We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

10

Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence

J. Planas-Iglesias, J. Bonet,
M.A. Marín-López, E. Feliu, A. Gursoy and B. Oliva
Structural Bioinformatics Lab. Universitat Pompeu Fabra, Catalunya
Spain

1. Introduction

Although the regulatory role of non-coding nucleic acids is currently being unraveled, the role of proteins is still a major issue as they mediate most biological functions. Thus, understanding how proteins fulfill their intricate functions is one of the most relevant current challenges in biology. It is well known that a protein's function is determined by its three-dimensional (3D) structure known as tertiary structure, which in turn is mainly dictated by its sequence (Thornton and cols. reviewed this issue in detail in (Watson et al., 2005)). Despite the exponential increase of available sequences and 3D structures, the number of sequences highly exceeds that of 3D structures. This difference in numbers is proportional to the disparity of the costs for experimentally obtaining either the sequence or the structure of a protein. Therefore, covering the gap between sequence and structure becomes a compelling requirement to achieve a molecular understanding of the protein function. Theoretical methods can help to bridge this gap by inferring the 3D structure from the sequence. These methods are classified into three different groups: comparative modeling, fold recognition and new fold or *ab initio* methods.

Besides the tertiary structure of a protein, other contextual factors may modulate its function. Among these, the ability of the proteins to interact with others and the particular partners with which they form complexes are one of the most important. This is because proteins rarely act alone; they rather constitute a mingled network of physical interactions, some times to form large macro-complexes and sometimes to produce transient interactions. In this context, understanding the function of a protein implies to recognize its partners and to grasp how they associate, even at the atomic level. The structure of these complexes is known as quaternary structure. To this end, computational techniques have been developed to dock one protein onto another (Janin, 2010; Vajda and Kozakov, 2009), and can help to infer 3D structure of a protein from the knowledge of the protein interactions (Fornes et al., 2009) and vice-versa (Stein et al., 2005). Furthermore, the combined use of data from multiple resources allows us to obtain an accurate model of large molecular complexes such as nucleopore (Alber et al., 2007).

There are two strategies for modeling the interaction between two proteins from sequence data. The first one is to model the unbound interactors and to dock them into the final

complex (i.e. solving first the tertiary structure of the proteins and afterwards the quaternary). The second is to model the interacting pair or complex from the scratch, using as template the structural knowledge of an available homologous interacting pair (interolog, (Matthews et al., 2001)). When the template is not available the strategy can only play with the docking of the unbound partners. Figure 1 summarizes these possibilities. Here, we will cover these strategies and methods to infer and assess the 3D structure of binary protein interactions, and we will review the existing techniques to model large cellular macro-complexes.

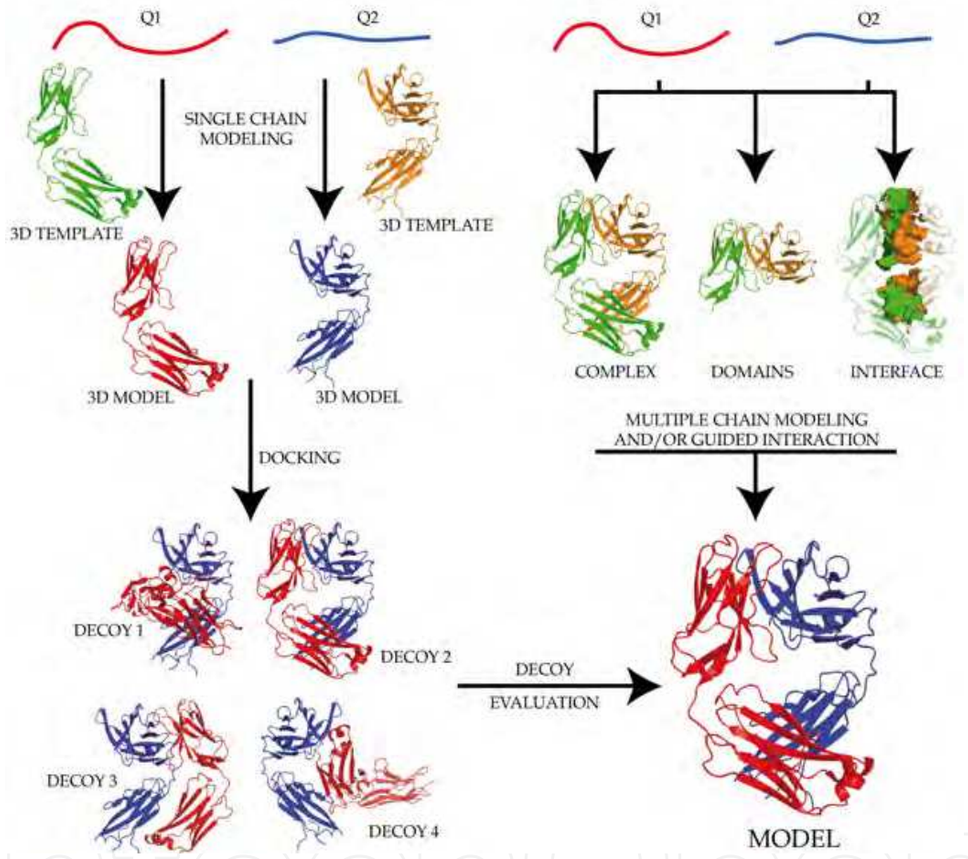


Fig. 1. Different strategies for modeling a protein interaction: The 3D structure of a binary protein interaction can be inferred by modeling individual interacting partners apart and subsequently docking them (left side) or modeling the interaction with one template, taking advantage of the available information of homologous complexes. Templates can be obtained from structural resources of information containing the full complex, a partial complex (in general formed by interacting domains) or with only the interacting interface (right side).

2. Modeling the tertiary structure of proteins

In order to obtain a complete model of a protein interaction, the interacting partners can be modeled separately and then docked into a functional complex. The first step of this approach is to obtain the 3D structure of each of the interacting partners. Comparative modeling, fold recognition, and *ab initio* (for new folds) are computational methods that

may overcome the lack of experimental structural data. Models obtained using these approaches may be further assessed, in order to ensure that the inferred 3D structure contains no errors (see section 4). In case of persistence, such errors would damper the deduction of further biological conclusions such as the mode the modeled protein can interact with others. Figure 2 summarizes the different steps and strategies that can be exploited to achieve these objectives.

2.1 Homology modeling

Homology or comparative modeling techniques are those devoted to the prediction and construction of the 3D conformation of proteins. These methods are based on the assumption that structural features in proteins are more conserved than its sequences. Thus, two proteins with enough sequence similarity will fold in a similar way and share the same conformation in space. The process through which a tertiary structure is assigned to a given sequence is carried out in three steps, namely: template identification, template alignment, and model building. Finally, the produced model should be assessed (see section 4).

Template identification is the key step in the molecular modeling process. Templates are defined as the set of known structures used to build the tertiary structure of the query (target or problem protein). Known 3D data of proteins is stored in the Protein Data Bank (PDB) (Berman et al., 2000). Thus, the identification of the template refers to the process of identifying the structure of the PDB whose sequence is the closest homolog of the target. Such sequence homology search can be performed using sequence alignment tools like BLAST and PSI-BLAST (Altschul et al., 1997), or Hidden Markov Model (HMM) profile methods like HMMER (Eddy, 1998). While BLAST will reveal if there is any relatively close homolog to our query, PSI-BLAST and HMMER will also reveal the possibility of remote homologues. The homology threshold that can be used to define whether or not a template assignation is correct may be fuzzy. Those templates assigned with low percentage of identity, low homology, or in short parts of the sequence fall into what is known as the twilight zone (Rost, 1999). Some rules have been described to shed some light into that twilight region in order to better describe the viability of a template for a given query (Fornes et al., 2009).

