Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W417–W422***
doi:10.1093/nar/gki459

ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters

Oleg V. Vishnevsky[1,2,]* and Nikolay A. Kolchanov[1,2]

[1]Institute of Cytology and Genetics, SB RAS, Lavrentyev Avenue, 10, Novosibirsk, 630090, Russia and [2]Novosibirsk State University, Pirogova Street, 2, Novosibirsk, 630090, Russia

Received February 14, 2005; Revised and Accepted April 13, 2005

ABSTRACT

Reliable recognition of the promoters in eukaryotic genomes remains an open issue. This is largely owing to the poor understanding of the features of the structural–functional organization of the eukaryotic promoters essential for their function and recognition. However, it was demonstrated that detection of ensembles of regulatory signals characteristic of specific promoter groups increases the accuracy of promoter recognition and prediction of specific expression features of the queried genes. The ARGO_Motifs package was developed for the detection of sets of region-specific degenerate oligonucleotide motifs in the regulatory regions of the eukaryotic genes. The ARGO_Viewer package was developed for the recognition of tissue-specific gene promoters based on the presence and distribution of oligonucleotide motifs obtained by the ARGO_Motifs program. Analysis and recognition of tissue-specific promoters in five gene samples demonstrated high quality of promoter recognition. The public version of the ARGO system is available at http://wwwmgs2.bionet.nsc.ru/argo/ and http://emj-pc.ics.uci.edu/argo/.

INTRODUCTION

The assembly of the basal transcription complex and the tissue-specific and stage-specific features of eukaryotic gene transcription depend on the context and structural organization of the promoter core and the presence of transcription factor binding sites (TFBSs) in the 5′ regulatory region of the gene (1,2).

Most approaches to promoter recognition use information about the distribution features of potential TFBSs (3,4) and of short oligonucleotides along promoters (5–10), derived from the analysis of Internet-accessible databases (11,12). It is suggested that oligonucleotide composition characteristics of various promoter regions may be determined not only by the presence of TFBSs in these regions but also by certain context-dependent specific features of promoter DNA local conformation (13,14).

To reveal such specific oligonucleotide signals, there are methods based on the detection of short conserved motifs that are significantly overrepresented in a sample of promoter sequences compared with the number expected by chance. The methods include analysis of the frequencies of *l*-mers (*l*-letter substrings) (15,16), suffix trees (17,18), finding of the largest cliques in the graph induced by the edit distance between the *l*-mers (19), local multiple alignment approaches using a greedy algorithm (20), expectation–maximization algorithm (21,22) and stochastic sampling strategy (23).

Genomic sequencing of an increasing number of eukaryotes (24,25) encouraged the development of methods that use multiple alignment of genomic sequences with expressed sequence tags and mRNA sequences during promoter recognition (26–30) and of those utilizing the information about the localization of promoters in orthologous genes (29,31). In addition, consensus-based methods are applied, which are based on the combined use of several independent methods for recognizing promoters (32). However, despite the diversity of the approaches, the reliable recognition of promoters in eukaryotic genomes remains an open issue (33,34). A great hindrance to the development of accurate methods for promoter recognition is the tremendous diversity of their structural–functional organization (11). This makes the search for general context regularities that would serve as the background for recognition of promoters difficult.

Recent data indicate a certain similarity in the promoter organization in genes with similar expression patterns (e.g. promoters of genes expressed in particular tissues). This manifests itself as the presence of similar TFBS sets in the promoters of such genes (1,2,35,36).

*To whom correspondence should be addressed. Tel: +7 3832 333119; Fax: +7 3832 331278; Email: oleg@bionet.nsc.ru

The detection of such ensembles of regulatory signals characteristic of specific groups of promoters increases promoter recognition accuracy and prediction of specific expression patterns of the analyzed genes.

The ARGO_Motifs package was developed for analysis of functional nucleotide sequences (37). It allows the recognition of oligonucleotide motifs with the following properties: (i) degeneracy, i.e. the use of the extended IUPAC code (A,T, G,C, R = G/A, Y = T/C, M = A/C, K = G/T, W = A/T, S = G/C, B = T/G/C, V = A/G/C, H = A/T/C, D = A/T/G, N = A/T/G/C); (ii) region-specificity, i.e. the preferential occurrence in a certain region of a functional sequence; (iii) quasi-invariance, i.e. the occurrence in certain sequence subgroups only; (iv) contrast, i.e. much more frequent occurrence in functional than random sequences.

