On the development of computational tools for the study of protein-protein interactions and protein-protein binding

Manuel Alejandro Marín López

DOCTORAL THESIS UPF / 2016

THESIS DIRECTOR:

Dr. Baldomero Oliva Miguel

Structural Bioinformatics Lab (SBI)

Research Program on Biomedical Informatics (GRIB)

Department of Experimental and Health Sciences (CEXS)

upf.
Universitat
Pompeu Fabra
Barcelona

A Isabel

Acknowledgements

En primer lugar me gustaría agradecer a toda mi familia, especialmente a mis padres Concepción y Manuel y a mi hermano Luís por su cariño y apoyo en todo momento. A mis abuelos, sé que les hubiese llenado de orgullo leer estas frases, especialmente a mi abuelo Ramón. También quiero agradecer a mis suegros Andrés e Isabel, y a la abuelita Marisol, por sus ánimos y sesiones de "coaching". ¡¡Gracias a todos!!

Esta tesis no habría sido posible sin mi mujer Isabel, que ha acabado por sufrirla tanto como yo. Gracias por estar siempre a mi lado y ayudarme a seguir adelante. ¡¡Eres lo mejor!! También, aunque es muy pequeño aún (de hecho se está gestando), quiero agradecer a mi hijo Andrés por ser una fuente de ilusión y motivación.

This journey would not have been the same without my historical SBI lab mates and now my friends: Jaume, Oriol, Javi, Joan, Daniel, David, Dani, Emre and Bernat. For sure I will remember with nostalgia the after lunch coffees at Gamar. I enjoyed a lot working with you. I would like to thanks also all the visitors: Narcís, Jasha, Attila, Billur… and all the GRIB members, particularly Xavi and I.T. staff.

Finally, I would like to express my gratitude to my PhD supervisor Prof. Baldo Oliva for his support and guidance during these years. He makes SBI lab a nice place to work. Thanks!

This thesis has been supported by the FI-DGR 2012 grant from AGAUR (Agència de Gestió d'Ajuts Universitaris i de Recerca).

Structural BioInformatics

GRIB RESEARCH PROGRAMME ON BIOMEDICAL INFORMATICS

upf. Universitat Pompeu Fabra Barcelona

Abstract

Proteins are involved in almost all cell processes, with physical interaction between them being key to their function and dictated by its 3D structure. Hence, the study of protein-protein interactions and protein-protein binding is crucial to fully understand biological systems. In this thesis, we present V-D^{2}OCK, a fast and accurate data-driven docking tool for high throughput prediction of the structure of protein complexes. We have also studied the conformational space of potential encounter complexes by means of non-specific decoys obtained by docking in order to develop BADock, an accurate binding affinity predictor from the unbound individual structures. Finally, we have published online an integrated and centralized resource (InteractoMIX) that allows to the research community an easy access to a compendium of bioinformatic web applications to study protein-protein interactions.

Resum

Les proteïnes estan implicades en gairebé tots els processos cel·lulars, amb la interacció física entre elles clau per la seva funció i dictada per la seva estructura 3D. Per tant, l'estudi de la unió i les interaccions proteïna-proteïna és crucial per entendre completament els sistemes biològics. En aquesta tesi, es presenta V-D^{2}OCK, una eina de "docking" dirigit ràpida i precisa per predir l'estructura de complexes de proteïnes a gran escala. També hem estudiat l'espai conformacional de possibles complexes transitoris per mitjà de resultats de "docking" no específics per tal de desenvolupar BADock, un predictor d'energia d'unió a partir de les estructures individuals per separat, Finalment, hem publicat online un recurs integrat i centralitzat que permet a la comunitat investigadora l'accés fàcil a un conjunt de aplicacions web de bioinformàtica per l'estudi de interaccions proteïna-proteïna.

Table of Contents

List of Publications

Publications are listed in reverse chronological order. Articles 1, 3 and 4 conform the main body of this thesis (results section). The contribution to each article is indicated before its presentation. The remaining articles are exposed in appendix section.

Articles

(*) Indicates that the authors have contributed equally to a given paper.

1. Marín-López MA, Planas-Iglesias J, Bonet J, Garcia-Garcia J, Fernandez-Fuentes N, Oliva B. **On the mechanisms of protein-protein binding: predicting their affinity from the unbound tertiary structures.** *Manuscript in preparation.*
2. Sieberts SK*, Zhu F*, Garcia-Garcia J*, […], Marín-López MA, […], Mangravite LM. (2016) **Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis.** *Nature Communications*, 7, 12460.
3. Poglayen D*, Marín-López MA*, Bonet J*, Fornes O, Garcia-Garcia J, Planas-Iglesias J, Segura J, Oliva B, Fernandez-Fuentes N. (2016) **InteractoMIX: a suite of computational tools to exploit interactomes in biological and clinical research.** *Biochemical Society Transactions,* 44(3), 917-924.
4. Segura J, Marín-López MA, Jones PF, Oliva B, Fernandez-Fuentes N. (2015) **VORFFIP-Driven Dock: V-D^{2}OCK, a Fast and Accurate Protein Docking Strategy.** *PLoS ONE*, 10(3): e0118107.
5. Bonet J, Planas-Iglesias J, Garcia-Garcia J, Marín-López MA, Fernandez-Fuentes N, Oliva B. (2014) **ArchDB 2014: structural classification of loops in proteins.** *Nucleic Acids Research*, 42(1), D315–9.
6. Planas-Iglesias J*, Marín-López MA*, Bonet J*, Garcia-Garcia J, Oliva B. (2013) **iLoops: a protein-protein interaction prediction server based on structural features.** Bioinformatics, 29(18), 2360-2362.

7. Planas-Iglesias J, Bonet J, Garcia-Garcia J, Marín-López MA, Feliu E, Oliva B. (2013) **Understanding Protein-Protein Interactions Using Local Structural Features.** *Journal of Molecular Biology*, 425(7), 1210-1224.

8. Planas-Iglesias J, Bonet J, Marín-López MA, Feliu E, Gursoy A, Oliva B. (2012) **Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence.** *InTech, In Protein-Protein Interactions - Computational and Experimental Tools*, Dr. Weibo Cai, (Ed.), ISBN: 978-953-51-0397-4.

Posters and oral presentations

(*) Indicates the presenter(s)

1. Marín-López MA*, Planas-Iglesias J, Bonet J, Oliva B. **Understanding protein recognition using structural features.** *Poster, 29th annual symposium of the protein society*, Barcelona, Spain, 2015.

2. Garcia-Garcia J*, Aguilar D, Poglayen D, Bonet J, Fornes O, Guney E, Planas-Iglesias J, Marín-López MA, Anton B, Oliva B. **Predicting response to arthritis treatments: regression-based Gaussian processes on small sets of SNPs**. *Poster, RECOMB/ISCB2014*, San Diego, USA, 2014.

3. Marín-López MA*, Planas-Iglesias J, Bonet J, Oliva B. **Understanding protein recognition using structural features.** *Poster, II Jornada de Bioinformàtica I Biologia Computacional (JdB2014)*, Barcelona, Spain, 2014.

4. Marín-López MA*, Planas-Iglesias J, Bonet J, Oliva B. **Understanding protein recognition using structural features.** *Poster, XII Spanish Symposium on Bioinformatics (JBI2014)*, Sevilla, Spain, 2014.

5. Poglayen D, Garcia A, Casadio J, Fornes O, Garcia-Garcia J, Zinman G, Marín-López MA*, Bar-Joseph Z, Hirt H, Klein-Seetharaman J, Oliva B. **High-throughput information integrated with in-silico predictions to identify key participants in host-pathogen interactions**. *Poster, XII Spanish Symposium on Bioinformatics (JBI2014)*, Sevilla, Spain, 2014.

6. Planas-Iglesias J*, Bonet J*, Garcia-Garcia J*, Marín-López MA*, Feliu E, Oliva B. **Understanding protein-protein interactions using local structural features**. *Poster, 12th CRG Symposium BCN2: Biological Control Networks in Barcelona*, Barcelona, Spain, 2013.

7. Marín-López MA*, Planas-Iglesias J, Oliva B. **Understanding protein recognition using structural features.** *Oral Presentation, IX Workshop on Genomics and Proteomics*, Barcelona, Spain, 2013.

8. Marín-López MA*, Bonet J, Planas-Iglesias J, Oliva B. **iLoops Server: A Protein-Protein Interaction Prediction Utility Based On Local Structural Features.** *Poster, XI Spanish Symposium on Bioinformatics (JBI2012)*, Barcelona, Spain, 2012.

9. Planas-Iglesias J*, Bonet J, Marín-López MA, Feliu E, Oliva B. **To Bind or not To Bind: Predicting protein-protein interactions from favouring and disfavouring local structural features.** *Oral presentation, XI Spanish Symposium on Bioinformatics (JBI2012)*, Barcelona, Spain, 2012.

10. Planas-Iglesias J*, Bonet J, Marín-López MA, Feliu E, Oliva B. **To Bind or not To Bind: Predicting protein-protein interactions from favouring and disfavouring local structural features.** *Oral presentation, 7th Annual RECOMB Systems Biology*, Barcelona, Spain, 2011.

1. INTRODUCTION

The central dogma of molecular biology established decades ago by Francis Crick (1) states that the inheritable information (genes) encoded in deoxyribonucleic acid (DNA) is transcribed to ribonucleic acid (RNA) and then translated to protein. Even thought many exceptions to this dogma are now known (2), such as non-coding RNAs, proteins are still considered the main effectors of biological activities. They are involved in virtually all cell processes in a living organism, from the catalysis of metabolic reactions to the DNA replication, including structural support, cell communication, signal transduction and gene regulation, which control the central dogma itself (3).

In this introduction we will first review protein structure and the principle experimental and computational methods for its determination and prediction. Then, we will explain protein-protein interactions (PPI) from different level of detail (taking a special attention to structure) and the determination and prediction methods, focusing on protein docking. Finally, we will deepen into PPI by discussing about protein association and recognition mechanisms.

1.1 Protein structure

It is well known that protein function is determined by its three-dimensional (3D) structure. Thus, determining the 3D structure of proteins is important to fully understand how a cell or an organism works (4). Protein structure is usually described hierarchically in four levels of complexity:

1) The primary structure refers to the sequence of amino acids that are covalently linked through peptide bonds into a chain. There are 20 different amino acids in the standard genetic code and when linked are referred as residues (5).

2) The secondary structure of a protein is defined by regular local conformations established by hydrogen bonding between the amide groups of the backbone chain. The main two types of secondary structure are the alpha helices and beta strands (6,7). The flexible non-regular regions between secondary structures are known as loops. The combination of few secondary structures with specific geometric arrangements frequently found in structures defines what is called structural motifs. Loops have been described to play an important role in folding, dynamics and function (8).

3) The tertiary structure is the global three-dimensional structure that a protein adopts in the space. It defines the relationships between the residues located far away in the primary structure that held together by non-covalent interactions (hydrogen bonds, van der Waals, ionic interactions and hydrophobic packing), disulphide bonds and metal ion coordination. A tertiary structure can be described by domains (9), which are independent stable 3D units that have some degree of functionality and can be combined to build more complex proteins (10).

It has to be noted that proteins are not static and their flexibility results in constant small conformational changes. In fact, some proteins do not actually reach neither secondary nor tertiary structure. These proteins, known as intrinsically unstructured proteins (IUPs), have the residues oriented randomly and in constant movement (11).

4) The quaternary structure, also known as protein complex, is the three-dimensional conformation that adopts multiple proteins when interacting. Some proteins can results in significant conformational changes in the tertiary structure upon interaction (12). Quaternary structure will be addressed in more detailed in section 1.2.

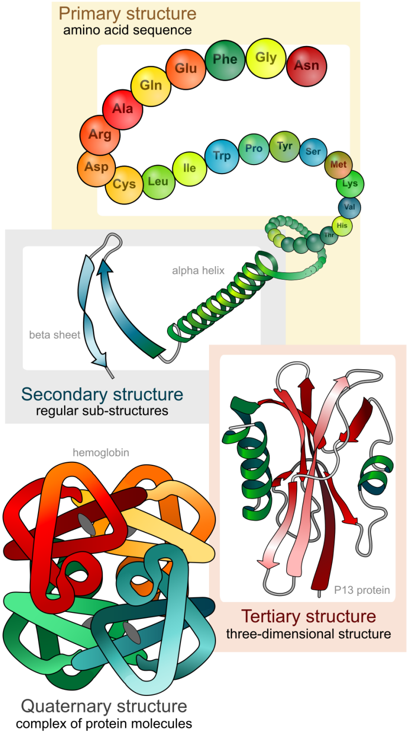


Figure 1.1 The four levels of protein structure. From primary structure (top) to quaternary structure (down). Image obtained from https://en.wikipedia.org/wiki/Protein_structure.

1.1.1 Protein structure determination

There are many experimental methods to determine the 3D structure of proteins. The two most successful and most used methods are the X-ray diffraction (13) and NMR spectroscopy (14), which cover 99% of structures deposited in the Protein Data Bank (PDB). The PDB is a repository created in the 1970s aimed to collect all protein structures in a unified format (15). From it, multiple database with different focus have been developed (e.g. SCOP (16), CATH (17), 3DID (18))

a) The X-rays diffraction is based on the scattering pattern that x-rays produces when interacting with the electrons. Then, the protein being studied can be fitted into the resulting electron density map. In order to obtain enough signal, multiple proteins are required to be arranged in a lattice forming a regular geometric pattern: a crystal.

