**SEARCH FOR DEGENERATE OLIGONUCLEOTIDE MOTIFS IN TRANSCRIPTION FACTOR BINDING SITES AND EUKARYOTIC PROMOTERS (THE SYSTEM ARGO).**

VISHNEVSKY O.V.+, PODKOLODNAYA O.A., BABENKO V.N.

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

+Corresponding author e.mail: oleg@bionet.nsc.ru

Keywords: oligonucleotide motifs, binding sites, eukariotic promoters, tissue specificity, globin genes

A new method for recognition of functional sites and gene regulatory regions has been developed. This method is based on construction of sets of degenerate oligonucleotide motifs and, unlike the existing recognition methods, such as consensus building or weight matrix, requires no preliminary alignment of the sequences to be analyzed. The approach developed was used to analyze sets of sequences of transcription factor binding sites and several groups of tissue-specific eukaryotic promoters. For each sequence sample analyzed, the set of oligonucleotide motifs significantly occurring in and specific for these sequence sample was found.

**Introduction**

Efficient and accurate recognition of the regulatory regions of genes and functional sites is important for both identification of genes and prediction of tissue-specificity of their .expression. The currently available methods for promoter region recognition [1] are based on the data concerning various transcription factor binding sites. These data are represented in the form of either consensuses [2] or weight matrices [3]. However, alignment of the sequences under study containing a given functional site is necessary to build both the consensus and weight matrix. The methods implementing the search for common oligonucleotide motifs, on the contrary, require no preliminary processing of the sample under study as well as no additional information concerning the sequences [4]. We have earlier developed a set of programs for revealing the oligonucleotides that are specific for the coding regions of the families of isofunctional genes [5]. Basing on the method of search for degenerate oligonucleotide motifs, we have developed a new software package ARGO, designed for analysis of functional sites and gene regulatory regions.

**Methods**

A set of unaligned nucleotide sequences of a certain regulatory genomic sequence (RGS) is the initial information for the analysis. Degeneracy of the motifs means that they are considered in an expanded 15 single letter-based code. The method of the search for significant motifs is based on consideration of the complete vocabulary of the length L for each RGS and subsequent clusterization of the oligonucleotides belonging to different RGSs. The oligonucleotides from different sequences, the Heming’s distance R between which is lower than the threshold value ro, are united in one class; the consensus in a 15 single letter-based code is created for each class as follows. The significance of each of the 14 sense signals at each position is evaluated by binomial criterion, and the signal with the minimal probability to appear by accidence is selected. The oligonucleotide motif obtained by this procedure is considered significant, if it meets the following conditions: (1) the fraction f of RGS, in which it occurs, is higher than a certain given level fo and (2) the binomial probability P(n,N) to observe this motif by accidence in n and more RGS of N RGS considered is lower than a given significance level a.

Results of the analyses performed are accumulated in the knowledge base of the system ARGO in its inner standard format. This approach requires no preliminary alignment of the RGS analyzed, representing an advantage compared to all available methods for constructing consensuses and weight matrices.

**Results**

This approach was implemented to analyze the sets of nucleotide sequences of transcription factor binding sites (GATA-1, etc.) and of several groups of tissue-specific promoters. The sets of significant oligonucleotide motifs were found in each of these cases. While studying the eukaryotic promoters, the patterns of location of the motifs revealed along the RGS were discovered in several cases. These oligonucleotide motifs can be used for construction of RGS-recognizing methods.



Fig. 1. Distribution of significant oligonucleotides within GATA-1 site of the promoter of human epsilon-globin gene. Experimentally observed footprint is bold-faced.

Table 1. Characteristics of significant oligonucleotide motifs for GATA-1 sites

|  |  |  |  |
| --- | --- | --- | --- |
| Motif | Fraction of GATA-1 site sequences containing a motif | Fraction of random sequences containing a motif | Significance level of motif occurrence in GATA-1 sites |
| Nhgakadn | 0.94 | 0.28 | 10 -14.49 |
| Mhkatvdv | 0.94 | 0.28 | 10 -13.35 |
| Vwkatvdv | 0.94 | 0.25 | 10 -13.59 |
| Vhgatrdn | 0.94 | 0.23 | 10 -17.69 |
| Hnwkatvd | 0.94 | 0.29 | 10 -11.05 |
| Hgakadnn | 0.94 | 0.28 | 10 -14.49 |
| Nnhgatrd | 0.97 | 0.27 | 10 -16.08 |