The ARGO_Viewer package was developed (37) for the recognition of tissue-specific gene promoters based on the presence and distribution of oligonucleotide motifs obtained by the ARGO_Motifs program.

Five samples of tissue-specific promoters from the Transcription Regulatory Regions Database of regulatory sequences (http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4) in the [−300; +100] region relative to the transcription start were studied using the ARGO_Motifs program (37). The resulting sets of motifs were used to construct methods for recognizing tissue-specific promoters by the ARGO_Viewer program. It was demonstrated that one overprediction error occurred per 100 000 bp with a false negative error in the 4–8% range for four of the five promoter samples.

METHODS AND ALGORITHMS

ARGO_Motifs description

The search for degenerate motifs in a sample of functional sequences using the ARGO_Motifs program (Figure 1) is detailed in (37) and is implemented by the grouping of similar perfect oligonucleotides from the oligonucleotide vocabularies corresponding to different sequences. Each oligonucleotide of the sequence vocabularies is considered, and the group for each oligonucleotide is formed. A group consists of oligonucleotides belonging to the vocabularies of other sequences differing from it by not more than R positions ($R < r_0$, where r_0 is the threshold similarity value). Then, the consensus in an extended IUPAC code is constructed for each oligonucleotide group using an iteration procedure. Each position of the consensus is occupied by the most significant of the 15 possible letters and their significance is estimated independently of each other using the binomial criterion. The obtained oligonucleotide motifs are regarded as significant if they meet the requirements in Equation 1a–c. The significant motif that has the smallest probability to occur by chance is deposited in the databank, while all the perfect oligonucleotides it describes are removed from the vocabularies of the oligonucleotide sequences. The procedure for the detection of the motif ranking next in significance is applied in the same way to the modified vocabularies. The procedure is iterated until the detection of common degenerate motifs that satisfy the condition in Equation 1a–c is still feasible.

The degenerate oligonucleotide motif obtained using this procedure is considered significant if it meets the following criteria:

$$F > f_0 \qquad \textbf{1a}$$

$$P(n,\ N) < p_0 \qquad \textbf{1b}$$

$$Q < q_0. \qquad \textbf{1c}$$

Here, F is the proportion of promoters containing the motif in the window under analysis; f_0 is the threshold level of the motif occurrence in the promoter sample; $P(n, N)$ is the probability of the accidental occurrence of the motif in the analyzed window in not less than n sequences of N; p_0 is the threshold probability level (see the estimation method below); Q denotes the proportion of sequences of the negative sample containing the motif; q_0 is the threshold level of the motif occurrence in the negative sample. A set of 1000 randomly generated sequences of the length L is used as the negative sample. Thus, an oligonucleotide motif is accepted as significant if (i) it occurs frequently in a promoter sample, (ii) infrequently in a sample of random sequences and (iii) its occurrence probability by chance in a sample of promoter sequences is significantly low.

