Dissertation

Submitted to the

Combined Faculties for the Natural Sciences and for Mathematics of the Ruperto-Carola University of Heidelberg, Germany

For the degree of

Doctor of Natural Sciences

Protein Complexes Structure Prediction by Combination of Binary Interactions Derived by Homology

presented by

Graduate Engineer: **Matthieu Pichaud**

Dissertation

Submitted to the

Combined Faculties for the Natural Sciences and for Mathematics of the Ruperto-Carola University of Heidelberg, Germany

For the degree of

Doctor of Natural Sciences

presented by

Graduate Engineer: Matthieu Pichaud

Born in: Nantes, France

Oral examination: _____

Protein Complexes Structure Prediction by Combination of Binary Interactions Derived by Homology

Referees: Dr. Elena Conti

Prof. Dr. Irmgard Sinning

Willst du ins Unendliche schreiten,
Geh nur im Endlichen nach allen seiten.
Johann Wolfgang von Goethe

Table of contents

Acknowledgments

First of all, I would like to thank Dr. Rob Russell, my supervisor, for introducing me to the fascinating world of protein structures. His vast knowledge, his enthusiasm for emerging ideas and his encouragements contributed immensely to the success of this Ph.D.

Many thanks to all the members of the Russell group for their great support, interesting discussions, inspiring advices and mainly for making the atmosphere so nice and stimulating. It is hard for me to know what I will miss the most from this time in room V115: “Privet”, “Sacrebleu” or “La Java Bleue”.

I am very grateful to the members of my Thesis Advisory Committee, Prof. Dr. Irmi Sinning, Dr. Elena Conti and Dr. Carsten Schultz for their invaluable help and care, not only scientific.

Thanks also to Prof. Dr. Kummer and Dr. Anne-Claude Gavin for making me the honor of joining my Thesis Defense Committee.

I would like to thank the people who kindly spent a great amount of their time reading and enriching this essay of their comments, Mirana, Julie, Stu, Erik, Damien, Victor, Matthew, Chad and Rob. More than connections, I have learned a lot from their remarks (‘for that’ is now banned from my speech).

Last but not least, I would like to thank my family, my ‘family’ from Heidelberg, Isabelle and my friends for their warm and kind support. Without them, the task would have been even harder and these years far less fun.

Publications

Publication 1:

A structural perspective on protein-protein interactions. Russell RB, Alber F, Aloy P, Davis FP, Korkin D, **Pichaud M**, Topf M, Sali A. Curr Opin Struct Biol. 2004 Jun; 14(3):313-24. Review.

Publication 2:

Protein complexes: structure prediction challenges for the 21st century. Aloy P, **Pichaud M**, Russell RB. Curr Opin Struct Biol. 2005 Feb; 15(1):15-22. Review.

Zusammenfassung

Proteine spielen eine Schlüsselrolle in den meisten, wenn nicht sogar allen, zellulären Prozessen. Sie üben ihre Funktion jedoch selten für sich allein aus und für gewöhnlich formen sie multimolekulare Komplexe. Die strukturelle Beschreibung der Bildung solcher Komplexe vermittelt viele Details über die biochemischen Prozesse, die schlussendlich zur Funktion des Komplexes führen. Da die Bestimmung der Bildung solcher Komplexe experimentell anspruchsvoll ist und bleibt, gibt es nur einen kleinen Teil bekannter Proteinekomplexe, die strukturell aufgeklärt sind. Somit werden alternative Methoden gesucht um Proteinstrukturen und ihre Komplexierung zu erschließen.

In meiner Arbeit habe ich ein Programm entwickelt, dass die Komplexierung von Proteinen anhand ihrer Struktur und der Struktur ihrer Untereinheiten vorhersagt. Dieses Programm sammelt die Vorhersagen der gekoppelten Anordnung der Untereinheiten, die von homologen Interaktionsvorlagen abgeleitet sind. Alle möglichen Anordnungen der Untereinheiten werden aus einem Graphen ausgelesen, der das Problem wiedergibt. Die Vorhersagen werden hinsichtlich der Sequenz- und Strukturhomologien der Untereinheiten mit den Vorlagen ausgewertet oder anhand der interagierenden Grenzflächen der Vorhersagen verglichen. Die Methode bezieht sich auf Drei-Domänen Komplexierung bekannter Strukturen und auf neun vollständige Strukturen, die auf unterschiedliche Weise aus ihren Untereinheiten zusammengesetzt werden können. Als Ziel der Arbeit wurde versucht, die Komplexierung der RNA Polymerase I und die Struktur des CDC48/Ufd1/Npl4 Komplexes aus dem Ubiquitin-Proteasom-Weg vorherzusagen.

Diese Vorgehensweise scheint angemessen wie die Ergebnisse, auf die sich meine Arbeit bezieht (auf denen meine Arbeit gründet), zeigen. Wir konnten die Orientierung derjenigen Untereinheiten der RNA Polymerase I vorhersagen, die

homolog zu den Untereinheiten der RNA Polymerase II sind. Dies zeigt, dass Strukturen mit direkter Homologie leicht vorherzusagen und zu bestimmen sind. Für den Cdc48/Ufd1/Npl4 Komplex zeigen wir drei Vorhersagen, die weitere Untersuchungen lohnenswert erscheinen lassen.

Dieses Programm kann für jedwede Art von Proteinkomplexen verwendet werden und ist im Besonderen nützlich, wenn sich die Strukturen experimentell schwer bestimmen lassen.

Abstract

Proteins are key participants in most cellular processes. However, they rarely function in isolation and usually they form multimolecular assemblies. The structural description of such an assembly provides critical details about the protein function. As the determination of such structures remains a great experimental challenge, only a small fraction of known protein complexes are currently available. This has created a need for alternative, predictive methods that can bridge the gap between complexes that are known to exist in the cell, and those for which structural information is available.

This thesis presents a program to predict the structure of protein assemblies from the structures of their subunits. The method combines predictions of pairwise arrangements derived from homologous interaction templates to consider all possible assemblies. The problem of finding the best arrangement is modeled as a graph to allow fast graph traversing algorithms to be exploited. Individual predictions are evaluated by sequence identity or structural similarity between the subunits and the templates or by evaluation of the interfaces in the predictions. The method is benchmarked on three-domain assemblies derived from known structures and on nine complete structures that could possibly be re-assembled in a non-trivial fashion from previously determined structures. The method was also applied to complexes determined from high-throughput complex determination procedures, including RNA polymerase I and the Cdc48/Ufd1/Npl4 complex from the ubiquitin-proteasome pathway.

The benchmark demonstrates that the approach can often work on small assemblies. For larger complexes, certain details can be predicted, and occassionaly large parts of the complex, though currently a lack of suitable templates limits applicability. Nevertheless, the method can now be applied to any protein complex and should be particularly useful when structures are difficult to

obtain by experiments, and where additional information, such as pairwise interactions or stoichiometry, is available.

Introduction

1. The protein complex, one level of biological organization

Living organisms are highly organized and consist of complex structures at many different resolutions. From the atomic level to the macroscopic there are many organizational interactions and processes: several chemical interactions form biomolecules (*e.g.* DNA, RNA, proteins, peptides) that eventually organize in assemblies (protein complexes, cell wall) and arrange further in sub cellular compartments (nucleus, proteasome, lysosome). Cells and organs finally collaborate to form the organism. Each level of organization is generally studied at specific resolutions that embrace their inherent specificity.

Protein complexes, or assemblies, are organizations of particular importance. They participate in all biological functions and are usually made of several proteins arranged in space via specific protein-protein interactions. The best descriptions of the structures of protein complexes come when a high-resolution structure is available by X-ray crystallography or Nuclear Magnetic Resonance, though key insights can also come from lower resolution structures that are increasingly available from electron microscopy. These techniques, however, remain time consuming meaning that there is now a large gap between complexes that are known in the sense that the proteins composing them have been determined, and those for which a 3D structure is available.

2. Protein structures

2.1. Tertiary structure

Proteins are the expressed form of genes. They are made of a chain of amino acids (or residues) that is a functional translation of the information encoded in the corresponding piece of deoxyribonucleic acid (DNA). Each of the 20 amino acids

has particular chemical properties: polar, non-polar, charged, extended, small, structurally constrained or flexible. To accommodate those amino acids in an energetically favorable manner in the context of the biological medium (the solvent, the cytosol, the membrane, etc.), the protein usually folds and acquires a precise tertiary structure in space. The primary structure of the protein refers to the sequence of amino acids, and the secondary structure consists of stretches of amino-acids that organize in helices called α-helices or in strands called β-sheets. To best satisfy their environmental preferences, hydrophobic residues in soluble proteins are normally buried at the core of the structure, whereas hydrophic residues normally prefer to be exposed to solvent. Different preferences apply to membrane proteins, where, for example, hydrophobic residues often reside in the membrane.

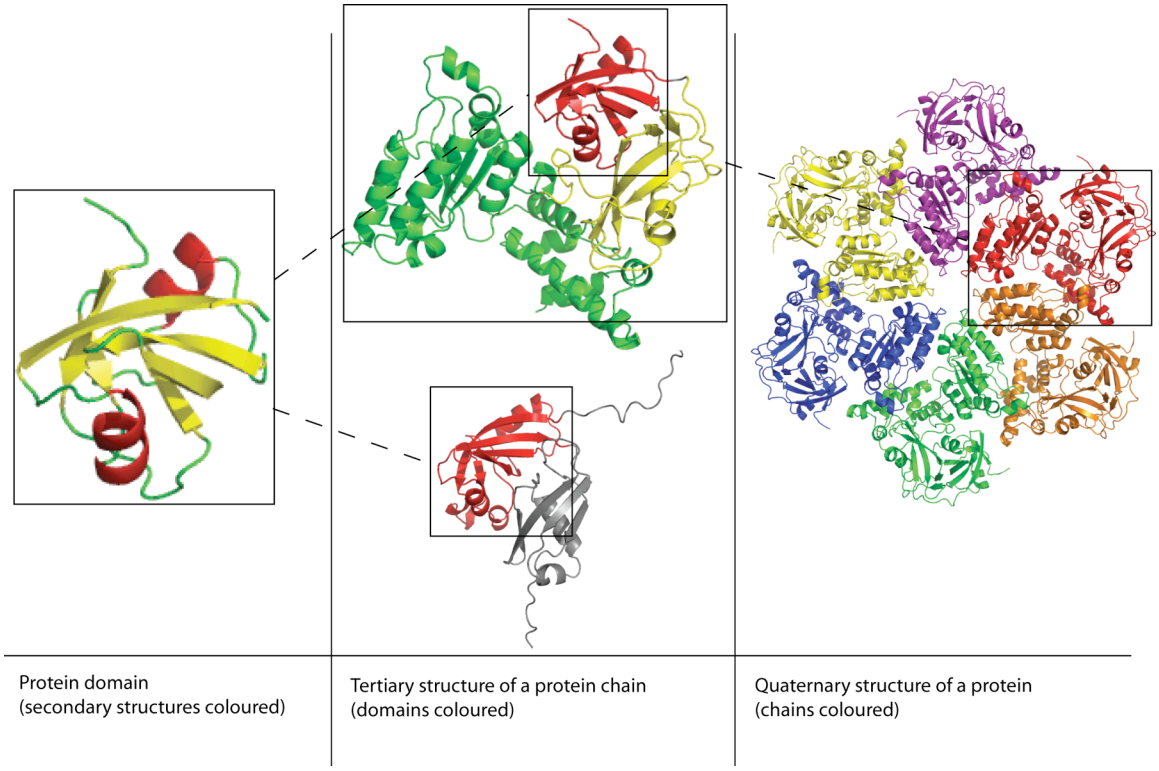

Figure 1: Protein domain, tertiary and quaternary structures (illustrated on the exosome)

Domains have been defined either as compact regions in a protein structure (Richardson 1981), segments of residues that conserve properties during evolution (Bork 1989) or parts of the protein that fold independently (Wetlaufer 1973). Today domains are generally referred to in the context of protein function, being discrete units that are normally performing a particular sub-function of the protein. For example, a catalytic domain will perform an enzymatic function, complemented by regulatory domains that might be responsible for substrate specificity, or localization. Because they are autonomous and often functional structural units, domains are often seen and used as basic functional and evolutionary entities of proteins (Apic *et al*. 2001; Copley *et al*. 2002; Vogel *et al*. 2004; Bornberg-Bauer *et al*. 2005).

Domains usually consist of about 150 amino acids. These normally come from a continuous fragment of the protein chain, but sometimes they span over several fragments. Most proteins are multi-domain, containing several domains (Murzin *et al*. 1995; Orengo *et al*. 1997) normally performing discrete sub-functions. It has long been observed that proteins can adopt a similar 3D structure even in the absence of clear sequence similarity. This ultimately led to the idea that nature was somehow limited in the number of folds that could be adopted by proteins. The number of domain folds is normally estimated to be limited to a few thousand (Chothia 1992; Blundell and Johnson 1993; Govindarajan *et al*. 1999; Koonin *et al*. 2002; Orengo *et al*. 2002) and indeed they are commonly re-used in several proteins with some variations (Bork 1991).

2.2. Quaternary structure

Proteins often act together to form stable, functional complexes, sometimes referred as protein assemblies. This spatial organization of different proteins chains is the quaternary structure of a protein. Polypeptide chain surfaces complement each other to support the formation of specific atomic interactions (hydrogen bonds, Van

der Waals, salt bridges) and favorably accommodate residues at the inter-subunit interfaces.

Frequently, several identical polypeptide chains assemble into homo-multimers, or assemblies made of several copies of the same protein chain. The most common of which is the homo-dimer that contains two copies of the same chain (Orlowski *et al.* 2007). These structures built on the repetition of the same proteins tend to be symmetric. Other assemblies are heteromeric, with two or more different proteins acting together.

3. Determining protein structure

The knowledge of the structure of a protein or a complex greatly aids the understanding of molecular function. The more precise the structure determination, the better the determination of the *modus operandi*.

Various biophysical methods can be used to determine protein tertiary or quaternary structure. The main differences between each method are the state and quantity of protein required and the resolution attainable (the ability to capture molecular details of the structure). High-resolution methods (<5Å) like X-ray crystallography and Nuclear Magnetic Resonance (NMR) are difficult because they require high quantities of highly purified homogeneous protein, which can be difficult to obtain and which is often unstable over the course of the experiment. Other techniques can operate on easier to obtain samples, but these normally provide only low-resolution structures.

3.1. Protein over-expression

Even if crucial for most cellular functions, most proteins are present in very small amounts in the cell. Moreover, in its natural state, it is impossible to distinguish the protein of interest from the others. Thus, a step of sample preparation is required in the two high-resolution structure determination methods. When working with X-ray

crystallography and NMR, the protein of interest is usually over-expressed and purified. An expression vector – usually a plasmid – designed to produce large amount of mRNAs coding for the protein of interest is transfected into a host cell (bacteria, yeast, insect or mammal). Once expressed, the protein is extracted from the lysed cell by normal chromatography or by affinity chromatography when the protein of interest was engineered to display a specific tag to help the purification step.

3.2. X-ray crystallography

In X-ray crystallography, the protein is first crystallized, meaning that specific chemical conditions are found in which the protein molecules arrange themselves into a regular lattice in space and form a crystal. This usually requires high concentrations of very pure protein (>97% purity at 2-50 mg/ml). This crystal is thereafter bombarded by X-rays, which are scattered by the molecules in a diffraction pattern captured on a photographic plate or recorded by other methods.

Only the amplitude of diffraction maxima can be read from the diffraction patterns and the phase that is crucial for the determination of the structure has to be determined by other means. Molecular Replacement is efficiently used when the structure of a homologous protein (>25% sequence identity) is known and can be used to get an initial estimate of the phases. However, it tends to bias the model obtained towards the structure of the homologue. Alternatively, the diffraction pattern can be disturbed by soaking heavy-atom derivatives in the native crystal (Multiple Isomorphous Replacement) or choosing radiation wavelengths that correspond to the absorption edge of certain atoms (Multiple Anomalous Dispersion) and phases can be determined by the comparison of several such spectra. Upon determination of good enough phases, an electron density map is calculated in which the main-chain of the protein and the side-chains are carefully fitted.

The quality of the crystal is crucial as it strongly influences the resolution at which the protein structure is solved and the difficulty to fit residues into the density map. X-ray crystallography may reach resolutions below 1Å and can cover molecules that have a wide range of molecular weights when the right conditions are found to crystallize the molecules. Sometimes during the formation of the crystal, proteins pack closely together in a manner that does not reflect any physiological affinity (crystal packing). Attempts have been made to detect such dubious interactions and identify the ones that do not to complete a symmetrical assembly (PQS (Henrick and Thornton 1998)).

X-ray crystallography has proven to be a very powerful technique and accounts for 87% of the structures solved to date. The structures of the ribosome (Ban *et al.* 2000), RNA polymerase II (Cramer *et al.* 2001), the exosome (Lorentzen *et al.* 2005; Liu *et al.* 2006) and the proteasome (Groll *et al.* 1997) are amongst X-ray crystallography's great achievements.

3.3. Nuclear Magnetic Resonance (NMR)

With Nuclear Magnetic Resonance, experimentalists work with high concentrations of proteins directly in solution. Atomic nuclei react differently to an electromagnetic field depending on their nature and their environment. Nuclear magnetic resonance from protons (^{1}H) and isotopically labeled molecules (usually ^{13}C or ^{14}N) is determined. Upon collection of the NMR spectrum, the peaks of resonance must be assigned to pairs of reactive atoms from the molecule in order to obtain their relative position. While X-ray crystallography determines one single structure for the sample, NMR usually provides an ensemble of atomic coordinates (20-30) since several structural arrangements may satisfy the spatial constraints derived from the spectra. NMR is usually applied to small molecules of less than 50 kDa because of the difficulty interpreting NMR spectra and the difference in reactivity of large samples (Yu 1999). NMR accounts for 13% of the molecules solved at a high-resolution. Despite difficulties with large molecules, some

relatively large complexes were solved such as that between GroEL and GroES (Fiaux *et al.* 2002).

When it is not possible to meet the requirements of high-resolution methods, techniques like electron microscopy (EM), small angle scattering (SAS) and electron tomography can be used to obtain lower resolution structures.

3.4. Electron Microscopy (EM)

In EM, a biological specimen is illuminated by electrons. Several projected images of the specimen are taken, aligned and cross-correlated in order to reconstruct the structure. Sample preparation is crucial for success. The specimen is usually stabilized to resist the high vacuum in the microscope column, stained to increase the contrast and sectioned to facilitate the penetration of electrons in the specimen. The main limitation of electron microscopy is the irradiation damage that inevitably affects and distorts the sample. More recent Cryo-EM methods decrease beam damage by collecting data at low-temperature and reduce the damage caused by removal of water from the specimen.

Although EM can reach resolutions around 8Å, which is much higher than the resolution obtained by light microscopy, it still provides far less structural details than X-ray crystallography or NMR. At this resolution, the position and the conformation of the residues cannot be determined precisely. Single particle EM was applied successfully on the 50S ribosomal subunit (7.5Å – (Matadeen *et al.* 1999)) and on the structure of GroEL (~10Å – (Ludtke *et al.* 2001; Ranson *et al.* 2001)).

3.5. Small-angle Scattering (SAS)

In small-angle scattering (SAS), neutrons or X-ray radiation are emitted and scattered by the sample, which does not need any special preparation (Svergun and Koch 2002). The resulting scattering pattern contains information about the

geometry of the sample. SAS was used for the determination of the structure of cAMP-dependent protein kinase (Tung *et al.* 2002).

