Promoter Analysis using Position Specific Scoring Matrix (PSSM)

P.K. Naik and S. Marla

Department of Bioinformatics, Jaypee University of Information Technology
Waknaghat, Solan - 173 215, Himachal Pradesh, India,
e-mail: pknaik73@rediffmail.com

Receiving Date: 26-11-04 **Accepted on: 31-01-05**

ABSTRACT: Binding of RNA polymerase to DNA sequence play a central role in transcription regulation, and the annotation of polymerase binding sites in upstream regions of genes is essential for building a genome-wide regulatory network. In this work we describe a methodology to refine the accuracy in predicting the polymerase binding site region of function specific genes. In order to increase the accuracy of promoter site prediction, we rely on available genome sequence data and the known polymerase binding-site matrices. We analyzed a majority of promoter sites from E. coli with evidence from existing literature. Obtained scores for exact TATAAT is 8.67 and score for exact TTGACA is 6.70. Any sequence deviated from it has score less than this score. These subsequence can be anywhere in upstream region. In large percentage of cases distance between subsequences varies between 11 to 20. Our tools and analysis provide a new resource for deciphering transcription regulation in different biological paradigms.

KEYWORDS: RNA polymerase, Binding sites, promoters, matrix, log odd score.

INTRODUCTION

Short, conserved sequence elements, or DNA motifs located upstream of the transcriptional start site are often the binding factors that play a major role in gene regulation. The availability of the draft sequence of the human genome and other organisms is an enormous achievement, but characterizing the entire set of functional elements encoded in the human and other genomes remains an immense challenge. Two of the most important functional elements in any genome are RNA polymerase and the promoter sites within the DNA to which they bind. These interactions between protein and DNA decide gene expression level which is critical steps in development and response of an organism to various environmental stresses. The development of computer algorithms to correctly recognize polymerase II promoter sequences in primary sequence data, however, is an extremely complex and difficult problem. While there have been many successful attempts to build algorithm to recognize prokaryotic promoter sequences using a variety of approaches. The method that has been most often been used to analyze *E.Coli* promoters is to align a set of promoter sequences by the position that marks the known transcription start sites (TSS) and then to search for conserved sequences in the upstream regions. The promoter region is found to contain three conserved sequence features: a region approximately 6 bp long with consensus TATAAT at position -10 (the Prebnow box); a second region approximately 6 bp long with consensus TTGACA at position -35, and distance between them approximately 17 bp that is relatively constant. A weaker region exists around +1, and designation given to the start of transcription, and an AT- rich region is found before the 1-35 region. These promoters sequences are frequently degenerate motifs and the sequence degeneracy has been selected through evolution and is found to be beneficial, since it confers different levels of activity upon different promoters, thus causing some genes to be transcribed at higher levels than others, as may be required by the cell.

Identifying candidate promoters in silico

Once a regulatory sequence motif has been identified, the next goal is frequently to identify candidate target genes that may be regulated through it, potentially by a RNA polymerase that may bind to it. Although degenerate consensus sequences are still frequently used to depict the binding specificities of polymerase, they do not contain precise information about the relative likelihood of observing the alternate nucleotides at the various positions of a promoters. Thus, a

common way of representing the degenerate sequence preferences of a DNA-binding protein is by a position weight matrix (PWM), also known as a position-specific scoring matrix (PSSM). Briefly, the elements of a PWM correspond to scores reflecting the likelihood of observing that particular nucleotide at that particular position of the known or candidate promoter. Although there are certain problems inherent in the use of PWMs, they are nevertheless a good approximation and a useful representation that can identify biologically interesting candidate sites.

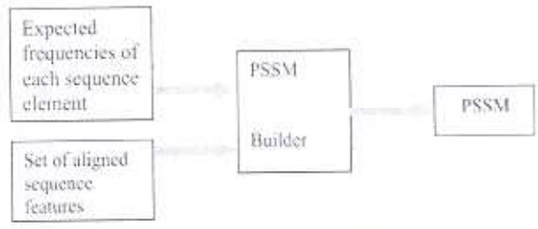
Position specific scoring matrix (PSSM)

The PSSM may be used to search a sequence to obtain the most probable location of a motif represented by PSSM. Alternatively PSSM may be used to search an entire database to identify the additional sequences that also have the same motif. Consequently, it is important to make the PSSM as the representative of the expected sites as possible. The quality and the quantity of the information provided by the PSSM also varies for each column in the motif, and this variation profoundly influences the matches found with sequence.