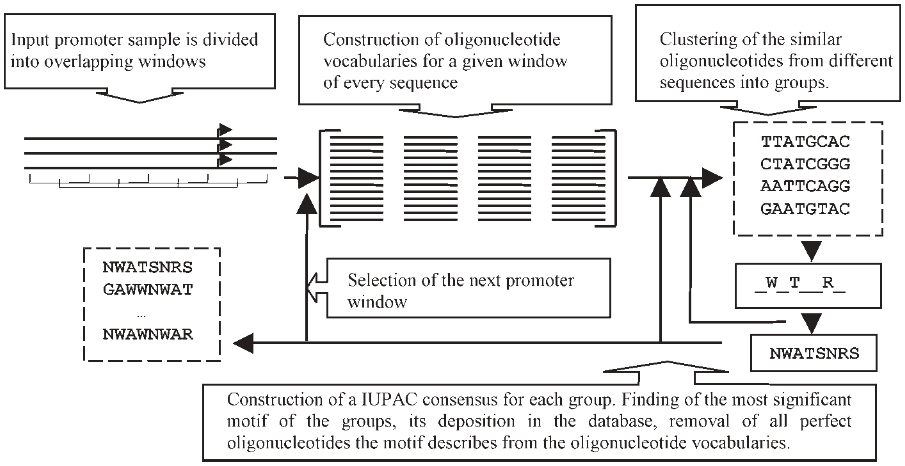

Input promoter sample is divided into overlapping windows
Construction of oligonucleotide vocabularies for a given window of every sequence
Clustering of the similar oligonucleotides from different sequences into groups.
TTATGCAC
CTATCGGG
AATTCAGG
GAATGTAC
_W_T__R_
NWATSNRS
Selection of the next promoter window
NWATSNRS
GAWWNWAT
...
NWAWNWAR
Construction of a IUPAC consensus for each group. Finding of the most significant motif of the groups, its deposition in the database, removal of all perfect oligonucleotides the motif describes from the oligonucleotide vocabularies.

Figure 1. Layout of the algorithm for the recognition of degenerate oligonucleotide motifs in a promoter sample.

The probability $P(n, N)$ is calculated as follows. Let us consider the oligonucleotide motif $M = m_1, m_2, \ldots, m_l$ of the length l in the extended 15 single letter-based IUPAC code. The occurrence probability of the motif at a certain position of a sequence of length L is estimated as:

$$P(M) = \prod_{i=1}^{l} P_i,$$

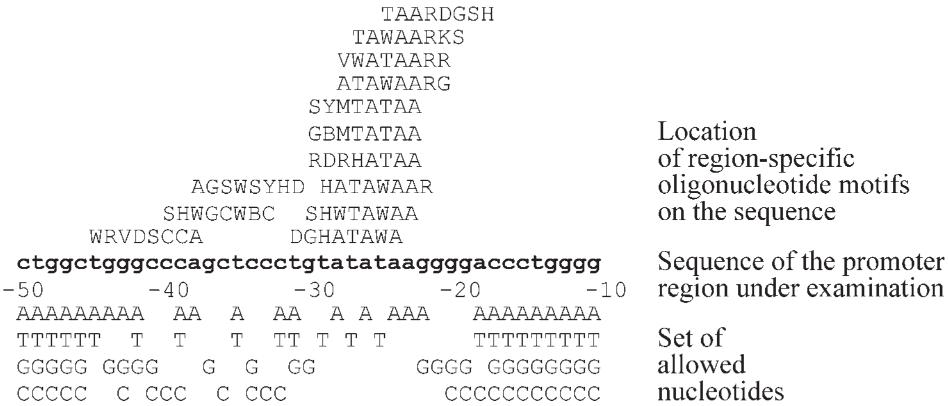
where P_i is the frequency of the letter m_i calculated from the mononucleotide composition of promoters.

The binomial occurrence probability of the motif M in $\geqslant n$ sequences of N, $P(n, N)$ is

$$P(n,N) = \sum_{i=n}^{N} C_N^i P^i (1-P)^{N-i}, \quad \text{where } P = 1 - \mathrm{e}^{-(L-l+1)\times P(M)}.$$

