




# universalmotif: An R package for biological motif analysis

Benjamin Jean-Marie Tremblay <sup>1</sup>

<sup>1</sup> Independent Researcher, Spain

DOI: [10.21105/joss.07012](https://doi.org/10.21105/joss.07012)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Susan Holmes](#)  

## Reviewers:

- [@chrisamiller](#)
- [@ajank](#)

Submitted: 17 June 2024

Published: 19 August 2024

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

Sequence motifs are an important concept in molecular biology, as specific repeating patterns in DNA, RNA and proteins form the basis of biological regulation. Identifying and characterizing these motifs is therefore a important part of studying various aspects of cellular processes, such as gene regulation, transcript stability, and protein function. Many programs have been developed over the years to tackle these tasks, though their interoperability remains poor. The universalmotif package has two main goals: to serve as a go-between for most common biological motif programs and Bioconductor packages used by the research community, and to provide a robust set of tools for basic motif analysis and manipulation in R. Tools for motif and sequence manipulation, scanning, enrichment, comparison, shuffling and P-value computation are included.

## Installation

The universalmotif project including its extensive documentation are hosted on [Bioconductor](#), with pre-built binaries available for macOS and Windows (and installation from source available for all platforms). Installation takes place from within R using the BiocManager package, which itself can be installed from [CRAN](#):

```
install.packages("BiocManager")
BiocManager::install("universalmotif")
```

## Statement of need

Identifying and characterizing biological sequence motifs is an important task in the field of molecular biology, especially for the understanding of gene, transcript and protein regulation. Over the year many programs have been created to tackle this, such as the hugely popular MEME suite ([Bailey & Elkan, 1994](#)) and the TFBSTools R/Bioconductor package ([Tan & Lenhard, 2016](#)), as well as various curated databases such as JASPAR ([Rauluseviciute et al., 2024](#)) containing thousands of published motifs maintained by the community. While collectively these efforts have been responsible for significant advances in the field, the proliferation of different formats and the resulting poor interoperability hinders their use. To solve this problem, the universalmotif R/Bioconductor package allows for the import and export of a large number of commonly used motif formats, including CIS-BP ([Weirauch et al., 2014](#)), HOMER ([Heinz et al., 2010](#)), JASPAR ([Rauluseviciute et al., 2024](#)), HOCOMOCO ([Vorontsov et al., 2024](#)), TRANSFAC ([Wingender et al., 1996](#)), UniPROBE ([Hume et al., 2015](#)), and any additional simple formats via the `read_matrix()` and `write_matrix()` functions. Furthermore, existing R/Bioconductor packages providing their own motif classes can be converted to and from the universal universalmotif motif class via the `convert_motifs()` function, which include the popular TFBSTools ([Tan & Lenhard, 2016](#)) and motifStack ([Ou et al., 2018](#)) packages,

among others. While other R/Bioconductor motif-related packages often provide functions to import external motif formats (such as `importMatrix()` from `motifStack`), their other functions still typically cannot be used with motif classes from different packages. By allowing for all R/Bioconductor packages to interoperate via the `universalmotif` class, the `universalmotif` package provides a way for users to pick and choose functions from all available R/Bioconductor packages. Various other projects have now made use of this extensive compatibility and flexible motif class, such as `memes` (Nystrom & McKay, 2021), `CollecTRI` (Müller-Dott et al., 2023), `circRNAprofiler` (Aufiero et al., 2020), `ASTK` (Huang et al., 2024), and the standalone RSAT matrix-clustering tool (Castro-Mondragon et al., 2017). The `universalmotif` project has been continuously developed over six years and will continue to add more formats and classes when requested by the community.

