**SAMPLES AND ALIGNED: DATABASES FOR FUNCTIONAL SITE SEQUENCES.**

VOROBIEV D.G., PONOMARENKO J.V., PODKOLODNAYA O.A.

Institute of Cytology and Genetics, (Siberian Branch of the Russian Academy of Sciences), 10 Lavrentieva ave., Novosibirsk, 630090 Russia

Keywords: database, transcription factor, binding sites, sequences, alignment

The majority of molecular processes in the cell connected with the storage, transmission, and realization of hereditary information is controlled by relatively short regions of the DNA molecule, the so-called functional sites. A great volume of data on their location, structure, and function has been so far accumulated in such databanks as EMBL and TRANSFAC. However, these databases contain only the data on individual sites, whereas no databases of site samples are available. Thus, we believe that development of a specialized database aimed to provide the data for recognition methods is a timely objective.

We have developed the database SAMPLES compiling sets of 42 transcription factor binding sites using the information contained in TRANSFAC, TRRD, and EMBL databases. We have included the homological sequences if they were extracted from different EMBL entries. Names of the sites represented in SAMPLES, their synonyms, number of sequences in each set with and without 100% homologies, and characteristics of the corresponding footprints are listed in Table. Locations of the sites were determined according to TRANSFAC and TRRD, that is, only the sites with the location determined experimentally were taken into account. The sites were aligned relative to the center of the footprints or TRANSFAC sequence elements so that the total length of the sequences was equal to 120 bp.

Table. The list of DNA functional sites compiled in the database SAMPLES.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Factor name | Synonymous factor name | Number of sequences in set with/without 100% homologies | Average size of footprint or TRANSFAC sequence |
|   |   |   | with | without | element, bp |
| 1 | AP-1 | PEA1; (Jun)2; AP1; Fos/Jun; yAP1; PAR1; PDR4; SNQ3 | 74 | 69 | 14.9 |
| 2 | c-Fos | p55(c-fos) | 21 | 19 | 13.9 |
| 3 | c-Jun | JunA; p39; p39 c-jun | 32 | 28 | 14.6 |
| 4 | NF-E2 | NFE2; NF-E2 p45; nuclear factor erythroid 2 p45 | 12 | 12 | 10.7 |
| 5 | CRE-BP1 | XBP4; ATF-2; HB16; TREB-7 | 26 | 22 | 12.7 |
| 6 | ATF | yATF; ATF-1; ATF-2; ATF-3; ATF-4; ATF-5; ATF-6; ATF-7; ATF-8; ATF-a; ATF-adelta; 75 kDa, 77 kDa; TREB-36; ATF/CREB; ATF-3deltaZIP | 28 | 25 | 12.9 |
| 7 | CREB | ATF-47; CREB-341; CREB-B; CREBalpha; CREB-2 | 46 | 37 | 13.6 |
| 8 | C/EBP | slbo; slow border cells; DmC/EBP; AcC/EBP; C/EBPalpha; CBP; EBP20;BPc | 126 | 108 | 19.1 |
| 9 | NF-IL6 | C/EBPbeta; LAP1; NF-M; AGP/EBP; ANF-2; CRP2; H-APF-2; IL-6DBP; LAP; | 23 | 21 | 19.9 |
| 10 | MyoD | XMyoD; Myf-3; MEF1; MyoD1; CMD1; | 17 | 16 | 18.5 |
| 11 | E2F | EIIF; E2F+E4; E2F-BF; E2F-I; E2F-1; E2F-2; E2F-3; E2F-4; E2F-5 | 12 | 9 | 13.2 |
| 12 | USF | UEF; MLTF; SpF1; gamma-factor; yMRF; MRF; pf51; USF43; USF1; eUSF; | 28 | 25 | 14.1 |
| 13 | NF-1 | NF-I; CTF; TGGCA-binding protein; RPF-A (hamster); NF1/L; NF-1A1.1 (chick); NF-I/L; NF-1/Red1; RPF-B; NFI-B1; NF-1B1 (chick); | 101 | 101 | 17.2 |
| 14 | RF-X | EF-C; RFX; RFX1; enhancer factor C; | 12 | 12 | 17.0 |
| 15 | CP1 | NF-Y; CBF;alpha-CP1; | 51 | 33 | 17.2 |
| 16 | ER | estrogen receptor; | 32 | 25 | 17.5 |
| 17 | GR | GCR; glucocorticoid receptor; GR-DBD;GR alpha; GR beta | 64 | 54 | 12.6 |
| 18 | PR | progesterone receptor; PR B; progesterone receptor form B; PR A | 23 | 20 | 11.7 |
| 19 | RAR | retinoic acid receptor; RAR-alpha1; RAR-alpha1; RAR-alpha2; RAR-beta; RAR-beta1; RAR-beta2; RAR-beta3; RAR-beta4; RAR-gamma; RAR-gamma1; RAR-gamma2; | 16 | 16 | 24.6 |
| 20 | RXR | RXR-alpha; retinoid X receptor alpha; RXR-beta ;H2RIIBP; RXR-gamma; RXR-beta2 | 21 | 21 | 20.9 |
| 21 | T3R | thyroid hormone receptor; c-ErbA; T3R-alpha; c-ErbAalpha; T3R-beta; betac-ErbA; c-ErbA-beta; T3R-beta1; T3R-alpha1; T3R-alpha2; c-ErbA-T; ear-7-2; T3R-beta2; | 22 | 21 | 20.5 |
| 22 | COUP | COUP-TF; ear3; COUP-alpha; | 18 | 17 | 20.6 |
| 23 | GATA-1 | NF-E1; NF-E1a; Eryf1; EF1; EFgamma; Eryf-1; GF-1 | 102 | 76 | 16.1 |
| 24 | Sp1 |   | 197 | 176 | 13.1 |
| 25 | YY1& | NF-E1, UCR, UCRBP, CF1, CSBP1, d , F-ACT1 | 27 | 27 | 13.6 |
| 26 | GAGA | Trl; Trithorax-like | 7 | 7 | 16.0 |
| 27 | GAL4 |   | 16 | 16 | 14.7 |
| 28 | EN | Engrailed; En-1; Engrailed 1; Gg-en.1; Hu-en.1 (human); Mo-en.1 (mouse); En-2; Engrailed 2; Gg-en.2; Hu-en.2 (human); Mo-en.2 (mouse); | 12 | 12 | 11.4 |
| 29 | HNF-1 | HNF-1A; APF; HNF1; HNF-1alpha; LF-B1; | 42 | 38 | 21.0 |
| 30 | TTF-1 | T/EBP; thyroid transcription factor 1; thyroid nuclear factor 1; | 7 | 7 | 18.0 |
| 31 | OCT | Oct-1; Oct-1A; Oct-2.1; Oct-2; Oct-4; Oct-5; Oct-8; Oct-9; Oct-2B; oct-B2; oct-B3; Oct-2C; Oct-1B; Oct-1C; Oct-2.1; Oct-2.3; Oct-2.4; Oct-2.6; Oct-2.7; Oct-2.8 | 101 | 73 | 14.9 |
| 32 | HNF-3 | HNF3; HNF-3A; HNF-3alpha; HNF-5; | 15 | 10 | 24.6 |
| 33 | HSF | HSTF; HTF; HSF1; mHSF1; heat shock transcription factor 1; | 7 | 7 | 27.7 |
| 34 | c-Myb |   | 19 | 19 | 18.9 |
| 35 | Ets | c-Ets-1; p54; Ets1; c-Ets-2; Ets2; Ets-2; c-Ets-2 58-64; p58-64c-Ets-2; v-Ets; v-Myb/v-Ets; | 16 | 15 | 20.9 |
| 36 | IRF-1 | ISGF2 | 11 | 7 | 16.1 |
| 37 | NF-kappaB | p50/p65 | 39 | 36 | 13.9 |
| 38 | MEF-2 | MEF-2A; SL-2; | 12 | 12 | 17.3 |
| 39 | SRF | SRFx; p67; p67SRF; CArG-binding factor; CBF (3) | 29 | 29 | 15.7 |
| 40 | E2 |   | 20 | 20 | 15.2 |
| 41 | TCF-1 | T cell factor 1; TCF-1alfa; TCF-1A; TCF-1B; TCF-1C | 6 | 6 | 14.7 |
| 42 | GATA-1 (from TRRD) |   | 81 | 81 | 16.5 |