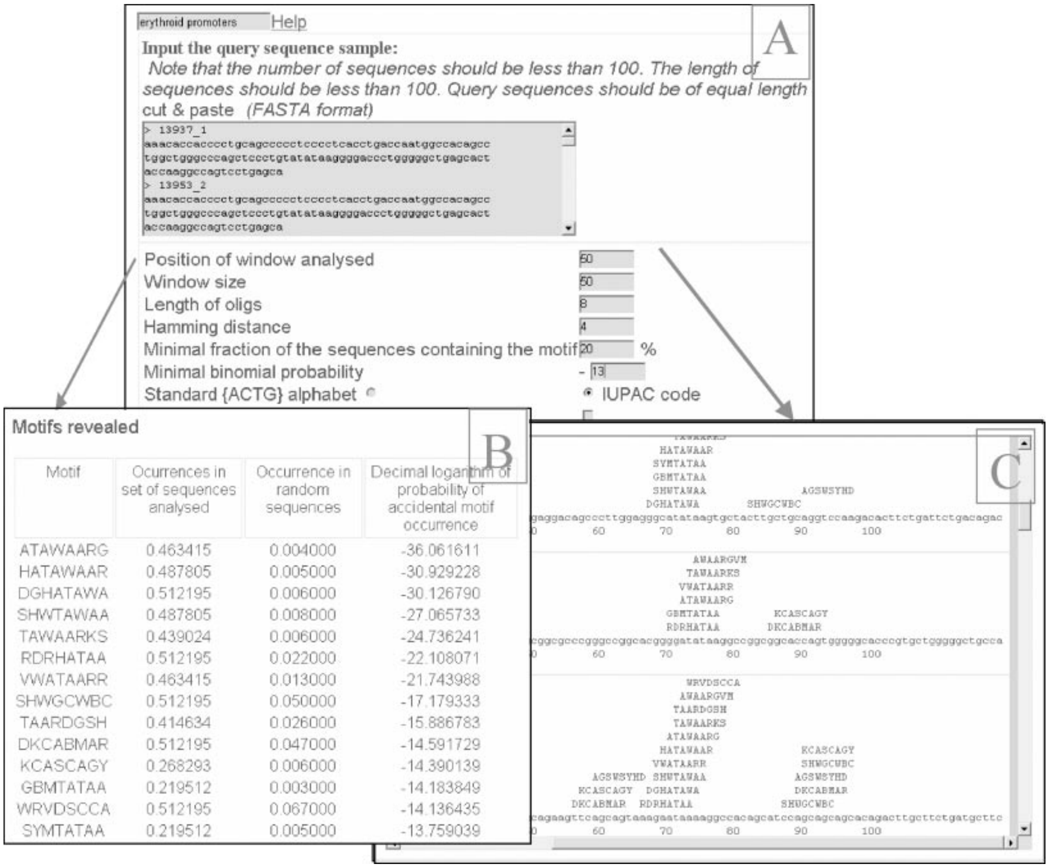
The probability $P(n, N)$ calculated in such a way is used to assess the significance of the motif by the significance criteria (Equation 1b).

Then, the oligonucleotides contained in this motif are removed from the oligonucleotide vocabularies of all the



```
                          TAARDGSH
                        TAWAARKS
                       VWATAARR
                       ATAWAARG
                     SYMTATAA
                     GBMTATAA
                     RDRHATAA
             AGSWSYHD HATAWAAR
           SHWGCWBC  SHWTAWAA
      WRVDSCCA      DGHATAWA
 ctggctgggcccagctccctgtatataaggggaccctgggg
-50       -40       -30       -20       -10
 AAAAAAAAA  AA  A  AA  A A AAA   AAAAAAAAA
 TTTTTT  T  T   T  TT T T T      TTTTTTTTT
 GGGGG  GGGG  G  G  GG       GGGG GGGGGGGG
 CCCCC  C CCC  C CCC           CCCCCCCCCCC
```

Figure 2. Example of determination of the set of permissible nucleotides for each position of the [−50; −10] region of an erythroid-specific promoter.



Input the query sequence sample:

Note that the number of sequences should be less than 100. The length of sequences should be less than 100. Query sequences should be of equal length

cut & paste *(FASTA format)*

```
> 13937_1
aaacaccaccccctgcagcccccctcccctcacctgaccaatggccacagcc
tggctgggcccagctccctgtatataaggggaccctgggggctgagcact
accaaggccagtcctgagca
> 13953_2
aaacaccaccccctgcagcccccctcccctcacctgaccaatggccacagcc
tggctgggcccagctccctgtatataaggggaccctgggggctgagcact
accaaggccagtcctgagca
```