The `universalmotif` package also provides a suite of functions for working with motifs and biological sequences, giving researchers the ability to perform most motif-related tasks from within R (which has been embraced by a large number of molecular biologists as the programming environment of choice for bioinformatic analyses). These include functions for manipulating motifs themselves (`create_motif()`, `convert_type()`, `filter_motifs()`, `motif_rc()`, `switch_alph()`, `trim_motifs()`), comparison and merging (`compare_motifs()`, `merge_motifs()`, `merge_similar()`), plotting (`view_motifs()`), motif P-values (`motif_pvalue()`), and sequence scanning, enrichment, and manipulation (`scan_sequences()`, `enrich_motifs()`, `create_sequences()`, `get_bkg()`, `shuffle_sequences()`, `sequence_complexity()`). Many additional utilities are included, all of which are extensively documented. No other R/Bioconductor motif-related package offers such a large set of functions for working with motifs, though for packages which contain specialized methods not provided by the `universalmotif` package (such as the advanced motif plotting of the `motifStack` package), they can be used seamlessly alongside the `universalmotif` package. These functions from the `universalmotif` package have now seen widespread use by researchers over the past several years, as evidenced by their appearance in a wide range of journals. For example: motif import into R (Li et al., 2023), motif comparison and merging (Hoge et al., 2024; Jores et al., 2021; Najle et al., 2023), sequence scanning and motif enrichment (Hawkins et al., 2024; Jores et al., 2021; Mikl et al., 2022), and motif plotting (Gao et al., 2024; Meeuse et al., 2023; Zeng et al., 2022). Future developments of the `universalmotif` project are aimed to increase the available functionality of the package.

## Acknowledgements

This software has been developed and maintained without any financial support. I am grateful to present and past mentors for encouraging me to pursue this project, including Barbara Moffatt, Andrew Doxey, and Julia Qüesta.

## References

- Aufiero, S., Reckman, Y. J., Tijssen, A. J., Pinto, Y. M., & Creemers, E. E. (2020). `circRNAprofiler`: An r-based computational framework for the downstream analysis of circular RNAs. *BMC Bioinformatics*, 21(1), 164. <https://doi.org/10.1186/s12859-020-3500-3>
- Bailey, T. L., & 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.
- Castro-Mondragon, J. A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., & Helden, J. van. (2017). RSAT matrix-clustering: Dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Research*, 45(13), e119. <https://doi.org/10.1093/nar/gkx314>
- Gao, L., Behrens, A., Rodschinka, G., Forcelloni, S., Wani, S., Strasser, K., & Nedialkova, D. D.

- (2024). Selective gene expression maintains human tRNA anticodon pools during differentiation. *Nature Cell Biology*, 26(1), 100–112. <https://doi.org/10.1038/s41556-023-01317-3>
- Hawkins, S., Mondaini, A., Namboori, S. C., Nguyen, G. G., Yeo, G. W., Javed, A., & Bhinge, A. (2024). ePRINT: Exonuclease assisted mapping of protein-RNA interactions. *Genome Biol*, 25(1), 140. <https://doi.org/10.1186/s13059-024-03271-1>
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol Cell*, 38(4), 576–589. <https://doi.org/10.1016/j.molcel.2010.05.004>
- Hoge, C., Manuel, M. de, Mahgoub, M., Okami, N., Fuller, Z., Banerjee, S., Baker, Z., McNulty, M., Andolfatto, P., Macfarlan, T. S., Schumer, M., Tzika, A. C., & Przeworski, M. (2024). Patterns of recombination in snakes reveal a tug-of-war between PRDM9 and promoter-like features. *Science*, 383(6685), eadj7026. <https://doi.org/10.1126/science.adj7026>
- Huang, S., He, J., Yu, L., Guo, J., Jiang, S., Sun, Z., Cheng, L., Chen, X., Ji, X., & Zhang, Y. (2024). ASTK: A machine learning-based integrative software for alternative splicing analysis. *Advanced Intelligent Systems*, 6(4), 2300594. <https://doi.org/10.1002/aisy.202300594>
- Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., & Bulyk, M. L. (2015). UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res*, 43, D117–122. <https://doi.org/10.1093/nar/gku1045>
- Jores, T., Tonnies, J., Wrightsman, T., Buckler, E. S., Cuperus, J. T., Fields, S., & Queitsch, C. (2021). Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat. Plants*, 7(6), 842–855. <https://doi.org/10.1038/s41477-021-00932-y>
- Li, M., Yao, T., Lin, W., Hinckley, W. E., Galli, M., Muchero, W., Gallavotti, A., Chen, J.-G., & Huang, S. C. (2023). Double DAP-seq uncovered synergistic DNA binding of interacting bZIP transcription factors. *Nat Commun*, 14(1), 2600. <https://doi.org/10.1038/s41467-023-38096-2>
- Meeuse, M. W. M., Hauser, Y. P., Nahar, S., Smith, A. A. T., Braun, K., Azzi, C., Rempfler, M., & Großhans, H. (2023). C. Elegans molting requires rhythmic accumulation of the grainyhead/LSF transcription factor GRH-1. *The EMBO Journal*, 42(4), e111895. <https://doi.org/10.15252/embj.2022111895>
- Mikl, M., Eletto, D., Nijim, M., Lee, M., Lafzi, A., Mhamedi, F., David, O., Sain, S. B., Handler, K., & Moor, A. E. (2022). A massively parallel reporter assay reveals focused and broadly encoded RNA localization signals in neurons. *Nucleic Acids Research*, 50(18), 10643–10664. <https://doi.org/10.1093/nar/gkac806>
- Müller-Dott, S., Tsirovouli, E., Vazquez, M., Ramirez Flores, R. O., Badia-i-Mompel, P., Fallegger, R., Türei, D., Lægreid, A., & Saez-Rodriguez, J. (2023). Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, 51(20), 10934–10949. <https://doi.org/10.1093/nar/gkad841>
- Najle, S. R., Grau-Bové, X., Elek, A., Navarrete, C., Cianferoni, D., Chiva, C., Cañas-Armenteros, D., Mallabiabarrena, A., Kamm, K., Sabidó, E., Gruber-Vodicka, H., Schierwater, B., Serrano, L., & Sebé-Pedrós, A. (2023). Stepwise emergence of the neuronal gene expression program in early animal evolution. *Cell*, 186(21), 4676–4693.e29. <https://doi.org/10.1016/j.cell.2023.08.027>
- Nystrom, S. L., & McKay, D. J. (2021). Memes: A motif analysis environment in r using tools from the MEME suite. *PLoS Comput Biol*, 17(9), e1008991. <https://doi.org/10.1371/journal.pcbi.1008991>

