Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw338
Advance Access Publication Date: 10 June 2016
Applications Note

OXFORD

Sequence analysis

RTFBSDB: an integrated framework for transcription factor binding site analysis

Zhong Wang[1], André L. Martins[1] and Charles G. Danko[1,2,*]

[1]Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA and [2]Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on February 9, 2016; revised on April 22, 2016; accepted on May 24, 2016

Abstract

Summary: Transcription factors (TFs) regulate complex programs of gene transcription by binding to short DNA sequence motifs. Here, we introduce rtfbsdb, a unified framework that integrates a database of more than 65 000 TF binding motifs with tools to easily and efficiently scan target genome sequences. Rtfbsdb clusters motifs with similar DNA sequence specificities and integrates RNA-seq or PRO-seq data to restrict analyses to motifs recognized by TFs expressed in the cell type of interest. Our package allows common analyses to be performed rapidly in an integrated environment.

Availability and Implementation: *rtfbsdb* available at (https://github.com/Danko-Lab/rtfbs_db).

Contact: dankoc@gmail.com

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

1 Introduction

Transcription factors (TFs) regulate complex programs of gene expression by modulating the rates of several steps early in transcription. TFs bind degenerate DNA sequence motifs, typically 3–20 bp, located in regulatory regions known as promoters, enhancers and insulators. Identifying the coordinates of TF binding motifs within the genome is a crucial step in many genomic analyses. However, motif discovery is a challenging computational problem owing to the short lengths and high degeneracy of TF binding motifs.

Using experimentally derived sources of TF binding is one strategy to improve accuracy by constraining the motif discovery problem to known binding sequences. This strategy requires extensive knowledge about the DNA sequence specificities of TFs, which have historically been time-consuming to measure experimentally. Recently, high throughput experimental approaches have allowed the systematic discovery of motifs for thousands of TFs (Jolma *et al.*, 2013; Mathelier *et al.*, 2014; Weirauch *et al.*, 2014). Moreover, strategies to impute binding motifs using TF amino-acid sequences extend these resources to most species with a sequenced genome (Weirauch *et al.*, 2014). These developments make the use of known TF binding motifs a powerful strategy in many common applications.

Here, we introduce *rtfbsdb*, an open-source pipeline for transcription factor binding site (TFBS) identification and analysis, which integrates experimentally derived TF binding data for thousands of TFs. Unlike other TFBS identification tools, *rtfbsdb* integrates high-throughput measurements of gene expression for TFs associated with each motif. For downstream TFBS scanning and identification, *rtfbsdb* uses the *rtfbs* package (Peterson *et al.*, unpublished), a highly flexible and efficient implementation of many TFBS scanning tasks. Many common and complex analyses can be solved by *rtfbsdb* in as little as a single line of R. We demonstrate *rtfbsdb* using genomic data from the ENCODE project.

2 Description

2.1 Description of the RTFBSDB package

The *rtfbsdb* package is an open-source package for R which automates many common tasks in TFBS discovery and analysis. The flowchart for a typical analysis using *rtfbsdb* is presented in Figure 1. To begin an analysis, users import a large database of experimentally defined TF binding motifs. We primarily support the Catalog of Inferred Sequence Binding Preferences (Cis-BP) database, which integrates more than 65 000 motifs from >25 distinct experimental

sources (Weirauch *et al.*, 2014). Cis-BP also includes imputed motifs for non-model organisms and for paralogs of well-characterized TFs. In addition to Cis-BP, *rtfbsdb* supports motifs from a variety of different sources and integrates data seamlessly in R.

DNA binding specificities are often highly similar between different TFs. Duplicate motifs reduce the interpretability of downstream analyses and can inappropriately decrease statistical power when using more stringent corrections for multiple hypothesis testing. We provide tools to focus analyses on motifs that are directly suitable for the user's application. Many analyses benefit from focusing on the subset of motifs for which the cognate TF is expressed in the cell type of interest. For this task, *rtfbsdb* integrates gene expression data collected using high-throughput sequencing approaches, including RNA-seq, PRO-seq, or related assays. *Rtfbsdb* estimates transcriptional activities of the TF associated with each motif in Cis-BP, which contains ENSEMBL IDs for each motif, using counting functions in the BedTools (Quinlan and Hall, 2010) and bigWig packages. *Rtfbsdb* implements a statistical test to identify TFs which are expressed in the cell type or tissue of interest relative to background (Core *et al.*, 2008). To remove remaining redundant entries from *rtfbsdb* we cluster motifs based on their DNA sequence specificities using an affinity propagation clustering algorithm provided in APCluster (Bodenhofer *et al.*, 2011). In most cases, downstream analyses will represent each cluster as a single motif which is a member of that cluster. After clustering, the similarities between each pair of motifs can be visualized as a heatmap (Supplementary Fig. S1) and images of the motif logos within each cluster can be visualized in R (Bembom, unpublished) or written to disk as a PDF file. The result of these pruning and clustering steps is to tailor the repertoire of motifs analyzed for the user's application.

