Computational prediction of protein-protein interactions

Alex Wilkinson
BIOC218
Spring 2012

Introduction:

Structural biology and molecular biology are complementary approaches that are able to inform our mechanistic understanding of biological systems. For technical reasons, our understanding of proteins and their interactions has primarily come by way of molecular biology. However, our knowledge of high-resolution crystal structures is growing as evidenced by the many entries into the Protein Data Bank (PDB). Also, some mechanistic questions can only be answered by looking directly at structural information. The need for structural information has outpaced by the ability of the scientific community to attain this information. Computational methods to attain this structural information *in silico* would increase the availability of this type of information and expedite scientific progress.

Simple pull-down experiments are a well-established way of determining protein-protein interactions. However, the way in which associated proteins interact with one another is not a trivial question. Computational algorithms to circumvent the need for a crystal structure would allow labs that do not specialize in crystallography a platform for identifying potential interaction surfaces to inform the mechanism of interaction. Similarly, the rise of systems biology would benefit immensely from being able to predict interaction networks based on high-confidence interaction networks (Figure 1). A community-wide effort has been established to assess the progress of the field. This program called CAPRI (Critical Assessment of PRotein Interactions) has been established with the goal to test the newest algorithms in a blind prediction of protein interactions (Janin et al., 2003).

Two general types of docking problems exist and are assessed at CAPRI: bound and unbound (Janin et al., 2003). Crystal structures can be taken from a complex of proteins, separated, and individually recombined in their proper orientation. This example of docking is called bound docking. Unbound docking takes structures that are already solved individually and then tries to find how those proteins interact with one another. The major different is that bound structures have already undergone conformational changes, if any, which helps the prediction. While both problems are important, the unbound docking presents the more physiologically relevant problem. Bound docking provides a nice background to troubleshoot new algorithms and changes to algorithms. The goal of this review is to examine strategies of protein interaction prediction and the challenges facing current algorithms. In the process, several algorithms will be mentioned that represent some different strategies for predicting protein-protein interactions.

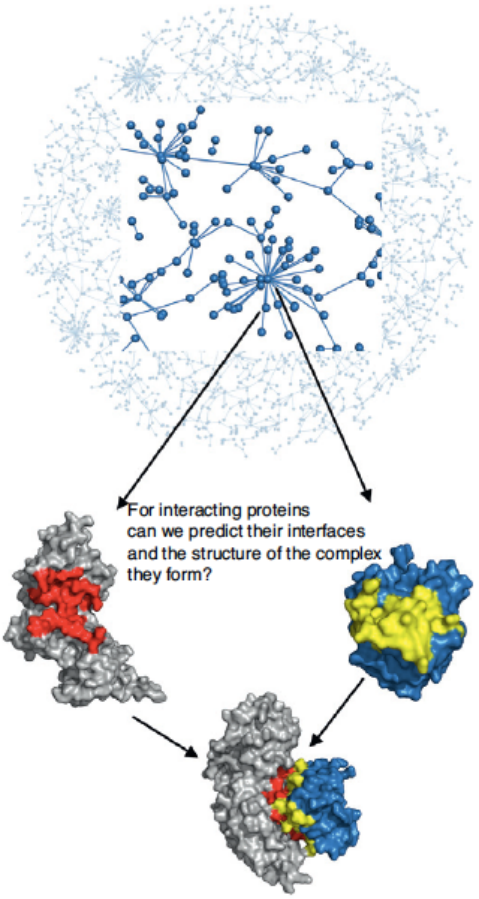

Figure 1. Representation of predictive protein interactions on a pairwise and proteome-wide scale. (Wass et al., 2011)

Strategies:

The computational determination of protein interactions is known as docking. The docking procedure has evolved from static assignment of known crystal structures of individual proteins to a more complex format that accounts for more variables. A generic workflow can be

summarized by 1) a docking step that provides an interface for two proteins to interact, 2) an optimization step and 3) a scoring step to determine how well the model fits. Before, during, or after any of these steps, energy optimization analyses can be performed to maximize the 'fit,' which will be explained in more detail later.