Position of window analysed 50

Window size 50

Length of oligs 8

Hamming distance 4

Minimal fraction of the sequences containing the motif 20 %

Minimal binomial probability - 13

Standard {ACTG} alphabet ○ ● IUPAC code

□

Motifs revealed

Motif	Ocurrences in set of sequences analysed	Occurrence in random sequences	Decimal logarithm of probability of accidental motif occurrence
ATAWAARG	0.463415	0.004000	-36.061611
HATAWAAR	0.487805	0.005000	-30.929228
DGHATAWA	0.512195	0.006000	-30.126790
SHWTAWAA	0.487805	0.008000	-27.065733
TAWAARKS	0.439024	0.006000	-24.736241
RDRHATAA	0.512195	0.022000	-22.108071
VWATAARR	0.463415	0.013000	-21.743988
SHWGCWBC	0.512195	0.050000	-17.179333
TAARDGSH	0.414634	0.026000	-15.886783
DKCABMAR	0.512195	0.047000	-14.591729
KCASCAGY	0.268293	0.006000	-14.390139
GBMTATAA	0.219512	0.003000	-14.183849
WRVDSCCA	0.512195	0.067000	-14.136435
SYMTATAA	0.219512	0.005000	-13.759039

```
                 HATAWAAR
                SYMTATAA
                GBMTATAA
                SHWTAWAA              AGSWSYHD
               DGHATAWA       SHWGCWBC
gaggacagcccttggagggcatataagtgctacttgctgcaggtccaagacacttctgattctgacagac
0        60        70        80        90        100

                      AWAARGVM
                     TAWAARKS
                    VWATAARR
                    ATAWAARG
                  GBMTATAA        KCASCAGY
                  RDRHATAA       DKCABMAR
cggcgcccgggccggcacggggatataaggccggcggcaccagtgggggcacccgtgctggggctgcca
0        60        70        80        90        100

                     WRVDSCCA
                    AWAARGVM
                   TAARDGSH
                   TAWAARKS
                  ATAWAARG
                 HATAWAAR             KCASCAGY
                VWATAARR              SHWGCWBC
        AGSWSYHD SHWTAWAA            AGSWSYHD
      KCASCAGY  DGHATAWA             DKCABMAR
     DKCABMAR  RDRHATAA            SHWGCWBC
cagaagttcagcagtaaagaataaaaggccacagcatccagcagcagcacagacttgcttctgatgcttc
0        60        70        80        90        100
```

Figure 3. Example of ARGO_Motifs input and output windows. (**A**) Input window. The region [−50; +1] of promoters of erythroid-specific genes is analyzed. (**B**) A table containing the motifs detected and their characteristics. (**C**) A distribution pattern of the motifs found.

sequences. The procedure of clustering and construction of the new motif is repeated for the current oligonucleotide vocabularies thus constructed until it is possible to construct new motifs meeting the criteria of significance (Equation 1a–c) using the continuously decreasing vocabularies.

ARGO_Viewer description

Tissue-specific promoters are recognized by the ARGO_ Viewer program [detailed in (37)], in a scanning window sliding with a specified step along the genomic sequence analyzed. In every window, the corresponding region-specific oligonucleotide motifs obtained by the ARGO_Motifs (Figure 2) are detected. Then, the similarity between the distributions of the motifs found in this window and in promoters of the groups studied is assessed. As a measure of similarity between the jth promoter and the sequence studied, the value $P_j = -\sum_{k=1}^{L} \log p_k/L$ is used, where L is the size of the window analyzed and p_k is the product of nucleotide frequencies consistent with the motifs covering the kth position (Figure 2).

The greater the value of P_j, the lower is the probability of chance occurrence of the motif set characteristic of the jth promoter in the sequence.

