A new method to analyze protein sequence similarity using dynamic time warping

Hou Wenbing, Pan Qiuhui, Peng Qianying, He Mingfeng

PII: S0888-7543(16)30131-8
DOI: doi:10.1016/j.ygeno.2016.12.002
Reference: YGENO 8854

To appear in: *Genomics*

Received date: 12 August 2016
Revised date: 6 December 2016
Accepted date: 10 December 2016

Please cite this article as: Hou Wenbing, Pan Qiuhui, Peng Qianying, He Mingfeng, A new method to analyze protein sequence similarity using dynamic time warping, *Genomics* (2016), doi:10.1016/j.ygeno.2016.12.002

A new method to analyze protein sequence similarity using dynamic time warping

Hou Wenbing[1], Pan Qiuhui[2,1], Peng Qianying[3], He Mingfeng[1,*]

[1]School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, PR China

[2]School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian, 116024, PR China

[3]Department of Academics, Dalian Naval Academy, Dalian, 116001, PR China

Abstract

Sequences similarity analysis is one of the major topics in bioinformatics. It helps researchers to reveal evolution relationships of different species. In this paper, we outline a new method to analyze the similarity of proteins by Discrete Fourier Transform (DFT) and Dynamic Time Warping (DTW). The original symbol sequences are converted to numerical sequences according to their physico-chemical properties. We obtain the power spectra of sequences from DFT and extend the spectra to the same length to calculate the distance between different sequences by DTW. Our method is tested in different data sets and the results are compared with that of other software algorithms. In the comparison we find our scheme could amend some wrong classifications appear in other software. The comparison shows our approach is reasonable and effective.

Keywords

Protein sequences similarity analysis; Discrete Fourier Transform; Dynamic Time Warping; phylogenetic tree;

1. Introduction

With the advance of sequencing techniques, the database of DNA, RNA and protein has been enlarged rapidly, promoting the development of bioinformatics effectively. It has been increasingly important to develop effecient ways to obtain the information hidden in the gene data. In the last few decades, several methods to classify the genes have been proposed. In 1983, Hamori and Ruskin proposed a visible 3-D curve with the name of H-curve to tell the relations between different DNAs[1]. As the first graphical representation, it motivates other researchers in the following years to develop more graphical representations of DNA sequences including 2D, 3D and even multidimensional representations [2-14]. Besides the graphical representations, researchers try to combine some techniques from other disciplines into the study of genes and have proposed novel methods. For example, the Discrete Fourier Transform, which is broadly applied in signal process, has been introduced into the process of genes [15, 16]. It is proved effective in the analysis of DNA sequences.

Methods for similarity analysis of proteins also have been proposed recently. Considering a protein sequence consists of 20 kinds of different amino acids while a DNA sequence only consists of four bases, it is much more complex to express a protein than a DNA sequence. However, there are some methods which are generalized from the ways of analyzing the DNA sequences [17-21]. Yau et al. propose a method with the name of protein map [22] following their previous work. They use the moment vectors to represent proteins and generate a universal protein map[23]. Motivated by the protein map, they also develop a novel method, with the name of protein space, to realize the nature of protein universe[24]. Their method is applied successfully in their following papers and proved effective [25, 26]. He et al. present a new way of