Neurocomputing xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers

Zhu-Hong You[a], Xiao Li[a,*], Keith CC Chan[b,*]

[a] Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China
[b] Department of Computing, Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Keywords:
Protein-protein interaction
Substitution matrix
Rotation forest
Protein sequence
Ensemble classifier

ABSTRACT

Protein-protein Interactions (PPIs) play important roles in a wide variety of cellular processes, including metabolic cycles, DNA transcription and replication, and signaling cascades. High-throughput biological experiments for identifying PPIs are beginning to provide valuable information about the complexity of PPI networks, but are expensive, cumbersome, and extremely time-consuming. Hence, there is a need for accurate and robust computational methods for predicting PPIs. In this article, a sequence-based approach is proposed by combining a novel amino acid substitution matrix feature representation and Rotation Forest (RF) classifier. Given the protein sequences as input, the proposed method predicts whether or not the pair of proteins interacts. When performed on the PPI data of *Saccharomyces cerevisiae*, the proposed method achieved 93.74% prediction accuracy with 90.05% sensitivity at the precision of 97.08%. Extensive experiments are performed to compare our method with the existing sequence-based method. Experimental results demonstrate that PPIs can be reliably predicted using only sequence-derived information. Achieved results show that the proposed approach offers an inexpensive method for computational construction of PPI networks, so it can be a useful supplementary tool for future proteomics studies.

1. Introduction

Protein-protein interactions (PPIs) play key roles in many cellular functions such as DNA transcription, metabolic cycles, and signaling cascades [1]. Until recently, several high-throughput experimental screens like Yeast two-hybrid screen (Y2H) [2], mass spectrometry [3], protein chip technology [4], and Tandem Affinity Purification tagging (TAP) [5] have been used for identifying PPIs for different species. Consequently, a large amount of PPI data for different organisms has been uncovered and recorded in public databases [6–9]. However, the biological experiments are not without shortcomings. Not only are they time-consuming and labor intensive, they are also extremely expensive. Another main disadvantage is that they yield relatively high rates of false-positive.

Hence, a reliable identification of PPIs with effective computational approach is of great significance. To date, a wide range of attempts have been made to develop computational methods for PPI identification. Broadly, these methods can be divided into several general categories: approaches based on evolutionary relationships, protein structure, protein domain, genomic information, and protein primary sequence [10]. Generally, the prediction accuracy of the first four is higher, but the deficiency of these approaches appears when they are exposed to problems without prior knowledge about proteins. Theoretically, the amino acids sequence of proteins contains all the necessary information to predict PPIs [11–13]. In addition, the advent of a complete genome sequence for many organisms has generated an enormous amount of protein sequence data for computational biologists [14–19]. Until now, a number of approaches for predicting PPIs directly from protein sequences are developed, and these works demonstrated that the information from protein sequences alone might suffice to predict PPI. There are two major challenges in the task of computational PPIs prediction from primary protein sequence, which are: a) how to effectively represent an amino acids sequence as a feature vector, and b) how to design a powerful computational model to accurately and fast predict the desired class (interacting or non-interacting) [20,21].

For the first issue, the major challenge is how to supply a classifier with the features containing the interaction information to distinguish the interacting and non-interacting protein pairs. Currently, many protein sequence descriptors and features were adopted to represent amino acid sequence of a protein. Shen*et al.* developed the conjoint

* Corresponding authors.
E-mail addresses: xiaoli@ms.xjb.ac.cn (X. Li), keith.chan@polyu.edu.hk (K.C. Chan).

http://dx.doi.org/10.1016/j.neucom.2016.10.042
Received 30 December 2015; Received in revised form 23 September 2016; Accepted 12 October 2016
Available online xxxx
0925-2312/

triad descriptor, which employs the frequency of three continuous amino acids to encoded protein sequences. When performed to predicting human PPIs, it achieves a high accuracy of 83.93% [22]. Guo et al. have developed a sequence-based method which yields a high prediction accuracy of 87.36%, when applied to predicting the PPIs of *S. cerevisiae*. This method considers the interactions between residues a certain distance apart in protein sequence [23]. You*et al.* presented an ensemble model to predict PPIs using the information of protein sequence. When applied on the yeast dataset, it achieves 87.00% prediction accuracy with 86.15% sensitivity. Xia et al. also proposed a novel method for PPIs prediction by using a set of distinctive descriptors from protein sequences [24]. Although existing methods for PPIs prediction have been applied successfully, they still have some limitations and disadvantages [25–27]. For example, the sequence-order information of protein sequences is not fully considered. In this study, we aim to propose a novel protein feature representation method via the approach of amino acid substitution matrix and low rand approximation transformation.

