A Fast de novo Genome-Wide Tandem Repeat Discovery Algorithm

Marcelo Gonçalves Narciso

narciso@cnpaf.embrapa.br

Embrapa Rice and Beans

Michel E. Beleza Yamagishi

michel@cnptia.embrapa.br

Embrapa Agriculture Informatics

Abstract

Tandem Repeats (TR) are sequences where the same pattern repeats consecutively. They have been used as genomic markers (microsatellite and minisatellite) mostly in plant and animal breeding programs or medical diagnosis. However, recently, new studies have associated TR to important regulatory processes which substantially increased the interest in their study. Furthermore, the advent of the new sequencing technologies turned possible to sequence complex genomes at a low price. Since then, novel organism genomes have been released at a daily basis, and their annotation had to be made from scratch. Very often, the first annotation step is the identification of genetic markers such as SNPs and TRs. As the SNPs identification may be a byproduct of the genome assembly, the real challenge resides in the genome-wide de novo TR discovery. Although there are several algorithms designed to address this task, genome-wide TR discovery is time consuming, and any improvement in computational performance may be significant. Nowadays, genotyping by re-sequencing is an actual option, and any genome-wide de novo TR discovery algorithm applied to several individuals must be as fast as possible. In this work, we propose a new fast de novo genome-wide TR discovery algorithm, called ConvolutionTR. Using a mathematical function called convolution, we were able to compare any sequence to itself in a reasonable computational time in order to find all the TRs. Our experiments show that on average ConvolutionTR is 30% to 50% faster than other algorithms;
moreover, ConvolutionTR finds all TR while the most popular algorithms do not as will be shown.