Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx223
Advance Access Publication Date: 13 April 2017
Applications Note

OXFORD

Sequence analysis

DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses

Jinyu Yang[1,2], Xin Chen[3], Adam McDermaid[1,2] and Qin Ma[1,2,4,5,*]

[1]Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture and Plant Science, [2]Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA, [3]Center for Applied Mathematics, Tianjin University, Tianjin 300000, China, [4]BioSNTR, Brookings, SD, USA and [5]Population Health group, Sanford Research, Sioux Falls, SD 57104, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on February 22, 2017; revised on April 5, 2017; editorial decision on April 10, 2017; accepted on April 12, 2017

Abstract

Motivation: Motif identification and analyses are important and have been long-standing computational problems in bioinformatics. Substantial efforts have been made in this field during the past several decades. However, the lack of intuitive and integrative web servers impedes the progress of making effective use of emerging algorithms and tools.
Results: Here we present an integrated web server, DMINDA 2.0, which contains: (i) five motif prediction and analyses algorithms, including a phylogenetic footprinting framework; (ii) 2125 species with complete genomes to support the above five functions, covering animals, plants and bacteria and (iii) bacterial regulon prediction and visualization.
Availability and Implementation: DMINDA 2.0 is freely available at http://bmbl.sdstate.edu/DMINDA2.
Contact: qin.ma@sdstate.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Regulatory DNA motifs (or *motifs*) are short, recurring patterns, which are usually composed of transcription factor binding sites (TFBSs) (D'haeseleer, 2006b). TFBSs play critical roles in regulating transcription rates and expression levels of their target genes. The knowledge of genome-scale TFBSs can greatly help the elucidation of gene regulatory mechanism in a cell (D'haeseleer, 2006a). Hence, *de-novo* motif identification and associated computational analyses (e.g. motif scanning and comparison) play an important role in regulatory network construction in all the organisms (Brohée *et al.*, 2011; Davidson and Levin, 2005). Despite a lot of algorithms and tools that have been proposed and developed in the past few decades, most mainly focused on motif identification without integrating associated motif analyses (Tompa *et al.*, 2005). Several web servers are available in the public domain, including the MEME Suite, PATLOC, AIMIE, Melina II, MotifSampler and STAMP (Bailey *et al.*, 2009; Mahony and Benos, 2007; Mrázek and Xie, 2006; Mrázek *et al.*, 2008; Okumura *et al.*, 2007; Thijs *et al.*, 2002). However, phylogenetic footprinting-based algorithms have not been fully considered. The identification and visualization of the relationship among identified motifs (or corresponding genes) remains unexplored. Hence, integrated web servers enabling reliable identification, comprehensive analyses and intuitive visualization of motifs are still needed.

We have developed an updated version of the DMINDA motif analysis web server (Ma *et al.*, 2014), DMINDA 2.0, which is available at http://bmbl.sdstate.edu/DMINDA2 and will be updated on a regular basis. Besides *de-novo* motif identification, motif scanning, motif comparison and motif co-occurrence analysis, DMINDA 2.0 integrates two newly-published algorithms (Liu *et al.*, 2016a,b), 2125 complete genome sequences, and visualization and interpretation functionalities. DMINDA 2.0 has several key features, namely,

(i) identification of motifs at a genome scale (for prokaryotes) along with estimated statistical significance values (Li *et al.*, 2011a); (ii) accurate scan for all motif instances of a query motif in specified genomic regions, and comparison and correlational analyses among the identified motifs to facilitate the inference of joint regulatory relationships among TFs (Ma *et al.*, 2013); (iii) 53 eukaryotic genomes downloaded from the Ensembl and JGI databases as of 01/12/2016 (including human, mouse and all fully sequenced plant genomes) and genome-scale operons for 2072 prokaryotes with complete genomes retrieved from the DOOR2 operon database (Mao *et al.*, 2014), in support of the above motif-based analysis; (iv) an integrative phylogenetic footprinting framework for *de-novo* motif identification in prokaryotic genomes based on a global orthologous gene mapping algorithm (Li *et al.*, 2011b; Liu *et al.*, 2016a); and (v) bacterial regulon (co-regulated operons by the same TF) prediction based on a new motif analysis framework and a novel graph model (Liu *et al.*, 2016b), along with a Cytoscape-like network interpretation and visualization function. A systematic comparison between DMINDA 2.0 and other six webservers indicates that DMINDA 2.0 and the MEME Suite can provide the most comprehensive motif finding and analysis functionalities (Fig. 1).

2 Functions and methods

There are six motif analysis functions in DMINDA 2.0 (Fig. 2A): (i) motif finding; (ii) motif scanning; (iii) motif comparison; (iv) motif co-occurrence analysis; (v) motif prediction by phylogenetic footprinting (namely *MP3*); and (vi) regulon prediction.

The input data for (i) and (v) are DNA sequences in the FASTA format; motif alignments (or their position weight matrices) are required for (ii), (iii) and (iv); and species name along with operon/gene IDs are needed in (vi). These input data can be uploaded manually or selected from our underlying database by users (Supplementary Example S1 and S2).