Provided that a good template has been selected, the sequence alignment between the query and the template can be directly extracted or easily inferred (in case of the HMM) from the template search. Depending on specific requirements, the alignments can be redone with other sequence alignment methods such as CLUSTALW (Chenna et al., 2003) or T-COFFE (Notredame et al., 2000). Additionally, some methods optimize the sequence alignment through a genetic algorithm protocol that iterates the alignment, model building and model evaluation in order to obtain the best possible alignment (Fernandez-Fuentes et al., 2007) .

Model building is the process by which the three-dimensional data of the template(s) is applied on the query sequence. MODELLER is one of the most used and comprehensive pieces of modeling software (Sali et al., 1995). Provided the sequence alignment the modeling process is practically automatic. As many other modeling tools, it is based on satisfying a set of spatial constraints. Specifically it satisfies three spatial constraints being (1) homology-derived constraints, (2) stereochemical constraints such as bond angles, and (3) statistical preferences for dihedral angles and non-bonded interatomic distances.

Optionally, manually curated restraints from secondary structure packing to site-directed mutagenesis can be added to the modeling process.

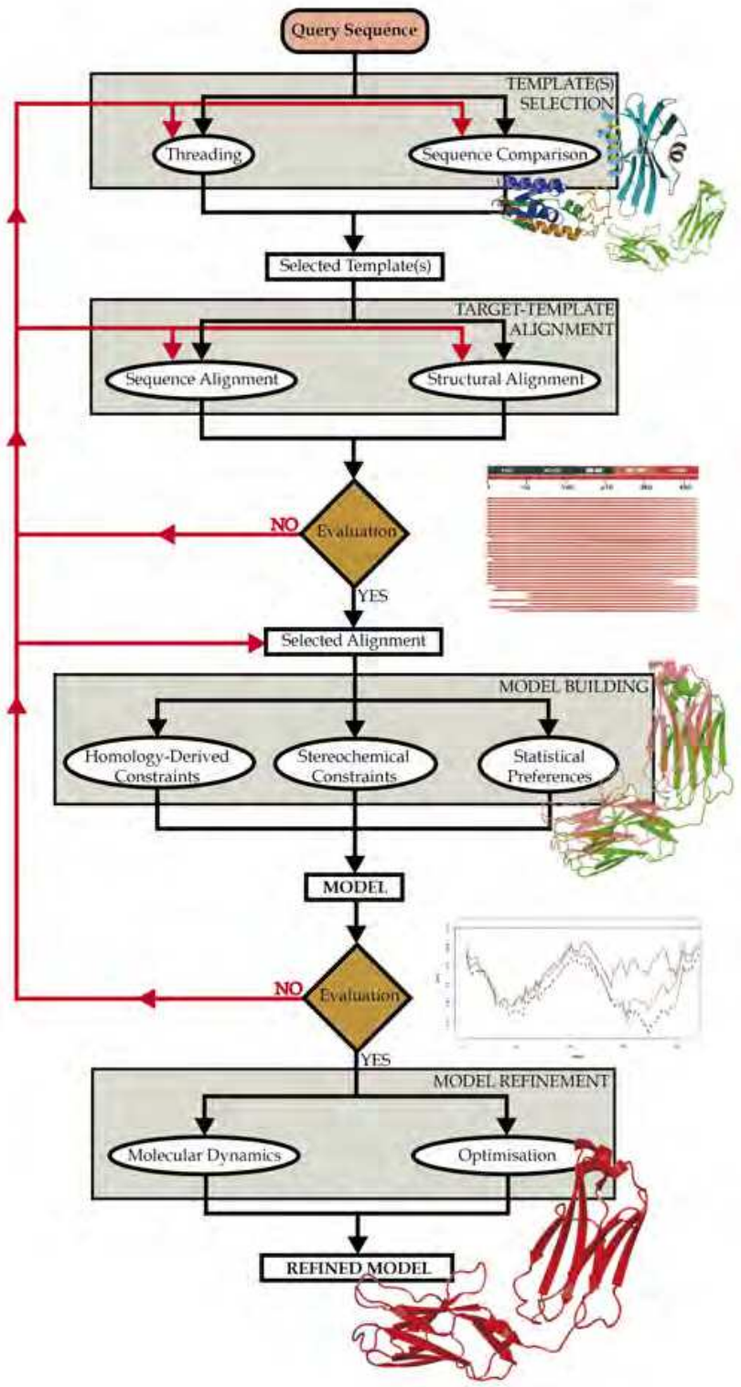


Fig. 2. Flowchart for single protein modeling: Scheme of the methods used for modeling, comprising template(s) selection, template-target alignment, model building, model evaluation, and model refinement steps.

2.2 Threading

In the same way as homology modeling, threading is a method to determine the tertiary structure of a protein based on the fairly small number of different folds contained in nature. The main difference resides in the fact that threading does not use specific protein 3D structures as templates but uses statistical knowledge extracted from all structures in PDB. Thus, threading is especially useful when no suitable template for the protein can be found. Basically, the prediction is made by aligning each amino acid of a given query sequence to a position in a set of structural templates. Once the optimal template is selected through this method the structural model is built according to the alignment between the query and the template. The threading process can be divided in four different steps: the template database construction, the scoring function, the threading alignment and the threading prediction. Among the most used programs on threading and fold prediction are GeneThreader (Jones, 1999) and Phyre (Kelley and Sternberg, 2009).

Gathering representative structures for the different folds avoiding redundancy creates the template database. This means extracting all the structures from PDB and picking one representative of each known fold, eliminating redundancy and sequence homology. The homology filtering is key to ensure that the predictions over the query sequence are not going to be biased because of the database composition.

Designing an optimal function to score the suitability of the templates for modeling the query protein is determinant. The scoring function should be based on the known relationships between structure and sequence. A good scoring function should contain as much information as possible such as pairwise potential, secondary structure compatibilities, environment fitness potential, and gap penalties. The accuracy of the alignment and the prediction will be directly related to the quality of the scoring function. During the threading alignment the query sequence is going to be tested against each given possible template. This part of the process is, by far, the most computationally costly as it takes into account pairwise contact potentials (see section 4.2) and cannot be substituted by the classic dynamic programming algorithm for sequence alignment. Finally the threading prediction uses the scoring function and all the provided alignments to select the better template and build the protein model by placing the backbone atoms of the query according to the position of their aligned counterparts in the template.

2.3 *Ab Initio* methods

Ab initio or *de novo* protein structure prediction tries to predict the tertiary structure of proteins directly from its sequence properties. The idea is that the structure of proteins can be determined without any explicit templates by means of applying the general principles that govern protein folding and the statistical tendencies of conformational features gathered from structural knowledge. Those predictions involve sampling the conformational space, which means that a large set of decoys (structural candidates) is likely to be generated. Scoring functions, either physics-based or knowledge based, are then used to select those decoys that can be identified as more native-like conformations. Optionally, high-resolution refinement is used to optimize those native-like structures. Few programs have been successful in this task, and among them the most flourishing are ROSETTA (Leaver-Fay et al., 2011) and TASSER (Chen and Skolnick, 2008).

3. Modeling the quaternary structure of proteins

As the structural data on protein complexes keep increasing steadily, using known protein complex structures has become an important approach for modeling protein interactions. This increase of structural knowledge of protein complexes (even if it is only partial) opens a new window of possibilities to infer the quaternary structure of proteins. However, for a large quantity of protein complexes this knowledge is still limited, and alternative techniques are required to infer their 3D structure. Docking methods surmount this lack of data providing predictions of the quaternary structure of the complex based on the physical, chemical, and biological known properties of protein complexes. New approaches have introduced the possibility to integrate different sources of experimental information, such as high-resolution electron-microscopy, SAXS, NMR, yeast-two-hybrid, and affinity purifications to extract restraints that can be applied to model the quaternary structure of macro-complexes (Alber et al., 2007).

3.1 Comparative modeling of protein binary complexes

Provided that a homolog structure of an interaction is known, homology modeling can be used to model a protein-protein interaction of interest. Two different approaches can be taken: (1) direct interaction modeling or (2) protein modeling and reorientation (see Figure 1).

