sample of couples “periodic gene – periodic protein” one can see that nearly perfect periodicity in gene is correspond to nearly perfect periodicity in protein (MRFGS 4569 – MRFPS 269, MRFGS 279 – MRFPS 645). Strongly eroded periodicity in protein is due to strongly eroded periodicity of gene (MRFGS 9884 – MRFPS 369). At last the latent periodicity in gene is transferred into the strongest eroded periodicity in protein (MRFGS 522 – MRFPS 653).

So, a couple of the MRFGS – MRFPS databases will facilitate an understanding of the limits of possible relationship between perfect, eroded and latent periodicity in the genes and proteins.


Opera 5 - [Periodicity information]
File Edit View Navigation Bookmarks E-mail Messaging News Window Help
Periodicity information

Periodicity information.

For the description of the periodicity information see the following page

Periodicity accession number: 646
SwissProt accession number: P14728
Period length: 34
Significance (score): 93.41
Left coordinate: 50
Right coordinate: 638

Sequence:

```
  1 MDVRSGIRRL VLLIKAGACG NPPPTASSAG RPPRQAKATR WSLVNAANAG SGRDNWATML
 61 SSAGLPPPLL AMATTCCGVR PWANHSTGNS RCTVSSACLP PSWLAMATTC SGVRPWANHS
121 TGNSRCTVSS AGLPPPLLAM ATTCCGVRPW ANHSTGNSRC TVSSACLPPS WLAMATTCSG
181 VRPWANHSTG NSRCTVSSAC LPPSWLAMAT TCSGVRPWAN HSTGNSRCTV SSACLPPSWL
241 AMATTCSGVR PWANHSTGNS RCTVSSACLP PLLLAMATTC SGVRPWANHS TGNSACTVSS
301 ACLPPLLLAM ATTCSGVRPW ANHSTGNSRC TVSSACLPPP LLAMATTCCG VRPWANHSTG
```

http://212.16.9.100/SPModulero.dll/PeriodicityInfo.html?Periodicity_Accession=646

Figure.11. A beginning of page “Periodicity information” for a record with Periodicity accession number equal to 646 in the MRFPS database.

Besides that both databases could be useful for study the characteristic periodicities for the known elements of spatial structure and functional sites in the proteins.

We have already prepared the second version of the MRFGS and MRFPS databases which includes all findings of periodicity in amino acid and nucleotide sequences with Z value greater than 5.0. The number of periodic nucleotide sequences in the GenBank release 116 is equal to a little more than 600000, and the number of periodic amino acid sequences in the SwissProt release 38 is equal to more than 20000. The questions on data access in the second version of the MRFGS and MRFPS databases should be send in address: katrin22@mtu-net.ru.

The MRFGS-MRFPS databases project has been funded by RFBR grant 99-07-90154.

REFERENCES.

1. Korotkov E.V. and Korotkova M.A. 1995, "DNA regions with latent periodicity in some human clones", DNA Seq., 5: 353-358.
2. Korotkov E.V. and Korotkova M.A. 1996, "Enlarged similarity of nucleic acid sequences", DNA Res., 3: 157-164.
3. Korotkov E.V., Korotkova M.A., Tulko J.S. 1997, "Latent sequence periodicity of some oncogenes and DNA-binding protein genes", CABIOS, 13: 37-44.
4. Chaley M.B., Korotkov E.V., Skryabin K.G. 1999, "Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples", DNA Res., 6: 153-163.
5. Korotkova M.A., Korotkov E.V., RudenkoV.M. 1999, "Latent periodicity of protein Sequences", J. Mol. Model., 5: 103-115.
6. Korotkov E.V., Korotkova M.A., Rudenko V.M., Skryabin K.G. 1999, "Latent Periodicity Regions in Amino Acid Sequences", Molecular Biology (Russian), 33(4): 688-695.
7. Mowbray S.L., Sandren M.O.J. 1998, “Chemotaxis Receptors: A progress report on structure and function”, J. Struct. Biol., 124: 257-275.