Rigid Docking Strategies.

Many docking strategies use rigid docking during some phase of the structural analysis. Some systems use the rigid docking as a more central component of the algorithm. The basic premise is that the three dimensional structures are used where a single protein is fixed and the other is rotated in six-dimensional translational space (Zacharias, 2010) . Such methods often use fast Fourier Transform (FFT) or geometric matching (Kozakov et al., 2006; Mintseris et al., 2007; Shen et al., 2007; Wiehe et al., 2007) to analyze the docking permutations.

One algorithm that takes advantage of FFT to align protein structures is ZDOCK. The early versions of this algorithm took into account six-dimensional orientations of the ligand around the fixed protein of interest and desolvation energy, the energy when a protein-water interaction is replaced with a protein-protein interaction (Chen and Weng, 2002). In this study, 16 complexes were able to be resolved within the top 20 matches where the desolvation energy played a significant role in determining correct structural relationships. It should be noted that this algorithm is returned hits that were near native state, structurally. These parameters were improved upon with the complementary ZRANK software that optimizes the energetic information for top hits to increase the number of true hits in the top 1000 structures (Pierce and Weng, 2007).

PIPER is another algorithm that uses FFT in a similar way. However, this algorithm was the first in incorporate structural information into the algorithm. Using this idea, PIPER uses

pairwise structure-based potentials to increase the number of near native structures that are determined (Kozakov et al., 2006). When ZDOCK and PIPER were compared against one another, PIPER provided a higher percentage of hits within 10Å root mean square deviation (RMSD) for many protein-protein interactions (Kozakov et al., 2006). ZDOCK was later modified to take pairwise structure-based potentials into account. In doing so, ZDOCK interface energies were highly correlated with those produced using PIPER, which suggests that these two algorithms independently produce similar interaction types (Mintseris et al., 2007). These interactions among others' work show how useful current experimental information can be.

Flexible Docking Strategies

A problem with the rigid-docking strategies is that they rely on structures that change very little upon binding to each other. Realistically, protein-protein interactions require a mechanism more akin to an induced-fit model. Proteins are dynamic structures where conformational changes are very much important for their function. To address these issues in a computational protein docking, several algorithms have been developed that are more successful than the rigid docking algorithms at dealing with these types of interactions. Generically, these methods do quick searches to find low energy conformations then take these hits then subsequently refine the conformations of the proteins more precisely.

The RosettaDock algorithm starts out with a low resolution docking stage to find interaction sites between the query proteins (Gray et al., 2003). Once these proteins are docked to one another, the side chains along the interface can be altered and manipulated to minimize the energy of the system. This method is performed using a Monte Carlo minimization technique that provides a compromise between speed and accuracy (Andrusier et al., 2008). This method was improved later to incorporate backbone flexibility using a fold-tree method that takes into

account torsional and rigid-body degrees of freedom (Wang et al., 2007a). This backbone flexibility did not improve docking predictions compared to previous CAPRI rounds, and the rigid-body docking algorithm ZDOCK actually outperformed the updated RosettaDock (Wang et al., 2007b).

ICM-DISCO (Docking and Interface Side-Chain Optimization) algorithm also takes advantage of Monte Carlo energy minimization when defining the input structures for docking (Fernandez-Recio et al., 2003; Grosdidier et al., 2007). However, this method was computationally very expensive as many different orientations and conformations of proteins were calculated before trying to dock the two proteins (Cheng et al., 2007). To relieve some of the computational time, this algorithm was modified to incorporate parameters such as electrostatic forces and desolvation energy to assist in finding a binding site. This method combines information learned in the ICM-DISCO algorithm and applies it to FFT to generate a much faster version of the previous ICM-DISCO method. pyDOCK can produce solutions within the top 100 models for 56% cases in a large benchmark dataset and within the top 20 solutions in 37% of the dataset (Cheng et al., 2007). This study highlights the importance of considering electrostatic interactions and desolvation energies. Updated versions of pyDOCK have resulted in algorithms such as pyDOCK*RST* that uses distance restraints from previously known data along the interface in addition to the electrostatic and desolvation parameters to help identify correct binding sites (Chelliah et al., 2006).