   To obtain a protein crystal, high concentrations of proteins and the appropriate experimental conditions are required. Thus, the crystallization becomes a serious limitation that difficult this technic to be used high throughput (19). In addition to this, the flexible regions of proteins such as loops may adopt different conformations in each cell in the lattice, causing an irregular dispersion of the x-rays resulting in a blurred area, leaving gaps in the 3D structure (20).

b) The RNM spectroscopy is a technique based on the magnetic moment (nuclear spin) that some atomic nuclei (including ^{1}H, ^{13}C, ^{15}N) possess, which give rise to different resonance frequencies in a magnetic field. The magnetic field is induced through short pulses of electromagnetic energy, which enhance the raise of the energetic level. Excited nuclei return to the basal state resonating in a frequency characteristic of the atom environment (chemical shift). Pulses of different radio frequencies provide different information about the environment and allow the reconstruction of the 3D structure of the protein (21).

NMR is applied to molecules in solution (natural environment), solving the problem of protein flexibility and giving a more dynamic view of the protein. In addition, it is less costly than X-rays diffraction since it is not necessary to make a crystal. On the other hand, the complexity of the obtained result spectra makes this technique difficult to apply to large proteins.

c) There are also other methods to determine 3D structures of proteins. These includes cryo-electron tomography (22), X-ray scattering (SAXS) (23) and solid-state NMR (24). Cryo-electron tomography allows the determination of the shape of large macromolecular complexes, but not the atomic details (25). SAXS is applied to large proteins in solution at cost of resolution (26) and solid-state NMR is useful to study membrane proteins at high resolution (27).

1.1.2 Protein structure prediction

The introduction of high throughput sequencing technologies resulted in the determination of a vast amount of new protein sequences (28). However, the determination methods of protein structure, as previously described, are a slow and costly. Thus, there is an important difference between known proteins and proteins with known structure. This difference is defined as the sequence-structure gap. Computational methods can help to bridge this gap by inferring protein structure from sequence (29). These methods can be classified in three different groups:

1. Homology modelling: Homology or comparative modelling is the preferred theoretical method to infer the tertiary structure of a protein from its sequence. This method is based on the principle that structure is more conserved than sequence (30). Hence, proteins with enough similar sequences (homologs) will acquire same conformation in the space. Therefore, homologs with known structure (templates) can be used to model a protein with unknown structure (query).

   Homology modelling is carried out in three steps: template identification, template alignment and model building. Template identification is the process of finding the closest homolog in a database of protein structures using applications such as BLAST (31) or HMMR (32). Once the template has been selected, depending on certain requirement the sequence has to be re-aligned using tools such as CLUSTALW (33) or T-COFFE (34). The last step creates the 3D structure. At this point, the structural information of the template is transferred to the query protein guided by the alignment (fits the target sequence upon the template structure) using applications such as MODELLER (35). Finally, the prediction obtained is refined and evaluated to asses de quality of the model. PROSA (36) is one of the most frequently used applications to evaluate the resultant models.

2. Threading: Threading or fold recognition is particularly useful when no appropriate template is found for homology modelling. Threading is based on the principle that the number of protein folds is limited in nature and the query protein must belong to one of them (37). This process is performed by aligning the query protein against a database of protein folds (non-redundant sequences) and using a scoring function (based on structural features) to find the best possible template. Then, the atoms of the query are arranged around the backbone of the

template. Examples of threading tools are GenTHREADER (38) or Phyre (39).

3. *Ab Initio* methods: Similarly to threading, when no homologs are found *ab initio* methods tries to elucidate the 3D structure of a protein from scratch. Those techniques explore the conformation space creating multiple candidates that are ranked using a scoring function based on statistical potentials or physic. One way to explore the conformational space is to reproduce the folding path through molecular dynamics (e.g. I-TASSER (40)). However, it requires a lot of computational efforts and is usually only applied to small peptides. Another way is to thread fragments of the sequence into knows structure fragments and subsequently assembly them (e.g. ROSETTA (41)). This approach can be applied to single domain proteins since it is less costly.

1.2 Protein-protein interactions

Cells are complex biological entities whose normal function revolves around a delicate interplay between different biomolecules. Among them, proteins are fundamental to carry out most of the molecular functions, acting as molecular machines, sensors, transporters and structural elements (among others), with physical interactions between proteins being key to their function (42). Hence, the study of protein-proteins interactions (PPI) is crucial to understand cellular processes and develop new therapies for the treatment of human diseases.

PPI networks have been used widely with different purposes, such as inferring protein function, identification of drug targets, identification of biomarkers or prioritizing disease-gene candidates in network medicine (43). One of the major challenges in current biomedical research is the determination of the whole network of PPI for a given organism (defined as the interactome) (44).

1.2.1 Levels of detail

Protein-protein interactions can be studied from different levels of details, depending on the research focus (3). Going from lower to higher, we distinguish the five different levels:

1) Proteins related by co-expression or co-localization: Proteins that are found at the same time or same location can be used to predict functional relationships and to validate experimental results.

2) Proteins that belong to the same macromolecular complex: Proteins belonging to the same complex do not necessary have physical interactions, but it is said they have indirect interactions.

3) Pairs of proteins that physically interact: This level recognizes physical interactions between two proteins (binary interactions), which can be used to create networks. Then, the topology of the network can be analysed using graph-theory algorithms.

4) The determination of the interacting region: The knowledge of a binary complex starts by knowing the binding region, which is essential to understand the molecular mechanism involved in the association.

5) The determination of the structural details of the PPI: The most detailed level of knowledge of a binary complex is obtained by focusing on the on the structural atomic details of the residues forming the interaction.

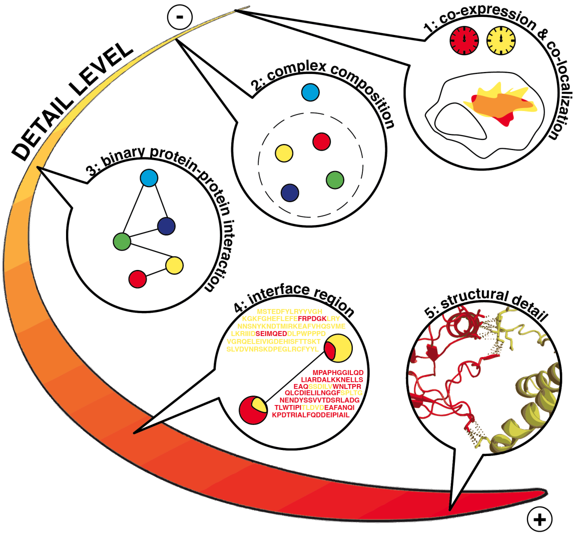


Figure 1.2 Five level details in PPI. A representation of each level is shown from low (-) to higher (+) detail. Image obtained from (3).

1.2.2 Protein-protein interaction determination

Over the years, the number of technologies for the identification and characterization of binary PPI has increased dramatically. Each method has its own strengths and weaknesses. Thus, the selection of the appropriate technique depends on the goal of the study (interactome-wide or small defined set of PPI), the nature of the PPI studied (transient or stable interactions), the time/costs constraints and the specialized equipment and expertise (45).

Protein Complementation Assays (PCA) (46) represents the group of the most widely used. In PCA protocols, the two proteins (bait and prey) are covalently linked at genetic level to incomplete fragments of a reporter protein. The reporter protein is usually a transcription factor that regulates a gene, which upon activation shows an observable phenotype. Thus, if the bait and prey interacts, the two fragments of the reporter are close enough to trigger the reporter activity. Among PCA, Yeast Two Hybrid (Y2H) remains one of the most popular methods. Y2H is simple, well established, low cost and can be used high-throughput (47).

Other approaches are the proximity-based methods, such as the Fluorescence Resonance Energy Transfer (FRET) (48). This technique is based on the transfer of energy from an exited donor fluorophore to a nearby acceptor molecule. The Bioluminescence Resonance Energy Transfer (BRET) (49) is a similar method that used luciferase as donor and a fluorescent protein as acceptor, dismissing the strong background signal of FRET. Both methods are suitable to monitor short-lived real-time PPI.

The previous methods involve the creation of fusion proteins, which may affect the binding ability of the targets. Thus, some methods use antibodies to surmount this problem, such as the Proximity Ligation Assay (PLA) (50) or the Affinity Purification-Mass Spectrometry (AP-MS) (51).

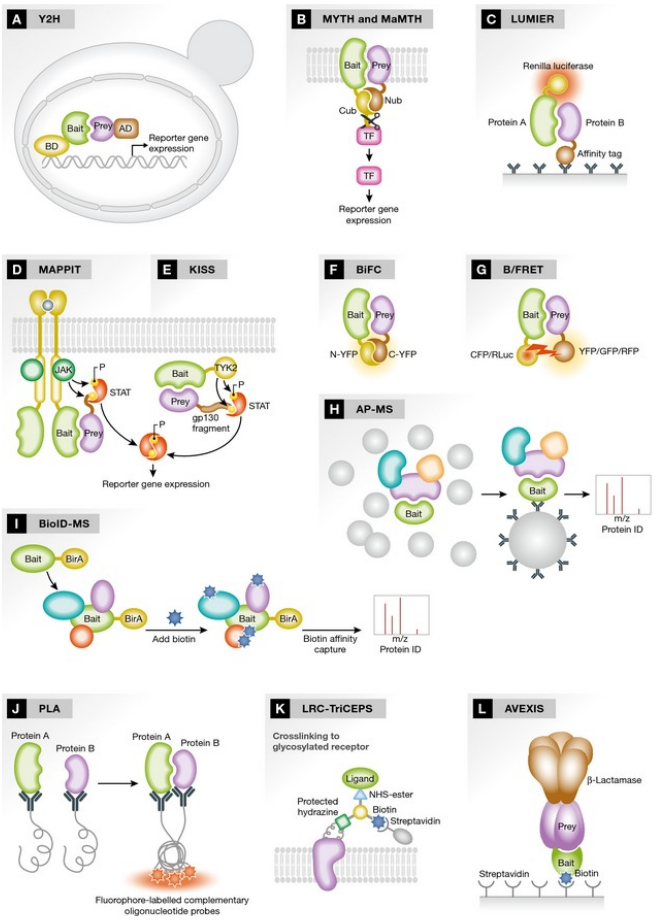


Figure 1.3 PPI detection methods. A: Yeast Two Hybrid. B: Membrane Yeast Two Hybrid. C: Luminescence-based Mammalian interactome mapping. D: Mammalian Protein-protein Interaction Trap. E: Kinase Substrate Sensor. F: Bimolecular Fluorescence Complementation. G: Bioluminescence/Fluorescence Resonance Energy Transfer. H: Affinity Purification-Mass Spectrometry. I: Proximity-dependent Biotin Identification Coupled to Mass Spectrometry. J: Proximity Ligation Assay. K: Ligand-Receptor Capture-Trifunctional Chemoproteomics Reagents. D: Avidity-based Extracellular Interaction Screen. Image obtained from (45).

Finally, the methods described for the determination of protein structures (X-rays crystallography, Nuclear Magnetic Resonance, Cryo-electron tomography) can also be used to detect PPI and provide further information about the structural details of the interaction.

1.2.3 Prediction of binary interactions

High-throughput experimental methods have produced a large amount of PPI data. However, the reliability and coverage of those techniques has been questioned (52). Thus, in order to improve data quality and fill the interactome gaps several computational methods have been developed.

PPI prediction methods of binary interactions are mainly used to infer new protein-protein interactions (regardless of the atomic details) but also to validate experimental results. These methods can be categorized by the data they use:

a) Genomic data: These methods analyse conserved operons, fusion domains or phylogenetic profiles, among others. Analysis of operon is base on the idea that genes in close proximity in the genome are more likely to encode proteins that interact (53). Similarly happens if two genes exist in a single fused gene in other species (54). Phylogenetic profiles identify genes pairs that tend to co-occur across genomes (55).

b) Protein sequence: Many approaches use directly or indirectly the sequence for the PPI predictions. Some methods directly analyse know sequences in order to find patters (typically using machine learning algorithms) that distinguish interacting proteins from non-interacting pairs (56,57). Other approaches use sequence homology to find interologs (two proteins that interacts in one specie are more likely to interact in another) (58) or signatures, such as domains, characteristics of interactions (10). Furthermore, some signatures may be related to structural features, such as in iLoops method (59,60). Correlated mutation is also another method based on sequence and evolution (61).

c) Protein structure: Similarly to protein sequence, structure can be used to do discern between interacting and non-interacting proteins using structural patterns or structural homology, such as in PRISM (62) or InterPreTS (63). Moreover, the utilization of the tertiary structures can be used to infer PPI through affinity predictions or docking techniques.

d) PPI networks: New interactions may be predicted from the network structure of a partially known interactome, based on the principle that interacting proteins tend to share interaction partners (64,65).

1.2.4 Prediction of quaternary structure

Once a binary interaction is known; a more detailed level of knowledge is obtained studying the structural details of the residues that contributes to the interaction. This information can be used, for example, to estimate the effect of mutations and for ration drug design. However, the experimental costs for obtaining a protein complex are much more higher than those for determining a binary interaction, so the number of known interactions highly exceeds the number of protein complexes with known structure. Therefore, computation approaches have been developed to cover this gap (66).

There are two strategies for predicting a quaternary structure from sequence or structures (67). The most reliable is comparative modelling, based on the same principles as for individual proteins but using as a template the structure of an available interacting complex (interolog) (68). This method requires global similarity between the query and the template interactions, seriously limiting the applicability. This drawback can be partially overcome by taking into account that protein interface architectures are reused frequently (69). In fact, several studies suggest that the number of possible interfaces is smaller than the possible number of protein interactions (70). Thus, it is possible to model a protein complex using only the similarity of protein interfaces and the unbound structures of the two proteins (71).

In contrast to comparative modelling methods (which are based on structural knowledge of the interaction), docking methods use the two unbound structures to sample orientations and produce several predictions that are ranked according to a scoring function to find near-native conformations (72). The docking process can be more accurate and fast if the interface is known, since the sampling space is reduced (73). However, docking remains an unresolved challenge in structural bioinformatics (74).

