

Prediction of Protein-Protein Interaction Sites Using Support Vector Machines

Yohei Minakuchi¹ Kenji Satou^{1,2} Akihiko Konagaya¹
y-minaku@jaist.ac.jp ken@jaist.ac.jp kona@jaist.ac.jp
Takashi Ito³
titolab@kenroku.kanazawa-u.ac.jp

- ¹ School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
² Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Corporation (JST)
³ Cancer Research Institute, Kanazawa University, 13-1 Takara-machi, Kanazawa, Ishikawa 920-0934, Japan

Keywords: protein-protein interaction sites, surface patch, support vector machines

1 Introduction

Protein-protein interactions play an important role in various biological processes. Over the past few years, several studies have been made on protein interface and those results enable us to obtain massive data on various aspects of protein interface. The problem that we have to consider next is predicting the interaction sites on the protein surface. Although several prediction methods are developed, there is still room for improvement. The purpose of this study is to develop a reliable prediction system of protein-protein interaction sites from their three-dimensional structure.

2 Materials and Methods

2.1 Dataset

We used a non-redundant dataset of heterodimers selected by Farizelli *et al.* [2]. This dataset consists of 226 interacting protein chains. There are two reasons for the use of this dataset. First, our aim is to learn general properties of protein interface. For this purpose, several kinds of protein, whose protein has strong signals on protein interface such as homodimers and protease-inhibitor complexes, must be eliminated from the dataset. Second, noncontact surface residues are used as negative data for the SVMs because there are no chains involved in more than one interaction in the dataset.

2.2 Surface and Contact Definitions

Most of protein interfaces are exposed to the solvent when the partner chain is removed. Therefore, we focused only on protein surface residue. The threshold for deciding whether a residue was a surface residue was set at 16% of the relative accessible surface area [2]. The accessible surface areas were calculated for each chain using the DSSP program [4]. We classified a residue to be an contact residue if it has at least three atoms whose distance with some atoms of the partner chain is less than 5Å [5]. According to the above definitions, there are 31,932 surface residues, including 6,515 contact residues and 25,417 noncontact residues, respectively.

2.3 Surface Patches and Feature Representation for SVMs

A surface patch [3] was defined as a central surface residue and its 10 nearest surface residues. Each residue of surface patch is not necessarily contiguous in the sequence and represent local protein surface centering on the target residue. In order to learn with support vector machines, surface patches should be represented as vectors. Here, each residue of a surface patch is represented by the 20-dimensional vector, whose values are taken from a sequence profile in the HSSP database [1].

3 Results

The accuracy resulting from the 3-fold cross-validation is 75.0%. Although it is difficult to perform direct comparison with other predictors because of different definitions of surface and contact residue, our system was definitely better than when compared with similar predictor based on neural networks. Judging from the above, we conclude that support vector machine approach is useful for prediction of protein-protein interaction sites.

Acknowledgments

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) “Genome Information Science” from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

- [1] Dodge, C., Schneider, R., and Sander, C., The HSSP database of protein structure-sequence alignments and family profiles, *Nucleic Acids Res.*, 26(1):313–315, 1998.
- [2] Fariselli, P., Pazos, F., Valencia, A., and Casadio, R., Prediction of protein–protein interaction sites in heterocomplexes with neural networks, *Eur. J. Biochem.*, 269(5):1356–1361, 2002.
- [3] Jones, S. and Thornton, J.M., Prediction of protein–protein interaction sites using patch analysis, *J. Mol. Biol.*, 272(1):133–143, 1997.
- [4] Kabsch, W. and Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 22(12):2577–2637, 1983.
- [5] Zhou, H.X. and Shan, Y., Prediction of protein interaction sites from sequence profile and residue neighbor list, *Proteins*, 44(3):336–343, 2001.