The outputs of each function are: (i) aligned motif instances along with their motif logos and related sequence details (Fig. 2B); (ii) query motif logo and identified motif instances (Supplementary Fig. S1); (iii) similarity score, heat-map and clustering tree of query motifs (Supplementary Fig. S2); (iv) identified co-occurrence motifs and their locational mapping to query genome sequences (Supplementary Fig. S3); (v) voting score curve and candidate binding regions along with same output in (i) (Supplementary Fig. S4) and (vi) identified regulons and their network visualization (Fig. 2C–E). All the outputs can be easily downloaded or converted for further computational analysis. The description of these six functions are shown below.

	Webservers	DMINDA 2.0	MEME Suite	PATLOC	AIMIE	Melina II	MotifSampler	STAMP
Database	Genome database for motif finding	√			√		√	
	Genome database for motif scanning	√	√	√		√		
	Built-in TFBS database		√					√
Motif finding	*De-novo* motif finding	√	√		√	√	√	
	Phylogenetic Footprinting Framework	√						
	Gapped-motif finding		√					
	Motif finding based on ChIP-seq		√					
Motif analyses	Motif scanning	√	√	√	√	√		
	Motif comparison	√	√					√
	Motif enrichment analysis	√	√					
	Motif co-occurrence analysis	√						
	Regulon prediction	√						
	Gapped-motif scanning		√					
Visualization	Cytoscape-like network for regulons	√						

Fig. 1. Comparison of DMINDA 2.0 and six motif analyses webservers. A check mark indicates that the corresponding functionality is provided by the specific webserver

i. *de novo motif finding* identifies a set of statistically significant motifs (if any) in a set of provided promoters (Supplementary Example S3). The backend algorithm, BOBRO (BOttleneck BROken) (Li *et al.*, 2011a), has been demonstrated on genome-scale datasets and does so more efficiently and accurately than the best available tools such as MEME (Bailey *et al.*, 2009).
ii. *motif scanning* scans for all motif instances of a query motif in given genomic sequences (Supplementary Example S4). The implemented tool, BBS (BoBro-based motif Scanning tool), has been shown to perform better than the MEME in accuracy on *E.coli* K12 and human genomes.
iii. *motif comparison* compares the similarity among the query motifs, and clusters similar motifs into groups (Supplementary Example S5). The implemented tool, BBC (BoBro-based motif Comparison and Clustering tool), identifies more accurate motif groups with a competitive sensitivity on synthetic datasets compared to MEME.
iv. *motif co-occurrence analysis* identifies co-occurring motifs which may regulate the same set of genes, in given regulatory sequences (Supplementary Example S6). The implemented tool, BBA (BoBro-based motif correlation Analysis tool), enables statistically significant TF pairs be identified among 12 561 pairs of *E.coli* K12, with some of them having been fully or partially proven in the published literature.

The integration of the phylogenetic footprinting strategy and the systematic combination of motif-associated analyses have been integrated into a phylogenetic footprinting framework for motif identification and bacterial regulon prediction in our server, respectively.

v. *MP3* identifies novel motifs (if any) in prokaryotic genomes based on an integrative phylogenetic footprinting framework (Supplementary Example S7). Compared with seven prevalent programs on *E. coli* K12 genomes, MP3 consistently achieved distinct improvement in motif identification accuracy. It mainly benefits from a new reference promoters preparation strategy, a promoter refining and pruning method and the integration of six widespread motif identification tools serving as a candidate TFBSs search engine (Fig. 2F).
vi. *regulon prediction* models and predicts regulons in given bacterial genomes (Supplementary Example S8). Evaluated through documented regulons and co-expressed modules derived from *E.coli*, this method outperforms other algorithms across a wide variety of experiments. This remarkable performance is mainly achieved through the use of a novel computational framework and a graph model, integrating motif identification, motif comparison and clustering (i.e. functions (i), (iii) and (v)). To intuitively illustrate the predicted regulons, a Cytoscape-like visualization method was also implemented in support of further studies (Supplementary Example S9).

3 Conclusions

Motif identification and analyses provide a solid foundation to infer gene regulatory mechanism in a genome. Our previously published studies showed that, compared to the best available tools such as MEME, our implemented methods can identify and analyze statistically significant motifs equally, sometimes even better at a genome scale. We believe that our web server provides a highly useful and easy-to-use platform for motif identification and analyses complementary to the existing web servers and tools, and benefits the genomic research community in general and prokaryotic genome researchers in particular.