3.6. Electron tomography

In electron tomography, an electron beam passes through the sample at different angles of rotation. The images are collected and a structure is reconstructed. Electron tomography achieves resolution of 30Å. At this resolution, neither the secondary structure elements nor the tertiary structure elements can be precisely localized. The representation obtained can still be informative as those of the Nuclear Pore Complex (Stoffler *et al.* 2003; Beck *et al.* 2004), virus assemblies (Grunewald *et al.* 2003), or even entire cells (Medalia *et al.* 2002; Hoog *et al.* 2007).

3.7. Hybrid approaches

The precise determination of a protein structure is a difficult exercise restricted to X-ray crystallography and NMR. Still, detailed structural insights into a protein complex can be obtained by fitting high-resolution structures into a low-resolution envelope.

When the EM structure of a complex is determined, it can be used as a framework to constrain the placement of its constitutive subunits if they are available individually (Topf *et al.* 2005). The problem consists in optimizing the fit of high-resolution structures in the low-resolution EM structure of the complex (Volkmann and Hanein 1999; Rossmann 2000; Chacon and Wriggers 2002; Ceulemans and Russell 2004; Topf *et al.* 2005). Such approaches have been successfully applied to the determination of *E. coli* 30S ribosomal subunit (Malhotra and Harvey 1994), the yeast exosome (Aloy *et al.* 2002) and *S. cerevisiae* 80S ribosome (Spahn *et al.* 2001). Moreover, similar approaches are used to determine the structure of single protein chains: Baker *et al.* determined the structure of the capsid protein of Herpesvirus VP26 (Baker *et al.* 2006) and Topf *et al.* developed Moulder (Topf *et al.*

2006), a method to optimize the prediction of protein structures by fitting in CryoEM pictures.

In a similar manner, small-angle scattering can be used to help determining high-resolution structures of large assemblies of a few kDa up to hundreds of MDa. Data obtained by SAS have for example led to the location of the subunits in the low-resolution structure (Krueger *et al.* 2000; Wall *et al.* 2000; Sun *et al.* 2004; Petoukhov and Svergun 2005) and was used for the prediction of the structure of cAMP-dependent protein kinase heterodimer (Zhao *et al.* 1998).

4. Towards a structural determination of protein complexes

When the experimental methods to determine the structure of the complex are either not available, or prove too difficult, then computational predictions can be considered. The prediction of the structure of a complex then requires knowledge of the constituents of the complex, the structure of the each of them (determined or predicted), the protein domains that are likely to serve as binding anchors and a mean to assemble separate constituents into sensible interactions. This process can also be aided by knowledge about how the subunits interact by non-structural techniques such as the two-hybrid system. Here only large-scale approaches are considered, as we search a method that can apply to as many complexes as possible.

4.1. Determination of the composition of a protein complex

4.1.1 Determination of protein interactions

Proteins usually achieve their various functions by interacting with other proteins. The number of interactions between two proteins in yeast has been estimated to be around 30 000 (Kumar and Snyder 2002; von Mering *et al.* 2002). When put in the perspective of the number of protein in the yeast proteome (~6200), it indicates that on average, one protein has 9 protein partners (Sali *et al.* 2003). Because of the

importance of protein neighborhood, extensive efforts have been carried to identify and characterize protein interactions and complexes.

4.1.1.1 Experimental methods

The yeast two-hybrid system (Y2H)

In yeast two-hybrid experiments, the potential interaction between two proteins is studied and reported by hijacking a transcription factor (Figure 2). The protein of interest is fused to one part of a split transcription factor, and potential interacting partners are fused to the other part of the transcription factor. If there is an interaction between the two proteins that are tested, the two parts of transcription factor are brought together leading to the activation of a reporter gene, which is then detected. If the two proteins investigated do not interact, the transcription factor remains split, the reporter gene is not transcribed and therefore no signal is detected. The 'bait' is the protein investigated and is usually tested against a library of potential binding partners, or 'prey' molecules.

Two extensive studies of protein interactions in *Saccharomyces cerevisiae* have been conducted (Uetz *et al*. 2000; Ito *et al*. 2001) and revealed 691 and 841 putative interactions respectively. Surprisingly, the two experimental sets did not overlap much and only 135 interactions were common to both sets (Ito *et al*. 2001). Large-scale experiments have also been performed using proteins from *H. pylori* (Rain *et al*. 2001), *C. elegans* (Li *et al*. 2004), *D. melanogaster* (Giot *et al*. 2003) and humans (Rual *et al*. 2005; Stelzl *et al*. 2005).

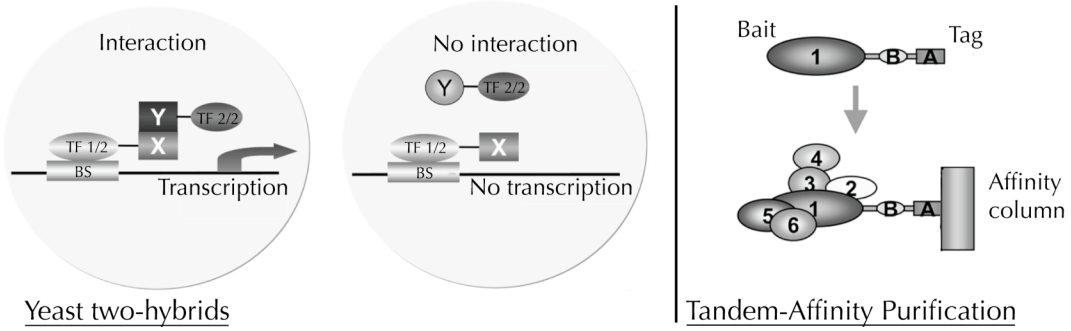

Figure 2: The two main methods for large-scale detection of protein interaction: Yeast Two-Hybrid and Tandem-Affinity Purification with Mass Spectrometry (adapted from (Cho *et al.* 2004))
Left: Description of the Yeast Two-Hybrid experiment. TF 1/2 and TF 2/2 are the two parts of the transcription factor. X and Y are the two proteins tested for interaction. BS is the binding site of the transcription factor. If X and Y interacts, the two parts of the transcription factor are close enough to trigger transcription. Right: Description of the Tandem-Affinity Purification setting. The Protein of interest (1) is fused to a tag and fished out of the cell with its interacting partners (only one purification step is shown).

Tandem-affinity Purification followed by Mass Spectrometry identification (TAP-MS)

Another strategy consists of fishing for the protein of interest and the proteins in contact with it by two rounds of purification and then identifying all the proteins that purify together. The first step is called tandem-affinity purification (TAP) and the identification step is tackled usually by mass spectrometry (MS) (Figure 2). Theoretically, one purification reveals all the binding partners of the protein of interest. However, proteins are sometimes involved in several complexes and thus direct inference of the components of a complex from a purification is uncertain. Moreover, only 70% of proteins are retrieved in the two sets when a purification is repeated (Gavin *et al.* 2002) and all the binding partners of a protein may not be retrieved in one purification. Thus, the method is applied on a large scale and most proteins are tagged and purified. The data from purifications are finally combined to determine the composition of the complexes. This approach was used to study the interactome (the space of interacting proteins) at large-scale in yeast (Gavin *et al.* 2002; Ho *et al.* 2002; Gavin *et al.* 2006; Krogan *et al.* 2006).

Direct determination of complex architecture by Mass Spectrometry

Alternatively, a novel approach consists of preserving the protein interactions in the gas phase that comes before mass spectrometry detection (Hernandez *et al*. 2006) so that complexes and subunits can be detected. It provides information about the composition of the complex, its stoichiometry and even the composition of the complex subunits. Mass spectrometry determination of intact complexes was successfully applied to the bacterial 20s proteasome (Sharon *et al*. 2007), p97-Ufd1-Npl4 (Pye *et al*. 2007) and the eukaryotic exosome (Robinson C, in preparation).

4.1.1.2 Bioinformatics methods

In parallel with experimental methods, bioinformatics, mostly using genomic data, contributed to a better understanding of protein interactions. Phylogenetic profiling groups together those proteins that occur in the same set of organisms, since this implies that they may tackle a common function and may also have the same binding partners. Obviously, this method does not apply to genes essential to the cell maintenance of the cell because they are present in most organisms. It indicates a probable co-evolution and a putative common function but does not indicate direct contact between proteins (Gaasterland 1998; Pellegrini *et al*. 1999). Bacterial genomes are organized in functional clusters called operons. Two proteins homologous to proteins from the same bacterial operon are likely to participate in the same function. More generally, conservation of the proximity of two genes may indicate a common function (Dandekar *et al*. 1998; Overbeek *et al*. 1999). Also, when two genes in one organism are fused in another organism, it often indicates that their gene products interact (Enright 1999; Marcotte *et al*. 1999). More interactions are detected by co-evolution of conserved regions (Pazos *et al*. 1997; Kann *et al*. 2007). The problem is then to distinguish functional co-evolution from the simple co-evolution in (Pazos *et al*. 2005; Sato *et al*. 2005).

Other methods compile the information from various sources and evaluate interactions by cross-validating several experiments. STRING (Search Tool for the Retrieval of Interacting Proteins (Snel *et al.* 2000; von Mering *et al.* 2007) for instance integrates results from the methods mentioned above with data from protein co-expression, literature and text mining. The combination of all the methods enables stronger assertions about the relationship between two proteins.

Gene evolution is another good indicator of protein functionally relationships. However, the information is mainly functional and cannot compare to details achieved by biophysical methods.

InterPreTS (Aloy and Russell 2003) or MULTIPROSPECTOR (Lu *et al.* 2002) are methods that use information derived from known structures to assess the interaction between proteins homologous to those in the known structure. More specifically, InterPreTS consists of threading two sequences in the structure of the two domains of an interaction and evaluating the likelihood of the interface created by comparing it to the likelihood of the interface formed with the same structure and a random sequence.

4.1.1.3 Repositories

Protein interaction data from various sources (experiments, literature mining, computational predictions) are collected and organized in dedicated databases (DIP (Xenarios *et al.* 2000; Salwinski *et al.* 2004), BIND (Alfarano *et al.* 2005), MPact (Guldener *et al.* 2006), IntAct (Kerrien *et al.* 2007), BioGRID (Stark *et al.* 2006), STRING (von Mering *et al.* 2003), MINT (Zanzoni *et al.* 2002), HPRD (Peri *et al.* 2003)). Similarly, structures of protein interactions are stored and annotated in several databases (3did (Stein *et al.* 2005), PIBASE (Davis and Sali 2005), SCOPPI (Winter *et al.* 2006), iPfam (Finn *et al.* 2005), SNAPPI-DB (Jefferson *et al.* 2007), PROTCOM (Kundrotas and Alexov 2007)). Most of these databases list interactions between protein domains from protein structures deposited in the Protein Data

Bank based on domain assignments from SCOP (Murzin *et al.* 1995) or Pfam (Orengo *et al.* 1997) whereas some attempt to detect similar interactions based on common orientation (3did, SNAPPI) or residue contacts and other interface specific features (PIBASE).

4.1.2 Difference between datasets – Error assessment

TAP-MS experiments do not reveal the direct binding of two proteins contrary to most yeast two-hybrid experiments (exceptions revealed by structural analysis are found in (Aloy and Russell 2002a)) Thus, models have to be used to extrapolate direct interactions from TAP-MS results and usually one of the two extreme models is adopted. In the 'spoke' representation, only the direct interactions between the bait and its preys are considered whereas in the more permissive 'matrix' model, the interactions between all the proteins co-purified are considered. The reality must lie somewhere between those two extremes but it was shown that the spoke model is three times more accurate than the matrix model (Bader and Hogue 2002). Even if the information about direct, physical interactions is not available as it is in the yeast two-hybrid system, TAP-MS experiments provide as well information about protein-protein interactions. Specifically, they provide collections of proteins that act together, of which some pairs invariably are in direct physical contact.

Major efforts have been made to evaluate the quality of large-scale interaction datasets and several types of false results can be studied. Within an experiment, interactions that do not exist to the same extent in the biological medium are sometimes detected because some proteins engage in non-specific interactions in the experimental framework (like heat-shock and ribosomal proteins (Gavin *et al.* 2002)). Such proteins are detected in a large number of purifications and are usually discarded. Moreover when experiments are repeated, the outcome is partly reproduced (~70% for TAP-MS (Gavin *et al.* 2006; Krogan *et al.* 2006), 80% for

Y2H (Rual *et al.* 2005)). It is usually assumed that the proteins detected repeatedly are correct when the others correspond to false detections (*false positives*).
Several studies have attempted to estimate the false-positive rates of these methods. For instance, genome-wide yeast two-hybrid scans were compared to more reliable individual experiments (Mrowka *et al.* 2001) and Von Mering *et al.* (von Mering *et al.* 2002) compared the ratio of interactions between proteins involved in different functions to their expected ratio. Overall, these studies estimate a false positive rate in interaction data around 50%.

It is also possible for interactions to be missed during a screen (*false negatives*). These can arise for multiple reasons: the complex may be transient, insoluble or disrupted by the modifications required to perform the experiments (Aloy and Russell 2002b). Moreover, each technique performs differently depending of the type of interaction (Aloy and Russell 2002a), the cellular localization (Yook *et al.* 2004) and the abundance of the protein (von Mering *et al.* 2002; Gavin *et al.* 2006; Krogan *et al.* 2006). For instance, membrane proteins are rarely retrieved in TAP-purifications and proteins involved in translation are hardly detected in yeast two-hybrid assays.

Another particularity is that most of the information collected about protein interactions and complexes does not account for spatial or temporal aspects of complex formation and existence. Only the superposition of all the possible interactions in which protein is engaged is detected. For instance, Cdc48, an ATPase, has various cellular functions which depend on the adaptor proteins it interacts with: it is involved in spindle assembly with the help of Ufd1 and Npl4 (Cao *et al.* 2003) and plays a role in ubiquitin-dependent protein degradation with another adaptor, Shp1 (Johnson *et al.* 1995). TAP purification data (Gavin *et al.* 2006) indicates clearly that Cdc48 binds the three adaptors. But from the sole purification of Cdc48, it is not possible to separate the two variants of the Cdc48-complex. This example illustrates the temporal integration achieved in TAP

purification experiments and large-scale studies in general: they circumvent the composition of complexes at one time-point and one place and everything that binds a protein at one point or one place is retrieved with little distinction.

Usually, interaction data are crossed with spatial and temporal data to clarify the definition of complexes. Spatial integration is untangled usually by the use of protein localization annotations (GO terms for instance (Ashburner *et al.* 2000)). Similarly, time-dependent expression data are used to reveal the dynamic aspect of complex formation (Jensen *et al.* 2006). Finally, the structure of one protein cannot always accommodate simultaneously the structures of all the complexes in which it is involved, and this information can also be used to untangle the effective composition of protein complexes (Kim *et al.* 2006).

The best way to work with a reliable set of interactions is to combine data from several experiments, even if this impacts the coverage (von Mering *et al.* 2002). More recent studies directly cross-validated their results with protein localization data and functional data (Gavin *et al.* 2006; Krogan *et al.* 2006) in order to filter dubious interactions and improve the detection. Still one must bear in mind that these methods almost always detect a mixture of multiple protein states. Moreover, the datasets normally used to filter for differences in space and time are themselves quite error-prone, and thus represent only a partial solution.

4.1.3 General characteristics of biological networks

When a set of protein interactions is known, they are usually represented in a graph in which the nodes represent proteins and the edges represent interactions. Many of the first analyses of these, and many other biological and real-world networks, revealed the recurrence of many network properties, the most common being 'small world' and 'scale-free'.

A 'small world' network (Watts 1999) is one in which most of the nodes are not direct neighbors and nevertheless remarkably few edges are needed to create a path from any node to any other node. The concept was introduced at the middle of the 20th century (Milgram 1967) in social networks where nodes and edges represented individuals and relationships between them. This concept has since been used to describe the internet, hence the word "hub" is used to define critical nodes that are involved in multiple interactions.

In the graphs derived from protein interactions, most proteins are involved in few interactions and few proteins are involved in many. All studies agree on this property, but they sometimes disagree on what mathematical distribution best describes the graph. Distributions suggested to date include scale-free (Barabasi and Albert 1999; Yook *et al.* 2004), hierarchical (Ravasz *et al.* 2002), and geometric random (Przulj *et al.* 2004). The yeast two-hybrid protein-protein interaction networks from Ito (Ito *et al.* 2001) and Uetz (Uetz *et al.* 2000) seem to follow a scale-free topology, but the portion of the interactome covered by those experiments is insufficient to extrapolate the scale-free property to the whole interactome (Han *et al.* 2005; Pereira-Leal *et al.* 2005).

The study of the interaction graph reveals the importance of certain nodes. 'Hubs' are particular proteins in the network that are involved in many more interactions than average. They tend to be long multi-domain proteins enriched in binding-associated domains (Ekman *et al.* 2006). The removal of such a protein is critical to cell survival (Jeong *et al.* 2001). Vidal and co-workers sub-divided hubs into two overlapping classes: 'party' hubs that interact with their different partners simultaneously and 'date' hubs that bind their partner at different time points or locations (Han *et al.* 2004). The distinction makes some biological sense: party hubs are often central components of large complexes, making several simultaneous interactions, whereas date hubs often correspond to enzymes such as kinases that act on many different substrates, but never at the same time (Aloy and

Russell 2006; Kim *et al.* 2006) The distinction between 'date' hubs and 'party' hubs is debated (Batada *et al.* 2007). Still it illustrates the various means of action of proteins, from high-specialization to promiscuity.

4.1.4 Mining interaction data to define protein complexes

One might expect interaction discovery experiments to uncover complexes unambiguously, but the interaction networks normally reveal fuzzy balls. Therefore, several methods are used to search and extract complexes from such data.

Complexes usually form dense regions in the network because proteins in direct contact or in the same complex are more prone to come together in experiments. In contrast, proteins are less likely to interact when they belong to different complexes. Methods from graph theory can directly be applied to search such dense regions that would correspond to a protein complex and are robust enough to be implemented successfully in this kind of noisy context (Bader and Hogue 2003; Spirin and Mirny 2003; King *et al.* 2004; Arnau *et al.* 2005). Markov clustering is currently amongst the best of these methods (Brohee and van Helden 2006).

In order to improve the detection of protein complexes, most recent analysis used a two-step approach (Gavin *et al.* 2006; Krogan *et al.* 2006). Gavin *et al.* devised a socio affinity score to estimate the propensity of two proteins to interact. Proteins are clustered iteratively using this score and can contribute to protein complexes in three manners (Dezso *et al.* 2003): 'Core' proteins are proteins found in the same set of purifications and for this reason, are likely to belong to the same complex, whereas 'attachments' proteins are found only in a subset of purifications. Finally 'modules' are groups of proteins shared by several complexes. This approach led to

the determination of 491 complexes. Krogran *et al.* estimated the confidence for each interaction by machine learning and then used a Markov clustering algorithm to delimit 547 complexes. Due to slight differences in the protocols, the two sets do not overlap and are complementary (Gagneur *et al.* 2006; Goll and Uetz 2006). The performance of such methods is usually assessed by comparing the complexes predicted to a set of manually curated complexes. The two methods recall around 275 complexes defined in MIPS (Mewes *et al.* 2000). More than a mere data-mining exercise, the problem is to capture the subtle variability in protein complexes and their various modes of organizations (Devos and Russell 2007).

Deriving definitions of complex from interaction data is made difficult by the lack of large-scale error-free data and the great complexity of protein complex arrangements. However, with the better coverage achieved by recent studies, the definition of complexes become more accurate and precise.

