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 and 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 contains 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 and 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 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 (Joel and David, 2001; Ofran and 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

model is applied to predict the final class value based on predicted values in SVM model, exploiting the fact that interfacial residues tend to form clusters.

Although classification methods have many advantages, for example, they are good at handling transient complexes which are tough in docking, they have several disadvantages. First, they suffer from unavailability and low quality of the negative data. Second, many classification algorithms apply fixed-length windows in coding, which conflicts with the basic fact that many interaction sites have variable length. Finally, the complicated coding often results in incomprehensibility to the interaction sites from a biological point of view.

Pattern Mining Methods

Pattern mining methods assume that interaction sites are highly conserved in protein homologous data and protein-protein interactions. These conserved patterns about interaction sites are quite different from random expectation and thus, can be revealed even in the absence of negative data. Typical patterns include binding motifs, domain-domain interaction pairs and binding motif pairs.

Binding Motifs from Homologous Proteins

Given a group of homologous proteins, the inherited patterns can be recovered by searching the locally over-represented patterns (so-called motifs) among their sequences, which is often referred to as motif discovery. It is a NP-hard problem and similar to sequential pattern mining but more complicated since its score function is often implicit. The majority of methods discover motifs from primary sequences and can be roughly categorized into pattern-driven, sequence-driven and the combined ones. Pattern-driven methods enumerate all possible motifs with a specific format and output the ones with enough occurrences. For instance, MOTIF (Smith et al., 1990) searches all frequent motifs with three fixed positions and two constant spacings. Sequence-driven methods restrict the candidate patterns to occur at least some time in the group of sequences, for instance, the widely used CLUSTALW (Thompson et al., 1994). Combined methods integrate the strength of pattern-driven and sequence-driven methods. They start from patterns at short lengths and extend them based on conservation of their neighbors in the sequences, for instance, PROTOMAT (Jonassen, 1997). Other motif discovery approaches have also been studied, for example, statistical models such as Hidden Markov models (HMM) and expectation maximization (EM) models.

Motifs can also be discovered from homologous protein structures, which are called structural motifs. Structural motif discovery has been studied from various perspectives, such as frequent common substructure mining (Yan et al., 2005) and multiple structure alignment (Lupyan et al., 2005), impelling by the rapid growth of solved protein structures in recent years.

In general, motif discovery only identifies individual motifs without specifying their interacting partners and thus, can't be considered as complete interaction sites. To reveal both sides of interaction sites, individual motifs can be randomly paired and their correlation can be evaluated by protein-protein interaction data, as done by Wang et al. (2005). Even though, the discovered motif pairs can not guarantee to be interaction sites or binding sites. The rationale is that binding and folding are often interrelated and they could not be distinguished only from homologous proteins. Since homologous proteins share more folding regions than bindings regions, the discovered motifs by sequence or structure conservation are more likely to be folding motifs rather than binding motifs (Kumar et al., 2000). To identify the complete interaction sites, protein interaction information should be taken into consideration in the early stage of learning.

Domain-Domain Pairs from Protein-Protein Interactions

Domain-domain interactions, which are closely related to protein interaction sites, have been widely studied in recent years, due to the well-acknowledged concept of domain and the abundantly available data about protein interactions. Many domains contain regions for interactions and involve in some biological functions.

From data mining perspective, each protein sequence is a sequence of domains and the target patterns are correlated domain pairs. The correlated pairs have been inferred by various approaches. Sprinzak and Margalit (2001) extracted all over-represented domain pairs in protein interaction data and initially termed them as correlated sequence-signatures. Wojcik and Schachter (2001) generated interacting domain pairs from protein cluster pairs with enough interactions, where the protein clusters are formed by proteins with enough sequence similarities and common interacting partners. Deng et

al. (2002) used maximum-likelihood to infer interacting domain pairs from a protein interaction dataset, by modeling domain pairs as random variables and protein interactions as events. Ng et al. (2003) inferred domain pairs with enough integrated scores, integrating evidences from multiple interaction sources.