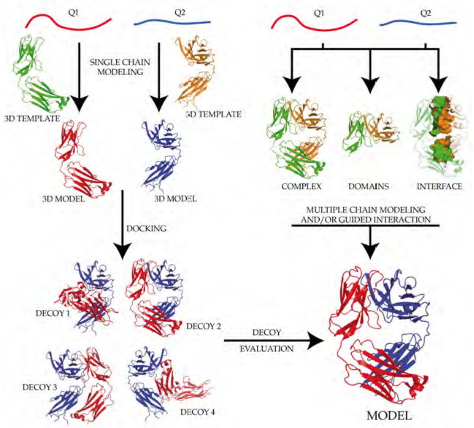


Figure 1.4 Strategies for predicting quaternary structure. Left, docking approach. Right, comparative modelling approach. Image obtained from (67).

1.2.4.1 Interface prediction

Computational methods aimed to predict the binding interface fall into two categories, depending on whether they require the knowledge of both interacting partners or not.

The identification of a binding site regardless of the protein partners is possible due to the fact that interacting regions have specific properties different from the rest of the protein surface (75). For example, binding regions are more conserved and present different composition of amino acids (e.g. more hydrophobic in obligate complexes) (76). This information can be combined using machine learning methods to improve the binding site predictions (77). For example, VORFFIP is a structure-based binding site predictor that uses residue features and voronoi diagrams to train a two-step random forest classifier (78). The method scores the residues and evaluates what regions are more likely to interact. Other structure-based machine learning predictors are: SPPIDER (79), PINUP (80), ProMate (81) and PIER (82).

On the other hand, the knowledge of the interacting partners can provide information through co-evolution constraints (83) and structural or sequence patterns (84), among others. Furthermore, network topology based methods have been successfully used to predict the interface region (85).

1.2.4.2 Docking

Protein docking is a computational strategy for elucidating the structure of a protein complex when the unbound structures are determined (or modelled) and no data regarding the structure of the complex is known. These methods where introduced in 1978 (86). Since then, docking algorithms have improved substantially. The docking process typically involves two main steps (87):

1) Sampling: During this stage a rigid search is performed. One of the proteins is kept fixed (receptor) while the other (ligand) is rotated and translated around the first. According to the search strategies, current docking programs can be grouped into three categories: Fast Fourier Transform (FFT) or Spherical Fourier Transform (SFT), local shape feature matching approaches and randomized search algorithms (88). Some programs allow the introduction of restrictions, such as the maximum distance between two residues or specifying the binding sites of both proteins. These restrictions reduce the sampling space, which increase the computational speed and accuracy.

2) Ranking: Rigid-body search yields a large number of output conformations, which includes many false positives. Thus, the different poses needs to be ranked typically by means of a scoring function to discriminate near-native structures (top best scored decoys). Finally, some decoys can be refined adding side chain movements and backbone flexibility. Examples of post-docking algorithms for flexible refinement are: MutiDock (89), RDOCK (90), FireDock (91), FiberDock (92,93) and EigenHex (94). The resultant candidates can be re-ranked and clustered to avoid redundant poses.

Selecting the appropriate scoring function is crucial to rank near-native conformations at top. They can be grouped into three basic categories: force-field based, knowledge based (statistical potentials) and empirical (67). The scoring function can be computed at residue or atomic level. Atomic level is more detailed but residue level is less sensible to small conformational changes.

Force-field scoring functions usually are atomistic-detailed and a linear combination of energetic terms such as Van der Waals, electrostatics, hydrogen bonding and desolvation energy (95). On the other side, knowledge-based potentials rely on the statistical analysis of specific properties observed in known protein-protein interactions stored in some database (96).

In statistical potentials, the interaction between two residues can be statistically described by a potential of mean force (PMF). The energy score of the interaction is obtained by summing the potential of mean force of each pair of interacting residues *a*, *b* of the two proteins:

$$E = \sum_{a,b} PMF(a,b)$$

The standard residue-pair potential of mean force (PMF) is obtained from the probability of finding a pair of residues (a, b) at a given distance (d_{ab}) (k_B denote the Boltzmann constant and T the standard temperature of 300K):

$$PMF_{pair}(a,b) = -k_B T log\left(\frac{P(a,b|d_{ab})}{P(a)P(b)P(d_{ab})}\right)$$

In a recent work, the standard residue-pair potential is decomposed into four new statistical potentials that reflect different level of detail of the residue-residue interactions (96). To do that, the surface accessibility and secondary structure of the residues are considered. Their success has been evaluated in different databases.

$$\theta = (Secondary\ structure, polar\ c\hbar aracter,\ degree\ of\ exposure\ )$$

$$PMF_{local}(a,b) = k_BTlog\left(\frac{P(a|\theta_a)P(\theta_a)}{P(a)}\right) + k_BTlog\left(\frac{P(b|\theta_b)P(\theta_b)}{P(b)}\right)$$

$$PMF_{3D}(a,b) = k_BTlog(P(d_{ab}))$$

$$PMF_{3DC}(a,b) = k_BTlog\left(\frac{P(\theta_a,\theta_b|d_{ab})}{P(\theta_a,\theta_b)}\right)$$

$$PMF_{S3DC}(a,b) = -k_BTlog\left(\frac{P(a,b|d_{ab},\theta_a,\theta_b)P(\theta_a,\theta_b)}{P(a,b|\ \theta_a,\theta_b)P(\ \theta_a,\theta_b|d_{ab})}\right)$$

Note that the statistical potential E$_{s3dc}$ is a refinement of the standard residue-pair statistical potential, since it takes into account not only the residues but also the condition in which each of them.

Even thought the majority of the docking programs include both steps (sampling and ranking), these can be performed by different algorithms. In fact, it is common to re-rank the docking poses using different scoring functions (e.g. RPScore (97), ZRANK (95), PyDock (98), EMPIRE (99), DARS (100), DECK (101), SIPPER (102), PIE (103), etc.) and even mix them (96). Table 1.1 shows current docking programs along with the search strategy and the default scoring function implemented.

Program	Search algorithm	Scoring function
FTDock (104)	FFT-based correlation	Shape complementary and electrostatics
GRAMM (105)	FFT-based correlation	Shape complementary and hydrophobic match
MolFit (106)	FFT-based correlation	Geometric complementary, hydrophobic complementarity and electrostatics
DOT (107)	FFT-based correlation	Van der Waals and electrostatics
ZDOCK 3.0.2	FFT-based correlation	Shape complementarity, electrostatics

(108)		and knowledge-based pair potentials
PIPER (109)	FFT-based correlation	Shape complementarity, electrostatics and knowledge-based pair potentials
SDOCK (110)	FFT-based correlation	Van der Waals attractive, geometric collision, electrostatics and desolvation energy
F2DOCK (111)	FFT-based correlation	Shape complementarity and electrostatics
ASPDock (112)	FFT-based correlation	Atomic solvation parameters
HEX (113)	SFT-based correlation	Surface complementarity and electrostatics
FRODOCK (114)	SFT-based correlation	Van der Waals, electrostatics and desolvation energy
GAPDOCK (115)	Local shape match (Genetic algorithm)	Surface and chemical complementarity
PatchDock (116)	Local shape match (Geometric hashing)	Geometric shape complementarity
SymmDock (116)	Local shape match (Geometric hashing)	Geometric shape complementarity
LZerD (117)	Local shape match (Geometric hashing)	Geometric complementarity based on the 3D Zernike shape descriptors
ATTRACT (118)	Randomized search (Monte Carlo search)	LJ-type effective potentials and electrostatics
RosettaDock (119)	Randomized search (Monte Carlo search)	Van der Waals, electrostatics, hydrogen-bonding, pair-wise interactions and solvation energy
ICM-DISCO (120)	Randomized search (Monte Carlo search)	Van der Waals, electrostatics, hydrogen-bonding, hydrophobic potential and desolvation energy
HADDOCK (121)	Randomized search (Monte Carlo search)	Van der Waals, electrostatics, BSA and desolvation energy
SwarmDock (122)	Randomized search (Swarm optimization)	Van der Waals and electrostatics

AutoDock (123)	Randomized search (Genetic algorithm)	Empirical free energy function

Table 1.1 Docking programs. Columns from left to right: name of the program and refs, search algorithm and scoring function. Adapted from (72) and (88).

A recent work compared the success rates (the number of protein for which at least an acceptable solution is found in the top *n* decoys) for 115 scoring functions, yielding top 10 success rates up to 58% (figure 1.5) (74). If we take into account the large number of docking poses obtained from the sampling stage, those scoring functions are quite successful. However, the reliability for practical use is still limited.

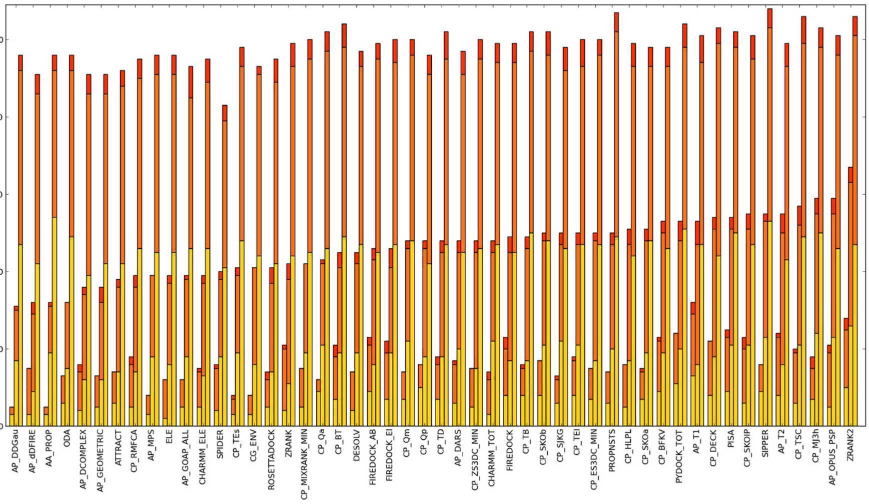


Figure 1.5 Success rates for best 40 scoring functions. Acceptable, medium and high quality solutions are shown in yellow, orange and red respectively. Top 1, 10 and 100 measures are in left, middle and right. Image obtained from (74).

1.2.5 Protein-protein interaction databases

The results obtained from the PPI determination experiments and prediction methods are deposited in public repositories, which allows the access to the information available for its further analysis and utilization.

Databases differ in different aspects, such as the level of detail recorded or the experimental/computational methods used for the acquisition. Moreover, reliability and completeness of the databases have been questioned (124,125). Therefore, the integration of such data in unified systems is still a challenging crucial task in biomedical research (126).

The following table (table 1.2) summarizes the principal databases on PPI with the detail deep according to the classification shown in section 1.2.1.

Databases	Information	Level of detail
STRING (127)	Functional relation between proteins, including co-localization and co-expression.	1,2, and 3
BIND (128) IntAct (129) DIP (130) BioGRID (131) HPRD (132) MINT (133) MPact (134) MIPS (135) HPID (136)	Complex composition and binary pairs obtained experimentally	2 and 3
PIPs (137) OPHID (138) POINT (139)	Predictions of PPI obtained from different methods	2 and 3

Domine (140) PSIbase (141)	Domain-domain interactions observed in the PDB database	3 and 4
PCRPi-DB (142) HotRegion (143) HotSprint (144) ASEdb (145)	Hot-spot databases (residues in the interface that greatly contributes to binding energy)	4
Affinity Benchmark (146)	Benchmark of protein-protein affinities (K_d).	4 and 5
iPfam (147) 3DID (148) SCOPPI (149) SCOWLP (150) PIBASE (151) InterPare (152) PRINT (69)	Structurally determined domain-domain interactions	5
PDBSUM (153) PROTCOM (154)	Databases of protein complexes	5
Docking Benchmark (155)	High-resolution protein complexes with bound and unbound structures.	5
InterEvol (156)	Evolution of protein-protein complex interfaces	4 and 5

Table 1.2 PPI repositories. Columns from left to right: name of the databases, information regarding the nature of the data and level of detail (1-5) according to classification of section 1.2.1. Adapted from (3)

1.3 Protein-protein recognition

The formation of a protein complex involves two steps: the translational and rotational diffusion of the proteins in the solvent environment that brings the interfaces to an orientation of high specificity and the conformational changes upon binding.

From the point of view of conformational changes, the three classical mechanisms proposed over the last century to describe the binding process are the (a) lock and key, (b) induced fit and (c) conformational selection.

a) The Lock and key model was first postulated in 1894 by Emil Fischer (157) and describe that for rigid proteins (the interface is nearly identical in bound and unbound states) the interaction is achieved through the complementarity of the binding sites.

b) A second recognition model is the induced fit, which postulate that the binding of one protein to another induces conformational changes, ranging from small side-chain or surface loops movements to large movement of domains, which result in the bound complex (158).

c) Finally, the third model is the conformational selection. This mechanism hypothesized that the unbound structures fluctuates in multiple conformations, the best fitting is which proceed to form the protein complex (159,160).

Both mechanisms, induced fit and conformational selection, are not mutually exclusive and has been observed to occur simultaneous (161) and in a sequential manner (162). In fact, the three described models have been observed experimentally.

