Vol.2016-MPS-108 No.12
Vol.2016-BIO-46 No.12
2016/7/4

Improving Accuracy for Predicting Heterodimeric Protein Complexes Using Combination Kernels

Peiying Ruan[†1] Morihiro Hayashida[†2]
Tatsuya Akutsu[†3] Jean-Philippe Vert [†4]

Abstract: Since many proteins become functional only after they interact with their partner proteins and form protein complexes, it is very important to identify the sets of proteins that form complexes. Therefore, several high-throughput methods have been proposed to predict complexes, such as MCL, MCODE, CMC, PCP and CFinder. These methods show higher performances on predicting protein complexes with size of more than three because they are mainly based on the topological structures of protein-protein interaction (PPI) network. However, for heterodimeric protein complexes, each complex involving only two proteins, their topological structures are too simple to analyze, but the majority of known protein complexes are heterodimeric protein complexes. In this paper, we use three promising kernel functions, Min kernel, Metric Learning Pairwise Kernel (MLPK) and Tensor Product Pairwise Kernel (TPPK). We also consider the normalized forms of Min kernel, which are MinMax kernel and Scaled Min kernel. Then, we combine one of the Min kernels (Min kernel, MinMax kernel and Scaled Min kernel) and one of the pairwise kernels by plugging. We applied kernels based on PPI, domain, phylogenetic profile and subcellular localization properties to predicting heterodimeric protein complexes. Then, we evaluate our method by employing C-Support Vector Classification (C-SVC) and carrying out 10-fold cross-validation, and calculating the average F-measures. The results suggest that our proposed method improved the performance of our previous work, which had been the best existing method so far.

Keywords: Heterodimeric protein complex, Support Vector Machine, Kernel

1. Introduction

Since many proteins become functional only after they interact with their partner proteins and form protein complexes, it is very important to identify the sets of proteins that form complexes. Therefore, several high-throughput methods have been proposed Many proteins carry out their biological functions by interacting with other proteins to form multi-protein structures, called protein complexes [1], which are crucial for a wide range of biological process. For example, the ribosome is an assembly of protein and RNA subunits responsible for protein translation. Therefore, understanding protein functions as well as biological processes requires identification of sets of proteins that form complexes. A large fraction of known protein complexes are heterodimeric, that is, formed by the assembly of two different proteins. For example, the two main protein complex catalogs CYC2008 [31] and MIPS [32] include respectively 172 (42%) and 64 (29%) heterodimeric protein complexes. Hence, it is necessary to develop accurate methods for predicting heterodimeric complexes. Here CYC2008 is a comprehensive catalogue of 408 manually curated yeast protein complexes reliably supported by small-scale experiments, and MIPS provides detailed information involving classification schemes for analysis of protein sequences, RNA genes and other genetic elements [19, 20, 21].

In this paper, our goal is to further improve the prediction accuracy for heterodimeric protein complexes. We investigate novel kernels to encode the domain composition of proteins involved in a complex, because the one used in the previous study was very crude, and employ C-Support Vector Classification (C-SVC) to predict heterodimeric protein complexes. We use Min kernel and its two normalized forms, MinMax kernel and Scaled Min kernel, as well as two pairwise kernels, metric learning pairwise kernel (MLPK) and tensor product pairwise kernel (TPPK). Compared to previous work that they only used a single kernel, we try to combine multiple kernels, since proper combination of kernels may achieve better performance than only using one. Besides the domain property, we also try to use phylogenetic profile property. Since if two proteins are present or absent in the same genome, then these two proteins are likely to have related functions. Moreover, protein subcellular localization property is considered as well. Proteins must be localized at their appropriate subcellular compartment to perform their function, so that proteins in the same location may have similar function. Kernels mentioned above are applied to these two properties as well. Then we perform ten-fold cross validation, and calculate the average F-measures. The computational experiments show that the combinations of multiple kernels outperform single kernel proposed in our previous method, and therefore is superior to other existing methods.

2. Methods

2.1 Problem

We formulate the problem of heterodimeric complex prediction as a supervised binary classification problem: given a training set of pairs of proteins known to form a complex (positive pairs) and pairs of proteins that do not form a complex (negative

†1 Hitachi Ltd.
†2 Kyoto University
†3 Nara Institute of Science and Technology

pairs), we learn a function f(x) to predict if a new pair x of proteins can form a complex or not. We use a support vector machine (SVM) classifier, with balanced loss penalty to compensate for the fact that the numbers of positive examples and negative examples can be very unbalanced.

