A Gaussian Evolutionary Method for Predicting Protein-Protein Interaction Sites

Kang-Ping Liu[1] and Jinn-Moon Yang[1,2,3,*]

[1] Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan
[2] Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan
[3] Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan
moon@faculty.nctu.edu.tw

Abstract. Protein-protein interactions play a pivotal role in modern molecular biology. Identifying the protein-protein interaction sites is great scientific and practical interest for predicting protein-protein interactions. In this study, we proposed a Gaussian Evolutionary Method (GEM) to optimize 18 features, including ten atomic solvent and eight protein 2nd structure features, for predicting protein-protein interaction sites. The training set consists of 104 unbound proteins selected from PDB and the predicted successful rate is 65.4% (68/104) proteins in the training dataset. These 18 parameters were then applied to a test set with 50 unbound proteins. Based on the threshold obtained from the training set, our method is able to predict the binding sites for 98% (49/50) proteins and yield 46% successful prediction and 42.3% average specificity. Here, a binding-site prediction is considered successful if 50% predicted area is indeed located in protein-protein interface (i.e. the specificity is more than 0.5). We believe that the optimized parameters of our method are useful for analyzing protein-protein interfaces and for interfaces prediction methods and protein-protein docking methods.

Keywords: Atomic solvation parameter, Gaussian evolutionary method, protein-protein interactions, protein-protein binding site.

1 Introduction

Protein-protein interactions play a pivotal role in modern molecular biology. Study of the energetics and mechanism of protein-protein association is a matter of great scientific and practical interest. Identifying the interface between two interacting proteins can reduce the search space required by docking algorithms to predict the structures of complexes and provides important information to identify the function of a protein.

It is widely accepted that protein structural knowledge on a residue and atom level share common properties that can be used to distinguish a protein-protein interacting

* Corresponding author.

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 143 – 154, 2007.

interface from the rest of a protein [1-4]. The hydrophobic interaction is one of the major contributors to the affinity of the association [5, 6]. Fernandez-Recio *et al*. [7, 8] successfully extracted the desolvation properties of protein surface to construct atomic solvation parameters for predicting protein-protein interaction sites. However, no single attribute absolutely identifies interface from the rest of a protein [3]. Combination of more physical–chemical properties [2, 9-12] and computational methods are needed for predicting protein-protein interaction sites.

In this study, according to the concept of atomic ASA-based model, we examine the difference in parameters of hydrophobic and structure between the protein interface and the rest of the protein surface for a set of 104 protein–protein interfaces. Briefly, we combine secondary structure information with atomic solvation parameters [7, 8] and optimize these parameters using the GEM [13-18] method for developing an interface prediction. These 18 parameters composed of 10 of the atom properties derived form Fernandez-Recio *et al*. [7, 8] and 8 of the secondary structure properties from DSSP [19]. Based on these visualized optimizing parameters, we are able to predict and analyze interface residues of proteins which are not included in the training set, without any prior knowledge of the binding partner.

2 Materials and Methods

Figure 1 shows the scheme of our method for predicting protein-protein interaction binding site. First, we prepared a training data set which consisted of 52 complexes (104 chains) from a widely used benchmark [20] . For each chain in this data set, the protein surface and interacting interface are derived from the protein structures collected in Protein Data Bank (PDB). Based on these characteristics, we calculated all surface residues scores and trained the GEM parameters to distinguish interacting residues from non-interacting residues by ranking scores of surface residues. The surface residue with the score lower than a given threshold was predicted as an interface residue. For the predicted area of a protein, the specificity and sensitivity [21] were applied to measure performance. Specificity was defined as number of interface residues in predicted area/number of predicted area residues. Sensitivity was defined as number of interface residues in predicted area/number of interface residues. A prediction was deemed a success if predicted area with over 50% specificity [11]. Based on these measure factors as the scoring function, the GEM method optimized atomic and structure parameters to find best solution of average specificity and success rate from 104 training proteins. These optimized parameters were used to predict the protein-protein binding sites for 50 testing proteins if the score of a interface was lower than a given threshold obtained from the training set.

2.1 Data Sets

The aim of this work is to study energetics and mechanism of protein-protein association and improve protein-protein docking algorithm by identifying protein-protein interfaces.