www.elsevier.com/locate/yjtbi

Protein Sequence Analysis by Incorporating Modified Chaos Game and Physicochemical Properties into Chou's General Pseudo Amino Acid Composition

Chunrui Xu, Dandan Sun, Shenghui Liu, Yusen Zhang

PII: S0022-5193(16)30168-0
DOI: http://dx.doi.org/10.1016/j.jtbi.2016.06.034
Reference: YJTBI8722

To appear in: *Journal of Theoretical Biology*

Received date: 6 June 2016
Revised date: 17 June 2016
Accepted date: 25 June 2016

Cite this article as: Chunrui Xu, Dandan Sun, Shenghui Liu and Yusen Zhang Protein Sequence Analysis by Incorporating Modified Chaos Game and Physicochemical Properties into Chou's General Pseudo Amino Acid Composition, *Journal of Theoretical Biology* http://dx.doi.org/10.1016/j.jtbi.2016.06.034

Protein Sequence Analysis by Incorporating Modified Chaos Game and Physicochemical Properties into Chou's General Pseudo Amino Acid Composition

Chunrui Xu, Dandan Sun, Shenghui Liu, Yusen Zhang *
School of Mathematics and Statistics, Shandong University at Weihai
Weihai 264209, China

Abstract. In this contribution we introduced a novel graphical method to compare protein sequences. By mapping a protein sequence into 3D space based on codons and physicochemical properties of 20 amino acids, we are able to get a unique P-vector from the 3D curve. This approach is consistent with wobble theory of amino acids. We compute distance between sequences by their P-vectors to measure similarities/dissimilarities among protein sequences. Finally, we use our method to analyze four datasets and get better results compared with previous approaches.

Key words: Sequences analysis; Codon; Isoelectric point; Hydropathy index; Protein vector

1 Introduction

With the rapid development of sequencing techniques, the discovery of biological sequences increases fast while the potential information of many sequences is not defined. In order to infer structure and function of protein sequences effectively, many valid approaches have been developed for analyzing biological sequences, which can be divided into two groups: alignment-based (Liu et al., 2015; Needleman and Wunsch, 1970; Gotoh, 1982) and alignment-free methods (Elloumi, 1998; Pham and Zuegg, 2004; Vinga and Almeida, 2003). Sequence alignment assigns scores to represent genetic changes in sequence comparison, but it is computationally difficult with large databases. Therefore, alignment-free methods, including methods based on (1) k-word frequency, (2) sub-string, (3) graphical representation and (4) information theory, have been introduced to overcome the limitations of sequence alignment (Hide et al., 1994; Kantorovitz et al., 2007; Zhang and Chen, 2011; Song et al., 2014).

Considering k-word frequency, the simplest model representing protein sequences is amino acid composition (AAC) (Nakashima et al., 1986). However, AAC model will miss information of sequence-order, which may limit prediction results. In order to overcome this limitation, a powerful and novel discrete model called 'pseudo amino acid composition' (PseAAC) has been introduced by Chou (2001), which has been increasingly and widely used into many biomedicine and drug development areas (Zhong and Zhou, 2014) and nearly all the areas of computational proteomics (Mohabatkar et al., 2013; Mondal and Pai, 2014; Du, 2014). As PseAAC has been widely used, three useful open access soft-wares, named 'PseAAC-Builder' (Du, 2012), 'propy' (Cao, 2013), and 'PseAAC-General' (Du, 2014), were established: the former two are able to generate a wide variety of modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC (Chou, 2011), including not only all the

*Corresponding author:zhangys@sdu.edu.cn