Fat-tail test of regulatory DNA sequence

Jian-Jun Shu

School of Mechanical and Aerospace Engineering,
Nanyang Technological University, Singapore
mjjshu@ntu.edu.sg

Content-based computational methods for distinguishing between *cis*-regulatory element (CRE) and non-CRE are valuable for predicting CRE that have not been observed experimentally. The fluffy-tail test is one of the content-based CRE prediction methods. This is a bootstrapping procedure to identify abundant similar words with statistical significance in regulatory DNA, and then differentiate the regulatory DNA from non-CRE. The fluffy-tail test focuses only on the most frequently occurring subsequences in CREs, thus providing only a measure of homotypic transcription factor binding site (TFBS) clusters; however, most transcriptional regulatory regions contain multiple types of binding motifs; therefore, the fluffy-tail test may fail to capture statistical features arising from heterogeneous TFBS clusters in regulatory regions. In this paper, a kurtosis-based fat-tail test is proposed to measure both homogeneous and heterogeneous clustering. The results show that the fat-tail test separates CRE from non-CRE better than the fluffy-tail test.

Keywords: *cis-regulatory element, DNA sequence, statistical approach, transcription factor binding site*

References

[1] J.-J. Shu, Q.-W. Wang, K.-Y. Yong, F. Shao, K. J. Lee, Programmable DNA-mediated multitasking processor, *Journal of Physical Chemistry B*, 119(17):5639-5644, 2015.

[2] J. R. Wong, K. J. Lee, J.-J. Shu, F. Shao, Magnetic fields facilitate DNA-mediated charge transport, *Biochemistry*, 54(21):3392-3399, 2015.

[3] J.-J. Shu, Q.-W. Wang, K.-Y. Yong, Programmable DNA-mediated decision maker, *International Journal of Bio-Inspired Computation*, 10(1):51-55, 2017.