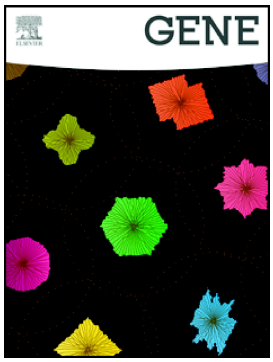
Accepted Manuscript



PATSIM: Prediction and analysis of protein sequences using hybrid Knuth-Morris Pratt (KMP) and Boyer-Moore (BM) algorithm

P. Manikandan, D. Ramyachitra

PII: S0378-1119(18)30217-8
DOI: doi:10.1016/j.gene.2018.02.069
Reference: GENE 42621

To appear in: *Gene*

Received date: 25 October 2017
Revised date: 26 February 2018
Accepted date: 27 February 2018

Please cite this article as: P. Manikandan, D. Ramyachitra , PATSIM: Prediction and analysis of protein sequences using hybrid Knuth-Morris Pratt (KMP) and Boyer-Moore (BM) algorithm. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Gene(2017), doi:10.1016/j.gene.2018.02.069

PATSIM: Prediction and Analysis of Protein Sequences using Hybrid Knuth-Morris Pratt (KMP) and Boyer-Moore (BM) Algorithm

P. Manikandan[1], Dr. D. Ramyachitra[2],
[1]Research Scholar, [2]Assistant Professor,
Department of Computer Science, Bharathiar University, Coimbatore-641 046, India.
manimkn89@gmail.com[1], jaichitra1@yahoo.co.in.

Abstract
In phylogenomic profiling, the genomic context based methods are based on the observation that two or more proteins having the same pattern of presence or absence in many diverse genomes most likely have a functional link. In this research work, a tool (PATSIM) has been developed to predict the protein patterns based on the SOPM tool. In this tool, the secondary structure for CATH database protein sequences, predicted by the SOPM (Self Optimized Prediction Method) server is passed as input to fulfill objectives such as, (i) Predict the Amino Acid Pattern using the proposed Hybrid KMP and BM algorithm, (ii) Predict the physiochemical properties such as Hydrophobic Non-Polar ALKYL Amino Acid groups, Hydrophobic Non-Polar AROMATIC Amino Acid groups, Hydrophilic Polar Neutral Amino Acid groups, Hydrophilic Polar Acidic Amino Acid groups and Hydrophilic Polar Basic Amino Acid groups of protein sequence, (iii) Predict the secondary structure of protein where the structure of protein sequence is unknown, and (iv) Similarity analysis of protein sequence (structure unknown) with the CATH database. From the results, it is inferred that this tool effectively predicts the similarity between the sequences and also identifies the protein patterns for four secondary structural classes, namely Alpha Helix (h), Beta Sheet (e), Turn (t) and Coil (c). Based on the experimental results, it is inferred that this tool identifies the physiochemical properties of the protein sequence in an effective manner. The source code and its documentation for the PATSIM tool is freely available in the GitHub public repository. (https://github.com/manimkn89/Protein-Sequence-Analysis).

Keywords: *CATH*, Protein Secondary Structure, Amino Acid Patterns, Physiochemical Properties, Similarity Analysis, Knuth-Morris Pratt (KMP), Boyer-Moore (BM).

1. Introduction

The prediction of protein secondary structure comprises a more or less precondition step that helps in model building. Numerous methods have been developed to predict the secondary structure of proteins from their amino acid sequences (Gamier and Levin, 1991 & Gamier, 1990). Most of these methods are statistical and they are based on the observed frequency with which individual residues or short sequences of residues are found in given structural states (Robson and Pain, 1971; Gibrat et al., 1987). The sequence similarity methods use the classical substitution matrix (Levin et al., 1986) or averaged amino acid physico-chemical properties (Sweet, 1986) for amino acid comparisons. The evolutions of neural network methods are applied to predict the secondary structure of proteins based on neural nets (Holley and Karplus, 1989; Muskal and Kim, 1992). Nevertheless, a lot of these methods were confirmed against a size-limited database of ~ 100 proteins which contained some interrelated members. The amount of proteins with well-known structure has augmented with an average rate of 150 new structures exposed per year. At the same rate, the volume of the database of secondary structures has not grown, since all the proteins should present <50% identity (Kabsch and Sander, 1983b) to be integrated into it. Secondary structure prediction has mapped into three-class problem of pattern classification such as helix, sheet and coil. Three distances based classifiers namely K-nearest neighbour, minimum distance and fuzzy K-nearest neighbour is used to analyze the secondary structure (Ashish Ghosh & Bijnan Parai, 2008).