

# Distribution of DNA Palindromes in short Genomes

Yong Kheng Goh<sup>1</sup>  
gohyk@mail.utar.edu.my

<sup>1</sup> Department of Mathematical and Actuarial Sciences, Universiti Tunku Abdul Rahman, No 13 Jalan Bersatu 13/4, 46200 Petaling Jaya, Selangor, Malaysia

## Keywords:

palindromes, plasmid, periodicity transform, signal processing

## 1 Introduction

The simplest non-coding sequences in genome that play significant roles in the genetic engineering and of biological importance are short palindromes. The term “palindromes” here refers to genetic sequences that with reverse-complement symmetry or the so-called inverted-repeat sequences. Palindromes are crucial for DNA replications, recognition sites of restriction enzymes, and recombinant DNA technologies [1, 2].

In this paper we have carried out a systematic exploratory investigation on the distributions of palindromes in short genomic sequences. We first constructed the palindromes distribution maps for various genomes. The palindrome distributions were then compared with distributions obtained from randomly generated sequence.

## 2 Method and Results

### 2.1 Identifying palindromes

Genome sequences of several organisms were downloaded from NCBI database. For each genome sequence, we scanned for palindromes by using a moving window with the length of 20 nucleotides. For each window, we applied a simplified version of periodicity transform [3, 4] to identify perfect palindromes with their length between 6 nucleotides to 20 nucleotides. In the periodicity transform, we divided the sequence into two sub-sequences with equal length. If the number of nucleotides in the sequence is an odd number, the center of the sequence will be discarded. The first sub-sequence was kept as it is, but the second sub-sequence was reverse-inverted to form a new sub-sequence. The reverse-inverted sub-sequence is then compared with the first sub-sequence. If the two sub-sequences are of exact match, we will have a perfect palindrome.

### 2.2 Palindrome distributions

We scanned the genomes obtained from NCBI as well as from the randomly generated DNA sequences for palindromes. The randomly generated sequences were constructed by drawing “A”, “T”, “C” and “G” with equal probabilities. Typical palindrome distribution maps for *Yersinia pseudotuberculosis* PB1/+ plasmid pYPTS01 genome and a randomly generated DNA sequence with the same length as *Yersinia* was shown in Figure 1.

### 2.3 Total palindromes count per genome size ratio

We also calculated the ratio of total palindrome count to its genome size (or, for short, P/G ratio). Figure 2 shows the P/G ratio against genome size for 26 different plasmid genomes and 16 randomly generated DNA sequences with different sizes.

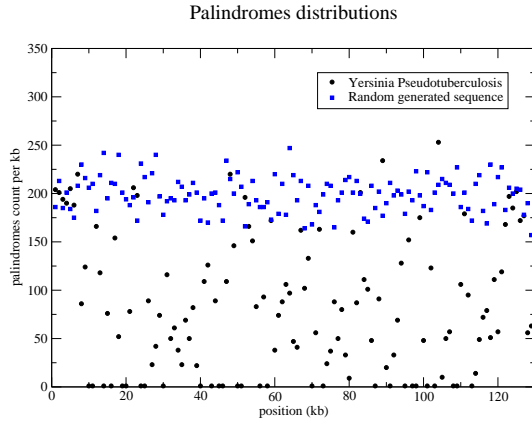


Figure 1: Palindrome distribution maps for Yersinia genome and a randomly generated sequence.

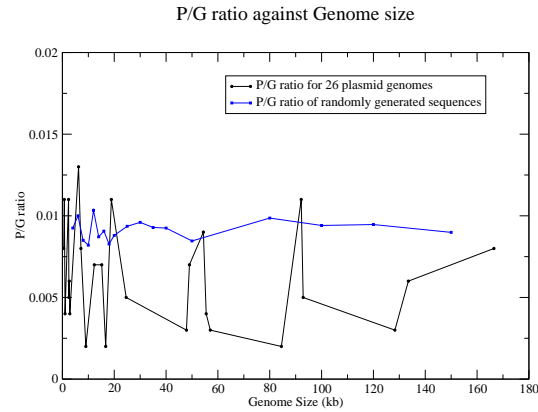


Figure 2: P/G ratio for various genomes.

## 3 Discussions

Most randomly generated DNA sequences have approximately 200 palindromes per 1000 nucleotides as shown in Figure 1. Figure 1 also shows that the palindrome distribution for various genomes, like Yersinia genome, are far from random. Further confirmation comes from Figure 2 which shows that most plasmid genomes have P/G ratios that is lower than those from randomly generated sequences.

## References

- [1] Cech, T.R. and Pardue, M.L., Electron microscopy of DNA crosslinked with trimethylpsoralen: Test of the secondary structure of eukaryotic inverted repeat sequences *Proc. Natl. Acad. Sci.*, 73(8):2644–2648, 1976.
- [2] Bzymek, M. and Lovett, S.T., Evidence for two mechanisms of Palindrome-stimulated deletion in *Escherichia coli*: Single-strand annealing and replication slipped mispairing *Genetics*, 158:527–540, 2001.
- [3] Sethares, W.A. and Staley, T.W., Periodicity Transforms *IEEE Trans. Signal Process.*, 47(11):2953–2964, 1999.
- [4] Gupta, R., Mittal, A. and Gupta, S., An efficient algorithm to detect palindromes in DNA sequences using periodicity transform *Signal Process.*, 86:2067–2073, 2006.