Computers in Biology and Medicine 43 (2013) 1196–1204

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Improving protein complex classification accuracy using amino acid composition profile

Chien-Hung Huang [a], Szu-Yu Chou [a], Ka-Lok Ng [b,*]

[a] Department of Computer Science and Information Engineering, National Formosa University, 64, Wen-Hwa Road, Hu-wei, Yun-Lin 632, Taiwan
[b] Department of Biomedical Informatics, Asia University, 500 Lioufeng Road, Wufeng Shiang, Taichung 41354, Taiwan

ARTICLE INFO

Article history:
Received 28 October 2012
Accepted 30 May 2013

Keywords:
Protein complex
Protein–protein interaction
Gene Ontology
Sequence alignment
Physicochemical property
Hydrophobic
Hydrophilic
Amino acid composition profile
Machine learning method

ABSTRACT

Protein complex prediction approaches are based on the assumptions that complexes have dense protein–protein interactions and high functional similarity between their subunits. We investigated those assumptions by studying the subunits' interaction topology, sequence similarity and molecular function for human and yeast protein complexes. Inclusion of amino acids' physicochemical properties can provide better understanding of protein complex properties. Principal component analysis is carried out to determine the major features. Adopting amino acid composition profile information with the SVM classifier serves as an effective post-processing step for complexes classification. Improvement is based on primary sequence information only, which is easy to obtain.

1. Introduction

It is known that protein complexes are involved in many biological processes. Some of the well-known protein complexes are: enzyme–inhibitor complexes, antibody–protein complexes, and protein–receptor complexes [1]. Enzyme–inhibitor complexes include trypsin-like serine proteinases and subtilisins (PDB code 2six), antibody–protein complexes include immunoglobulin FAB complexed with lysozyme (PDB code 2hfl), and protein–receptor complexes include human growth hormone, hGHbp (PDB code 3hhr). The spoke model hypothesizes that all the subunits inside the complex directly interacts with the bait protein, whereas the matrix model assumes all possible interacting pairs among the complex's subunits [2–5]. The correctness of these two models is still an open question and needs further investigation. Subunits refer to the protein constituents of a protein complex.

Recent experimental studies indicate that a protein complex can be visualized as a unit composed of cores, modules and attachments [6,7]. Core proteins are proteins that have comparatively more interactions among themselves and belong to a unique protein complex [8,9]. Attachment proteins bind to the core proteins with relative fewer interactions among them. Module proteins are a subset of the attachment, which are always present together, and module proteins can be present in more than one complex. A recent study has suggested that the prediction of protein complexes based on the core-attachment model can achieve better performance than graphical approaches [10]. Furthermore, it is reported that subunits of a complex tend to have highly correlated gene expression patterns [11].

In this study we propose to characterize human and yeast protein complexes by adopting protein–protein interaction (PPI) data. This allows us to quantify the interaction topology among the subunits of a complex. It is known that many protein complex prediction calculations are based on the identification of pseudo-cliques [12–15], and dense PPI regions [16–19]. Our study may serve as a test of whether a protein complex is composed of highly interacting subunits.

Secondly, the average of the pairwise sequence similarity, i.e. the bit score, of subunits inside a protein complex will be computed. This number can be used to characterize the overall sequence similarity of a complex. Thirdly is the *Jaccard index* (*JI*) of the Gene Ontology (GO) annotation; here the molecular function descriptions for protein subunits are used.

In our previous study [20] it is conjectured that prediction approaches based on the assumption that complexes are composed of highly PPI dense regions can predict a rather limited number of complexes. In this study we propose characterizing protein complexes by considering their physicochemical properties. Amino acids' physicochemical properties are used in characterizing PPI interfaces for protein complexes [21]. Also, there has

[*] Corresponding author. Tel.: +886 423 394541.
E-mail address: ppiddi@gmail.com (K.-L. Ng).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.05.026

been an attempt to use physicochemical properties in detecting remote protein homology [22], with successful results. In a previous work [23], it was suggested that p*I* and sequence length could be used to help predict the probability that a protein belongs to a particular complex. AAindex is a database [24] that collected various physicochemical and biochemical properties of amino acids. AAindex (version 9.1) documented a list of 544 amino acid indices. A recent work [25] proposed the use of fuzzy clustering techniques to categorize these 544 indices into three high quality subsets, and demonstrated the effectiveness of their approach for prediction of post-translational modification sites.

Many physicochemical property calculations required secondary structure or tertiary structure information, which limited the usefulness of such approach. Here we propose to consider the following physicochemical properties of a complex; the composition profile of the 20 amino acids, hydrophobicity, hydrophilicity, p*I* value, and subunit sequence length. The numerical value of these properties was derived from the primary sequence information, which is much easier to access. For instance, the ExPASy tool, ProtParam [26], computes the numeric physicochemical properties of a protein using sequence data only.

