BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 23 2009, pages 3191–3193
doi:10.1093/bioinformatics/btp570

Systems biology

W-ChIPMotifs: a web application tool for *de novo* motif discovery from ChIP-based high-throughput data

Victor X. Jin[1,*], Jeff Apostolos[1], Naga Satya Venkateswara Ra Nagisetty[2] and Peggy J. Farnham[3]

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, [2]Bioinformatics Program, The University of Memphis, Memphis, TN 38152 and [3]The Genome Center, The University of California, Davis, CA 95616, USA

Received on August 3, 2009; Revised on September 28, 2009; Accepted on September 29, 2009

Advance Access publication October 1, 2009

Associate Editor: Martin Bishop

ABSTRACT

Summary: W-ChIPMotifs is a web application tool that provides a user friendly interface for *de novo* motif discovery. The web tool is based on our previous ChIPMotifs program which is a *de novo* motif finding tool developed for ChIP-based high-throughput data and incorporated various *ab initio* motif discovery tools such as MEME, MaMF, Weeder and optimized the significance of the detected motifs by using a bootstrap resampling statistic method and a Fisher test. Use of a randomized statistical model like bootstrap resampling can significantly increase the accuracy of the detected motifs. In our web tool, we have modified the program in two aspects: (i) we have refined the P-value with a Bonferroni correction; (ii) we have incorporated the STAMP tool to infer phylogenetic information and to determine the detected motifs if they are novel and known using the TRANSFAC and JASPAR databases. A comprehensive result file is mailed to users.

Availability: http://motif.bmi.ohio-state.edu/ChIPMotifs. Data used in the article may be downloaded from http://motif.bmi.ohio-state.edu/ChIPMotifs/examples.shtml.

Contact: victor.jin@osumc.edu

1 INTRODUCTION

DNA motifs are short sequences varying from 6 to 25 bp and can be highly variable and degenerated. Understanding how transcription factors usually selectively bind to these motifs is important for understanding the logic and mechanisms of gene regulation. One major approach is using position weight matrices (PWMs; Stormo *et al*., 1982) to represent information content of regulatory sites. However, when used as the sole means of identifying binding sites suffers from the limited amount of training data available (Roulet *et al*., 1998) and a high rate of false positive predictions (Tompa *et al*., 2005). Many *de novo* motif finding tools have been developed to detect these unknown motifs. Typical tools include hidden Markov models (Pedersen and Moult, 1996), Gibbs sampling (Lawrence *et al*., 1993), exhaustive enumeration (i.e. detecting the set of all nucleotide n-mers, then reporting the most frequent or overrepresented; e.g. Weeder (Pavesi *et al*., 2004), greedy alignment algorithms [e.g. CONSENSUS (Hertz and Stormo, 1999)], expectation-maximization (MEME) (Bailey and Elkan, 1995) and probabilistic mixture modeling (NestedMica; Down and Hubbard, 2005).

ChIP-based high-throughput techniques such as ChIP-chip (Ren *et al*., 2000; Weinmann *et al*., 2002), ChIP-seq (Barski *et al*., 2007; Robertson *et al*., 2007) and ChIP-PET (Loh *et al*., 2006) have been used to interrogate protein–DNA interactions in intact cells and is well-documented in many comprehensive reviews (Hanlon and Lieb, 2004). The identified enrichment DNA sequences usually ranging from ∼150 to ∼1500 bases from these techniques are currently considered to be highly reliable datasets for detecting the novel motif. Many computational tools including ours (Ettwiller *et al*., 2007; Gordon *et al*., 2005; Hong *et al*., 2005; Jin *et al*., 2007) have been recently developed to *de novo* find the motifs for the data generated from these techniques.

There exist many kinds of available computational tools. However, most of them are platform-dependent stand-alone executable programs, and not easily used by biologists. In this application, we have built a web-based *de novo* motif discovery tool for identifying novel motifs for ChIP-based high-throughput techniques. Although the web tool is based on our previous program, ChIPMotifs, we have significantly modified the program with a refined P-value computation using Bonferroni correction and incorporated a new STAMP tool (Mahony and Benos, 2007) to find the phylogenetic information and similar motifs in TRANSFAC (Wingender *et al*., 2000) and JASPAR (Sandelin *et al*., 2004) databases. The web interface is friendly and accessible by this research community.

2 DESCRIPTION OF W-ChIPMotifs

Usage of W-ChIPMotifs web service is simple and does not require any knowledge of the underlying software. The structure of W-ChIPMotifs is shown in Figure 1. There are three required inputs from the user: the DNA sequence data, contact information and a transcription factor name. DNA sequences are required to be in the FASTA format. They can be uploaded either by selecting an existing file, or by directly copying the data into the form. Results will be emailed to the address given in the contact information.

[*]To whom correspondence should be addressed.

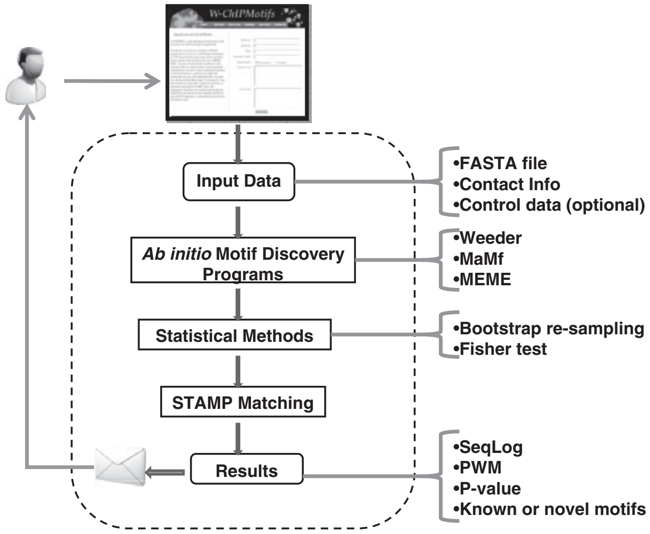

Fig. 1. A schematic view of W-ChIPMotifs.

