 Nucleic Acids Research, 2000, Vol. 28, No. 1 **205–208**

SELEX_DB: an activated database on selected randomized DNA/RNA sequences addressed to genomic sequence annotation

Julia V. Ponomarenko*, Galina V. Orlova, Mikhail P. Ponomarenko, Sergey V. Lavryushev, Anatoly S. Frolov, Svetlana V. Zybova and Nikolay A. Kolchanov

Laboratory of Theoretical Genetics, Institute of Cytology and Genetics, 10 Lavrentyev Avenue, Novosibirsk 630090, Russia

Received September 1, 1999; Revised September 10, 1999; Accepted September 30, 1999

ABSTRACT

SELEX_DB is a novel curated database on selected randomized DNA/RNA sequences designed for accumulation of experimental data on functional site sequences obtained by using SELEX and SELEX-like technologies from the pools of random sequences. This database also contains the programs for DNA/RNA functional site recognition within arbitrary nucleotide sequences. The first release of SELEX_DB has been installed under SRS and is available through the WWW at http://wwwmgs.bionet.nsc.ru/mgs/systems/selex/

INTRODUCTION

Functional site recognition is one of the key aspects of genomic DNA annotation (1). A huge number of methods have been developed so far to address this problem. The most widely used are the matrix methods (2–6) based on the evolutionarily conservative nucleotides of functional sites and used by various Internet-available tools for promoter and transcription factor binding site recognition, i.e., object-oriented Transcription Factors Database (ooTFD) (7), PromFD (6), TESS (8), the TRANSFAC-based expert system (9), ConsInd and Cons-Inspector (10), MatInd and MatInspector (4), CoreSearch (11), MATRIX SEARCH (12), SIGNAL SCAN (13), FunSiteP (14), etc. These programs refer to consensuses and weight matrices for DNA–protein binding sites accumulated in the specialized databases such as TRANSFAC (9), IMD (12), RegulonDB (15), PLACE (16), PlantCARE (17), etc.

However, over the last 10 years, the novel technologies have been designed for identification of high affinity DNA and RNA sequences (ligands) to a wide variety of different targets, including nucleic acid binding proteins, peptides and small organic molecules (for review, see 18–20). Among these technologies are the following: SELEX (Systematic Evolution of Ligands by EXponential enrichment) (21,22), SAAB (Selected And Amplified Binding site imprint assay) (23), REPSA (Restriction Endonuclease Protection Selection and Amplification) (24), CASTing (Cyclical Amplification and Selection of Targets) (25), and other binding site selection procedures (26,27). In general, genetic analysis *in vitro* of the structural and functional properties of many nucleic acids was enhanced by the availability of methods for the amplification of nucleic acid sequences.

Selected affinity-enriched sequences from combinatorial libraries are widely used for functional site recognition and site activity prediction. For example, analyses of the selected randomized ribosome binding sites in *Escherichia coli* with determined translational yield of each site has enabled the calculation of a weight matrix for prediction of translational yield on the sequence of ribosome binding sites (28). The novel class of exonic splicing enhancers recognized by SR protein were identified by using SELEX-like technologies (29).

The matrices resulting from the analysis of selected affinity-enriched sequences are also stored in the databases TRANSFAC (9), IMD (12) and others. These matrices are used by the programs for site recognition along the matrices based on the analysis of naturally occurring sites. However, the samples of real sites are more heterogeneous than selected *in vitro* sequences. For instance, all *in vitro* selected YY1 binding sites contain the CAT motif (30,31), whereas among those occurring in nature, sites lacking the CAT sequence occur frequently (32; TRANSFAC: R03177, R00688). The particular conditions of an experiment are also of importance. For example, HEN1 protein produced *in vitro* and *in vivo* has different consensuses (33). In some cases, the optimal targets expressed in different tissues are not identical. For instance, targets for MEF2 expressed in brain are not observed in skeletal and cardiac muscle (34). Thus, differences in the DNA binding specificities of MEF2 proteins might be a mechanism by which these factors differentially regulate gene expression during myogenesis and neurogenesis (34).

Given current advances in sequencing whole genomes, combinatorial methods will be important in the next generation of studies, thus making the bridge between raw sequence data and actual biological processes. At present, enormous starting libraries are used in different SELEX processes and contain up to 10^{14}–10^{15} sequences (19). Naturally, this information needs to be collected into public databases available via the Internet.

