Chapter VII
Discovering Interaction Motifs from Protein Interaction Networks

Hugo Willy
National University of Singapore, Singapore

ABSTRACT

Recent breakthroughs in high throughput experiments to determine protein-protein interaction have generated a vast amount of protein interaction data. However, most of the experiments could only answer the question of whether two proteins interact but not the question on the mechanisms by which proteins interact. Such understanding is crucial for understanding the protein interaction of an organism as a whole (the interactome) and even predicting novel protein interactions. Protein interaction usually occurs at some specific sites on the proteins and, given their importance, they are usually well conserved throughout the evolution of the proteins of the same family. Based on this observation, a number of works on finding protein patterns/motifs conserved in interacting proteins have emerged in the last few years. Such motifs are collectively termed as the interaction motifs. This chapter provides a review on the different approaches on finding interaction motifs with a discussion on their implications, potentials and possible areas of improvements in the future.

INTRODUCTION

Protein interaction plays an essential role in a vast number of known biological processes. It is responsible in the formation of multimeric protein complexes, signal transduction, cell regulation and immune response processes. The interaction can be permanent, with relatively high binding affinity, and usually lasts throughout the protein's lifetime. This type of interaction can often be seen in the binding of different subunits of a permanent protein complex (the *obligate* interaction). The second type of interaction is a temporary, mostly of lower affinity (the *transient* interaction). An example of this type of interaction is the protein binding cascade commonly seen in the cellular signal transduction process (Jones & Thornton, 1996; Ofran & Rost, 2003). A good understanding of the mechanism underlying these protein-protein interactions is essential to the understanding of the biological system as a whole and potentially aid research on diseases and finding novel drug targets.

Protein interaction can be identified using quite a number of different biological experiments. Some are of high accuracy but has low throughput – which makes such methods too costly but for a few very important interactions. An example of such experiments would be the methods to determine the structure of protein complexes such as X-Ray crystallography or NMR spectroscopy. Recent years had witnessed the breakthrough of high throughput protein interaction identification technique like the Yeast Two Hybrid (Y2H) techniques and the Tandem Affinity Purification coupled with Mass Spectrometry (TAP-MS). But, as discussed in (Sprinzak *et. al.*, 2003), the higher throughput come with the cost of a significant reduction in accuracy. For yeast-two-hybrid, the expected number of false positive interaction could reach 50% and beyond. Nevertheless, given the wealth of newly generated interaction data, a number of computational methods had been devised and shown to perform reasonably accurate interaction prediction in-silico. Their potential in discovering novel interactions and cleaning up the noisy high-throughput interaction data using relatively inexpensive computational methods has generated a lot of work on this problem. Most of the approaches are machine learning based, combining multiple information source, each of which gives a certain degree of confidence on the protein interaction being investigated. For a broader review on works done on computational approach of protein interaction prediction, readers are referred to the excellent reviews in (Valencia & Pazos, 2002) and a more recent one in (Skrabanek *et. al.*, 2008). However, most of the approaches described above, both biochemical and computational based, are not able to elucidate the mechanism underlying the observed protein interaction. Specifically, most of them are only able to say if two proteins are interacting, but not on how or which parts of the proteins cause them to interact. Such details could only be seen by solving the 3D structure of the interacting protein. To date, the amount of structural data is rapidly increasing but their coverage is still limited in comparison to the existing high-throughput interaction data.

This chapter will discuss several computational approaches emerging over the years on mining the interaction motif. These approaches attempt to mine motifs/signatures within protein sequences which are highly indicative of protein interaction. The interaction motif could be motifs on protein binding sites, enzyme active sites or some structurally important sites which indirectly influence protein interaction. By knowing such motifs, we could possibly predict protein interaction based on the occurrence of interaction motifs in the candidate protein interaction in question. The main underlying assumption on interaction motif mining is that interaction motifs are conserved throughout evolution given its role in protein interaction. Thus one should be able to see such motifs enriched in an adequately large protein interaction data. The motifs can be of two general types; one is the structural, non-linear (e.g. the protein domains) and the other is the linear peptide motif. Since domain interaction is already covered in the other chapter of this book, we would only cover methods which are not solely based on interacting domains (i.e., the domain-linear motifs and linear motifs-linear motifs based methods). It is known that not all protein interactions are mediated by pairs of protein domains. Many of them involve the binding of a protein domain to a short peptide sequence with a defined linear sequence motif, which is the SLiMs (Pawson & Scott, 1997; Sudol, 1998; Neduva & Russell, 2006). The listing of all known SLiMs to date could be found in ELM (Puntervoll *et. al.*, 2003) or MiniMotif (MnM) database (Balla *et. al.*, 2006) .

There are multiple challenges in finding interaction motifs. The most fundamental one would be the limitation of the source interaction data. To be able to confidently say that a motif is essential for interaction, one would expect the motif to be enriched in interacting pair of proteins but not in the non-interacting ones. However, as we mentioned earlier, the available high throughput interaction data is highly noisy. Furthermore, to be able to design a protein interaction prediction system, we must have both interacting and non-interacting protein data. The latter is known to be scarce and most of the time, it is approximated using random pairs or pairs of protein with different cellular localization (Guo *et. al.,* 2008). The second problem is on the nature of linear motifs in particular. Since most of the time linear motifs are degenerate, one could find a huge number of candidate instances of them all over the interaction data. Hence, we would need to design a careful filtering and scoring methodology to fish out the true motif instances to be able to make meaningful inference of an interaction event based on the existence of linear motifs.

The following section will discuss the existing methodologies to find interaction motifs from the protein interaction data. Then we present some discussions on the challenges, issues and possible future work in this area.