9

Flexible Protein-Protein Docking

Sebastian Schneider and Martin Zacharias
Technische Universität München/Physik Department T38,
85747 Garching,
Germany

1. Introduction

Biological processes almost always involve protein-protein interactions. Understanding the function of protein-protein interactions requires knowledge of the structure of the corresponding protein-protein complexes. The experimental structure determination by X-ray crystallography requires purification of large amounts of proteins. In addition, it is necessary to crystallize the proteins in the native complex which may not be feasible for all known interacting proteins. Multi-protein complexes mediate many cellular functions and are in a dynamic equilibrium with the isolated components or sub-complexes (Gavin et al., 2002; Rual et al., 2005). In particular, complexes of weakly or transiently interacting protein partners are often not stable enough to allow experimental structure determination at high (atomic) resolution. Experimental studies on detecting all protein-protein interactions in a cell indicate numerous possible interactions ranging from few to several hundred possible binding partners for one protein (Gavin et al., 2002; Rual et al., 2005). A full understanding of cellular functions requires structural knowledge of all these interactions. In the foreseeable future it will not be possible to determine the structure of all detected protein-protein interactions experimentally at high resolution. Structural modeling and structure prediction is therefore of increasing importance to obtain at least realistic structural models of complexes (Bonvin, 2006; Andrusier et al., 2008; Vajda & Kozakov, 2009; Zacharias, 2010). If the structure of the isolated protein partners or of closely related proteins is available it is possible to use a variety of computational docking methods to generate putative complex structures.

The driving force for the protein binding process corresponds to the associated change in free energy which depends on the structural and physicochemical properties of the protein partners. The "lock and key" concept of binding proposed by E. Fischer (Fischer, 1894) emphasizes the importance of optimal sterical complementarity at binding interfaces as a decisive factor to achieve high affinity and specificity. However, proteins and other interacting biomolecules are not rigid but can undergo a variety of motions even at physiological temperatures. The induced fit concept has evolved based on the observation that binding can result in significant conformational changes of partner molecules (Koshland, 1958). Within this concept protein partners induce conformational changes during the binding process that are required for specific complex formation. It should be emphasized, that in principle all possible molecular recognition processes require a certain degree of conformational adaptation. In recent years extensions of the induced-fit concept,

based on ideas from statistical physics, emerged. A pre-existing ensemble of several inter-convertible conformational states being in equilibrium has been postulated (Tsai et al., 1999). Among these states are structures close to the bound and unbound forms. Binding of a partner molecule to the bound form shifts the equilibrium towards the bound form. Since every conformation is, in principle, accessible albeit with a potentially low statistical weight already in the unbound state the original induced fit concept is a special case of ensemble selection where only the presence of a ligand gives rise to an appreciable concentration of the bound partner structure.

Progress in protein-protein docking prediction methods has been monitored with the help of the community wide Critical Assessment of Predicted Interactions (CAPRI) experiment (Janin et al., 2003; Lensink et al., 2007). In this challenge participating groups test the performance of docking methods in blind predictions of protein-protein complex structures. The results of the CAPRI challenge indicate that for protein partners with minor conformational differences between unbound and bound conformation and some experimental hints on the interaction region often quite accurate predictions of complex structures are possible (Bonvin, 2006; Andrusier et al., 2008; Zacharias, 2010). However, the docking problem becomes much more difficult when protein partners undergo significant conformational changes upon association or for protein structures based on comparative modeling (Andrusier et al., 2008; Zacharias, 2010). The magnitude of possible conformational changes during association can range from local alterations of side chain conformations to global changes of domain geometries and can even involve refolding of protein segments upon association. Computational approaches to realistically predict protein-protein binding geometries need to account for such conformational changes. Often, protein-protein complex structures obtained from protein-protein docking but also in case of comparative modeling are of limited accuracy and require further structural refinement to achieve the generation of a realistic structural model. Since rigid docking is computationally much faster compared to flexible docking, the majority of current protein-protein docking approaches distinguishes between a first exhaustive systematic docking search followed by a second refinement step of pre-selected putative complexes (Bonvin, 2006; Vajda & Kozakov, 2009). Docking protocols may even consist of several consecutive refinement and rescoring steps (Andrusier et al., 2008). In the present contribution recent progress in the area of protein-protein docking with an emphasis on modeling conformational changes and adaptation during protein binding processes will be discussed.

2. Protein protein docking

The purpose of computational protein-protein docking methods is to predict the structure of a protein-protein complex based on the structure of the isolated protein partners. If the structure of the isolated partner proteins is not known it is often possible to build structures based on sequence homology to a known structure using comparative modelling methods.

Receptor and ligand proteins are discretized on three-dimensional grids and are portioned into inside, surface and outside regions, respectively. Matching of surfaces is measured by the overlap of surface regions. For each ligand rotation with respect to the receptor the correlation problem is solved using Fast-Fourier-Transformation (FFT). After filtering and possible refinement steps solutions with high overlap of surface regions (high surface complementarity) are collected as putative solutions.
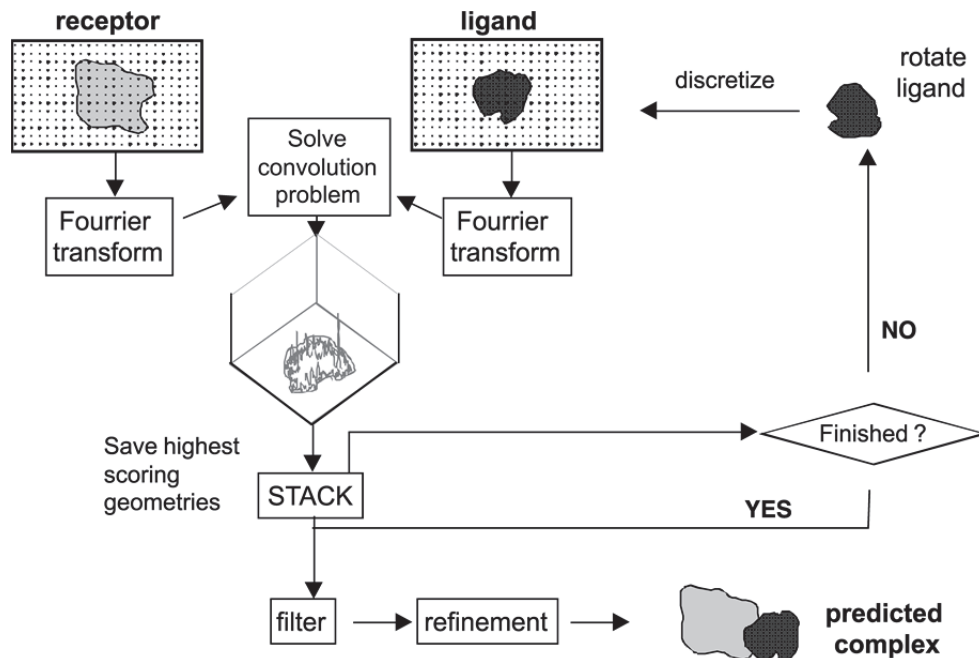
Fig. 1. Rigid protein-protein docking using Fast Fourier Transformation to solve a correlation problem.