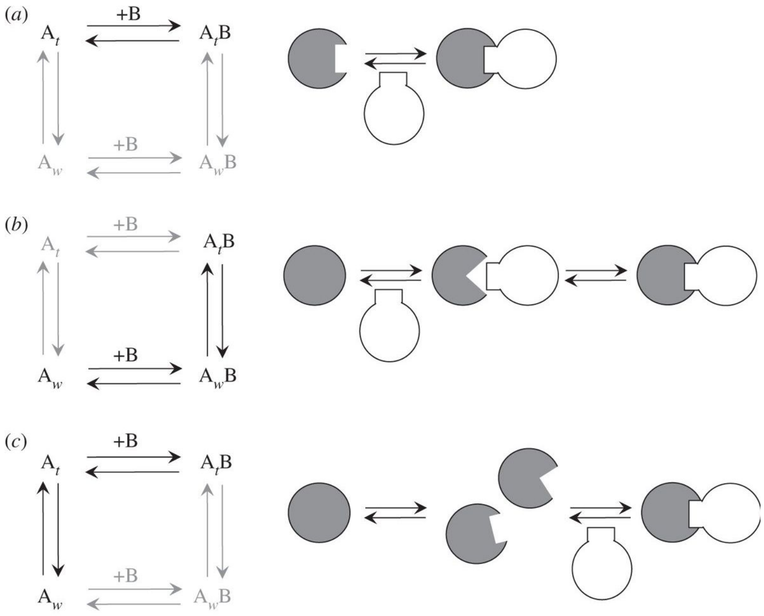


Figure 1.6 Basic mechanisms for protein binding. (a) Lock and Key, (b) induced fit, (c) conformational selection. A_t and A_w denote protein at bound and unbound conformation respectively. The chemical reactions that take place are indicated black arrows. Protein B does not undergo conformational changes for simplicity. Image obtained from (163).

1.3.1 Diffusion in protein-protein association

The association of protein molecules is characterized by a second order kinetic rate constant (k_{on}) that typically ranges between 10^5 and 10^6 $M^{-1}s^{-1}$. Antibody-protein association rate constants are commonly observed in this narrow range. This range appears to represent the typical rate for associating without any special steering forces in a diffusion rate-limiting step (164).

When considering the steric specificity of the bonds connecting the two proteins, this rate seems to be very fast. For example, for two spherical proteins of 18 Å radius that form complex in every collision regardless of the orientation, the association rate constant would be given by the Smoluchowski rate constant: $k_{on} = 7*10^{9}\ M^{-1}\ s^{-1}$ (165). As average binding interface constitutes about 10-15% of total surface area (166), the probability of that a random collision will result in a productive orientation is extremely small, implying a theoretical 10^{6} fold decrease ($k_{on} = 7*10^{3}\ M^{-1}\ s^{-1}$), 10^{3} less than observed values experimentally (164). Thus, for a long time is has been recognised that some additional forces must be responsible for accelerating the macromolecular association in solution (167).

The first mechanism to explain this discrepancy was proposed by Sommer *et al.* and was called the "lengthy collisions between proteins" (168). The hypothesis suggested that interacting partners form weakly bonded nonspecific complexes, which are held closely for a long time together but are free to rotate. Consequently, the rotational diffusion would eventually bring the binding regions in contact. However, this hypothesis is not supported experimentally since the nonspecific complexes would have a $K_d=10^{-4}$ and a lifetime of microseconds, 100 times larger than observed.

In 1985, Berg further developed these hypotheses (169). He proposed that the proteins sustain multiple collisions that upon dissociation of an unproductive complex the dissociated proteins would remain spatially arranged increasing the chance to a second productive collision. Further Brownian dynamics studies provided a quantitative description of this rate enhancement mechanism (164). It was shown that two neutral spherical proteins surrounded by water undergo multiple collisions and rotational reorientations before they separate caused by a diffusive entrapment effect. This studies concluded that in the absent of any biasing force, the "basal" value of k_{on} ranges between 10^5 and 10^6 $M^{-1}s^{-1}$ in agreement with the typical association rates observed experimentally.

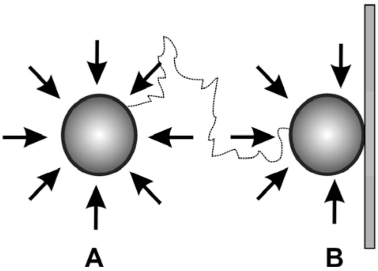


Figure 1.7 Brownian diffusion through liquids. Macromolecule pushed by water moving in a Brownian way. When it collides with another molecule (a wall in the schema) water push against the other molecule. Image obtained from (170).

1.3.2 Electrostatic rate enhancement in protein-protein association

Many protein associations occur at much higher rate than basal estimations of 10^6 $M^{-1}s^{-1}$ (arriving up to $>10^9$ $M^{-1}s^{-1}$), suggesting that intermolecular forces must be present in order to extend the time that proteins remains spatially close (171). This rate enhancement is crucial in a wide range of biological processes. For example, rapid binding of cytotoxic nucleases with inhibitors is essential to avoid damage in the host cell.

Site-directed mutagenesis and Brownian dynamics simulations have suggested that long-range electrostatic interactions greatly enhance association rates (172). Indeed, proteins with high complementary electrostatic surfaces have been observed in transient complexes with fast association rate constants (173). Furthermore, studies in protein design shown that association rate can be enhanced by optimizing the electrostatic attraction between proteins (174). Figure 1.7 shows the example of the interaction between TEM1 β-lactamase and its protein inhibitor BLIP. In this example, 6 mutations were introduces in BLIP to increase the electrostatic complementary between specific surface patches (outside but near the interface), resulting in a 200-fold increase of the association constant.

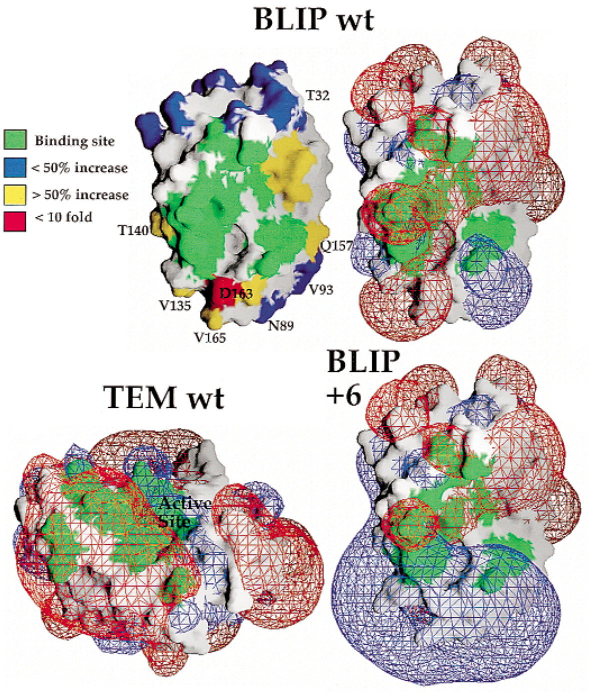


Figure 1.8 Design of a superactive BLIP inhibitor. Electrostatic potentials on the TEM1 wt, BLIP wt and BLIP enhanced mutant surfaces (blue for positive and red for negative charge). Green patch denotes the binding interface. Image obtained from (174).

1.3.3 Encounter complex

The complete ensemble of collisions happening during the rotational search is commonly called the "encounter complex". Recent studies using paramagnetic relaxation enhancement RMN allowed the visualization of the encounter complexes for relatively week protein-protein complexes (175,176). These studied have confirmed that the formation of encounter complexes is predominantly driven by weak non-specific electrostatic attractions. Moreover, the non-specific interfaces are more planar and small than the stereospecific complex (the buried surface area is an order of magnitude smaller), indicative of the absence of lock and key binding (175).

Once the non-specific encounter complex is formed by electrostatic interactions, both proteins carry out a two-dimensional search on the surface, eventually falling into a narrow energy funnel that leads directly to the stereospecific complex. Figure 1.8 shows of the encounter complexes between the amino-terminal domain of enzyme I (EIN) and the phosphocarrier protein HPr. It has to be noted that the region on EIN comprising the specific interaction surface for HPr is only minimally occupied by encounter complexes, suggesting that once he HPr reaches this region the formation of the stereospecific complex occurs with high probability.

It is becoming clear that the formation of an encounter complex is crucial to reduce the time necessary for formation of the productive complex (170). This means that the residues outside the binding site but relevant for the encounter complex may well have been optimized by evolution. Only by studying both the productive complex and the dynamic encounter complex will it be possible to fully understand protein complexes.

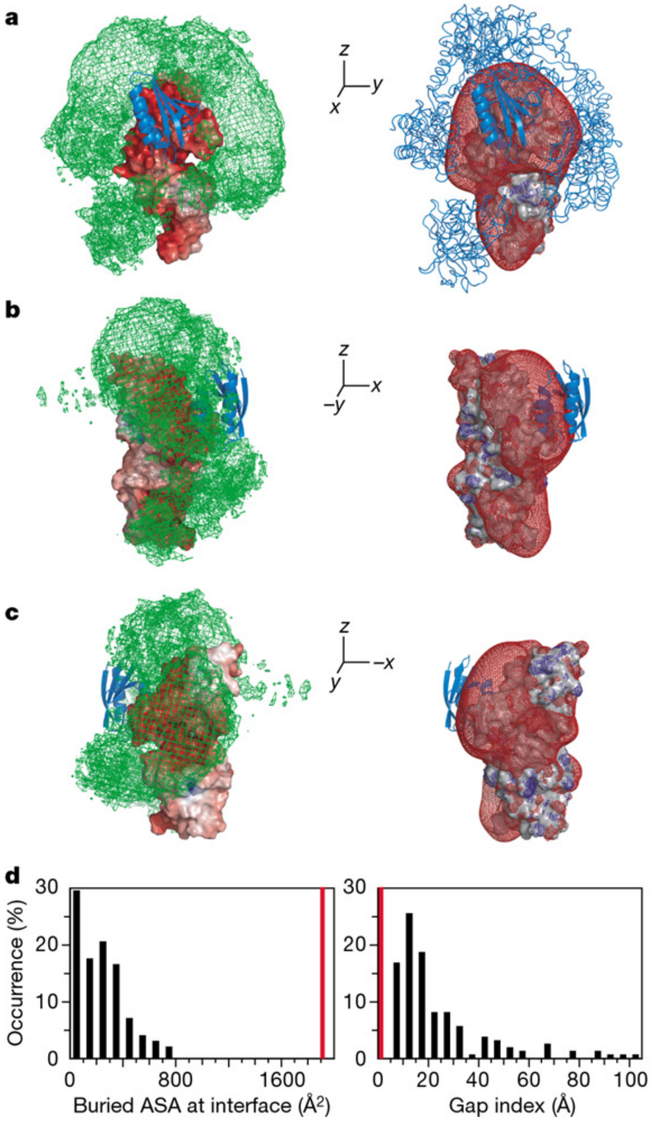
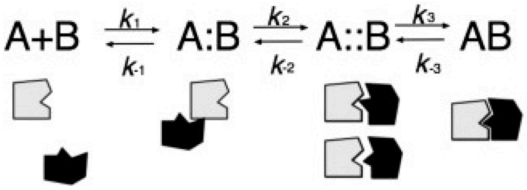


Figure 1.9 Visualization of EIN-HPr encounter complexes. Left, atomic probability density map (green) of HPr on the surface of EIN. The stereospecific complex is shown in blue. Right, electrostatic potential isosurface of EIN. Histograms of BSA and gap index distributions of the encounter complexes. Red line indicates the values for the stereospecific complex. Image obtained from (175).

1.3.4 Transition state

Association rates also occur at slower rates than basal ($k_{on} = <10^5\ M^{-1}s^{-1}$), suggesting that in some proteins energy obstacles limits the association process. Some studies suggest that large conformational changes upon binding (the induced-fit process) require the proteins to pass over a transition state that can act as an energy barrier (177).

To better understand this phenomenon it is useful to consider the association process as going through three steps. First of all, the encounter complexes are formed very fast from the free proteins in solution. Then, the two-dimensional search of both surfaces results in a transient complex that can lead to the naïve complex through conformational rearrangements (178). Figure 1.9 shows the formation of a productive protein through the three-step model, being the last step the transition state.



$$A+B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} A{:}B \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} A{::}B \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} AB$$

Figure 1.10 Three-step complex formation. From the left to the right: Proteins in solution, encounter complex, transient complex and native complex. Image obtained from (178).

In figure 1.10 3D funnel energy diagrams are shown to illustrate various association regimes for two proteins that interact (170). Left funnel shows proteins highly transient and dynamic that does not achieve high specificity, such as electron transfer complexes. In the middle, all the encounter complexes proceed to produce the productive complex, being the diffusion the limiting step. Conversely, on the right, many encounter complexes dissociate before the productive complexes is formed. Therefore, for these complexes the k_{on} is limited by the transition state.

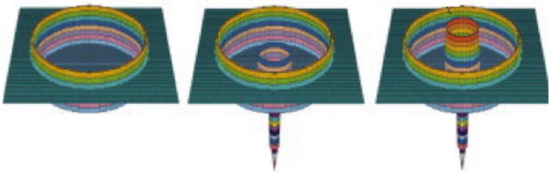


Figure 1.11 Energy diagrams. From the left to the right: Proteins with low specificity, proteins limited by diffusion, proteins limited by transient state. Image obtained from (170).

In fact, a simple classification for the wide spectrum of experimental association rate constants has been recently reported (171,177). This study describes two limiting regimes: the conformational change-limiting regime (below basal $k_{on} = 10^5$ $M^{-1}s^{-1}$) and the diffusion-limiting regime. If there is no long-range force to bias the $k_{on,}$ it falls between 10^5 and 10^6 $M^{-1}s^{-1}$. Proteins with complementary electrostatic surfaces can enhance association to > 10^6 $M^{-1}s^{-1}$.

