BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 5 2011, pages 715–717
doi:10.1093/bioinformatics/btq707

Genome analysis

Advance Access publication December 23, 2010

CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments

Lakshmi Kuttippurathu[1,†], Michael Hsing[1,2,†], Yongchao Liu[3], Bertil Schmidt[3], Douglas L. Maskell[3], Kyungjoon Lee[4], Aibin He[2], William T. Pu[2] and Sek Won Kong[1,2,*]

[1]Children's Hospital Informatics Program at the Harvard-MIT, Division of Health Sciences and Technology, Department of Medicine, Children's Hospital Boston, Boston, MA 02115, [2]Department of Cardiology, Children's Hospital Boston, Boston, MA 02115, USA, [3]School of Computer Engineering, Nanyang Technological University, Singapore 639798 and [4]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: CompleteMOTIFs (cMOTIFs) is an integrated web tool developed to facilitate systematic discovery of overrepresented transcription factor binding motifs from high-throughput chromatin immunoprecipitation experiments. Comprehensive annotations and Boolean logic operations on multiple peak locations enable users to focus on genomic regions of interest for *de novo* motif discovery using tools such as MEME, Weeder and ChIPMunk. The pipeline incorporates a scanning tool for known motifs from TRANSFAC and JASPAR databases, and performs an enrichment test using local or precalculated background models that significantly improve the motif scanning result. Furthermore, using the cMOTIFs pipeline, we demonstrated that multiple transcription factors could cooperatively bind to the upstream of important stem cell differentiation regulators.

Availability: http://cmotifs.tchlab.org

Contact: sekwon.kong@childrens.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 26, 2010; revised on November 28, 2010; accepted on December 14, 2010

1 INTRODUCTION

Gene expression is a complex process that is coordinately regulated by interactions of multiple transcription factors (TFs) and other proteins that form promoter and enhancer complexes. The recent development of next-generation sequencing technologies has made it possible to determine *in vivo* and *in vitro* bindings of diverse TFs on a genomic scale (i.e. ChIP-Seq). One immediate question is whether the peak regions are functionally associated with the transcriptional regulation of target genes. An indication of such functional associations is the presence of loci occupied by multiple TFs or highly conserved sequences across species (Wasserman and Sandelin, 2004). Hence, an important step towards understanding functional binding events is to study the presence of DNA sequence motifs in the context of nearby TFs, genomic features and epigenetic modifications.

There are several challenges in finding motifs in peak regions identified from TF binding experiments. In the case where only a subset of ChIP-Seq peaks has common motifs, *de novo* discovery tools might not be optimal. Scanning for known motifs is faster and capable of detecting motifs that appear less frequently; however, the relatively short (4–25 bp) nature of protein-binding motifs can produce non-significant results. Many tools such as GimmeMOTIFs (van Heeringen and Veenstra, 2010), Tmod (Sun *et al.*, 2010) and MoAn (Valen *et al.*, 2009) have been deployed to identify sequence motifs, while RSAT (Thomas-Chollier *et al.*, 2008) and Galaxy (Goecks *et al.*, 2010) provide more comprehensive means to analyse regulatory sequences in general. Each method has its own strength in identifying potential transcription factor binding site (TFBS); however, it is crucial to combine results from complementary tools (Tompa *et al.*, 2005) and provide an intuitive interface that streamlines peak annotation, filtering, motif discovery and summary.

CompleteMOTIFs (cMOTIFs) was developed to provide biologists an easy to use yet comprehensive web tool for ChIP-seq data analysis. First, peak regions are annotated with comprehensive genomic information such as known genes, conservation scores and repeating sequence elements. Second, Boolean logic operations such as intersection and union can combine multiple datasets from different TF bindings or histone modification states. Using the annotation and Boolean operations, users can filter and combine peak lists from one or multiple experiments, facilitating the study of cooperative TF binding events (Farnham, 2009). Third, combining the top 10 ranked motifs from three *de novo* methods and known motif scanning using background models helps to identify novel or known TFBS of interest as well as possible co-factor binding sites.

2 IMPLEMENTATION

The cMOTIFs pipeline was built using MySQL, Perl, CGI and *R* statistical language, and the overview of workflow is illustrated in Figure 1.

Input format: the pipeline accepts DNA sequences in FASTA format or genomic coordinates in BED or GFF formats. Format conversion tools are provided for human (hg18, hg19), mouse (mm9) and rat (rn4) genomes.

Genomic annotation: cMOTIFs provides annotations from UCSC Genome Bioinformatics including gene annotations (RefSeq and

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

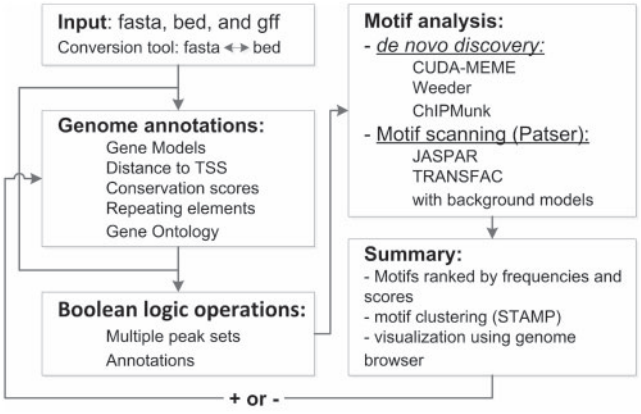

Fig. 1. Workflow of analysis pipeline. A subset of peak sequences based on the presence (+) or absence (−) of specific motifs can be easily defined for in-depth iterative analysis.

