Journal of Fiber Bioengineering and Informatics 7:3 (2014) 387–395
doi:10.3993/jfbi09201408

Similarity Analysis of Protein Sequences Based on the EMD Method

Jihong Zhang [a], Junsheng Zheng [b,*], Fenglan Bai [a], Liwei Liu [a]

[a] *School of Science, Dalian Jiaotong University, Dalian 116028, China*

[b] *Department of Computer Science and Technology, Dalian Neusoft University of Information Dalian 116023, China*

Received 26 February 2014; accepted (in revised version) 23 June 2014; available online 23 September 2014

Abstract

An Empirical Mode Decomposition (EMD) method to analyze the similarities of protein sequences is proposed. The EMD method was used to divide a signal sequence converted from a protein sequence into a group of well-behaved Intrinsic Mode Functions (IMFs) and a residue which is monotonic or a trend. This is so that the similarities can be compared among protein sequences by the corresponding residues conveniently and intuitively. This work verifies the method's suitability by using the cytochrome c protein sequences of seven different species.

Keywords: MQ-RBF Quasi-interpolant; EMD; IMFs; Similarity Analysis; DNA Sequences

1 Introduction

Protein sequence analysis have been widely used in structure and function prediction such as phylogenesis studies and the study on different conservation pattern recognition on the basis of molecular biological analysis.

Various approaches have been explored for protein sequences. Recently, graphical representations of proteins have emerged as a powerful tool for sequence analysis. The advantage of these representations is that it creates special patterns of protein sequences that can be recognized intuitively by the human eyes. These methods provide a simple way for viewing, sorting, and comparing various structures, and making the analysis of similarity among protein sequences. It has been successfully applied to genome sequences of many species as demonstrated [1-14].

The protein sequences could be converted into nonlinear signal sequences by graphical representations of proteins, while the EMD method is a nonlinear, non-stationary data processing method proposed by Norden Huang [15, 16] et al. in 1998. With this method, a complicated

*Corresponding author.
Email address: iamzjs@126.com (Junsheng Zheng).

data set can be decomposed into a small number of intrinsic mode functions (IMFs) that admit well-behaved Hilbert transforms, with an additional residue. As one of the most popular tools in signal processing, the EMD method [17] has been used to divide the DNA sequences into a set of IMFs and residues, this helps to establish the similarity of the different DNA sequences by comparing their residues. Although the similarity of proteins have been studied in the paper [18], they don't focus on the residues but the cross-correlation functions between the IMFs.

In this paper, we extend our method [17] to protein sequences. Cytochrome c is a highly conserved protein across the spectrum of species, found in plants, animals, and many unicellular organisms which is used to transport oxygen around the body for respiration. When the specimens of cytochrome c from different organisms are compared, it is found that the more widely separated two species are in their macroscopic features, the greater the degree of difference in their protein sequences, therefore we select cytochrome c protein sequences of seven species [19] shown in Table 1 - the human, the pig, the dog, the kangaroo, the wheat germ, the green bean and the sunflower seed -as our research objects. Then the EMD method was used to carry out the similarity analysis. In other words, cytochrome c was firstly transformed from protein sequences into nonlinear signal sequences on the basis of graphical representations; after that the EMD method was used to obtain the corresponding IMFs and the residues of the protein sequences which was then compared. Compared with other methods, our EMD method based on graphical

Table 1: The cytochrome c protein sequences of seven species

Species	protein sequence
Human	GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGY SYTAANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADL IAYLKKATNE
Pig	GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGF SYTDANKNKGITWGEETLMEYLENPKKYIPGTKMIFAGIKKKGEREDL IAYLKKATNE
Dog	GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGF SYTDANKNKGITWGEETLMEYLENPKKYIPGTKMIFAGIKKTGERADL IAYLKKATKE
kangaroo	GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLNGLFGRKTGQAPGF TYTDANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFAGIKKKGERADL IAYLKKATNE
wheat germ	ASFSEAPPGNPDAGAKIFKTKCAQCHTVDAGAGHKQGPNLHGLFGRQ SGTTAGYSYSAANKNKAVEWEENTLYDYLLNPXKYIPGTKMVFPGLX KPQDRADLIAYLKKATSS
green bean	ASFBEAPPGBSKSGEKIFKTKCAQCHTVDKGAGHKQGPNLNGLFGRQ SGTTAGYSYSTANKNMAVIWEEKTLYDYLENPXKYIPGTKMVFPGLX KPQDRADLIAYLKESTA
sunflower seed	ASFAEAPAGDPTTGAKIFKTKCAQCHTVEKGAGHKQGPNLNGLFGRQ SGTTAGYSYSAANKNMAVIWEENTLYDYLENPXKYIPGTKMVFPGLX KPQERADLIAYLKTSTA