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., 2002). 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.
A bootstrap resampling method is then used to infer the optimized motifs. We also apply the Bonferroni correction by adjusting the 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 newick format is created with the DRAWTREE tool. The PHP mailer 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 EP34, 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