4.2. Prediction of protein structure

The difficulties in determining protein structure experimentally long ago prompted attempts to predict protein structure from sequence information. This field has now matured to the point where many approaches can be applied more or less systematically to make useful predictions (Pieper *et al.* 2004). The best way to predict protein structure is to exploit the fact that proteins sharing similar sequences most often adopt a similar 3D structure. When the structure of a protein homolog (*i.e.* a protein that diverged from the same ancestor) is known, the structure of a protein can be predicted from the structure of the homolog, called template. The accuracy of these models depends on the degree of sequence similarity between the two proteins, and the quality of the alignment between them. The best predictions are obtained when the conformation of residues is directly deduced from the structure of the template using the alignment. Weak similarities between the template sequence and the model and unaligned residues make the predictions more error-prone. When two protein sequences share more than 30% sequence

identity, the prediction of the structure is accurate (the difference between the structure and the prediction is around 4Å Root Mean Square Deviation (RMSD)) (Sternberg *et al.* 1999), whereas the predictions using templates with less than about 20% identity are likely to contain many more errors.

When no protein of known structure is homologous to the protein of interest, another method called fold recognition or threading can be used. It is based on the idea that the number of folds taken by protein structures is limited to several thousands (Chothia 1992) of which around 1100 are determined at the moment (Greene *et al.* 2007). When compared to the number of proteins of known structure (40000 in August 2007), it is clear that one fold must account for many hundreds or thousands of structures. Fold recognition exploits the idea that there is a very high chance that a certain protein adopts the fold of a protein of known structure. The protein polypeptide chain is computationally threaded into several possible folds and an energy function evaluates the "goodness of fit" (Jones *et al.* 2000; Zhang *et al.* 2005) *i.e.* the suitability of the fold to accommodate the protein residues. Recent improvements the fields of homology modeling and threading methods are limited and in both cases it is rare that the model predicted is closer to the real structure than the template (Tress *et al.* 2005).

Both homology modeling and threading methods use known protein structures to predict the structure of new proteins. When none of these methods provide satisfying results (*i.e.* when no homolog is found or when no fold accommodates accurately the protein), other methods can be tried to predict the protein structure with information about a complete structure that can be used as reference. Rosetta (Rohl *et al.* 2004) is the most successful amongst such methods (Vincent *et al.* 2005). It is based on the fact that short peptides are limited in the number of conformations they can take and that the same structure of small peptide fragments is used in several proteins. Thus, given a library of small peptide conformations, the Rosetta method explores combinations of fragments using a Monte-Carlo procedure

to search for compact and energy favorable conformations. In their current state, these methods are limited to small peptide fragments (Dill *et al.* 2007) and predictions are still often far from the right answer.

4.3. Detection of domains

4.3.1 Methods for the detection of protein domains

Since domains are protein segments that have been conserved during evolution, they can often be detected and identified using protein sequence comparison and in fact, the first detections of protein domains were based on sequence consensus (Bork 1991). Direct alignment of two protein sequences is informative but rarely sufficient and it is always better to use multiple sequence alignment in the detection of common ancestry (*e.g.* (Altschul *et al.* 1997; Eddy 1998)). The most conserved residues are almost always the most informative during the alignment process. More refined methods are complemented by the use of secondary structure information to detect more remote similarities and achieve better alignments (Soding 2005). In the case of domain assignment, a protein is carefully aligned to all the known domains. If a successful alignment is found, a domain can be assigned to the protein.

Protein structures, of course, allow for a much more rational way to deduce domains. There have been a number of automated approaches to assign domains from structure (*e.g.* DomainParser (Guo *et al.* 2003) and PDP (Alexandrov and Shindyalov 2003), reviewed in (Holland *et al.* 2006)). These mostly attempt to partition the structure into fragments that have the characteristics of domains in terms of compactness, length and radius of gyration. Efficient structure alignments methods on the other hand can be used to compare a structure to those of known domains (SSM (Krissinel and Henrick 2004) or fastSCOP (Tung and Yang 2007)), and the principle of domain *recurrence* has been systematically incorporated into many domain detection schemes (e.g. (Holm and Sander 1998)). A good structural

alignment indicates a new instance of the domain. However, since such methods only apply to known protein structures, new domains cannot be discovered, save for those compact regions that are distinct and do not resemble any known structure.

4.3.2 Database of protein domains

The information relative to protein domains is usually stored in repositories (Pfam (Sonnhammer *et al.* 1997), SMART (Schultz *et al.* 2000), ProDom (Corpet *et al.* 1998), Conserved Domain Database (CDD) (Marchler-Bauer *et al.* 2005), Prosite (Mulder *et al.* 2003), InterPro (Apweiler *et al.* 2001)). Domains are manually annotated or retrieved from the literature and multiple sequence alignments are used to infer new domains.

There are three main collections of domain structures: FSSP (Families of Structuraly Similar Protein (Holm and Sander 1994)), CATH (Class, Architecture, Topology and Homologous superfamily (Orengo *et al.* 1997)) and SCOP (Structural Classification Of Proteins (Murzin *et al.* 1995)). Hierarchies within all of these classifications reflect the different degrees of similarity between domain or protein structures.

The SCOP classification is built manually with support from automated tools. At the fold level, domains have the same secondary structure arrangement with the same connections between secondary structure elements. Domains with the same superfamily, the next category, are thought to have a common ancestry despite little sequence identity is (usually below 30%). The evolutionary relationship is deduced by the presence of common structural or functional features unlikely to arise by convergence. At the family, domains share a clear evolutionary relationship usually with detectable sequence similarities.

In CATH, structures are first divided into domains automatically or by experts in ambiguous cases. They are then classified in an hierarchy with 4 main categories.

The Class-level, which is the highest category, describes the secondary structure composition of the domains. The second category is the manually determined Architecture-level that relates the domain secondary structure to known architectures (*e.g.* beta-propeller). The third level, the Topology level, accounts for the connectivity between secondary structure elements. Finally domains with the same Homology assignment have been grouped because they are thought to have a common ancestor (as evaluated by a high sequence identity or a high structure comparison score).

FSSP is a fully automated and discontinued database of protein folds based on a hierarchical clustering of structures superimposed using Dali (Holm and Sander 1997). The main difference is that FSSP is based purely on automatic structural comparisons of domain structures when CATH and SCOP are annotated by experts.

These databases of protein domains can be used as references to find domains in proteins that lack domain annotation.

4.4. Prediction of the structure of a protein assembly

Different experiments contribute to a better understanding of the composition of protein complexes and also reveal the interactions within them. This information can be used for the prediction of structural features of a protein complex.

4.4.1 Predicting the structure of a protein-protein interaction

4.4.1.1 Predicting sites of protein interaction

In order to predict the structure of a protein interaction, it is sensible to locate the parts of the protein that mediate it. Several methods have been developed to do this, which typically exploit the specific properties of known interfaces.

Protein binding sites usually involve large surface areas (several hundreds of square angstrom) which are either contiguous or which span several patches of residues (Janin and Chothia 1990). Although some protein binding sites are flat, most of the interactions between proteins occur either in large cavities on the protein surface (Hubbard *et al.* 1994) or at protruding loops (Jones and Thornton 1996). To favor space and charge complementarities, such interfaces tend to have on average more hydrophobic groups exposed than the rest of the protein surface (Ringe 1995). Some *in silico* methods search locations on the surface of a protein structure with interface-like topological features (Goodford 1985; Miranker and Karplus 1991) while others probe for positions on the surface predisposed to the binding of protein (Silberstein *et al.* 2003). However structural and chemical considerations are often not enough to pinpoint the site of interaction.

As one would expect, the residues of a protein involved in binding are more conserved (del Sol Mesa *et al.* 2003). As a consequence, when an interaction involving homologous proteins is known, the binding sites can be transferred since it is likely to occupy the same position in the homologs (Bork *et al.* 1998). The evolutionary trace method looks for similarities that are conserved within a family but which differ from other families in the same superfamily. This can help to determine interaction-specific residues that are characteristic of a family (Aloy *et al.* 2001; Landgraf *et al.* 2001; del Sol Mesa *et al.* 2003). Moreover, it is possible to display conservation data directly on the structure of a protein and thus locate the binding site on the molecule and assess visually the quality of the interface between two structures (Consurf (Landau *et al.* 2005), ProFunc (Laskowski *et al.* 2005), Evolutionary Trace Viewer (Morgan *et al.* 2006)).

Hot spots are particular residues on a surface that are critical to an interaction. They are alternatively defined as surface residues in an unfavorable environment (Elcock 2001), high-energy surface residues (Clackson and Wells 1995; Bogan and Thorn 1998) or residues that disrupt the binding of the protein when mutated. Hot

spots are enriched in Tryptophane, Tyrosine and Arginine and usually surrounded by a hydrophobic ring (Bogan and Thorn 1998). Energy calculations are used to determine the position of hot-spot residues, often given the structure of the protein-protein interactions. The problem is in estimating the importance of a residue on the stability of the interaction. The stabilizing effect is evaluated by the energy of side-chain/side-chain interactions (Li *et al.* 2006) or the disruption induced by substituting the residue for an Alanine (Verkhivker *et al.* 2002; Kortemme *et al.* 2004), alternatively any other residue (Guerois *et al.* 2002) or the evaluation of shape specificity and biochemical contacts (Darnell *et al.* 2007)). Another approach consists in searching the protein surfaces for spots where binding affinities are the highest (Gao *et al.* 2004).

4.4.1.2 Prediction of the structure of protein-protein interactions

The prediction of the structure of a complete protein assembly usually begins with the prediction of the interaction between two proteins. In the two methods presented here, two structures are given and means to put them together are sought.

Docking is a procedure that searches for a conformation in which the arrangement of two structures optimizes some criteria. To be successful, the search has to be as exhaustive as possible and the criteria have to be accurate. Every configuration cannot be studied due to the expensive computation it would require. Various criteria are used to evaluate the binding (*e.g.* shape complementarity, free energy, interface assessment). The backbone of the structures is usually not rearranged (in rigid body docking), but in cases where structures are modified upon interaction, computer-expensive flexible docking is applied. The knowledge of the possible location of the interface (called 'modes') on each of the two structures is used to limit the search space and contributes to the achievement of better predictions (Korkin *et al.* 2006).

The applications of docking are unlimited and it is employed to predict the structure of the interaction between any pair of structures. However it cannot discriminate at the moment real interacting proteins from artifacts, structures of interactions that are not real. The performance of docking methods is steadily increasing as reported during the CAPRI meeting (Critical Assessment of Predicted Interactions (Mendez *et al.* 2005)). For the time being, it is most successfully applied to binary interactions between small proteins with known monomer structures, proteins, with high affinity one for each other or with no conformational change upon binding (Gray 2006). It is also not currently possible to use docking to say *whether* proteins interact or not. Instead, it is normally applied in situations where an interaction is known, and a conformation is sought.

When domains are similar (in sequence or structure), they tend to interact the same way (Aloy *et al.* 2003). It is thus feasible to predict the arrangement of two domains when they are homologous to two interacting domains for which the structure of the interaction has been solved. Although the method is limited by the number of interaction templates currently available, the quality of such a prediction is high when performed in the right conditions.

Docking can be generally applied to all structures and is particularly efficient with small tightly-bound structures. The number of templates available limits homology-based modeling, but the quality of predictions can be accurately evaluated by homology between the structures and the templates. A hybrid approach has been developed where docking is constrained by the knowledge of binding regions derived by homology (Korkin *et al.* 2006).

4.4.1.3 The variety of protein-protein interaction structures

In order to estimate the variety of protein-protein interaction structures, a measure that captures the structural differences between two protein-protein interactions was derived previously (Aloy *et al.* 2003) from which the relation between protein sequence and interaction similarity could be deduced. As this measure is used extensively in this thesis, it is discussed here in further details.

iRMSD (Aloy *et al.* 2003) is a measure of the structural similarity between two protein-protein interactions. The two interacting domains must share enough structural resemblance that it is possible to superpose the two protein structures using the trace of their backbone (*e.g.* with STAMP (Russell and Barton 1992) or DALI (Holm and Sander 1993)). Assume that we compare the interaction A1-B1 and the interaction A2-B2 with A1 and A2 being two instances of the protein/domain type A and B1 and B2 being two instances of protein/domain type B. Each protein/domain is represented by a set of 7 coordinates: the center of mass and one point +/- 5 angstroms along the X, Y, and Z axis. A2-B2 is transformed in A2′-B2′ by superposing A2 on A1 and in A2′′-B2′′ by superposing B2 on B1. iRMSD is the root-mean-square distance between the coordinate sets of A2′-B2′ and A2′′-B2′′ (Figure 3). It accounts for both translational and rotational differences between the interactions. Below 10, the structural similarity of the interactions is good; above 10, the similarity is difficult to see by eye. Above a threshold of 20-30% sequence identity, domains are likely to interact in the same way, whereas below this threshold, they are more likely to interact differently. Finally, if domains belong to the same family, whatever the sequence identity, the structures of the interactions are often similar. Obviously, there are exceptions, whereby highly similar sequence interact differently (*e.g.* different antibodies to the same lysozyme) and those where seemingly unrelated protein pairs sharing only a common fold show a similar interacting structure. Many of the former exceptions, such as lectins

(Prabu *et al.* 1999), bacterial chemotaxis-related proteins (Park *et al.* 2004) and domains from different families (Kim and Ison 2005) have been highlighted in the literature.

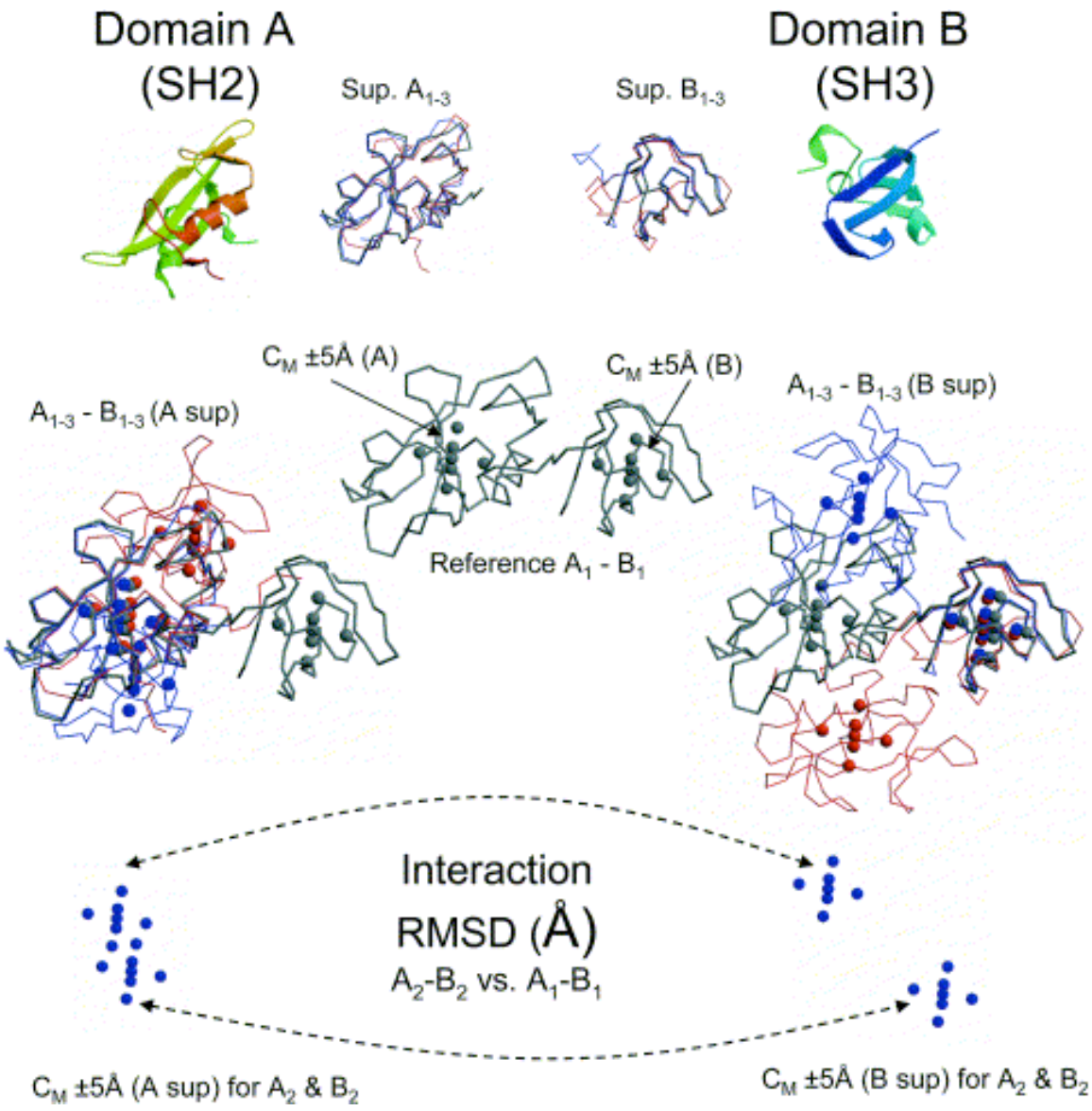

Figure 3: Interaction RMSD (from (Aloy *et al.* 2003))

As more and more proteins are discovered, the number of ways in which proteins could interact could in principle become enormous. However, in a study similar to

that of Chothia who estimated the number of possible protein folds (Chothia 1992), a study of 7 datasets across 4 species showed that the number of possible structures of protein-protein interactions would likely be around 10000 (Aloy and Russell 2004). Thus, the repertoire of protein folds and the repertoire of protein-protein interaction structures are finite and it is meaningful to list them. However, according to the same study, two decades will have to pass before we get structural knowledge of all the interactions. A systematic comparison of all protein-protein interaction structures revealed that there are currently 5677 distinct interfaces (Jefferson *et al.* 2007).

4.4.2 Prediction of the structure of protein complexes

When no biophysical methods apply, all the information available on a protein complex can help for the prediction of its structure. At the moment, only two such strategies have been developed.

4.4.2.1 Serial docking

Inbar *et al.* (Inbar *et al.* 2005) have designed the first method to predict the structure of a protein assembly using only the structures of its constituents. They show that the problem is NP-hard, a class of unsolved mathematical problems, and cannot be solved without some heuristic. First, they try to dock all the structures of all possible pairs of components. They then iteratively assemble the structures of the binary interactions to generate the structure of the most complete and accurate models. The solutions are then re-ranked to estimate the geometrical and biochemical fit of each interaction in the model. They developed two methods based on the same principles: one for combining intra-chain structures, the other for inter-chain structures. Their benchmark set consists of 5 structures of protein complexes selected from the Protein Data Bank that they separate into subunits (from three to ten) before attempting to predict the structure of the assembly. Moreover they tried the same exercise with structures of homologous subunits.

They could predict near-native structures (RMSD<5) and best predictions were amongst the 10 first structures predicted in all the cases.

They show that their method that combines several structures at once outperforms methods that combine domains in a pairwise manner. Thus, Inbar *et al.* demonstrate the validity of their new bottom-up approach for the prediction of structures of protein complexes.

4.4.2.2 Homology based prediction

In another attempt to predict the structure of protein complexes, Aloy *et al.* (Aloy *et al.* 2004) combined interaction templates inferred by homology to predict the structure of new interactions. They assessed the quality of the structure of the interactions predicted by considering sequence similarity between the proteins of interest and the templates, the quality of the interface (determined by InterPreTS) and conservation of functional classes. They could arrange most of the domains for the exosome (Aloy *et al.* 2002), the RNA polymerase II with two interactors TFG2 and SPT5, the Ski complex, the CCT chaperonin in complex with a phosducin (PLP2) and G protein {gamma} homolog (VID27) and the POP complex.

In the case of the exosome, the overall structure was predicted correctly by homology but the arrangement of the proteins within the exosome ring was incorrect. The two proteins that could be bound on RNA polymerase II could not be seen in the available EM map and thus, no assessment of the prediction was possible. The overall shape of CCT could be predicted and was confirmed by EM. Afterwards, the structure of the two proteins was added to the CCT using the EM grid without any homology inference. Finally despite some minor clashes, two remote templates could be used to accommodate the 3 proteins from the Ski complex.

