Abstract

Genome banks contain precious biological information that is mostly not discovered yet. Biologists in turn are keen to precisely explore these banks in order to discover effective patterns (such as motifs and retro-transposons) that have a real impact on the function and evolution of living creatures. Because the modern genome sequencing technologies produce genomes in high throughputs, many techniques have emerged to store genomes in the lowest possible space. Reference-based Compression algorithms (RbCs) efficiently compress the sequenced genomes by mainly storing their differences with respect to a reference genome. Therefore, RbCs give very high compression ratios compared to the traditional compression algorithms. However, in order to search a compressed genome for specific patterns, it has to be totally decompressed, wasting both time and storage. This paper introduces searching for either exact or incomplete patterns inside the referentially compressed genomes without their complete decompression. The introduced search methodology is based on instantly searching subsequent sequences that are partially decompressed from the compressed genome. Moreover, the same search process is allowed to simultaneously search for multiple patterns, thus saving more resources. The experimental results showed noticeable performance gains.
compared to traditionally searching the same compressed genomes after their complete referential decompression.

References


Index Terms

Computer Science  Artificial Intelligence
Keywords
Reference-based Genome Compression; Partial Genome Decompression; Searching Compressed Genome; Incomplete Patterns; inCompressi.