



Available online at www.sciencedirect.com

ScienceDirect

Procedia
Computer Science

Procedia Computer Science 62 (2015) 73 - 80

The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015)

High speed database sequence comparison

Talal Bonny*, Bassel Soudan

Department of Electrical and Computer Engineering, University of Sharjah, Sharjah, UAE

Abstract

Database sequence comparison applications compare a query sequence with each sequence in a database to find the closest match. These applications are high consumers of computation time because they use dynamic programming algorithms to perform the large number of required sequence comparisons. Traditional methods perform the comparisons on the entire set of sequences in the database. In this work, we introduce a novel high-speed technique that reduces the number of database sequences to which the time-consuming matching algorithm is applied. The selection of the target database sequences is based on similarity measures that will be introduced in this contribution as well. Using the proposed technique and the proposed similarity measures, we are able to accelerate the database sequence comparison by 65% compared to traditional exhaustive methods.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of organizing committee of The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015)

Keywords: Database, sequence comparison, Sequence Analysis, Needleman-Wunsch

1. Introduction

Database sequence comparison applications are widely used in different research fields. In Biology^{10, 12} for example, rapid analysis of DNA and Protein sequences are performed to search a large database of sequences for

* Corresponding author. Tel.: +971-6-5053940. *E-mail address:* tbonny@sharjah.ac.ae close matches to particular sequence of interest, typically a recently discovered protein. If correlations are found, new drugs may be developed or better techniques invented to treat the disease.

Database matching applications consume large amounts of computation time because they are based on comparing a particular sequence with large database of sequences. To perform this comparison process, the query sequence needs to be compared with each sequence in the database individually. For each pairwise comparison, a similarity score SS (explained later) is computed which refers to the metric distance between the two compared sequences (see Fig. 1). The highest score refers to the database sequence that is the closest match to the query sequence. Finding the minimum distance (highest similarity score) between two sequences is accomplished through dynamic programing algorithms such as the Needleman-Wunsch¹ and Smith-Waterman² algorithms. These techniques provide optimal alignment in a time that is proportional to the product of the lengths of the two sequences being compared. If n is the length of the query sequence and m is the length of the database sequence, then the previous algorithms provide the optimal alignment in n x m steps. Therefore, the computation time grows linearly with the size of the database.

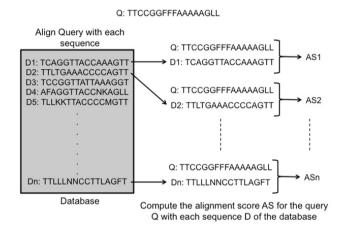


Fig. 1. Exhaustive sequence comparison. The matching algorithm is applied to each database sequence to compute the similarity score

Several techniques have been proposed to process these large amounts of data in a reasonable time ¹⁴. In reference 6, the authors presented a CUDA-based implementation of the Smith-Waterman Algorithm using GPU cards in a common workstation. In reference 14, the authors used the GPU and CPU cooperatively to improve the sequence alignment speed by running long sequences on the GPU and short ones on the CPU. Other methods were introduced based on heuristic sequence alignment algorithms. In reference 15, the authors presented fast algorithm for sequence alignment using Bloom filters that are used in web searching. They implemented their algorithm and compared it with the Needleman-Wunsch method. The Results showed that the average response time was improved by 40%. In reference 16, the authors proposed similarity measures between two web pages and a method of clustering the web sessions using a developed fast sequence-comparing algorithm. Their method aligns the sessions in an average time gain of 35.84% over the conventional dynamic programming Needleman-Wunsch method. In all previous work and applications, similarity measures are used to measure how close is one object to the other. The object might be a database sequence, a string file, a video stream, a website page, etc.

In this work, we propose new similarity measures that are based on the mathematical parameters: frequency, mean and standard deviation of the codes of each database sequence. Using our similarity measures, we expect to reduce the time required to measure the similarity between a pair of database sequences. In addition, we introduce a novel technique to find the most promising sequences in the database. Our technique computes the similarity scores for all database sequences and excludes sequences that have low scores from the detailed alignment process. The

Download English Version:

https://daneshyari.com/en/article/487253

Download Persian Version:

https://daneshyari.com/article/487253

<u>Daneshyari.com</u>