5. The problem

Biophysical methods are the most reliable means to achieve detailed structural determination of protein assemblies. However the larger the protein assembly, the harder the resolution of a high-resolution structure and alternatives have to be used to compensate for the limited application of purely biophysical methods. Hybrid approaches where high-resolution sub-complexes are fitted in low-resolution templates for larger complexes are valuable and when the composition of the complex is known and the structure of its parts known or predicted, iteration of docking can be used to predict the structure of the complete assembly. However, docking approaches do not exploit the potential structural similarity of interactions between homologous proteins.

Here we use homology modeling, a fast and reliable method for the prediction of the structure of protein interactions. The method automatically combines homology-predicted interaction structures to assemble the structure of complexes as inspired by the pioneering work of Aloy *at al* (Aloy *et al.* 2004).

The procedure is benchmarked using elementary arrangements of three domains and few complete structures that are predicted from pieces. Potential applications of the method are sought amongst complexes. We show how it fares on two large assemblies and propose three possible candidates for one of them.

Material and Methods

1. Overview of the method

The procedure takes as input a set of sequences from a complex that is the target for prediction. A series of sequence comparisons using HHsearch (Soding 2005) identifies all possible matches to known structures as determined by biophysical methods. These matches are then parsed for those that permit two or more parts of the target proteins to be modeled in an interaction (*interaction templates*). Interaction templates are collected and stored in a database. Some redundancy is removed from the set of templates by comparing the interactions using an interaction-specific distance (iRMSD) and keeping the distinct ones within each structure.

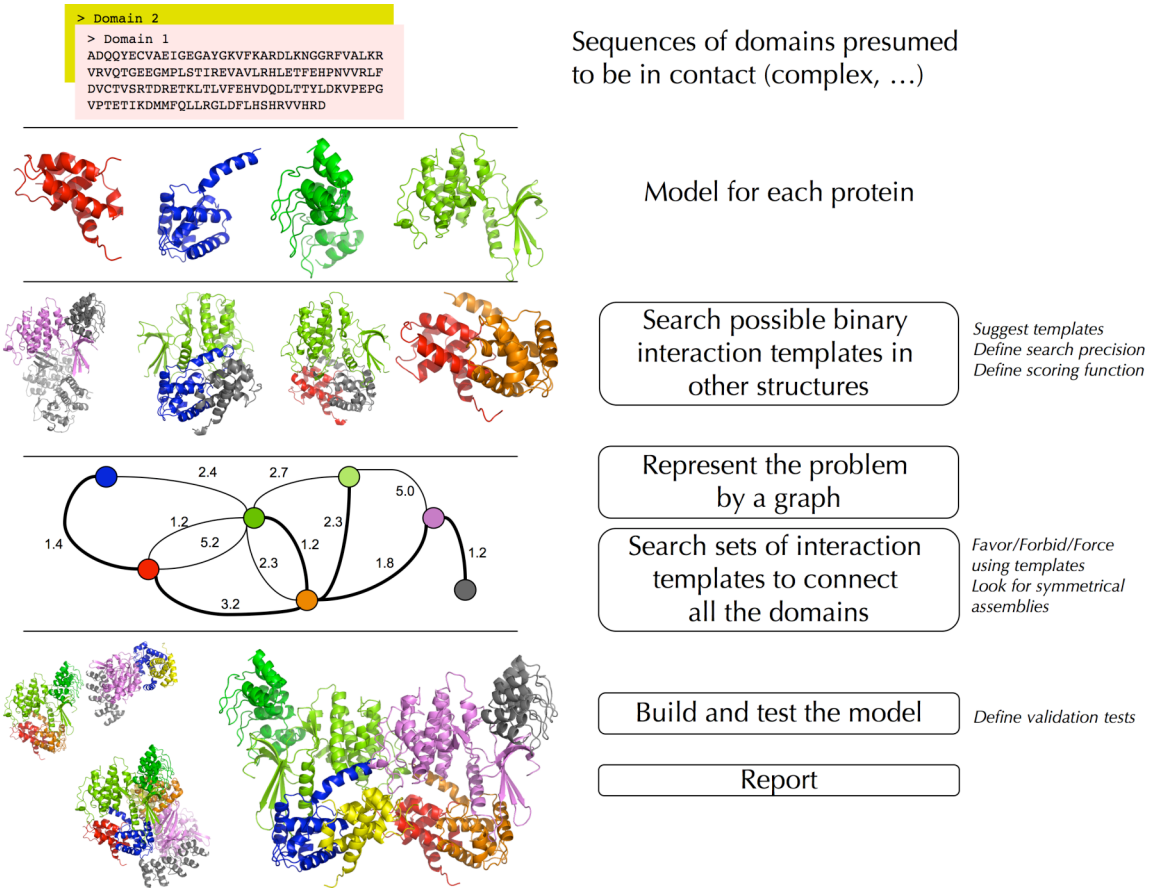

Figure 4: Global description of the procedure

The templates are used to predict the orientation of putatively interacting protein domains. Then, orientations are represented as a graph, and a graph searching algorithm is used to find combinations of interaction templates that can accommodate as many domains as possible in a single structure (Figure 4). Each combination represents a candidate structure for the assembly that is evaluated by three alternative criteria: (i) the sequence or (ii) structural similarity between the domains assembled and the domains from the templates or (iii) by evaluating the interface of each interaction predicted. The performance of the procedure is demonstrated on several benchmark sets derived from previously solved complex structures.

2. Collecting interaction templates

Interaction templates from solved structures are used to predict possible orientations of domain pairs from the target complex. First all possible interaction templates are collected from solved protein structures and organized in a database. Then within each structure, redundant interactions are filtered using an interaction-specific distance, the iRMSD.

The filtering step is important, as structures sometimes contain many copies of the same interaction. For instance, viral capsids form football-like structures made of many domains, and like footballs subunits are combined in the same way to form the final structures. Thus, despite numerous interactions in the structures, most are redundant and only the few non-redundant interaction templates are kept by this procedure.

2.1. Comparison of interaction templates using iRMSD

iRMSD is a measure for comparing the structures of two interactions. It is used to detect the similarity of two interaction templates so that only single representatives are kept. The procedure assumes that each interaction in the pair being compared consists of similar domains or proteins in contact. That is, to compare the interaction A1-B1 and A2-B2 it must be the case that one can structurally equivalence domains/proteins A1 & A2, and B1 & B2.

Here a matrix interpretation of the iRMSD computation is presented. A transformation in space that conserves distances and oriented angles, as is the case when finding the optimal transformation from one structure to another, is defined by a rotation and a translation. The inverse transformation is:

$$(ri, ti) = (r^{-1}, -r^{-1}.t)$$

Then, iRMSD is computed by the following procedure:

- Call (rA, tA) the rotation and translation from domain A in structure 1 to domain A in structure 2. (rB, tB) are named analogously.

- Compute a set of 7 coordinates for each of domain A and B in structure 1 as explained in Aloy *et al.* (Aloy *et al.* 2003) and collect them in 3x7 matrices called modelA and modelB

- Compute iRMSD directly by applying the following formula that is a matrix form of the procedure described originally by Aloy *et al*:

$$(rAi, tAi) = (rA^{-1}, -rA^{-1}.tA) \quad (rBi, tBi) = (rB^{-1}, -rB^{-1}.tB)$$
$$(r1, t1) = (rA.rBi, rA.tBi + tA) \quad (r2, t2) = (rB.rAi, rB.tAi + tB)$$

$$iRMSD(A1-B1, A2-B2) = max\begin{pmatrix} RMSD(r1.modelA + t1, modelA) \\ RMSD(r2.modelB + t2, modelB) \end{pmatrix}$$

Here the set of 7 coordinates of the two interacting domains and the two superpositions of the domains from one interaction to the domains from the other are sufficient to compute iRMSD. This matrix formulation of iRMSD is the one used throughout this work, in particular when comparing interaction templates to filter the database.

2.2. Inventory and selection of interaction templates

Here is the procedure to go over each protein structure, extract interactions between domains and select those that are most distinct:

- Assign domain types to each structure from the Protein Data Bank (PDB) using a

manually curated database of protein domains (SCOP preview version 1.71 - with curated domain assignments for structures released before January 2005 and automatic assignments for structures released between January 2005 and April 2006).

- In each structure, search and list interacting domains, or those that have more than 5 residues within 10Å.
- Keep only one template amongst redundant interactions of the same type (iRMSD<1Å) within one structure

The structures that are not classified in the most recent version of the database of protein domains are not used. In addition, domain types that do not belong to real classes of domains are discarded (*e.g.* small proteins, peptides, low resolution structures, designed proteins). We treat intramolecular (within one protein chain) and intermolecular (between protein chains) domain interactions the same way in the procedure.

In general, we limit the number of interaction templates by removing redundant interactions in each protein structure: within one structure we compare all the interactions between proteins of the same family type by computing the iRMSD score. When the iRMSD score is below 1Å, the two interactions are similar and only one is kept in the database.

2.3. Database schema

To retrieve the information related to interaction templates, all the data are collected and stored in a MySQL database represented in the following schema (Figure 5).

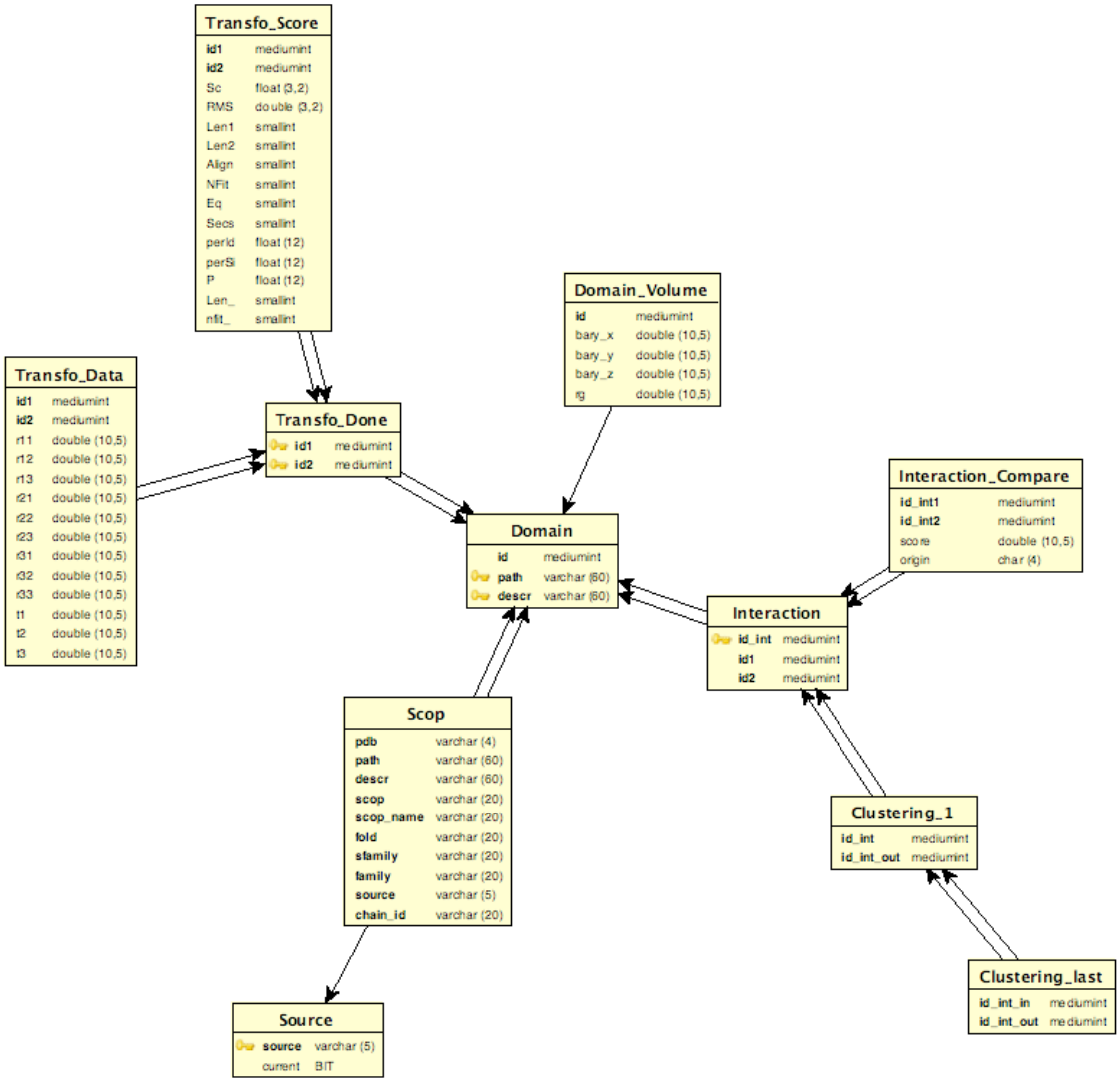

Figure 5: Description of the database

Domain definitions are updated regularly according to the SCOP version (SCOP version 1.71 was released in October 2006, the previous version, SCOP version 1.69, was released in 2005) and with each update, new structures are annotated and some annotations from the previous version are changed, created or removed. As such, this database is robust to any modifications of the domain definitions (SCOP and Domain tables) and preserves the information already computed about domain interactions (Interaction table). Data about superpositions of domains and comparison of protein interactions are stored as well (superposition data are stored

in the Transfo_Done, Transfo_Score and Transfo_Score tables and the comparison of protein interactions is stored in the Interaction_Compare table).

The current version of the database (based on SCOP pre-release from 2006) contains information about 99779 domains, amongst which 95737 interacting pairs are found. There are 3705 different types of domain family to domain family interaction. Some types of interactions are not well represented whereas some others are observed frequently. For instance, the interaction between the alpha chain (b.34.4.3) and the catalytic beta chain (g.36.1.1) of a ferrodoxin thioredoxin reductase is only seen in one structure (PDB code: 1dj7)), whereas the interaction between an antibody variable domain-like (b.1.1.1) and an antibody constant domain-like (b.1.1.2) occurs 1556 times.

After removing repetitions of the same orientation of interacting domains within each structure, 65561 interactions remained. Again, some interaction types are represented many times, whereas some others are unique. Antibody-related domains are seen very often in comparison to other interactions. Consequently, to orient some pairs of domains, few interaction templates, if any, are available, while for other pairs there are many hundreds. The number of templates influences the number of predictions possible for the orientation of a given pair of domains and does not relate to the quality of the prediction: a good template may not be found amongst many interaction templates whereas sometimes the correct orientation is the only possible one.

2.4. Maintenance

In order to keep the database updated with data from new structures, the following steps are followed:

-Update the tables containing the definitions of domain ranges (Scop and Domain tables)

- Search domains that have more than 5 residues within 10Å
- Compare pairwise interactions of the same type
- Update the list of interaction templates with non-redundant interactions in the Interaction table

3. Getting annotated structures for each domain

The composition of protein complexes can be determined by experiments independently of the structures of proteins and subunits contained in it. In our attempt to predict the structure of protein assemblies from individual proteins, models are first determined for each component that lacks a structure and domain types are assigned to components whose structure is known.

3.1. From sequence to structural models

In order to get structures for all the domains whose structure has not been solved the following procedure is applied:

- Search for homologues amongst proteins of known structure and/or protein domains of known structure using a fine-grained profile hidden Markov Model procedure (Soding 2005)
- Upon success, clean and format the alignment to use MODELLER (Sali and Blundell 1993) (as predicting the structure of unaligned residues is prone to error, stretches of residues where more than 5 residues from the sequence of the protein of unknown structure are not aligned are removed from the alignment and residues are renumbered so that they match in the alignment and in the structure)
- Run MODELLER to generate an interaction model

When the alignment between the protein and the potential template is poor (E-value >0.1), the chance of achieving a good model prediction for the whole structure or part of the structure is low and such models are ignored.

We do not expect the accuracy of the models to impact greatly on the accuracy of the procedure, since none of the methods currently used to assess structure quality are drastically affected by typical limitations of homology modeling, such as loops or side-chain orientations. In the future, however, this might become a more critical part of the procedure.

3.2. Assigning domains to a protein

When the domain composition of a protein is not known, several options are considered to locate domains on the protein structure.

3.2.1 Sequence-based

We attempt to align the protein sequences to domain sequences of domains of known structure, as done when predicting homology models. With a good match (E-value<0.001), we can confidently assign a domain to a portion of the protein encompassed in the alignment. Ambiguous cases, where part of a protein matches several distinct domain types, are rare.

3.2.2 Structure-based

Assigning protein domains to a structure can be done using the SSM server (Secondary Structure Matching) (Krissinel and Henrick 2004). It superposes a protein structure to any set of protein structures, in particular, structures of protein domains. If part of the structure of the protein superposes well to the structure of a known protein domain, that part is assumed to be another instance of the domain.

In favorable cases, the structure is known to be homologous to a protein whose structure is solved and whose domain composition is already determined. The assignment of protein domains is then done by superposing the two structures and assigning domains by eye.

4. Program

4.1. The basic search procedure

Given a set of components forming a complex (e.g. as determined by an experimental procedure such as a TAP purification), we obtain all known or predicted structures for all domains in each component as mentioned above. Next suitable templates are searched in the database of interactions and eventually used to predict possible orientations of domain pairs. This is achieved by superposing structures for each pair of separate domains from the complex onto those from each possible template. Finally, those orientations are combined using a graph representation of the problem in an attempt to find the best spatial arrangement of all protein domains.

4.1.1 A reference structure is needed for each domain

Structure prediction is error-prone, especially when the protein shares little homology with a protein of known structure. To account for this, the program uses a reliable SCOP representative structure for each domain when determining conformation is critical.

For instance, when comparing the domain structure to that of a potential template, the two structures are superimposed. When using a model for the protein structure, the superposition score drops because the model is imperfectly predicted and structural changes penalize the superposition score. In such a situation, the SCOP referent is used instead of the model as it represents a real structure that is most similar to the structure of the protein.

4.1.2 Searching for interaction templates

Possible interaction templates to accommodate each pair of domains are searched in the database or directly used when provided. Domains in the SCOP classification are assigned a family, a superfamily and a fold. Each level

corresponds to a degree of structural similarity. When searching for interaction template candidates, candidates amongst interactions of domains with very similar structures (structures with the same family) are searched through first, if not templates are found, the search is extended to structures of interactions with more remote features (structures with the same superfamily).

Different scoring schemas can be applied to evaluate the fit between two domains and an interaction template. Here three scoring schemas were considered. The default schema accounts for the accuracy of the superposition of each putative interacting domain on its counterpart in the interaction template. In the STAMP package (Russell and Barton 1992), the sc score evaluates the quality of fit between two structures. This score ranges from 0 to 10. Above 3, structural similarities between two structures are strong enough to achieve a good fit. Below 3, limited structural similarity is found and the superposition is dubious. When the two domains are superposed on corresponding domains in the interaction template, the worst sc score is kept to score the interaction template.

The similarity in sequence between the proteins in the query and the proteins from the interaction template was also used. In this case, the template is scored by using the worst sequence identity obtained when aligning the sequences of the two domains on the corresponding domains from the interaction template.

Finally, the likelihood, or goodness of fit, for the new interface built when using each template can be scored. Given the structure of an interaction between two domains and two sequence alignments between the sequences of the domains and the query domain sequences, InterPreTS (Aloy & Russell, 2002) uses pair potentials to assess the possibility that the structure accommodates also the two proteins from the query. The evaluation is made by comparing the affinity of the domains in the original structure to the affinity of the domains when the query proteins are threaded onto the template structure.