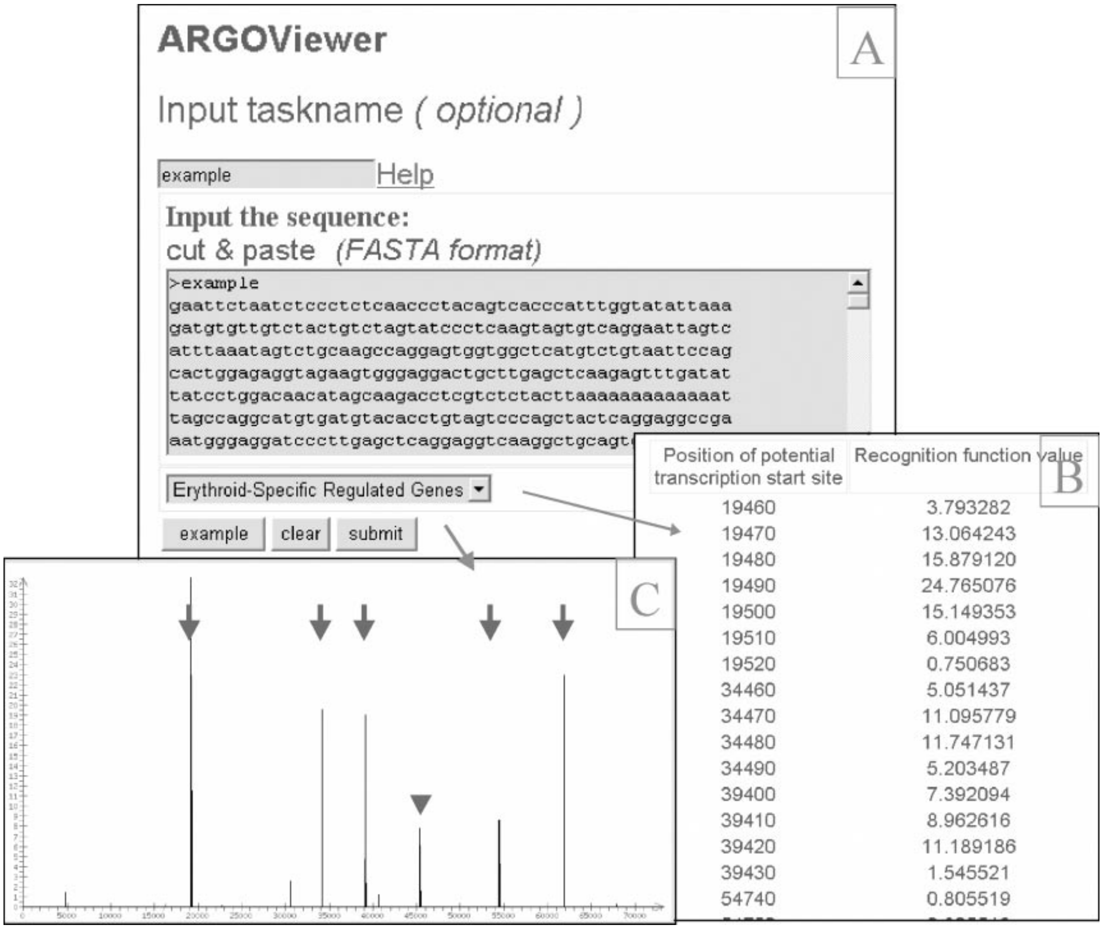
Thus, the promoter displaying the maximum value of the similarity function is found. If this value exceeds a certain threshold value, it is thought that the promoter of the considered group is identified in the window.

IMPLEMENTATION

Description of the web-interface of the ARGO_Motifs program

The public version of ARGO_Motifs (Figure 3A) is available at http://wwwmgs2.bionet.nsc.ru/argo/ and http://emj-pc.ics.uci.edu/argo/.

The user can paste a set of analyzed sequences of equal lengths in FASTA format via the sequence input. All the parameters needed for analysis are specified in the lower part of the window. The program was designed to search for region-specific motifs. Therefore, once the sample of DNA sequences is input, the user can analyze consecutively the regions of interest. In addition, the length of the motifs detected, the Hamming's distance and the degree of similarity between the perfect oligonucleotides clustered in a motif are indicated. The user can search for both perfect oligonucleotide motifs in the 4 single letter-based (A, T, G and C) code and



ARGOViewer

Input taskname *(optional)*

example

Input the sequence:

cut & paste *(FASTA format)*

```
>example
gaattctaatctccctctcaaccctacagtcacccatttggtatattaaa
gatgtgttgtctactgtctagtatccctcaagtagtgtcaggaattagtc
atttaaatagtctgcaagccaggagtggtggctcatgtctgtaattccag
cactggagaggtagaagtgggaggactgcttgagctcaagagtttgatat
tatcctggacaacatagcaagacctcgtctctacttaaaaaaaaaaaaat
tagccaggcatgtgatgtacacctgtagtcccagctactcaggaggccga
aatgggaggatcccttgagctcaggaggtcaaggctgcagt
```

Position of potential transcription start site	Recognition function value
19460	3.793282
19470	13.064243
19480	15.879120
19490	24.765076
19500	15.149353
19510	6.004993
19520	0.750683
34460	5.051437
34470	11.095779
34480	11.747131
34490	5.203487
39400	7.392094
39410	8.962616
39420	11.189186
39430	1.545521
54740	0.805519

Figure 4. Profile of the promoter recognition function for the sequence of the human β-globin gene clusters (EMBL ID: HSHBB). Values of the recognition function (ordinate) are plotted versus positions of the sequence (abscissa). Arrows indicate the positions of the transcription starts of the genes in this cluster. The triangle shows the position of the 5′-terminal region of the pseudogene corresponding to the transcription start point.

degenerate motifs in the 15 single letter-based IUPAC code. The program allows the motifs meeting the significance criteria to be found in both DNA strands. It is possible to specify for the motifs detected both the boundary value of binomial probability of their random occurrence in the examined sample and the threshold occurrence rate (%) of a motif, i.e. the fraction of analyzed sequences containing the motif.