When directly modeling a protein interaction, it has to be taken into account that both query proteins need to have a couple of acceptable templates that share the same crystal structure. If that is the case, MODELLER is able to directly model the protein-protein interaction. However, in not only each separate structure needs to be evaluated but also the interface created between them (see section 4.2).

An alternative is to model each protein separately and afterwards use a known interaction as a guide to reposition each structure in the way the interaction is supposed to be taking place. To do so it is required to perform a structural alignment between the model and the template for the interaction. That can be done with strictly devoted tools such as STAMP (Russell and Barton, 1992) or through a variety of protein structure graphical interfaces such as PYMOL (http://www.pymol.org). This approach should be selected if the resolution of the structure of the templates in the interaction is largely worse than the unbound templates. However, it has to be taken into account the need to introduce some structural flexibility produced to construct the interaction. Considering the principal motions and intrinsic fluctuations to accommodate the unbound structures (Dobbins et al., 2008) may help to this purpose. The final structure needs to be refined, and additional restraints are applied to keep the partners on the orientation defined by the template of the binary complex.

3.2 Modeling of protein binary complexes from partial structural information

The sequence and structural homology methods described in the previous section require global similarity (sequence or structure). However, recent research shows that the binding sites of proteins are somewhat more distinguishable from the rest of the protein surface. The binding site of two interacting proteins is called a protein-protein interface. If the structure of a protein complex is available, determining the interface is fairly simple. The interface can

be found either by finding contacting residues (distance based) or by calculating the accessible surface area of the residues. Since proteins interact through interfaces, physico-chemical properties of interfaces are important to study protein interactions. Statistical studies of known protein complexes have revealed general characteristics of interfaces. Interfaces in general have electrostatic and shape complementary. Compared to the rest of the protein surface, interfaces are found to be slightly more conserved (Caffrey et al., 2004). Depending on the interaction type, properties of interfaces display variation. Homo-oligomeric complexes have more hydrophobic and larger interfaces than the hetero complexes. Homo-oligomers are usually permanent and their interfaces resemble interior of globular proteins. Transient interactions, on the other hand, are mediated by smaller interfaces (less than 1500 Å^2) and have more polar and charged amino acids than the interfaces of permanent interactions (Nooren and Thornton, 2003). The small surface-area of transient interfaces are partly due to requirement of individual partners of the interaction to fold independently and to be soluble. The secondary structure content of interfaces shows differences between permanent and transient interfaces as well. For example, turns are observed more frequently in non-obligatory interfaces since flexibility is required to repeatedly associate/disassociate (De et al., 2005). Even within an interface, the properties and organization of residues are not uniform. The interface area may be dissected into regions where a set of buried residues forming a core region is surrounded by a rim of residues that are partially solvent accessible. The composition of residues are distinct between these two regions (Guharoy and Chakrabarti, 2005). Alanine scanning mutagenesis of interface residues has also revealed that some residues contribute more to the binding energy (Clackson and Wells, 1995). These areas, called hot spots, are particularly enriched in Trp, Tyr, and Arg residues and are structurally conserved, which can be used to differentiate binding sites from the rest of the surface (Ma et al., 2003).

All these characteristics can be used to identify binding sites of proteins either from sequence (Ofran and Rost, 2007) or from unbound structures (Neuvirth et al., 2004), and potentially for modeling protein interactions. Therefore, a systematic collection and categorization of protein interfaces play important role. Several databases of interfaces have been compiled along this direction, including PiBASE (Davis and Sali, 2005), SCOWLP (Teyra et al., 2006), SCOPPI (Winter et al., 2006) and PRINT (Tuncbag et al., 2008). These databases, in general, present interfaces extracted form known protein complexes together with features of the interfaces such as change in accessible surface area, conservation, or residue composition (reviewed in (Tuncbag et al., 2009)). PRINT database presents all interfaces from PDB (as of 2006) clustered by structural similarity where each cluster represents a different interface architecture. Some interface architectures are observed to be more favorable and reused frequently. These interface architectures are found to be similar to domain folds, consistent with earlier studies indicating that the folding and binding are similar processes (Tsai et al., 1997).

The analysis of protein interactions and interfaces has suggested that the number of possible interfaces is much smaller than the possible number of protein interactions (Aloy and Russell, 2004; Tuncbag et al., 2008). In addition, interfaces are observed to be reused in different protein interactions that are not globally similar (the same interface used by proteins with different fold architectures) (Keskin et al., 2004). This information can be used to overcome the global similarity of requirement of the homology based modeling methods.

That is, modeling protein interactions using only the similarity of protein interfaces. PRISM (Aytuna et al., 2005) is one of the first approaches that has used interface similarity along this direction. It was originally developed to predict protein interactions between proteins (target set) from a set of known protein interfaces (template set). If the two complementary sides of a template interface are found to be structurally similar to two target proteins (one side on one protein, the other on another protein), then the proteins are predicted to be interacting and modeled using the binding site dictated by the template interface. A schematic description of the method is illustrated in Figure 3. Putative interactions are then re-ordered by flexible refinement. PRISM protocol is a collection of scripts that performs a) preparing a set of template interfaces from known complexes, b) preparing surfaces of target proteins that interactions among them to be predicted, c) structural alignment of templates to targets, d) scoring with flexible refinement. The method can be used to model a protein interaction by selecting the two potentially interacting proteins as targets, and using all non-redundant interfaces as the template set. Although the method is limited by the availability and coverage of known protein-protein interfaces from PDB, the continuous growth of the PDB database will increase the applicability of the method. In fact, a recent study on the structural coverage of known protein interfaces already points out that the coverage is close to complete (Gao and Skolnick, 2010).

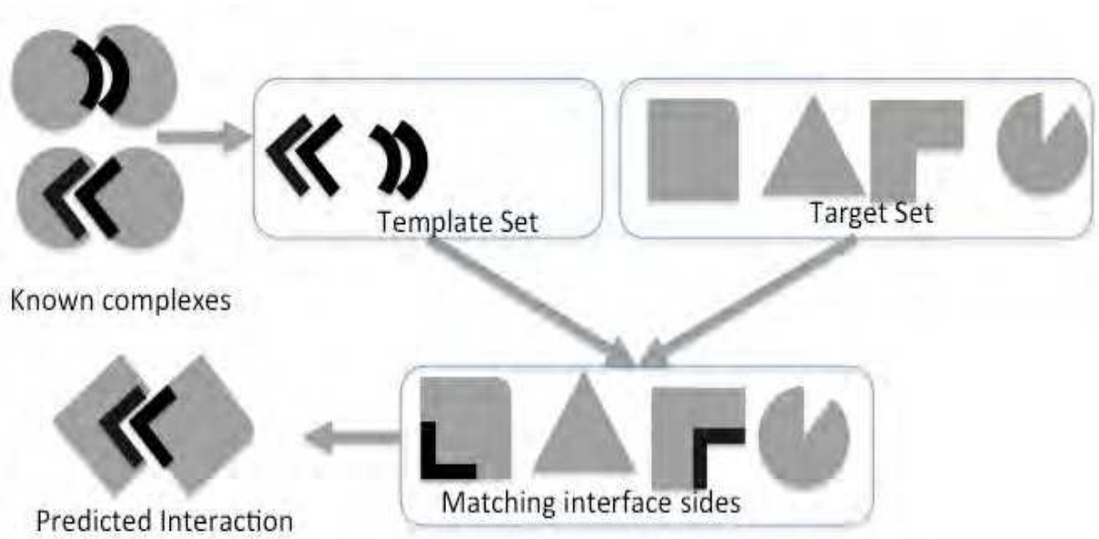


Fig. 3. Schematic representation of PRISM: Two target proteins are predicted to interact if the two complementary sides of a template interface are found to be structurally similar to them (a different side on each protein).