Fig.1. Example of a SAMPLES card.



Fig. 2. Example of a card from the ALIGNED database.

The database SAMPLES has a EMBL-like format, is under the SRS5, and is available via the Internet at URL: <http://sgi.sscc.ru/srs5/>. An example of the SAMPLES entry is shown in Fig. 1. The card includes the following fields: FI (short name of the set); NM (full name of the set supplemented with explanation of its biological meaning); DA and LU (dates of creation and last update); ST (formalized description of the sites of a given type including the names of the site and the corresponding transcription factor); WW (hyperlink to the corresponding X-ray structure image); WA, WP, and WF (hyperlinks to the databases on significant contextual, conformational, and physico-chemical properties of the sites Matrix, Consensus, Features, and Profiles). A description of the site sequence includes identifier (ID); accession number (AC); phylogenetic classification information (OS and OC); references to source databases (DR); boundary positions of the site in the sequence, its orientation, and the method, experimental or computer-assisted, used to identify this site (FT). Abbreviations EXP. (experimental) and GBS (Gibbs sampler) are used to specify the method used.

To reveal the contextual peculiarities of the sites, the sequences from SAMPLES were aligned by a modified method of local multiple alignment Gibbs sampler (Lawrence, 1993). Only the sequences of footprints with three additional nucleotides from either sides were employed to increase the accuracy of alignment. The results of this alignment are compiled in the database ALIGNED (URL: <http://sgi.sscc.ru/srs5/>). The format of the database ALIGNED is similar to the format of SAMPLES (Fig. 2).

Further development of the databases SAMPLES and ALIGNED will be performed by increasing the number of both the functional sites involved and the sequences in each set of the already described sites. In addition, we plan to supplement SAMPLES with the tool for automatic extraction of necessary data from the EMBL, TRRD, and other relevant databases via the query language MGL (at the moment, the SAMPLES database is filled manually). To resolve problems with homological sequences we plan to include into the database interface an on-line tool for homolog eliminating. We also plan to develop the tools providing an Internet user the on-line input of his own information, its automatic aligning, and generation of knowledge (programs for site recognition).

We thank Grigoriy Kolesov for the program for alignment kindly provided by him.

**References**

1. Lawrence Ch. E., Altshul S. F., Boguski M. S., Liu S. L., Neuwald A. F., Wootton J. C., Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 1993, V. 262, p. 208-214