Genome analysis

**TFBSTools: an R/bioconductor package for transcription factor binding site analysis**

**Ge Tan and Boris Lenhard***

Computational Regulatory Genomics, MRC Clinical Sciences Centre, Imperial College London, London W12 0NN, UK

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received and revised on 10 November 2015; accepted on 13 January 2016

**Abstract**

**Summary:** The ability to efficiently investigate transcription factor binding sites (TFBSs) genome-wide is central to computational studies of gene regulation. **TFBSTools** is an R/Bioconductor package for the analysis and manipulation of TFBSs and their associated transcription factor profile matrices. **TFBSTools** provides a toolkit for handling TFBS profile matrices, scanning sequences and alignments including whole genomes, and querying the JASPAR database. The functionality of the package can be easily extended to include advanced statistical analysis, data visualization and data integration.

**Availability and implementation:** The package is implemented in R and available under GPL-2 license from the Bioconductor website (http://bioconductor.org/packages/TFBSTools/).

**Contact:** ge.tan09@imperial.ac.uk

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

---

1 Introduction

Transcription factor binding sites (TFBSs) on DNA play a central role in gene regulation via their sequence-specific interaction with transcription factor (TF) proteins (Wasserman and Sandelin, 2004). Most individual TFBSs are 4–30 base-pairs (bp) wide, but are generally located in larger cis-regulatory regions of 50–200 bp. Analysis and identification of TFBSs is crucial for understanding the regulatory mechanisms of gene regulation.

At present, the TFBS analysis functionality in R/Bioconductor (Gentleman et al., 2004) is limited and scattered across multiple packages. Here we introduce an R package **TFBSTools**, which provides a unified and efficiently implemented suite of TFBS analysis tools. The package provides a number of functions for manipulating TFBS profile matrices and searching DNA sequence and pairwise alignments using them. We have ported all of the functionality of our popular TFBS Perl modules (Lenhard and Wasserman, 2002), retaining the equivalent class structure where possible, and expanded the functionality to provide efficient genome-wide analysis of TFBSs. Our implementation is tightly integrated with the existing Bioconductor core packages, enabling high-performance sequence and interval manipulation. A database interface for JASPAR2014 (Mathelier et al., 2014), JASPAR2016 (Mathelier et al., 2015) and wrapper function for de novo motif discovery software are also provided.

2 Methods

2.1 S4 classes defined in TFBSTools

To provide easy data storage, manipulation and exchange, we created several novel S4 classes (Fig. 1), and also defined an aggregate version of each class (e.g. PFMatrixList) to help manipulate sets of the corresponding objects. The design of these classes corresponds to classes in TFBS Perl modules, while remaining extensible in an object-oriented manner, adding new functionality and taking advantage of functional programming capabilities of R.

2.2 Operations with TFBS matrix profiles

To characterize the binding preference of a TF, the aligned sequences bound by the TF are aggregated into a position frequency matrix (PFM). From this matrix, another two matrices can be derived: position weight matrix (PWM, the most commonly used kind of position-specific scoring matrix) and information content matrix (ICM). PWM is a matrix of positional log-likelihoods normally used for sequence scanning and scoring against the motif, while ICM is mostly used in motif visualization, e.g. for drawing sequence logos which can be easily done by the package seqLogo (Fig. 1A). As a novel feature, in addition to matrix profiles, TFBSTools also supports the manipulation of transcription factor...
2.3 Sequence/alignment scanning with PWM profiles

TFBSTools includes facilities for screening potential TFBSs present in a DNA sequence (searchSeq), or conserved in a pairwise alignment.

When a pairwise alignment is available, it can be used to combine the TFBSs prediction with phylogenetic footprinting, which can in many cases reduce the false discovery rate whilst retaining a sufficient level of sensitivity (Wasserman and Sandelin, 2004). Alternatively, it many cases reduce the false discovery rate whilst retaining a sufficient level of sensitivity (Wasserman and Sandelin, 2004). Alternatively, it many cases reduce the false discovery rate whilst retaining a sufficient level of sensitivity (Wasserman and Sandelin, 2004).

For genome-wide phylogenetic footprinting, TFBSTools can accept two BSgenome objects, and a chain file for liftover from one genome to another (searchPairBSgenome) or a novel S4 class Axt from our CNEr package (available from the Bioconductor website) for representing the axt alignments (searchAxt). It can take up to 50 CPU hours to run searchAxt on human-mouse pairwise alignment with the possibility of parallel computation, while searchSeq or searchPairBSgenome only needs several minutes. The computationally predicted putative TFBSs can be returned in GFF format or GRanges for downstream analysis.

2.4 JASPAR database interface

Since the release of JASPAR2014 (Mathelier et al., 2014), we have provided Bioconductor data packages, JASPAR2014 and JASPAR2016, holding the profile matrices and associated metadata. To accompany the use of this data package for TFBS analysis, TFBSTools provides functions to enable efficient database querying and manipulation.

2.5 Use of de novo motif discovery software

TFBSTools provides wrapper functions for de novo motif discovery softwares and seamlessly integrates the results back into R objects. Currently, support for MEME is implemented and reported motifs are stored in MotifSet object.

3 Conclusions and further information

The Bioconductor TFBSTools package provides a full suite of TFBS analysis tools. The package allows the efficient and reproducible identification and analysis of TFBSs. In combination with other functionality in Bioconductor, it provides a powerful way to analyze TF binding motifs on genome-wide scale. Further development will include an efficient implementation of scanning sequence/alignment with TFFM. A tutorial and additional use cases are available at Bioconductor website.

Acknowledgement

We thank Nathan Harmston for his comments on the manuscript.

Funding

G.T. is funded by the EU FP7 grant 242048 (ZF-HEALTH). B.L. is funded by Medical Research Council UK.

Conflict of Interest: none declared.

References

Gentleman,R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol., 5, R80.

Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. Bioinformatics, 18, 1133–1136.

Mathelier,A. and Wasserman,W.W. (2013) Integrated analysis of yeast regulatory sequences for transcription factor binding site prediction. PLoS Comput. Biol., 9, e1003214.

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.

Mathelier,A. et al. (2015) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res., gkv1176.

Nishida,K. et al. (2009) Pseudocounts for transcription factor binding sites. Nucleic Acids Res., 37, 939–944.

Sandelin,A. et al. (2003) Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. Funct. Integr. Genomics, 3, 125–134.

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