3.3 Protein-protein docking

In contrast to previously described methods (which are based on the structural knowledge of the interaction), protein docking is one of the computational techniques for elucidating the structures of binary bio-molecules (e.g. two proteins) when experimental data regarding the structure of the complex is lacking but the structures of the interacting proteins are known. Docking techniques sample the orientation of two unbound protein

structures to produce several predictions about their interaction, followed by a scoring step to rank the predictions. These methods were introduced in 1978 (Wodak and Janin, 1978). Since then, docking algorithms have substantially improved, with a breakthrough in algorithm speed given by the introduction of the Fast Fourier Transform (FFT) (Katchalski-Katzir et al., 1992) (e.g. FTDock (Gabb et al., 1997), ZDock (Mintseris et al., 2007), PIPER (Kozakov et al., 2006)), and by some other very successful geometry-based methods (e.g. FRODOCK (Garzon et al., 2009), Hex (Ritchie and Kemp, 2000), MolFit (Katchalski-Katzir et al., 1992)). A docking procedure usually involves several steps (Vajda and Kozakov, 2009). First, a rigid-docking search is performed by treating the two proteins as rigid bodies. One of the proteins, called the receptor, is kept fixed while the other protein, the ligand, is rotated and translated around the first. Next, further refinement of some structures takes place, allowing changes in conformation of the two unbound structures upon binding (Dobbins et al., 2008; Shen et al., 2008); this step may or may not be supported by experimental evidence.

3.3.1 The ranking problem

Docking methods yield a large number of output conformations (ranging from 10000 to more than 50000), which include a large number of false positives. Thus, a crucial point after a rigid-docking search is the discrimination of near-native structures for further consideration and refinement. The number of selected conformations typically spans from 10 to 2000. There are two non-excluding strategies to perform such selection. The first strategy is to re-rank the docked conformations with a scoring function, which is supposed to rank near-native structures at the top (i.e. describe the molecular environment of the molecular interaction). Scoring functions are usually built upon different properties of protein-protein interactions observed in known binary complexes. These properties include physical and chemical characteristics of the binding site, at the level of residue or atomic contacts (Z-rank (Pierce and Weng, 2007), Fold X (Guerois et al., 2002)). Among these scoring functions, statistical potential is a term that refers to a knowledge-based scoring function that depends on specific properties of known protein-protein interactions stored in some database. Initially, statistical potentials were derived in order to distinguish a correct protein fold (i.e. near-native) of a model from a plethora of generated solutions (see section 4.2). In contrast to atomistic-detailed scoring functions, statistical potentials represent a much faster approach to solve this problem. It has been recently shown that the performance of split statistical potentials to rank docking poses (see following sections) may surpass that of scoring functions encoding atomistic energy terms or other statistical potentials (Feliu et al., 2011).

The atomistic scoring potentials of Z-rank and FoldX split the score into a linear combination of energetic terms and further obtained the best parameterization. In FoldX (1) the energy terms were split in the van der Waals (Gvdw), electrostatic (Gel), solvation (Gsolv) and hydrogen bonding (GHbond) contributions, and the entropy was also included. Some of these terms were split with different weights (i.e. the solvation of hydrophobic residues had a different weight than the solvation of polar residues, and the entropy of the main-chain had different weight than the entropy of side-chains). The parameters optimizing the final score were obtained using single-point mutations of nine different proteins and the corresponding free energies obtained with their 3D conformations.

$$\begin{aligned}
\Delta G &= \alpha_{vdw}\Delta G_{vdw} + \Delta G_{solv} + \Delta G_{Hbond} + \Delta G_{el} + T\Delta S \\
\Delta G_{solv} &= \alpha_{sh}\Delta G^{solv}_{hydrophobic} + \alpha_{sp}\Delta G^{solv}_{polar} \\
\Delta G_{Hbond} &= \Delta G_{water-bridge} + \left(G^{prot}_{Hbond} - G^{water}_{Hbond}\right) \\
\Delta S &= \alpha_{mc}\Delta S_{main-chain} + \alpha_{sc}\Delta S_{side-chain}
\end{aligned} \tag{1}$$

In Z-rank the energies were also split in van der Waals, electrostatic and solvation terms, but the weights of van der Waals and electrostatic interactions were different for attractive (a) and repulsive (r) interactions, and also different for short-range (<5Å) and long-range (>5 Å) interactions (sr and lr, for short and long ranges, respectively):

$$\begin{aligned}
score &= E_{vdw} + E_{el} + E_{solv} \\
E_{vdw} &= \alpha^{lrr}_{vdw}E^{lrr}_{vdw} + \alpha^{lra}_{vdw}E^{lra}_{vdw} + \alpha^{srr}_{vdw}E^{srr}_{vdw} + \alpha^{sra}_{vdw}E^{sra}_{vdw} \\
E_{el} &= \alpha^{lrr}_{el}E^{lrr}_{el} + \alpha^{lra}_{el}E^{lra}_{el} + \alpha^{srr}_{el}E^{srr}_{el} + \alpha^{sra}_{el}E^{sra}_{el}
\end{aligned} \tag{2}$$

The second strategy follows the rationale that near-native structures will show a broader and deeper well in the energy landscape compared to non–near-native structures. This assumption is the basis of clustering a collection of output conformations (around 1000–2000) as a function of the number of similar structures. Clustering is performed using as the similarity measure either the Cα binding site root mean square deviation (named I-RMSD) (Comeau et al., 2004) or the ligand Cα RMSD (Ritchie and Kemp, 2000). Selection based on the clustering methodology has proved to be better for determining near-native conformations than selection based solely on scoring functions (Ritchie and Kemp, 2000; Vajda and Kozakov, 2009). Consequently, the clustering method has become popular, mainly in combination with a re-ranking given by a scoring function that guides the selection of structures to cluster (Comeau et al., 2004; Shen et al., 2008).

3.3.2 Knowledge based potentials

In knowledge-based potentials, also named statistical potentials, the interaction between two residues is scored by the potential of mean force (PMF) obtained from the probability of finding a pair of residues at a given distance (Sippl, 1990). Let k_B denote the Boltzmann constant and let T be the standard temperature (300K). If A and B are the two interacting chains and a,b are two residues in chains A and B (respectively) at distance d, the potential of mean force is given by:

$$PMF(a,b,d) = PMF_{std}(d) - k_B T \log\left(\frac{P(a,b \mid d)}{P(a)P(b)}\right) \tag{3}$$

where $PFM_{std}(d)=k_B T\log(P(d))$; P(a), P(b) are the individual probabilities of residues a, b; P(a,b | d) is the conditional probability of residues a,b at distance smaller or equal to d and P(d) the probability of any pairs of residues at distance smaller or equal to d. All probabilities correspond to the observed frequencies of the events in the reference database (i.e. 3DID (Stein et al., 2005))

The score of the interaction is then defined as the sum over all interacting pairs of the pair residue scores. Formally, if $a_1,...a_s$ is the residue sequence of chain A, $b_1,...,b_r$ is the residue

sequence of chain B, Γ is the set of pair position indices (i,j) of interacting residues a_i,b_j at distance d_{ij}, then the statistical potential E_{pair} is:

$$E_{pair} = \sum_{(i,j)\in\Gamma} PMF(a_i, b_j, d_{ij}) \tag{4}$$

As energy can usually be split in independent terms from which different forces are derived, the statistical potential can also be split in terms that would describe the different parts of the interaction as particular forces. Particularly, considering a residue *condition* θ as the triplet formed by (secondary structure, polarity, degree of exposure), then the PMF in (3) can be decomposed using:

$$\begin{aligned}
PMF_{pair}(a,b) &= -k_B T \log\left(\frac{P(a,b \mid d_{ab})}{P(a)P(b)P(d_{ab})}\right) \\
PMF_{local}(a,b) &= k_B T \log\left(\frac{P(a \mid \theta_a)P(\theta_a)}{P(a)}\right) + k_B T \log\left(\frac{P(b \mid \theta_b)P(\theta_b)}{P(b)}\right) \\
PMF_{3D}(a,b) &= k_B T \log(P(d_{ab})) \\
PMF_{3DC}(a,b) &= k_B T \log\left(\frac{P(\theta_a,\theta_b \mid d_{ab})}{P(\theta_a,\theta_b)}\right) \\
PMF_{S3DC}(a,b) &= -k_B T \log\left(\frac{P(a,b \mid d_{ab},\theta_a,\theta_b)P(\theta_a,\theta_b)}{P(a,b \mid \theta_a,\theta_b)P(\theta_a,\theta_b \mid d_{ab})}\right)
\end{aligned} \tag{5}$$

Finally, the split statistical potentials E_{pair}, E_{local}, E_{3D}, E_{3DC}, and E_{S3DC} can be obtained by applying the formula (4) to the decomposed PMFs (5), with corresponding subindexes between E_ and PMF_. It was shown (Aloy and Oliva, 2009) that Epair admits a decomposition of the form:

$$E_{pair} = E_{S3DC} - E_{3DC} + E_{3D} - E_{Local} + E_{cmp} \tag{6}$$

where E_{cmp} is a residual energy term depending only on the conditions of the interacting residues and accounts for the reference state (first term in PMF equations). This equation was initially derived for the scoring of protein folds, but it remains valid when applied to the residues in the interface between two interacting proteins (Feliu et al., 2011).