The HADDOCK (high ambiguity driven protein-protein docking) algorithm takes advantage of molecular dynamics to identify a correct fit (Dominguez et al., 2003). First energy minimization of rigid body conformations is performed followed by an annealing and refinement step are performed (Dominguez et al., 2003). In order to do this, the ambiguous interaction

restraints and the PDB crystal files are required. The updated version of HADDOCK2.0 can support data from NMR structures but can also perform interactions *ab initio*, where no experimental information is available (de Vries et al., 2007). HADDOCK2.0 was able of solving one-star level structure for CAPRI targets 100% of the time compared to the 65% that the original HADDOCK algorithm was able to achieve (de Vries et al., 2007). Examples of HADDOCK predictions can be seen in figure 2. While this algorithm is very successful, the molecular dynamics is very computationally expensive (Andrusier et al., 2008).

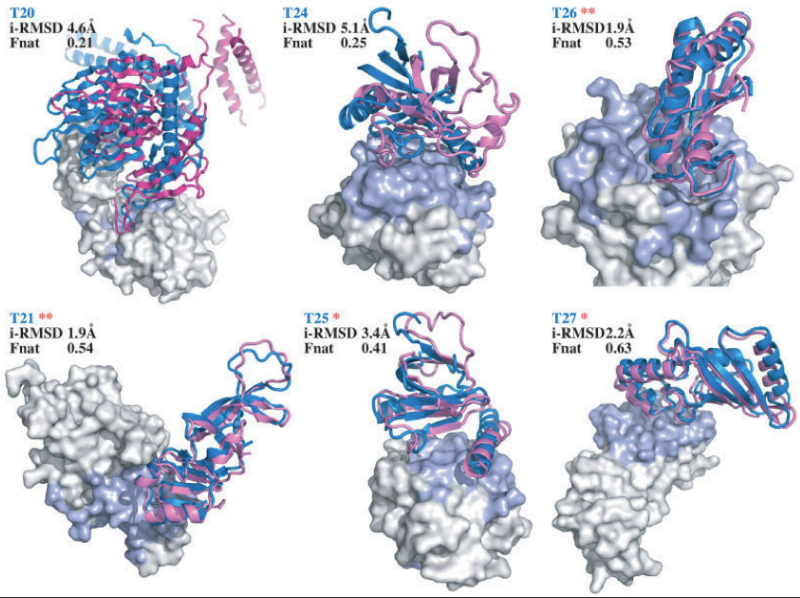

Figure 2. Protein interaction predictions for CAPRI targets T20, T21, and T24-27. A single asterisk represents a good match to the actual structure and two asterisks represents very good matches based on RMSD when compared to the actual structures (Mendez et al., 2003). The pink structures represent the prediction and the blue represents the target structure. Adapted from de Vries et al., 2007.

Conclusion

The algorithms for discovering protein-protein interactions are still very much in their infancy and are improving with each round of CAPRI. This trajectory will only increase as the technology and computing power increases over time. The consistent performance of the HADDOCK algorithm at CAPRI highlights the abilities of algorithms that can take advantage of molecular dynamics to accurately predict the structures of protein-protein interactions. It is interesting that the modification of the ICM-DISCO algorithm to pyDOCK provided results

comparable to other rigid-docking programs. Similarly, the addition of a backbone flexibility to the RosettaDock algorithm did not necessarily improve its ability predict proper docking. These examples suggest that finding the optimal set of parameters to emphasize will be an important factor as the field develops.

The way different parameters are incorporated into the algorithms is also interesting as most of the actual computing is fairly similar between algorithms (Vajda and Kozakov, 2009). For example, RosettaDock uses Monte Carlo minimization to optimize energy levels once a docking position is found. However with ICM-DISCO, the Monte Carlo minimization was performed to find templates for docking. Innovation in computer science and mathematics will also stimulate our predictive abilities in the long run. One limiting resource may be our recycling of methodology. The growing field of bioinformatics will only help this problem as more people are directed into the realm of computational structural and molecular biology.