Ensembl), multispecies conservation scores (PhastCons) and repeating sequence elements (RepeatMasker), which enable users to select regions based on any combination of annotation criteria. The GOstats R library (Falcon and Gentleman, 2007) was integrated to the pipeline to summarize enriched Gene Ontology categories for the peak-associated genes. For instance, users can select peaks, which are highly conserved, contain no repeating sequence elements, and are located within 10 kb of transcription start sites of known genes, for subsequent motif discovery processes.

Boolean logic operations with multiple datasets: eight operations are provided to perform intersection, union, subtraction, common merge, common region, merge by distance, unique and extension on multiple genomic-region files (more details on the web site). For instance, to identify possible multiple transcription factor binding loci (MTL), users can merge neighbouring motif-enriched regions from different TF experiments.

Motif discovery: we incorporated three existing *de novo* discovery algorithms [MEME (Bailey and Elkan, 1994), Weeder (Pavesi *et al*., 2004) and ChIPMunk (Kulakovskiy *et al*., 2010)] and a motif scanning method [Patser (Hertz and Stormo, 1999)] with position-specific scoring matrices (PSSM) from databases including TRANSFAC (public version 7.0) (Matys *et al*., 2003) and JASPAR (version October 12, 2009) (Bryne *et al*., 2008) to prioritize overrepresented motifs. The pipeline also accepts user-defined PSSM for motif scanning. The Computer Unified Device Architecture (CUDA) library enhanced MEME performance (Liu *et al*., 2010). A suffix-tree-based exhaustive enumeration algorithm, Weeder (v1.4.2), and an iterative algorithm that combines greedy optimization with bootstrapping, ChIPMunk, have been included. The Patser [version 3b, (Hertz and Stormo, 1999)] is used to scan motifs from JASPAR and TRANSFAC databases. After the original Patser scanning, each target sequence is shuffled N times (default N = 1000) while maintaining (mono, di or tri) nucleotides frequency to create a random background model based on user input sequences or pre-compiled upstream sequences for each species. A permutation P-value is calculated from the frequency of random motif occurrences with the false discovery rate calculation (Storey and Tibshirani, 2003). This procedure allows estimating the significance of a motif occurrence in a target sequence compared with a set of randomly shuffled sequences. The current version of the pipeline allows up to a total of 500 000 bases for motif discovery, and this can be increased to 5 million bases if users choose to run ChIPMunk alone.

Motif summary: the top 10 motifs from the above four methods are ranked with corresponding scores and frequencies, and summarized with the STAMP tool by clustering motifs based on similarities (Mahony and Benos, 2007), and loci can be saved as BED format for further visual inspection and iterative analysis. Results are confidentially stored in users' personal accounts.

3 RESULTS AND DISCUSSION

We demonstrated our pipeline with a previously published ChIP-Seq dataset, which mapped binding locations of 13 TFs involved in embryonic stem cell differentiation (Chen *et al*., 2008). For 12 of the 13 TFs (except for E2f1), the pipeline successfully identified the reported binding motifs as the top-ranked motifs. Notably, the dinucleotide shuffling improved motif rankings from Patser (see Supplementary Material). The pipeline facilitates identification of multiple TF-binding loci. For instance, after processing the top 500 peaks from Nanog, Oct4 (encoded by Pou5f1 gene) and Sox2 TF profiles, we found a subset of peaks enriched with known binding motifs in JASPAR. Using Boolean logic, we merged neighbouring motif '+' regions within 500 bp and identified 35 MTL, each of which was occupied by Nanog, Oct4 and Sox2, suggesting collaborative binding from all three. These illustrate some examples of complex gene expression regulations that can be analysed efficiently with the pipeline (see Supplementary Material for the detail).

ACKNOWLEDGEMENTS

We thank Dr. Kulakovskiy for sharing ChIPMunk source codes for our pipeline.

Funding: Charles H. Hood Foundation and NHLBI (HL098166) to S.W.K.

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Bryne,J.C. *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.

Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.

Goecks,J. *et al*. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

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

Kulakovskiy,I.V. *et al.* (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.

Liu,Y. *et al.* (2010) CUDA-MEME: Accelerating motif discovery in biological sequences using CUDA-enabled graphics processing units. *Pattern Recognit. Lett.*, **31**, 2170–2177.

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

Matys,V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

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.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Sun,H. *et al.* (2010) Tmod: toolbox of motif discovery. *Bioinformatics*, **26**, 405–407.

Thomas-Chollier,M. *et al.* (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–W127.

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

Valen,E. *et al.* (2009) Discovery of regulatory elements is improved by a discriminatory approach. *PLoS. Comput. Biol.*, **5**, e1000562.

van Heeringen,S.J. and Veenstra,G.J. (2010) GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics*, [Epub ahead of print, November 15, 2010].

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.