Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 258 (2015) 101–116

FUZZY sets and systems

www.elsevier.com/locate/fss

Mining of protein–protein interfacial residues from massive protein sequential and spatial data

Debby D. Wang [*], Weiqiang Zhou, Hong Yan

Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

Received 16 September 2013; received in revised form 17 December 2013; accepted 26 January 2014

Available online 21 March 2014

Abstract

It is a great challenge to process *big data* in bioinformatics. In this paper, we addressed the problem of *identifying protein–protein interfacial residues from massive protein structural data*. A protein set, comprising 154 993 residues, was analyzed. We applied the three-dimensional alpha shape modeling to the search of surface and interfacial residues in this set, and adopted the spatially neighboring residue profiles to characterize each residue. These residue profiles, which revealed the sequential and spatial information of proteins, translated the original data into a large matrix. After vertically and horizontally refining this matrix, we comparably implemented a series of popular learning procedures, including neuro-fuzzy classifiers (NFCs), CART, neighborhood classifiers (NECs), extreme learning machines (ELMs) and naive Bayesian classifiers (NBCs), to predict the interfacial residues, aiming to investigate the sensitivity of these massive structural data to different learning mechanisms. As a consequence, ELMs, CART and NFCs performed better in terms of computational costs; NFCs, NBCs and ELMs provided favorable prediction accuracies. Overall, NFCs, NBCs and ELMs are favourable choices for fastly and accurately handling this type of data. More importantly, the marginal differences between the prediction performances of these methods imply the insensitivity of this type of data to different learning mechanisms.

Keywords: Protein–protein interface prediction; 3D alpha shape modeling; Residue sequence profile; Joint mutual information (JMI); Neuro-fuzzy classifiers (NFCs); Neighborhood classifiers (NECs); CART; Extreme learning machines (ELMs); Naive Bayesian classifiers (NBCs)

1. Introduction

Analysis of big data is required in almost every sector of our world [31]. How to analyze the massive data or information produced in a scientific area, such as computer science, physics or bioinformatics, is a major but challenging task [10]. Generally, these challenges may include preservation, access and computation of different types of data [30].

[*] Corresponding author.
E-mail addresses: danwang6-c@my.cityu.edu.hk (D.D. Wang), kenandzhou@hotmail.com (W. Zhou), h.yan@cityu.edu.hk (H. Yan).

http://dx.doi.org/10.1016/j.fss.2014.01.017
0165-0114/© 2014 Elsevier B.V. All rights reserved.

Bioinformatics is a representative area that involves analysis or processing of substantial biological data, which are in an exponential growth [20]. More discussions on big data in biology can be found in [13,4,39]. In recent years, Protein Data Bank (PDB) [7] provides a platform for researchers to share biological data (three-dimensional molecular structures), and this platform has become a major database for studies in bioinformatics and structural biology [7,8,43]. The extremely rapid growth of structural data in PDB throws an intractable situation to bioinformaticists.

Specifically, our studies focused on predicting protein–protein interfacial residues, which is important for the prediction of complex structures [48,49], the construction of protein–protein interaction (PPI) networks [29,40] and the understanding of inner working mechanisms of cells [38]. Protein structures in PDB were our mainly studied objects, and their sequential and spatial information was adopted as principal features for the component residues. As a major headache, we were facing abundant samples (vertical) decoded from the structures and a large feature set (horizontal) extracted from the sequential and spatial data of proteins. Therefore, how to handle this large matrix and mine interfacial residues from the massive protein structural data was our major task.

In our previous studies [45], we applied extreme learning machines (ELMs) to the prediction of interfacial residues and comparably implemented support vector machines (SVMs) as proposed in [26]; ultimately, similar prediction performances were obtained. In this paper, we introduced more state-of-the-art learning techniques for the prediction, aiming to investigate whether the massive protein sequential and spatial data are sensitive to any specific learning mechanisms. For example, Neuro-Fuzzy classifiers [42] were implemented. Fuzzy modeling [44,41] and fuzzy techniques have become more and more popular in handling classification and regression problems, and the boundary assumption in fuzzy classification [35,47] gains more attentions for its benefits in improving classification accuracy. Other techniques such as decision trees [11], rough sets [21] and probabilistic models [25,36] were discussed as well.

Here, we provide an outline of our major works. Firstly, we compiled a representative protein set from PDB [26], and all the component residues of these proteins were collected. Novelly, the three-dimensional (3D) alpha shape modeling was applied to labeling interfacial and non-interfacial surface residues from the ensemble of residues. A subsequent feature extraction procedure, which revealed the structural information of the observed proteins [26,50], was performed to characterize these residues. After refining the obtained residue features [12], we conducted a number of learning mechanisms to predict the interfacial residues. Those learning mechanisms included Neuro-Fuzzy Classifiers (NFCs), Classification And Regression Trees (CART), Neighborhood Classifiers (NECs), Extreme Learning Machines (ELMs) and Naive Bayesian Classifiers (NBCs).

2. Materials and methods

Residues, composing proteins, were our input samples. For each observed protein, its interfacial and non-interfacial surface residues were regarded as positive and negative samples, respectively. We adopted the sequential and spacial information of proteins as main features for the samples, and comparably predicted them based on different learning mechanisms.

2.1. Data collection

A comprehensive protein set, proposed in [26], was compiled from the Protein Data Bank (PDB) [7] and employed in our studies. This set was composed of 563 non-redundant protein chains that originated from multi-protein complexes (resolution <3.5 Å). All these chains survived a series of filtering operations in [26]. These operations encompassed eliminating small molecules (<100 residues), removing sequence redundancy for each sequence pair by BLAST using a 25% similarity cut-off of >100 residue regions, and abandoning proteins with small interfaces (<20 residues for hetero complexes, <30 residues for homo complexes). Subsequently, these 563 chains were remained, with 271 chains from hetero complexes and 292 chains from homo complexes [26]. We collected all the component residues of these chains, and finally derived an original data set of 154 993 residues.

2.2. Labeling

In [26,50], surface residues were determined through their exposure ratios of the solvent accessible surface areas (ASA) to the nominal maximum areas, based on a threshold of 10%; the interfacial residues were detected according to the relative positions of the observed chains and their partners, with a relative distance threshold defined (0.5 nm).