[BIOINFO 2023]

Continuous wavelet transforms of nucleotide sequences simplify repeated regions in the genome

Hyunsu Lim[1] and Youbong Hyun[1,*]

[1]*Department of Biological Sciences, Seoul National University*
**Corresponding author: youbong.hyun@snu.ac.kr*

The development of long-read sequencing and advanced assemblers have revolutionized the field of genomics, opening new avenue for investigating biological messages that were previously inaccessible through short-read sequencing. These advancements have enabled the study of repeated regions in genomes, which were considered missing messages due to the limitations of *de Bruijn* or overlap graphs generated from short reads. We present a novel approach for identifying emergent repeat contexts in genomes. Our method involves applying continuous wavelet transform, a well-studied operation in the signal processing field, to reveal hidden repeat contexts. Unlike conventional genome observation methods, which only consider the positional domain of the genome, our approach also considers the frequency domain of the genome, allowing us to identify emergent, targeted repeat elements. The repeat sequences manifested as a two-dimensional matrix that could be visualized intuitively, with a position and frequency axis. The method was applied and tested in several simulated tandem arrays and dispersed repeats. We also operated on whole genome, centromeric tandem arrays and dispersed gene families on *Arabidopsis thaliana*, which genome sequences are fully revealed, suggesting the method can gain insights by simplifying the structure of repeated regions.