bioRxiv preprint doi: https://doi.org/10.1101/2021.09.02.458722; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

XSTREME: Comprehensive motif analysis of biological sequence datasets

Charles E. Grant[1] and Timothy L. Bailey[2,*]

1) Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065
2) Department of Pharmacology, University of Nevada, Reno, NV 89557, USA

* timothybailey@unr.edu

Abstract

XSTREME is a web-based tool for performing comprehensive motif discovery and analysis in DNA, RNA or protein sequences, as well as in sequences in user-defined alphabets. It is designed for both very large and very small datasets. XSTREME is similar to the MEME-ChIP tool, but expands upon its capabilities in several ways. Like MEME-ChIP, XSTREME performs two types of *de novo* motif discovery, and also performs motif enrichment analysis of the input sequences using databases of known motifs. Unlike MEME-ChIP, which ranks motifs based on their enrichment in the *centers* of the input sequences, XSTREME uses enrichment *anywhere* in the sequences for this purpose. Consequently, XSTREME is more appropriate for motif-based analysis of sequences regardless of how the motifs are distributed within the sequences. XSTREME uses the MEME and STREME algorithms for motif discovery, and the recently developed SEA algorithm for motif enrichment analysis. The interactive HTML output produced by XSTREME includes highly accurate motif significance estimates, plots of the positional distribution of each motif, and histograms of the number of motif matches in each sequences. XSTREME is easy to use via its web server at `https://meme-suite.org`, and is fully integrated with the widely-used MEME Suite of sequence analysis tools, which can be freely downloaded at the same web site for non-commercial use.

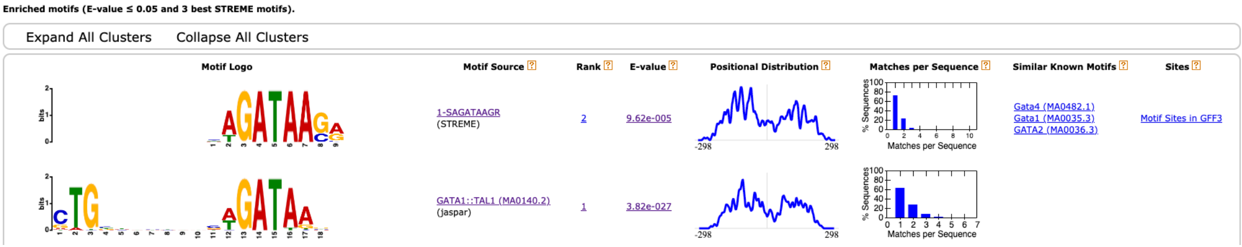
1 Introduction

Short, approximate sequence patterns (motifs) are known to encode functional biological signals in DNA, RNA and protein sequences. In genomic DNA, motifs capture the preferred binding sites of transcription factors (TFs) and promoter elements. In RNA motifs describe the binding preferences of RNA-binding proteins (RBPs). Motifs can also represent many protein features such as the targets of enzymes.

XSTREME is a web-based tool for comprehensive motif-based sequence analysis of DNA, RNA or protein data sets. It provides computationally efficient algorithms for discovering and analyzing the sequence motifs. Given a set of biological sequences, XSTREME first executes two different motif discovery algorithms: MEME (multiple EM for motif elicitation) [2], and STREME (Sensitive, Thorough, Rapid Enriched Motif Elicitation) [1] to discover novel sequence motifs. Next it then uses a motif enrichment analysis algorithm, SEA (Simple Enrichment Analysis) [3] to detect enrichment of previously characterized functional motifs and to rank the enrichment of discovered and known motifs on the same scale. Finally, to ease interpretation of the results, XSTREME applies a clustering algorithm to group the discovered and enriched motifs by similarity to each other. XSTREME returns its results as an interactive HTML document that provides visual representations of each motif and its locations within the input sequences, as well as estimates of the statistical significance of its enrichment. All results are presented in groups sorted according to statistical significance, and the HTML document provides clickable links to all details of the individual analyses. The XSTREME web-server provides numerous databases of known motifs for use in the motif enrichment analysis step, including motifs associated with given TFs (e.g, the JASPAR database [5]), motifs bound by RBPs (e.g., the CISBP-RNA database [9]), and protein motifs (ELM database [7]).

The XSTREME algorithm is similar to the existing MEME-ChIP [8], but XSTREME is applicable to a wider range of motif analysis problems. The primary distinction is that XSTREME makes no assumptions about the positional distribution of motif

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.02.458722; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Enriched motifs (E-value ≤ 0.05 and 3 best STREME motifs).

