Methods 110 (2016) 26–34

Contents lists available at ScienceDirect

Methods

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

Prediction of Protein–Protein Interaction via co-occurring Aligned Pattern Clusters

Antonio Sze-To *, Sanderz Fung, En-Shiun Annie Lee, Andrew K.C. Wong

Systems Design Engineering, University of Waterloo, Waterloo, Canada

ARTICLE INFO

Article history:
Received 16 April 2016
Received in revised form 25 June 2016
Accepted 26 July 2016
Available online 27 July 2016

Keywords:
Protein–Protein Interaction
Co-occurring Aligned Pattern Cluster
Supervised learning
Random Forest

ABSTRACT

Predicting Protein–Protein Interaction (PPI) is important for making new discoveries in the molecular mechanisms inside a cell. Traditionally, new PPIs are identified through biochemical experiments but such methods are labor-intensive, expensive, time-consuming and technically ineffective due to high false positive rates. Sequence-based prediction is currently the most readily applicable and cost-effective method. It exploits known PPI Databases to construct classifiers for predicting unknown PPIs based only on sequence data without requiring any other prior knowledge. Among existing sequence-based methods, most feature-based methods use exact sequence patterns with fixed length as features — a constraint which is biologically unrealistic. SVM with Pairwise String Kernel renders better predicting performance. However it is difficult to be biologically interpretable since it is kernel-based where no concrete feature values are computed. Here we have developed a novel method WeMine-P2P to overcome these drawbacks. By assuming that the regions/sites that mediate PPI are more conserved, WeMine-P2P first discovers/locates the conserved sequence patterns in protein sequences in the form of Aligned Pattern Clusters (APCs), allowing pattern variations with variable length. It then pairs up all APCs into a set of Co-Occurring APC (cAPC) pairs, and computes a cAPC-PPI score for each cAPC pair on all PPI pairs. It further constructs a feature vector composed of all cAPC pairs with their cAPC-PPI scores for each PPI pair and uses them for constructing a PPI predictor. Through 40 independent experiments, we showed that (1) WeMine-P2P outperforms the well-known algorithm, PIPE2, which also utilizes co-occurring amino acid sequence segments but does not allow variable lengths and pattern variations; (2) WeMine-P2P achieves satisfactory PPI prediction performance, comparable to the SVM-based methods particularly among unseen protein sequences with a potential reduction of feature dimension of 1280×; (3) Unlike SVM-based methods, WeMine-P2P renders interpretable biological features from which we observed that co-occurring sequence patterns from the compositional bias regions are more discriminative. WeMine-P2P is extendable to predict other biosequence interactions such as Protein–DNA interactions.

1. Introduction

1.1. Background

Protein–Protein Interaction (PPI) is important for various biological processes and functions in living cells such as metabolic cycles, DNA transcription and replication, and signaling cascades [1]. Predicting PPI is thus critical for better understanding the molecular mechanisms inside the cell [1]. It is particularly useful for discovering unknown functions of a protein [2]. Following [3,4], we refer a PPI as an interaction that brings two different proteins A and B into direct physical contact, i.e. heterodimeric interactions. In contrast, most homodimeric interactions, where proteins A and B are identical, are for maintaining the stability of the interacting complex but not for regulating cellular processes [5].

A number of experimental techniques, such as the two–hybrid systems [6], mass spectrometry [7] tandem affinity purification (TAP) [1], and microarray analysis [8], have been developed for systematic and large-scale prediction of PPIs. However, these experimental methods are costly, labor-intensive and time-consuming [9,10]. Thus, existing PPI data obtained by these methods covers only a small fraction of the complete PPI networks [11,12]. Moreover, these experimental methods usually suffer from high rates of both

* Corresponding author.
E-mail addresses: hy2szeto@uwaterloo.ca (A. Sze-To), s3fung@uwaterloo.ca (S. Fung), annie.lee@uwaterloo.ca (E.-S.A Lee), akcwong@uwaterloo.ca (A.K.C. Wong).

http://dx.doi.org/10.1016/j.ymeth.2016.07.018
1046-2023/

false positive and false negative predictions [13,14]. Hence, developing effective and reliable computational methods based on sequence data alone to facilitate PPI prediction is of fundamental importance [15].

1.2. Related work

Existing computational methods for PPI prediction can be divided into four types depending on the input data. The first type such as Computational docking [16] requires three-dimensional structures of the target proteins. It can be applied to the target proteins to simulate if they can interact based on physiochemical properties such as shape complementarity, electrostatics, and biochemcial information [17]. The second type requires genomic information of the target proteins, e.g. gene fusion events [18], the conservation of gene-order [19], and the calculation of prior probabilities of genomic features between interacting proteins [20]. The third type requires prior biological knowledge of the target proteins, e.g. phylogenetic profiles [21], domain knowledge of proteins [22–24] and topological properties of proteins in PPI networks [11]. All these methods have limited applicability because the required data/information is not always available. The last type of methods require only sequence data. It uses the coded information inherent in sequences to predict if a protein pair interacts. For this reason, sequence-based methods are becoming popular, since sequence data is more readily available nowadays [2].

