A simple and efficient method for predicting protein-protein binding sites

Roberto H. Higa[1,2]
Clésio Luiz Tozzi[1]
[1]Departamento de Computação e Automação Industrial, Faculdade de Engenharia Elétrica e de Computação,
Universidade Estadual de Campinas.
[2]Centro Nacional de Pesquisa em Informática Agropecuária, Empresa Brasileira de Pesquisa Agropecuária.

Abstract

Computational methods for predicting protein-protein binding sites based on structural data are useful for assisting on the identification of putative function for proteins in the PDB [2] having unknown function and for important applications like rational drug design. Existing methods may be divided in two groups: those based on patch analysis of structural and chemical parameters [5][6][9][11] and those based mostly on sequence profile information [12][8][10][4]. Concerning the classification approach, there is a prevalence of neural networks and SVM [4][6][8][10][12], even though statistical score and naive bayes classifier have also been used [5][9][11].

In addition, binding sites are not uniform with only a few key residues contributing to a large fraction of the binding energy [3]. These residues are referenced in the literature as binding hot spots and tend to form clusters in the central part of the binding site. In particular, an interface model for binding sites having a core region surrounded by a rim region has been proposed by Chakrabarti and Janin [7] and Bahadur et al. [1]. In this model, a core region is formed by residues totally buried in complex while the rim region is formed by those residues, which still partly exposed in complex. In terms of amino acid composition, the core region can be better discriminated from the remainder of the protein surface while the rim region presents an intermediate characteristic.

Despite this fact, in order to predict binding sites, methods based on patch analysis of structural and chemical parameters consider the binding site as an uniform region and evaluate hundred, sometimes thousand, of approximately circular patches sampled from the surface of a protein.

In this work, we propose a strategy for predicting binding sites by exploiting the characteristic of core and rim regions of binding sites mentioned above, while using simple and well-known pattern recognition techniques. Specifically, first the interface core residues are predicted and then used for building surface patches, which are defined as predicted binding sites. A linear classifier is used in the first stage and the k-means cluster algorithm is used in the second.

We use a previously validated dataset containing 180 non-redundant proteins involved in both obligatory and non-obligatory interactions, built by Bradford and Westhead [5]. A set of 28 properties was computed for each residue exposed to solvent and used to form a 56-dimensional vector of attributes, with 28 of them corresponding to the residue properties and 28 to the average among its neighbor residues of the same set of properties.

In order to evaluate the performance, the success rate was estimated by a leave-one(protein)-out cross validation method, where the success rate is defined as the number of successful predictions over the entire dataset and a prediction is defined as successful if at least one predicted patch presents coverage ≥ 20% and precision ≥ 50% [5].

Despite its simplicity, the proposed predictor achieved a success rate of 82%. This is a quite competitive performance compared to other methods, described in the literature, using the same dataset: 76% from Bradford and Westhead [4] and 82% from Bradford et al. [5]. In addition, different from other methods based on patch analysis, where hundred of patches per protein are evaluated, our method evaluated only 2.7 patches per protein on average. This represents a considerable reduction of computational effort, showing that our strategy of building patches parsimoniously is effective.

Currently we are investigating the extending of the vector of attributes with structural parameters reported in the literature and correlated to hot spots. In particular, we expect to improve the performance for non-obligate interactions.

References

[1] Bahandur RP, Chakrabarti P, et al. (2003). Dissecting subunit interfaces in homodimeric proteins. Proteins, 53:708-719.

[2] Berman HM, Westbrook J, et al. (2000). The protein data bank. Nucleic Acids Res. 28: 235-242.

[3] Bogan AA, Thorn KS (1998). Anatomy of hot spots in protein interfaces. JMB, 280:1-9.

[4] Bordner A, Abagyan R (2005). Statistical analysis and prediction of protein-protein interfaces. Proteins, 60:353-366.

[5] Bradford JB, Needham CJ, et al. (2006) Insights into protein-protein interfaces using a bayesian network prediction method. JMB, 362:365-386.

[6] Bradford JR, Westhead DR (2005). Improved prediction of protein-protein binding sites using support vector machines approach. Bioinformatics, 21(8):1487-1494.

[7] Chakrabarti P, Janin J (2002). Dissecting protein-protein recognition sites. Proteins, 47:334-343.

[8] Fariselli P, Pazos F, Valencia A, et al. (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. EJB, 269:1356-1361.

[9] Jones S, Thornton JM (1997). Prediction of protein-protein interaction sites using patch analysis. JMB, 272:133-143.

[10] Koike A, Takagi T (2004). Prediction of protein-protein interaction sites using support vector machines. Prot. Eng. D& S, 17(2):165-173.

[11] Neuvirth H, Raz R, et al. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. JMB, 338:181-199.

[12] Zhou HX, Shan Y (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins, 44:336-343.