Non-random Clusters of Palindromes in Herpesvirus Genomes

Louis Chen
Institute for Mathematical Sciences
National University of Singapore
imsdir@nus.edu.sg

Abstract

Palindromes are symmetrical words of DNA in the sense that it reads exactly the same as its reverse complimentary sequence. Palindromes in DNA are involved in a variety of biological processes. For example, the recognition sites for bacterial restriction enzymes are mostly palindromes. The transcriptional regulatory regions of many genes contain palindromes and the origins of replication contain a high concentration of palindromes.

The scan statistic has been used to identify regions of unusually high concentration of palindromes on herpesvirus genomes, thereby locating the likely sites of gene regulators and origins of replication. This method assumes that the palindromes, which may be represented by points on the unit interval, are i. i. d. uniformly distributed. In this talk, we provide a mathematical basis for making this assumption by showing that the palindromes, as a point process, can be approximated by a Poisson process. Stein's method and coupling are applied to obtain an upper bound on the Wasserstein distance between the palindrome process and the Poisson process. This bound is then used as a guide to choose optimal lengths of palindromes for the approximation. The optimal length depends on the length of the genome sequence and its nucleotide base composition. The scan statistic is then applied to identify unusual clusters of palindromes for a number of herpesviruses.

This talk is based on a paper joint with K. P. Choi, M. Y. Leung and A. Xia