Therefore, instead of trying to predict complexes from PPI data only, the major objective of the present study is to identify important physicochemical parameters for protein complex classification. It is proposed that the results of this work will be helpful in improving the accuracy of protein complex classifications. This is achieved by post-processing protein complex prediction results.

For the purpose of protein complex classification, physicochemical parameters are used to construct the feature vectors that are trained by support vector machine (SVM) [27], neural networks (NN), decision tree (DT) and a naïve Bayes classifier (NBC) [28].

Principal component analysis (PCA) is an useful technique in bioinformatics, it reduces the dimensionality of the original data set [29], improves performance by removing correlations among the feature variables. To identify the major features or capture the contribution due to physicochemical properties, PCA [30] is adopted to determine the major feature spaces (a space spanned by the linear combination of the original features) before using the machine learning classifiers. PCA had been presented as a feature selection method [31–33] for extracting a reduced set of feature variables, which preserve the main features of the whole data set. This approach found applications in corn fungi detection [34], machine defect classification [35], and image classification [36,37].

Once the major feature spaces are determined they will be used and trained by the above four machine learning methods. This will be followed by ten-fold cross-validation test to validate the classification accuracy based on the major feature spaces.

2. Methods

2.1. Interaction topology of protein complex subunits

A total of 1818 protein complex data were retrieved from MIPS [38] for human data, and for yeast data a total of 1643 protein complexes were retrieved from Bond [39], and 491 complexes from a database maintained by a group of scientists at Cellzome AG and the European Molecular Biology Laboratory in Heidelberg, Germany [http://yeast-complexes.russelllab.org/]. This database is denoted as 'Yeast' in our study.

Protein subunits' accession numbers are labeled according to the gene index for the protein. A topological parameter is defined to test whether protein complexes are found in PPI dense regions, or not [20]. This parameter is the density of interaction, which describes the experimentally recorded PPI among the subunits of a protein complex relative to the maximum possible PPI (i.e. clique).

Given a protein complex with N subunits, there can be $N*(N+1)/2$ possible PPIs, including self-interaction. The density of PPI, ρ, among the subunits of a protein complex, is then given by

$$\rho = \frac{2s}{N*(N+1)} * 100\% \quad (1)$$

where s is the observed number of PPIs among the subunits. PPI data are obtained from the BioGrid database [40].

2.2. Sequence bit score of protein complex subunits

An all-against-all pairwise sequence alignment is performed using the BLAST program. Output files reported by the BLAST program for all the protein complexes are parsed and the bit score value for each complex is kept for further analysis. The average of the bit score of a complex D, I_D, is defined by

$$I_D = \frac{2}{N(N-1)} \sum_{i<j\in D} I_{ij} \quad (2)$$

where I_{ij} denotes the bit score values reported by BLAST, i and j are labels ($i=1, \ldots, N-1$) which denote the complex subunits.

Since the average bit score value I_D varies from complex to complex, a normalized index V is introduced to represent the complex. Given a property, complex D has an average value $\overline{d}$ over its subunits; in other words, $\overline{d}$ is a generalized symbol for the average value of a property, such as the bit score or any other physicochemical property. Let $V(D)$ represent the normalized computed index for a complex D, which is defined by

$$V(D) = \frac{\overline{d} - min(D)}{max(D) - min(D)} \quad (3)$$

where *max* and *min* correspond to the maximum and minimum operations respectively. The *max* and *min* operations do not run over subunit's index i and j but over the complete set of protein complexes. It is noted that $V(D)$ lies between 0 and 1 for a property.

There is concern that the use of normalized value may filter out information, we demonstrated that the use of normalized value resulted in better classification accuracy (see Appendix Table 1).

2.3. Gene Ontology of protein complex subunits

It is suggested that a protein complex is a biologically functional module composed of subunits performing similar functions [41]. Although evolutionary mechanisms drive the emergence of functional modules, the function of the core component of the complex appears to be more conserved among duplicate complexes; hence each complex remains functionally similar.

Molecular function (MF) annotations for the subunits are carried forward from GO, which is used to characterize the whole complex. *JI* is a quantity that is used to quantify the similarity between two sets, hence, given two subunits i and j, the *JI* is given by

$$JI^{MF}(i,j) = \frac{|i\cap j|}{|i\cup j|} \quad (4)$$

where $|i\cap j|$ and $|i\cup j|$ denote the cardinality of $i\cap j$ and $i\cup j$ respectively. It is noted that *JI* lies between 0 and 1. For example, given that the MF annotation for subunit $i=\{a, b, c\}$ and subunit $j=\{b, c, d\}$, then $JI(i,j)=2/4=0.50$. An all-against-all pairwise subunits' JI^{MF} is computed, and the average *JI* score for a complex D, $JI^{MF}(D)$, with N