With respect to the problems mentioned above, we have developed SELEX_DB, a database storing selected affinity-enriched sequences from different combinatorial libraries. The site sequences listed within the database may be used as

*To whom correspondence should be addressed. Tel: +7 383 2333 119; Fax: +7 383 2331 278; Email: jpon@bionet.nsc.ru

independent control data in developing both novel methods for functional site recognition within gene sequences and recognition under concrete experimental conditions documented in the database. Additionally, information on functional site sequences and experimental conditions for their determination is useful for planning novel experiments applying SELEX technology.

DATA REPRESENTATION

A database entry corresponds to a single experiment. Each line of an entry begins with a two-character line code indicating the type of information contained in the line and denoting some informational field in SELEX_DB. As an example, an entry containing the information on *in vitro* selected YY1 transcription factor binding sites from a pool of 18 bp random sequences (30) is shown in Figure 1.

The entry description is based on 27 fields: **AC**, an accession number of an experiment; **ID**, identifier; **DA**, date of creation; **DT**, date of the last update; **FV**, release number; **MN**, name of an entry; **CR**, name of an annotator (linked to SCIENTIST database); **NF**, name of a ligand; **OS**, organism; **OC**, taxon; **TE**, templates for amplification; **EX**, type of an experiment; **EC**, experimental conditions (*in vitro* or *in vivo*); **RF**, reference to the literature source (link to SELEX_BIB database); **KW**, keywords; **NS**, sequence quantity; **AA**, aligned sequences as they are represented in the original paper; **WA**, **WT**, **WG**, **WC**, weight impacts of the letters A, T, G and C, respectively, at functionally important positions; **CN**, consensus; **DR**, links to the other database entries if any; **WW**, a link to recognition tools; **NM**, number of sequences in the set; **SQ**, sequences**.** The field **CC** contains different annotator's comments concerning the functional role of a factor or peculiarities of consensus evaluation.

CONTENT OF THE DATABASE

The first release, SELEX_DB 1.0, contains 105 entries with description of selected DNA/RNA sequences from 85 original papers.

The sequences contained in SELEX_DB could be classified into groups according to the type of the binding molecule (proteins, ligands, organic dyes, small molecules, pharmaceuticals, etc), the type of the nucleic acid molecule (DNA or RNA) or the type of SELEX technology. Mostly, SELEX_DB contains the sequences of different proteins binding to DNA, they comprise up to 85% of the database content. The binding sites for proteins causing various disorders, such as B-cell acute lymphoblastic leukemias (35), breast cancer (36) or myeloid leukemia (37) are described. Among RNA binding proteins there are those influencing splice site selection (38), post-transcriptional regulation (39) or recombination (40).

Among the organisms for which the target sequences were selected are human, mouse, chicken, *Drosophila*, rat, rabbit, some plants and others.