The results of the sequence analysis are displayed as a table containing the motifs detected and their characteristics. As an example, Figure 3B shows the motifs found in the [−50; +1] region of promoters of erythroid-specific genes. The motifs of length $l = 8$ meeting the parameters below of Equation 1a–c were considered significant: $P(n, N) < 10^{-13}$, $f_0 = 20\%$ and $q_0 = 100\%$. As an example, let us consider the first oligonucleotide listed in Figure 3B: ATAWAARG = (A)(T)(A)(A/T)(A)(A)(A/G)(G), found in the [−50: +1] region relative to the transcription start. This motif was found in 19 of 41 promoters (46%), exceeding the threshold (20%) by ~2-fold. The random occurrence probability of this motif in 19 or more of the 41 promoters is 10^{-36}. In the negative sample, this motif occurred in the queried region only in 4 random sequences of 1000 (0.4%). Hence, this motif meets the significance criteria (Equation 1a–c).

In addition to the table output mode, the user can get a distribution pattern of the motifs found in the selected window of the sample analyzed (Figure 3C). This representation may be useful for the detection of ensembles of mutually present motifs and subgrouping of the sequences of the total sample.

Web interface of the ARGO_Viewer program

The ARGO_Viewer package was developed for the recognition of tissue-specific gene promoters on the basis of the presence and distribution of oligonucleotide motifs obtained by the ARGO_Motifs program. The public version of the ARGO_Viewer (Figure 4A) is available at http://wwwmgs2.bionet.nsc.ru/argo/ and http://emj-pc.ics.uci.edu/argo/.

The user can paste the genomic sequence analyzed in FASTA format into the sequence input box. The class of promoters to be searched for is specified at the bottom of the window. The program provides the search for promoters in both the direct and the complementary DNA strands. Furthermore, two modes of output recognition results are provided. In the case of text mode, the user gets a list of positions of potential transcription starts. In graphic mode, the program constructs the profile of recognition function. The program implementation is illustrated (Figure 4) using the example of the human β-globin region (ID HSHBB), of 73 308 bp in length, mapped on chromosome 11. This sequence contains five experimentally detected transcription start sites at positions 19 487, 34 478, 39 414, 54 740 and 62 137 together with the promoter region of a pseudogene in the vicinity of position 45 557.

Predicted positions of the transcription starts in five genes of this cluster differed from the real starts by not more than 20 bp. Therefore, the proposed procedure provides high efficiency of promoter recognition.

ACKNOWLEDGEMENTS

The author is grateful to Dr A. S. Poplavsky for interface design. This work was supported by the Russian Foundation for Basic Research (grants nos 03-04-48555-a, 03-07-90181-b, 03-04-48469-a, 03-04-48829, 03-07-96833); grant PCB RAS (no. 10.4); NATO (grant LST.CLG 979815); Ministry of Education of Russian Federation program 'Development of science potential of high school' (grant No. 8319); the Federal Agency for Science and Innovation Studies and Projects in priority areas of development in science and technology. Funding to pay the Open Access publication charges for this article was provided by Institute of Cytology and Genetics, SB RAS.

Conflict of interest statement. None declared.