2.1 Rigid docking methods

A variety of computational methods have been developed in recent years to efficiently generate a large number of putative binding geometries. The initial stage consists typically of a systematic docking search keeping partner structures rigid (Bonvin, 2006; Vajda & Kozakov, 2009). Subsequently, one or more refinement and scoring steps of a set of preselected rigid docking solutions are added to achieve closer agreement with the native geometry and to recognize near-native docking solutions preferentially as the best or among the best scoring complexes. In the initial search some unspecific sterical overlap between docking partners is typically tolerated to implicitly account for conformational adjustment of binding partners (e.g. Pons et al., 2009). Among the most common are geometric hashing methods to rapidly match geometric surface descriptors of proteins (Norel et al., 1994) and fast Fourier transform (FFT) correlation techniques to efficiently locate overlaps between complementary protein surfaces (Katchalski-Katzir et al., 1992). In the latter approach the two protein partners are represented by cubic grids, the grid points are assigned discrete values for inside, outside and on the surface of the protein. A geometric complementarity score can be calculated for the two binding partners by computing the correlation of the two grids representing each protein. Instead of summing up all the pair products of the grid entries one can make use of the Fourier correlation theorem. The corresponding correlation integral can easily be computed in Fourier space. The discrete Fourier transform for the receptor grid needs to be calculated only once. Due to the special shifting properties of Fourier transforms the different translations of the ligand grid with respect to the receptor

grid can be computed by a simple multiplication in Fourier space. This process is repeated for various relative orientations of the two proteins. A disadvantage of standard Cartesian FFT-based correlation methods is the need to perform FFTs for each relative orientation of one protein molecule with respect to the partner. This can be avoided by correlating spherical polar basis functions that represent, for example, the surface shape of protein molecules. It has been successfully applied in the field of protein-protein docking (Ritchie et al., 2008). Recently, new multidimensional correlation methods have been developed that allow the correlation of multi-term potentials. Each function needs to be expressed in terms of spherical basis functions characterizing the surface properties of the protein partners (Ritchie et al., 2008; Zhang et al., 2009).

Geometric hashing is another common approach to identify possible protein-protein arrangements. It has been originally used as a computer visualization technique to match complementary substructures of one or several data sets (Norel et al., 1994). In protein-protein docking each protein surface is discretized as a set of triangles, which are stored in a hash table. By means of a hash key similar matching triangles on the surface of protein partners can be found quickly. During docking, these triangles comprise points on a molecular surface, having a certain geometrical (concave, convex) or physico-chemical (polar, hydrophobic) character. By matching triangles belonging to different molecules and being of complementary character, putative complex geometries can be generated.

A third class of methods uses either Brownian Dynamics (Schreiber et al., 2009; Gabdoulline & Wade, 2002), Monte Carlo, or multi-start docking minimization to generate large sets of putative protein-protein docking geometries (Zacharias, 2003; Fernandez-Recio et al., 2003; Gray et al., 2003). These methods have in principle the capacity to introduce conformational flexibility of binding partners already at the initial search step. Since these approaches are computational more expensive compared to FFT based correlation methods or geometric hashing a search is frequently limited to predefined regions of the binding partners (Bonvin, 2006). Alternatively, it is possible instead of atomistic models to employ coarse-grained (reduced) protein models to perform systematic docking searches. With such reduced protein models it is possible to optimize docking geometries starting from tens of thousands of protein start configurations (Zacharias, 2003; May & Zacharias, 2005). In order to limit the number of putative complex structures generated during an initial docking search cluster analysis is typically employed to reduce the number to a subset of representative complex geometries. Recently, the limitations of rigid docking strategies combined with a rescoring step have been systematically investigated by Pons et al. (2009). The authors applied a combination of rigid FFT-correlation based docking and re-scoring using the pyDock approach (Cheng et al., 2007). PyDock combines electrostatic Coulomb interactions with a surface-area-based solvation term (and an optional van der Waals term). The protocol showed very good performance for most proteins that undergo minor conformational changes upon complex formation (<1 Å Rmsd between unbound and bound structures) but unsatisfactory results for cases with significant binding induced conformational changes or applications that involved homology modelled proteins. A conclusion is that more specific scoring requires at the same time an improvement of the prediction accuracy of proposed binding modes in terms of deviation from the experimental binding interface. It also indicates the coupling between realistic scoring and accurate prediction of the complex structure.
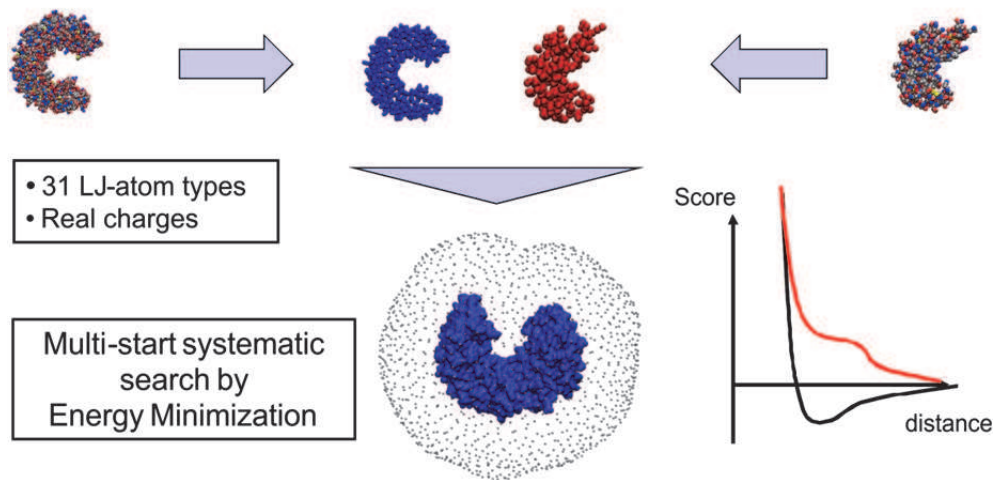
Fig. 2. Illustration of the ATTRACT docking methodology In the ATTRACT docking approach (Zacharias, 2003) atomic resolution partner structures are first translated (arrows) into a reduced (coarse-grained) representation based on pseudo atoms representing whole chemical groups. The smaller (ligand) protein is placed at various orientations on many starting placements around the receptor protein (in the middle of the Figure) followed by energy minimization to find an optimal docking geometry. In case of an attractive pseudo atom pair (black line in the plot on the right) an r^{-8}/r^{-6}-Lennard-Jones-type potential is used (r is the distance between atoms). For a repulsive pair (red curve) the energy minimum is replaced by a saddle point. The mathematical form of the scoring function is given in (Fiorucci & Zacharias, 2010b).