SELEX_DB ACTIVATION

To activate SELEX_DB information, the supplementary database SELEX_TOOLS has been developed by analogy to technology applied by the authors in the databases MATRIX (41), ACTIVITY (42) and B-DNA-FEATURES (43). For a fixed functional site, by using the nucleotide occurrence matrix

```
ID S00J0008
XX
AC BS_YY1
XX
DA 13/04/99
DT 13/04/99
FV 1.0
XX
MN Selected YY1 binding sites
XX
CR Ponomarenko JV; SCI00002
XX
NF YY1
OS human
OC EUKARYOTA
XX
TE 5'-AACGGTCCCTGGCTAAAC-18(N)-CAGTGTGTGGACTATTAG-3'
EX PCR-assisted binding site selection
EC in vitro
XX
RF Yant SR et al, 1995; RFSJ0008
XX
KW YY1, globin gene, binding site (Medline,GenBank)
XX
CC CCAT binding core
XX
NS SEQUENCE QUANTITY: 175
XX
AA Aligned sequences from paper
    A2;......CAGAGACACAGACGCCAT
   A17;..............TACAGCCATTATTCCCCA
   A22;......CAGACTACAATCTACCAT
   A30;...........TGACCGGCGCCATTGTTA
.....
   G27;..............TACAGCCATATTACTGCA
   G54;......TATCNTACGTACCTCCAT
XX
WA 34 28 25 14 20 14   0   0 100   0 15 21  4 18 25 16
WT 26 27 18 16  6 18   0   0   0 100 42 71 81 33 29 39
WG 15 27 39 14 14 67   0   0   0   0 23  5  3 31 22 18
WC 25 18 18 56 60  1 100 100   0   0 20  3 12 18 24 27
XX
CN N N N N V D C C A T N W Y N N N
XX
DR SELEX_TOOLS: S00J0008a
WW RECOGNITION: http://wwwmgs.../selex/YY1a_selex.html
XX
CC ACAT binding core
XX
NS SEQUENCE QUANTITY: 14
XX
AA Aligned sequences from paper
   A12;............CGGAGACATTTTTGAGTA
   A14;............GGTAGACATATTCGGGTA
.....
   F41;...CATCAGGACGGCAGACAT
   G53;....CAGATTAAGGCCGACATT
XX
WA 25 33  8 15 43   0 100   0 100   0 17   0   0 10 11 29
WT 12 17  8 15  0   0   0   0   0 100 67 100 100 30 22 14
WG 38 42 83 23  7 100   0   0   0   0  8   0   0 40 56 29
WC 25  8  0 46 50   0   0 100   0   0  8   0   0 20 11 29
XX
CN N D D N M G A C A T N T T N N N
XX
DR SELEX_TOOLS: S00J0008b
WW RECOGNITION: http://wwwmgs.../selex/YY1b_selex.html
XX
NM A2
SQ CAGAGACACAGACGCCAT
NM A17
SQ TACAGCCATTATTCCCCA
.....
NM G53
SQ CAGATTAAGGCCGACATT
//
```

Figure 1. An example of a SELEX_DB entry.

stored within four SELEX_DB fields WA, WT, WG and WC, the C-encoded procedures recognizing this site were generated and stored within the SELEX_TOOLS database accompanying SELEX_DB. For each matrix extracted from SELEX_DB, the total number of the recognition procedure variants equals 15. Namely, seven procedures calculate the matrix recognition scores [eg., homology score (44), matrix similarity (4), etc.], seven procedures weighting consensus match scores [i.e., by Mahalanobis distance, by information content (45), etc.], and

an integrated procedure averaging 14 partial scores described above. So, we follow the 'impartiality' principle to accumulate a variety of recognition score approximations without any preference. Thus, a user may choose an approximation which better suits the particular biological problem. For an appropriate choice, each recognition procedure is documented by (i) false positive and negative error rates (fields ST and NT, respectively), and (ii) by the histogram of the score calculated over the site sequences versus 8000 random sequences (the field FG). A user may exploit the chosen procedure in two modes: (i) on-line mode, by clicking the field 'WW RECOGNITION' to load the Web-tools implementing this procedure; and (ii) off-line mode, by extracting the C-codes of this procedure (the field C-CODE) in order to incorporate them into a user's software. This is the novelty of our approach.

For example, the SELEX_DB entry S00J0008 describing the randomized/selected DNAs binding the transcription factor YY-1 contains the field 'DR SELEX_TOOLS; S00j008a' as shown in Figure 1. By clicking this field, the SELEX_TOOLS entry S00J0008a is loaded (Fig. 2A). Then the C-procedures for recognition of transcription factor binding site YY-1 with the core 'CCAT' are seen in the window. In addition, the entry S00J0008a contains the field 'WW RECOGNITION', which links to the Web-based tools implementing these C-procedures for an arbitrary DNA sequence. The input window for these Web-tools is shown in Figure 2B. The output window for the fragment inbetween positions 7805 and 7924 of the Moloney murine leukemia virus complete genome (EMBL: J02255, REMLM) input with the option 'from Screen' is shown in Figure 2C. In this window, the YY-1 recognition score profile is shown. The pick marked by the arrow corresponds to the experimentally identified YY-1 transcription factor binding site (positions 7860–7868) documented within the entry R01149 of TRANSFAC database (9). The successful recognition of the natural YY-1 site can be considered as an independent control, because neither natural YY-1 site has been documented in SELEX_DB for development of the YY-1 site recognition tools. Thus, SELEX_DB is directly applicable in the course of genomic sequence analysis.

