Fat-Tail Test of Regulatory DNA Sequences

Authors : Jian-Jun Shu

Abstract : The statistical properties of CRMs are explored by estimating similar-word set occurrence distribution. It is observed that CRMs tend to have a fat-tail distribution for similar-word set occurrence. Thus, the fat-tail test with two fatness coefficients is proposed to distinguish CRMs from non-CRMs, especially from exons. For the first fatness coefficient, the separation accuracy between CRMs and exons is increased as compared with the existing content-based CRM prediction method – fluffy-tail test. For the second fatness coefficient, the computing time is reduced as compared with fluffy-tail test, making it very suitable for long sequences and large data-base analysis in the post-genome time. Moreover, these indexes may be used to predict the CRMs which have not yet been observed experimentally. This can serve as a valuable filtering process for experiment.