2.2 Flexible docking methods

A significant fraction of experimentally known protein-protein complexes belongs to the class that show only little conformational change upon complex formation. As indicated above in such cases it is possible to separate the initial rigid search from a subsequent flexible refinement and re-scoring step (see below). However, for many interesting docking cases with large associated conformational changes explicit consideration of conformational flexibility during the entire docking procedure or at an early refinement step appears to be necessary. Furthermore, in order to enhance the impact of docking in structural biology it is highly desirable to be able to use protein structures obtained by comparative (homology) modeling based on a known template structure with sufficient sequence similarity to the target protein. The accuracy of such comparative models depends on the correct alignment of target and template sequence. Even in cases of significant average target-template similarity the quality of the alignment is often not uniform along the whole protein sequence for example due to insertions or deletions in the aligned sequences which can result in structural inaccuracies. Overlap of such inaccurate structural segments with the protein region in contact with binding partners may interfere with the possibility to produce near-native complexes using rigid docking methods. This is also reflected in the fact that

docking cases that involve homology modelled protein partners belong to the most difficult cases in the CAPRI docking challenge (Lensink et al., 2007).

One possibility to directly use computationally rapid rigid docking algorithms is to indirectly account for receptor flexibility by representing the receptor target as an ensemble of structures. The structural ensemble can, for example, be a set of structures obtained experimentally (e.g. from nuclear magnetic resonance (NMR) spectroscopy) or can be formed by several structural models of a protein. It is also possible to generate ensembles from MD simulations (Grunberg et al., 2004) or from distance geometry calculations (de Groot et al., 1997). Docking to an ensemble increases the computational demand and due to the large number of protein conformations may also increase the number of false positive docking solutions. In the field of small-molecule docking a variety of ensemble based approaches have been developed in recent years (reviewed in Totrov & Abagyan, 2008). Cross docking to ensembles from MD simulations have also been used to implicitly account for conformational flexibility in protein docking (Krol et al., 2007). Mustard & Ritchie (2005) generated protein structures deformed along directions compatible with a set of distance constraints reflecting large scale sterically allowed deformations. Subsequently, the structures were used in rigid body docking searches to identify putative complex structures. Conformer selection and induced fit mechanism of protein-protein association have been compared by ensemble docking methods using the RosettaDock approach (Chaudhury & Gray, 2008). The RossettaDock approach includes the possibility of modelling both side chain as well as backbone changes for a set of starting geometries obtained from a low-resolution initial search (Wang et al., 2007). The method was able to successfully select binding-competent conformers out of the ensemble based on favourable interaction energy with the binding partner (Chaudhury & Gray, 2008). It was recently shown that the Rosetta approach can also be used to simultaneously fold and dock the structure of symmetric homo-oligomeric complexes starting from completely extended (unfolded) structures of the partner proteins (Das et al., 2009).

For a limited number of start configurations (in case of knowledge of the binding sites) it is possible to combine docking with molecular dynamics (MD) or Monte Carlo (MC) simulations. This allows, in principle, for full atomic flexibility or flexibility restricted to relevant parts of the proteins during docking. The HADDOCK program employs MD simulations including ambiguous restraints to drive the partner structures towards the approximately known interface (Dominguez et al., 2003). The success of HADDOCK in many Capri rounds for targets where some knowledge of the interface region was available underscores also the benefits of treating flexibility explicitly during early stages of the docking process. For protein-protein docking it is always helpful to include some knowledge on the putative interaction region. In these cases the docking problem can often be reduced to the refinement of a limited set of docked complexes close to the known binding site. Fortunately, for proteins of biological interest and with experimentally determined structure there is often also some biochemical (e.g. mutagenesis) data available on residues involved in binding to other proteins. Alternatively, bioinformatics techniques to predict putative protein interaction regions can often be used to limit or restraint the docking search to relevant protein surface parts. Several new techniques to locate putative binding sites based on physico-chemical properties or evolutionary conservation have been developed in recent years (e.g. de Vries & Bonvin, 2008).