2.2 Properties and their kernels

We explain kernels involving PPI properties, domain property, phylogenetic profile property and subcellular localization property. Then, we use Min kernel's normalization form and propose combination kernels for predicting heterodimeric protein complexes. For the kernels between proteins, we mentioned 3 kernels: Min kernel, and two normalized versions (MinMax and scaled). We also mentioned two ways to make a pairwise kernel: MLPK and TPPK. So we consider all possible combinations (3*2 = 6) of these kernels.

3. Experiments

In order to compare our proposed method with the method in [26], we used the same dataset WI-PHI including 49607 interacting protein pairs except self-interactions. The weights of interactions were calculated in the following way. (1) They used the high-throughput yeast two-hybrid data by Ito [3] and Uetz [2] as well as several databases such as BioGRID [6], MINT [7] and BIND [8] to build the literature-curated physical interaction (LCPH) dataset. (2) They constructed a benchmark dataset to evaluate high-throughput data. The interactions of the dataset were obtained by two independent methods from LCPH-LS, which was a low-throughput dataset in LCPH. (3) They calculated a log-likelihood score (LLS) to each dataset except LCPH-LS. (4) They computed the weight of each interaction by multiplying the socioaffinity (SA) indices [1] and the LLSs from different datasets. Note that SA index is the log-odds score of the number of times that we observed two proteins interact to each other to the expected value in the dataset. Also, we prepared the same dataset from CYC2008 [31] for training and testing as the previous study. We defined a positive example as a pair of proteins included in WI-PHI as well as a heterodimeric protein complex included in CYC2008. A negative example was defined as a pair of proteins included in WI-PHI, which meanwhile should not be any heterodimeric protein complex but be in a subset of some other complexes in CYC2008. As a result, we had 152 positive examples and 5345 negative examples.

3.1 Results

We present a comparison of the previously described combination kernels to the best existing method [26]. These three figures indicate that all the combination kernels except kernels that combined with TPPK kernel perform better than previously proposed Domain Composition kernel for every value of alpha. However, TPPK combined kernels are just a little lower than Domain Composition kernel. This result is not surprising because we mentioned before that Domain Composition kernel is either 1 or 0, and our proposed combination kernels are not binary, but are allowed to have many possible values. Therefore, the result of them may be expected not worse than that of Domain Composition kernel. In addition, we observed that Normalized Min-MLPK kernel and MinMax-MLPK kernel had better performances in most cases. Their curves crossed over each other at several points but almost better than other combination kernels. This observation shows that the combination of Min kernel and MLPK kernel works well, and MinMax kernel or Normalization is necessary for improving prediction accuracy. The result shows the performance of each combination kernel on their best average precision, recall and F-measure. Normalized Min-MLPK kernel had the best performance on precision and Min-TPPK kernel had the best performance on recall. Normalized Min-MLPK kernel achieved the best performance (F-measure increased from 63.1% [26] to 68.6%) and all the proposed methods that combined with MLPK kernel outperform Domain Composition kernel.

4. Concluding Remarks

We applied several kernels based on PPI, domain, localization and profiles information to predicting heterodimeric protein complexes. To evaluate our proposed method, we performed ten-fold cross-validation computational experiments for the combination kernel of Min kernel, MinMax kernel and MLPK kernel by plugging and TPPK kernel by summation. The results suggest that our proposed method improved the performance of our previous work, which had been the best existing method so far. In particular, the combination kernel Normalized Min-MLPK has the best performance. The paper [27] showed that combination of MLPK and TPPK together almost always had best results.

Reference

[1] A. C. Gavin, P. Aloy, P. Grandi, R. Krause and M. Boesche. Proteome survey reveals modularity of the yeast cell machinery. Nature, 440:631-636, 2006.
[2] P. Uetz, L. Giot, G. Cagney, T. Mansfield and R. Judson. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403(6770):623-627, 2000.
[3] T. Ito, T. Chiba, R. Ozawa, M. Yoshida and M. Hattori. A comprehensive two-hybrid analyzes to explore the yeast protein interactive. Proceedings of the National Academy of Science of the United States of America, 98(8):4569-4574, 2001.
[4] P. L. Bartel and S. Fields. The Yeast Two-Hybrid System. Oxford University Press, 1997.
[5] L. Kiemer, S. Costa, M. Ueffing and G. Cesareni. WI-PHI: A weighted yeast interactive enriched for direct physical interactions. Proteomics, 7(6):932-943, 2007.
[6] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher and A. Breitkreutz. BioGRID: a general repository for interaction datasets. Nucleic Acids Research, 34(Database issue):D535-D539, 2006.