Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Web Server**

D-MATRIX: A web tool for constructing weight matrix of conserved DNA motifs

Naresh Sen, Manoj Mishra, Feroz Khan*, Abha Meena, Ashok Sharma

Bioinformatics & In Silico Biology Division (BISB), Central Institute of Medicinal & Aromatic Plants (CIMAP), (Council of Scientific & Industrial Research), Lucknow-226015 (UP), India, Feroz Khan – Email: f.khan@cimap.res.in; *Corresponding author

Received December 20, 2008; revised March 02, 2009; accepted April 16, 2009; published July 27, 2009

Abstract:
Despite considerable efforts to date, DNA motif prediction in whole genome remains a challenge for researchers. Currently the genome wide motif prediction tools required either direct pattern sequence (for single motif) or weight matrix (for multiple motifs). Although there are known motif pattern databases and tools for genome level prediction but no tool for weight matrix construction. Considering this, we developed a D-MATRIX tool which predicts the different types of weight matrix based on user defined aligned motif sequence set and motif width. For retrieval of known motif sequences user can access the commonly used databases such as TFD, RegulonDB, DBTBS, Transfac. D-MATRIX program uses a simple statistical approach for weight matrix construction, which can be converted into different file formats according to user requirement. It provides the possibility to identify the conserved motifs in the co-regulated genes or whole genome. As example, we successfully constructed the weight matrix of LexA transcription factor binding site with the help of known *sos-box cis*-regulatory elements in *Deinococcus radiodurans* genome. The algorithm is implemented in C-Sharp and wrapped in ASP.Net to maintain a user friendly web interface. D-MATRIX tool is accessible through the CIMAP domain network.

Availability: http://203.190.147.116/dmatrix/

Keyword: Weight matrix, motif prediction, file format, motif databases

Background:
An important task in molecular biology is to identify DNA regulatory elements for transcription factors. These binding sites are short regions and called as 'motifs'. Despite considerable efforts to date, DNA motif finding in whole genome remains a challenge for researchers. There are several approaches to identify the conserved motifs but the recent one is through weight matrix based. So far no such tool is available to construct the different types of weight matrices according to user defined set. Earlier tools uses promoter sequences of co-regulated genes from single genome and search for statistically over-represented motifs. However, most of these motif finding tools have been shown to work successfully in yeast and other lower organisms, but perform significantly worse in higher organisms. Over the past few years, numerous tools have become available for the prediction of TF binding sites **[1-3]**. Especially popular are those tools which use information of known binding sites that are collected in databases such as TRANSFAC **[4]**, EpoDB **[5]**, TRANSCompel **[6]**. More sophisticated approaches include consideration of nucleotide correlation in different positions of the sites, HMM, taking into account flanking regions and others **[7-14]**. But usually, complex approaches require large training sets, which is rather problematic since, only small sets of binding patterns are known for a motif (*i.e.* up to 10 sites). Currently the genome wide motif prediction tools required either direct pattern sequence (for single motif) or weight matrix (for multiple motifs). Although there are known motif pattern databases and tools for genome wide prediction but no tool for weight matrix construction. Considering this, we have developed D-MATRIX tool which constructs the different types of weight matrices based on user defined motif sequences and width. D-MATRIX can use both orthologous and co-regulated genes upstream sequences as input data set. For demonstration, we used the known LexA transcription factor binding site of *Deinococcus radiodurans* (a radiation digestive bacterium), to construct the weight matrix similar to earlier reported one **[15]**. Predictions performance showed promising results, as on comparison of weight matrix with known one, we found 90% accuracy with aligned motifs of same width. D-MATRIX can generate different types of matrices *i.e.*, alignment, frequency and weight matrix. D-MATRIX also offers weight matrix conversion into different file formats as per user ease. These converted files can than be used as input files by genome wide motif prediction tools *e.g.* PoSSuMsearch **[16]** and RSAT-Patser **[17]**. Aligned motif sequences can be retrieved through available motif discovery tools *e.g.* SIGNAL SCAN **[7]**, MATRIX SEARCH **[8]**, MatInspector **[9]**, Fuzzy clustering tool **[10]**, FUNSITE **[11]**, Gibbs Sampling tool **[13]**, AliBaba2 **[14]** *etc*. D-MATRIX differs from existing tools by providing liberty to design user defined weight matrix model & signature.

Methodology:
D-MATRIX takes aligned DNA motif sequences 'N' and motif width 'w' as input, searches for nucleotide frequency at each position '$F_{(ij)}$' and outputs the found consensus patterns/motifs according to conservation priority based on nucleotide frequency '$F_{(ij)}$', constructed frequency matrix,

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 415-418 (2009)

alignment matrix and weight matrix along with motif signature and degenerate consensus sequence according to IUPAC/IUB convention. Scoring of the weight matrix was done through following equation **(see equation 1 in supplementary material)** as described elsewhere **[15, 18].**

Implementation:

The D-MATRIX web tool is implemented in C-Sharp and wrapped in ASP.Net to maintain a user friendly web interface. The D-MATRIX user interface is shown in snapshots (**Figure 1**). It has been designed so that the user has all necessary parameters available on one screen. The top panel is used to paste the input sequences (or aligned known TF binding sites) and to specify the name and width of motif to be search. The results panel contains five major sections: consensus pattern/motif sequence, frequency matrix, alignment matrix, weight matrix and signature sequence as per IUPAC code. Along with these results a tool for matrix transformation is also associated in right panel, which can transform the derived matrix according to input file format of various genomic motif discovery tools. Since input sequence set required is experimental one, thus all weight matrices constructed through D-MATRIX tool can be considered as a source of well supported hypotheses for further experimental verification.

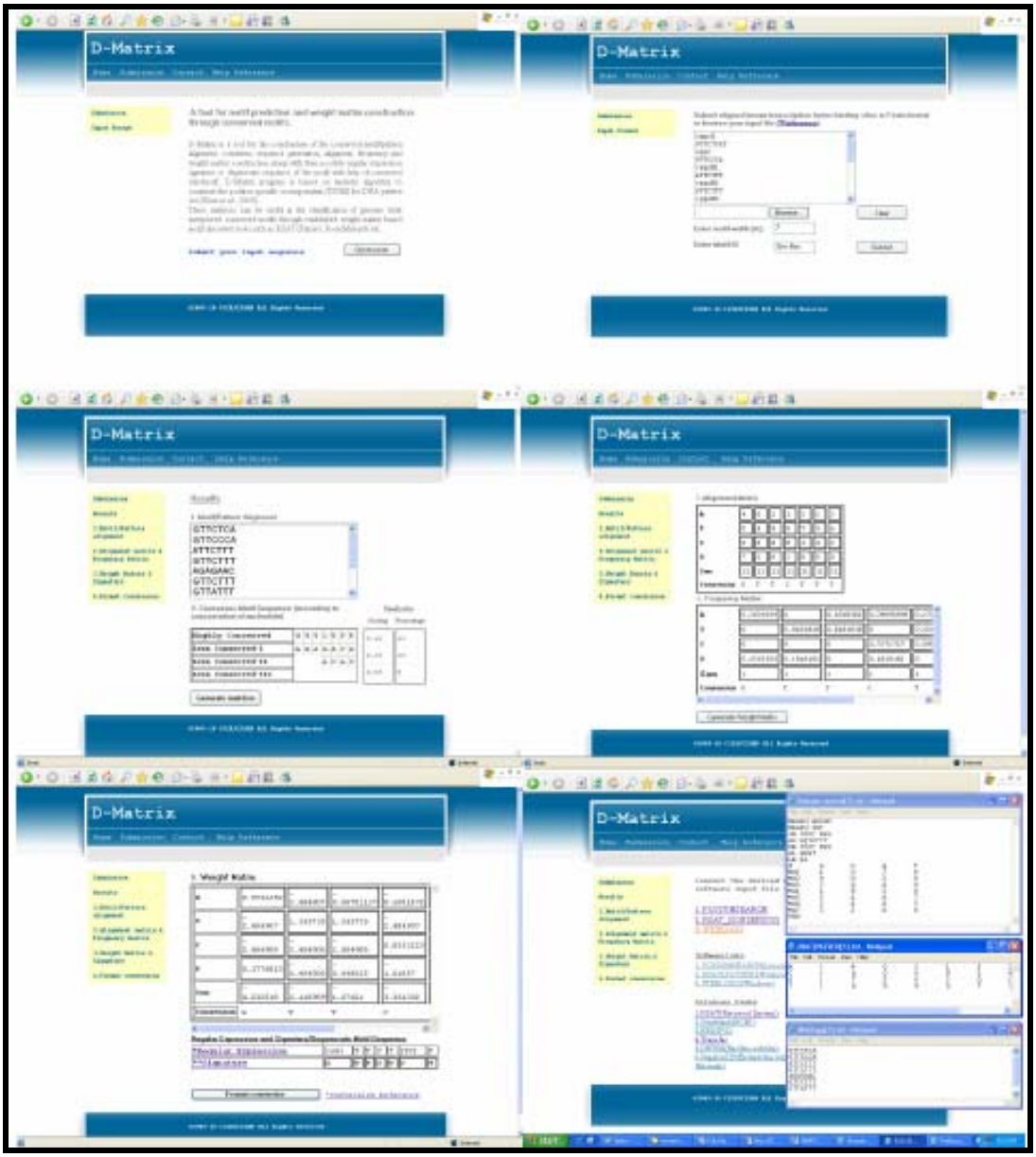

Figure 1: Snapshots of D-Matrix tool.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 415-418 (2009)

Acknowledgment:
We would like to thank experts of ICT division, CIMAP, for web hosting support. We also acknowledge Department of Biotechnology, New Delhi for financial support as Bioinformatics Centre at Central Institute of Medicinal & Aromatic Plants (Council of Scientific & Industrial Research), Lucknow (UP) INDIA.

References:

[1] GD Stormo, *Bioinformatics*, (2000) **16**:16–23.
[2] P Bucher, *Curr. Opin. Struct. Biol*., (1999) **9**:400–407.
[3] J.W. Fickett & W.W. Wasserman, *Curr. Opin. Biotechnol*., (2000) **11**:19–24.
[4] V. Matys, *et al.*, *Nucleic Acids Res*., (2003) **31**:374–378.
[5] C. J. Stoeckert, *et al.*, *Nucleic Acids Res*., (1999) **27**:200–203.
[6] O. V. Kel-Margoulis, *et al.*, *Nucleic Acids Res*., (2002) **30**:332–334.
[7] D. S. Prestridge, *Comput. Appl. Biosci*., (1996) **12**:157–160.
[8] Q. K. Chen, *et al.*, *Comput. Appl. Biosci*., (1995) **11**:563–566.
[9] K. Quandt, *et al.*, *Nucleic Acids Res*., (1995) **23**:4878–4884.
[10] L. Pickert, *et al.*, *Bioinformatics*, (1998) **14**:244–251.
[11] A. E. Kel, *et al.*, *Proc. Int. Conf. Intell. Syst. Mol. Biol*., (1995) **3**:197–205.
[12] M.D. Conkright, *et al.*, *Mol. Cell*, (2003) **11**:1101–1108.
[13] C. E. Lawrence, *et al.*, *Science*, (1993) **262**:208–214.
[14] N. Grabe, *In Silico Biol*., (2000) **1**:0019.
[15] F. Khan *et al.*, *Journal of Integrative Bioinformatics*, (2008) **5**(1):86.
[16] M. Beckstette, *et al.*, *BMC Bioinformatics*, (2006) **7**(1):389.
[17] G. Z. Hertz & SD Stormo, *Bioinformatics*, (1999) **15**:563–577.
[18] T. D. Schneider, *et al.*, *J. Mol. Biol*., (1986) **188**:415–431.

Edited by P. Kangueane
Citation: Sen *et al*, Bioinformation 3(10): 415-418 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 415-418 (2009)

Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Web Server**

Supplementary material

$$\mathbf{F_{(ij)} = \sum_{(w)} [C_{(ri)} / N]}$$

To calculate the nucleotide frequency '$F_{(ij)}$' within data set of known motif sequences 'N', program uses aligned set of motif sequences for calculating nucleotide repeat conservation '$C_{(ri)}$' at specific nucleotide position. In this equation, 'i' refers to nucleotide (e.g. A, T, G and C) at position 'j' in the motif width of 'w'. This approach finds optimized local alignments in related sequences in order to detect short conserved regions or motifs that may not be in the same positions. The matrix length can be set from 4 to 25, which allowed the detection of very short and also longer and more complex conserved sequences. Matrices generated, which represent the nucleotide conservation in each position of the motifs, can be used to search for the motifs positions, copy number and conservation using the RSAT-Patser (http://rsat.ulb.ac.be/rsat/) **[17]** and PoSSuMsearch **[16]** algorithms. The algorithm can be applied to the matrices by transforming it in required format. Moreover, user can convert the matrix into the input file format for Web-Logo program (http://weblogo.berkeley.edu/logo.cgi), so that to generate the visual representation of the conserved predicted motifs. For details see the web-demonstrated example of LexA transcription factor binding site weight matrix construction through D-Matrix web tool, for the prediction of *sos-box* in *D. radiodurans* whole genome **[15]**.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 415-418 (2009)