The other way of SELEX_DB activation is the usage of SRS-formatted (46) keywords. For example, by the standard SRS-indexed keyword 'DNA-binding', the entry S00J0008 shown in Figure 1 may be retrieved and subsequently used for the YY-1 site recognition described previously (Fig. 2). In addition, by exploiting keyword query generator (47), the search of terms contained in SELEX_DB may be automatically provided in the MEDLINE or GenBank databases. For this purpose, it is necessary to click the database name at the end of SELEX_DB field 'KW A, B, …, Z'. Then the query 'A&B&…&Z' is generated and addressed to the corresponding database search machine. As a result, the current list of SELEX_DB-related papers or sequences is retrieved.

Thus, SELEX_DB is (i) a database, (ii) the Web-tools for genomic sequence analysis, and (iii) the query access to MEDLINE and GenBank for extracting related papers and sequences. Hence, SELEX_DB is called an 'activated database'.

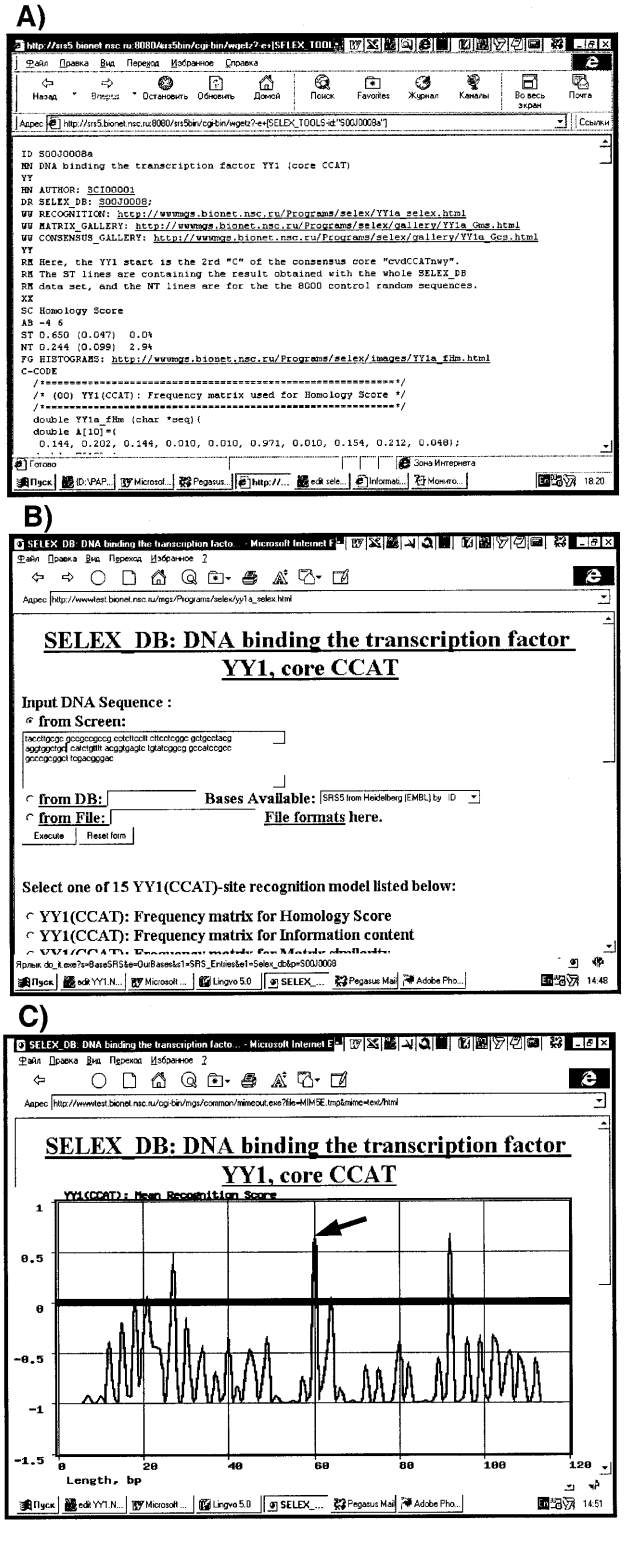

Figure 2. An example of a SELEX_TOOLS entry. (**A**) The C-encoded procedures recognizing a fixed functional DNA/RNA site. (**B**) The input Web-form of the tools implementing these procedures for an arbitrary sequence analysis. (**C**) The output Web-form of the tools showing the site recognition score under the sequence positions (the input sequence start position is '0').