Whereas the first two scores (sc and sequence identity) evaluate the resemblance between the two query domains and the two domains from the interaction template, in terms of structure and sequence respectively, and disregard the resulting interaction, the InterPreTS score evaluates the quality of the interface generated when an interaction template is used to model the orientation of the two query domains. These three scoring schemas are evaluated using arrangements of three domains.

While for certain pairs of domains, interaction templates are abundant, in some other cases there are few. In order to limit and keep control of the number of combinations generated, the number of interaction templates used for each pair of domains is limited (with a user-defined parameter, default value 3).

Some interaction templates are redundant (in the database, interaction templates are clustered within a structure and not across structures) and consequently less informative. In order to select only relevant interaction templates, they are ordered according to the score chosen, from best to worst, then each interaction template is compared, using iRMSD, to candidates already selected and removed from the set if the it is not sufficiently distinct from those already selected. Thus, a limited best-scoring set of interaction templates is obtained for each pair of interacting domains.

When a potential orientation for a pair of domains has been determined by other means, for example by successful application of docking, it can be added to the set of interaction template candidates. Any structure of interaction can be used to complement the data from the interaction database. Finally, when the orientation between two domains is known, no search is made and the orientation is used directly.

Overall, interaction templates are searched as follow:

Iteration

For each pair of domains:

- Use the known orientation, if any:

OR

- Search potential interaction template candidates in the database or in user-suggested structures
- Score each interaction template using superposition score, sequence identity, and InterPreTS
- Sort interaction templates by means of score
- Keep the x best-scoring distinct interaction templates (the difference between interaction templates being estimated by iRMSD)

4.1.3 Modeling the problem as a graph

Once all potential interaction templates are collected, the problem is modeled as a graph in which nodes represent domains and edges represent interaction templates. Edges are undirected and an edge weight is the score of the corresponding interaction template. Note that not all pairs of nodes are connected by an edge in the graph as there may be no suitable interaction template candidate for a pair of domains. On the contrary, some pairs of vertices may be connected by many edges, as when there are several possible orientations for a pair of domains.

In order to search combinations of interaction templates that can be used to accommodate the domains, all the spanning trees of the graph are explored. Spanning trees are minimum sets of edges that connect all vertices. In this case, they correspond to minimum sets of orientations that can be used to model the structure of the assembly.

Although we translate the problem into a graph, we record the edges in the graph that use the same interaction templates and the edges in the graph taken from the same structure. This information will be used when searching arrangements with specific features.

4.1.4 Solving the spanning tree problem

4.1.4.1 Feasibility

We determine first if it is possible to find a set of edges that connects all vertices of the graph. If not, either only a single subset of vertices can be connected or several subsets of vertices are connected independently. In the former case, we search spanning trees for the set of vertices that can be connected, while vertices that cannot be connected are discarded. In the latter case, the program indicates to the user the different subsets of vertices that can be formed and it proceeds with the search for spanning trees in the largest subset. If the largest subset is not the one of interest to the user, the procedure can be run again with domains of interest. Finally, the procedure stops when no possible connections are available.

4.1.4.2 Estimating the number of solutions

Here we address the problem of the number of possible spanning trees generated for a given problem. A precise calculation is difficult since many spanning trees correspond to bad models that cannot be detected by considering the graph alone. Thus, we evaluate the maximum number of spanning trees found in a given graph.

The real number of possible spanning trees for a graph can be computed by iterations of a deletion-contraction step where, given an edge e, two simpler graphs are produced: one where the edge e is removed, the other where the vertices bound by the edge e are merged. The procedure is time-consuming: it generates

and counts all the spanning trees in the graph. Thus, the method is inappropriate for estimating the number of spanning trees.

We assume that there are n domains to orient and that whenever it is possible to model an interaction between two pairs of domains, there are systematically k interaction templates available. Thus, the corresponding graph is made of n vertices and there are k edges between each pair of vertices which there are templates for.

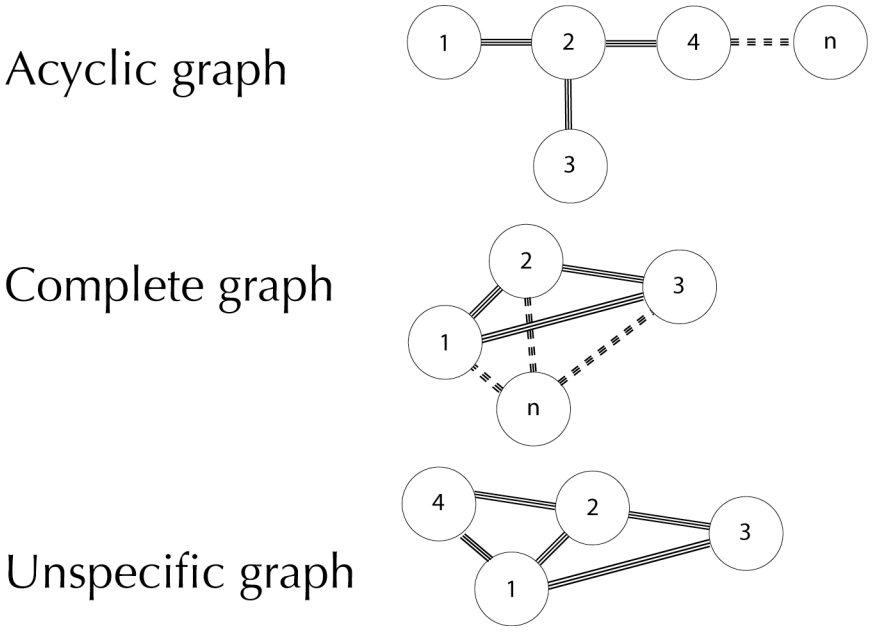

Figure 6: Three types of graph
For an acyclic graph, the number of spanning-tree is 1; for a complete graph, the number of spanning-tree is given by Cayley's formula; for any graph, the number of spanning-tree is attained using the Kirchhoff's theorem

If the graph contains no cycle of any kind (Figure 6 top), the number of spanning trees can be computed directly: $t(G)=k^{n-1}$ While for a complete graph (Figure 6 middle) where the n domains are all directly connected and spanning trees most numerous (for a graph containing n nodes), the number of spanning tree can be estimated by adapting the Cayley's formula: $t(G)=k^{n-1}.n^{n-2}$

When the graph is not acyclic or complete (Figure 6 bottom**),** the number of spanning trees is constrained between those two values. Let G′ be the graph with the same vertices as G and one edge between two vertices, if they are connected in G. The number of spanning trees in G′ is given by the Kirchhoff's theorem:

Let A be the admittance matrix of graph G′, and v_1, …. v_l be the non-null eigenvalues of A, then: $t(G')=(v_1.v_2. \dots v_l)/n$

Finally, $t(G)= k^{n-1}.t(G')$

In this approach, computing the number of spanning trees of the graph amounts to computing eigenvalues in a n by n symmetrical matrix.

The three estimations of the number of possible arrangements increase at least exponentially with the number of vertices. In fact, Inbar *et al*. demonstrated that this problem is NP-hard (Inbar *et al*. 2005) and developed a heuristic to generate solutions to the problem.

Returning to the problem of complex assembly predictions, this complexity means that the number of possible arrangements of domains increases at least exponentially with the number of domains and that a good trade-off must be found between the number of domains in the assembly and the number of possible interaction templates allowed.

4.1.4.3 Algorithm

To search for spanning trees in the graph, an adaptation of Kruskal's algorithm is used. The main difference is that the original algorithm searches the minimum spanning tree of a given graph, while this adaptation searches all possible spanning trees and builds them with the best-scoring set of edges first (Figure 7).

Initialization

- Create a forest F (a set of trees) where each vertex is in a separate tree
- Order edges by score from greatest to least and list them in S

Iteration

If S is empty: (in the case where the set of edges is not explored further)

- Roll back to the previous state of the set of edges and forest, if any, else end the search procedure
- Remove from S the first available edge with maximum score
- Continue the search

Else:

- Pick the edge with maximum score from S
- If the edge connects two distinct trees, remove the two trees and add to the forest the tree resulting from the combination of the two trees
- If not, discard the edge from S

This procedure enables the determination of all possible spanning trees in the graph in an order where spanning trees built with the highest-scoring edges are retrieved first. Moreover, preliminary constructions can be controlled at any step, meaning that the study of the set of edges can be continued or aborted if necessary. Here, the set of edges is translated into a model that is tested and, depending of the validity of the construction, the set of edges is explored further.

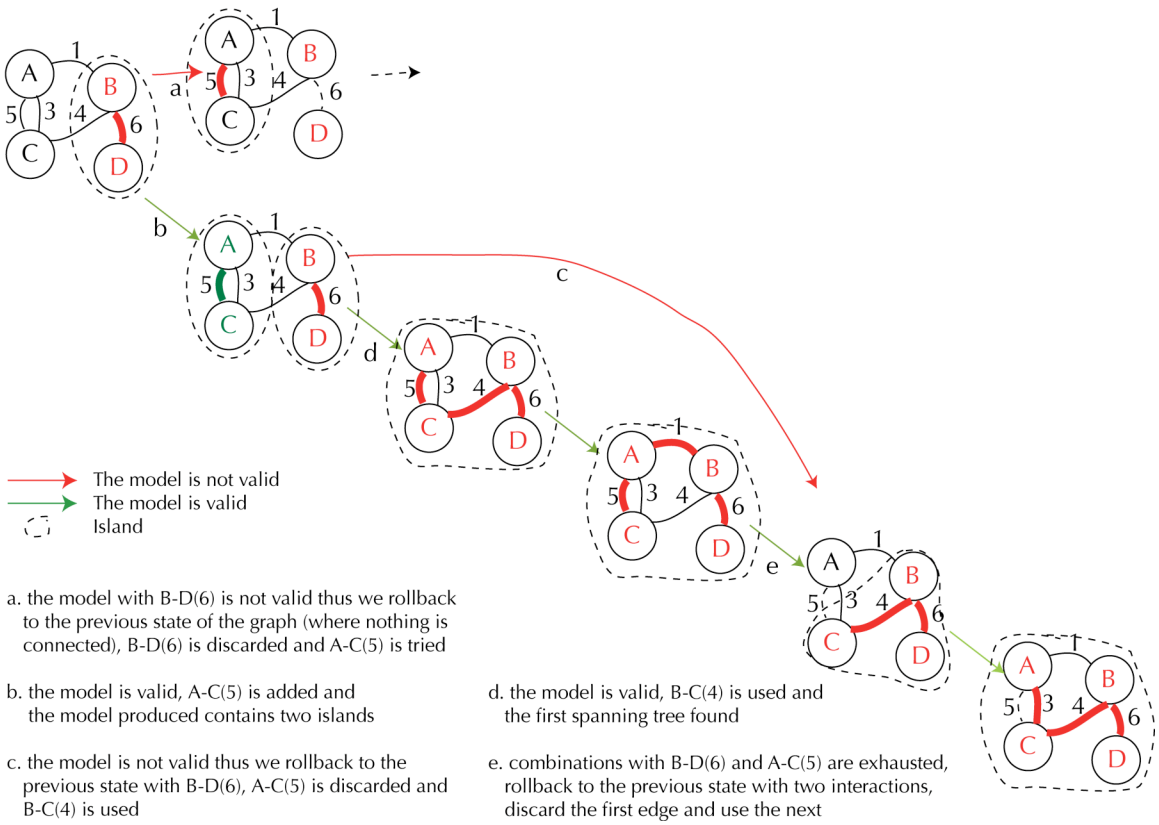

Figure 7: Spanning-tree search procedure

4.1.5 Converting sets of edges into transformations

To translate a set of edges into a spanning tree, the following steps are taken:

Initialization

- Each domain from the query is assigned a transformation that tracks the transformations undergone by the domain whenever an interaction template is used. Each transformation is initialized with the identity transformation (*i.e.* 3x3 identity matrix for the rotation, 3x1 null matrix for the translation).
- Each domain is assigned an island, an island being a set of domains oriented in the same coordinate system. All domains start in separate islands.

Iteration

For each edge in the set:

- Identify the two domains oriented by the transformation
- Apply the transformation represented by the edge to all domains that are in the same island as the domain being oriented by the transformation
- Update islands for the domains that were transformed

Ending

The procedure ends when there are no more edges in the list. Each domain is assigned a transformation to place it into its final orientation and domains oriented in the same coordinate system are listed.

Upon completion of the procedure, there is one island left only when there were enough edges to orient all the domains (*i.e.* n_{domain}-1 edges). When there are several islands, then it is not possible to group all the domain structures in one single structure and each island corresponds to a subset of coherently oriented domains.

4.1.6 Evaluating the predictions

When we search for spanning trees, the first models predicted are the models built using the best scoring interaction templates. As a consequence, the first model is expected to be the most accurate.

It is possible that some interaction templates in the graph mutually exclude each other. For instance, given a domain A and two domains B′ and B′′, it is not possible to use the same interaction template to model A-B′ and A-B′′ as it would result in B′ and B′′ occupying the very same place. So, before proceeding to any test involving the structure of the prediction, the set of edges is checked for validity.

To test the models quickly before proceeding to their combinations, a simple representation of the structure is obtained as follows: each domain is abstracted as

a sphere centered on the center of mass of the domain with a radius that is the radius of gyration of the domain. We apply the transformations to the center of mass of each domain and check that the distance between two centers of mass is higher than 0.2 times the sum of the radii of gyration of the two domains considered. This criterion was determined by studying all the interactions contained in our dataset: 99% of the interactions obeyed this simple rule. With this procedure, we detect, before construction, cases where domains should overlap each other.

Upon success of the quick validation procedure, the transformations computed are applied to the domains in order to build the prediction. It is possible that several separate structures are created due to the fact that the set of edges may not connect all the domains together and, therefore, belong to separate islands. The procedure checks that all the domains in each island of the prediction are interconnected and that there is no obvious close contact (or *bump*) between the domains. Optionally, it computes an InterPreTS score for all the interactions in the structure. If the structure is valid, the set of edges is further explored; if not, it is skipped and the search is continued with the next set of edges.

The validation process and its influence on the spanning tree search are summarized as follows:

- Obtain a set of edges from the spanning tree search procedure
- Check that edges used are compatible
- Check that the spherical abstraction of the structure is valid
- Use the set of edges to assemble the structure
- Check that the structure is valid
- Report to the spanning tree search procedure to continue or skip the study of the set of edges

4.2. Making the best use of prior information

4.2.1 Information about direct contacts

As mentioned previously, various experimental techniques can be used for the determination of direct interactions between proteins (the yeast two-hybrid system, FRET, etc.). Such information can be used to guide the search procedure and limit the combinations explored to those that are the most relevant. More generally, the program can account for any prior information related to the organization of the complex as it restricts the search space to the most accurate predictions.

If two proteins are known to interact directly, the user has two options to constrain the program. In the first, interaction templates that bind the two domains are favored (the score is scaled-up), the program runs normally, predicts all the possible structures that accommodate the set of proteins and checks after each prediction if the structure generated satisfies the constraints. If not, the arrangement is not further explored. Obviously, such constraints can only be checked after the two proteins known to interact are oriented in the same coordinate system (*i.e.* they have to belong to the same island). In the second setup, the user forces the program to bind the two domains directly from the search procedure, *i.e.* all predictions will be made with an interaction template to accommodate the two domains. This constraint of being much more stringent and restrictive increases the speed of the search, as the variety of combinations explored is reduced. The direct connections specified by the user are also checked in the predictions, as in the first procedure.

Sometimes, all the constraints cannot be satisfied at once. For instance, if three domains form a ring 1-2-3-1 and the user forces the program to use interaction templates to connect directly 1-2, 2-3 and 3-1, the graph search procedure will not be able to satisfy the three constraints at once (as two interaction templates are sufficient for the orientation of three domains, three interaction templates will never be used at once in this case). Thus, a routine breaks down user-defined constraints

into attainable constraints. In the example, because of the constraints, 1, 2 and 3 are directly bound. Whenever a set of edges from the graph will somehow arrange 1, 2 and 3, the program checks that two (3 (*i.e.* the number of nodes) – 1) constraints are satisfied ensuring that the constraints are as satisfied as possible.

4.3. Looking for specific features

4.3.1 Structure largely similar to another structure

In a situation where there is a remarkable similarity between the set of domains from the query and the domains in a solved structure, the program finds the structure that is most similar to the query and maps directly the domains from the query to their putative corresponding domains in the solved structure. This is done as follows:

- Describe each structure from the Protein Data Bank as a collection of domains
- Search amongst structures those that contains more than three domains of the same type as the domains from the query
- Map domains from the query onto domains from the candidate structure
- When several domains have the same assignment, report them as being ambiguous
- Compute sequence identity between each domain from the query and the domain it is mapped to from the structure
- Keep the mapping that involved the most domains and where the sequence similarity between the query domains and the domains from the structure are the greatest
- Create the corresponding set of constraints

One domain from the query can sometimes be mapped onto several domains from the solved structure. For instance, if the query contains 2 family-A domains and a

structure contains 3 family-A domains, there are 6 (3x2) possible correspondences to draw from one set to the other. In such a case, we try all possible arrangements and keep the arrangement with the best fit as scored using the sequence identity between two corresponding domains. Domains that were assigned ambiguously are reported.

This procedure enables the quick recognition of similar structures that can accommodate the query domains. However, it comes with several drawbacks: when several mappings have scores in the same range or when one domain is ambiguously assigned, one map from the set of domains to the structure is arbitrarily kept when the others are possibly relevant. So even if this procedure is quick, it is used with caution.

4.3.2 Untangle the search procedure by preprocessing

Many structures in the PDB contain several occurrences of one sub-complex (for instance, the CDK-cyclin complex PDB code: 1g3n contains two copies of a sub-complex composed of three identical chains). All domains from each sub-complex usually adopt the same conformation. For a quick estimation of the number of structures contained in the PDB that form potentially multimeric organizations of sub-complexes, the domains contained in each structure are listed, grouped by type, and we assess if it is possible that such a set of domains forms a multimer of sub-complex structures. More specifically, the greatest common divisor (*i.e.* the largest positive integer that divides both numbers without remainder) of the number of occurrences of each domain is computed, which gives an indication of the number of times a sub-complex can be repeated in the structure. Amongst the structures that contain more than two domains, 15849 have a domain whose domain composition is compatible with such repeated patterns, 3006 do not. Even though, this is a rough estimate, it indicates a clear tendency for such multimeric assemblies.

The disadvantage of using our implementation of the spanning tree search is that it treats all potential sub-complexes in a structure independently (Figure 9). For instance, if there are two distinct interactions between a domain A and a domain B in the complex, the default procedure searches an interaction template to model the interactions between the first (A, B) pair, then independently searches for an interaction template for the second (A, B) pair. However, it seems more relevant to treat the two (A, B) pairs as potential sub-complexes and use the same interaction template to model each of the two interactions. Because repetitions of sub-complexes are very frequent, we implement a method to account for those cases (Figure 8):

Initialization

- When encoding the graph representing the interaction templates, list the edges where the same interaction is used as a template more than once (for several pairs of domains)
- For each such list, search valid combinations of edges: the basic seeds

Iteration

- Search the cases where seeds can be combined in principle
- Combine the seeds from two sets of seeds, keep the valid ones and add them to the set of seeds

Ending

- When possible combinations of seeds are exhausted, rank seeds by number of edges and number of distinct interaction templates

In principle, each seed contains the set of transformations needed to create all the sub-complexes of a structure. Seeds are then used as starting points in the spanning-tree search. Best seeds (those that span over the greatest number of

domains with the minimum number of groups) are used first and the search is stopped when a seed is successfully used.

