Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Web Server**

A web server for transcription factor binding site prediction

Gang Su, Binchen Mao and Jin Wang*

The State Key Laboratory of Pharmaceutical Biotechnology and Model Animal Research Center, School of Life Science, Nanjing University, Nanjing - 210093, China;

Jin Wang* - Email: jwang@nju.edu.cn; *Corresponding author

received April 30, 2006; revised May 9, 2006; accepted May 27, 2006; published online June 15, 2006

Abstract:

Promoter prediction has gained increased attention in studies related to transcriptional regulation of gene expression. We developed a web server named PMSearch (Poly Matrix Search) which utilizes Position Frequency Matrices (PFMs) to predict transcription factor binding sites (TFBSs) in DNA sequences. PMSearch takes PFMs (either user-defined or retrieved from local dataset which currently contains 507 PFMs from Transfac Public 7.0 and JASPAR) and DNA sequences of interest as the input, then scans the DNA sequences with PFMs and reports the sites of high scores as the putative binding sites. The output of the server includes 1) A plot for the distribution of predicted TFBS along the DNA sequence, 2) A table listing location, score and motif for each putative binding site, and 3) Clusters of predicted binding sites. PMSearch also provides links for accessing clusters of PFMs that are similar to the input PFMs to facilitate complicated promoter analysis.

Availability: PMSearch is available for free at http://www.nicemice.cn/bioinfo/PMS

Keyword: position frequency matrix; motif; transcription factor binding site; web server

Background:

Transcription factors play a pivotal role in the regulation of gene expression by sequence specific binding to the promoters of target genes. Prediction of putative transcription factor binding sites has become an important resource to explore genome organization and regulatory mechanisms. The binding specificity of transcription factors are usually represented by known sequence motifs (consensus sequences) or matrices (PFM (Position Frequency Matrix) or PWM (Position Weight Matrix)). **[1]** High throughput analyses using SELEX (Systematic Evolution of Ligands by Exponential Enrichment) and CHIP-Chip (Chromatin Immunoprecipitation-Microarray) along with computational sampling methods have generated thousands of PFMs. These data together with the related transcription factor information are curated in online databases such as Transfac **[2]**, JASPAR **[3]**, etc.

Online applications, such as MatInspector **[4, 5],** MATCH **[6]** and ConSite **[7],** have been built to utilize PFMs to predict transcription factor binding sites (TFBSs) embedded in promoter sequences. Many of the servers are comprehensive but lack the information on transcription factors whose binding specificities are similar to the input PFMs, which could effectively assist regulatory module finding. To provide such information and to meet the needs for efficient prediction tasks in the study of gene regulation network, we developed a server named PMSearch (Poly Matrix Search) for predicting TFBSs in DNA sequences with the novel functions of fetching similar PFMs and processing multi-forms of input motifs. PMSearch made the following improvements: 1) It has a more succinct and friendly user-interface. 2) It provides user with the convenience of customizing any set of PFMs from the local dataset. 3) It generates a resizable plot that shows the distribution and scores of predicted binding sites in variable scales. 4) It provides clusters of PFMs similar to the input PFMs.

Methodology:

We have implemented a scoring scheme adapted from previous algorithms. **[4, 6]** The uninformative nucleotides (ambiguous letters: N, B, D, H, and V) on either end of an input PFM are discarded before searching to enhance efficiency and specificity. We downloaded 507 PFMs from JASPAR and Transfac Public 7.0 to construct a local dataset from which the user may take any subset of PFMs, or along with user-defined PFMs, for a prediction task. In addition to PFMs, the server could also accept sequence motifs (modeled in the format of consensus sequences), which will then be converted into pseudo PFMs and applied in subsequent prediction. We have set a default global cutoff value of 0.85 previously adopted by TFSEARCH. **[9]** The user may adjust this cutoff if required.

