Kim WC, Park S, Won JI. CORE: Common region extension based multiple protein structure alignment for producing multiple solution. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 28(4): 647–656 July 2013. DOI 10.1007/s11390-013-1365-x

CORE: Common Region Extension Based Multiple Protein Structure Alignment for Producing Multiple Solution

Woo-Cheol Kim[1], Sanghyun Park[2], and Jung-Im Won[3,*]

[1] *College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, U.S.A.*

[2] *Department of Computer Science, Yonsei University, Seoul 120-749, Korea*

[3] *Research Center of Information and Electronic Engineering, Hallym University, Chuncheon, Gangwon 200-702, Korea*

E-mail: wxk11@psu.edu; sanghyun@cs.yonsei.ac.kr; jiwon@hallym.ac.kr

Received September 8, 2012; revised May 10, 2013.

Abstract Over the past several decades, biologists have conducted numerous studies examining both general and specific functions of proteins. Generally, if similarities in either the structure or sequence of amino acids exist for two proteins, then a common biological function is expected. Protein function is determined primarily based on the structure rather than the sequence of amino acids. The algorithm for protein structure alignment is an essential tool for the research. The quality of the algorithm depends on the quality of the similarity measure that is used, and the similarity measure is an objective function used to determine the best alignment. However, none of existing similarity measures became golden standard because of their individual strength and weakness. They require excessive filtering to find a single alignment. In this paper, we introduce a new strategy that finds not a single alignment, but multiple alignments with different lengths. This method has obvious benefits of high quality alignment. However, this novel method leads to a new problem that the running time for this method is considerably longer than that for methods that find only a single alignment. To address this problem, we propose algorithms that can locate a common region (CORE) of multiple alignment candidates, and can then extend the CORE into multiple alignments. Because the CORE can be defined from a final alignment, we introduce CORE* that is similar to CORE and propose an algorithm to identify the CORE*. By adopting CORE* and dynamic programming, our proposed method produces multiple alignments of various lengths with higher accuracy than previous methods. In the experiments, the alignments identified by our algorithm are longer than those obtained by TM-align by 17% and 15.48%, on average, when the comparison is conducted at the level of super-family and fold, respectively.

Keywords structure alignment, similarity search, protein structure

1 Introduction

Over the past several decades, biologists have conducted numerous studies examining both general and specific functions of proteins[1]. However, due to the time and effort required for the experimental methods employed by biologists, there are numerous limitations to these approaches in analyzing protein functions. As such, automated computer systems have recently been developed to analyze the functions of multiple proteins simultaneously[2].

Generally, if similarities in either the structure or sequence of amino acids exist for two proteins, then a common biological function is expected[3]. Protein function is determined primarily based on the structure rather than the sequence of the amino acids that compose it[4]. The fact that the sequence of amino acids can be mutated into other forms during evolutionary processes while the structure generally remains intact is evidence of this[5]. In this context[6], it has been shown that replicate proteins can be artificially constructed from existing proteins when the two proteins have similar functions and structures even with different amino acids sequences. That is, the structure of functionally related proteins provides additional insight into their functional mechanisms and this knowledge has been successfully applied to the functional annotation of proteins whose structure has been previously identified[7-8].

Regular Paper

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology of Korea under Grant No.2012R1A1A3013084.

The preliminary version of the paper was published in the Proceedings of EDB2012.

*Corresponding Author