After reading and filtering motifs, an *rtfbsdb* object can then be used to solve two classes of problem that are common in genomics.

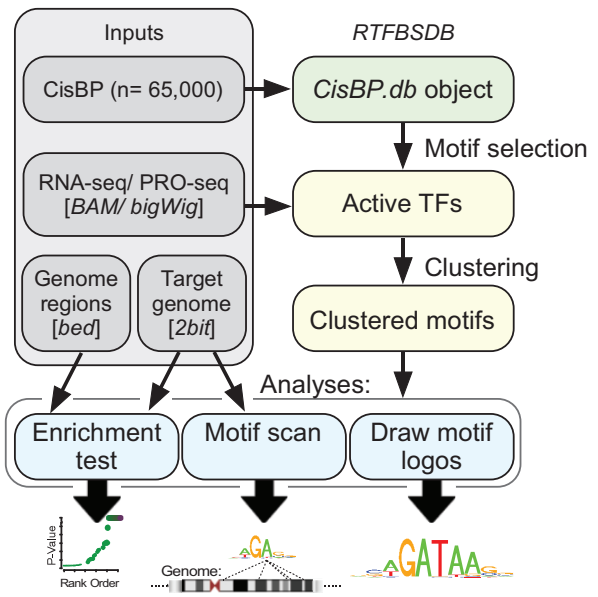

Fig. 1. Schematic illustrating the *rtfbsdb* workflow. Motifs are loaded into a *CisBP.db* object in R using an automated web scraper that imports data directly from the Cis-BP database. The set of motifs is reduced to those most applicable for analysis (right side) by removing TFs that are not expressed in the cell system of interest, and subsequently grouping motifs recognizing similar DNA sequences by clustering. The final set of motifs can be used to complete several common tasks in genomics (bottom row), including testing a set of DNA sequences for enriched motifs, scanning a target genome, or visualizing motifs as sequence logos

First, a common analysis is to identify the coordinates of known motifs in a target genome. During this task the user provides a target genome file (in UCSC 2-bit format) and *rtfbsdb* will return the coordinates of motif matches to the user. Although this challenge is addressed by FIMO (Grant *et al.*, 2011) and other applications, a notable advantage of *rtfbsdb* is that a database of experimentally derived motifs is integrated directly within the package. Moreover, our pipeline provides users with the option to write the coordinates of each motif directly to disk in a highly efficient compressed file format using bedops (Neph *et al.*, 2012), enabling thousands of motifs to be scanned efficiently across large genomes.

Second, another common analysis task is to identify motifs that are enriched in a common set of DNA sequences compared to background. This task can provide insight into which TFs putatively cause changes between two or more biological conditions. For example, differences in ATAC-seq or dREG peaks between two biological samples often reflect differences in TF binding. Motifs that are enriched in peaks that change between conditions relative to constitutive peaks can provide insight into which TFs are responsible for causing these changes. However, these analyses can be challenging to implement in practice. The most common challenge with such an analysis is a systematic difference in nucleotide composition between test and background sequences. By default, *rtfbsdb* uses a resampling approach to identify background sequences with a similar distribution of GC content as test sequences. Additionally, *rtfbsdb* identifies motifs that are robustly enriched at several motif match score cutoff thresholds. Together, these innovations result in more reliable discriminative TFBS identification. To our knowledge, HOMER is the only other package that allows discriminative TFBS identification using experimentally derived TF binding motifs (Heinz *et al.*, 2010). Compared to HOMER, *rtfbsdb* provides a larger repertoire of motifs, rigorously integrates TF expression levels using genomic data, and supports clustering motifs with similar DNA sequence specificities. Together, these advantages are likely to make *rtfbsdb* a more powerful and reliable tool for discriminative TFBS identification in many applications.