Another roadblock that keeps appearing is the challenge of dealing with conformational changes when using two unbound proteins as inputs. Large conformational changes appear to favor the algorithms such as HADDOCK that incorporate flexible backbones and not the rigid-docking algorithms. Again, this suggests the need for increased computational power. This problem is particularly important in the setting of a biology lab. Publication-quality structures would need to have an extremely high confidence in accuracy. Obviously, we are far away from that point. However, useful information can still be gleaned from the structures that we are getting today. Many times, structural information is important for the identification of important protein sites that can be mutated to abrogate a protein-protein interaction. The predictive powers of today's algorithms may be sufficient to gain some information in this manner. There is

certainly a place for computational predictions of protein-protein interactions and when the algorithms become more reliable, it will become an invaluable tool.

The state of the field today leaves much to be desired but is making steady progress. The co-evolution with *de novo* protein structure prediction using computational methods will very important for informing prediction of protein-protein interactions and vice versa. This is particularly true because structural data is needed to even begin trying to predict protein-protein interactions. Regardless, this era of protein-protein interaction predictions has just started and we good reason to be excited about and have high expectations for the field.

References

Andrusier, N., Mashiach, E., Nussinov, R., and Wolfson, H.J. (2008). Principles of flexible protein-protein docking. Proteins *73*, 271-289.

Chelliah, V., Blundell, T.L., and Fernández-Recio, J. (2006). Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patters with those predicted from local environment. Journal of molecular biology *357*, 1669-1682.

Chen, R., and Weng, Z. (2002). Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins *47*, 281-294.

Cheng, T.M., Blundell, T.L., and Fernandez-Recio, J. (2007). pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins *68*, 503-515.

de Vries, S.J., van Dijk, A.D.J., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A.M.J.J. (2007). HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins *69*, 726-733.

Dominguez, C., Boelens, R., and Bonvin, A.M.J.J. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. Journal of the American Chemical Society *125*, 1731-1737.

Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. Proteins *52*, 113-117.

Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A., and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of molecular biology *331*, 281-299.

Grosdidier, S., Pons, C., Solernou, A., and Fernández-Recio, J. (2007). Prediction and scoring of docking poses with pyDock. Proteins *69*, 852-858.

Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J.E., Vajda, S., Vakser, I., Wodak, S.J., and Interactions, C.A.o.P. (2003). CAPRI: a Critical Assessment of PRedicted Interactions. Proteins *52*, 2-9.

Kozakov, D., Brenke, R., Comeau, S.R., and Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. Proteins *65*, 392-406.

Mendez, R., Leplae, R., De Maria, L., and Wodak, S.J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins *52*, 51-67.

Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R., and Weng, Z. (2007). Integrating statistical pair potentials into protein complex prediction. Proteins *69*, 511-520.

Pierce, B., and Weng, Z. (2007). ZRANK: reranking protein docking predictions with an optimized energy function. Proteins *67*, 1078-1086.

Vajda, S., and Kozakov, D. (2009). Convergence and combination of methods in protein-protein docking. Current opinion in structural biology *19*, 164-170.

Wang, C., Bradley, P., and Baker, D. (2007a). Protein–Protein Docking with Backbone Flexibility. Journal of molecular biology *373*, 503-519.

Wang, C., Schueler-Furman, O., Andre, I., London, N., Fleishman, S.J., Bradley, P., Qian, B., and Baker, D. (2007b). RosettaDock in CAPRI rounds 6-12. Proteins *69*, 758-763.

Wass, M.N., David, A., and Sternberg, M.J.E. (2011). Challenges for the prediction of macromolecular interactions. Current opinion in structural biology *21*, 382-390.

Zacharias, M. (2010). Accounting for conformational changes during protein-protein docking. Current opinion in structural biology *20*, 180-186.