Note that the statistical potential E_{S3DC} is a refinement of the residue-pair statistical potential E_{pair}, in the sense that it takes into account not only the residues that interact but also the condition in which each of them sits. On the contrary, the statistical potential E_{3DC} depends on the occurrence of interacting conditions, disregarding the specific interacting residues. The score E_{local} reflects the probability of placing a residue on a specific condition. Moreover, it splits into two terms, each of them depending only on the probability of placing a certain residue in some condition for each chain separately. The energy term E_{3D} concerns only the distance at which pairs of residues interact, and increases together with the number of interacting residue-pairs, thus being proportional to the number of residues implied in the interface.

3.3.3 Using split statistical potentials to rank docking poses

To test the scoring functions, the benchmark decoy dataset of Weng and cols. (Hwang et al., 2008) is widely used as gold standard. This dataset is based on a set of non-redundant real interactions for which both the complex 3D structure and the individual chain structures are available. It consists of a collection of binary complexes (124) with known structure (named targets) and a set of decoys for each of them (named target set). The 54,000 decoys generated using the rigid-body docking algorithm ZDock3.0 (Mintseris et al., 2007) from the individual chain structures were considered. The set of binary-complex conformations of a rigid-body prediction are classified according to the expected difficulties to obtain a near-native solution of the target. They deal with three types named: easy, medium and difficult cases. In total, the dataset consists of 124 cases, 88 of which are straight forward for rigid-body docking, 19 are medium and 17 are difficult cases for which further conformational changes are required upon binding. Only 97 of them (88 rigid-body and 9 medium) fit into the common near-native decoy criterion of structures differing from the native one at most 2.5Å (computed in terms of I-RMSD from the native structure). For difficult cases it is not possible to have near-native poses because of the deformation suffered by one or two of the protein partners. Thus, a different definition of a successful prediction is required in these cases. A selected pose was considered good if its I-RMSD differs less than 0.5Å from the lowest I-RMSD among all the decoys in the target set. This measure enables to determine if the scoring function top-ranks the best available decoys of the set. Figure 4 shows the success curves for the split potentials, revealing the relative importance of E_{local} and E_{3D} in the composition of the residue-pair statistical potential E_{pair}.

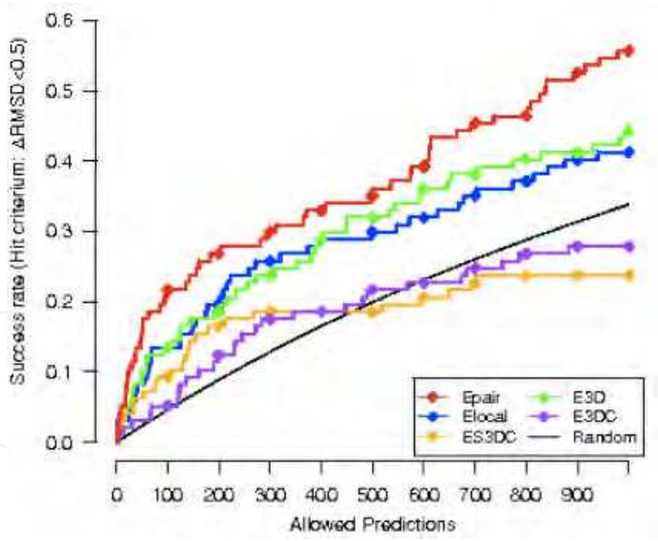


Fig. 4. Success curves for the split potentials: Success curves on the whole benchmark dataset are plotted for the five statistical potentials E_{pair} (red), E_{S3DC} (orange), E_{local} (blue), E_{3D} (light green) and E_{3DC} (purple), plus the success curve expected by random (black).

Based on the observation that E_{pair} and E_{S3DC} provided a fairly amount of non-overlapping hits, a new ranking strategy was defined: "MixRank". This strategy consists of first considering the lists of decoys ranked by different scoring functions separately, and then alternatively selecting one decoy from each list. Then, in order to avoid repetitions, we

apply a removal of redundant predictions (Feliu and Oliva, 2010). That is, we do not include decoys that are less than 5Å of I-RMSD from an already selected decoy. This way of removal of redundancies was analyzed (Feliu and Oliva, 2010) and was proved to provide better selection of near-native decoys. This ranking strategy proved to be able to compete with other ranking strategies based on atomistic-detailed scoring functions if large conformational changes of the interacting partners are required for the interaction. These are the cases typically included in the medium and difficult categories of the benchmark data set. This is shown in Figure 5, where E_{pair} and MixRank surpass ranking system based either on a reference statistical potential (RPScore (Moont et al., 1999)) or on an atomistic-detailed scoring function (ZRank) when predicting near-native poses within the medium and difficult categories of the benchmark data set.

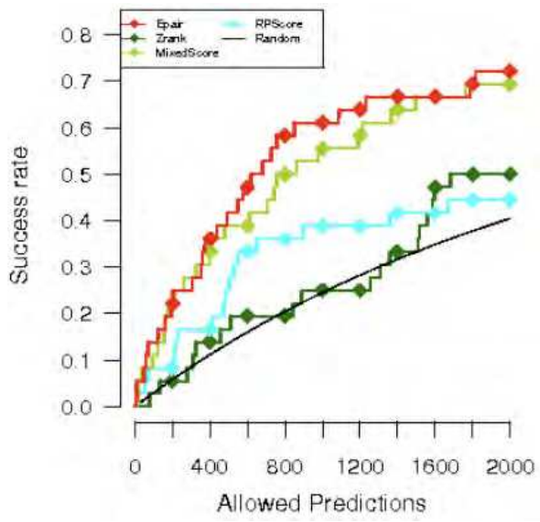


Fig. 5. Different ranking approaches compared for difficult cases of the benchmark data set: Success curves are plotted after removal of redundant solutions for the MixRank strategy (light green), Epair (red), Zrank (dark green) and RPScore (blue) scoring functions, and also compared with the success curve expected by random (black), only with the medium and difficult cases of the benchmark dataset.

4. Errors in models

The quality of the obtained model establishes the limits of the biological information that can be safely extracted from it. Although all structural models may enclose errors, these become less of a problem if correctly detected and assessed: once an error is identified, it is possible to discriminate whether it affects key structural or functional regions. Therefore an essential step in any structural modeling process is the detection of the wrongly modeled regions.

4.1 Sources of errors

In comparative modeling (homology modeling and threading), wrongly modeled regions are expected to be more frequent as the sequence identity between the query protein and the template decreases. Errors can be expected to occur at any step of the process, thus, they can

be catalogued according to the step in which they can be found and, therefore, the step in which they can be corrected or compensated. Docking techniques may incorporate similar errors during the step of molecular refinement.

Wrong template selection is the most costly error that can be found in a modeling process. Being a key step in the process, the selection of a wrong template cannot be overcome at any other part of the process and will inevitable yield to a wrong model. Correcting such error implies going back to the beginning of the modeling process and start all over again. The selection of a wrong template usually derives from the lack of a sequence homologous enough to sequence the query protein. A lot of effort is being put into trying to describe the optimal thresholds of identity and similarity to decide whether or not a sequence can be chosen as template.