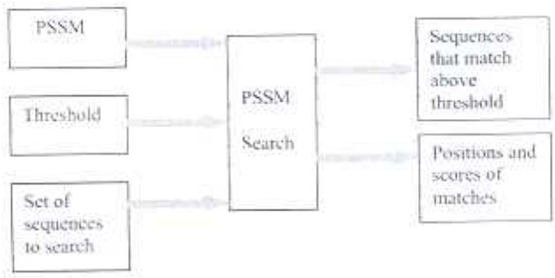
The PSSM is constructed by a simple logarithmic transformation of a matrix giving the frequency of each amino acid in the motif. Two considerations arise in trying to tune the PSSM so that it adequately represents the training sequences. First, if number of sequences with found motif is large and reasonably diverse, the sequences represent a good statistical sampling of all sequences that are ever likely to be found with the same motif. Given a good sampling of sequences is available, the number of sequences is sufficiently large, and the motif structure is not too complex, it should in principle be possible to obtain frequencies highly representative of all similar motifs in other sequences too.

Application of PSSM in promoter analysis

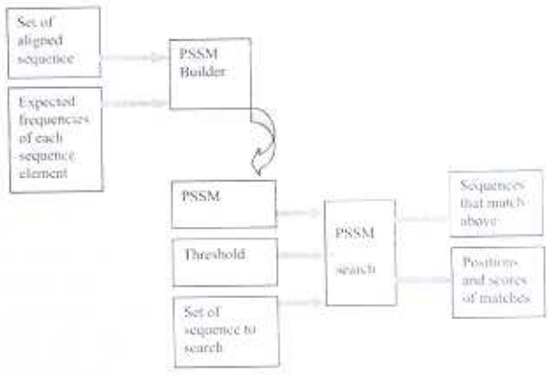
A more complex type of promoter analysis is used for both prokaryotic and eukaryotic sequences in a scoring or weight matrix. For this purpose we do multiple sequence alignment to find additional sequence of same pattern in the database search using Prosite and CLUSTALW. In prokaryotes a scoring matrix for representing the 1-10 region of promoters is to construct in the promoter sequence. Followed using block analysis of these conserved sequences we can calculate fraction of each base at each column of the aligned promoter in the 1-10 region. Using this frequency log odds can be generated. The same approach is followed for 1-35 region of promoter sequences. Given a new sequence it needs to scan against these matrices. Best scored subsequence predicted as promoter sequence. Gap between these sequences are also very important to predict real promoter sequences.



Block Diagram for building PSSM



Block diagram for searching with PSSM



For sequences related block diagram for searching to a family with PSSM

PROTOCOLS AND METHODS

- About 470 E. coli promoter sequences from E. coli database was taken and each sequence has a length form -75 to +25 nucleotides.
- All the promoter sequences were aligned using CLUSTALW according to region corresponding to TATAAT (TATA box) at -10 region and TTGACA region at -35 and sequence profile were generated for these consensus subsequences.
- Log odd scoring matrix was developed from the sequence profile.

  Odds score = Observed frequency / Expected frequency

  Score (i) = log m(j,i)/f(j)

  Where m(j,i) is the frequency of character j observed at position i and f(j) is the overall frequency of character j (usually in some large set of sequence)

  Log odds score = log (observed frequency/expected frequency)/(log2)
- The query sequences were scanned against the matrices and the subsequence of best score were selected. In order to score the subsequence log odds score was converted into odds score.

  Odds score = 2 (log odds score) = e (log odds score)
- Same procedure was used for -35 region of query sequences and best subsequence in it was located.
- The distance between these consensus sequences was calculated.
- If the sequences which have -10, -35 region and distance between them are found adequate enough then they can be identified as promoter sequences.

Consensus consideration for promoter

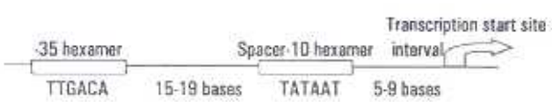


Table 1. Log odds score for -10 and -35 region.

	1	2	3	4	5	6
A	0.00	1.41	0.00	1.58	-1.22	0.00
T	1.22	0.41	1.58	-8.23	0.73	1.22
G	-0.58	-8.23	-8.23	-0.58	-8.23	-8.23
C	-8.23	-8.23	-8.23	-1.58	-8.23	-0.58

-35 region

	1	2	3	4	5	6
A	-0.58	8.23	-1.58	1.00	-1.58	0.73
T	1.00	1.73	0.00	-0.58	0.00	0.00
G	-1.58	-1.58	1.22	0.00	-0.50	0.00
C	0.00	-1.58	-1.58	-1.58	1.00	-1.58

-10 region

RESULTS AND DISCUSSION

Consensus sequences are generated by calculating the frequency at each point of subsequences of multiple local alignment. The consensus sequences are TATAAT and TTGACA. Obtained Score for exact TATAAT is 8.67 and for TTGACA is 6.70. Any sequence deviated from it has score less than this score. These subsequence can be located any where in upstream region. Difference in distance between these subsequences is crucial for finding actual sequences.

Table 2. Predicted score of the different subsequences using PSSM tool.