Protein partner structures can undergo not only local adjustments (e.g. conformational adaptation of side chains and backbone relaxation at the interface) during association but also more global conformational changes that involve for example large loop movements or domain opening-closing motions. Proteins in solution are dynamic and the question to what extend the accessible conformational space in the unbound form overlaps with the bound conformation has been at the focus of several experimental and computational studies. Elastic Network Model (ENM) calculations are based on simple distance dependent springs between protein atoms and despite its simplicity are very successful to describe the mobility of proteins around a stable state (Bahar et al., 1997; Bahar et al., 2006). Systematic applications to a variety of proteins indicate that there is often significant overlap between observed conformational changes and a few soft normal modes obtained from an ENM of the unbound form (Keskin, 1998; Tobi & Bahar, 2005; Bakan & Bahar, 2009). ENM-based normal mode analysis has been used to identify hinge regions in proteins (Emekli et al., 2008) and can also be used to design conformational ensembles.
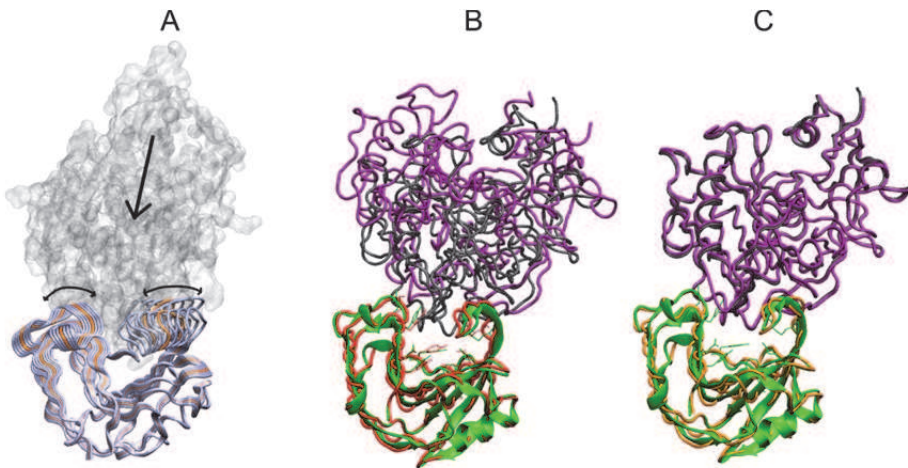

Fig. 3. Docking including minimization in soft flexible normal modes (A) Illustration of the flexible docking process of the taxi-inhibitor protein (pdb3HD8) to the xylanase target receptor protein (pdb1UKR) using the ATTRACT program (May & Zacharias, 2005). Putative translational motion of the inhibitor during docking approach is indicated by an arrow and the deformability of the xylanse by the superposition of several structures deformed in the softest normal mode (grey backbone tube representation). Best possible docking solutions (in pink) of the inhibitor relative to the bound (green cartoon) and unbound xylanase (red tube) are shown for rigid (B) and flexible (C) docking employing minimization along the 5 softest normal mode directions of the xylanase receptor protein. The placement of the inhibitor in the experimental structure is shown as grey tube. For flexible docking the root-mean-square deviation (Rmsd) from the inhibitor placement in the experimental structure was < 2 Å compared to > 8 Å in case of rigid docking.

It is also possible to use soft collective normal mode directions as additional variables during docking by energy minimization (Zacharias & Sklenar, 1999; May & Zacharias, 2005).

This allows the rapid relaxation of protein structures on a global scale involving much larger collective displacements of atoms during minimization then conventional energy minimization using Cartesian or other internal coordinates. The application of refinement in normal mode variables has been applied successfully in a number of studies (Lindahl & Delarue, 2005; May & Zacharias, 2005; Mashiach et al., 2010). Based on a coarse-grained protein model in the ATTRACT docking program (Zacharias, 2003) it has also been used in systematic docking searches to account approximately for global conformational changes already during the initial screen for putative binding geometries (May & Zacharias, 2008). In cases where protein partners undergo collective changes that overlap with the NM variables the approach can result in improved geometry and ranking of near-native docking solutions and can also lead to an enrichment of solutions close to the native complex structure (illustrated for an example case in Figure 3).

It should be emphasized that the inclusion of pre-calculated flexible degrees of freedom obtained from the unbound partners assumes that the collective directions of putative conformational change do not change upon binding to a partner protein. Although it has been shown that in many cases one can indeed describe a significant part of the observed conformational changes upon binding by a few collective degrees of freedom calculated for the unbound protein partners this does not need to be generally correct (Keskin, 1998; Tobi & Bahar, 2005; Bakan & Bahar, 2009). The binding partners may induce structural changes that are not possible for the isolated partner. In such cases pre-calculated flexible degrees of freedom cannot account for the true conformational change upon binding.

2.3 Prediction of putative binding regions prior to docking

If no experimental data on binding sites is available, binding site prediction methods can provide useful data for information driven docking. This type of information can be very helpful in order to limit the docking search or to evaluate and filter docking results. Docking approaches like HADDOCK (Dominguez et al., 2003) are based on applying restraints derived from experimentally known binding sites or predicted binding regions. Several different approaches exist to identify putative protein-protein binding sites. These methods focus on different characteristics of protein interaction sites like solvent accessibility (Chen & Zhou, 2005) or desolvation properties (Pons et al., 2009; Fiorucci & Zacharias, 2010a) and in many cases on combining different surface properties (Neuvirth et al., 2004; Liang et al., 2006). De Vries and Bonvin (2008) divided the properties of binding sites into three groups: 1. Properties of residues; 2. Evolutionary conservation; 3. Data obtained from atomic coordinates. The latter property includes, for example, secondary structure or solvent accessibility of residues or protein regions.

The data generated by predictors using one or more binding site features is presented either as a list of residues (Qin & Zhou, 2007) or as a patch on the proteins surface (Jones & Thornton, 1997a,b). Patch methods generate one or more patches of circular shape which can be found close to each other or distributed on the surface, sometimes additionally centre coordinates of these spots are given. In the other case residues from residue list predictors do not have to be nearby each other but are often clustered afterwards to receive a joined prediction at one or more spots on the proteins surface. Since proteins often have more than one binding site, prediction tools can indicate a correct binding site but maybe for the wrong binding partner.