Although domain-domain interactions imply abundant information about interaction sites, domains are usually very lengthy, in which only small parts involve binding while most regions contribute to folding. On the contrary, some interaction sites may not occur in any domain. Therefore, the study of domain-domain interactions is not enough to reveal interaction sites, although helpful.

Binding Motif Pairs from Protein-Protein Interactions

To fill the gap left by all above techniques, a novel concept of binding motif pairs has been proposed to define and capture more refined patterns at protein interaction sites hidden among abundant protein interaction data (Li and Li, 2005a). Each binding motif pair consists of two traditional protein motifs, which are usually more specific than domains. Two major methods were developed for the discovery of binding motif pairs.

The first method is based on a fixed-point theorem, which describes the stability under the resistance to some transformation at some special points; that is, the points remain unchanged by a transformation function. Here the stability corresponds to the biochemical laws exhibited in protein-protein interactions. The points of function are defined as protein motif pairs. This transformation function is based on the concept of occurrence and consensus. The discovery of fixed points, or the stable motif pairs, of the function is an iterative process, undergoing a chain of changing but converging patterns (Li and Li, 2005b).

The selection of the starting points for this function is difficult. An experimentally determined protein complex dataset was used to help in identifying meaningful starting points so that the biological evidence is enhanced and the computational complexity is greatly reduced. The consequent stable motif pairs are evaluated for statistical significance, using the unexpected frequency of occurrence of the motif pairs in the interaction sequence dataset. The final stable and significant motif pairs are the binding motif pairs finally targeted (Li and Li, 2005a).

The second method is based on the observation of frequently occurring substructures in protein interaction networks, called interacting protein-group pairs corresponding to maximal complete bipartite subgraphs in the graph theory. The properties of such substructures reveal a common binding mechanism between the two protein sets attributed to all-versus-all interaction between the two sets. The problem of mining interacting protein groups can be transformed into the classical problem of mining closed patterns in data mining (Li et al., 2007). Since motifs can be derived from the sequences of a protein group by standard motif discovery algorithms, a motif pair can be easily formed from an interacting protein group pair (Li et al., 2006).

FUTURE TRENDS

Current approaches to discover interaction sites from protein-protein interactions usually suffer from ineffective models, inefficient algorithms and lack of sufficient data. Many studies can be done in the future exploring various directions of the problem. For example, new machine learning methods may be developed to make full use of existing knowledge on the successfully docked protein complexes. More effective models may be created to discover binding motifs from homologous proteins or from domain-domain interacting pairs, through effective distinguishing of binding patterns from folding patterns. Other diverse models such as quasi-bipartite may be developed to discover binding motif pairs. Overall, the focus of studying interactions will then move from protein-protein interactions and domain-domain interactions to motif-motif interactions.

CONCLUSION

In this mini review, we have discussed various methods for the discovery of protein interaction sites. Protein-protein docking is the dominant method but it is constrained by the limited amount of protein structures. Classification methods constitute traditional machine learning methods but suffer from inaccurate negative data. Pattern mining methods are still incapable to distinguish binding patterns from folding patterns, except a few work based on binding motif pairs. Overall, current biological data mining methods are far from perfect. Therefore, it is valuable to develop novel methods in

the near future to improve the coverage rate, specificity and accuracy, especially by using the fast growing protein-protein interaction data, which are closely related to the interaction sites.

REFERENCES

Aloy, P., Pichaud, M., & Russell, R. (2005). Protein complexes: structure prediction challenges for the 21st century. *Current Opinion in Structural Biology*, 15(1), 15-22.

Aloy, P., & Russell, R. (2004). Ten thousand interactions for the molecular biologist. *Nature Biotechnology, 22*(10), 1317-1321.

Agrawal, R & Srikant, R. (1995). Mining sequential Patterns. *Proceedings of International Conference on Data Engineering (ICDE)* (pp. 3-14).

Bonvin, A. (2006). Flexible protein-protein docking. *Current Opinion in Structural Biology, 16*(2), 194-200.

Bordner. AJ & Gorin, AA (2007). Protein docking using surface matching and supervised machine learning. *Proteins*, 68(2), 488-502.