For the second issue, various machine learning methods such as Support Vector Machine (SVM), neural networks, decision tree classifier, Naïve Bayes, and ensemble classifier have been employed to construct the prediction model [28,29]. Jansen et al. constructed a Naïve Bayes classifier to predict the interactions of proteins by integrating multiple features. [30]. You et al. propose a method for PPI prediction using Extreme Learning Machine model combined with a local protein sequence representation [26]. Bock et al. used a SVM model to classify the protein pairs using protein primary structure and associated physicochemical properties [31]. Other classification methods used in the prediction of PPIs are the k-Nearest Neighbor (kNN), Logistic Regression (LR), Extreme Learning Machine (ELM), and Decision Trees (DT) [32,33]. Among them, it has been shown that Random Forest model (RFD) consistently ranks as a top classifier, with SVM being in second place [34–39].

Rotation Forest is a recently proposed ensemble classifier, which is based on the idea of Random Forest. It was found to be more accurate than Random Forest classifier across a number of benchmark data sets [40]. In this article, we report a novel method for predicting protein-protein interactions using the rotation forest (RF) algorithm in conjunction with information of protein sequences. First, we transform each protein sequence into a substitution matrix, in which protein mutation information are contained. Then, a novel method termed as low-rank approximation (LRA) is introduced to find its approximate matrix, from which the row vector is extracted to numerically characterize each protein sequence. Finally, to improve the overall accuracy and robust for PPI prediction, rotation forest predictor is designed to carry out the PPI prediction. The proposed method was tested upon two PPI datasets. The experimental results demonstrate that the proposed approach is indeed feasible and effectual. Consequently, it is a new promising and powerful tool for large-scale PPI prediction.

2. Materials and methodology

In this section, we describe the proposed approach for predicting protein interactions from protein sequences. Our method for predicting PPIs depends on three steps: (1) Map a protein sequence into a numerical matrix by using the amino acids substitution matrix feature representation; (2) The mapped matrix is analysed using Low-Rank Approximation technique to find its approximate matrix, from which a fixed length row vector is extracted to numerically characterize each protein sequence; (3) A Rotation Forest predictor is performed using the fixed length feature vector to predict the PPIs.

2.1. Data source

In this study, three large, real public PPIs data sets are employed to evaluate the performance of the proposed method. The first dataset of physical protein interactions is collected from *Saccharomyces cerevisiae* core subset of database of interacting proteins (DIP), which is also used in the study of Guo et al. [41]. After the redundant protein pairs which contain a protein with fewer than 50 residues or have ≥40% sequence identity were remove, the remaining 5594 protein pairs comprise the final positive dataset. The 5594 non-interacting protein pairs were generated from pairs of proteins whose sub-cellular localizations are different. The whole dataset consists of 11188 protein pairs, where half are from the positive dataset and half are from the negative dataset. The second dataset is composed of 2916 *Helicobacter pylori* protein pairs (1458 non-interacting pair and 1458 interacting pairs) as described by Martin *et al.* Other five species-specific PPI dataset including *E.coli*, *C.elegans*, *H.sapiens*, *M.musculus*, and *H.pylori* are employed in our experiment to verify the effectiveness of the proposed method. All the protein sequences in the seven datasets are extracted from database of SWISS-PROT. The dataset used in this study can be accessed at https://sites.google.com/site/zhuhongyou/data-sharing.

2.2. Feature vector extraction

2.2.1. Substitution matrix for protein sequence

To develop a powerful predictor for PPIs prediction, the key is how to design an effective feature representation method that can truly reflect the intrinsic information of protein sequences. The previous works show that the interacting proteins demonstrate similarity in the molecular phylogenetic tree because of the co-evolution through the interaction. Here we propose a novel feature representation method by incorporating the evolutionary information into protein sequence for PPIs prediction.

In the process of evolution, each amino acid is more or less likely to mutate into various other amino acids. For example, a hydrophilic residue such as arginine is more likely to be replaced by another hydrophilic residue such as glutamine. Formally, substitution matrix (SM) provides a description of the rate at which one residue in a sequence changes to other amino acid over time. The element in the *i*th row and *j*th column of a SM denotes the probability of the *i*th amino acid being mutated to the *j*th amino acid in the evolutional process. A 20 by 20 substitution matrix M is shown in Fig. 1, and the element $M_{i,j}$ (i, j=1, 2,...,20) represents the probability of amino acid i mutating to amino acid j during the evolution process. The matrix M could be denoted as a 20 dimensional vector $M = (V_1, V_2,...,V_{20})$, where $V_i = (M_{1,i}, M_{2,i},...,M_{20,i})^T$.

Given a protein sequence S of length L, $S = s_1 s_2 \cdots s_L$ over the 20-letter amino acids alphabet {*A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y*}. Let s_i represents the *ith* amino acid of the protein sequence S and could be replaced by a vector V_{si} of the substitution matrix M. Then, we can easily obtain a 20 by L matrix $D = (V_{s1}, V_{s2},...,V_{sL})$, where L

	A	C	D	E	F	G	H
A	6	-1	-3	-1	-3	0	-2
C	-1	13	-5	-5	-4	-4	-4
D	-3	-5	9	2	-5	-2	-2
E	-1	-5	2	7	-5	-3	0
F	-3	-4	-5	-5	9	-5	-2
G	0	-4	-2	-3	-5	8	-3
H	-2	-4	-2	0	-2	-3	11

Fig. 1. An illustration of the amino acid substitution matrices.