Homework 3

2024-10-29

Data Description:

The DNA sequence of CMV was published in 1990 by Chee et al. A CMV DNA molecule has 229,354 complementary pairs of letters or base pairs. They are in search of special patterns in the virus' DNA that contains instructions for its reproduction: origins of replication. They discovered a total of 296 palindromic sequences, each of which are at least 10 pairs long.

The longest ones found were 18 letters long and occurred in locations 14719, 75812, 90763 and 173893 along the sequence.

Objective

From the perspective of analyzing the structure in the data, we aim to assess how the distribution of palindromes deviates from a uniform scatter across the DNA sequence. We need to study if the clusters are due to chance.

Dataset

The file hcmv.txt contains the DNA locations of the aforementioned 296 palindromes.

Questions to be aswered for Analysis Section of the Report

1. [Random scatter] Use a computer simulation to see what random scattering looks like. Simulate 296 palindrome sites chosen at random along a DNA sequence of 229,354 bases using a pseudo random number generator. Do this several times, by making sets of simulated palindrome locations. Perform a quantitative and qualitative comparison between random scatters and real data.
2. [Locations and spacings] Using graphical methods to analyze the patterns in:
    a. Spacing between consecutive palindromes
    b. Sums of palindrome pairs
    c. Sums of palindrome triplets
    Compare observed patterns to expected uniform random distributions to identify any significant clusters or unusual spacing in palindrome locations
3. [Counts] Use graphical methods and more formal statistical tests to examine the counts of palindromes in various regions of the DNA. Split the DNA into nonoverlaping regions of equal length to compare the number of palindromes in an interval to the number of that would you expect from uniform random scatter. Study the impact of having shorter and longer region.
4. [The biggest cluster] Does any interval with the greatest number of palindromes indicate a potential origin of replication? Validate your results (be careful in making your intervals).
5. How would you advise biologist who is about to start experimentally searching for the origin of replication?