ID 67

A Fast Algorithm to "de novo" Genome Wide Tandem Repeats Discovery

Marcelo Narciso, Embrapa Rice and Bean

Michel Yamagishi, Embrapa Informatica

Tandem Repeats (TR) are sequences where the same pattern repeats consecutively. They have been used as genomic markers (microsatellite and minisatéllite) since the begining of the genomic era. Recently, new studies have associated TR to important regulatory processes which substantialy increased the interest in TR. The exponential reduction cost of sequencing caused by the new technologies, resulted in the proliferation of genome projects, and particularly of novel model organisms. Very often, the first sequence analysis is the identification of genetic markers such as SNPs and TRs. As the former is a by product of the assembly phase, the real chalenge resides in the latter since the TRs identification must be done de novo. This scenario requires a faster and more efficient algorithms to perform de novo TR discovery. In this paper, we propose a new strategy to address this problem. Our algorithm is able to deal with large genomes in a reduced computational time (on average 30% to 50% faster than other the approaches). Furthermore, our algorithm finds all TR in a genome while some popular algorithms do not as will be shown. Consequently, as our algorithm is faster and find all TR, it may be used in new genomes and old genomes as well to discover eventually missed TR.