Carter, P., Lesk, V., Islam, S., & Sternberg, M. (2005). Protein-protein docking using 3d-dock in rounds 3,4, and 5 of CAPRI. *Proteins, 60*(2), 281-288.

Deng, M., Mehta, S., Sun, F., & Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Research, 12*(10), 1540-1548.

Dziembowski, A., & Seraphin, B. (2004). Recent developments in the analysis of protein complexes. *FEBS Letters, 556*(1-3), 1-6.

Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *European Journal of Biochemistry, 269*(5), 1356-1361.

Joel, R., & David, A. (2001). Predicting protein-protein interactions from primary structure. *Bioinformatics, 17*(5), 455-460.

Jonassen, I. (1997). Efficient discovery of conserved patterns using a pattern graph. *Computer Applications in Biosciences, 13*(5), 509-522.

Keskin, O., & Nussinov, R. (2005). Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Engineering Design & Selection, 18*(1), 11-24.

Kumar, S., Ma, B., Tsai, C., Sinha, N., & Nussinov, R. (2000). Folding and binding cascades: dynamic landscapes and population shifts. *Protein Science, 9*(1), 10-19.

Li, H., & Li, J. (2005a). Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction datasets. *Bioinformatics, 21*(3), 314-324.

Li, H., Li, J., & Wong, L. (2006). Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics, 22*(8), 989-996.

Li, J., & Li, H. (2005b). Using fixed point theorems to model the binding in protein-protein interactions. *IEEE transactions on Knowledge and Data Engineering, 17*(8), 1079-1087.

Li, J., Liu, G. Li, H., & Wong, L. (2007). A correspondence between maximal biclique subgraphs and closed pattern pairs of the adjacency matrix: a one-to-one correspondence and mining algorithms. *IEEE transactions on Knowledge and Data Engineering*, 19(12), 1625-1637.

Lorenzen, S & Zhang, Y (2007). Identification of near-native structures by clustering protein docking conformations. *Proteins*, 68(1), 187-194.

Lupyan, D., Leo-Macias, A., & Ortiz, A. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics, 21*(15), 3255-3263.

Mendez, R., Leplae, R., Lensink, M., & Wodak, S. (2005). Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins, 60*(2), 150-169.

Ng, S., Zhang, Z., & Tan, S. (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics, 19*(8), 923-929.

Ofran, Y., & Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. *FEBS Letters, 544*(3), 236-239.

Smith, H., Annau, T. M., & Chandrasegaran, S. (1990). Finding sequence motifs in groups of functionally related proteins. *Proceedings of National Academy of Sciences, 87*(2), 826-830.

Sprinzak, E., & Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology, 311*(4), 681-692.

Thompson, J., Higgins, D., & Gibson, T. (1994). ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research, 22*(22), 4673-4680.

Wang, H., Segal, E., Ben-Hur, A., Koller, D., & Brutlag, D. (2005). Identifying protein-protein interaction sites on a genome-wide scale. *Advances in Neural Information Processing Systems 17* (pp. 1465-1472). USA.

Wojcik, J., & Schachter, V. (2001). Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics, 17*(Suppl 1), S296-S305.

Yan, C., Dobbs, D., & Honavar, V. (2004). A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics, 20*(Suppl 1), I371-I378.

Yan, X., Yu, P. S., & Han, J. (2005). Substructure similarity search in graph databases. In *proceedings of 2005 ACM-SIGMOD International Conference on Management of Data* (pp. 766-777). Baltimore, Maryland.

KEY TERMS

Binding Motif Pairs: A pair of binding motifs which interact with each other to determine a type of protein interaction.

Binding Motifs: The patterns which describe a group of sequences or structures that bind or interact with a specific target.

Bioinformatics: The research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those used to acquire, store, organize, analyze, or visualize such data.

Domain-Domain Interactions: The binding or association among two or more domains due to their sequence or structure preference.

Protein Interaction Sites: The regions of proteins associated with the other interacting partner during protein-protein interactions.

Protein-Protein Docking: The determination of the molecular structures of complexes formed by two or more proteins without the need for experimental measurement.

Protein-Protein Interactions: The association of protein molecules and the study of these associations from the perspective of biochemistry, signal transduction and networks.