Discovery of Protein Interaction Sites

Haiquan Li
The Samuel Roberts Noble Foundation, Inc., USA

Jinyan Li
Nanyang Technological University, Singapore

Xuechun Zhao
The Samuel Roberts Noble Foundation, Inc., USA

INTRODUCTION

Physical interactions between proteins are important for many cellular functions. Since protein-protein interactions are mediated via their interaction sites, identifying these interaction sites can therefore help to discover genome-scale protein interaction map, thereby leading to a better understanding of the organization of living cell. To date, the experimentally solved protein interaction sites constitute only a tiny proportion among the whole population due to the high cost and low-throughput of currently available techniques. Computational methods, including many biological data mining methods, are considered as the major approaches in discovering protein interaction sites in practical applications. This chapter reviews both traditional and recent computational methods such as protein-protein docking and motif discovery, as well as new methods on machine learning approaches, for example, interaction classification, domain-domain interactions, and binding motif pair discovery.

BACKGROUND

Proteins carry out most biological functions within living cells. They interact with each other to regulate cellular processes. Examples of these processes include gene expression, enzymatic reactions, signal transduction, inter-cellular communications and immunoreactions.

Protein-protein interactions are mediated by short sequence of residues among the long stretches of interacting sequences, which are referred to as interaction sites (or binding sites in some contexts). Protein interaction sites have unique features that distinguish them from other residues (amino acids) in protein surface. These interfacial residues are often highly favorable to the counterpart residues so that they can bind together. The favored combinations have been repeatedly applied during evolution (Keskin and Nussinov, 2005), which limits the total number of types of interaction sites. By estimation, about 10,000 types of interaction sites exist in various biological systems (Aloy and Russell, 2004).

To determine the interaction sites, many biotechnological techniques have been applied, such as phage display and site-directed mutagenesis. Despite all these techniques available, the current amount of experimentally determined interaction sites is still very small, less than 10% in total. It should take decades to determine major types of interaction sites using present techniques (Dziembowski and Seraphin, 2004).

Due to the limitation of contemporary experimental techniques, computational methods, especially biological data mining methods play a dominated role in the discovery of protein interaction sites, for example, in the docking-based drug design. Computational methods can be categorized into simulation methods and biological data mining methods. By name, simulation methods use biological, biochemical or biophysical mechanisms to model protein-protein interactions and their interaction sites. They usually take individual proteins as input, as done in protein-protein docking. Recently, data mining methods such as classification and clustering of candidate solutions contributed the accuracy of the approach. Data mining methods learn from large training set of interaction data to induce rules for prediction of the interaction sites. These methods can be further divided into classification methods and pattern mining methods, depending on whether negative data is required. Classification methods require both positive and negative data to develop discriminative features for interaction

sites. In comparison, pattern mining methods learn from a set of related proteins or interactions for over-presented patterns, as negative data are not always available or accurate. Many homologous methods and binding motif pair discovery fall into this category.

MAIN FOCUS

Simulation Methods: Protein-Protein Docking

Protein-protein docking, as a typical simulation method, takes individual tertiary protein structures as input and predicts their associated protein complexes, through simulating the confirmation change such as side-chain and backbone movement in the contact surfaces when proteins are associated into protein complexes. Most docking methods assume that, conformation change terminates at the state of minimal free energy, where free energy is defined by factors such as shape complementarity, electrostatic complementarity and hydrophobic complementarity.

Protein-protein docking is a process of search for global minimal free energy, which is a highly challenging computational task due to the huge search space caused by various flexibilities. This search consists of four steps. In the first step, one protein is fixed and the other is superimposed into the fixed one to locate the best docking position, including translation and rotation. Grid-body strategy is often used at this step, without scaling and distorting any part of the proteins. To reduce the huge search space in this step, various search techniques are used such as, Fast Fourier transformation, Pseudo-Brownian dynamics and molecular dynamics (Mendez et al., 2005). In the second step, the flexibility of side chains is considered. The backbone flexibility is also considered using techniques such as principal component analysis in some algorithms (Bonvin, 2006). Consequently, a set of solutions with different local minima is generated after the first two steps. These solutions are clustered in the third step and representatives are selected (Lorenzen & Zhang, 2007). In the fourth step, re-evaluation is carried out to improve the ranks for nearly native solutions, since the nearly native solutions may not have the best free energy scores due to the flaws of score functions and search algorithms. Supervised data mining techniques have been applied in this step to select the near-native solution, using the accumulative confirmation data for benchmark protein complexes (Bordner and Gorin 2007). Note that in all steps, biological information may be integrated to aid the search process, such as binding sites data (Carter et al., 2005). In the interaction site determination problem, without the guidance of binding sites in docking, the top-ranked interfaces in the final step correspond to the predicted interaction sites. With the guidance of binding sites, the docking algorithms may not contribute remarkably to the prediction of interaction sites since the above steps may be dominated by the guided binding sites.

Although protein-protein docking is the major approach to predict protein interaction sites, the current number of experimentally determined protein structures is much less than that of protein sequences. Even using putative structures, ~ 40% proteins will be failed in protein structure prediction (Aloy et al., 2005), especially for transmembrane proteins. This leaves a critical gap in the protein-protein docking approach.

Classification Methods

Classification methods assume that the features, either in protein sequence or in protein spatial patches, distinguish positive protein interactions from negative non-interactions. Therefore, the distinguishing features correspond to protein interaction sites. The assumption generally holds true but not always.

The first issue in protein interaction classification is to encode protein sequences or structures into features. At least two encoding methods are available. One transforms continuous residues and their associated physicochemical properties in the primary sequence into features (Yan et al., 2004). The other encodes a central residue and its spatially nearest neighbors one time, which is so called spatial patches (Fariselli et al., 2002). The latter encoding is more accurate than the first one because protein structures are more related to interaction sites.

After encoding the features, traditional classification methods such as support vector machine (SVM) and neural networks can be applied to predict interaction sites (Bock & Gough, 2001; Ofran & Rost, 2003). Recently, a two-stage method was proposed (Yan et al., 2004). In the learning phase, both SVM and Bayesian networks produce a model for the continuously encoded residues. In the prediction phase, the SVM model is first applied to predict a class value for each residue, then the Bayesian