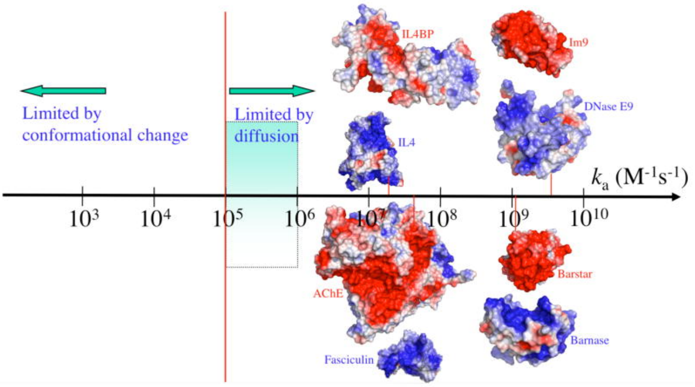


Figure 1.12 Association rate limiting mechanisms. Vertical line separates the rate limiting regimes. Green region shows the basal k_{on}. Four complexes with higher rate constants are shown with the electrostatic surfaces (blue for positive and red for negative charge). Image obtained from (171)

It is worth highlighting that recent proteomic studies in protein-protein association show that the variation in basal rates is wider and larger than was estimated previously, from 10^4 $M^{-1}s^{-1}$ to 10^7 $M^{-1}s^{-1}$ (179). Hence, the emerging picture is that electrostatics has a limited role in dictating association in most complexes, increasing their rate by up to 100-fold. Only in a few extreme cases, electrostatic has a major role. The basal rates were slightly correlated with the size of the protein and the shape of the interface. However, a second and unknown mechanism should be present to explain these high basal rates. Some studies suggest that short-range electrostatics and desolvation defines weakly specific pathways leading to a low free energy attractor embedded in a repulsive environment guiding the proteins towards well oriented kinetic intermediates (180). Nevertheless, more work has to be done to study this phenomenon.

1.4 Binding affinity

The binding between two proteins can be viewed as a reversible and rapid process in an equilibrium that is governed by the law of mass action. The strength of the interaction is defined as the binding affinity and may be influenced by temperature, ionic strength, pH and posttranslational modifications (163). Binding affinity spans more than 10 orders of magnitude, from high μM to fM, and changes caused by mutations or post-traductional modification may be responsible of many protein dysfunctions and disease (181,182).

The binding equilibrium dissociation constant (K_d) is a function of the rates of association (k_{on}) and dissociation (k_{off}), with the simple relation:

$$K_d=k_{off}/k_{on}$$

The rate constants k_{on} and k_{off} determine the timescale of association and dissociation, providing a “dynamic” view of the protein complex. This kinetic nature is a crucial aspect in diverse biological processes (171). For example, for proteins involved in cell signalling slow dissociation is not a good option since it implies a long-lasting bound state and a permanent on switch. Thus, fast dissociation is partially compensated through fast association.

The dissociation equilibrium constant can be empirically translated into the Gibbs free energy of binding:

$$\Delta G_d = -RT\ln K_d = \Delta H_d\text{-}\Delta TS_d$$

R is the gas constant (8.3144 J K^{-1} mol^{-1} equal to 1.9872 cal K^{-1} mol^{-1}) and T the absolute temperature in kelvin.

The free energy reflects all the chemical and energetic factors involved in the binding reaction. These factors can be decomposed between chemical forces acting on the two proteins mainly through non-covalent bonds (analogous to enthalpy) and entropy changes (183).

A clear distinction between forces affecting k_{on} and k_{off} has been observed (184). Dissociation is a first order reaction whose rate is dictated by the strength of short-range interactions between proteins (van der Waals interactions, hydrogen bonds, hydrophobic interactions and salt bridges). Conversely, association rate is dictated by long-range electrostatic forces along the protein surface, as we described previously.

1.4.1 Binding affinity determination

For the determination of the binding affinity of a biological reaction between two proteins several methods have been developed (185). Overall, these methods can be categorized in two general classes: separative (direct) or non-separative (indirect) (186).

a) Direct methods measure the actual concentration of the bound and free proteins. These methods are only appropriate for proteins exhibiting slow dissociation rates, since the process of separating the bound and unbound proteins might disturb the equilibrium if the dissociation and separation occur on similar time-scales. Gel filtration, ultracentrifugation, ultrafiltration and equilibrium dialysis are examples of direct methods.

b) Indirect methods infer the concentration of the bound and free proteins from an observed signal. Optical methods such as absorbance, resonance or fluorescence spectroscopy techniques belong to this class, where the measured signal is proportional to the concentration of the product.

The current three most frequently used methods to experimentally determine the binding affinity between two proteins are: the isothermal titration calorimetry (ITC) (187), the surface plasmon resonance (SPR) (188) and the flurocence-based methods (189).

1.4.2 Binding affinity prediction

Experimental techniques for measuring binding affinity (ΔG_d) are expensive and time-consuming. For this reason, many computational methods have been developed for more than 20 years to predict the binding affinity (190–199). Those approaches usually consider properties of the complex interface and calculate empirical scoring functions based on statistical potentials, thermodynamic equations and scoring functions used in docking. Thus, these methods are based on the energetic forces affecting k_{off}. Recently, few methods also incorporate properties of the non-interacting surface into the predictive models (200). However, most of them still have poor performance when tested against large datasets (201). Moreover, the applicability of those methods is limited to the determination of the quaternary structure of the interaction.

It is noteworthy to mention the Buried Surface Area (BSA), since was the primary descriptor to be related to binding affinity (183). BSA is a macroscopic descriptor for the hydrophobic interactions and expresses the gain of entropy of the water molecules upon binding. Its magnitude has been estimated to be 0.025 kcal mol^{-1} per 1 Å^2 of hydrophobic surface removed from contact with water. All the other non-covalent interactions are theorized as negligible, sine proteins are highly solvated when unbound. However, interface ions, hydrogen bonds and van der Waals must be complementary to avoid complex destabilization. This model falls when explaining hot-spots (residues that destabilize the bound state by more than 2 kcal mol^{-1} when mutating to alanine) (202). Thus, hydrophobic interactions are not the absolute determinants for the binding process.

1.5 Thesis motivation

Bioinformatics is an interdisciplinary scientific field aimed to develop software for storing and analysing biological data. As we have previously described, in structural bioinformatics is of great relevance the development of predictive tools to bridge the sequence-structure gap, either for single protein or for protein-protein interactions. In fact, several community-wide experiments in modelling structures, such as CASP (for tertiary structure) or CAPRI (for protein docking), take place regularly to deliver an assessment of the state of the art to the research community (203–206). Most docking tools are computationally costly and result in a large amount of false positives. Thus, it is important to develop new tools to improve speed and accuracy in order to make it useful and affordable even for high throughput. In this thesis we address this problem by developing V-D^{2}OCK (73).

In addition, the determination of the binding affinity and association mechanisms is key for understanding association and dysfunction of protein complexes. Nowadays, the binding affinity predictors rely on the quaternary structure to develop scoring functions based on the characteristics of the interface. Despite all the efforts regarding the development of accurate docking programs, the reliability of those methods is still limited. Thus, current binding affinity predictors can only be applied to few cases. Here, we present a novel strategy to predict binding affinity of soluble globular proteins from the unbound individual protein structures, increasing the applicability over the existent methods. This strategy is based on the study of non-specific poses obtained by docking to scout the conformational space of potential encounter complexes formed during the association process, which contributes to the binding affinity.

Finally, the integration of computational programs in centralized resources is vital to collect and put into value all these tools. In this thesis we present InteractoMIX (207), a website that not only allows an easy access to a compendium of bioinformatics web applications to exploit interactomes, but also allows a better understanding of the connectivity and relations between them.

2. OBJECTIVES

This PhD thesis aims to fulfil the following objectives:

a. Develop and analyse a new data-driven docking strategy with improved speed and accuracy for high-throughput, genome wide, protein docking.

b. Study the non-specific decoys obtained by docking (potential encounter complexes) to develop a binding affinity predictor for globular soluble proteins from the unbound proteins structures (unknown quaternary structure).

c. Describe and develop a complete, integrated and centralized resource (i.e. a web site) to facilitate the access to a compendium of web applications for the analysis of protein-protein interactions and interactomes at both sequence and structure level.

The first objective (a) has been addressed by contributing to develop V-D^{2}OCK, a fast and accurate data-driven protein docking strategy. V-D^{2}OCK publication is presented in **section 3.1**. The second objective (b) is accomplished by developing a novel strategy to predict the binding affinity of globular protein-protein interactions from non-specific docking poses. The manuscript of this work (in preparation) and is presented in **section 3.2**. Finally, the third objective (c) is achieved by developing InteractoMIX, a suite of computational tools to exploit interactomes in biological and clinical research. It is worth mentioning that V-D^{2}OCK is included in InteractoMIX. This resource is published in a scientific journal and constitutes **section 3.3**.

In the appendix, I present several published works I was involved during the PhD that are related to the objectives of this thesis.

3. RESULTS

3.1 VORFFIP-Driven Dock: V-D^{2}OCK, a Fast and Accurate Protein Docking Strategy

Large-scale interactomic experiments usually do not provide atomic details, resulting in an expanding gap between interactomic data and determined structures of protein-protein interactions. Protein docking is a computational approach aimed to derive structural models of protein complexes. Docking can be data-driven to restrict the sampling to selected regions. In this paper we present V-D^{2}OCK, a data-driven docking strategy that uses functional sites predicted by VORFFIP to direct the sampling process using PacthDock. The resultant docking candidates are clustered to remove redundant poses and ranked according to several scoring functions, including the new ES3DC statistical potential. The results obtained shows that V-D^{2}OCK efficiently samples the docking space finding poses near to native, performing similar, or even better in flexible cases (using ES3DC), to other state-of-art methods (ZDOCK) in a faster way. Thus, the speed and accuracy justify the usage of this application for high throughput. V-D^{2}OCK web server is accessible at:

http://www.bioinsilico.org/cgi-bin/VD2OCK/staticHTML/home.

Contribution: In this project I performed the experimental part regarding the implementation of the scoring function ES3DC. I also contribute to analyse the data (making the success rate curves and examples) and writing the article.

Segura J, Marín-López MA, Jones PF, Oliva B, Fernandez-Fuentes N. VORFFIP-Driven Dock: V-D2OCK, a Fast and Accurate Protein Docking Strategy. PLoS One. 2015 Mar 12;10(3):e0118107. DOI: 10.1371/journal.pone.0118107

3.2 On the mechanisms of protein-protein binding: predicting their affinity from unbound tertiary structures

Binding affinity (BA) is what defines if two proteins will interact or not and the strength of the interaction. Therefore, determining and predicting the BA of protein complexes is key for comprehending protein association and dysfunction caused by mutations or post-transcriptional modifications. Currently, most BA predictors rely on the atomic details of the complex interface, reducing the applicability of those methods to proteins with a quaternary structure elucidated. In this work, we present a novel strategy for predicting the binding affinity of soluble globular proteins from individual protein structures, partially solving the coverage problem. This strategy exploits the non-specific decoys obtained from docking to scout the conformational space of potential encounter complexes. These complexes are formed during the association process contributing to the binding affinity. Our results show that the developed method performs comparable to other state-of-art methods that require the native structure. In addition, our analysis of the non-specific docking poses shows a recognition path from non-productive poses to native and suggests that docking is a suitable method for studying the encounter complex. The manuscript of this work is in preparation. A web application implementing this strategy is available at: http://sbi.upf.edu/BADock

Contribution: In this project I contribute to conceive and coordinate the integration process. I also contribute to write the article.

Marín-López MA, Planas-Iglesias J, Aguirre-Plans J, Bonet J, Garcia-Garcia J, Fernandez-Fuentes N, et al. On the mechanisms of protein interactions: predicting their affinity from unbound tertiary structures. Bioinformatics. 2018 Feb 15;34(4):592–8. DOI: 10.1093/bioinformatics/btx616

3.3 InteractoMIX: a suite of computational tools to exploit interactomes in biological and clinical research

In the post-genomic era the new high throughput technologies resulted in an exponential increase of biological data. Computational tools, in combination with experimental data, are crucial to analyse these data in order to better understand biological systems. In this paper, we describe 11 published tools and databases for the analysis of protein-protein interactions (PPI). These tools have been integrated and centralized in a website: InteractoMIX, providing a comprehensive and easy access to the different tools. InteractoMIX is organized in two level of detail (interactomic and atomic level) indicating a workflow that can be followed for a particular study involving PPI. Interactomic-level tools are: BIANA for the integration interactomic data; iLoops and BIPS for the prediction of PPI and GUILDify for discovering new disease-related genes. Atomic-level tools are: ModLink+ for structure modelling; VORFFIP and M-VORFFIP for predicting functional sites in protein structures; V-D^{2}OCK for data-drive docking; PCRPi for predicting critical residues in protein interfaces and PiPreD for modelling orthosteric peptides. InteractoMIX is available at http://interactomix.com. Contribution: In this project I conceived, designed and performed all the experiment. I also analyse the data and write the article.

Poglayen D, Marín-López MA, Bonet J, Fornes O, Garcia-Garcia J, Planas-Iglesias J, et al. InteractoMIX: a suite of computational tools to exploit interactomes in biological and clinical research. Biochem Soc Trans. 2016 Jun 15;44(3):917–24. DOI: 10.1042/BST20150001

4. DISCUSSION

4.1 The docking problem

Protein-Protein docking algorithms have been developed to assist experimentalist in investigating how two proteins of known structure interact and form a three-dimensional complex. *Ab initio* docking is currently used to predict the quaternary structure of proteins when no data regarding their complex is known, so we can go beyond homology modelling (the preferred method) to theoretically predict the structure of all known protein complexes. Paradoxically, in this thesis (section 3.2) we present a binding affinity predictor whose main advantage is precisely that do not need the quaternary structure to predict the free energy. This seems contradictory; if docking is a valid and widely used approach to predict the structure of protein complexes it should be correct to rely binding affinity predictors on quaternary structure predictions. The answer to this issue is the large amount of false positives resulted from exploring the whole conformational space that hinders its practical use, especially as input for other predictor methods that requires reliable atomic details. Therefore, how good is docking? Can we consider the problem solved?

Docking is usually evaluated using success rate curves. These curves illustrate the percentage of benchmark cases with at least a near-native solution (the RMSD of the ligand is lower than 10Å) among the top N predictions. State of the art docking algorithms, such as ZDOCK, provides usually a ~10%, ~20% and ~30% success rate at top 1, 10 and 100 solutions respectively (figure 4.1). If we take into account the large number of docking candidates obtained from the conformational search step (54.000 using ZDOCK 6 degree sampling), the results are pretty accurate. However, if we select the top, top 10 or even top 100 solutions only for between 10% and 30% of cases we will obtain a near-native solution. That is a clear limitation in many research projects where precise atomic details are required. Thus, despite *ab* *initio* docking is better than nothing in many cases; it must be improved to be really useful for scientific community.

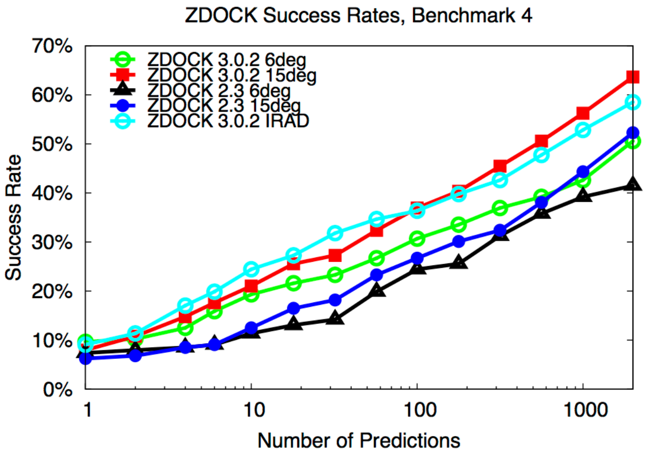


Figure 4.1 ZDOCK success rates. Success rates are shown for different versions and sampling degrees of ZDOCK (obtained from: https://zlab.umassmed.edu/zdock/perf_decoys.shtml).

Even thought *ab initio* docking is moving ahead at staggering speed, there is a trend to incorporate experimental information in the docking procedure in order to improve the accuracy. This information can be used either *a priori*, to drive the process, or *a posteriori*, to filter the obtained solutions. In addition, the *a priori* methods take advantage of the dramatic reduction the conformational space to be sample to increase the speed of the procedure. In this thesis have presented V^{2}DOCK, a data-driven docking tool that use the binding site predictions obtained by VORFFIP to drive the docking application PatchDock, increasing the speed of the process making it affordable for large-scale analysis.

The principal characteristic of V^{2}DOCK is the usage of binding site predictions, instead of experimental information. The main advantage of that feature is that we do not need previous knowledge about the protein complex. However, the first concern we arise is: how can affect the reliability of the predictions to the accuracy of the method? If we fail predicting the residues involved in binding site it won't be possible to find near-native solutions, even a worse approach that using *ab initio* docking. Therefore, it is crucial to keep enough sensitivity to allow the docking program to explore the interface while maximizing specificity. The analysis of V^{2}DOCK shows that beyond 20% of correct binding site residues predicted, PatchDock is able to find near-native solutions. Thus, it is not necessary to fully predict the binding site. The success rates of V^{2}DOCK show similar performance, or even better for flexible cases using the statistical potential ES3DC, to the state-of-art *ab initio* docking ZDOCK, so VORFFIP predictions are sensible enough. This validates V^{2}DOCK as a fast docking tool that can be used not only when speed is required, such as for high-throughput docking, but also as a good substitute to traditional *ab initio* methods. In addition, as V^{2}DOCK key value proposal is in the search step, this method can be improved as new and better scoring functions are developed.

4.2 Protein-protein binding mechanisms: lessons from iLoops and docking

During the research period of the PhD I collaborated in the iLoops project (Appendix 6.3 and 6.4), a protein-protein prediction method based on local structural features (group of super-secondary structures) defining characteristic patterns of interaction or non-interaction (positive and negative interaction signatures) by combining groups of structural features from both proteins. A notable trait of this system is that interaction signatures can be distributed along the protein surface, not only in the protein-protein interface, supporting previous works showing that not only the binding region is involved in the interaction process.

Hence, we used such features in order to study the differences between the binding interface and the rest of the protein surface in known protein-protein interactions (the PRISM database). Particularly, we study three different groups of protein-protein interfaces: i) Native interfaces (the actual binding patches of the interacting pairs), referred as Face-Face in the manuscript of section 3.2; ii) Partial interfaces (the putative interfaces between the binding patch of one protein and a non-interacting patch of the interacting partner), referred as Face-Back in the manuscript of section 3.2, and iii) Back-Back interfaces (the putative interfaces between non-interacting patches for both of the interacting proteins). To do so, we classify the interacting signatures into these three categories depending the on the origin (inside or outside the interface) of each super-secondary structures (see Figure 4.2).

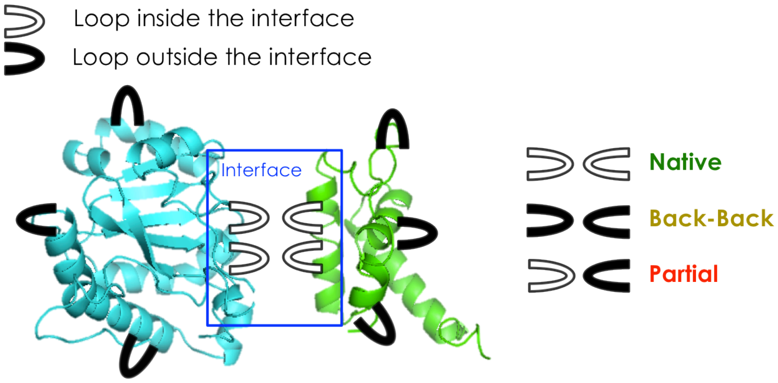


Figure 4.2 Interaction signatures classification. Native: white-white super-secondary structures; Partial: white-black super-secondary structures; Back-Back: black-black super-secondary structures.

We calculate the score of each class by averaging the minus logarithm of the p-value of the interacting signatures within the class for a given protein. In Figure 4.3 (left boxplots) we show the comparison of the distributions of the positive and negative interacting signatures for each class. We can observe that native and back-back classes have a better score distributions in positive signature than for negative (not in partial). Native class is a productive class, so this result was expected. Nevertheless, it was a surprising the difference between back-back and partial. We hypothesise that back-back interfaces preserve the exposure of both binding sites, while in a partial interface one interacting patch is sequestered and becomes unavailable to form a native interaction. According to this reasoning, partial interfaces represent a major obstacle in the formation of the real interaction. In comparison, although back-to-back interfaces also represent a wrong interacting conformation, they still expose the binding patches of both interacting partners and may represent an opportunity for the native conformation to occur.

We perform the same procedure between domains in proteins consisting of two domains from SCOP. In Figure 4.3 we can observe that we loose the previous pattern (right boxplots), back-back and partial shows the same profile. We suggested that since domains are connected by a linker, the previous reasoning do not apply. Thus, these results supported the previous hypothesis.

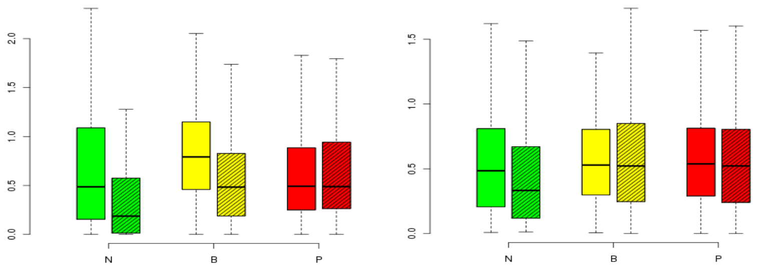


Figure 4.3 Boxplots for each interface class. Left figure: PRISM protein complexes; right figure: SCOP two domain proteins. Solid boxes show the distribution of the average of the -log(p-values) of the positive interacting signatures, (dashed boxes for negative signatures). Green: Native; yellow: Back-Back, red: Partial.

From here, we decided to use PatchDock and the Docking Benchmark 5.0 to work with real docking poses in order to validate the previous hypothesis in soluble globular proteins. However, as can be observed from BADock (section 3.2), we obtained the contrary results: partial (or face-back) was better than back-back. As expected from previous works about the encounter complex, we find high correlations between electrostatic energies and binding affinity for non-productive docking poses, whereas productive candidates better correlate with shape and Van der Waal forces. Thus, these experiments supported the theory of the 2D search of both surfaces (rotational diffusion) upon a collision of the protein partners. Moreover, we suggested that proteins move from back-back to face-back and then to productive conformations. Finally, we used these finding to create a binding affinity predictor from non-specific decoys, allowing us to perform the prediction without any knowledge regarding the native complex structure.

We know from recent proteomic studies that there is a high variability of association rate constants in absence of electrostatic forces. So, additional mechanism should exist in order to enhance association rate. Is the mechanisms observed with iLoops a secondary mechanisms that affects proteins without electrostatic forces? The high variability between types of proteins (high-low charged, permanent-transient, membrane-soluble, fast-slow dissociation, etc.) allows the coexistence different association enhancing mechanisms. Further work must be done to elucidate for each type of protein what are the mechanisms involved in their binding.

4.3 Developing bioinformatics tools

Currently, one of the main objectives of bioinformatics is the development of computational tools to store and analyse biological data. Most of these tools are implemented in web services and web applications, so the user does not need to install the software and all the dependencies and allows non-expert users to use these tools in an easy and friendly way. However, bioinformatics servers have many drawbacks: computational resource limitations, privacy, and usually after few years of the publication the web application stop working because lack of maintenance.

Another important issue about computational tools is the ever-growing number and redundant applications dispersed through different publications. Thus, it is very difficult for scientist to have a wide picture about the available tools for a specific problem. A suitable solution is to create web portals that aggregate available servers, so scientist just have to know the name of those resources in order to explore what is available. In addition, web portals can be designed to better comprehend the relation between the different tools.

In this thesis we have developed InteractoMIX (www.interactomix.com), a web portal that integrates 11 bioinformatics web applications. InteractoMIX interface is a bullseye that indicates a possible pipeline for the analysis of protein-protein interactions, so user can start at any position and follow the pipeline and exit when the desired results are obtained. InteractoMIX has demonstrated in our lab that helps new members to catch up faster about what tools have been previously developed and understand its functionality.

5. CONCLUSIONS

This section summarizes the main contributions presented in this thesis:

a) We developed a new data-driven docking approach with improved speed and accuracy using binding site predictions and several scoring functions.

b) We demonstrate that ES3DC statistical potential is an accurate score that can be used to rank docking poses with conformational changes upon binding (difficult cases).

c) We were able to use non-specific docking poses to explore the conformational space of potential encounter complexes and relate the different energetic components of productive and non-productive decoys with the binding affinity.

d) We show that non-productive decoys can be classified into two classes with different characteristics: Face-Back and Back-Back

e) We suggest a new recognition model in which globular proteins move towards to contact their binding sites (first one binding site and then the other) following a subtle energy funnel.

f) We can use the non-specific decoys to predict the binding affinity of two globular proteins, so we do not need the complex structure. We implemented this strategy using the ES3DC statistical potential with comparable results to other state-of-art methods.

g) We developed InteractoMIX, a new centralized resource that put into context 11 computational web applications for studying protein-protein interactions that facilitates its access and comprehension.

6. APPENDIX

Planas-Iglesias J, Bonet J, Marín-López MA, Feliu E, Gursoy A, Oliva B. Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence. In Cai W. Protein-Protein Interactions - Computational and Experimental Tools. Intech; 2012

Bonet J, Planas-Iglesias J, Garcia-Garcia J, Marín-López MA, Fernandez-Fuentes N, Oliva B. ArchDB 2014: structural classification of loops in proteins. Nucleic Acids Res. 2014 Jan;42(Database issue):D315-9. DOI: 10.1093/nar/gkt1189

Planas-Iglesias J, Bonet J, García-García J, Marín-López MA, Feliu E, Oliva B. Understanding Protein–Protein Interactions Using Local Structural Features. J Mol Biol. 2013 Apr 12;425(7):1210–24. DOI: 10.1016/j.jmb.2013.01.014

Planas-Iglesias J, Marin-Lopez MA, Bonet J, Garcia-Garcia J, Oliva B. iLoops: a protein–protein interaction prediction server based on structural features. Bioinformatics. 2013 Sep 15;29(18):2360–2. DOI: 10.1093/bioinformatics/btt401

Sieberts SK, Zhu F, García-García J, Stahl E, Pratap A, Pandey G, et al. Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. Nat Commun. 2016 Aug 23;7:12460. DOI: 10.1038/ncomms12460

7. REFERENCES

References

1. Crick F. Central dogma of molecular biology. Nature. 1970 Aug 8;227(5258):561–3.
2. Li G-W, Xie XS. Central dogma at the single-molecule level in living cells. Nature. 2011 Jul 21;475(7356):308–15.
3. Garcia-Garcia J, Bonet J, Guney E, Fornes O, Planas J, Oliva B. Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details. Mol Inform. 2012 May;31(5):342–62.
4. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. 2013 Jan;10(1):47–53.
5. Gutteridge A, Thornton JM. Understanding nature's catalytic toolkit. Trends Biochem Sci. 2005 Nov;30(11):622–9.
6. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A. 1951 Apr;37(4):205–11.
7. Eisenberg D. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. Proc Natl Acad Sci U S A. 2003 Sep 30;100(20):11207–10.
8. Bonet J, Planas-Iglesias J, Garcia-Garcia J, Marín-López MA, Fernandez-Fuentes N, Oliva B. ArchDB 2014: structural classification of loops in proteins. Nucleic Acids Res. 2014 Jan;42(Database issue):D315–9.
9. Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci U S A. 1973 Mar;70(3):697–701.
10. Sprinzak E, Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol. 2001 Aug 24;311(4):681–92.
11. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci. 2002 Oct;27(10):527–33.
12. Bonet J, Caltabiano G, Khan AK, Johnston MA, Corbí C, Gómez A, et al. The role of residue stability in transient protein-protein interactions involved in enzymatic phosphate hydrolysis. A computational study. Proteins. 2006 Apr 1;63(1):65–77.
13. Crowfoot D. X-ray crystallographic studies of compounds of biochemical interest. Annu Rev Biochem. 1948;17:115–46.
14. Wüthrich K. The way to NMR structures of proteins. Nat Struct Biol. 2001 Nov;8(11):923–5.
15. Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank archive as an open data resource. J Comput Aided Mol Des. 2014 Oct;28(10):1009–14.
16. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D226–9.
17. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, et al. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res. 2013 Jan;41(Database issue):D490–8.
18. Mosca R, Céol A, Stein A, Olivella R, Aloy P. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2014 Jan;42(Database issue):D374–9.
19. Ochi T, Bolanos-Garcia VM, Stojanoff V, Moreno A. Perspectives on protein crystallisation. Prog Biophys Mol Biol. 2009 Nov;101(1-3):56–63.
20. Bonet J, Segura J, Planas-Iglesias J, Oliva B, Fernandez-Fuentes N. Frag'r'Us: knowledge-based sampling of protein backbone conformations for de novo structure-based protein design. Bioinforma Oxf Engl. 2014 Jul 1;30(13):1935–6.
21. Braun W, Wider G, Lee KH, Wüthrich K. Conformation of glucagon in a lipid-water interphase by 1H nuclear magnetic resonance. J Mol Biol. 1983 Oct 5;169(4):921–48.
22. Frank J. Single-particle imaging of macromolecules by cryo-electron microscopy. Annu Rev Biophys Biomol Struct. 2002;31:303–19.
23. Neylon C. Small angle neutron and X-ray scattering in structural biology: recent examples from the literature. Eur Biophys J EBJ. 2008 Jun;37(5):531–41.
24. Judge PJ, Watts A. Recent contributions from solid-state NMR to the understanding of membrane protein structure and function. Curr Opin Chem Biol. 2011 Oct;15(5):690–5.
25. Maimon T, Elad N, Dahan I, Medalia O. The human nuclear pore complex as revealed by cryo-electron tomography. Struct Lond Engl 1993. 2012 Jun 6;20(6):998–1006.
26. Bernadó P, Svergun DI. Analysis of intrinsically disordered proteins by small-angle X-ray scattering. Methods Mol Biol Clifton NJ. 2012;896:107–22.
27. Krepkiy D, Mihailescu M, Freites JA, Schow EV, Worcester DL, Gawrisch K, et al. Structure and hydration of membranes embedded with voltage-sensing domains. Nature. 2009 Nov 26;462(7272):473–9.
28. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008 Oct;26(10):1135–45.
29. Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct. 1996;25:113–36.
30. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999 Feb;12(2):85–94.

31. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402.
32. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011 Oct;7(10):e1002195.
33. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinforma Oxf Engl. 2007 Nov 1;23(21):2947–8.
34. Magis C, Taly J-F, Bussotti G, Chang J-M, Di Tommaso P, Erb I, et al. T-Coffee: Tree-based consistency objective function for alignment evaluation. Methods Mol Biol Clifton NJ. 2014;1079:117–29.
35. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol. 2003;374:461–91.
36. Melo F, Feytmans E. Assessing protein structures with a non-local atomic interaction energy. J Mol Biol. 1998 Apr 17;277(5):1141–52.
37. Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification and classification of protein fold families. Protein Eng. 1993 Jul;6(5):485–500.
38. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol. 1999 Apr 9;287(4):797–815.
39. Kelley LA, Sternberg MJE. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc. 2009;4(3):363–71.
40. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. 2007;5:17.
41. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545–74.
42. Aloy P, Russell RB. Ten thousand interactions for the molecular biologist. Nat Biotechnol. 2004 Oct;22(10):1317–21.
43. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011 Jan;12(1):56–68.
44. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, et al. Estimating the size of the human interactome. Proc Natl Acad Sci U S A. 2008 May 13;105(19):6959–64.
45. Snider J, Kotlyar M, Saraon P, Yao Z, Jurisica I, Stagljar I. Fundamentals of protein interaction network mapping. Mol Syst Biol. 2015 Dec;11(12):848.
46. Michnick SW. Exploring protein interactions by interaction-induced folding of proteins from complementary peptide fragments. Curr Opin Struct Biol. 2001 Aug;11(4):472–7.
47. Ratushny V, Golemis E. Resolving the network of cell signaling pathways using the evolving yeast two-hybrid system. BioTechniques. 2008 Apr;44(5):655–62.
48. Day RN, Periasamy A, Schaufele F. Fluorescence resonance energy transfer microscopy of localized protein interactions in the living cell nucleus. Methods San Diego Calif. 2001 Sep;25(1):4–18.
49. Xu Y, Piston DW, Johnson CH. A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins. Proc Natl Acad Sci U S A. 1999 Jan 5;96(1):151–6.
50. Koos B, Andersson L, Clausson C-M, Grannas K, Klaesson A, Cane G, et al. Analysis of protein interactions in situ by proximity ligation assays. Curr Top Microbiol Immunol. 2014;377:111–26.
51. Dunham WH, Mullin M, Gingras A-C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. Proteomics. 2012 May;12(10):1576–90.
52. Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? J Mol Biol. 2003 Apr 11;327(5):919–23.
53. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 1998 Sep;23(9):324–8.
54. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature. 1999 Nov 4;402(6757):86–90.
55. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999 Apr 13;96(8):4285–8.
56. Roy S, Martinez D, Platero H, Lane T, Werner-Washburne M. Exploiting amino acid composition for predicting protein-protein interactions. PloS One. 2009;4(11):e7813.
57. Nanni L, Lumini A. An ensemble of K-local hyperplanes for predicting protein-protein interactions. Bioinforma Oxf Engl. 2006 May 15;22(10):1207–10.
58. Garcia-Garcia J, Schleker S, Klein-Seetharaman J, Oliva B. BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W147–51.

59. Planas-Iglesias J, Bonet J, García-García J, Marín-López MA, Feliu E, Oliva B. Understanding protein-protein interactions using local structural features. J Mol Biol. 2013 Apr 12;425(7):1210–24.

60. Planas-Iglesias J, Marin-Lopez MA, Bonet J, Garcia-Garcia J, Oliva B. iLoops: a protein-protein interaction prediction server based on structural features. Bioinforma Oxf Engl. 2013 Sep 15;29(18):2360–2.

61. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol. 1997 Aug 29;271(4):511–23.

62. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A. PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W285–9.

63. Aloy P, Russell RB. InterPreTS: protein interaction prediction through tertiary structure. Bioinforma Oxf Engl. 2003 Jan;19(1):161–2.

64. Goldberg DS, Roth FP. Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci U S A. 2003 Apr 15;100(8):4372–6.

65. Liu G, Li J, Wong L. Assessing and predicting protein interactions using both local and global network topological metrics. Genome Inform Int Conf Genome Inform. 2008;21:138–49.

66. Aloy P, Russell RB. Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol. 2006 Mar;7(3):188–97.

67. Planas-Iglesias J, Bonet J, Marín-López MA, Feliu E, Gursoy A, Oliva B. Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence. In: InTech, In Protein-Protein Interactions - Computational and Experimental Tools. 2012.

68. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs." Genome Res. 2001 Dec;11(12):2120–6.

69. Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O. Architectures and functional coverage of protein-protein interfaces. J Mol Biol. 2008 Sep 5;381(3):785–802.

70. Keskin O, Tsai C-J, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. Protein Sci Publ Protein Soc. 2004 Apr;13(4):1043–55.

71. Tuncbag N, Keskin O, Nussinov R, Gursoy A. Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement. Proteins. 2012 Apr;80(4):1239–49.

72. Huang S-Y. Exploring the potential of global protein-protein docking: an overview and critical assessment of current programs for automatic ab initio docking. Drug Discov Today. 2015 Aug;20(8):969–77.

73. Segura J, Marín-López MA, Jones PF, Oliva B, Fernandez-Fuentes N. VORFFIP-driven dock: V-D2OCK, a fast and accurate protein docking strategy. PloS One. 2015;10(3):e0118107.

74. Moal IH, Torchala M, Bates PA, Fernández-Recio J. The scoring of poses in protein-protein docking: current capabilities and future directions. BMC Bioinformatics. 2013;14:286.

75. Valdar WS, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins. 2001 Jan 1;42(1):108–24.

76. Hoskins J, Lovell S, Blundell TL. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. Protein Sci Publ Protein Soc. 2006 May;15(5):1017–29.

77. Ofran Y, Rost B. ISIS: interaction sites identified from sequence. Bioinforma Oxf Engl. 2007 Jan 15;23(2):e13–6.

78. Segura J, Jones PF, Fernandez-Fuentes N. Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. BMC Bioinformatics. 2011;12:352.

79. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins. 2007 Feb 15;66(3):630–45.

80. Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. Nucleic Acids Res. 2006;34(13):3698–707.

81. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol. 2004 Apr 16;338(1):181–99.

82. Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R. PIER: protein interface recognition for structural proteomics. Proteins. 2007 May 1;67(2):400–17.

83. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins. 1994 Apr;18(4):309–17.

84. Henschel A, Winter C, Kim WK, Schroeder M. Using structural motif descriptors for sequence-based binding site prediction. BMC Bioinformatics. 2007;8 Suppl 4:S5.

85. Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B. Characterization of protein hubs by inferring interacting motifs from protein interactions. PLoS Comput Biol. 2007 Sep;3(9):1761–71.

86. Wodak SJ, Janin J. Computer analysis of protein-protein interaction. J Mol Biol. 1978 Sep 15;124(2):323–42.

87. Vajda S, Kozakov D. Convergence and combination of methods in protein-protein docking. Curr Opin Struct Biol. 2009 Apr;19(2):164–70.

88. Huang S-Y. Search strategies and evaluation in protein-protein docking: principles, advances and challenges. Drug Discov Today. 2014 Aug;19(8):1081–96.

89. Jackson RM, Gabb HA, Sternberg MJ. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol. 1998 Feb 13;276(1):265–85.

90. Li L, Chen R, Weng Z. RDOCK: refinement of rigid-body protein docking predictions. Proteins. 2003 Nov 15;53(3):693–707.

91. Andrusier N, Nussinov R, Wolfson HJ. FireDock: fast interaction refinement in molecular docking. Proteins. 2007 Oct 1;69(1):139–59.

92. Mashiach E, Nussinov R, Wolfson HJ. FiberDock: Flexible induced-fit backbone refinement in molecular docking. Proteins. 2010 May 1;78(6):1503–19.

93. Mashiach E, Nussinov R, Wolfson HJ. FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W457–61.

94. Venkatraman V, Ritchie DW. Flexible protein docking refinement using pose-dependent normal mode analysis. Proteins. 2012 Aug;80(9):2262–74.

95. Pierce B, Weng Z. ZRANK: reranking protein docking predictions with an optimized energy function. Proteins. 2007 Jun 1;67(4):1078–86.

96. Feliu E, Aloy P, Oliva B. On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. Protein Sci Publ Protein Soc. 2011 Mar;20(3):529–41.

97. Moont G, Gabb HA, Sternberg MJ. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins. 1999 May 15;35(3):364–73.

98. Cheng TM-K, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins. 2007 Aug 1;68(2):503–15.

99. Liang S, Liu S, Zhang C, Zhou Y. A simple reference state makes a significant improvement in near-native selections from structurally refined docking decoys. Proteins. 2007 Nov 1;69(2):244–53.

100. Chuang G-Y, Kozakov D, Brenke R, Comeau SR, Vajda S. DARS (Decoys As the Reference State) potentials for protein-protein docking. Biophys J. 2008 Nov 1;95(9):4217–27.

101. Liu S, Vakser IA. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. BMC Bioinformatics. 2011;12:280.

102. Pons C, Talavera D, de la Cruz X, Orozco M, Fernandez-Recio J. Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking. J Chem Inf Model. 2011 Feb 28;51(2):370–7.

103. Ravikant DVS, Elber R. PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. Proteins. 2010 Feb 1;78(2):400–19.

104. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol. 1997 Sep 12;272(1):106–20.

105. Vakser IA. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. Proteins. 1997;Suppl 1:226–30.

106. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A. 1992 Mar 15;89(6):2195–9.

107. Roberts VA, Thompson EE, Pique ME, Perez MS, Ten Eyck LF. DOT2: Macromolecular docking with improved biophysical models. J Comput Chem. 2013 Jul 30;34(20):1743–58.

108. Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PloS One. 2011;6(9):e24657.

109. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. Proteins. 2006 Nov 1;65(2):392–406.

110. Zhang C, Lai L. SDOCK: a global protein-protein docking program using stepwise force-field potentials. J Comput Chem. 2011 Sep;32(12):2598–612.

111. Bajaj C, Chowdhury R, Siddavanahalli V. F2Dock: fast Fourier protein-protein docking. IEEEACM Trans Comput Biol Bioinforma IEEE ACM. 2011 Mar;8(1):45–58.

112. Li L, Guo D, Huang Y, Liu S, Xiao Y. ASPDock: protein-protein docking algorithm using atomic solvation parameters model. BMC Bioinformatics. 2011;12:36.