Misalignment errors tend to appear under a 40% sequence identity. Their abundance rapidly increases below 30% of identity, as the occurrence of local regions with very low sequence identity makes wrong alignments more feasible. These errors are specially focused on gap misplacements in the alignment, and are one of the major sources of problems in homology modeling. As with the detection of the correct template, the sequence-template alignment is key, and correcting it requires redoing the alignment.

Structural distortions can be found both in well-aligned and in unaligned regions. Those in aligned regions appear when the sequence identity is too low in a local region and the sequence does, in fact, acquire a different secondary structure than that of the template. This problem can be overcome by using several templates in low identity regions in order to explore the possibilities. The regions that, even with multiple templates, are not aligned to any template have to be predicted by energy-based methods of database searching. The alignments at the sequence boundaries and 3D boundaries of such regions will determine the accuracy of the prediction.

Finally, side chain packing needs to be optimized especially as sequence identity decreases. Such optimization can be a major issue, specifically when it involves residues implicated in the protein's function and mostly in the interface of interacting proteins.

4.2 Detecting the errors

Automated methods for detecting errors in 3D models rely on the knowledge of previously solved structures in the PDB. This knowledge has lead to identify stereochemical and energy-related restrictions in the final 3D conformation of a protein. Considering stereochemical restrictions, perhaps the most obvious is that two amino-acids cannot clash (i.e. they cannot occupy the same spatial region). In addition, not all possible relative orientations of two correlative amino-acids in the protein sequence are allowed. These orientations are defined by the Φ and Ψ angles of the amino-acidic bond and the applicable restrictions are summarized in the Ramachandran diagram (Ramachandran et al., 1963), which represents the allowed conformations as a function of the Φ and Ψ angles. PROCHECK program (Laskowski et al., 1996) assess the overall quality of a protein model based on these parameters.

Besides stereochemistry, there are other protein spatial features in the proteins that could be used as indicators of errors in the models: packing, creation of a hydrophobic core,

residue and atomic solvent accessibilities, spatial distribution of charged groups, distribution of atom-atom distances, and main-chain hydrogen bonding structures (Sali, 1995). These are key features to understand the mechanisms by which a protein finds its native state. This mechanism is known as the folding pathway and the possibilities space for the folding of a protein is vast (Levinthal, 1968). Solving this problem requires an accurate potential describing the interactions among different amino-acid residues (Dinner et al., 2000). However, the use of such atomistic-detailed potentials (Brooks et al., 2009) is quasi-prohibitive and it does not ensure the native and biologically active conformation.

An alternative approach to the full atomistic description is to construct a coarse grained potential. The aim of such potential would be to approximate the function: a) whose global minimum corresponds to the native structure (Sippl, 1990), and b) capable to drive the structure from incorrect folding states toward native-like conformations (i.e. the having a correlation with native structure similarity (Keasar and Levitt, 2003)) describing a funnel-like energy surface. This scoring function, termed knowledge-based or statistical potential, works as a coarse-grained descriptor of the environment of the protein, and can be used to assess the quality of a protein 3D model. Based on this approach PROSAII (Sippl, 1993) is probably the most widely used program to assess the quality of a protein 3D model. Similarly, specific potentials have been derived for the interaction between macromolecules in order to assess protein-protein interactions (e.g., M-TASSER (Chen and Skolnick, 2008), MULTIPROSPECTOR (Lu et al., 2002) or InterPreTS (Aloy and Russell, 2003). Nevertheless, a funneling theory such as the Levinthal paradox in protein folding is still under development and some explanations are recently found (Wass et al., 2011).

5. Integrative modeling

The previous detailed methods could be useful in small complexes, where the docking of few subunits can solve the quaternary structure. However, the assembly of large macromolecular complexes such as the nucleopore complex, which contains more than 450 proteins, is unaffordable. In these cases, the presence of such amount of subunits forces the necessity to find methods that could manage the assembly problem in terms of costs and time.

During the last years, the integration of the maximum amount of structural information available about the structurally unknown macromolecular complex has become the state of the art solution to this problem. The main idea of this methodology is to use particular characteristics of the complex that can be synergistically combined in order to restrict the possible solutions to only those consistent with these features.

Electron microscopy has been established as a crucial technique for studying the structure of macromolecular assemblies (Alber et al., 2007). The resolution is insufficient to construct an atomic model but reveals insights into the shape and size of the whole complex. Thus, fitting atomic-resolution structures into the electron density maps is a suitable method for determining not only large macromolecular assemblies but also small ones.

Several methods have been developed for simultaneously fitting the individual protein subunits into the density map of their assembly. MultiFit (Lasker et al., 2010b) solve the position and orientation of each component within a protein structure using a function score that maximizes the quality of fit in the electron density map, the protrusion from the density map envelope, and the complementary shape between subunits. An optimizer algorithm DOMINO (Discrete Optimization of Multiple Interacting Objects) (Lasker et al., 2009) searches like a puzzle the positions of the subunits within a discrete sampling space. Each subunit is placed in a particular position inside the density map, conditioning the position of the rest of the subunits. This algorithm is used to efficiently find the global minimum in an affordable way.

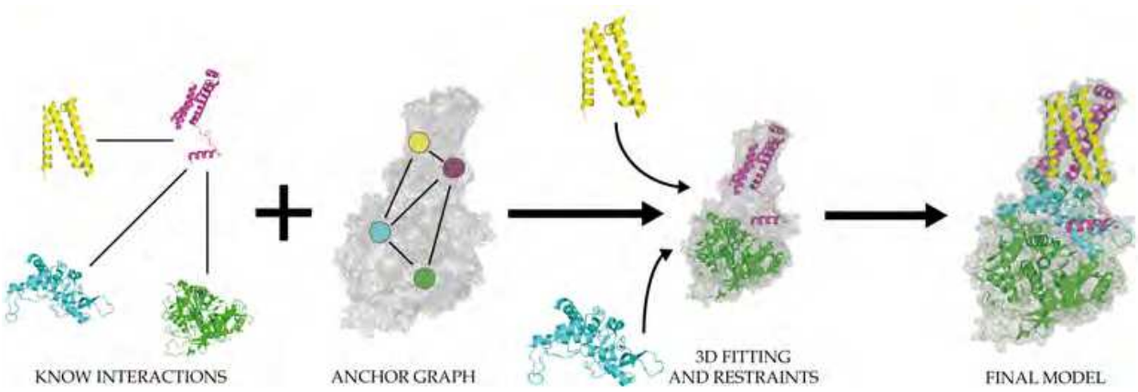


Fig. 6. Schematic representation of integrative modeling.

Often, the electron density map or the high-resolution structure of the subunits is not available. In these cases, it is not possible to apply the fitting procedure mentioned above. However, the integrative approach is not restricted to this data. There are different techniques that provide different types of information that can be used to understand particular features of the assembled complex. Table 1 highlights a list of proteomics, biophysical and computational methods used to obtain this valuable data.

In this way, Sali and collaborators developed an integrative modeling platform (IMP) (Lasker et al., 2010a) that collect this information and consider them simultaneously to generate models consistent with the data. This platform was used to describe the nuclear pore complex (Alber et al., 2007) and the structure of chromatin at megabase scale (Bau et al., 2010). Moreover, this platform can be used to solve any kind of 3D structure when enough data is provided.

IMP performs its function in an iterative series of four different steps. Below, a brief description of each step gave us an insight into how this heterogeneous data can be combined to deliver such large complex models.

