Classification of Protein-protein Interaction Types Using Support Vector Machines

Hongbo Zhu, Francisco S. Domingues, Ingolf Sommer and Thomas Lengauer

1 Introduction

Protein-protein interactions are of vital importance to many biological processes. However, not all the interactions presented in a certain protein complex structure determined by x-ray crystallography are biologically relevant. Many of them are formed during the crystallization process and would not appear *in vivo*. Such crystal packing interactions are non-specific crystal artefacts which have no biological functionality [1]. The determination of the oligomeric state of protein complexes remains a non-trivial problem [2].

The types of biological interactions are also diverse [3]. Protomers from obligate complexes do not exist as stable structures *in vivo*, whereas protomers of non-obligate complexes (e.g. transient complexes) may dissociate from each other and stay as stable and functional units *in vivo*.

We present a two-stage support vector machine (SVM) classifier for discriminating three types of protein-protein interactions: obligate, non-obligate and crystal packing interactions. Firstly, we analyzed five protein-protein interface properties for our interaction data. Then these properties were combined using a support vector machine algorithm to help determine the types of protein-protein interactions. We achieved a total accuracy of 91.1% with a leave-one-out cross-validation (LOOCV) procedure.

2 Interface Properties

Based on previously defined sets of protein-protein interaction data [4, 5], we compiled a non-redundant dataset composed of 302 interactions. In this new dataset 94 interactions are obligate, 88 are non-obligate and 120 are crystal packing. A protein-protein interface is defined as the ensemble of all interface residues[1]. Five interface properties are investigated based on this dataset: interface area, ratio of interface area to protein surface area, area-based amino acid composition in protein-protein interface, correlation between amino acid compositions of interface and protein surface, and degrees of conservation of interface residues.

We found that interface area by itself is the best discriminant for identifying non-biological interactions. It failed for only around 10% of all instances when doing a binary classification (biological versus non-biological interaction). However, interface area is biased against complexes containing small protomers. Therefore we normalized it with the solvent accessible surface area (SASA) of the smaller protomer in the complex, and defined it as ratio of interface area to protein surface area. When combined with interface area, the misclassification rate in the binary classification of biological and non-biological interations could be reduced to one half,

[1] Interface residues are those that lose more than 1 Å^2 of their solvent accessible surface area (SASA) during the formation of complex.

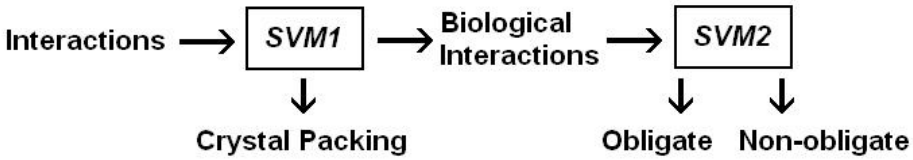

Figure 1: Two-stage SVM classifier.

Table 1: Performance of SVM classifiers

	OB	NO	CP
Precision	90.4%	86.7%	94.4%
Sensitivity	90.4%	85.5%	95.3%
Specificity	95.4%	95.2%	96.1%

Note: OB: Obligate; NO: Non-obligate; CP: Crystal Packing. Precision = TP/TP+FP; Sensitivity=TP/TP+FN; Specificity=TN/TN+FP.

compared to the results using only interface area. Amino acid composition can be either number-based or area-based. Number-based amino acid composition is computed based on the frequency of the 20 standard amino acids at protein-protein interface. Area-based amino acid composition additionally takes into account the loss of solvent accessible surface areas of interface residues. We found that area-based amino acid composition is able to differentiate between the three types of interactions better than number-based amino acid composition. The correlation coefficient between interface and protein surface amino acid compositions were calculated to measure the randomness of interface patch. The average value of this property in obligate interations is the smallest in all the three types of interactions, while it is the largest in crystal packing interactions. Degrees of conservation of interface residues were calculated using ConSurf [6]. In biological interfaces, including both obligate and non-obligate interfaces, most conserved residues were found to contribute almost twice (36 $Å^2$) as much area as those in non-biological interfaces do (19 $Å^2$) on average. In general, all the interface properties of non-obligate interactions are in between of those of obligate and crystal packing interactions.

3 SVM classifier

Our problem is a multi-class classification problem. We chose a support vector machine algorithm to solve it, and implemented a two-stage SVM classifier for discriminating the three classes of interaction data. In the first stage, crystal packing interactions were separated from biological interactions. All biological interactions were further divided into obligate and non-obligated interactions in the second stage (Figure 1).

We tested our classifier with all combinations of the five interface properties. Interface area and ratio of interface area to protein surface area are the two most powerful features in SVM classification. Correlation between interface and surface amino acid composition showed weakest prediction power. When using all five interface properties, we achieved an overall accuracy for our SVM classifier of 91.1% with a leave-one-out cross-validation procedure. The accuracies for SVM classifiers in the first stage and the second stage are 95.7% (biological versus non-biological interactions) and 88.5% (obligate versus non-obligate interations) respectively. The detailed performance of our SVM classifier is reported in Table 1. Generally speaking, our SVM classifier performed best for identifying crystal packing interactions

and worst for non-obligate iterations.

4 Acknowledgements

This research was performed in the context of the EU Network of Excellence BioSapiens (EU grant No. LHSG-CT-2003-503265).

References

[1] J Janin and F Rodier. Protein-protein interaction at crystal contacts. *Proteins*, 23(4):580–7, Dec 1995.

[2] Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joel Janin. A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, 336(4):943–55, Feb 2004.

[3] Irene M A Nooren and Janet M Thornton. Diversity of protein-protein interactions. *EMBO J*, 22(14):3486–92, Jul 2003.

[4] Hani Neuvirth, Ran Raz, and Gideon Schreiber. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*, 338(1):181–99, Apr 2004.

[5] James R Bradford and David R Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–94, Apr 2005.

[6] A Armon, D Graur, and N Ben-Tal. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*, 307(1):447–63, Mar 2001.