113. Ritchie DW, Kemp GJ. Protein docking using spherical polar Fourier correlations. Proteins. 2000 May 1;39(2):178–94.

114. Garzon JI, Lopéz-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J, et al. FRODOCK: a new approach for fast rotational protein-protein docking. Bioinforma Oxf Engl. 2009 Oct 1;25(19):2544–51.

115. Gardiner EJ, Willett P, Artymiuk PJ. Protein docking using a genetic algorithm. Proteins. 2001 Jul 1;44(1):44–56.

116. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W363–7.

117. Venkatraman V, Yang YD, Sael L, Kihara D. Protein-protein docking using region-based 3D Zernike descriptors. BMC Bioinformatics. 2009;10:407.

118. Zacharias M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci Publ Protein Soc. 2003 Jun;12(6):1271–82.

119. Chaudhury S, Berrondo M, Weitzner BD, Muthu P, Bergman H, Gray JJ. Benchmarking and analysis of protein docking performance in Rosetta v3.2. PloS One. 2011;6(8):e22477.

120. Fernández-Recio J, Totrov M, Abagyan R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. Proteins. 2003 Jul 1;52(1):113–7.

121. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc. 2003 Feb 19;125(7):1731–7.

122. Moal IH, Bates PA. SwarmDock and the use of normal modes in protein-protein docking. Int J Mol Sci. 2010;11(10):3623–48.

123. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem. 1998 Nov 15;19(14):1639–62.

124. Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, et al. An experimentally derived confidence score for binary protein-protein interactions. Nat Methods. 2009 Jan;6(1):91–7.

125. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis A-R, Simonis N, et al. Literature-curated protein interaction datasets. Nat Methods. 2009 Jan;6(1):39–46.

126. Garcia-Garcia J, Guney E, Aragues R, Planas-Iglesias J, Oliva B. Biana: a software framework for compiling biological interactions and analyzing networks. BMC Bioinformatics. 2010;11:56.

127. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011 Jan;39(Database issue):D561–8.

128. Bader GD, Betel D, Hogue CWV. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 2003 Jan 1;31(1):248–50.

129. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, et al. The IntAct molecular interaction database in 2010. Nucleic Acids Res. 2010 Jan;38(Database issue):D525–31.

130. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D449–51.

131. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2011 Jan;39(Database issue):D698–704.

132. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009 Jan;37(Database issue):D767–72.

133. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012 Jan;40(Database issue):D857–61.

134. Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes H-W, et al. MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D436–41.

135. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, et al. The MIPS mammalian protein-protein interaction database. Bioinforma Oxf Engl. 2005 Mar;21(6):832–4.

136. Han K, Park B, Kim H, Hong J, Park J. HPID: the Human Protein Interaction Database. Bioinforma Oxf Engl. 2004 Oct 12;20(15):2466–70.

137. McDowall MD, Scott MS, Barton GJ. PIPs: human protein-protein interaction prediction database. Nucleic Acids Res. 2009 Jan;37(Database issue):D651–6.

138. Brown KR, Jurisica I. Online predicted human interaction database. Bioinforma Oxf Engl. 2005 May 1;21(9):2076–82.

139. Huang T-W, Tien A-C, Huang W-S, Lee Y-CG, Peng C-L, Tseng H-H, et al. POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. Bioinforma Oxf Engl. 2004 Nov 22;20(17):3273–6.

140. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. Nucleic Acids Res. 2011 Jan;39(Database issue):D730–5.

141. Gong S, Yoon G, Jang I, Bolser D, Dafas P, Schroeder M, et al. PSIbase: a database of Protein Structural Interactome map (PSIMAP). Bioinforma Oxf Engl. 2005 May 15;21(10):2541–3.

142. Segura J, Fernandez-Fuentes N. PCRPi-DB: a database of computationally annotated hot spots in protein interfaces. Nucleic Acids Res. 2011 Jan;39(Database issue):D755–60.

143. Cukuroglu E, Gursoy A, Keskin O. HotRegion: a database of predicted hot spot clusters. Nucleic Acids Res. 2012 Jan;40(Database issue):D829–33.

144. Guney E, Tuncbag N, Keskin O, Gursoy A. HotSprint: database of computational hot spots in protein interfaces. Nucleic Acids Res. 2008 Jan;36(Database issue):D662–6.

145. Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinforma Oxf Engl. 2001 Mar;17(3):284–5.

146. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AMJJ, et al. A structure-based benchmark for protein-protein binding affinity. Protein Sci Publ Protein Soc. 2011 Mar;20(3):482–91.

147. Finn RD, Marshall M, Bateman A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinforma Oxf Engl. 2005 Feb 1;21(3):410–2.

148. Stein A, Céol A, Aloy P. 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2011 Jan;39(Database issue):D718–23.

149. Winter C, Henschel A, Kim WK, Schroeder M. SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D310–4.

150. Teyra J, Doms A, Schroeder M, Pisabarro MT. SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. BMC Bioinformatics. 2006;7:104.

151. Davis FP, Sali A. PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinforma Oxf Engl. 2005 May 1;21(9):1901–7.

152. Gong S, Park C, Choi H, Ko J, Jang I, Lee J, et al. A protein domain interaction interface database: InterPare. BMC Bioinformatics. 2005;6:207.

153. Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res. 2001 Jan 1;29(1):221–2.

154. Kundrotas PJ, Alexov E. PROTCOM: searchable database of protein complexes enhanced with domain-domain structures. Nucleic Acids Res. 2007 Jan;35(Database issue):D575–9.

155. Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. Proteins. 2010 Nov 15;78(15):3111–4.

156. Faure G, Andreani J, Guerois R. InterEvol database: exploring the structure and evolution of protein complex interfaces. Nucleic Acids Res. 2012 Jan;40(Database issue):D847–56.

157. Emil Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. Ber Dtsch Chem Ges. 1894;27:2984–93.

158. Koshland DE. Enzyme flexibility and enzyme action. J Cell Comp Physiol. 1959 Dec;54:245–58.

159. Brocklehurst K, Willenbrock SJ, Salih E. Effects of conformational selectivity and of overlapping kinetically influential ionizations on the characteristics of pH-dependent enzyme kinetics. Implications of free-enzyme pKa variability in reactions of papain for its catalytic mechanism. Biochem J. 1983 Jun 1;211(3):701–8.

160. Okazaki K-I, Takada S. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. Proc Natl Acad Sci U S A. 2008 Aug 12;105(32):11182–7.

161. Hammes GG, Chang Y-C, Oas TG. Conformational selection or induced fit: a flux description of reaction mechanism. Proc Natl Acad Sci U S A. 2009 Aug 18;106(33):13737–41.

162. Wlodarski T, Zagrovic B. Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. Proc Natl Acad Sci U S A. 2009 Nov 17;106(46):19346–51.

163. Kastritis PL, Bonvin AMJJ. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. J R Soc Interface R Soc. 2013 Feb;10(79):20120835.

164. Northrup SH, Erickson HP. Kinetics of protein-protein association explained by Brownian dynamics computer simulation. Proc Natl Acad Sci U S A. 1992 Apr 15;89(8):3338–42.

165. Smoluchowski M. Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Losungen. Z Phys Chem. 1917;92:129–68.

166. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci Publ Protein Soc. 1997 Jan;6(1):53–64.

167. Berg OG, von Hippel PH. Diffusion-controlled macromolecular interactions. Annu Rev Biophys Biophys Chem. 1985;14:131–60.

168. Sommer J, Jonah C, Fukuda R, Bersohn R. Production and subsequent second-order decomposition of protein disulfide anions lengthy collisions between proteins. J Mol Biol. 1982 Aug 25;159(4):721–44.

169. Berg OG. Orientation constraints in diffusion-limited macromolecular association. The role of surface diffusion as a rate-enhancing mechanism. Biophys J. 1985 Jan;47(1):1–14.

170. Ubbink M. The courtship of proteins: understanding the encounter complex. FEBS Lett. 2009 Apr 2;583(7):1060–6.

171. Schreiber G, Haran G, Zhou H-X. Fundamental aspects of protein-protein association kinetics. Chem Rev. 2009 Mar 11;109(3):839–60.

172. Zhou HX. Brownian dynamics study of the influences of electrostatic interaction and diffusion on protein-protein association kinetics. Biophys J. 1993 Jun;64(6):1711–26.

173. Alsallaq R, Zhou H-X. Electrostatic Rate Enhancement and Transient Complex of Protein-Protein Association. Proteins. 2008 Apr;71(1):320–35.

174. Selzer T, Albeck S, Schreiber G. Rational design of faster associating and tighter binding protein complexes. Nat Struct Biol. 2000 Jul;7(7):537–41.

175. Tang C, Iwahara J, Clore GM. Visualization of transient encounter complexes in protein-protein association. Nature. 2006 Nov 16;444(7117):383–6.

176. Fawzi NL, Doucleff M, Suh J-Y, Clore GM. Mechanistic details of a protein-protein association pathway revealed by paramagnetic relaxation enhancement titration measurements. Proc Natl Acad Sci U S A. 2010 Jan 26;107(4):1379–84.

177. Zhou H-X, Bates PA. Modeling protein association mechanisms and kinetics. Curr Opin Struct Biol. 2013 Dec;23(6):887–93.

178. Harel M, Cohen M, Schreiber G. On the dynamic nature of the transition state for protein-protein association as determined by double-mutant cycle analysis and simulation. J Mol Biol. 2007 Aug 3;371(1):180–96.

179. Shaul Y, Schreiber G. Exploring the charge space of protein-protein association: a proteomic study. Proteins. 2005 Aug 15;60(3):341–52.

180. Camacho CJ, Kimura SR, DeLisi C, Vajda S. Kinetics of desolvation-mediated protein-protein binding. Biophys J. 2000 Mar;78(3):1094–105.

181. Vidal M, Cusick ME, Barabási A-L. Interactome networks and human disease. Cell. 2011 Mar 18;144(6):986–98.

182. Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, et al. Widespread macromolecular interaction perturbations in human genetic disorders. Cell. 2015 Apr 23;161(3):647–60.

183. Chothia C, Janin J. Principles of protein-protein recognition. Nature. 1975 Aug 28;256(5520):705–8.

184. Zhou HX. Disparate ionic-strength dependencies of on and off rates in protein-protein association. Biopolymers. 2001 Nov;59(6):427–33.

185. Vuignier K, Schappler J, Veuthey J-L, Carrupt P-A, Martel S. Drug-protein binding: a critical review of analytical tools. Anal Bioanal Chem. 2010 Sep;398(1):53–66.

186. Wilkinson KD. Quantitative analysis of protein-protein interactions. Methods Mol Biol Clifton NJ. 2004;261:15–32.

187. Ladbury JE, Chowdhry BZ. Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. Chem Biol. 1996 Oct;3(10):791–801.

188. Willander M, Al-Hilli S. Analysis of biomolecules using surface plasmons. Methods Mol Biol Clifton NJ. 2009;544:201–29.

189. Masi A, Cicchi R, Carloni A, Pavone FS, Arcangeli A. Optical methods in the study of protein-protein interactions. Adv Exp Med Biol. 2010;674:33–42.

190. Horton N, Lewis M. Calculation of the free energy of association for protein complexes. Protein Sci Publ Protein Soc. 1992 Jan;1(1):169–81.

191. Ma XH, Wang CX, Li CH, Chen WZ. A fast empirical approach to binding free energy calculations based on protein interface information. Protein Eng. 2002 Aug;15(8):677–81.

192. Audie J, Scarlata S. A novel empirical free energy function that explains and predicts protein-protein binding affinities. Biophys Chem. 2007 Sep;129(2-3):198–211.

193. Su Y, Zhou A, Xia X, Li W, Sun Z. Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. Protein Sci Publ Protein Soc. 2009 Dec;18(12):2550–8.

194. Bai H, Yang K, Yu D, Zhang C, Chen F, Lai L. Predicting kinetic constants of protein-protein interactions based on structural properties. Proteins. 2011 Mar;79(3):720–34.

195. Moal IH, Agius R, Bates PA. Protein-protein binding affinity prediction on a diverse set of structures. Bioinforma Oxf Engl. 2011 Nov 1;27(21):3002–9.

196. Tian F, Lv Y, Yang L. Structure-based prediction of protein-protein binding affinity with consideration of allosteric effect. Amino Acids. 2012 Aug;43(2):531–43.

197. Erijman A, Rosenthal E, Shifman JM. How structure defines affinity in protein-protein interactions. PloS One. 2014;9(10):e110085.

198. Vangone A, Bonvin AMJJ. Contacts-based prediction of binding affinity in protein-protein complexes. eLife. 2015;4:e07454.

199. Marillet S, Boudinot P, Cazals F. High-resolution crystal structures leverage protein binding affinity predictions. Proteins. 2016 Jan;84(1):9–20.

200. Kastritis PL, Rodrigues JPGLM, Folkers GE, Boelens R, Bonvin AMJJ. Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. J Mol Biol. 2014 Jul 15;426(14):2632–52.

201. Kastritis PL, Bonvin AMJJ. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res. 2010 May 7;9(5):2216–25.

202. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998 Jul 3;280(1):1–9.

203. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) - progress and new directions in Round XI. Proteins. 2016 May 12;

204. Lensink MF, Velankar S, Kryshtafovych A, Huang S-Y, Schneidman-Duhovny D, Sali A, et al. Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. Proteins. 2016 Apr 28;

205. Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. Proteins. 2013 Dec;81(12):2082–95.

206. Bonvin A. Coming to peace with protein complexes? 5th CAPRI evaluation meeting, April 17-19th 2013--Utrecht. Proteins. 2013 Dec;81(12):2073–4.

207. Poglayen D, Marín-López MA, Bonet J, Fornes O, Garcia-Garcia J, Planas-Iglesias J, et al. InteractoMIX: a suite of computational tools to exploit interactomes in biological and clinical research. Biochem Soc Trans. 2016 Jun 15;44(3):917–24.