2.2 Using multiple GC content groups decreases accuracy

A common step in TFBS identification is to divide loci into separate groups based on GC content and use a separate background model for each group. This strategy is assumed to accommodate systematic differences in GC content across the genome, and thus improve the specificity of motif matches. However, whether this practice results in superior TF binding site predictions has not been tested directly. We created an empirical test using publicly available data from the ENCODE project to determine whether dividing sequences into multiple GC content groups improves the accuracy of TFBS predictions. We used motif match scores to classify high-confidence DNase-I hypersensitive sites (DHS), defined as the intersection of DHS discovered using Duke and UW assays (Danko et al., 2015), as bound or unbound to its cognate TF. Chromatin immunopercipitation and sequencing (ChIP-seq) peaks from 21 TFs were used as a gold-standard set. Surprisingly, a background model constructed using all available sequences performs more accurately than dividing sequences into four separate groups by GC content in almost 90% of cases (median AUC= 0.771 [1 GC group], 0.741 [4 GC groups] Supplementary Fig. S2; Supplementary Table S1). The largest differences were observed for ZBTB7A and E2F6 for which binding site discrimination was 12.9 and 11.1% more accurate with only one GC content group. Thus, we conclude that using a single GC

content group results in superior performance for the majority of TFs, though individual differences were relatively small. A single GC content group is the global default in *rtfbsdb*.

3 Example

To demonstrate the utility of *rtfbsdb* we used motifs in Cis-BP to search ChIP-seq peaks discovered by ENCODE for TFBS. We focused on ChIP-seq data profiling 97 TFs and co-factors in K562 cells. Forty-one of these are not represented by a motif in available databases, and these are largely comprised of either transcriptional co-repressors (e.g. HDAC1, EZH2 and KAP1) or general transcription factors (e.g. GTF3C2 and TAF1) without intrinsic sequence-specific DNA binding. Of the remaining 56 TFs the expected motif was recovered in 52 cases (93%), and was the most strikingly enriched motif in 41 (73%; Supplementary Table S2). For example, the motif corresponding to the transcriptional repressor REST was more than 50-fold enriched in ENCODE REST ChIP-seq peaks (Supplementary Fig. S3). In the 27% of cases where the expected motif was not the most enriched, *rtfbsdb* often recovered a motif corresponding to a known cofactor which may recruit the expected TF to ChIP-seq peaks by protein–protein interactions, in a process known as tethering. For example, peaks binding SP1 and SP2 were enriched for NFYA and NFYB binding motifs, which represent a known TF tethering interaction (Roder *et al.*, 1999). Similarly, although EP300 contains a motif in CisBP, it is a transcriptional co-activator which is recruited to DNA by other TFs (Goodman and Smolik, 2000). We thus conclude that *rtfbsdb* returns the expected motif in real world test cases.

Acknowledgements

We thank L. Choate for reading initial versions of the manuscript and documentation. Work in this publication was supported by NHLBI (National Heart, Lung and Blood Institute) UHL129958A to C.G.D. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

Conflict of Interest: none declared.

Funding

Work in this publication was supported by National Institutes of Health (NIH) and the National Heart, Lung and Blood Institute (NHLBI grant number: UHL129958A) to C.G.D. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

References

Bodenhofer,U. *et al.* (2011) APCluster: an R package for affinity propagation clustering. *Bioinformatics*, **27**, 2463–2464.

Core,L.J. *et al.* (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.

Danko,C.G. *et al.* (2015) Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods*, **12**, 433–438.

Goodman,R.H. and Smolik,S. (2000) CBP/p300 in cell growth, transformation, and development. *Genes Dev.*, **14**, 1553–1577.

Grant,C.E. *et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Heinz,S. *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

Jolma,A. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.

Mathelier,A. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.

Neph,S. *et al.* (2012) BEDOPS: high-performance genomic feature operations. *Bioinformatics*, **28**, 1919–1920.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Roder,K. *et al.* (1999) Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1. *Gene*, **234**, 61–69.

Wang,J. *et al.* (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.*, **41**, D171–D176.

Weirauch,M.T. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.