Zhou & Qin (2007) and de Vries & Bonvin (2008) analysed existing predictors which are available as Web servers and evaluated the performance of these servers using 25 structures

from the CAPRI targets and several other datasets. The binding site predictions can be used to evaluate possible predicted docking geometries but also to generate artificial binding sites around the prediction to bias the docking run towards a desired region. On the other hand predictions can be used to discard complexes with a low overlap of predicted contacts after a systematic docking run. Examples of predicted binding regions compared to the known binding sites are illustrated for two cases in Figure 4.
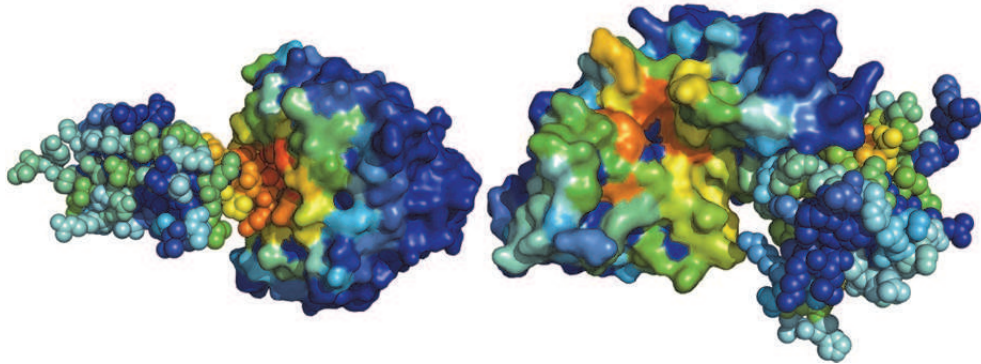

Fig. 4. Prediction of putative protein binding interfaces. Predictions were performed with the meta-PPISP server (Qin & Zhou, 2007) on the partner proteins of an enzyme inhibitor complex (pdb2SIC, left panel) and partners of a second complex (pdb1BUH, right panel). In each case one partner is represented as surface or collection of spheres, respectively. Protein partners are slightly displaced from the complexed state to indicate the native binding interface. Red indicates high predicted probability for a residue to be in the binding site and dark blue represents a low probability. Left example: The results match the real binding site. Right panel: The prediction for the smaller protein overlaps with the real binding site while for the larger protein residues quite far apart from the correct binding site are marked as putative binding site residues.

2.4 Flexible refinement and rescoring of docking solutions

As indicated in the two previous paragraphs protein-protein docking solutions obtained from an initial systematic docking run require typically a refinement and possibly also a rescoring step (Bonvin, 2006; Andrusier et al., 2008). This is not only necessary in case of rigid docking but also often if flexibility has been included approximately in the initial search by methods described in the previous paragraph (e.g. minimization in normal mode directions). The success of a multistep docking strategy requires that the set of initially docking structures contains solutions sufficiently close to the native structure in order to allow for further improvement during the refinement process. Hence, the initial scoring needs to recognize and preselect a binding mode sufficiently close the native placement and it has to simultaneously tolerate possible inaccuracies (atomic overlaps) at the interface. Before refinement the docking solutions are clustered to reduce the number of distinct docking geometries. Only one (the best scoring) solution from each cluster is typically used for further refinement and possible rescoring.

Refinement of a docked complex can be achieved by energy minimization based on a force field description of the proteins at atomic resolution. However, this results typically only in

small displacements of atoms to minimize overlap and to optimize locally a hydrogen bonding network. Frequently, molecular dynamics (MD) simulations are employed to achieve larger conformational adjustments compared to energy minimization during docking refinement. MD simulations are based on numerically solving Newton's equation of motion in small time steps (1-2 fs = 1-2 10^{-15} s) based on a molecular mechanics force field description of the protein-protein complex (Karplus & McCammon, 2002). Due to the kinetic energy of every atom of the proteins, it is in many cases possible to overcome energy barriers and to move the structure significantly farer away from the initial docking geometry. Depending on the simulation temperature and length displacements up to several Angstroms from the initial atom positions are possible. However, if the displacements during MD simulations indeed move the proteins towards a more realistic complex structure depends on the accuracy of the force field and on a realistic representation of the aqueous solution. Refinement simulations on a given protein-protein complex should, ideally, include surrounding aqueous solvent and ions. This, however, increases the computational demand for such refinement simulations. In addition, the equilibration of explicit solvent molecules around a solute molecule requires significant simulation times (currently limited to tens or in some cases hundreds of nanoseconds). Nevertheless, during the final stages of some protein-protein docking protocols explicit water molecules can be added to the simulation system (van Dijk & Bonvin, 2006). Explicit solvent MD simulations can also be used to investigate the flexibility of protein structures prior to docking (Rajamani et al., 2004; Camacho, 2005). It is for example possible to identify the alternative or most likely side chain conformations. Using principal component analysis of the motions extracted from MD simulations it is also possible to analyse the global conformational flexibility of binding partners prior to docking (Amadei et al., 1993; Smith et al., 2005).

The possibility to implicitly account for solvent effects can be used to accelerate the refinement process. A variety of implicit solvation models has been developed (reviewed in Bashford & Case, 2000; Baker, 2005; Chen et al., 2008). Only a brief description of the most relevant concepts for protein-protein docking and scoring will be given. A macroscopic solvation concept describes the protein interior as a medium with a low dielectric permittivity embedded in a high dielectric continuum representing the aqueous solution (Baker, 2005). The effect of the solvent is then calculated as a reaction field from a solution of Poisson's equation for the charges assigned to each atom of the molecule. The mean effect of a salt atmosphere can be included by solving the Poisson-Boltzmann equation. The most common method to solve the Poisson-Boltzmann equation is the finite-difference method on a grid representation of the protein system. However, the method cannot easily be combined with MD refinement due to the difficulty to extract accurate solvation forces from grid solutions of the Poisson-Boltzmann equation (Gilson et al., 1993).

