BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 6 2014, pages 887–888
doi:10.1093/bioinformatics/btt651

Sequence analysis

Advance Access publication November 8, 2013

A MATLAB-based tool for accurate detection of perfect overlapping and nested inverted repeats in DNA sequences

Sutharzan Sreeskandarajan[1], Michelle M. Flowers[2], John E. Karro[2,*] and Chun Liang[1,2,3,*]

[1]Department of Biology, [2]Department of Computer Science and Software Engineering, Miami University, Oxford, OH 45056, USA and [3]State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100193, China

Associate Editor: John Hancock

ABSTRACT

Summary: Palindromic sequences, or inverted repeats (IRs), in DNA sequences involve important biological processes such as DNA–protein binding, DNA replication and DNA transposition. Development of bioinformatics tools that are capable of accurately detecting perfect IRs can enable genome-wide studies of IR patterns in both prokaryotes and eukaryotes. Different from conventional string-comparison approaches, we propose a novel algorithm that uses a cumulative score system based on a prime number representation of nucleotide bases. We then implemented this algorithm as a MATLAB-based program for perfect IR detection. In comparison with other existing tools, our program demonstrates a high accuracy in detecting nested and overlapping IRs.

Availability and implementation: The source code is freely available on (http://bioinfolab.miamioh.edu/bioinfolab/palindrome.php)

Contact: liangc@miamioh.edu or karroje@miamioh.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 15, 2013; revised on October 23, 2013; accepted on November 5, 2013

1 INTRODUCTION

Palindromic sequences, or inverted repeats (IRs), are sequences that can form complementary pairing with themselves. Found in abundance throughout both eukaryotic and prokaryotic genomes, the ability of IRs to self-pair allows for the formation of DNA secondary structures that serve as intermediates in biological reactions such as DNA replication and transposition (Berg *et al*., 1982; Lu *et al*., 2007; Rice, 2005): cruciforms can initiate DNA replication by changing the supercoiling level or chromatin structure, affecting the binding of DNA replication regulatory factors (Cozzarelli and Wang, 1990; Pearson *et al*., 1996; White and Bauer, 1987); DNA hairpins are a result of the palindromic nature of perfect and imperfect IRs, allowing them to act as the binding sites of dimeric regulatory factors (LeBlanc *et al*., 2000; Lonskaya *et al*., 2005) and aid in the transposition process (Kuang *et al*., 2009; Linheiro and Bergman, 2008, 2012; Pray, 2008). Both perfect and imperfect IRs also have an impact on genomic structure: transposons prefer palindromic insertion sites (Linheiro and Bergman, 2012). Transposons can be present in eukaryotic genomes in a nested manner by occurring within other transposons (Gao *et al*., 2012). Hence, analysis of nested and overlapping perfect IRs can aid in a variety of research areas including the study of regulatory factor binding sites and of nested transposons.

Currently available tools in the detection of palindromes do not perform well in terms of result quality (Gupta *et al*., 2006). To make genome-wide studies of palindromic sequences feasible, we need to develop efficient and accurate tools for detecting both perfect and quasi-palindromic sequences (palindromes contain mismatches and/or non-palindromic spacers). Here we describe a novel algorithm for a tool detecting perfect IRs, implemented as a MATLAB function, capable of running on genome-scale inputs.

2 ALGORITHM

Different from conventional string comparison algorithms adopted in existing tools for palindrome detection, the major steps of our algorithm are shown in Figure 1 (see the graphic representation and pseudo-code of our algorithm in Supplementary Figs S1 and S2).

The algorithm detects the boarder positions of perfect IRs using a cumulative scoring system. The scoring system is based on assigning prime number scores to nucleotides such that the scores of complimentary bases cancel each other out. A perfect IR is defined as a DNA sequence that satisfies the following two conditions: (i) the numbers of complementary nucleotide bases are equal and (ii) the number of nested palindromes within the sequence, which share the same center as with the sequence, is equal to half the number of the total bases present.

Our detection scheme is based around the assignment of scores to each nucleotide: A and C are each assigned different prime number scores, whereas T and G are each assigned a score equal to the negative of the score of their complementary base, as shown in Supplementary Figure S1. We then scan the sequence and keep a running cumulative score, noting that if a score repeats itself at two positions, the intervening section must have a balanced number of complementary bases (a result following from our choice of prime number score values), and hence is potentially a perfect IR. We filter out all substrings larger than 1000 bp, as doing so significantly increases the efficiency of our algorithm, and our data testing using chromosome one of human, *Arabidopsis* and maize genome indicates that the presence of palindromes larger than 1000 bp are exceedingly rare. We

*To whom correspondence should be addressed.

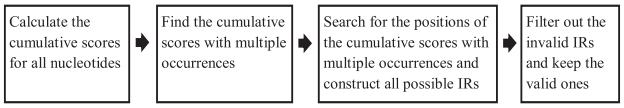

Fig. 1. The major steps of the cumulative prime-number scoring system algorithm for the detection of overlapping IRs

then process each remaining substring with a filtering step to eliminate non-IRs that happen to be balanced in nucleotide numbers.