- Ou, J., Wolfe, S. A., Brodsky, M. H., & Zhu, L. J. (2018). motifStack for the analysis of transcription factor binding site evolution. *Nat Methods*, 15(1), 8–9. <https://doi.org/10.1038/nmeth.4555>
- Rauluseviciute, I., Riudavets-Puig, R., Blanc-Mathieu, R., Castro-Mondragon, J. A., Ferenc, K., Kumar, V., Lemma, R. B., Lucas, J., Chèneby, J., Baranasic, D., Khan, A., Fornes, O., Gundersen, S., Johansen, M., Hovig, E., Lenhard, B., Sandelin, A., Wasserman, W. W., Parcy, F., & Mathelier, A. (2024). JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 52, D174–D182. <https://doi.org/10.1093/nar/gkad1059>
- Tan, G., & Lenhard, B. (2016). TFBSTools: An R/bioconductor package for transcription factor binding site analysis. *Bioinformatics*, 32(10), 1555–1556. <https://doi.org/10.1093/bioinformatics/btw024>
- Vorontsov, I. E., Eliseeva, I. A., Zinkevich, A., Nikonov, M., Abramov, S., Boytsov, A., Kamenets, V., Kasianova, A., Kolmykov, S., Yevshin, I. S., Favorov, A., Medvedeva, Y. A., Jolma, A., Kolpakov, F., Makeev, V. J., & Kulakovskiy, I. V. (2024). HOCOMOCO in 2024: A rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Research*, 52, D154–D163. <https://doi.org/10.1093/nar/gkad1077>
- Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., Bakel, H. van, Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., ... Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6), 1431–1443. <https://doi.org/10.1016/j.cell.2014.08.009>
- Wingender, E., Dietze, P., Karas, H., & Knüppel, R. (1996). TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 24(1), 238–241. <https://doi.org/10.1093/nar/24.1.238>
- Zeng, Y., Fair, B. J., Zeng, H., Krishnamohan, A., Hou, Y., Hall, J. M., Ruthenburg, A. J., Li, Y. I., & Staley, J. P. (2022). Profiling lariat intermediates reveals genetic determinants of early and late co-transcriptional splicing. *Mol Cell*, 82(24), 4681–4699.e8. <https://doi.org/10.1016/j.molcel.2022.11.004>