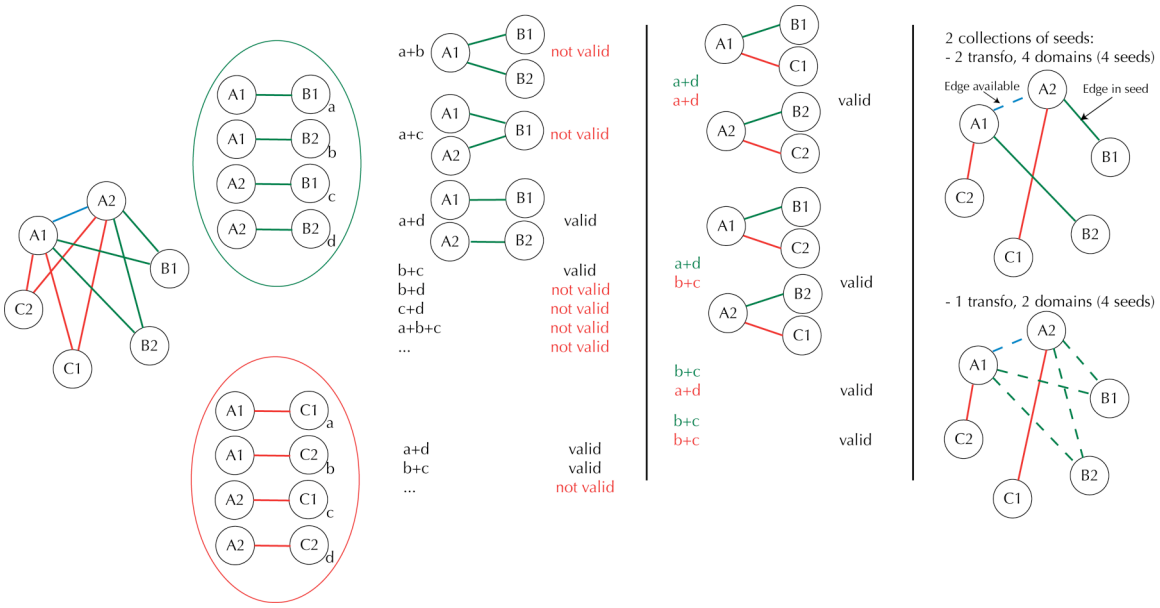

Figure 8: Procedure to search for seeds
Seeds are combinations to beginning the search for spanning-trees with where the same interaction templates is used to model several interactions

Similarly, the program can search for models built using interaction templates from few structures. Then, the same procedure applies, the only difference being that in this situation we use groups of edges representing interaction templates from the same structure file instead of groups of edges representing the same interaction templates as used in the symmetry-search method.

Even if the two seed searches do not account for the same properties, they do not seem easy to combine and they cannot be run simultaneously in the current implementation.

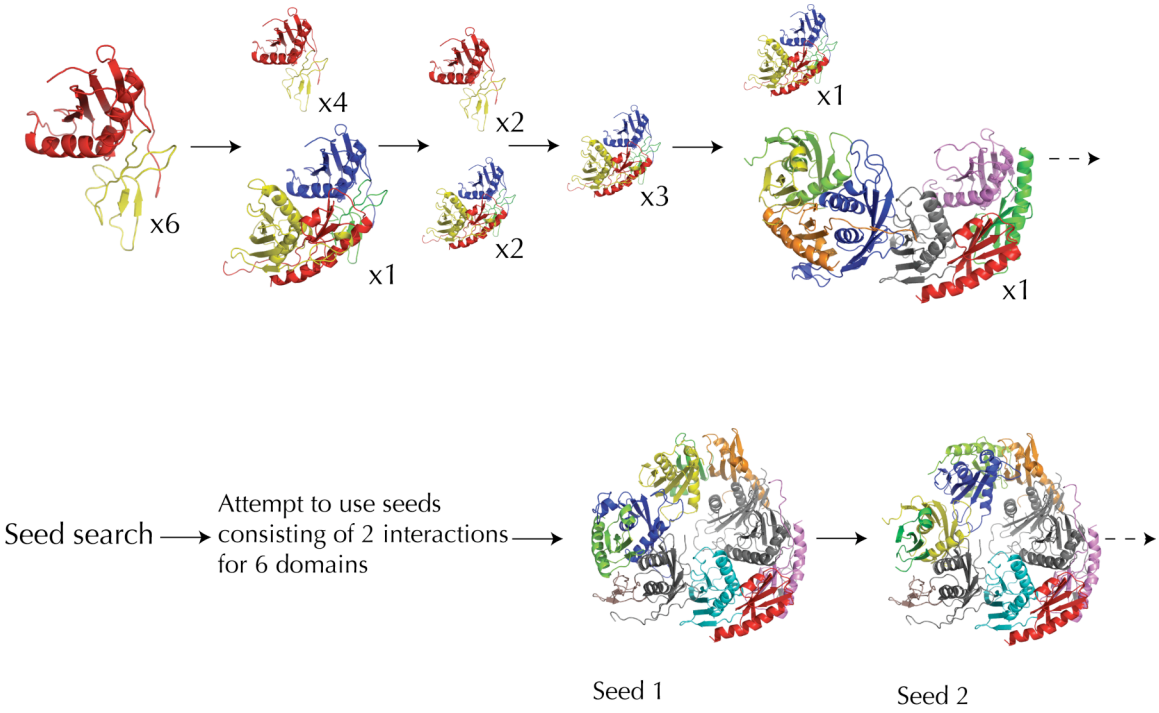

Figure 9: Comparison of the spanning-tree search without and with search for seeds

4.4. Exploring and understanding the predictions

The more domains in the query, the more possible arrangements there are. In order to be able to distinguish very dissimilar arrangements from variations around the same arrangement, we designed a procedure that directly compares the transformations and clusters them. This procedure operates as follows:

- Collect the transformations of all valid predictions
- Compare the predictions one by one by computing iRMSD for each pair of domains and keep the worst iRMSD score to evaluate the difference between two structures
- Cluster predictions that are close enough (by default the cutoff is1Å)
- Collect iRMSD scores in a lower-rectangular matrix
- Perform a single-linkage clustering of models

- Describe the hierarchy of clusters in a tree format (here, Newick format – such tree can be displayed using dedicated programs like iTOL nora.embl.de/ivica/).

With this procedure, information about the degree of similarity between the predictions is obtained. It is important to notice that the evaluation is made on final assemblies and not on the interaction templates used to build them. One could think that comparing interaction templates used to predict a model is sufficient to compare models, though it is not: if the sets of interaction templates used in two predictions are similar, the predictions will also be similar. The inverse assertion does not hold: all similar predictions are not built from similar interaction templates.

To illustrate this point, let us consider a structure in which three domains A, B and C are arranged at the three corners of a triangle. It is possible to predict the assembly with an interaction template for A-B and an interaction template for B-C. It is also possible to arrange the domains with an interaction template for A-B and an interaction template for C-A. The two predictions could be the same, even if the set of interaction templates used are not similar. This shows why it is necessary to use final transformations to compare the predictions and why the study of interaction templates is not enough.

At this stage, all the assemblies have been searched and related transformations have been computed. Models can be created upon request. But providing the structure is not enough and it is necessary to inform the user about the constitutive interaction templates and the resulting structure.

Therefore, for each model output, a file is created with details about the interaction templates used, the score of the templates, other evaluations (sequence identity, individual superposition scores, possibly the InterPreTS score of the resulting

interaction) and information about the structure it comes from. Moreover this file contains information about the resulting prediction that consists of the scores obtained during the evaluation step, *i.e.* a check that all the domains are connected and do not bump each other, the number of connections in the prediction (possibly greater than the number of interaction templates used) and InterPreTS scores for each interaction, if required.

Finally, a procedure can be used to generate a picture describing the assembly process that led to the prediction. The model is oriented by the user and the program generates the pictures of the prediction with each domain in a different color, the pictures of the oriented interaction templates in the context of their original structure with the same color schema, and finally the structure of each domain separately and oriented using the PyMol ray tracing function (DeLano, W.L. The PyMOL Molecular Graphics System (2002) http://www.pymol.org). If the prediction is to be compared to another structure (*e.g.* for benchmarking, where the prediction is compared to the native structure), the first domain of the prediction is used to orient the other structures and domains are colored again with the same color schema.

5. Benchmark sets

To benchmark the method, two datasets are employed: first an abstract and large-scale set where three-domain assemblies are isolated from all the complex structures and the program tries to see how many of those triplets could be predicted from parts using information from other structures. Second and more concretely, complete structures that can be built from parts were used. The predictions were compared to the original structure using a method that compares two structures by comparing all the interactions from one structure to the interactions from the other.

5.1. Comparison of multi-domain structures

This method is used when comparing predictions to native structures during the benchmark and can be used to compare two structures in general to evaluate how similar they are.

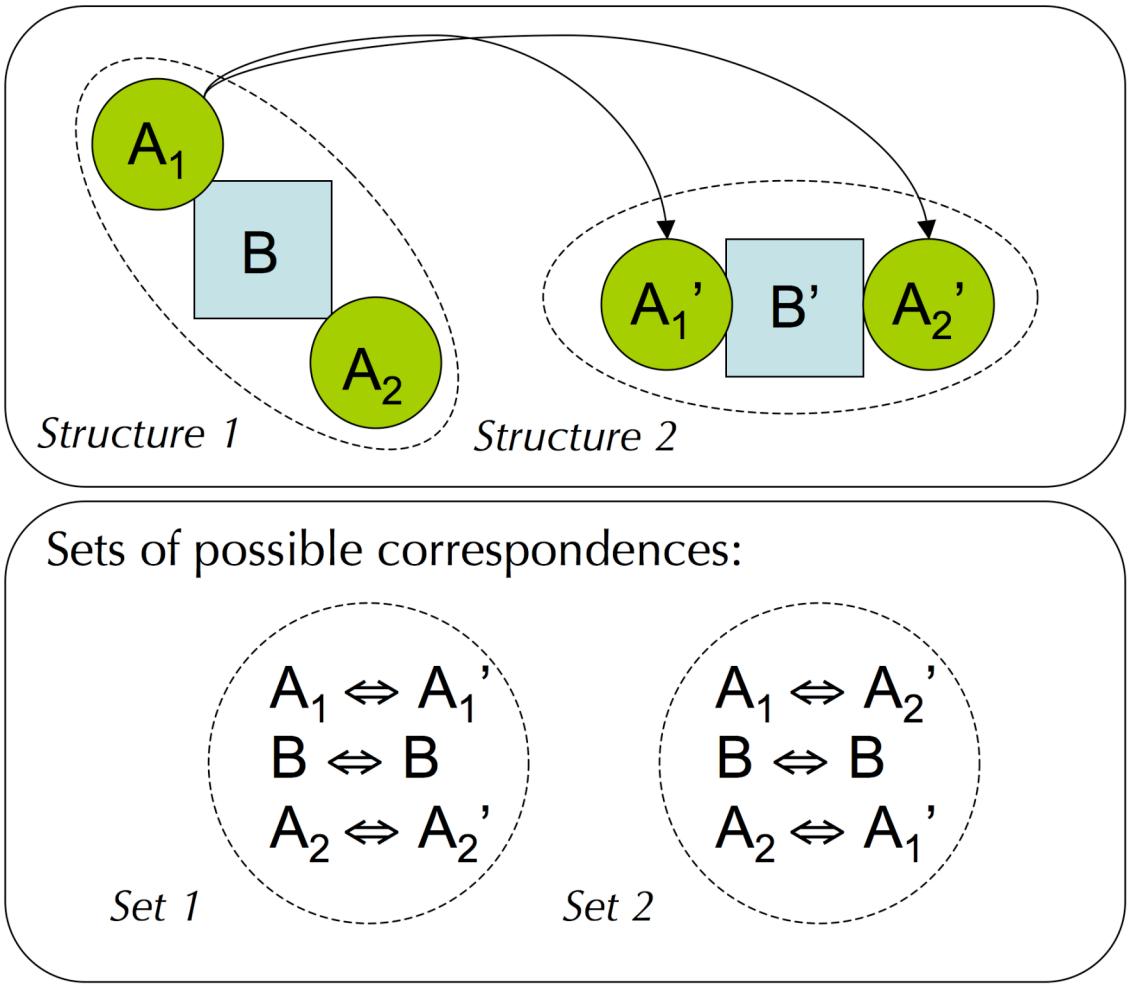

Figure 10: Procedure to compare the structure of two multi-domain assemblies

The method runs as follows:

- List the domains in each structure with their SCOP family and list interactions in each structure

- Search all possible sets of correspondences between domains from one structure and the other (a correspondence is drawn when a domain from structure 1 has the same SCOP classification as a domain from structure 2, they possibly correspond to one another in the two structures) (Figure 10)
- For each set of correspondences, compute iRMSD on all interactions
- Score each set of correspondences by the worst value of iRMSD computed amongst interactions
- Compare all the set of correspondences and keep the one with the least score

Beyond a mere evaluation of the structural similarity between two complex assemblies, the method also determines which domain from one structure corresponds to that from the other.

5.2. Triplets

Assemblies of three interacting domains are the most elementary complexes. Two-domain assemblies are simply interactions. From three domains on, the problems is to combine the correct pair of interaction templates to predict the assembly of the trimer. Interesting triplets are listed as follows:

- Collect all assemblies of three domains from known structures (219166) (Figure 11)
- For each triplet, list domains and keep track of the domain that binds the other two, the 'pivot' (if three domains are interconnected, there are three triplets, each with a different pivot)
- Group triplets by category (*i.e.* the list of the family of each domain and the family of the 'pivot')
- Search structures that can contribute to model an interaction from the triplet in which the pivot is involved

- If there are such potential interaction templates for the two interactions from the triplet involving the pivot, the triplet is added to the benchmark set
- Compare triplets of the same category and group similar triplets

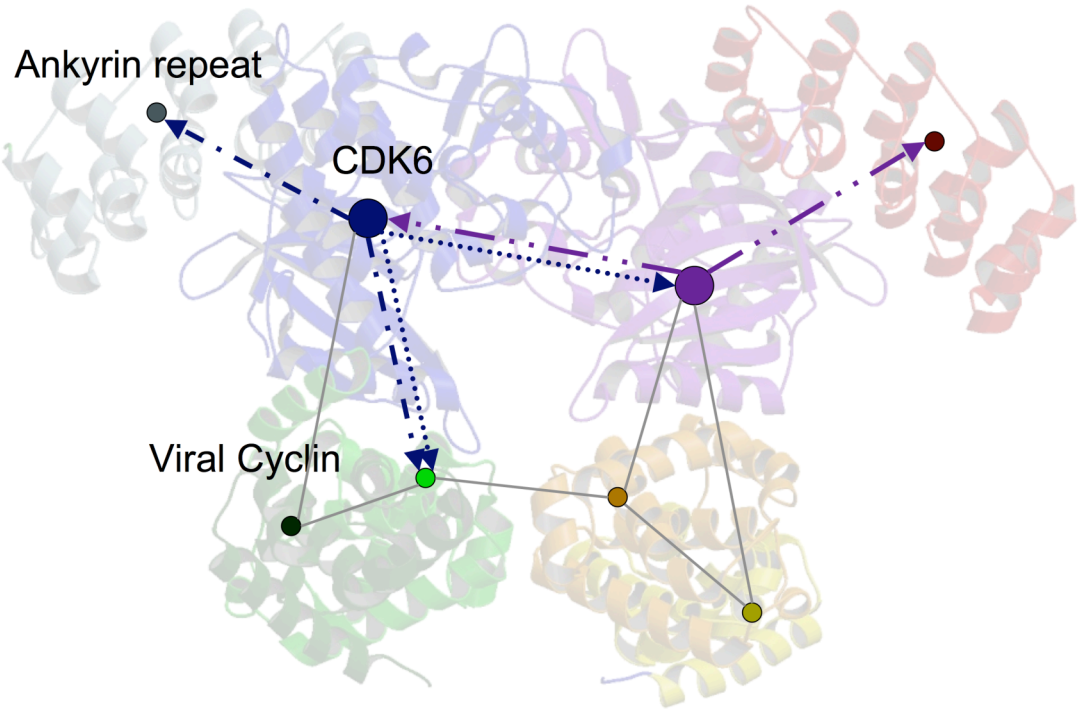

Figure 11: Extraction of triplets from structures
Domains are represented as dots and interactions are represented by lines. Two triplets are shown (dark blue, purple) and larger dots represent the 'pivot' for each triplet, *i.e.* the domains that where the domain common to the two interactions in the triplet.

425 such triplets were found and used to test the procedure. The information from identical triplets, identified by computing iRMSD on triplets of the same category, is masked along with interaction templates that are 90% sequence identical to the domains from the query. We varied several parameters to test their influence on the success of the construction: the lower limit for sequence identity between domains

and domains from the template was varied from 10% to 90%; the lower superposition score limit was varied from 2 to 10; and finally the standard deviation of the InterPreTS score for the structure predicted for the interaction was varied from -40 to 7.5.

Predicting assemblies while varying the three parameters enabled us to draw receiver operator characteristic (ROC) curves for each case. Those plots are broadly used to characterize the performance of a classifier. Here the goal of the procedure is to separate relevant predictions from those that are likely to be incorrect. In such a test, the results can be classified in four categories:

		Reference result	
Test result		Positive	Negative
	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Sensitivity (TP/(TP+FN)) is displayed on the Y-axis of the ROC plot. It represents the capacity of the test to identify the true assemblies as such. On the X axis, 1–specificity (1-TN/(FN+TN)) is displayed. Specificity represents the capacity to reject incorrect assemblies.

In this work, we attempt to detect good models. We consider the outcome of the method to be positive when the prediction is less than 20Å iRMSD distant from the original structure of the triplet. The different categories are counted as follow:

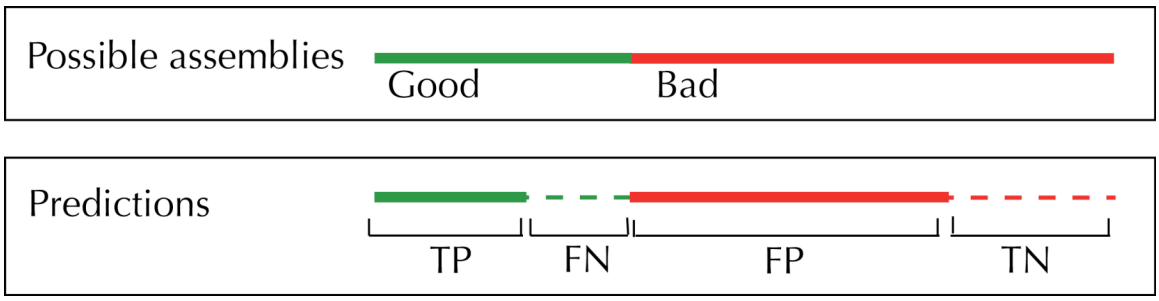

Figure 12: ROC categories assignment
On top, the line represents all the possible assemblies that can be achieved with interaction templates. Below, the assemblies in the range of detection with the parameter value are represented by the thick line and those rejected are represented by the dashed line.

When all triplet cases are considered together, there are a certain number of good and bad assemblies that can be constructed from the interaction templates (Figure 12). For each value of the parameter, we assess how many assemblies of each type are generated and deduce the number of True Positive, True Negative, False Positive and False Negative. Sensitivity and Specificity in the ROC plot.

When given a set of numerous predictions, one may wonder how many good predictions there are amongst them. For a perfect classifier, there are as many good predictions as items in the set of predictions. In contrast, for a bad predictor, there will be only incorrect predictions. The False Discovery Rate (FDR) evaluates this notion. It estimates the enrichment of good predictions within the set of predictions. It is equal to: FDR = (FP/(FP+TP)). We estimate the value of this parameter for different values of the parameters. When only good predictions are made, FDR=0.