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	1-SAGATAAGR (STREME)	2	9.62e-005			Gata4 (MA0482.1) Gata1 (MA0035.3) GATA2 (MA0036.3)	Motif Sites in GFF3
	GATA1::TAL1 (MA0140.2) (jaspar)	1	3.82e-027				

Figure 1: **Screenshot of (a portion of) the HTML output of XSTREME run on Gata1 Cut&Run in erythroid precursor cell data.**

instances in the input sequences. XSTREME ranks motifs by how enriched they are in a primary set of sequences compared with a control set. (XSTREME will create a control set by shuffling the primary sequences if no control set is provided.) In contrast, MEME-ChIP ranks motifs by how enriched they are in the central regions of the input sequences. When it is believed (or known) that motifs may not be concentrated centrally in the sequences, XSTREME will provide more useful results than MEME-ChIP. Types of datasets where XSTREME is more appropriate than MEME-ChIP include sets of promoters, sets of accessible chromatin regions from ATAC-seq [4] and Cut&Run datasets using TF antibodies [6]. In each of these cases the assumption that motif motif sites will be near the centers of the input sequences does not always hold.

2 Results

Here, we illustrate using XSTREME for motif analysis of a Cut&Run dataset for the transcription factor Gata1 in erythroid precursor cells [11]. Cut&Run is an antibody-targeted cleavage method for identifying protein binding to chromatin from as few as 1000 cells [10]. In some experiments, however, Cut&Run results in a majority of bound sequences where the TF was near one end. With such datasets, XSTREME is a better fit than MEME-ChIP, which assumes that motifs tend to concentrate around the midpoint of the input sequences.

Fig. 1 shows the top motif cluster reported by XSTREME on the sequences with lengths from 400bp to 600bp specified in the Zhu *et al.* 2019 [11] dataset contained in file GSM4043375_GATA1_D7_S11_peaks.narrowPeak.gz (GEO accession number GSM4043375). The bimodal distribution of the Gata1 binding sites is clearly apparent in the XSTREME output. The output shows that XSTREME has discovered a close match to the known Gata1 motif (the STREME motif 1-SAGATAAGR), and has associated that motif with the known Gata1 motif (MA0035.3) from the JASPAR database of motifs [5]. XSTREME also clusters both discovered and known motifs, and the known motif for the GATA1-TAL1 is shown aligned with the STREME motif. The figure also illustrates how XSTREME reports a histogram of the predicted number of motif matches per input sequence with at least one match. In this example, approximately 25% of the input sequences with one or more sites have multiple sites.

3 Funding

This work was supported by NIH award R01 GM103544.

References

[1] T. L. Bailey. STREME: accurate and versatile sequence motif discovery. *Bioinformatics (Oxford, England)*, Mar. 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.02.458722; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[2] T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, Cambridge, United Kingdom, July 16-19, 1995*, 3:21–29, 1995.

[3] T. L. Bailey and C. E. Grant. SEA: Simple Enrichment Analysis of motifs. *bioRxiv*, 2021.

[4] M. Bysani, R. Agren, C. Davegårdh, P. Volkov, T. Rönn, P. Unneberg, K. Bacos, and C. Ling. Author correction: ATAC-seq reveals alterations in open chromatin in pancreatic islets from subjects with type 2 diabetes. *Scientific reports*, 10:1744, Jan. 2020.

[5] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, W. Santana-Garcia, G. Tan, J. Chèneby, B. Ballester, F. Parcy, A. Sandelin, B. Lenhard, W. W. Wasserman, and A. Mathelier. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, Nov. 2019.

[6] D. H. Janssens, S. J. Wu, J. F. Sarthy, M. P. Meers, C. H. Myers, J. M. Olson, K. Ahmad, and S. Henikoff. Automated in situ chromatin profiling efficiently resolves cell types and gene regulatory programs. *Epigenetics & chromatin*, 11:74, Dec. 2018.

[7] M. Kumar, M. Gouw, S. Michael, H. Sámano-Sánchez, R. Pancsa, J. Glavina, A. Diakogianni, J. A. Valverde, D. Bukirova, J. Čalyševa, N. Palopoli, N. E. Davey, L. B. Chemes, and T. J. Gibson. ELM-the eukaryotic linear motif resource in 2020. *Nucleic acids research*, 48:D296–D306, Jan. 2020.

[8] P. Machanick and T. L. Bailey. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697, Jun 2011.

[9] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, Jul 2013.

[10] P. J. Skene, J. G. Henikoff, and S. Henikoff. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature protocols*, 13:1006–1019, May 2018.

[11] Q. Zhu, N. Liu, S. H. Orkin, and G.-C. Yuan. CUT&RUNTools: a flexible pipeline for CUT&RUN processing and footprint analysis. *Genome biology*, 20:192, Sept. 2019.