AVAILABILITY

SELEX_DB is available through the WWW at http://wwwmgs.bionet.nsc.ru/mgs/systems/selex/ . It is integrated into GeneExpress System devoted for studying eukaryotic gene expression (48). Email correspondence concerning SELEX_DB usage should

be addressed to the Administrator, J. V. Ponomarenko at jpon@bionet.nsc.ru . For distribution of the flat-files and for storing unpublished experimental data on a collaborative basis within SELEX_DB, contact the Supervisor, Prof. N.A. Kolchanov at kol@bionet.nsc.ru . No inclusion of SELEX_DB into other databases may be made without explicit permission of the authors. Please send comments, corrections and requests for additional information to us by Email or Fax (+7 3832 331 278). We kindly ask users to cite this article in reporting results based on SELEX_DB usage.

ACKNOWLEDGEMENTS

The work is supported by the Russian Foundation for Basic Research (grant nos. 98-07-910126 and 98-07-91078), Integration Program of Siberian Branch of Russian Academy of Sciences (IGSBRAS-97/13), Russian Human Genome Project and Russian Ministry of Sciences.

REFERENCES

1. Haussler,D. (1998) *Trends Guide Bioinformatics*, **1**, 12–15.
2. Bucher,P. (1990) *J. Mol. Biol.*, **212**, 563–578.
3. Karlin,S. and Brendel,V. (1992) *Science*, **257**, 39–49.
4. Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) *Nucleic Acids Res.*, **23**, 4878–4884.
5. Uberbacher,E.C., Xu,Y. and Mural,R.J. (1996) *Methods Enzymol.*, **266**, 259–281.
6. Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1997) *Comput. Appl. Biosci.*, **13**, 29–35.
7. Ghosh,D. (1998) *Nucleic Acids Res.*, **26**, 360–362. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 308–310.
8. Stoeckert,C.J.,Jr, Salas,F., Brunk,B. and Overton,G.C. (1999) *Nucleic Acids Res.*, **27**, 200–203.
9. Heinemeyer,T., Chen,X., Karas,H., Kel,A.E., Kel,O.V., Liebich,I., Meinhardt,T., Reuter,I., Schacherer,F. and Wingender,E. (1999) *Nucleic Acids Res.*, **27**, 318–322. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 316–319.
10. Frech,K., Dietze,P. and Werner,T. (1997) *Comput. Appl. Biosci.*, **13**, 109–110.
11. Wolfertstetter,F., Frech,K., Herrmann,G. and Werner,T. (1996) *Comput. Appl. Biosci.*, **12**, 71–80.
12. Chen,Q., Hertz,G. and Stormo,G. (1995) *Comput. Appl. Biosci.*, **11**, 563–566.
13. Prestridge,D.S. (1996) *Comput. Appl. Biosci.*, **12**, 157–160.
14. Kondrakhin,Y.V., Kel,A.E., Kolchanov,N.A., Romashchenko,A.G. and Milanesi,L. (1995) *Comput. Appl. Biosci.*, **11**, 477–488.
15. Salgado,H., Santos,A., Garza-Ramos,U., van Helden,J., Diaz,E. and Collado-Vides,J. (1999) *Nucleic Acids Res.*, **27**, 59–60. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 65–67.
16. Higo,K., Ugawa,Y., Iwamoto,M. and Korenaga,T. (1999) *Nucleic Acids Res.*, **27**, 297–300.
17. Rombauts,S., Dehais,P., Van Montagu,M. and Rouze,P. (1999) *Nucleic Acids Res.*, **27**, 295–296.
18. Werstuck,G. and Green,M.R. (1998) *Science*, **282**, 296–298.
19. Gold,L., Brown,D., He,Y., Shtatland,T., Singer,B. and Wu,Y. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 59–64.
20. Gold,L., Polisky,B., Uhlenbeck,O. and Yarus,M. (1995) *Annu. Rev. Biochem.*, **64**, 763–797.
