Palindrome Analysis: Distribution of Complete-Matched Inverted Repeats with a Moderate Length Spacer

Masao Fukagawa fukagawa@ims.u-tokyo.ac.jp
Tatsuhiko Tsunoda tatsu@ims.u-tokyo.ac.jp
Tosihisa Takagi takagi@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

1 Introduction

Repetitive sequences on nucleic acids sometimes provide functional sites for reproduction, transcription, translation, etc. They are important for discriminating each condition by combining proteins, e.g. transcription factors, or nucleic acid sequence itself. Also palindromes, i.e. inverted repeat base sequences of nucleic acids, produce such special conditions. Furthermore, they often imply smart mechanisms. First, some pair of proteins symmetrically recognizes the palindrome since each protein binds to either the sense strand or the antisense strand of the DNA. They include the recognition sites of restriction enzymes, the binding sites of dimers of transcription factors, prokaryotic operators, etc. Secondly, a palindrome forms a 3D structure of nucleic acids by forming a stem loop. This contributes to the interaction among nucleic acids, and the interaction between proteins and nucleic acids, which affects reproduction, translation termination of prokaryotes, etc. Finally, an inframe palindrome in a coding region may produce a hydrophobic complementary structure in a peptide chain, which stabilizes the protein molecule.

However, the precise contribution of each repetitive sequence and each palindromic sequence in the genomic sequences is not clear. We are still blind to how specific such sequences appear, and how we can discriminate them from a meaningless match produced by duplication.

At the first step of palindrome analysis, we have proposed an time and memory efficient algorithm that detects prefect palindromes with spacers of arbitrary length. Here we provide a analysis of the distribution of large perfect palindromes, whose half-site (repeat unit) length is more than 12 bases and whose spacer length is less than 100 bases, as applied to various genomic sequences.

2 Materials and Methods

2.1 Algorithm

We use an algorithm to identify the maximum match repeat sequence at each position with a calculation cost of $O(N \log N)$ and memory space of $O(N)$ [2]. In our algorithm, first, we modified the Nagao and Mori's n-gram statistics algorithm [1] such that the frequency statistics of the character string in a given text has been expanded to an algorithm by which the substrings between two texts can be compared. Next, we resolved the problem of detecting repetitive subsequences such as palindromes into a problem of comparing between two texts.

2.2 Database

Complete genome sequences, such as viral genomes were obtained from GenBank. Microorganism intergenic sequences and coding region sequences were obtained from GSDB (Genome Sequence Database). Eukaryotic promoter region sequences were obtained from EPD release 50. Upstream region sequences of yeast genes were obtained from TFCD (Transcription Factor Combination Discoverer). Human coding region sequences were obtained from VTS (Virtual Transcribed Sequence) of Kazusa DNA Research Institute.

3 Results and Discussion

Since large genome sequences requires much time, we analysed them by dividing it into a gene (ORF) and the domain between the genes. Using this method, it will be processed in a reasonable time.

There are many hits of simple repetition sequences in a eukaryotic genome, and there is virtually none in prokaryotes. Hence, some filter process to remove repetition sequence is required. In this analysis, large palindromes are found, in viral or retrotransposon related sequences. For example, the maximum length palindrome in the yeast, constituted by 117bp half-sites and a 66bp spacer, is located at upstream of the retrotransposon TY1A-ER1. Goose parvovirus virulent B is a single stranded DNA virus, and has a extremely large terminal palindrome, constituted by 182bp half-sites and 43bp spacer. This is a reproduction origin. Herpes simplex virus (HSV) type 1 is a double stranded DNA virus. It has a large complete palindrome without any spacer, constituted by 72bp half-sites. This palindrome is believed to have some functions, such as the stopper of the supercoil. Because of the compactness of viral genomes, some functional nucleic acid structure may arise without assistance of proteins.

Under these analysis conditions, the difference remarkable between the coding region (ORF) and the intergenic region was not found out. Although a palindrome in the ORF may have taken a complementary structure within the molecule, large ones were not found in this perfect palindrome analysis. Since real sequences may include mismatches, the possibility of not having been detected is high. In future research, we plan to analyze these sequences taking into account the mismatches.

Most palindromes which are attributed to reproduction or transcriptional regulation are believed to exist in the intergenic regions. We have confirmed that there are no perfect large palindromes in these regions in this analysis. In future work, we will analyze mid-sized palindromes taking mismatch into consideration. Further, we will analyze the tendency of the location from the transcription start site.

4 Acknowledgements

This work is partially supported by Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science" from the Ministry of Education, Science, Sport and Culture, Japan.

References

[1] Nagao, M. and Mori, S., A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, In *Proceedings of International Conference on Computational Linguistics (1994)*, 611–615, 1994.

[2] Tsunoda, T., Fukagawa, M. and Takagi, T., Time and Memory Efficient Algorithm for Extracting Palindromic and Repetitive Subsequences in Nucleic Acid Sequences, In *Proceedings of the Pacific Symposium on Biocomputing 4 (PSB-99)*, Jan 1999.