MoCA: Tool for Motif Conservation Analysis

Saket Choudhary*and Anton Valouev †

June 2, 2016

Website: http://saketkc.github.io/moca/
Repository: https://github.com/saketkc/moca
License: ISC license

Motif discovery tools for inferring transcription factor binding sites(TFBS) often predict multiple motifs. However, determining the quality of a reported motif is often difficult. Motifs discovered by tools such such as MEME[1] and other *de-novo* motif finders can often be 'false positives', even if they are reported to have have significant E-values.

An approach that has often been used and has proven to improve the accuracy of predictions involves phylogenetic footprinting. TFBS can be expected to be conserved evolutionarily, since they are functional sequences and hence will tend to evolve slowly as compared to the neighboring sequences. We hence expect that a 'true' motif would exhibit high PhastCons[3] and Gerp[2] scores.

We developed MoCA, a tool to perform conservation analysis of reported motifs. MoCA makes use of GERP and PhastCons scores to assess the conservation profile of TFBS and compares it with neighboring sequence. Input to MoCA is a bed file indicating ChIP-seq peak summits. The summits are then slopped to obtain flanking sequences. MoCA makes use of MEME to perform *de-novo* motif discovery. The motif logos are then plotted against their conservation scores. Other diagnostic plots include a regression plot indicating the correlation between most frequent base frequency at each position with its conservation score as compared to a motif found using randomly selected sequences(GC-controlled regions). MoCA also provides a centered-enrichment plot indicating how far the motif lies from the peak summit and the relevant statistics to asses the quality of reported motif. A motif representing 'higher' conservation scores as compared to its neighboring sequence can be expected to be a 'true' motif.

MoCA is different from other probabilistic approaches for inferring TFBS that take into account phylogenetic conservation for motif discovery as it provides more of a diagnostic framework to assess the quality of the motif based on this biological motivation that they should exhibit higher conservation scores rather than using this information apriori for motif discovery. However MoCA's API allows using phylogenetic footprinting based or any other motif finders besides MEME. MoCA is implemented in Python programming language and has an extensible API.

We performed analysis on various ENCODE ChIP-Seq datasets and found that the 'true motifs', validated experimentally do exhibit high conservation scores. A list of analyzed ENCODE data ChIP-Seq datasets is available on http://moca.usc.edu/

References

[1] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble. The meme suite. *Nucleic acids research*, page gkv416, 2015.

[2] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS Comput Biol*, 6(12):e1001025, 2010.

[3] A. Siepel, K. S. Pollard, and D. Haussler. New methods for detecting lineage-specific selection. In *Research in Computational Molecular Biology*, pages 190–205. Springer, 2006.

*Computational Biology and Bioinformatics, University of Southern California, Los Angeles, USA. Email: skchoudh@usc.edu
†Department of Preventive Medicine, University Of Southern California, Los Angeles, USA. Email: valouev@usc.edu