21. Tuerk,C. and Gold,L. (1990) *Science*, **249**, 505–510.
22. Ellington,A.D. and Szostak,J.W. (1990) *Nature*, **346**, 818–822.
23. Blackwell,T.K. and Weintraub,H. (1990) *Science*, **250**, 1104–1110.
24. Hardenbol,P., Wang,J.C. and Van Dyke,M.W. (1997) *Nucleic Acids Res.*, **25**, 3339–3344.
25. Wright,W.E., Binder,M. and Funk,W. (1991) *Mol. Cell. Biol.*, **11**, 4104–4110.
26. Pollock,R. and Treisman,R. (1990) *Nucleic Acids Res.*, **18**, 6197–6204.
27. Kinzler,K.W. and Vogelstein,B. (1989) *Nucleic Acids Res.*, **17**, 3645–3653.
28. Barrick,D., Villanueba,K., Childs,J., Kalil,R., Schneider,T.D., Lawrence,C.E., Gold,L. and Stormo,G.D. (1994) *Nucleic Acids Res.*, **22**, 1287–1295.
29. Liu,H.-X., Zhang,M. and Krainer,A.R. (1998) *Genes Dev.*, **12**, 1998–2012.
30. Yant,S.R., Zhu,W., Millinoff,D., Slightom,J.L., Goodman,M. and Gumucio,D.L. (1995) *Nucleic Acids Res.*, **23**, 4353–4362.
31. Hyde-DeRuyscher,R.P., Jennings,E. and Shenk,T. (1995) *Nucleic Acids Res.*, **23**, 4457–4465.
32. Klug,J. and Beato,M. (1996) *Mol. Cell. Biol.*, **16**, 6398–6407.
33. Brown,L. and Baer,R. (1994) *Mol. Cell. Biol.*, **14**, 1245–1255.
34. Andres,V., Cervera,M. and Mahdavi,V. (1995) *J. Biol. Chem.*, **270**, 23246–23249.
35. Van Dijk,M.A., Voorhoeve,P.M. and Murre,C. (1993) *Biochemistry*, **90**, 6061–6065.
36. Buckanovich,R.J. and Darnell,R.B. (1997) *Mol. Cell. Biol.*, **17**, 3194–3202.
37. Morris,J.F., Hromas,R. and Rauscher,F.J. (1994) *Mol. Cell. Biol.* **14**, 1786–1795.
38. Tacke,R. and Manley,J.L. (1995) *EMBO J.*, **14**, 3540–3551.
39. Bai,C. and Tolias,P.P. (1998) *Nucleic Acids Res.*, **26**, 1597–1604.
40. Tracy,R.B., Baumohl,J.K. and Kowalczykowski,S.C. (1997) *Genes Dev.*, **11**, 3423–3431.
41. Ponomarenko,M.P., Ponomarenko,J.V., Frolov,A.S., Podkolodnaya,O.A., Vorobyev,D.G., Kolchanov,N.A. and Overton,G.C. (1999) *Bioinformatics*, **15**, 631–643.
42. Ponomarenko,M.P., Ponomarenko,J.V., Frolov,A.S., Podkolodny,N.L., Savinkova,L.K., Kolchanov,N.A. and Overton,G.C. (1999) *Bioinformatics*, **15**, 687–703.
43. Ponomarenko,J.V., Ponomarenko,M.P., Frolov,A.S., Vorobyev,D.G., Overton,G.C. and Kolchanov,N.A. (1999) *Bioinformatics*, **15**, 654–668.
44. Mulligan,M.E., Hawley,D.K., Entriken,R., McClure,W.R. (1984) *Nucleic Acids Res.*, **12**, 789–800.
45. Schneider,T.D., Stormo,G.D., Gold,L., Ehrenfeucht,A. (1986) *J. Mol. Biol.*, **188**, 415–431.
46. Etzold,T. and Argos,P. (1993) *Comput. Appl. Biosci.*, **9**, 49–57.
47. Kolchanov,N.A., Ananko,E.A., Podkolodnaya,O.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margulis,O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busygina,T.V., Kolpakov,F.A., Podkolodny,N.L., Naumochkin,A.N. and Romashchenko,A.G. (1999) *Nucleic Acids Res.*, **27**, 303–306. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 298–301.
48. Kolchanov,N.A., Ponomarenko,M.P., Frolov,A.S., Ananko,E.A., Kolpakov,F.A.E., Ignatieva,E.V., Podkolodnaya,O.A., Goryachkovskaya,T.N., Stepanenko,I.L., Merkulova,T.I., Babenko,V.V., Ponomarenko,J.V., Kochetov,A.V., Podkolodny,N.L., Vorobiev,D.G., Lavryushev,S.V., Grigorovich,D.A., Kondrakhin,Yu.V., Milanesi,L., Wingender,E., Solovyev,V.V. and Overton,G.C. (1999) *Bioinformatics*, **15**, 669–686.