PIPE [25]/ PIPE2 [26,27] is a well-established sequence-based method. Given a protein A, a protein B and a database of positive PPIs, PIPE simply counts how frequently all fixed-length protein sequence segments in Proteins A and B found co-occurring in the database. To achieve such task, all combinations of 20-mers between Protein A and Protein B are first enumerated using a sliding window with a width of 20. Then, the co-occurrence of each combination, e.g. MGIRRLVSVITRPIINKVNS from Protein A and GPEAIILTGTFDDWKGTLPM from Protein B, is searched in the database, and the frequency of their co-occurrence is counted. The sum of all counts is then computed. If the sum is larger than or equal to a threshold, the algorithm then predicts that protein A and B would interact. PIPE2 is a much faster version of PIPE. However, in spite of the satisfactory prediction performance, we observe that there is room for improvement. The key drawback of PIPE/PIPE2 is their use of a fixed-window of 20 amino acids. This is biologically unrealistic since functional regions such as the Short Linear Motifs (SLiMs [28]) have variable length from 3 to 15 amino acids [28]). Most of them are less than 10 amino acids [29]. Recently, a similar algorithm called VLASPD [2] that allows variable length of protein sequence segments is proposed. Nevertheless, it still uses exact patterns, which are neither realistic nor useful for biological analysis since it does not accept variants. Furthermore, it adopts a threshold-based prediction model, which does not allow nonlinear relationship between features and class outputs. Nevertheless, since PIPE2 is well benchmarked [3], we would compare our newly proposed algorithm with it.

Another well-established sequence-based method involves the use of Support Vector Machine (SVM) with Pairwise String Kernel [30–32,15,33,34]. They encode a PPI pair into a feature vector composed by the co-occurrence of k-mer (a sequence of k residues) and train the SVM to predict if a protein pair can interact. For example, assume $k = 3$, a selected feature could be the number of counts of how often the 3-mers, say WTG and LGA co-occur in a protein pair along the entire sequence. Since all possible 3-mers are considered, the feature space could be as large as $20^3 \times 20^3$ (i.e. 64 millions) [4]. With SVM, even with such a high dimensionality, by using the kernel trick, neither computing nor storing the feature vector is needed. As no feature vectors are computed, in spite of achieving satisfactory prediction performance, it is hard to use SVM results to reveal or interpret why the feature space leads to its good performance. Thus, since the feature space is hardly interpretable, not much biological knowledge can be gained. Hence, to overcome this hurdle encountered in SVM is another key motivation of our proposed method. It should be noted that it is possible to generalize k-mer counting strategies allowing for gaps and mismatches [35]. However, these methods still do not allow a variable length. For example, if k is set to be 5, these methods would still consider all the 5-mers, while in WeMine-P2P, there could be 5-mers, 6-mers and 7-mers. In WeMine-P2P, we utilize the locally conserved sequence pattern clusters [36,37] and their co-occurrence [38] to obtain biologically realistic and interpretable features that are flexible in pattern length while allowing variants. Experiments showed that our prediction results based on these features are comparable to those achieved by the SVM with Pairwise String Kernel approaches. In addition, the presence of concrete feature values makes the feature analysis of our models (and the subsequent biological interpretation) easier for biologists, comparing to the SVM with Pairwise String Kernel approaches, which have no concrete features and thus make feature analysis (and the subsequent biological interpretation) of the models difficult.

1.3. Motivations and objectives

Motivated by the majority acceptance of sequence-based methods and the realization their drawbacks, the objective of our research as reported in this paper is to develop a new sequence-based prediction method which is (1) based on biologically interpretable features, (2) generating features to be more biologically realistic such as allowing variable lengths and pattern variations, and (3) achieving satisfactory prediction performance with biologically interpretable features. In this study, we propose a new algorithm WeMine-P2P, as illustrated in Fig. 1, to accomplish these objectives.

1.4. Paper layout

The remaining sections are outlined as follows. Section 2 explains in detail the WeMine-P2P prediction algorithm. Section 3 describes the dataset used and its pre-processing involved. Section 4 shows the design of the experiments and reports the results. Section 5 discusses the experimental results. Section 6 concludes the whole study.

2. Methods

2.1. Overview

We discover and locate APCs, then cAPC pairs, the "what" and "where" of the conserved regions, using them as discriminative features to construct the PPI classifier. This is elaborated in steps 1 to 6 in Fig. 1.

2.2. Problem definition

A protein pair, or a PPI pair is defined as a pair of protein sequences that can either be interacting or not interacting with one another. A Protein–Protein Interaction pair, referred to as a positive PPI pair, is defined as a pair of protein sequences that can interact with each other. A protein–protein non-interaction pair, or a negative PPI pair, is defined as a pair of protein sequences that cannot (or is not yet known to) interact with each other. A PPI database includes protein sequences, as well as both positive and negative PPI pairs. We use it to train a model for predicting