Type of Structural Information	Techniques
Composition	Mass spectrometry and quantitative immunoblotting
Interactions	Genetic interactions and bioinformatics predictions
Connectivity	Affinity purification and surface plasmon resonance (SPR)
Interaction partners	Yeast to hybrid, protein microarrays, protein-fragment complementation assay (PCA) and calorimetry
Interaction distances	Fluorescence resonance energy transfer (FRET), bioluminescence resonance energy transfer (BRET) and cross-linking
Complex shape	X-ray scattering (SAXS) Cryo-electron microscopy, Cryo-electron tomography and Negative stain electron microscopy
Protein positions	High resolution electron microscopy, gold-labelling, green fluorescence protein (GFP) labelling and Docking
Residue positions	Crosslinking, hydrogen/deuterium exchange, Limited Proteolysis and Footprinting
Atomic positions	X-rays crystallography and nuclear magnetic resonance (NMR)

Table 1. Proteomics, biophysical and computational methods used to obtain information for modelling macromolecular complexes.

5.1 Data gathering

The collection of structural information is the first requirement needed to start the assembly process. The techniques listed in table 1 are appropriate generators of this data. In addition, a large amount of biological information is available through databases. Table 2 lists some databases with structural relevant information.

5.2 System representation and data translation into spatial restraints

One of the most characteristic features of the integrative modeling process is the ability to use structures that are not solved in high-resolution. In those cases, it is necessary to find an appropriate representation of the system. For example, on one hand, an atomic-resolution structure can be represented with particles corresponding to atoms and, on the other hand, in a low-resolution structure a single particle can represent a sphere corresponding to a group of atoms, residues or domains. Consequently, the resolution of the final complex is dictated by the resolution of the available data.

The raw data gathered in the first step must to be translated into spatial restraints, which specify values for the encoded data in order to decide if the model satisfies or not the experimental information about it. A restraint is a scoring function that reaches its minimum if the feature is consistent with the experimental data. A 0 indicates a model that is perfectly consistent with the restraint, whereas the result of the function is higher when the restraint is violated. In Table 3, most common types of restrains are reviewed.

Database	Description
PDB	PDB (Protein Data Bank) is the worldwide repository of information of 3D biological molecules structures
ModBase	ModBase is a relational database of protein structure models calculated by comparative homology modelling of known structures
SCOWLP	SCOWLP (Structural Characterization Of Water, Ligands and Proteins) is a relational database for detailed structural analysis of PDB protein interfaces at atomic level. The SCOWLP includes proteins, ligands and water as descriptors of interfaces
3DID	3did (3D interacting domains) is a collection of domain-domain interactions extracted from atomic-resolution structures. Each domain is associated to a Pfam domain and the database is GO term functional annotated
EMdataBank	EM Data Bank is a database of cryo-electron microscopy maps, models and associated metadata
BioGRID	BioGRID (Biological General for Interaction Datasets) is a database that archives genetic and proteomic interactions curated from high-throughput datasets and individual studies
PRISM	PRISM (Protein Interactions by Structural Matching) is a web-served compilation of protein-protein interaction interfaces
SCOPPI	SCOPPI (Structural Classification of Protein-Protein Interactions) is a database of all domain-domain interactions and their interfaces derived from PDB structure files and SCOP domain definitions

Table 2. Databases of structural information suitable for the integrative modelling process.

Type of restraint	Description of the restraint
Distance restraints	Restraints the distance between two particles
Connectivity restraints	Restraints all proteins in a set to interact or not.
Quality of fit restraint	Restraints the overlapping of the particles in an electron density map
Excluded volume	Restraint steric clashes
Geometric complementary	Maintains the geometric complementary between two particles interfaces
Statistical potential restraint	Restraint depending on the frequencies of contacts in previous solved complexes
Angle restraint	Restraint the angle between three particles
Protein localization restraint	Restraints a particle in a specific position
Complex diameter restraint	Restraints the maximum distance between the two most distance particles
Symmetry restraint	Maintains the same configuration of equivalent particles across multiple symmetry units
Radial distribution function restraint	Restraints the correlation between experimentally measured and computed radial distribution functions

Table 3. Most common types of spatial restrains obtained from structural data.

5.3 Calculation of an ensemble consistent with the restraints

At this point, the different restraints are combined into a final scoring function, which is commonly the sum of the singular scoring functions corresponding to each restraint. Then, the configuration of the constituent protein beads is determined by optimizing this scoring function.

The optimization process consists of searching through the configuration space the positions and orientations of the structural subunits that minimizes this function. It starts from random positions and iteratively moves them to minimize the violation of the restraints. In essence, a kind of 'force' pulls the proteins together to the native complex configuration. For this task, it is possible to use methods that explore the scoring function landscape in an efficient manner, such as conjugate gradient, molecular dynamics with simulated annealing or personalized optimizers, such as DOMINO (Lasker et al., 2009).

5.4 Analysis of the ensemble

Assuming a unique native state of the complex, the optimization process it is supposed to give a single model that satisfies all restraints. However, if the data used to encode the restraints is insufficient, more than one solution might be obtained. This problem could be solved introducing new restrains and running the process again. Conversely, in case of incorrect restraints, it is possible that no solution is obtained because there is not a model that satisfies all the restraints. In conclusion, the integrative method is a very powerful tool but it is clearly conditioned by the quality of the gathered information. Finally, the structure of the complex needs to be evaluated using similar approaches as in modelling, but adding the quality of the accomplishment of the restraints applied to construct the macro-complex.

6. Conclusions

Protein sequences are totally valueless if meaningful information about their biological function is not reported. In the past 30 years clear relationships between proteins sequence, structure, and function have been proven. Thus, the knowledge of a protein's 3D structure is normally required to completely understand its function. Since many proteins act in association with others, the knowledge of the structure of the complex formed by this association (named quaternary structure) is crucial to understand how proteins perform their functions. In this chapter we have attempted to establish the capabilities and limitations of currently available computational methods for predicting the tertiary and quaternary structure of proteins. Different strategies can be followed depending on the data available, and this review hopefully could serve as a practical guide for modelling the tertiary structures of proteins and its association into complexes.

When known structures of homologous proteins are available, these can be used as a template to model the structure of a target protein or a protein complex in a process termed comparative modelling. Being the current knowledge on the structure of protein complexes much more limited than that of single proteins, several databases of protein-protein interfaces such as PRISM (Aytuna et al., 2005) have been developed to overcome this problem. In comparative modelling, the percentage of sequence identity between the problem proteins and the templates is crucial. Below a certain threshold of sequence identity

(~30%) comparative modelling becomes a difficult task even for experts. In any case, models must be critically evaluated to be sure that they are correct, devoting most efforts to the region involved in the function (usually implying its interaction with other proteins or compounds).

On the lack of experimental data of the structure of a complex of proteins, protein docking is one of the computational techniques for elucidating the structures of binary interactions. We have shown that the use of split knowledge-based statistical potentials to score and rank the different docking solutions can be as accurate as atomistic-detailed potentials (Feliu et al., 2011) in any type of docking. Furthermore, these statistical poteintals surpass atomistic detailed scores when the complex requires large conformational changes of the interacting partners upon the interaction and we apply a rigid docking protocol.

Finally, we reviewed how different sources of experimental data are synergistically used to model large macromolecular complexes by the Integrative Modelling Platform (Lasker et al., 2010a). This approach has successfully been used to elucidate the structure of the nucleopore complex or the structure of chromatin at megabase scale.

7. References

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., *et al.* (2007). Determining the architectures of macromolecular assemblies. *Nature 450*, 683-694.

Aloy, P., and Oliva, B. (2009). Splitting statistical potentials into meaningful scoring functions: testing the prediction of near-native structures from decoy conformations. *BMC Struct Biol 9*, 71.

Aloy, P., and Russell, R.B. (2003). InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics 19*, 161-162.

Aloy, P., and Russell, R.B. (2004). Ten thousand interactions for the molecular biologist. *Nat Biotechnol 22*, 1317-1321.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res 25*, 3389-3402.

Aytuna, A.S., Gursoy, A., and Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics 21*, 2850-2855.

Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. (2010). The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol 18*, 107-114.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res 28*, 235-242.

Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., *et al.* (2009). CHARMM: the biomolecular simulation program. *J Comput Chem 30*, 1545-1614.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci 13*, 190-202.