User interface:

The web interface of PMSearch helps in promoter analysis. The inputs are composed of PFMs and sequences under analysis. The user may submit PFMs, motifs or retrieve desired PFMs from our dataset using keywords (e.g., name of the transcription factor), its source database accession number (e.g. Transfac Matrix Accession number M00001) or our local accession number (e.g. X00001). Sequences in FASTA format or Genbank format are acceptable and will be auto-parsed. When the user initiates a

predication task, each sequence will fork an independent task in which corresponding results can be retrieved respectively. In the output, PMSearch plots a figure that illustrates binding sites that are scored above the cutoff value (Figure 1). The user may specify a sub-region on the sequence that will be plotted for more detailed view, such as a specific promoter region in a sequence file. Other results including scores, sequence motifs and closely located putative binding sites (clusters) are also listed. In addition, the server provides a hyperlink for the user to fetch PFMs that are related to the input PFMs by implementing a gap-allowed alignment algorithm. A comprehensive instruction for users is available online.

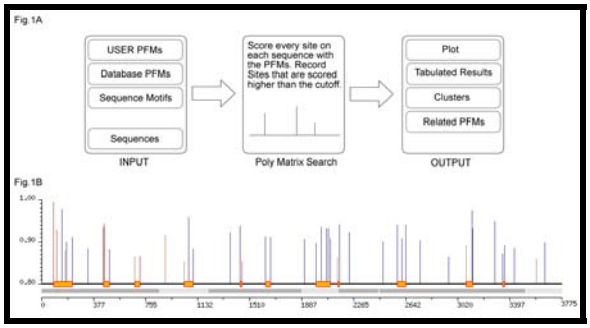

Figure 1: The scoring scheme is illustrated. 1B: An example of binding site search with PFMs for NF-Kappa B, AP-1 and Sp1 on the sequence of adenovirus early E3 gene (NCBI nucleotide accession: X03002). Orange rectangles represent clusters of putative binding sites. Dark grey bar indicates coding sequence (CDS). The first cluster located on the 5' end contains verified binding sites. This cluster belongs to E3 promoter, which overlaps with the adjacent CDS. High scoring sites located in non-coding sequences or promoters are considered significant

Conclusion:

PMSearch is an easy-to-use and efficient tool that utilizes PFMs to predict transcription factor binding sites in DNA sequences. It offers user the flexibility to search for putative TFBS with any set of PFMs. PMSearch outputs a plot demonstrating distribution of predicted binding sites and a table of the locations, scores, motifs and clusters of predicted binding sites. In addition, PMSearch provides the user access to PFMs that are relative to an input PFM for more sophisticated promoter analysis, as the predicted binding sites of one transcription factor could also be bond with other transcription factors that share the similar binding specificity. Such information may give hints to untangle the composite transcription regulatory network. We propose to update PFMs in our local dataset regularly. The source codes are available from the authors upon request.

Caveats:

It has been proposed that many of the predicted TFBSs lack biological function in vivo. **[10]** This could result from 1) The predicted site is located in a context which is insufficient to facilitate transcription factor binding, 2) The low specificity of the input PFM gives rise to large portion of false positive predictions. We suggest the users prepare a list of certain transcription factors that are suspected to regulate the target gene to specify a prediction task.

Acknowledgement:

This work is supported by the National Science Foundation of China (No.90208021) and 973 Project Grant No. 2003CB715905 founded by MOST of China. The implementation was done in IBM-NJU Laboratory of Bioinformatics.

References:

[1] G. D. Stormo, *Bioinformatics,* 16:16 (2000) [PMID: 10812473]

[2] E. Wingender, *In Silico Biol.,* 4:55 (2004) [PMID: 15089753]

[3] D. Vlieghe, *et al., Nucleic Acids Res.,* 34:D95 (2006) [PMID: 16381983]

[4] K. Cartharius, *et al., Bioinformatics,* 21:2933 (2005). [PMID: 15860560]

[5] T. Werner, *Methods Mol Biol.,* 132:337 (2000) [PMID: 10547845]

[6] A. E. Kel, *et al., Nucleic Acids Res.,* 31:3576 (2003) [PMID: 12824369]

[7] A. Sandelin, *et al., Nucleic Acids Res.,* 32: W249 (2004) [PMID: 15215389]

[8] J. Wang, *et al., J Mol Biol.,* 286:315 (1999) [PMID: 9973553]

[9] http://www.cbrc.jp/research/db/TFSEARCH.html

[10] W. W. Wasserman, *et al., Nat Rev Genet.,* 5:276 (2004) [PMID: 15131651]

Edited by P. Kangueane

Citation: Su *et al.,* Bioinformation 1(5): 156-157 (2006)