Abstract

In this paper, an approach for efficiently extracting the repeating patterns in a biological sequence is proposed. A repeating pattern is a subsequence which appears more than once in a sequence, which is one of the most important features that can be used for revealing functional or evolutionary relationships in biological sequences. The algorithm does a rapid scan of the string to find repeating regions where the repeating substring has been marked using length, occurrence positions, and occurrence frequency. The algorithm execute in linear time and space independent of alphabet size. The algorithm also has the capability to restrict output complete repeats in which length (period) $p \geq p_{\text{min}}$, where $p_{\text{min}} \geq 1$ is a user-specified minimum. The algorithm outputs complete repeats, and can be extended or applied to other situations, for example computing maximal repeats, or finding common motifs in a set of biological sequences.


Index Terms

Computer Science
Artificial Intelligence

Keywords

Complete repeats, Biological sequence, Suffix array, Motif finding