The transcription factor name is used as a label in the results. Also, control data can be specified as an optional input, which is used to infer the statistical significance for detected motifs. In case of no control data input from users, we will use default control datasets where we randomly selected 5000 promoter sequences per run from all human or mouse promoter sequences depending on the user selected species.

After the server validates and retrieves the input, the DNA sequences are processed by a group of existing *ab initio* motif discovery programs. This group is currently composed of MEME (Bailey and Elkan, 1995), MaMF (Hon and Jain, 2006) and Weeder (Pavesi *et al*., 2004). These three are frequently used by the community, and have proven to be relatively accurate in detecting motifs. The programs are included in a modular fashion which enables the easy addition of other components in the future. Using these programs, we identified a set of n candidate motifs (usually <10 motifs), then constructed n PWMs for each candidate motif. A bootstrap resampling method is then used to infer the optimized PWM scores. In this method, a new dataset is created by randomizing the user input's sequences of each with 100 times. This new set no longer corresponds to the original ChIP identified binding sequences, but shares the same nucleotide frequencies and therefore can be used as a negative control set. The negative control is used for scanning the identified motifs at a minimal core score of 0.5 and a minimal PWM score of 0.5. Then, we retrieve core and PWM scores at the top 0.1, 0.5 and 1% percentiles. A Fisher test was applied and the P-value was used to define the significant cutoff for these scores. We also apply the Bonferroni correction by adjusting the P-value multiplying by the number of samples being input. If the adjusted P-value ended up >1.0, it would be rounded down to 1.0.

To provide users with more flexible and useful information about detected motifs, W-ChIPMotifs also uses the STAMP tool (Mahony and Benos, 2007) to determine if the motifs are known or novel by finding phylogenetic information and motif similarity matches in the TRANSFAC and JASPAR databases. Phylogenetic information implemented in STAMP tool is based on two tree-building algorithms: an agglomerative method and a divisive method. Both take input motifs' PWMs aligned by multiple alignment strategies, and iteratively build tree nodes until reaching each leaf node containing a single PWM.

The results from W-ChIPMotifs are composed of two files. The first file contains detected motifs with their SeqLOGOs, PWMs, core and PWM scores, P-values and Bonferroni correction P-value at different percentile levels. The second file contains matched similar motifs from the STAMP tool. These files are in PDF format.

In the future, we plan on adding more accurate and efficient motif detecting programs, and optimizing the running time of the statistical methods.

3 IMPLEMENTATION

W-ChIPMotifs is written in Perl, and uses a web interface developed with PHP. Multiple scripts are used to produce output from the included motif discovery programs, parse this output and apply statistical techniques. The sequence logos for the motifs are generated using the WEBLOGO tool (Crooks *et al*., 2004). The open-source HTMLDOC program is used to convert these logos to PDF format (http://www.htmldoc.org/). A tree from the newicks format is created with the DRAWTREE tool. The PHPGmailer package is used for sending results to the user from the W-ChIPMotifs email account.

4 SAMPLE TESTS

The W-ChIPMotif server is tested with different well-known datasets from the ChIP-seq and ChIP-chip experiments with different sizes of inputs. Some of such datasets include E2F4, FOXA1, NRSF and OCT4, the test data and results are available online at http://motif.bmi.ohio-state.edu/ChIPMotifs/examples.shtml.

ACKNOWLEDGEMENTS

We thank the members of our teams for discussion and comments.

Funding: Department of Biomedical Informatics, The Ohio State University.

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21–29.

Barski,A. *et al*. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Crooks,G. *et al*. (2004) WebLogo: A sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Down,T.A. and Hubbard,T.J. (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res.*, **33**, 1445–1453.

Ettwiller,L. *et al*. (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods*, **4**, 563–565.

Gordon,D.B. *et al*. (2005) TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics*, **21**, 3164–3165.

Hanlon,S. and Lieb,J. (2004) Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Curr. Opin. Genet. Dev.*, **14**, 697–705.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Hon,L.S. and Jain,A.N. (2006) A deterministic motif finding algorithm with application to the human genome. *Bioinformatics*, **22**, 1047–1054.

Hong,P. *et al*. (2005) A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics*, **21**, 2636–2643.

Jin,V.X. *et al*. (2007) Identication of cis-regulatory modules for OCT4 using de novo motif discovery and integrated computational genomics approaches. *Genome Res.*, **17**, 807–817.

Lawrence,C. *et al*. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Loh,Y.-H. *et al*. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, [Epub ahead of print, March 5, 2006]

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

Pavesi,G. *et al*. (2004) Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.

Pedersen,J.T. and Moult.,J. (1996) Genetic algorithms for protein structure prediction. *Curr. Opin. Struct. Biol.*, **6**, 227–231.

Ren,B. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Robertson,G. *et al*. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

Roulet,E. *et al*. (1998) Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. *In Silico Biol.*, **1**, 21–28.

Sandelin,A. *et al*. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Stormo,G.D. *et al*. (1982) Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.*, **10**, 2997–3011.

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

Weinmann,A.S., *et al*. (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.*, **16**, 235–244.

Wingender,E. *et al*. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.