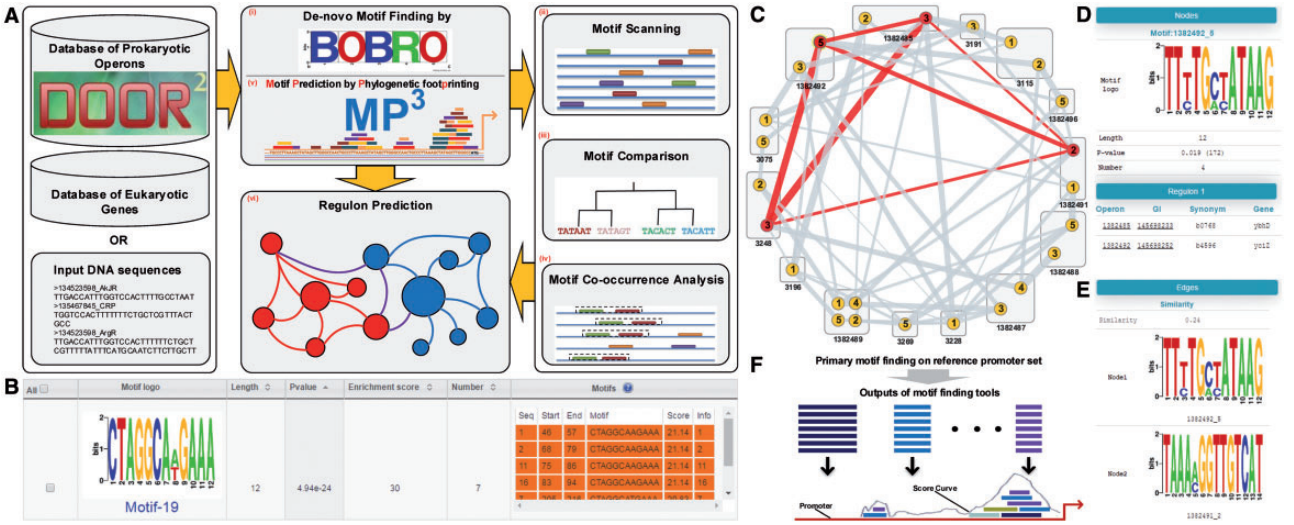

Fig. 2. (**A**) Workflow of DMINDA 2.0, including (i) *de-novo* motif finding, (ii) motif scanning, (iii) motif comparison, (iv) motif co-occurrence analysis, (v) *de-novo* motif finding based on phylogenetic footprinting strategy and (vi) regulon prediction; (**B**) A predicted motif and its output details; (**C**) A Cytoscape-like network visualization of predicted regulons. The rounded rectangles indicate operons, orange circles represent identified motifs, and network in red highlights the selected regulon; (**D**) Details of the selected node (1382492_5) in (C); (**E**) Details of the selected edge (1382492_5-1382491_2) in (C); and (**F**) The voting strategy in MP3 for generation of reliable TF binding regions

Acknowledgements

We thank Hanyuan Zhang, Kevin Brandt and Alan Carter for their technical assistance in web server development, Juan Xie, Yiran Zhang, Xijin Ge and Yanbin Yin for their assistance in web server testing.

Funding

This work was supported by the State of South Dakota Research Innovation Center, the Agriculture Experiment Station of South Dakota State University, and National Science Foundation of United States (1546869). This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

Conflict of Interest: none declared.

References

Bailey,T.L. *et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, gkp335.

Brohée,S. *et al.* (2011) Unraveling networks of co-regulated genes on the sole basis of genome sequences. *Nucleic Acids Res.*, gkr264.

D'haeseleer,P. (2006a) How does DNA sequence motif discovery work? *Nat. Biotechnol.*, **24**, 959–961.

D'haeseleer,P. (2006b) What are DNA sequence motifs? *Nat. Biotechnol.*, **24**, 423–425.

Davidson,E. and Levin,M. (2005) Gene regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 4935–4935.

Li,G. *et al.* (2011a) A new framework for identifying cis-regulatory motifs in prokaryotes. *Nucleic Acids Research*, **39**, e42.

Li,G. *et al.* (2011b) Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes. *Nucleic Acids Res.*, **39**, e150.

Liu,B. *et al.* (2016a) An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes. *BMC Genomics*, **17**, 578.

Liu,B. *et al.* (2016b) Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses. *Sci. Rep.*, **6**, doi:10.1038/srep23030.

Ma,Q. *et al.* (2013) An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. *Bioinformatics*, **29**, 2261–2268.

Ma,Q. *et al.* (2014) DMINDA: an integrated web server for DNA motif identification and analyses. *Nucleic Acids Res.*, gku315.

Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.

Mao,X. *et al.* (2014) DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res.*, **42**, D654–D659.

Mrázek,J. and Xie,S. (2006) Pattern locator: a new tool for finding local sequence patterns in genomic DNA sequences. *Bioinformatics*, **22**, 3099–3100.

Mrázek,J. *et al.* (2008) AIMIE: a web-based environment for detection and interpretation of significant sequence motifs in prokaryotic genomes. *Bioinformatics*, **24**, 1041–1048.

Okumura,T. *et al.* (2007) Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions. *Nucleic Acids Res.*, **35**, W227–W231.

Thijs,G. *et al.* (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, **9**, 447–464.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.