Range of score	No. of sequences (nt difference)	Percentage of Sequences (%)
0-10	5	14
11-20	13	37
21-30	6	17
31-40	7	20
41 to above	4	11

In large percentage of cases distance between subsequences varies between 11 to 20 (Table 2). This range is most suited to experimental data. For scoring E.coli sequences for the presence of promoters, scoring matrices for a -35 region, a 19-bp region encompassing the -10 region, a 12-bp region encompassing the +1 region. Each matrix will produce a distribution of odds score that predicted possible location for matches to itself in query sequence.

There are several regions that matrix methods do not always achieve a better

prediction of *E.coli* promoters. The first is that the matrix methods add to the score for each sequence position, where in reality; one position is the 1-10 region, for example, may play a role in one stage of transcription such as promoter recognition by RNA polymerase. Whereas another may play a role in the subsequent stage of transcription, such as initiation of transcription or elongation of mRNA. Matching positions with these types of functional separations are not expected to be additive, as assumed by matrix method. A second difficulty that the matrix method shares with most other methods of promoter prediction is that all promoters are treated as being in the same class , where different form of RNA polymerase that are complex with a set of transcriptional activation may have a particular preference for different sequence positions in the promoter region.

In recent years, a number of efforts have been focused on attempting to predict promoter sites and other transcriptional regulatory site using structural information on the protein or related protein-DNA complexes. Some of these studies have attempted to determine what 'recognition rules' or 'recognition code' may exist that stipulate which DNA base-pairs are likely to be bound by which amino acids, in the context of a particular structural class of DNA binding proteins. These approaches have come either from analysis of databases of well-characterized DNA-protein interactions, from computer modeling, or from experiments employing in vitro selection from a randomized library, either of the DNA base pairs or the amino-acid residues implicated in sequence-specific binding. There is no obvious, simple code like the genetic code, however, and any recognition rules that might exist are likely to be quite degenerate and highly dependent upon the docking arrangement of the protein with its DNA binding site. This area of work, including the possibility of deciphering a 'probabilistic code', is discussed by Benos *et al*. Such efforts will be greatly aided by the further development of high-throughput technologies for identifying interactions between polymerase and their DNA binding sites, so that much larger datasets can be generated for analyses required to decipher any 'degenerat probability codes' or to be used as training set for developing improved DNA binding-sit prediction algorithms. Studies using the variou high-throughput technologies described earlie will permit a better understanding of th locations and organization of regulatory DN elements in and the regulatory complexit resulting from combinatorial interactions o polymerases.

OUTPUT OF THE PROGRAM

Position -10 = -12 to -7, Score = 7.67

TTTATA

Position -35 = -21 to -16 , Score= 3.11

CTGTTA

Difference of position = 9.

Position -10 = -8 to -3 , Score = 7.67

TTTAAT

Position -35 = -43 to -38, Score = 4.11

TTGTTA

Difference of position = 35.

Position -10= -12 to -7, Score= 5.37

TTTAAT

Position -35= -40 to -35, Score= 4.37

TTGAGC

Difference of position = 28.

Position -10= -26 to -21, Score= 7.67

TTTAAT

Position -35= -61 to -56, Score= 4.11

TTGTTA

Difference of position = 35.

Position -10= -16 to -11, Score= 7.44

TTTAAT

Position -35 = -21 to -16, Score = 4.11

TTGTTA

Difference of position = 35.

REFERENCES

- Collins, F., Green, E., Guttmacher, A., Guyer, M. 2003: US National Human Genome Institute: A vision for the future of genomics research. Nature. 422:835-847.
- Stormo, G. 2000: DNA binding sites: representation and discovery. Bioinformatics 16:16-23
- Man, T.K., Stormo, G.D. 2001: Non-independence of Mnt repressoroperator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res 29:2471-2478.
- Bulyk, M., Johnson, P., Church, G. 2002: Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res 30:1255-1261.
- Benos, P., Bulyk, M., Stormo, G. 2002: Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res 30:4442-4451.
- Lee, M.L., Bulyk, M., Whitmore, G., Church, G.2002: A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. Biometrics 58:981-988.
- Jacobs, G. 1992: Determination of the base recognition positions of zinc fingers from sequence analysis. EMBO J 11:4507-4517.
- Desjarlais, J., Berg, J. 1992: Redesigning the DNA-binding specificity of a zinc finger protein: a data base-guided approach. Proteins 12:101-104.
- Desjarlais, J.R., Berg, J.M. 1992: Toward rules relating zinc finger protein sequences and DNA binding site preferences. Proc Natl Acad Sci USA 89:7345-7349.
- Suzuki, M., Yagi, N. 1994: DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. Proc Natl Acad Sci USA 91:12357-12361.
- Mandel-Gutfreund, Y., Baron, A., Margalit, H. 2001: A structure-based approach for prediction of protein binding sites in gene upstream regions. Pac Symp Biocomput 139-150.
- Pabo, C., Nekludova, L. 200: Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J Mol Biol 301:597-624.
- Benos, P., Lapedes, A., Stormo, G. 2002: Is there a code for protein-DNA recognition? Probabilistically. Bioessays 24:466-475.