REFERENCES

1. Ignatieva,E.V., Merkulova,T.I., Vishnevskii,O.V. and Kel,A.E. (1997) Transcription regulation of lipid metabolism genes as described in the TRRD database. *Mol. Biol. (Mosk.)*, **31**, 684–700.
2. Krivan,W. and Wasserman,W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res*., **11**, 1559–1566.
3. Prestrige,D.S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol*., **249**, 923–932.
4. Kondrakhin,Y.V., Kel,A.E., Kolchanov,N.A., Romashchenko,A.G. and Milanesi,L. (1995) Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Appl. Biosci*., **11**, 477–488.
5. Hutchinson,G.B. (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput. Appl. Biosci*., **12**, 391–398.
6. Solovyev,V. and Salamov,A. (1997) The Gene-Finder computer tools for analysis of human and model organism genome sequences. *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, ISMB-97*, 21–25 June, Halkidiki, Greece. AAAI Press, Melno Park, CA, pp. 294–302.
7. Scherf,M., Klingenhoff,A. and Werner,T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol*., **297**, 599–606.
8. Zhang,M.Q. (1998) Identification of human gene core promoters *in silico*. *Genome Res*., **8**, 319–326.
9. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet*., **29**, 412–417.
10. Bajic,V.B. and Seah,S.H. (2003) Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Res*., **13**, 1923–1929.
11. Kolchanov,N.A., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Stepanenko,I.L., Merkulova,T.I., Pozdnyakov,M.A., Podkolodny,N.L., Naumochkin,A.N. and Romashchenko,A.G. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res*., **30**, 312–317.
12. Schmid,C.D., Praz,V., Delorenzi,M., Périer,R. and Bucher,P. (2004) The Eukaryotic Promoter Database EPD: the impact of *in silico* primer extension. *Nucleic Acids Res*., **32**, D82–D85.
13. Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1999) The biology of eukaryotic promoter prediction—a review. *Comput. Chem*., **23**, 191–207.
14. Babenko,V.N., Kosarev,P.S., Vishnevsky,O.V., Levitsky,V.G., Basin,V.V. and Frolov,A.S. (1999) Investigating extended regulatory regions of genomic DNA sequences. *Bioinformatics*, **15**, 644–653.
15. Zhang,M.Q. (1999) Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem*., **23**, 233–250.
16. Pesole,G., Liuni,S. and Dsouza,M. (2000) PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, **16**, 439–450.
17. Marsan,L. and Sagot,M.F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comput. Biol*., **7**, 345–362.
18. Pavesi,G., Mauri,G. and Pesole,G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17**, 207–214.

19. Pevzner,P.A. and Sze,S.H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB-00)*, 20–23 August, La Jolla, CA. AAAI Press, Melno Park, CA, pp. 269–278.
20. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
21. Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.
22. Grundy,W.N., Bailey,T.L. and Elkan,C.P. (1996) ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. *Comput. Appl. Biosci.*, **12**, 303–310.
23. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
24. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
25. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R. and Alexandersson,M. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
26. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
27. Trinklein,N.D., Aldred,S.J., Saldanha,A.J. and Myers,R.M. (2003) Identification and functional analysis of human transcriptional promoters. *Genome Res.*, **13**, 308–312.
28. Halees,A.S., Leyfer,D. and Weng,Z. (2003) PromoSer: a large-scale mammalian promoter and transcription start site identification service. *Nucleic Acids Res.*, **31**, 3554–3559.
29. Zhang,T. and Zhang,M. (2001) Promoter Extraction from GenBank (PEG): automatic extraction of eukaryotic promoter sequences in large sets of genes. *Bioinformatics*, **17**, 1232–1233.
30. Chong,A., Zhang,G. and Bajic,V.B. (2003) FIE2: a program for the extraction of genomic DNA sequences around the start and translation initiation site of human genes. *Nucleic Acids Res.*, **31**, 3546–3553.
31. Solovyev,V.V. and Shahmuradov,I.A. (2003) PromH: promoters identification using orthologous genomic sequences. *Nucleic Acids Res.*, **31**, 3540–3545.
32. Liu,R. and States,D. (2002) Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res.*, **12**, 462–469.
33. Fickett,J.W. and Hatzigeorgiou,A.C. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
34. Bajic,V.B., Tan,S.L., Suzuki,Y. and Sugano,S. (2004) Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.*, **22**, 1467–1473.
35. Ananko,E.A., Bazhan,E.A., Belova,O.E. and Kel,A.E. (1997) Mechanisms of transcription of the interferon-induced genes: a description in the IIG-TRRD information system. *Mol. Biol. (Mosk.)*, **31**, 592–605.
36. Podkolodnaya,O.A. and Stepanenko,I.L. (1997) Mechanisms of transcription regulation of erythroid -specific genes. *Mol. Biol. (Mosk.)*, **31**, 671–683.
37. Vishnevsky,O.V., Anan'ko,E.A., Ignatieva,E.V., Podkolodnaya,O.A. and Stepanenko,I.V. (2004) Argo_viewer: a package for recognition and analysis of regulatory elements in eukaryotic genes. In Kolchanov,N. and Hofestaedt,R. (eds), *Bioinformatics of Genome Regulation and Structure*. Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 71–81.