Abstract

Discovery of motifs in biological sequences and various types of subsequences in commercial databases have varied applications and interpretations. This paper proposes a new approach to
solve the Combinatorial Pattern Matching (CPM), search for continuous and gapped rigid subsequences and discover Longest Common Rigid Subsequences (LCRS) from the given sequences using DNA operations and modified Position Weight Matrix (PWM). The algorithm and its variations have been tested with both real and simulated databases. The proposed work can be applied to genetic, scientific as well as commercial databases. Implementation results shown the correctness of the algorithms. Finally, the validity of the algorithms are checked and their time complexity is analyzed.

Reference

- D. Lo, S.C. Khoo, and C. Liu. 2007, Efficient mining of iterative patterns for software specification discovery. Int. Conf. on Knowledge Discovery and Data Mining, pages 460-469.


- R. Agarwal and R. Srikant, 1995, Mining sequential patterns. Int. Conf. on Data Engineering.


- T. D. Wu and D. L. Brutlag, 1995, Identification of protein motifs using conserved...
Discovering Sequence Motifs of Different Patterns Parallely using DNA Operations

- X.Yan, J.Han, and R.Afhar, 2003, Colspan: Mining closed sequential patterns in large datasets. SIAM Int. Conf. Data Mining, pages 166-177.

Index Terms

Computer Science
Bioinformatics

Key words
DNA operations
Motifs
LCRS
CPM
PWM
Molecular Computing