Chen, H., and Skolnick, J. (2008). M-TASSER: an algorithm for protein quaternary structure prediction. *Biophys J 94*, 918-928.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res 31*, 3497-3500.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science 267*, 383-386.

Comeau, S.R., Gatchell, D.W., Vajda, S., and Camacho, C.J. (2004). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics 20*, 45-50.

Davis, F.P., and Sali, A. (2005). PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics 21*, 1901-1907.

De, S., Krishnadev, O., Srinivasan, N., and Rekha, N. (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct Biol 5*, 15.

Dinner, A.R., Sali, A., Smith, L.J., Dobson, C.M., and Karplus, M. (2000). Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem Sci 25*, 331-339.

Dobbins, S.E., Lesk, V.I., and Sternberg, M.J. (2008). Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci U S A 105*, 10390-10395.

Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics 14*, 755-763.

Feliu, E., Aloy, P., and Oliva, B. (2011). On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. *Protein Sci.*

Feliu, E., and Oliva, B. (2010). How different from random are docking predictions when ranked by scoring functions? *Proteins 78*, 3376-3385.

Fernandez-Fuentes, N., Rai, B.K., Madrid-Aliste, C.J., Fajardo, J.E., and Fiser, A. (2007). Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics 23*, 2558-2565.

Fornes, O., Aragues, R., Espadaler, J., Marti-Renom, M.A., Sali, A., and Oliva, B. (2009). ModLink+: improving fold recognition by using protein-protein interactions. *Bioinformatics 25*, 1506-1512.

Gabb, H.A., Jackson, R.M., and Sternberg, M.J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol 272*, 106-120.

Gao, M., and Skolnick, J. (2010). Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proc Natl Acad Sci U S A 107*, 22517-22522.

Garzon, J.I., Lopez-Blanco, J.R., Pons, C., Kovacs, J., Abagyan, R., Fernandez-Recio, J., and Chacon, P. (2009). FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics 25*, 2544-2551.

Guerois, R., Nielsen, J.E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol 320*, 369-387.

Guharoy, M., and Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proc Natl Acad Sci U S A 102*, 15447-15452.

Hwang, H., Pierce, B., Mintseris, J., Janin, J., and Weng, Z. (2008). Protein-protein docking benchmark version 3.0. *Proteins 73*, 705-709.

Janin, J. (2010). Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst 6*, 2351-2362.

Jones, D.T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol 287*, 797-815.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., and Vakser, I.A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A 89*, 2195-2199.

Keasar, C., and Levitt, M. (2003). A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J Mol Biol 329*, 159-174.

Kelley, L.A., and Sternberg, M.J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc 4*, 363-371.

Keskin, O., Tsai, C.J., Wolfson, H., and Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci 13*, 1043-1055.

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

Lasker, K., Phillips, J.L., Russel, D., Velazquez-Muriel, J., Schneidman-Duhovny, D., Tjioe, E., Webb, B., Schlessinger, A., and Sali, A. (2010a). Integrative structure modeling of macromolecular assemblies from proteomics data. *Mol Cell Proteomics 9*, 1689-1702.

Lasker, K., Sali, A., and Wolfson, H.J. (2010b). Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins 78*, 3205-3211.

Lasker, K., Topf, M., Sali, A., and Wolfson, H.J. (2009). Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J Mol Biol 388*, 180-194.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR 8*, 477-486.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., *et al.* (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol 487*, 545-574.

Levinthal, C. (1968). Are there pathways for protein folding? *J Chem Phys*, 44-45.

Lu, L., Lu, H., and Skolnick, J. (2002). MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins 49*, 350-364.

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A 100*, 5772-5777.

Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res 11*, 2120-2126.

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

Moont, G., Gabb, H.A., and Sternberg, M.J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins 35*, 364-373.

Neuvirth, H., Raz, R., and Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol 338*, 181-199.

Nooren, I.M., and Thornton, J.M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol 325*, 991-1018.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol 302*, 205-217.

Ofran, Y., and Rost, B. (2007). ISIS: interaction sites identified from sequence. *Bioinformatics 23*, e13-16.

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

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol 7*, 95-99.

Ritchie, D.W., and Kemp, G.J. (2000). Protein docking using spherical polar Fourier correlations. *Proteins 39*, 178-194.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng 12*, 85-94.

Russell, R.B., and Barton, G.J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins 14*, 309-323.

Sali, A. (1995). Modeling mutations and homologous proteins. *Curr Opin Biotechnol 6*, 437-451.

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. *Proteins 23*, 318-326.

Shen, Y., Paschalidis, I., Vakili, P., and Vajda, S. (2008). Protein docking by the underestimation of free energy funnels in the space of encounter complexes. *PLoS Comput Biol 4*, e1000191.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol 213*, 859-883.

Sippl, M.J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins 17*, 355-362.

Stein, A., Russell, R.B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res 33*, D413-417.

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

Tsai, C.J., Xu, D., and Nussinov, R. (1997). Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes. *Protein Sci 6*, 1793-1805.

Tuncbag, N., Gursoy, A., Guney, E., Nussinov, R., and Keskin, O. (2008). Architectures and functional coverage of protein-protein interfaces. *J Mol Biol 381*, 785-802.

Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. (2009). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform 10*, 217-232.

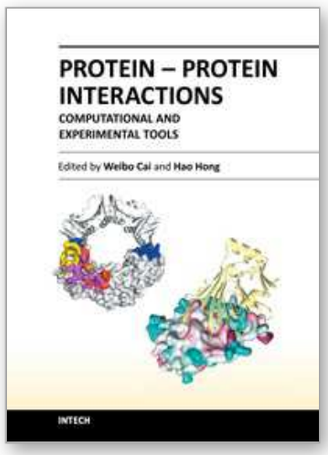
Vajda, S., and Kozakov, D. (2009). Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol 19*, 164-170.

Wass, M.N., Fuentes, G., Pons, C., Pazos, F., and Valencia, A. (2011). Towards the prediction of protein interaction partners using physical docking. *Molecular systems biology 7*, 469.

Watson, J.D., Laskowski, R.A., and Thornton, J.M. (2005). Predicting protein function from sequence and structural data. *Curr Opin Struct Biol 15*, 275-284.

Winter, C., Henschel, A., Kim, W.K., and Schroeder, M. (2006). SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res 34*, D310-314.

Wodak, S.J., and Janin, J. (1978). Computer analysis of protein-protein interaction. *J Mol Biol 124*, 323-342.



Protein-Protein Interactions - Computational and Experimental Tools

Edited by Dr. Weibo Cai

ISBN 978-953-51-0397-4
Hard cover, 472 pages
Publisher InTech
Published online 30, March, 2012
Published in print edition March, 2012

Proteins are indispensable players in virtually all biological events. The functions of proteins are coordinated through intricate regulatory networks of transient protein-protein interactions (PPIs). To predict and/or study PPIs, a wide variety of techniques have been developed over the last several decades. Many in vitro and in vivo assays have been implemented to explore the mechanism of these ubiquitous interactions. However, despite significant advances in these experimental approaches, many limitations exist such as false-positives/false-negatives, difficulty in obtaining crystal structures of proteins, challenges in the detection of transient PPI, among others. To overcome these limitations, many computational approaches have been developed which are becoming increasingly widely used to facilitate the investigation of PPIs. This book has gathered an ensemble of experts in the field, in 22 chapters, which have been broadly categorized into Computational Approaches, Experimental Approaches, and Others.

How to reference

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

Baldo Oliva, Joan Planas-Iglesias, Jaume Bonet, Manuel A. Marín-López, Elisenda Feliu and Attila Gursoy (2012). Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence, Protein-Protein Interactions - Computational and Experimental Tools, Dr. Weibo Cai (Ed.), ISBN: 978-953-51-0397-4, InTech, Available from: http://www.intechopen.com/books/protein-protein-interactions-computational-and-experimental-tools/structural-bioinformatics-of-proteins-predicting-the-tertiary-and-quaternary-structure-of-proteins-f

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen