Predicting Protein Complex Structures: A Review of the Docking Process

Adam Perez
BIOC218 Final Project
12/11/2011

Introduction

Proteins carry out enzymatic reactions and participate in cellular processes that are a necessary component of biological systems. A key aspect to the integrity of many proteins' functional roles is their interaction with other proteins. Indeed, identification of all the protein interactions within the cell indicates that proteins associate with a few to hundreds of partners (Gavin et al., 2002; Rual et al., 2005). However, these studies fail to indicate the functional biological role, if any, of these protein interactions. Understanding the function of these interactions, and ultimately the processes within a cell, requires knowledge of the three-dimensional structure of the interacting complex. However, experimentally obtaining the structure of a protein complex at high-resolution via x-ray crystallography and NMR is technically difficult and low-throughput, and very few of these structures have been solved. Thus, computationally modeling the structure of predicted protein-protein interaction complexes has emerged as a significant biological challenge.

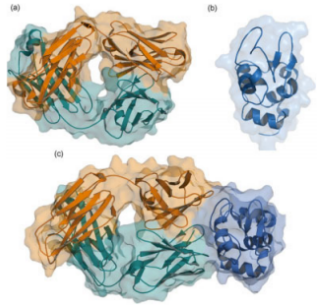

Fig. 1 Protein Docking X-ray structure of a) FAB Hyhel63 b) HEW c) complex (Moreira et al., 2010)

Computationally predicting the structure of two bound proteins is known as docking. The problem involves starting with the coordinates obtained from individual structures of each protein in the interacting pair, and then modeling the structure of the bound complex (Andrusier et al., 2008; Ritchie, 2008)(Fig. 1). The results of Critical Assessment of Predicted Interactions (CAPRI), a community wide experiment to evaluate the myriad methods used to dock proteins, show an improvement of protein docking methods over the past decade (Janin et al., 2003; Lensink et al., 2007). However, the problem is far from solved, and many challenges still remain. This review covers the general procedure and theory used in docking, offers an assessment of selected docking programs, and gives an outlook of the current progress towards overcoming the difficulties encountered during docking.

General Docking Overview

The original view of protein-protein interactions is characterized by Emil Fischer's "lock and key" model that emphasized the importance of steric effects at the binding interface towards achieving binding specificity. In this model, the structure of one protein provides a pocket that allows the complementary structure of a binding protein to fit tightly into. This would be ideal in the application of protein docking, as the shapes of the unbound and bound molecules would not significantly differ. In this case, programs could move one structure rotationally and translationally around the six dimensions of the cognate protein's stationery structure until the two meet at a physicochemical and geometrical complementary interface.

However, the view of protein binding has evolved to show that it does not occur between two extremely rigid structures. In fact, binding can and often does result in conformational changes of the interacting molecules, and specificity is achieved as the two molecules conformationally adapt to each other as they bind (Koshland, 1958). This complicates the

docking procedure, as conformational changes must be introduced into the interacting cognate structures in order to produce an accurate docked structure.

In light of these views, many of the developed docking programs all essentially share common computational steps (Vajda and Kozakov, 2009). These steps aim to produce general models that are systematically refined until near native models are produced. The first step is a rigid body search that globally scans the bodies of the proteins for a potential interaction site; the second step is to select a region of interest, the conformations generated from step one that are close to a native structure; the third step is structure refinement to increase the fidelity of the structure; and the last step is to select accurate models. This process is outlined in Fig. 2 (Vajda and Kozakov, 2009).

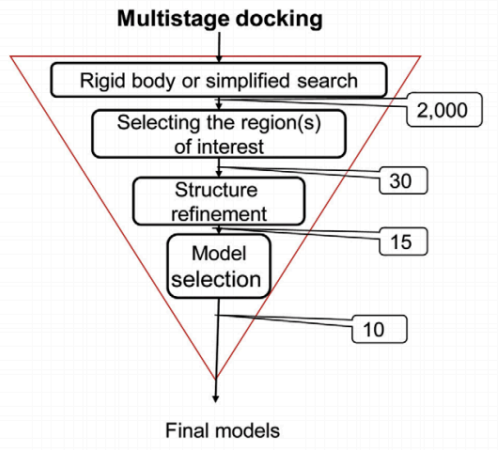

Fig 2. Steps in the Docking Protocol
The number of models typically retained at each step is indicated (Zacharias, 2010)

Step 1: Rigid Body/Simplified Geometry Search

The first step in protein docking is to globally search the entire rotational and translational space of the proteins to identify potential interaction sites. Due to the high computational cost associated with such a search, these methods utilize simplified protein models and base identification of putative interaction sites largely on geometric shape, electrostatic, and hydrophobic complementarily (Janin, 2010; Moreira et al., 2010). Simplified protein models "soften" the atomic details so that the protein can be represented as a discrete geometric shape. Geometric matching of the two proteins to be docked can then be performed on these simplified models to indentify putative interaction sites. Fast Fourier transforms (FFTs) are frequently utilized to perform these searches. In this method, the protein models are made up of sets of cubes. Calculation of a correlation function via FFT indentifies regions where the two geometrical surfaces overlap significantly while excluding regions of geometrical interpenetration (Katchalski-Katzir et al., 1992).

However, geometric constraints are not the only factors dictating interaction events. Chemical properties between interacting residues as well as changes in free energy upon binding are also critical determinants of binding. The ZDOCK program includes desolvation and electrostatic terms during the FFT that improves the accuracy of predicting interaction sites (Chen and Weng, 2002).

In addition to FFTs, geometric hashing is also utilized in this first step of protein docking. As an example, the PatchDock program creates a Connolly-style representation of the protein that is restricted to three critical points: concave, convex, and flat patches. The representations are scanned against each other to indentify configurations that contain high degrees of geometric matching (Schneidman-Duhovny et al., 2003).

The methods described above perform docking based on rigid structures. However, potential interactions can be missed if binding induced conformational changes occur within either structure that would otherwise occlude geometric matching. As such, the algorithms must allow some amount of steric overlap to account for conformational change (Zacharias, 2010). A delicate balance must be achieved in this restraint, however, as allowing too much overlap will result in the generation of many more false positive configurations.

Step 2: Selection of the Region of Interest

As shown in Fig. 2, rigid body docking typically results in the generation of around 2,000 candidate structures. These structures are generated using soft models and often include steric overlap. As such, the atomic resolution of the structure does not closely match the native configuration. Modeling side chain and backbone conformational changes that occur upon induced fit binding is required to generate near native structures. Due to the high number of degrees of freedom in modeling these changes, the process is computationally expensive. Thus the candidate models generated in step one must be reduced to a manageable number that can be further refined (Vajda and Kozakov, 2009).

Scoring functions must be able to correctly choose candidate solutions that are close to the native structure, but must also allow for steric clashes that occur during rigid docking. A common approach is to retain the lowest energy structures or a set of structures that cluster into discrete energy bins (Lorenzen and Zhang, 2007; O'Toole and Vakser, 2008; Vajda and Kozakov, 2009). In addition, terms representing physicochemical properties at the binding interface, empirical evidence of residue contacts at the interface, and residue conservation can be included in and increase the specificity of the scoring program (Janin, 2010).

Step 3: Structure Refinement

Structural refinement is highly focused on changes made to the binding interface of the complex. Energy minimization by removal of steric overlap and optimization of electrostatic and hydrophobic interactions is a straightforward method to refine structures, and works well to increase the number of near native structures in a candidate pool (Liang et al., 2009; Wiehe et al., 2005; Wiehe et al., 2007).

Monte Carlo methods are also widely used during refinement. In this approach, a random residue is selected and its rotamer switched for another. The overall energy change is calculated and the switch is retained if this change minimizes the energy of the bound structure (Holm and Sander, 1992). The process is repeated in an iterative manner until energy minimization is achieved. As expected, this process is lengthy and computationally expensive. To restrict the search space, some methods utilize side-chain rotamer libraries. These libraries are curated from statistical analysis of known protein structures and show that side-chains tend to adopt certain configurations based upon the orientation of the backbone (Dunbrack and Karplus, 1993). Programs can effectively utilize this information during docking by placing all side chains into favorable starting positions. The RosettaDock program takes further advantage of this procedure by causing small perturbations to the backbone before adding in a side-chain rotamer library (Gray et al., 2003). In this way, both the backbone and side-chain atoms are refined to simulate induced fit. The HADDOCK is another program that models these changes, but uses molecular dynamics in place of Monte Carlo methods (Dominguez et al., 2003). While these two refinement strategies are powerful, they typically require experimental information about the known binding site in order to be successful.

Step 4: Model Selection

This last step must choose candidate models that most closely resemble the native structure. This part of the procedure is similar to step 2, but as it is scoring refined models the programs do not have to allow for the inaccuracies that occur during soft docking.

Overview of Selected Docking Programs

Many docking programs utilize a global FFT or geometric hashing search followed by refinement using energy minimization (i.e. optimizing electrostatic interactions, hydrogen bonds, and van Der Waals forces) at the region of interest. These programs include FFT search programs ZDOCK and MolFit as well as the geometric hashing program PatchDock (Berchanski et al., 2004; Chen and Weng, 2002; Schneidman-Duhovny et al., 2003). MolFit handles refinement by allowing small rigid body rotations near the binding face (Berchanski et al., 2004; Kowalsman and Eisenstein, 2007). Success for ZDOCK and PatchDock depends on running the outputs of these programs through the refinement of the side chains by energy minimization via the programs RDOCK and FireDock, respectively (Mashiach et al., 2008; Wiehe et al., 2007).

The advantage of this method is that no information other than the structures themselves is required to generate models. However, these methods do not account for conformational changes induced upon binding, and thus fail at predicting structures when the unbound and bound configurations of the proteins differ significantly. Also, weakly interacting complexes are difficult to identify during model selection (Zacharias, 2010). Interestingly, these programs boost their chances of success by accepting biochemical, physical, and/or sequence data to either limit the search to a specific region or to aid in the ranking of models (Moreira et al., 2010; Zacharias, 2010).

To account for the conformational changes that frequently occur during binding, programs have been developed to allow increased flexibility throughout the docking process. In order to reduce the computational cost of sampling the entire conformational space during binding, these programs must limit their searches to selected regions.

One program that employs this strategy is RosettaDock, which is able to model both backbone and side-chain flexibility (Chaudhury and Gray, 2008; Wang et al., 2007a). The procedure begins with a low-resolution global search. Low energy ranked structures then undergo a Monte Carlo minimization cycle that allows side-chains to repack in parallel to backbone movement (Janin, 2010). Due to the higher resolution of the Monte Carlo method during the refinement stage, RosettaDock provided more accurate structures when compared to FFT/geometric hashing methods at CAPRI (Vajda and Kozakov, 2009). However, the high number of degrees of freedom allowed during docking in RosettaDock may result in the generation of more false positives (Wang et al., 2007b). The program also still fails at predicting complex structures whose component proteins undergo large conformational changes upon binding (Moreira et al., 2010).

The HADDOCK program has consistently performed well at CAPRI (de Vries et al., 2007). HADDOCK uses ambiguous restraints to approximately identify the native binding interface of two interacting proteins (Dominguez et al., 2003). These restraints arise from the experimental identification of residues implicated in binding of the two proteins from biochemistry and/or physical data. It is important to note that HADDOCK is the only program that uses such restraints in an ambiguous manner, as residues that are experimentally implicated in binding could be due to secondary factors and not a direct effect. During the refinement stage, simulated annealing occurs that allows movement of both the backbone and side-chains at the interface, and is followed by energy minimization and molecular dynamics (Dominguez et al., 2003). In contrast to other methods, HADDOCK models complexes whose individual components undergo significant conformational changes upon binding extremely well, the only disadvantage being that the program requires additional and accurate information to use as restraints along with the structures of the docked proteins (Moreira et al., 2010).

Results from CAPRI show that while the above programs substantially differ, they have similar success rates in predicting near-native structures (Vajda and Kozakov, 2009). It is also of note that each program handles certain problems well, while no program is able to correctly predict all structures. The FFT/geometric hashing methods can utilize just the structures to provide valid models that can be tested experimentally. Monte Carlo based methods such as RosettaDock can produce models that highly resemble native structures, but are not able to globally sample the entire conformational space. If reliable experimental evidence is available to provide restraints, HADDOCK is preferred to generate very accurate models (Vajda and Kozakov, 2009).

Docking Difficulties

CAPRI has shown that highly accurate models are generated for docking complexes whose interacting proteins undergo little conformational change when the programs are given experimental hints of the interaction interface (Janin et al., 2003; Lensink et al., 2007). However, the programs lose predictive power when interacting proteins undergo large conformational changes upon binding or when one of the protein structures (Andrusier et al., 2008; Bonvin, 2006; Lensink et al., 2007). Thus, a major challenge for the docking programs is to be able to accurately model conformational changes of proteins upon induced fit binding.

Programs such as RosettaDock are able to accurately model backbone and side chain flexible at the binding interface of a complex, but fail to model global protein movement and relaxation that occurs during binding. One way to model large-scale conformational changes is to represent a flexible protein as an ensemble of structures during the initial rigid docking step. These structures are gathered from experimental structural data or generated by structural modeling, and are cross-docked against the interacting protein (Andrusier et al., 2008; Zacharias, 2010). A key reason why HADDOCK generates accurate structures of complexes that undergo large conformational changes during formation is that it generates an ensemble of structures via a Monte Carlo method that are cross docked during the first rigid body scan (van Dijk et al., 2005). Molecular dynamics, genetic algorithms, and normal mode analysis can also be implemented to generate an ensemble of structures (Andrusier et al., 2008). This strategy, along with most strategies that model flexibility throughout the docking process, has the disadvantage in that it is computationally expensive and results in the generation of many false positives.

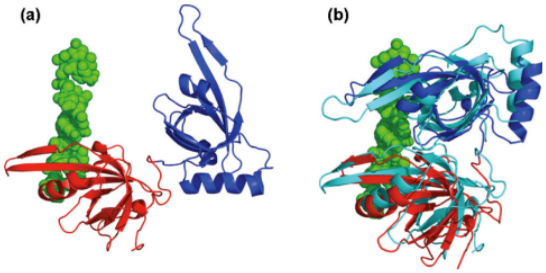

Fig. 3 Hinge Motion During Binding a) Unbound conformation of Replication Protein A (red and blue structure) b) Bound conformation of Replication Protein A (red and blue) and the predicted structure by FlexDock (cyan) (Andrusier et al., 2008)

Many times, a hinge bending motion occurs in a protein during docking (Andrusier et al., 2008)(Fig. 3). To take advantage of this observation, programs such as FlexDock have been developed. FlexDock identifies hinge points within a protein using a Gaussian Network Model (Emekli et al., 2008; Schneidman-Duhovny et al., 2005; Schneidman-Duhovny et al., 2007). The protein is divided into subdomains that are docked as rigid structures and subsequently models the continuous docked structure. While large scale conformational changes have been accurately modeled, these cases still present a significant challenge in CAPRI, indicating that future

docking programs will need to make progress on the current methods used to model these types of interactions.

Discussion

Many methods exist to dock proteins. Each method contains strengths and weaknesses that are highly dependent upon an individual docking problem. HADDOCK is very successful in generating highly accurate models when sufficient and reliable experimental data is available. FFT/geometric hashing methods can generate good predictions in many cases that can be used to design biochemical experiments. These insights reveal that at this point, a combination of computational modeling as well as biological experiments is necessary in order to truly understand the functional role of protein interactions.

A main limitation faced by virtually all docking programs is their inability to accurately and consistently predict the structure of complexes whose components undergo large conformational changes during binding. While ensemble docking (utilized by HADDOCK) and rigid body docking (utilized by FlexDock) have been used to accurately predict the structures of some of these cases, they cannot predict all cases and are still limited by computational power and the generation of false positives. Extreme conformational changes, such as refolding of proteins upon binding, would seem almost impossible to predict via the existing methods. Since this observation has been uncovered by CAPRI, the docking community will no doubt respond with new insights in solving the problem. Interestingly, the Baker group has utilized the Rosetta technique to simultaneously model protein refolding and docking (Das et al., 2009). Improvements in modeling protein folding can thus be assumed to only aid in the docking problem. Also, improvements in homology modeling can also be of use to predict the conformation of complexes whose individual structures are not yet solved.

The main goal of docking is to understand the function of protein-protein interactions in biological processes. Even though no current program can accurately solve all docking programs, it is likely that scientists can use the available methods to help design experiments and/or aid in the understanding of unique biological questions. The docking problem has made significant progress in the goals that it originally set out to reach. Advancements to the available methods should only make this realization much more clear.

References

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

Berchanski, A., Shapira, B., and Eisenstein, M. (2004). Hydrophobic complementarity in protein-protein docking. Proteins *56*, 130-142.

Bonvin, A.M. (2006). Flexible protein-protein docking. Curr Opin Struct Biol *16*, 194-200.

Chaudhury, S., and Gray, J.J. (2008). Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. J Mol Biol *381*, 1068-1087.

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

Das, R., Andre, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., Arrowsmith, C.H., Szyperski, T., and Baker, D. (2009). Simultaneous prediction of protein folding and docking at high resolution. Proc Natl Acad Sci U S A *106*, 18978-18983.

de Vries, S.J., van Dijk, A.D., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A.M. (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. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc *125*, 1731-1737.

Dunbrack, R.L., Jr., and Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol *230*, 543-574.
Emekli, U., Schneidman-Duhovny, D., Wolfson, H.J., Nussinov, R., and Haliloglu, T. (2008). HingeProt: automated prediction of hinges in protein structures. Proteins *70*, 1219-1227.
Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M.*, et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141-147.
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. J Mol Biol *331*, 281-299.
Holm, L., and Sander, C. (1992). Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. Proteins *14*, 213-223.
Janin, J. (2010). Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst *6*, 2351-2362.
Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J., Vajda, S., Vakser, I., and Wodak, S.J. (2003). CAPRI: a Critical Assessment of PRedicted Interactions. Proteins *52*, 2-9.
Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., and Vakser, I.A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A *89*, 2195-2199.
Koshland, D.E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci U S A *44*, 98-104.
Kowalsman, N., and Eisenstein, M. (2007). Inherent limitations in protein-protein docking procedures. Bioinformatics *23*, 421-426.
Lensink, M.F., Mendez, R., and Wodak, S.J. (2007). Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins *69*, 704-718.
Liang, S., Meroueh, S.O., Wang, G., Qiu, C., and Zhou, Y. (2009). Consensus scoring for enriching near-native structures from protein-protein docking decoys. Proteins *75*, 397-403.
Lorenzen, S., and Zhang, Y. (2007). Identification of near-native structures by clustering protein docking conformations. Proteins *68*, 187-194.
Mashiach, E., Schneidman-Duhovny, D., Andrusier, N., Nussinov, R., and Wolfson, H.J. (2008). FireDock: a web server for fast interaction refinement in molecular docking. Nucleic Acids Res *36*, W229-232.
Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2010). Protein-protein docking dealing with the unknown. J Comput Chem *31*, 317-342.
O'Toole, N., and Vakser, I.A. (2008). Large-scale characteristics of the energy landscape in protein-protein interactions. Proteins *71*, 144-152.
Ritchie, D.W. (2008). Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci *9*, 1-15.
Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N.*, et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.
Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H.J. (2005). Geometry-based flexible and symmetric protein docking. Proteins *60*, 224-231.
Schneidman-Duhovny, D., Inbar, Y., Polak, V., Shatsky, M., Halperin, I., Benyamini, H., Barzilai, A., Dror, O., Haspel, N., Nussinov, R.*, et al.* (2003). Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. Proteins *52*, 107-112.
Schneidman-Duhovny, D., Nussinov, R., and Wolfson, H.J. (2007). Automatic prediction of protein interactions with large scale motion. Proteins *69*, 764-773.
Vajda, S., and Kozakov, D. (2009). Convergence and combination of methods in protein-protein docking. Curr Opin Struct Biol *19*, 164-170.
van Dijk, A.D., Boelens, R., and Bonvin, A.M. (2005). Data-driven docking for the study of biomolecular complexes. Febs J *272*, 293-312.
Wang, C., Bradley, P., and Baker, D. (2007a). Protein-protein docking with backbone flexibility. J Mol Biol *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.

Wiehe, K., Pierce, B., Mintseris, J., Tong, W.W., Anderson, R., Chen, R., and Weng, Z. (2005). ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. Proteins *60*, 207-213.
Wiehe, K., Pierce, B., Tong, W.W., Hwang, H., Mintseris, J., and Weng, Z. (2007). The performance of ZDOCK and ZRANK in rounds 6-11 of CAPRI. Proteins *69*, 719-725.
Zacharias, M. (2010). Accounting for conformational changes during protein-protein docking. Curr Opin Struct Biol *20*, 180-186.