It is possible to use more approximate methods like the Generalized Born (GB) method (Still et al., 1990; Hawkins et al., 1995; Bashford & Case, 2000). In the GB approach an effective solvation radius is assigned to each atom. This effective radius can be thought of as an average distance of the selected atom from the solvent or from the solvent accessible surface of the molecule. With the effective Born radii calculated for each atom the electrostatic solvation and its derivative (solvation forces) can be calculated very rapidly (Schaefer & Karplus, 1996; Onufriev et al., 2002). The GB method and related implicit solvent approaches are frequently used during refinement of docked protein-protein complexes.

Once a set of docked and structurally refined complexes has been obtained a rescoring step can be used to finally select the most realistic predicted complex. An ideal scoring function

should recognize favourable native contacts as found in the bound complex and discriminate those from non-native contacts with lower scores. Scoring can be based on a physical force field with optimized weights on the energetic contributions (Dominguez et al., 2003; Bonvin, 2006; Audie, 2009) or can involve knowledge-based statistical potentials derived from known protein protein complex structures (Gottschalk et al., 2004; Zhang et al., 2005; Huang & Zou, 2008). Often a single descriptor (e.g. surface complementarity) or a single binding energy component (e.g. van der Waals or electrostatic energy) is non-optimal to distinguish non-native from near-native solutions. A combination of different surface and interface descriptors has been shown to better enrich near-native solutions in the pool of best scoring docking solutions (Murphy et al., 2003; Duan et al., 2005; Liu et al., 2006; Martin & Schomburg, 2008; Pierce & Weng, 2008; Audie, 2009;Liang et al., 2009).

The experimentally determined protein-protein complex structures allow the extraction of data on the statistics of residue-residue and atom-atom contact preferences at interfaces. Based on these statistics it is possible to design knowledge-based scoring functions which in general compare the frequency of contact pairs in known interfaces with the expected frequency if residues or atoms would randomly distributed at interfaces. Effective knowledge-based potentials have been developed that are based on contact preferences of amino acids at known interfaces compared to interfaces of non-native decoy complexes (Huang & Zou, 2008; Ravikant & Elber, 2009; Kowalsman & Eisenstein, 2009). The resulting contact or distance dependent pair-potentials can improve the scoring of near-native complexes. The distribution of amino acids in the core region of protein-protein interfaces differs on average from the whole interface and the rim region which is partially exposed to water even in the presence of the binding partner. This observation has also been explored to improve the recognition of near-native binding geometries and has been demonstrated on several test cases (Kowalsman & Eisenstein, 2009).

3. Conclusion

The rational modifications of protein surfaces are increasingly being used to design new protein-protein binding interfaces. Another ultimate aim of protein-protein docking approaches is the application on a systematic proteomic scale. Methods of protein-protein docking and interface refinement could help to predict possible protein interaction geometries and guide such protein interaction design. The realistic prediction of binding geometries of protein-protein complexes is highly desirable to provide structural models for the many important protein-protein interactions in a cell. Progress in both the efficiency and in the development of new docking algorithms has been achieved in recent years. Still a major challenge is the appropriate inclusion of possible conformational changes during the docking searches. This is of great importance since for the many protein interaction cases only homology modelled structures of the partners are available. Employing an appropriate ensemble of protein conformations or, alternatively, the efficient explicit consideration of conformational changes during docking are possible routes of progress. For many protein-protein interactions experimental data (e.g. low resolution structural or biochemical data) is available that restricts the range of possible complex structures. Here, restraint driven docking techniques that include flexibility of the binding partners at early refinement stages are promising. In recent years it has become clear that many protein-protein interactions involve coupled folding of disordered parts of proteins upon association. The possibility of

structure prediction and modelling of such interactions is at a very early stage. Progress in this area will require many new algorithms and method developments.

4. Acknowledgment

5. References

Amadei, A., Linssen, A. B. M. & Berendsen H. J. C. (1993). Essential dynamics of proteins, *Proteins* 17, 412-425.

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

Audie, J. (2009). Development and validation of an empirical free energy function for calculating protein-protein binding free energy surfaces, *Biophysical Chemistry* 139, 84-91.

Bahar, I., Atilgan, A.R., Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, *Folding Design* 2, 173-181.

Bahar, I., Chennubhotla, C. & Tobi D. (2006). Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation, *Current Opinion Structural Biology* 16, 194-200.

Bakan, A. & Bahar, I. (2009). The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding, *Proceeding of the National Academy of Sciences USA* 106, 14349-14354.

Bashford, D. & Case, D. A. (2000). Generalized born models of macromolecular solvation effects, *Annual Review of Physical Chemistry* 51 , 129-152.

Baker, N. A. (2005). Improving implicit solvent simulations: a Poisson-centric view, *Current Opinion Structural Biology* 15, 137-143.

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

Camacho, C. J. (2005). Modeling side-chains using molecular dynamics improve recognition of binding region in CAPRI targets, *Proteins: Structure, Function, and Bioinformatics* 60, 245-251.

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

Chen, J., Brooks, C. L. 3rd & Khandogin J. (2008). Recent advances in implicit solvent-based methods for biomolecular simulations, *Current Opinion Structural Biology* 18 , 140-148.

Chen, H. & Zhou, H.-X. (2005). Prediction of interface residues in protein–protein complexes by a consensus neural network method: Test against NMR data, *Proteins: Structure, Function, and Bioinformatics* 61, 21–35.

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

Das, R., Andre, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., Arrowsmith, C.H., Szyperski, T., Baker, D. (2009). Simultaneous prediction of protein folding and docking at high resolution, *Proceeding of the National Academy of Sciences USA* 106, 18978-18983.

de Groot, B. L., van Aalten, D. M. F., Scheek, R. M., Amadei, A., Vriend, G., and Berendsen, H. J. C. (1997). Prediction of protein conformational freedom from distance constraints. *Proteins: Structure, Function, and Bioinformatics* 29, 240-251.

De Vries, S. J. & Bonvin, A. M. J. J. (2008). How proteins get in touch: interface prediction in the study of biomolecular complexes,. *Current Protein Peptide Science* 9, 394-406.

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

Duan, Y., Reddy, B. V. B. & Kaznessis, Y. N. (2005). Physicochemical and residue conservation calculations to improve the ranking of protein-protein docking solutions, *Protein Science* 14, 316-328.

Emekli, U., Schneidman-Duhovny, D., Wolfson, H.J., Nussionov, R., Haliloglu, T. (2008). HingeProt: automated prediction of hinges in protein structures, *Proteins: Structure, Function, and Bioinformatics* 70, 1219-1227.

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

Fiorucci, S. & Zacharias, M. (2010a). Prediction of Protein-Protein Interaction Sites Using Electrostatic Desolvation Profiles, *Biophysical Journal* 98, 1921-1930.

Fiorucci, S. & Zacharias, M. (2010b). Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins*. 78, 3131-3139.

Fischer, H. (1894). Einfluß der Configuration auf die Wirkung der Enzyme, *Chemische Berichte* 27, 2985-2993.

Gabdoulline, R. R. & Wade, R. C. (2002). Biomolecular diffusional association, *Current Opinion Structural Biology* 12, 204-213.

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.

Gilson, M. K., Davis, M. E., Luty, B. A. & McCammon, J. A. (1993). Computation of Electrostatic Forces on Solvated Molecules Using the Poisson-Boltzmann Equation, *Journal of Physical Chemistry* 97, 3591-3600.

Gottschalk, K. E., Neuvirth, H. & Schreiber G. (2004). A novel method for scoring of docked protein complexes using predicted protein-protein interaction sites, *Protein Engneering* 17 183-189.

Gray, J. J., Moughon, S. E., Kortemme, T., Schueler-Furman, O., Misura, K. M. S., Morozov, A. V. & Baker, D. (2003). Protein-protein docking predictions or the CAPRI experiments, *Proteins: Structure, Function, and Bioinformatics* 52, 118-122.

Grunberg R., Leckner, J. & Nilges, M. (2004). Complementarity of structure ensembles in protein-protein docking, *Structure* 12, 2125-2136.

Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. (1995). Pairwise solute descreening of solute charges from a dielectric medium, *Chemical Physics Letters* 246, 122-129.

Huang, S-Y., Zou, X. (2008). An iterative knowledge-based scoring function for protein-protein recognition, *Proteins* 71, 557-579.

Janin, J., Henrick, K., Moult, J., Eyck LT, Sternberg, M. J., Vajda, S., Vasker, I., Wodak, S. J. (2003). CAPRI: a critical assessment of predicted interactions, *Proteins : Structure, Function, and Bioinformatics* 52, 2-9.

Jones, S. & Thornton, J. M. (1997a). Analysis of protein-protein interaction sites using surface patches, *Journal of Molecular Biology* 272, 121-132.

Jones, S. & Thornton, J. M. (1997b). Prediction of protein-protein interaction sites using patch analysis, *Journal of Molecular Biology* 272, 133-143.

Karplus, M. & McCammon, J. A. (2002). Molecular dynamics of biomolecules. *Nature Structural Biolog,* 9 646-652.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C. & Vakser I. A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques, *Proceeding of the National Academy of Sciences USA* 89, 2195-2199.

Keskin, O. (1998). Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: a case study of antibodies. *BMC Structural Biology* 7, 31-42.

Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis, *Proceeding of the National Academy of Sciences USA* 17, 1145-1150.

Kowalsman, N. & Eisenstein M. (2009). Combining interface core and whole interface descriptors in postscan processing of protein-protein docking models, *Proteins: Structure, Function, and Bioinformatics* 77, 297-318.

Krol, M., Chaleil, R. A., Tournier, A. L. & Bates P. A. (2007). Implicit flexibility in protein docking : cross docking and local refinement, *Proteins: Structure, Function, and Bioinformatics* 69, 7-18.

Lensink, M. F., Mendez, R., Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd Edition, *Proteins Structure, Function, and Bioinformatics*, 69, 704-718.

Lindahl, E. & Delarue, M. (2005). Refinement of docked protein-ligand and protein-DNA structures using low-frequency normal mode amplitude optimization, *Nucleic Acids Research* 33, 4496-4506.

Liang, S., Zhang, C., Liu, S. & Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function, *Nucleic Acid Research* 34, 3698-3707.

Liang, S., Meroueh , S. O., Wang, G., Qiu, C., Zhou, Y. (2009). Consensus scoring for enriching near-native structures from protein-protein docking decoys, *Proteins: Structure, Function, and Bioinformatics* 75, 397-403.

Liu, S., Li, Q. & Lai, L. (2006). A combinatorial score to distinguish biological and non-biological protein-protein interfaces, *Proteins: Structure, Function, and Bioinformatics* 64, 68-78.

Martin, O., Schomburg, D. (2008). Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. *Proteins: Structure, Function, and Bioinformatics* 70, 1367-1378.

Mashiach, E., Nussinov, R. & Wolfson. H. J. (2010). FiberDock: Flexible induced-fit backbone refinement in molecular docking, *Proteins: Structure, Function, and Bioinformatics* 78, 1503-1519.

May, A. & Zacharias, M. (2005). Accounting for global protein deformability during protein–protein and protein–ligand docking, *Biochimica et Biophysica Acta* 1754, 225-231.

May, A. & Zacharias, M. (2008). Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking, *Proteins* 70, 794-809.

Murphy, J., Gatchell, D. W., Prasad, J. C. & Vajda S. (2003). Combination of scoring functions improves discrimination in protein-protein docking, *Proteins* 53, 840-854.

Mustard, D. & Ritichie, D. W. (2005). Docking essential dynamics eigenstructures, *Proteins: Structure, Function, and Bioinformatics* 60, 269-274.

Neuvirth, H., Raz, R. Schreiber, G. (2004). ProMate: A Structure Based Prediction Program to Identify the Location of Protein–Protein Binding Sites, *Journal of Molecular Biology* 338, 181-199.

Norel, R., Fischer, D., Wolfson, H. J. & Nussinov R. (1994). Molecular surface recognition by a computer vision-based technique, *Protein Engneering*. 7, 39-46.

Onufriev, A., Case, D. A. & Bashford, D. (2002). Effective Born radii in the generalized Born approximation: the importance of being perfect, *Journal of Computational Chemistry* 23, 1297-1304.

Pierce, B., Weng, Z. (2008). A combination of rescoring and refinement significantly improves protein docking performance, *Proteins: Structure, Function, and Bioinformatics* 72, 270-279.

Pons, C., Grosdidier, S., Solernou, A., Perez-Cano, L. & Fernandez-Recio, J. (2009) Present and future challenges and limitations in protein-protein docking, *Proteins: Structure, Function, and Bioinformatics*, 78, 95-108.

Qin, S. & Zhou, H.-X. (2007). meta-PPISP: a meta web server for protein-protein interaction site prediction, *Bioinformatics 23, 3386-3387*.

Rajamani, D., Thiel, S., Vajda, S. & Camacho, C. J. (2004). Anchor residues in protein- protein interactions, *Proceeding of the National Academy of Sciences USA* 101, 11287-11292.

Ravikant, D. V. S, Elber, R, (2009). PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 75, 133-145.

Ritchie, D., Kozakov, D., Vadja, S. (2008). Accelerating and focusing protein-protein docking correlations using Multi-dimensional rotational FFT generating functions. *Bioinformatics* 24, 1865-1873.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network, *Nature*, 437, 1173-1178.

Schaefer, M. & Karplus, M. (1996). A comprehensive analytical treatment of continuum electrostatics, *Journal of Physical Chemistry* 100, 1578-1599.

Schreiber, G., Haran, G., Zhou, H.-X (2009) Fundamental Aspects of protein-protein association kinetics, *Chemical Reviews* 1009, 839-860.

Smith, G. R., Sternberg, M. J. E. & Bates P. A. (2005). The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *Journal of Molecular Biology* 347, 1077-1101.

Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrikson, T. (1990). Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics, *Journal of the American Chemical Society* 112, 6127-6129.

Totrov M. & Abagyan, R. (2008). Flexible ligand docking to multiple receptor conformations: a practical alternative, *Current Opinion Structural Biology* 18:178-184.

Tsai, C. J., Kumar, S., Ma, B. (1999). Nussinov R: Folding funnels, binding funnels, and protein function, *Protein Science* 8, 1181-1190.

Tobi, D., & Bahar, I. (2005). Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state, *Proceeding of the National Academy of Sciences USA* 2005, 102:18908-18913.

van Dijk, A. D. J. & Bonvin, A. M. J. J. (2006). Solvated docking: introducing water into the modelling of biomolecular complexes. *Bioinformatics*. 22, 2340-2347.

Vajda, S. & Kozakov D, (2009) Convergence and combination of methods in protein-protein docking. *Current Opinion Structural Biology* 19:164-170.

Wang, C., Bradley, P. & Baker, D. (2007). Protein-protein docking with backbone flexibility, *Journal of Molecular Biology* 373, 503-519.

Zacharias, M. & Sklenar, H. (1999). Harmonic modes as variables to approximately account for receptor flexibility in ligand-receptor docking simulations: Application to a DNA minor groove ligand complex, *Journal of Comput Chemistry* 20, 287-300.

Zacharias, M. (2003). Protein-protein docking using a reduced model, *Protein Science* 12, 1271-1282.

Zacharias, M. (2010). Accounting for conformational changes during protein-protein docking,. *Current Opinion Structural Biology* 16, 194-200.

Zhang, C., Liu, S. & Zhou, Y. (2005). Docking prediction using biological information, ZDOCK sampling technique, and clustering guided by the DFIRE statistical energy function,. *Proteins: Structure, Function, and Bioinformatics* 60, 314-318.

Zhang, Q., Sanner, M. & Olson A. J. (2009). Shape complementarity of protein-protein complexes at multiple resolutions, *Proteins: Structure, Function, and Bioinformatics* 75, 453-467.

Zhou, H.-X. & Qin, S. (2007). Interaction-site prediction for protein complexes: a critical assessment, *Bioinformatics* 23, 2203-2209.
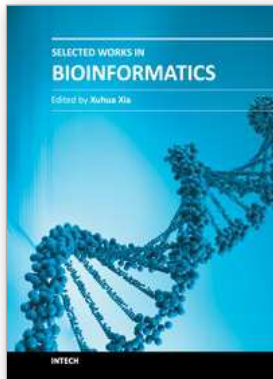
Selected Works in Bioinformatics

Edited by Dr. Xuhua Xia

This book consists of nine chapters covering a variety of bioinformatics subjects, ranging from database resources for protein allergens, unravelling genetic determinants of complex disorders, characterization and prediction of regulatory motifs, computational methods for identifying the best classifiers and key disease genes in large-scale transcriptomic and proteomic experiments, functional characterization of inherently unfolded proteins/regions, protein interaction networks and flexible protein-protein docking. The computational algorithms are in general presented in a way that is accessible to advanced undergraduate students, graduate students and researchers in molecular biology and genetics. The book should also serve as stepping stones for mathematicians, biostatisticians, and computational scientists to cross their academic boundaries into the dynamic and ever-expanding field of bioinformatics.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Sebastian Schneider and Martin Zacharias (2011). Flexible Protein-Protein Docking, Selected Works in Bioinformatics, Dr. Xuhua Xia (Ed.), ISBN: 978-953-307-281-4, InTech, Available from: http://www.intechopen.com/books/selected-works-in-bioinformatics/flexible-protein-protein-docking

INTECH
open science | open minds