The list of the significant motifs of the length **L = 8** revealed while analyzing the GATA-1 sites of erythroid-specific genes is shown in Table 1. The set analyzed contained 35 sequences of GATA-1 site. In this search, occurrence of the motif in at least 85% of the sequences was required; significance interval for a symbol in a position, 0.01; significance interval for a motif revealed, 10-10. Nondegenerate or low degenerate symbols are located at the central positions of these motifs. Projection of the motifs revealed on the actual sequences studied demonstrated that in the majority of cases they are clustered around the center of the footprint (Fig. 1), although some motifs were located outside the experimentally revealed region (may be, due to the noise).



Fig. 2. Distribution of significant motifs along the sample of GATA-1 sites. Sites are aligned relative to the center of footprints.

The integrated distribution of the motifs revealed over all the sequences of the site is shown in Fig. 2. Note that the maximum of this distribution corresponds to the centers of the experimentally determined regions of GATA-1 sites, indicating the utility of the motifs revealed by this approach for functional site recognition.

The simplest variant of the procedure of site recognition basing on the motifs revealed is as follows. All positions of the sliding window with a length equal to the RGS under study are considered. The motifs from the corresponding list are searched for at each position of the window. If the fraction of the motifs from the total list found in this window exceeds a threshold level, than it is considered that a site, the position of which is determined by the integral location of these motifs, is contained in this window. For example, the type I error (false negative) for GATA sites amounts to 9%; type II error (false positive), 13% .

Sets of significant motifs were revealed for different regions of globin gene promoters (Table 2).

Table 2. Significant oligonucleotide motifs in the core promoter region of globin genes (EPODB). Each motif is present at least in 70% of the sequences of the corresponding GATA-1 site regions, whereas the probability of its random occurrence in a given number of sequences is less than 10-12.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| -100 –60 | -80 –40 | -60 -20 | -40 +1 | -20 +20 |
| NrrccarnNnrccaryWttkgnnnNrrccamnRrccarnnRrnnaatnRrccamnnNrccamynTnnncymaTtgryynnRccamynnRsmaawnnMytsrcynNycantnk | NcwgncmnCcwgnynnNcnksccyNncwgncmNccwgnynYmannkgmGnmnntkgNnncarkgAnkggnynNnntgrcyNcnytgnyNcmnkgrsRscwgrnn | RnmwnwaaNtntwtnyNnawawanNnnrtawaRnrnwwaaRnnwawarRknwnwaaNnwtwwanYtwtwwnnTwwawnnnNrnrwawaNntwtawnNawawannNnyawaarNntwwawn | wwwaarnnyawaarnnawaarnsnnawaarnsnsnwtwwawwawarnnsnwtwwanwwawarsnnyawaarnnnntwtawnatanannnnyawaaryttwyrynnntwwawrnnnatanantntatnn | nctnctsnncwkntgnnswgmwgngnkkcwgnnstkctsnksknctgnnsasmagnnnctnctsnnstkctsngnkkcwgnksknctgnnctkstsstkctsnnctkstsnnnnsakmwgnycwganntcwgnnncgnrncwgmtcwgrnnnntcwgrnn |

This analysis demonstrates that the occurrence of groups of site-specific motifs in definite regions relative to the transcription start is typical of promoters. The results obtained demonstrate the utility of the approach described for development of methods for recognition of gene-specific eukaryotic promoters.

**Acknowledgments.**

**References**

1. Fickett J.W. and Hatzigeorgiou A.G., “Eukaryotic promoter recognition”, submitted.
2. Day W.H.E. and McMorris F.R., “Critical comparison of consensus methods for molecular sequences”, Nucleic Acids Res., 20, 1093-1099, (1992).
3. Bucher P., “Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences”. *J. Mol. Biol.*, 212, 563-578, (1990).
4. Kondrakhin Yu.V., Babenko V.N., Milanesi L., Lavryushev S.V. and Kolchanov N.A., “Recognition groups: a new method for description and prediction of transcription factor binding sites”, submitted.
5. Kolchanov N.A., Vishnevsky O.V., Babenko V.N., Kel A.E., “Oligonucleotide Sets. Computer Tool and Application”. Proceedings Third International Conference on Intelligent Systems for Molecular Biology. AAAI Press. Menlo Park.California (1995).