Osong Public Health Res Perspect 2014 5(4), 211–218
http://dx.doi.org/10.1016/j.phrp.2014.06.006
pISSN 2210-9099 eISSN 2233-6052

ORIGINAL ARTICLE

A Novel Approach for Predicting Disordered Regions in A Protein Sequence

Meijing Li [a], Seong Beom Cho [b], Keun Ho Ryu [a,*]

[a]*Database/Bioinformatics Laboratory, Chungbuk National University, Cheongju, Korea.*
[b]*Division of Bio-Medical Informatics, Center for Genome Science, Korea National Institute of Health, Cheongju, Korea.*

Received: June 19, 2014
Revised: June 24, 2014
Accepted: June 24, 2014

KEYWORDS:
amino acid sequence,
disordered protein,
emerging subsequence,
protein structure

Abstract
Objectives: A number of published predictors are based on various algorithms and disordered protein sequence properties. Although many predictors have been published, the study of protein disordered region prediction is ongoing because different prediction methods can find different disordered regions in a protein sequence.
Methods: Therefore we have used a new approach to find the more varying disordered regions for more efficient and accurate prediction of protein structures. In this study, we propose a novel approach called "emerging subsequence (ES) mining" without using the characteristics of the disordered protein. We first adapted the approach to generate emerging protein subsequences on public protein sequence data. Second, the disordered and ordered regions in a protein sequence were predicted by searching the generated emerging protein subsequence with a sliding window, which tends to overlap. Third, the scores of the overlapping regions were calculated based on support and growthrate values in both classes. Finally, the score of predicted regions in the target class were compared with the score of the source class, and the class having a higher score was selected.
Results: In this experiment, disordered sequence data and ordered sequence data was extracted from DisProt 6.02 and PDB respectively and used as training data. The test data come from CASP 9 and CASP 10 where disordered and ordered regions are known.
Conclusion: Comparing with several published predictors, the results of the experiment show higher accuracy rates than with other existing methods.

1. Introduction

The study of protein structure for the prediction of function using data mining has always been known as an important research topic in Bioinformatics. Disordered proteins, referred to as naturally unfolded proteins or intrinsically unstructured proteins, are characterized by a lack of stable tertiary structure when the protein exists as an isolated polypeptide chain under physiological conditions *in vitro*. However, all the analyses of protein

*Corresponding author.
E-mail: khryu@dblab.chungbuk.ac.kr

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © 2014 Korea Centers for Disease Control and Prevention. Published by Elsevier Korea LLC. All rights reserved.

are based on protein primary structure denoted amino acid sequences. Protein sequences decide protein structure, and protein structures concern protein function. In the study of protein structures, prediction of disordered regions in a protein sequence is an important topic [1]. The reasons are as follows: (1) Proteins can function when protein disordered sequences fold with other protein sequences. Therefore, finding the protein disordered regions helps to study functions of proteins [2]. Moreover, most of the hub proteins cannot highly interact with proteins compared with nonhub proteins [3] (disordered proteins) except cancer proteins [4]. (2) When we analyzed the similarity between proteins by protein alignments, identification of disordered regions could avoid disordered regions compared with ordered regions, which therefore improved the accuracy of analysis. (3) Eukaryotic Linear Motifs (ELMs) which are short linear peptide regions containing independent functions not related to protein structures. However, the 70% of ELMs are located in disordered regions [5]. (4) In sequence data, division between disordered regions and ordered regions are more beneficial to study three-dimensional protein structures and properties from protein sequences [6].

In early 1997, Romero et al. proposed the first protein disordered region predictor which applied data mining algorithms to protein sequence data without fixed protein three-dimensional structures [7]. To date, a number of predictors of protein disordered regions have been published. From a view of algorithms which were used to construct the prediction model, several data mining and machine learning algorithms were applied, such as nearest neighbor algorithm [8], support vector machines (SVMs) [9−14], neural networks (NNs) [15−23], artificial neural network (ANNs), regression [24−26], sliding window [27,28], random forest [29], Bayesian Markov chain model [30] and so on.

Many protein properties were used to study protein disordered regions, for example low hydrophobicity, the content of B-factor (residues with high B-factor loops) [31], position-specific score [32−35], high net charge and low hydrophobicity [27], low contact density (average amino acid contact propensity scores with or without pairwise interaction energy matrices) [37−39] and so on. Recently, two predictors [29] were proposed which are based on the profiles of amino acid indices representing various physiochemical and biochemical properties of the 20 amino acids. DISOclust [40] used a different method from other methods which was based on the analysis of three-dimensional structural models using ModFOLDclust [41].

In addition, to increase the accuracy of prediction, several meta-predictors were developed which were combined with several predictors [10,18,21,42−46]. Apart from these methods, multiple sequence alignment with proteins of known protein domains is used to analyze protein structures.

Although many meta-predictors are proposed for increasing prediction accuracy, the increase of accuracy is limited to published based models. We also need to propose new basic prediction methods to search the disordered regions which have specific characteristics using different methods. According to the characteristics of disordered proteins, the regions which are predicted are different from each other [47]. Most of the protein disordered region predictors applied characteristics of disordered proteins to identify the disordered region in a protein sequence. In this study, a novel approach was proposed which did not apply the characteristics of disordered protein. In this paper, we modified and applied an emerging substring generation algorithm which was based on a suffix tree to derive the protein emerging subsequences [36]. These protein emerging subsequences were used to predict disordered regions in a protein sequence sliding window.

The predictor is based on emerging subsequences (ESs) which have high discriminating power, and it is more suitable to use ESs in classification analysis. Comparing with most existing disorder predictors which use a sliding window to map individual residues into a certain feature space, the ES-based predictor decreases the useless patterns for classification. The predictor using sliding window applies the feature selection for selecting more useful patterns. However, the ES-based predictor does not need to change the window size and prunes the generated patterns using feature selection methods.

The rest of the paper is organized as follows. Section 2 presents the method applied to the ES-based predictor using some examples. Section 3 shows the performance of the predictor and discusses the experiment results. Finally, we give some concluding remarks in Section 4.

2. Materials and methods

2.1. Emerging Subsequence

Sequence data are special data which have ordering properties. To discover the emerging pattern from sequence data, an emerging substring and a suffix tree-based framework for generating emerging substring were proposed by Sarah Chan in 2003 [36]. In this paper, to apply the emerging sequential patterns to protein sequence data, the emerging pattern was called an Emerging Subsequence (ES) and defined as being a part of a protein sequence that has a higher frequency of occurrence in the target class than in the source class. Emerging subsequences are more suitable for classifying protein sequences to the disordered sequences and ordered sequences than frequent sequential patterns which are often used in subsequence mining, because of the high discriminating power of emerging subsequences. Frequent sequential patterns only depend on the frequency of the subsequence in the target class.

Download English Version:

https://daneshyari.com/en/article/4202060

Download Persian Version:

https://daneshyari.com/article/4202060

Daneshyari.com