# Application of Scan Statistics on Genomic Data:

# Searching Palindrome Clusters

I-Ping Tu
Institute of Statistical Science, Academia Sinica

## Abstract

A DNA palindrome is a segment of letters along a DNA sequence with inversion symmetry that one strand is identical to its complementary one running in the opposite direction. Searching non-random clusters of DNA palindromes, an interesting bioinformatic problem, relies on the estimation of the null palindrome occurrence rate. The most commonly used approach for estimating this number is the average rate method. However, we observed that the average rate could exceed the actual rate by 50% when inserting 5,000 bp hot-spot regions with 15-fold rate in a simulated 150,000 bp genome sequence. Here, we propose a Markov based estimator to avoid counting the number of palindromes directly, and thus to reduce the impact from the hot-spots. Our simulation shows that this method is more robust against the hot-spot effect than the average rate method. Furthermore, this method can be generalized to either a higher order Markov model or a segmented Markov model, and extended to calculate the occurrence rate for palindromes with gaps. We also provide a p-value approximation for various scan statistics to test non-random palindrome clusters under a Markov model.

This is a joint work with Shao-Hsuan Wang and Yuan-Fu Huang.