3 IMPLEMENTATION AND EVALUATION

The proposed algorithm was implemented as a MATLAB function (using MATLAB R2012a) and was compared for accuracy against existing IR detection tools including EMBOSS (Rice *et al.*, 2000), BioPHP (http://www.biophp.org/minitools/find_palindromes) and the MATLAB built-in function *palindromes* on several simple test cases (Supplementary Table S1). The tool IRF (Inverted Repeat Finder) (Warburton *et al.*, 2004) was excluded from the comparison because it cannot detect overlapping IRs. Our MATLAB tool was able to detect all nested and overlapping IRs, whereas the other competing tools were unable to function with 100% accuracy in many cases. As indicated in the Supplementary Table S1, MATLAB's and EMBOSS's algorithms were unable to detect some perfect IR instances of TATA sequences, whereas BioPHP's algorithm missed overlapping perfect IRs starting in same position. Guglietta *et al.* (2010) used BioPHP in their analysis of IRs in HIV-1 gp120 sequences. In their analysis in the C3 region on patient number three, they missed the detection of perfect IRs: ATAT and TATA, whereas our tool was able to detect them (see *Testcase_HIV.fa* in our supplementary data).

For validation on real data, our MATLAB program was run on the HIV-1 genome (NC_001802.1) to search for all perfect IRs, including the nested and overlapping patterns, of lengths 4–1000 nt. Consequently, we detected 649 perfect IRs in 0.1339 s. Moreover, our tool proves to be well-suited for the detection of perfect IRs in larger genome data sets like chromosome 1 of *Arabidopsis thaliana*, *Homo sapiens* and *Zea mays*: we detected 17 030 043 perfect IRs in *H.sapiens*, 25 181 675 in *Z.mays* and 2 788 106 in *A.thaliana*, using 1.035 h, 1.050 h and 7.943 min, respectively.

Although our tool has considerably higher accuracy than other tools, that accuracy comes with some cost in runtime. To determine whether the program was still practical to use on large sequence inputs, we looked at the average execution time over 100 randomly generated sequences at sequence length ranging from 4 to 1000 bp and found it to increase linearly with input size (Supplementary Fig. S3).

4 CONCLUSION

Different from conventional string-comparison approaches adopted in the existing tools, our MATLAB program uses a novel prime number-based algorithm and can accurately detect nested and overlapping IRs of lengths ranging from 4 to 1000 nt. This tool is practically feasible for perfect IR detection in large DNA sequences. Hence, this tool will assist in the effective and accurate detection of perfect IRs in a genome-wide scale.

Funding: NIH-AREA (1R15GM94732-1 A1 to C.L.) and NSF (No. O953215 to J.K.) (in part).

Conflict of Interest: none declared.

REFERENCES

Berg,D.E. *et al.* (1982) Inverted repeats of Tn5 are transposable elements. *Proc. Natl Acad. Sci. USA*, **79**, 2632–2635.

Cozzarelli,N.R. and Wang,J.C. (1990) *DNA topology and its biological effects*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Gao,C. *et al.* (2012) Characterization and functional annotation of nested transposable elements in eukaryotic genomes. *Genomics*, **100**, 222–230.

Guglietta,S. *et al.* (2010) Long sequence duplications, repeats, and palindromes in HIV-1 gp120: length variation in V4 as the product of misalignment mechanism. *Virology*, **399**, 167–175.

Gupta,R. *et al.* (2006) An efficient algorithm to detect palindromes in DNA sequences using periodicity transform. *Signal. Process.*, **86**, 2067–2073.

Kuang,H. *et al.* (2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res.*, **19**, 42–56.

LeBlanc,M.D. *et al.* (2000) An annotated catalog of inverted repeats of Caenorhabditis elegans chromosomes III and X, with observations concerning odd/even biases and conserved motifs. *Genome Res.*, **10**, 1381–1392.

Linheiro,R.S. and Bergman,C.M. (2008) Testing the palindromic target site model for DNA transposon insertion using the Drosophila melanogaster P-element. *Nucleic Acids Res.*, **36**, 6199–6208.

Linheiro,R.S. and Bergman,C.M. (2012) Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLoS One*, **7**, e30008.

Lonskaya,I. *et al.* (2005) Regulation of Poly(ADP-ribose) polymerase-1 by DNA structure-specific binding. *J. Biol. Chem.*, **280**, 17076–17083.

Lu,L. *et al.* (2007) The human genome-wide distribution of DNA palindromes. *Funct. Integr. Genomics*, **7**, 221–227.

Pearson,C.E. *et al.* (1996) Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *J. Cell. Biochem.*, **63**, 1–22.

Pray,L.A. (2008) Transposons: the jumping genes. *Nat. Educ.*, **1**, 1.

Rice,P.A. (2005) Visualizing Mu transposition: assembling the puzzle pieces. *Genes Dev.*, **19**, 773–775.

Rice,P. *et al.* (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.

Warburton,P.E. *et al.* (2004) Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res.*, **14**, 1861–1869.

White,J.H. and Bauer,W.R. (1987) Superhelical DNA with local substructures. A generalization of the topological constraint in terms of the intersection number and the ladder-like correspondence surface. *J. Mol. Biol.*, **195**, 205–213.