Dr. Khalid Mohammad Jaber
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Al-Zaytoonah University of Jordan

Adapting Decision Tree-Based Method to Index Large DNA-Protein Sequence Datasets

Abstract

Currently, the size of biological databases has increased significantly with the growing number of users and the rate of queries where some databases are of terabyte size. Hence, there is an increasing need to access databases at the fastest possible rate. Where biologists are concerned, the need is more of a means to fast, scalable and accuracy searching in biological databases. This may seem to be a simple task, given the speed of current available gigabytes processors. However, this is far from the truth as the growing number of data which are deposited into the database are ever increasing. Hence, searching the database becomes a difficult and time-consuming task. Here, the computer scientist can help to organize data in a way that allows biologists to quickly search existing information. In this paper, a decision tree indexing model for DNA and protein sequence datasets is proposed. This method of indexing can effectively and rapidly retrieve all the similar proteins from a large database for a given protein query. A theoretical and conceptual proposed frameworks is derived, based on published works using indexing techniques for different applications. After this, the methodology was proved by extensive experiments using 10 data sets with variant sizes for DNA and protein. The experimental results show that the proposed method reduced the searching space to an average of 97.9% for DNA and 98% for protein, compared to the Top Down Disk-based suffix tree methods currently in use. Furthermore, the proposed method was about 2.35 times faster for DNA and 29 times for protein compared to the BLAST+ algorithm, in respect of query processing time. It could be concluded from these results that the proposed DTIM is appropriate for large data sets, in terms of computation time searching algorithm and searching space.