



اَبُو سَيِّدِي نَبِي كَوْنِي مَارَا
UNIVERSITI
TEKNOLOGI
MARA

THE DOCTORAL RESEARCH ABSTRACTS

Volume: 7, Issue 7 May 2015

SEVENTH ISSUE

INSTITUTE of GRADUATE STUDIES

Leading You To Greater Heights, Degree by Degree

IGS Biannual Publication

Name :

Hazaruddin Harun

Title :

Linear-PSO With Binary Search Algorithm for DNA Motif Discovery

Supervisor :

Associate Prof. Dr. Sharifah Lailee Syed Abdullah (MS)**Associate Prof. Dr. Naimah Mohd Hussin (CS)**

Motif Discovery (MD) is the process of identifying meaningful patterns in DNA, RNA, or protein sequences. In the field of bioinformatics, a pattern is also known as a motif. Numerous algorithms had been developed for MD, but most of these were not designed to discover species specific motifs used in identifying a specifically selected species where the exact location of these motifs also needs to be identified. Evaluation of these algorithms showed that the results are unsatisfactory due to the lower validity and accuracy of these algorithms. At present, DNA sequencing analysis is the most utilised technique for species identification where patterns of DNA sequences are determined by comparing the sequence to comprehensive databases. However, several false and gap sequences had been identified to be present in these databases which lead to false identification. Therefore, this study addresses these problems by introducing a hybrid algorithm for MD. In this study, the MD is a process to discover all possible motifs that existed in DNA sequences whereas Motif Identification (MI) is a process to identify the correct motif that can represent a selected species. Particle Swarm Optimisation (PSO) was selected as the base algorithm that needs improvement and integration with other techniques. The Linear-PSO algorithm was the first version of improvement. However due to the longer time required for complete execution of this algorithm, the Binary Search technique was integrated and a new version of the algorithm was developed, namely the Linear-PSO with Binary Search (LPBS) algorithm. A total of 11 experiments were conducted in this research, where the aim of the first four experiments was algorithm improvement; the next four experiments were for identifying suitable input data, while the final three experiments were for algorithm validation. Several DNA sequences from different species were collected from the GenBank and TRansCompel databases and used as input for the algorithm. The collected DNA sequences were from the Mitochondrial Cytochrome C Oxidase Subunit I (COX1) gene. Due to the limitation of available data, only four species were collected for Motif Discovery, namely pig, cow, yak, and chicken. Another five species were used for Motif Identification, which were human, sheep, dog, frog, and rat. The algorithm was run on an Intel(R) Core(TM) Duo CPU 1.73 GHz notebook with 3 GB RAM. The results showed that the LPBS algorithm was able to discover possible correct motifs that can represent a species with higher validity and accuracy as compared to previous algorithms. The motifs discovered were consistent for each execution with higher calculated fitness values.