5.3. Sets of complexes of known structure that can theoretically be built from pieces

In order to illustrate how the method performs on complete concrete structures, the Protein Data Bank was browsed for a non-redundant set of structures that could potentially be built from parts as follows:

- List interactions in every structure and identify the type of each interaction (*i.e.* the SCOP family of each of the two interacting domains)
- Discard structures that contain only one interaction
- Compare structure interaction types one by one, this reveals four situations:
    - no interaction type in common
    - several interaction types in common
    - one structure contains all interaction types from the other
    - the two structures have the same interaction types
- Discard structures that are included in one another and those identical
- Compare interactions from the 174 remaining structures to interactions from other structures by means of sequence identity
- Keep the structures in which at least 70% of the interactions can be modeled using interaction templates with sequence identity between 30% and 70%

All the possible examples were studied manually and dismissed when the structures could be built with the sole use of interaction templates from a single structure. Nine structures remained after this subjective selection procedure.

6. Potential applications in unsolved complexes

To estimate how applicable the method is on complexes of unknown structure, we considered the complexes found in a large scale complex screen of the Yeast

genome using TAP (Gavin *et al.* 2006). Those complexes are made of three types: 'core' proteins belonging only to the complex, 'attachments' are proteins seen in other complexes as well and 'modules' are sets of proteins present in several complexes. Here complexes were defined by their core, attachment and module components. This definition of complex is the most permissive of all. However in this study, it is relevant because the list of interaction templates is scarce and there may be cases where the interactions between two domains cannot be modeled directly. However, a third protein can sometimes be used to bridge the two domains together; an idea similar to the third-party mediation discussed in (Aloy and Russell 2002a).

The following procedure is used to estimate the proportion of a complex that can be predicted:

- Assign SCOP domains to each protein of each complex
- Search putative interaction templates from known structures
- Compute sequence identity between domains from the complex and domains from the interaction template candidate
- Estimate the ratio of interactions needed to build the complex that can possibly be modeled (ratio between the number of domains that can be oriented together – 1 on the number of interactions needed to orient all the domains ($n_{domains}$ - 1))

The data are represented in a boxplot. A boxplot is an informative way to display a distribution of discrete data. The thick bar in the middle of the box represents the value of the median (*i.e.* the middle value in a list of ranked numbers). The two parallel lines delimiting the box represent the first quartile and third quartile value respectively (the first quartile cuts-off 25% of the data, the third 75%). The InterQuartile Range (IQR) is the difference between the value of the third quartile and the first. Extreme values are values that are higher than the third quartile + 1.5 IQR or less than the first quartile – 1.5 IQR. The short horizontal line indicates the

lowest/highest value that is not extreme. This graph shows how values are distributed: if the first quartile and the median are at 0, it means that for 50% of the samples, the value is 0. Still, some samples may perform better as shown by the third quartile and the points representing the extreme values. The boxplot is more informative than the mere mean and standard deviation values as it captures the repartition of the values and reveals extreme cases. An example of these plots can be found in Figure 19.

Results

1. Evaluation of the procedure – Benchmark

1.1. Results from the triplet dataset

Three domain assemblies are the minimal units to test the procedure, as two domains form only a single interaction that is not sufficient to test the assembly procedure. We devised a simple benchmark consisting of the 425 distinct triplets that can presumably be predicted using information from other structures out of the 219166 arrangements of three interacting domains (Methods). We then tested the approach using only templates lacking very close sequence similarity (sequence identity <= 90%). We varied 3 parameters in order to test the ability of the method to retrieve the right arrangement in different setups: two parameters account for the similarity between the domains from the query and the domains in the template (sequence identity and superposition quality), while the last parameter estimates the likelihood of the interaction surface formed (InterPreTS score (Aloy and Russell 2003)) This reflects the real situation when, given a set of domains, there is a limited set of interaction templates from which to derive orientations, and we want to estimate the predicted model. A good model is a model for which the worst predicted interaction is less than 20Å iRMSD different from the original structure (at around 10Å or less the similarity between two interactions can be seen by eye). The results of the prediction ability of each parameter are summed up in ROC plots (Figure 13).

For low sequence identity, the method is sensitive but poorly specific (e.g. sequence identity: 15%, sensitivity: 97% and specificity: 12% (1-0.88)). In this setting, few interaction templates are filtered out (sequence identity has to be more than 15%) and most of them are tried in the assembly process. In this case, the chance of assembling a good model is high (high sensitivity) but many wrong

models are produced (low specificity). On the contrary, when the sequence identity cutoff is high, only templates with high sequence identity are used, the method is highly specific but not sensitive (e.g. sequence identity: 70%, sensitivity: 2% and specificity: 90% (100-10)). In this setting, only the very best interaction templates are selected, so few predictions are made and many interactions are rejected (low sensitivity) but the method is specific (the few interaction templates used are good). Using sequence identity as a criterion to select good interaction templates is relevant but not sufficient (the ROC plot is distant from the ideal curve close to the top-left corner).

Similarly, the structural similarity between the domains from the query and the domains in the interaction templates is estimated and used as a score (sc score developed in STAMP (Russell and Barton 1992)). The sc score ranges from 0 to 10, 10 being a perfect superposition. Above 3, the structural similarities are good enough that an accurate superposition can be achieved. In this analysis, the cutoff for the sc score was varied from 2 to 10 with 0.4 increments. Results are similar to those obtained with sequence identity and actually the two evaluation methods perform very similarly. Again, the ROC curve shows the impact of different values of the parameter on the success of the method. The better the structural similarity, the fewer and more accurate predictions.

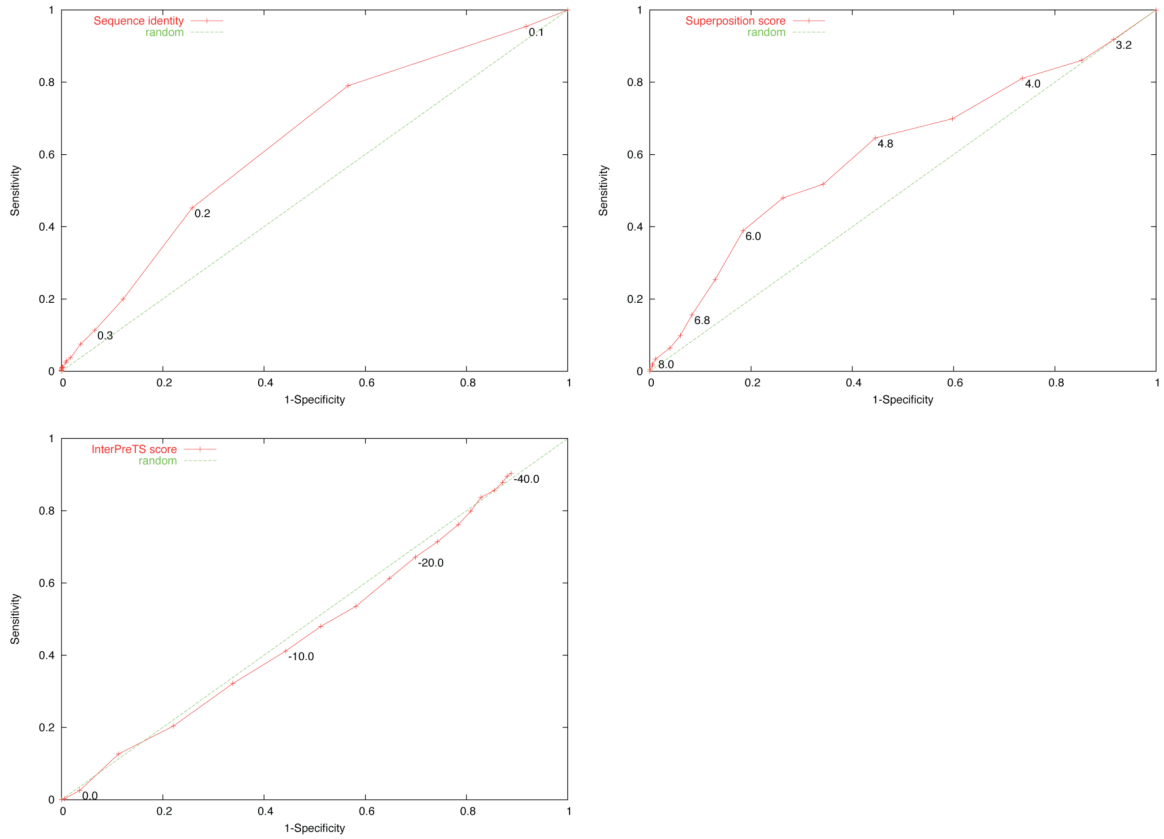

Figure 13: ROC plots using three different parameters to assess the quality of predictions

The three parameters used to assess the quality of the models are:

1. The sequence identity between the domains from the query and the domains in the interaction templates (upper left)
2. The structural similarity between the domains from the query and the domains in the template ((Russell and Barton 1992)) (upper right)
3. The likelihood of the interface assessed by InterPreTS ((Aloy and Russell 2003)) (bottom)

The last parameter estimates the confidence in the resulting interface as computed by InterPreTS. This study shows that assemblies built using interaction templates selected by the InterPreTS score of the resulting interface are not better than predictions made by picking interaction templates at random. The results are often deceiving, and the way it is currently used and implemented, InterPreTS does not provide any useful information regarding interaction templates to use for better predictions.

parameter	value	FDR
InterPreTS	-40	0.86
InterPreTS	-37.5	0.86
InterPreTS	-35	0.86
InterPreTS	-32.5	0.86
InterPreTS	-30	0.86
InterPreTS	-27.5	0.86
InterPreTS	-25	0.87
InterPreTS	-22.5	0.87
InterPreTS	-20	0.87
InterPreTS	-17.5	0.87
InterPreTS	-15	0.87
InterPreTS	-12.5	0.87
InterPreTS	-10	0.87
InterPreTS	-7.5	0.87
InterPreTS	-5	0.87
InterPreTS	-2.5	0.85
InterPreTS	0	0.90
InterPreTS	2.5	0.93
InterPreTS	5	0.88
InterPreTS	7.5	1.00

param.	value	FDR
Seq.Id.	0	0.86
Seq.Id.	0.05	0.86
Seq.Id.	0.1	0.86
Seq.Id.	0.15	0.82
Seq.Id.	0.2	0.78
Seq.Id.	0.25	0.79
Seq.Id.	0.3	0.78
Seq.Id.	0.35	0.76
Seq.Id.	0.4	0.75
Seq.Id.	0.45	0.68
Seq.Id.	0.5	0.69
Seq.Id.	0.55	0.53
Seq.Id.	0.6	0.48
Seq.Id.	0.65	0.32
Seq.Id.	0.7	0.32
Seq.Id.	0.75	0.10
Seq.Id.	0.8	0.08
Seq.Id.	0.85	0.08

param.	value	FDR
Superposition	2	0.86
Superposition	2.4	0.86
Superposition	2.8	0.86
Superposition	3.2	0.86
Superposition	3.6	0.86
Superposition	4	0.85
Superposition	4.4	0.84
Superposition	4.8	0.81
Superposition	5.2	0.81
Superposition	5.6	0.77
Superposition	6	0.75
Superposition	6.4	0.76
Superposition	6.8	0.77
Superposition	7.2	0.79
Superposition	7.6	0.79
Superposition	8	0.68
Superposition	8.4	0.63
Superposition	8.8	0.66
Superposition	9.2	0.51

Table 1: False Discovery Rate for different cut-off values of the three parameters

The ROC plot is the typical means to assess the performance of a predictor. Here we evaluated the False Discovery Rate (FDR) (Table 1) as a means to evaluate how the set of predictions can become enriched in 'good' predictions with different values of a parameter. If all the predictions are bad, FDR is 1 while FDR is 0 if all predictions are good.

The FDR for InterPreTS is high and almost constant (FDR=0.8), meaning that more stringent InterPreTS score cutoffs do not contribute to enrich the set of predictions with good predictions. However, using the superposition score (sc) has an impact on the FDR: for the most stringent value (sc=9.2) the FDR is 0.5 whereas for less

stringent values of Sc, the FDR is above 0.80. Thus, the use of more stringent sc score cutoffs helps to enrich the set of predictions with good predictions. Finally, when the cutoff for Sequence Identity is raised above 50%, the enrichment in good predictions increases very rapidly until FDR reaches 0.01 (*i.e.* there are on average 9 good predictions out of 10 predictions) for sequence identity 85%. Thus, Sequence Identity is the most efficient parameter to increase the ratio of good predictions.

With the ROC plots and the FDR values, it is possible to tune the parameters used by the program to obtain sets of predictions with specific characteristics: we decided not to use InterPreTS to evaluate interaction templates as the version used does not seem help making good predictions. Instead, the superposition score was used to estimate and rank interaction templates. For each analysis, if interaction templates with good superposition scores are available then we keep only these high-scoring templates and expect good predictions to be made, while for cases where interaction templates are scarce, we may allow the use of interaction templates with more structural differences and subsequently evaluate the constructions individually. Obviously, these parameters can be changed to satisfy specific requirements.

1.2. Evaluation of known complexes that can presumably be built from pieces

Maximal structures are defined as those that contain more than three domains reported in SCOP and that are not included in any structure when structures are abstracted to the list of interaction types they contain (Methods). Amongst a list of 55 maximal structures, we searched for those that can be reproduced with a clear and detectable fidelity (sequence identity between 30% and 70%) using interaction templates from other structures.

In some cases, no interaction template could be found within the range of sequence identity to reproduce the structure or, alternatively, all the interaction templates came from the same structure (thus, there is no combination of templates from different structures that can be used to assemble models and direct mapping of domains on domains of the similar structure suffices). For instance, there are 10 domains in four protein chains in the structure of the flavocytochrome C sulfide dehydrogenase (PDB code: 1fcd). They arrange themselves in a dimer of dimers. Without any filter, there are interaction templates to accommodate 6, 2 and 2 domains in separate structures with interaction templates from one structure for each group (three 'islands'). If we filter out interaction templates with the sequence identity criterium, there is no interaction template left for the prediction.

In other cases, part of the structures can be modeled on one existing structure and a few interaction templates could theoretically be used to complement the trivial structure. If the procedure fails to use those interaction templates (*e.g.* it is not possible to superpose efficiently one domain to its template), we are left with a trivial prediction. The structure of G-protein receptor kinase 2 with Galpha-q and Gbetagamma subunits (PDB code: 2bcj) is made of 7 domains. There are 24 structures that can be used to model the interaction between the transducin alpha-subunit and the G-protein domains but there is only one (the structure of the complex between G protein-couples receptor kinase 2 and G protein beta 1 and gamma 2 subunits, PDB code: 1omw) that helps to accommodate the PH-domain of the kinase with the WD40-repeat domain from the transducin. If the latter template cannot be used (e.g. when the domains cannot be superposed to the domains from the templates), the example becomes trivial and is discarded from the benchmark set because one structure template is enough to accommodate the structures of the subunits.

Description	code	No. tmpl. used	No. int. <=10Å	No. int. >10Å	greatest iRMSD original Vs predicted
Gelatinase A	1ck7	5	4	2	26.9
Gelatin binding domain of Fibronectin	1e88	2	0	2	17.8
Elongation factor EF-Tu/EF-Ts	1efu	11	10	5	47.6
Tissue facor + coagulation factor VIIa	1fak	5	2	3	41.6
CDK6/Cyclin/INK4	1g3n	7	5	3	20.3
POU/HMG/DNA	1gt0	2	0	2	33.2
Bovine factor Xa	1kig	2	2	0	5.3
Blood coagulation factor Xa + Ecotin	1p0s	7	7	0	6.7
G-protein coupled receptor Kinase 2 + Galpha-Q and Gbetagamma subunits	2bcj	3	1	2	70.7

Table 2: Results obtained when assembling the structures of nine known complexes using non-trivial templates

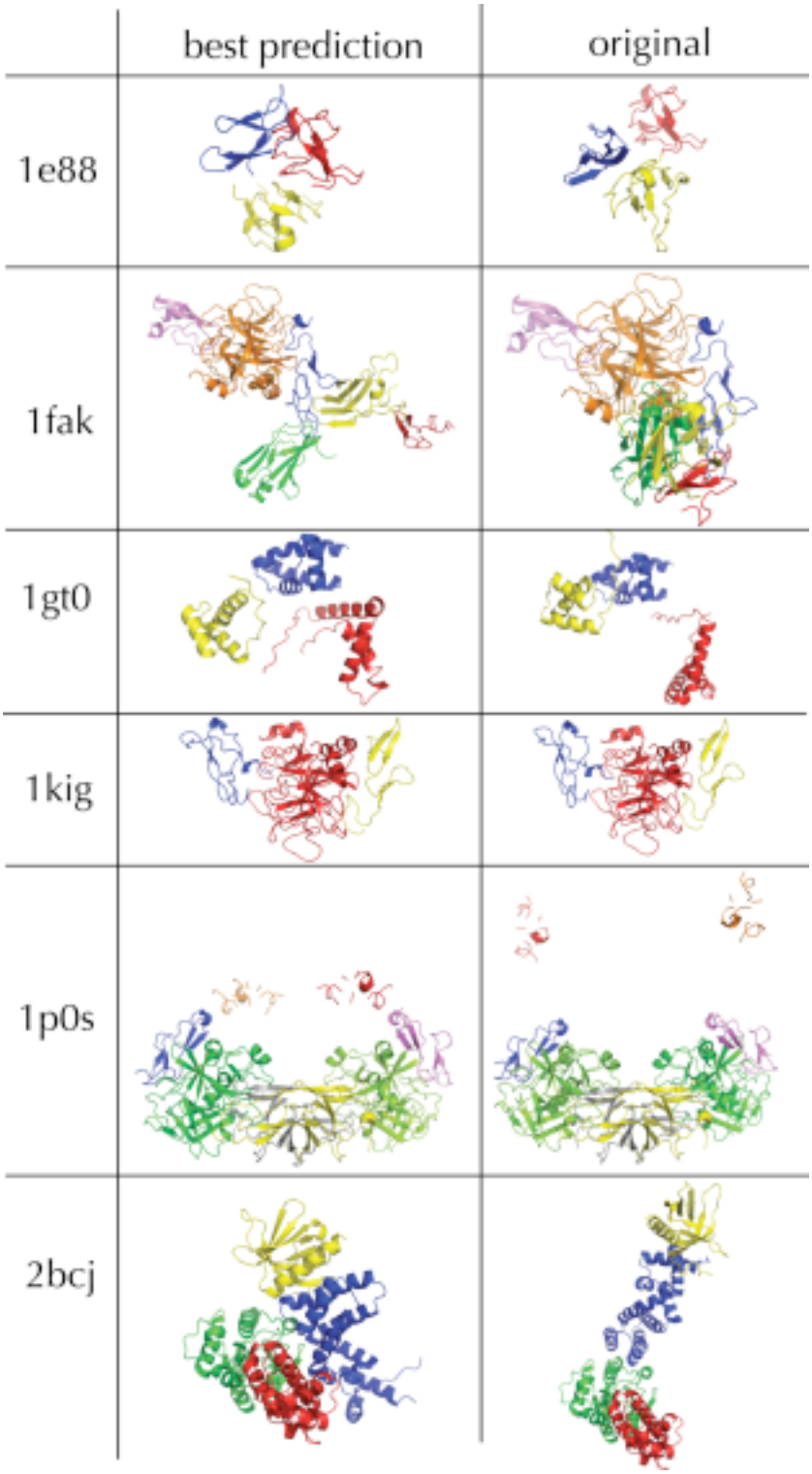

Figure 14: Results on six of the nine constructions

Each domain is colored differently. In each case, one structure is superposed in the prediction and in the original structure to make the comparison easier.

There were 9 complexes that matched our selection criteria. These structures are listed in (Table 3) together with details of how the predictions fared. In the 9 cases, we were able to predict structures that accommodate most domains of the original structure. In most cases, two structures provided enough interaction templates to build theoretically a prediction. However using the interaction templates contained in two structures did not necessarily lead to the best assembly.

The results are shown briefly in a gallery (Figure 14). Some examples are studied in more details because they show how the method performs with interactions of different kinds: intra- chain interactions (gelatinase A – PDB code: 1ck7), a dimer of dimers (PDB code: 1efu) and a dimer of trimers (e.g. CDK6/Cyclin D/INK4 – PDB code: 1g3n).

1.3. Multidomain polypeptide chain: Gelatinase A

Gelatinase A is an extra-cellular matrix metalloproteinase (MMP). It degrades type IV collagen (a component of basement membranes) and denatured collagen. The structure of gelatinase A was solved by X-ray crystallography at 2.8Å (Morgunova *et al.* 1999).

Gelatinase A is a single protein chain made of 6 domains: a MMP N-terminal domain, a MMP catalytic domain split in two parts, three Fibronectin type II domains, and a Hemopexin-like domain (beta-propeller). This simple example illustrates how the method performs on multi-domain chains.

Gelatinase B is used for the orientation of 5 of the 6 domains of Gelatinase A (Figure 15). Domains from one structure are very similar in sequence to domains from the other (sequence identity: 66%, 62%, 59%, 54%, 39%) and the two structures are 1.85Å RMSD apart. The missing Hemopexin-like domain is modeled in the structure with a template from the structure of proMMP-1 (RMSD: 1.5Å,

sequence identity: 51%, 38%). The worst interaction modeled in our prediction is 26.9Å iRMSD distant from the original interaction.

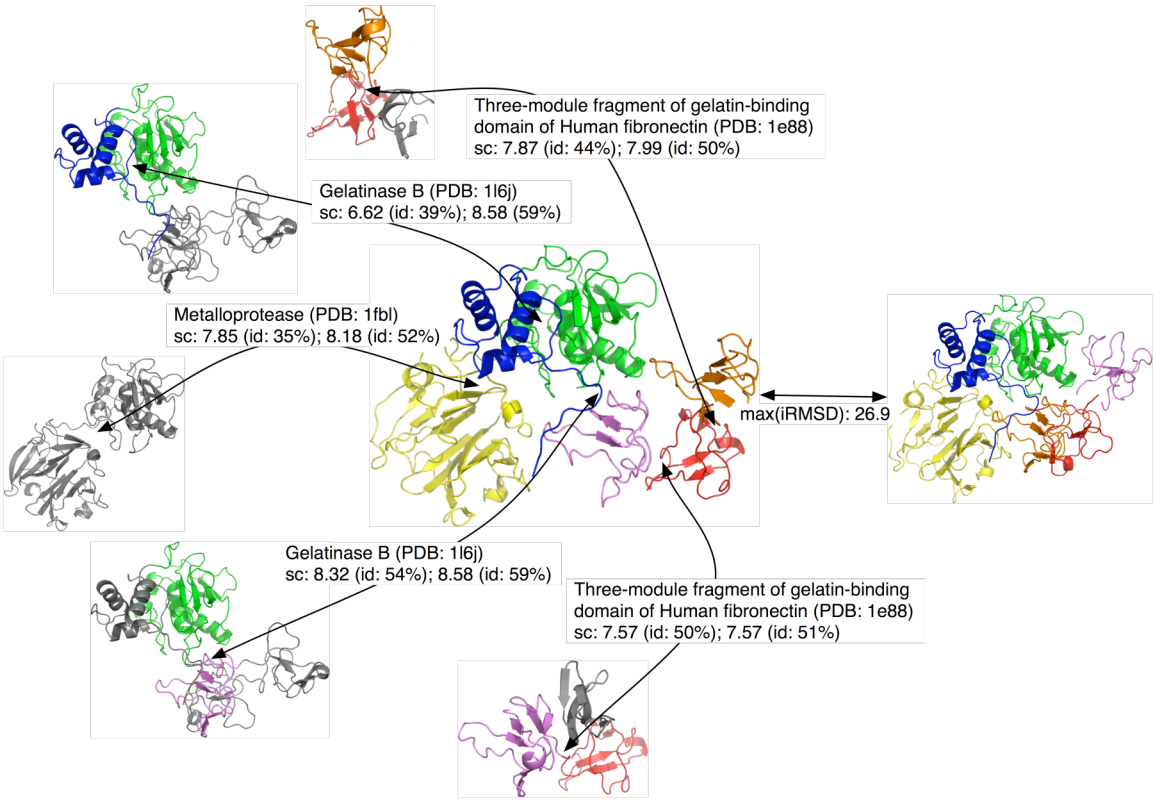

Figure 15: The reconstruction of Gelatinase A
On the left, the interaction templates used for the reconstruction; in the middle, the structure predicted; on the right, the structure of the original complex.

Assuming domains are in contact within the chain, the procedure can model the structure of multi-domain proteins.

1.4. Dimerisation: EF-Tu/EF-Ts

EF-Tu is a G protein (guanine-nucleotide-binding protein) and is involved in a wide range of metabolic processes. EF-Ts, a guanine-nucleotide exchange factor recycles inactive EF-Tu-GDP in active EF-Tu-GTP complex. The structure of the EF-Tu/EF-Ts complex was solved at 2.5Å resolution (Kawashima *et al.* 1996).

The structure is a dimer of sub-complexes and each sub-complex contains 6 domains. This example shows how the procedure deals with multimers of multimeric structures.

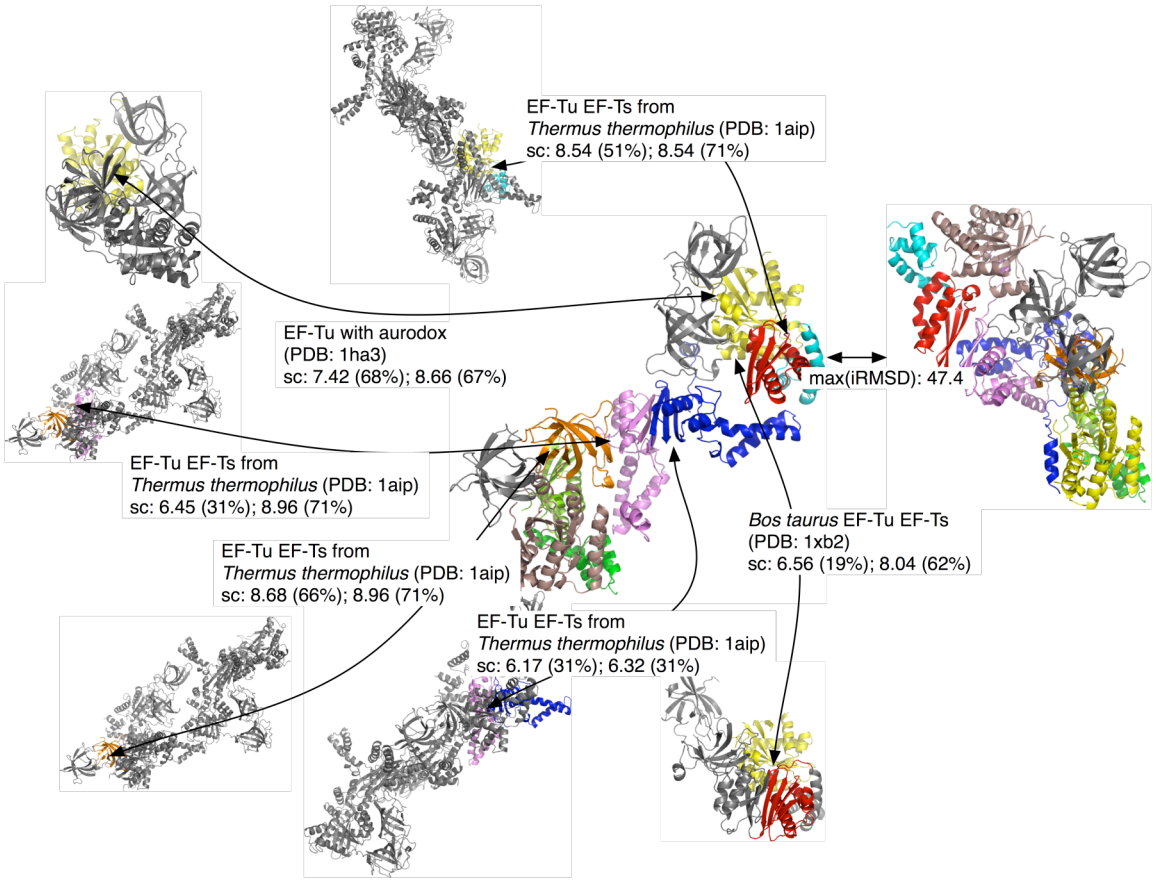

Figure 16: The reconstruction of EF-Tu/EF-Ts
On the left, the interaction templates used for the reconstruction; in the middle, the structure predicted, on the right, the structure of the original complex.

The prediction that is the closest to the original structure and contains no bumps is 47.6Å iRMSD distant from the native structure (Figure 16). Each sub-complex is built in parallel using interaction templates from the structure of three different complexes: the complex EF-Tu EF-Ts from *Thermus thermophilus* (PDB code: 1aip, sequence identity: 52%, 71%, 71%, 31%, 66%, 71%), the mitochondrial factor Tu/Ts complex from *Bos taurus* (62%, 19%) and elongation factor Tu in complex with aurodox in *T. thermophilus* (PDB code: 1ha3, sequence identity: 67%, 68%). Once the two sub-complexes are built, they are assembled using another template

from the structure of EF-Tu EF-Ts from *T. thermophilus* (sequence identity: 31%, 31%). The dimerisation in the prediction occurs via the elongation factor's Ts (EF-Ts) dimerisation domain. However, in the real structure the dimerisation is made by the EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain, explaining the difference between the predictions obtained and the original. We searched for predictions where the dimerisation structure resembles the original. The first complete assembly that we found uses an interaction template from elongation factor TU in complex with aurodox (sequence identity: 71%) to orientate the two EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domains. It is 13.8Å iRMSD from the original structures and was rejected in the first place for containing bumping domains.

1.5. Creation of interactions not in original structure: CDK6/cyclin D/INK4 complex

Cyclins bind and activate cyclin dependent kinases (CDKs). There are also a number of other molecules affecting CDK function, including the inhibitor INK4 (Review (Sherr and Roberts 1999)). A ternary complex of CDK6, the INK4 inhibitor and a viral D-type cyclin was solved by X-ray crystallography (Jeffrey *et al.* 2000). The structure assembles in a dimer of trimers and the two substructures interact at the level of the kinase domains.

We can assemble the 8 domains of the structure and obtain a prediction that is 20.3Å iRMSD distant from the original structure (Figure 17). Each subunit is accurately predicted (iRMSD 14.0Å). The structure of the subunit is predicted using interaction templates from 3 structures: the structure of an INK4-inhibited cyclin-dependent kinase (PDB code: 1bi8, sequence identity: 90%, 45%), the structure of CDK6 in complex with a flavonol inhibitor (PDB code: 1xo2, sequence identity: 26%; 96%) and the structure of the viral cylin from *Herpesvirus saimiri* (PDB code: 1bu2, sequence identity: 25%, 37%). Finally, the two sub-complexes are

assembled via the CDK domains using a template from glycogen synthase kinase 3 beta (PDB code: 1h8f, sequence identity: 25%).

When no information is provided about the direct contacts between chains, the best prediction is ranked 13 amongst the complete predictions (*i.e.* predictions that contains all domains from the query) that do not contain any bumps, and it is the 268th node explored during the search procedure. However, the rank of the best prediction can be improved by adding constraints to the procedure and binding the two subunits via the two CDKs. In this context, the same prediction ranks 3rd amongst complete predictions with no bumps and is the 6th combination considered during the graph exploration.

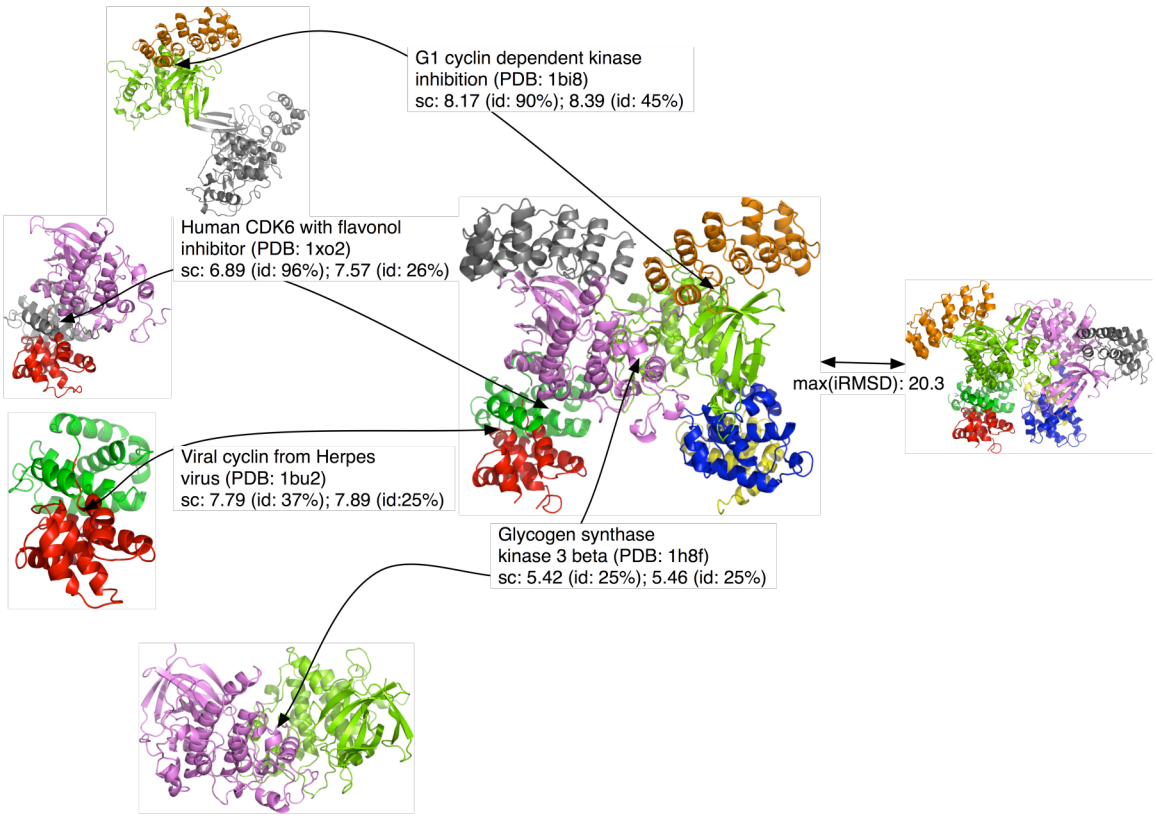

Figure 17: Reconstruction of the CDK6/cyclin D/INK4 complex
On the left, the interaction templates used for the reconstruction; in the middle, the structure predicted; on the right, the structure of the original complex.

These three cases showed firstly that the procedure works alike on inter- and intra-chain domains as long as they are in direct contact, secondly that structures with higher order of organization (multimer of multimers) can be accurately constructed, thirdly that accurate structures are sometimes rejected because they fail some validation tests by narrow margins, and finally that constraints can increase the speed and accuracy of the prediction.

1.6. Highly symmetrical structures

1.6.1 Eukaryotic exosome

The exosome is a protein complex involved in the degradation of mRNAs. The structure of the archaeal exosome core (Lorentzen *et al*. 2005) was first solved by X-ray crystallography and recently the structure of the human exosome was determned by Liu *et al*. (Liu *et al*. 2006).

Before the structure of the exosome was solved, our group attempted to predict its structure (Aloy *et al*. 2002) using the best template available at the time (the bacterial PNPase complex), negative stain EM and a battery of computational procedures (which including InterPReTs and methods of active site prediction). This study met with mixed results when compared to later two-hybrid data (Raijmakers *et al*. 2002), and most recently to the crystal structure of the human exosome (Liu *et al*. 2006). Although the overall model was broadly correct in shape, specific details of the assembly were not predicted correctly as later revealed by experiments using the yeast two-hybrid system (Raijmakers *et al*. 2002) and mass spectrometry (Hernandez *et al*. 2006).

The exosome ring is composed of 6 proteins of two different kinds (with three proteins each) and simple combinatorics shows that there are 120 possibilities to place 6 proteins in a ring.

The only interactions in the structure are interactions between two proteins of distinct kinds and there are two different orientations for such interactions that have been determined previously. Assuming that the proteins are evenly arranged in an alternation of proteins of each type, 24 possible arrangements of the proteins are left. Moreover, if two proteins are known to interact directly, there are 16 possibilities and if two such direct interactions are known (*i.e.* involving three of four proteins), there are 4 or 6 possible arrangements.

Thus, the addition of constraints untangles drastically the number of possible arrangements covered. On the other hand, using bad constraints will ensure the production of bad predictions. In the work of Aloy *et al.* (Aloy *et al.* 2002), the alignments produced did not capture the real separation of the 6 proteins in two classes and from then on it was impossible to generate the right assembly.

The exosome is a valuable case to comprehend various aspects of the method: the final structure being a ring, is it possible, using a method that arranges two structures, to retrieve this higher-level of organization? The ring of the exosome is built with a succession of subunits of similar structures. Do we retrieve all the possible combinations given constraints? Can we separate good constructions from bad? Does the method perform the same with the structures of the subunits and with models?

When combining binary interaction templates with no memory of those used in the previous steps of the construction, we predict several buckled-up assemblies regularly (*i.e.* in the models we predict, the interactions between proteins of the same type are not always identical). We used the protocol to search specifically for regular assemblies (*i.e.* structures built with several uses of the same interaction template – structures in which a sub-complex is repeated as in the case of 1efu (Elongation factor EF-Tu/EF-TS) and 1g3n (CDK6/Cyclin/INK4) described earlier). In a first test, the six real subunit structures from Human exosome were assembled

into models based on multiple usages of two interaction templates from PNPase (PDB code: 1e3h) for predicting the orientation of the chains. The program was forced to directly bind Rrp45 with Ski6 and Rrp42 with Mtr3 as it was known from yeast two-hybrid experiments at the time of the prediction made by Aloy *et al*. With the default parameters (3 interaction template candidates per interaction to model), the procedure generated 32 models of which 24 were built with interaction templates that score remarkably better (sc score of 9 vs 5). Those 24 models corresponded to all the possible regular ring arrangements of 6 structures of two different sorts. However, amongst all these predictions, we could not detect the native-like arrangement because we did not have any good mean to distinguish the good interfaces from the false ones.

We then tried to model the exosome ring using structural models predicted for each protein. The alignment method to find candidates for the modeling of the subunits does not separate correctly the 6 proteins into the two classes observed in the real structure (Ski6/Mtr3/Rrp46 and Rrp42/ Rrp43/Rrp45). Instead we obtained three groups derived from the SCOP domains aligned: (Rrp42, Ski6), (Rrp45, Mtr3) and (Rrp43, Rrp46). With the models derived from these alignments, we were able to accommodate only four of the six structures together.

Sequence alignment reveals that three proteins bound to the ring (Rrp4, Csl4 and Rrp40) contained a Cold shock DNA-binding-like domain. We tried to add these three structures to one of the 24 ring structures predicted (Figure 18). With the default parameters, we were not able to add the three Cold shock-like (*i.e.* RNA binding) domains to the structure of the ring. However, when lowering the requirements for the superposition score, we generated 120 predictions. The cold-shock DNA-binding-like domains can be orientated relative to the ribonuclease PH domain 1-like domain of each protein in the ring. Out of 120 predictions, only 12 exposed the cold-shock DNA-binding-like domains on the same side of the ring and corresponded to all permutations possible.

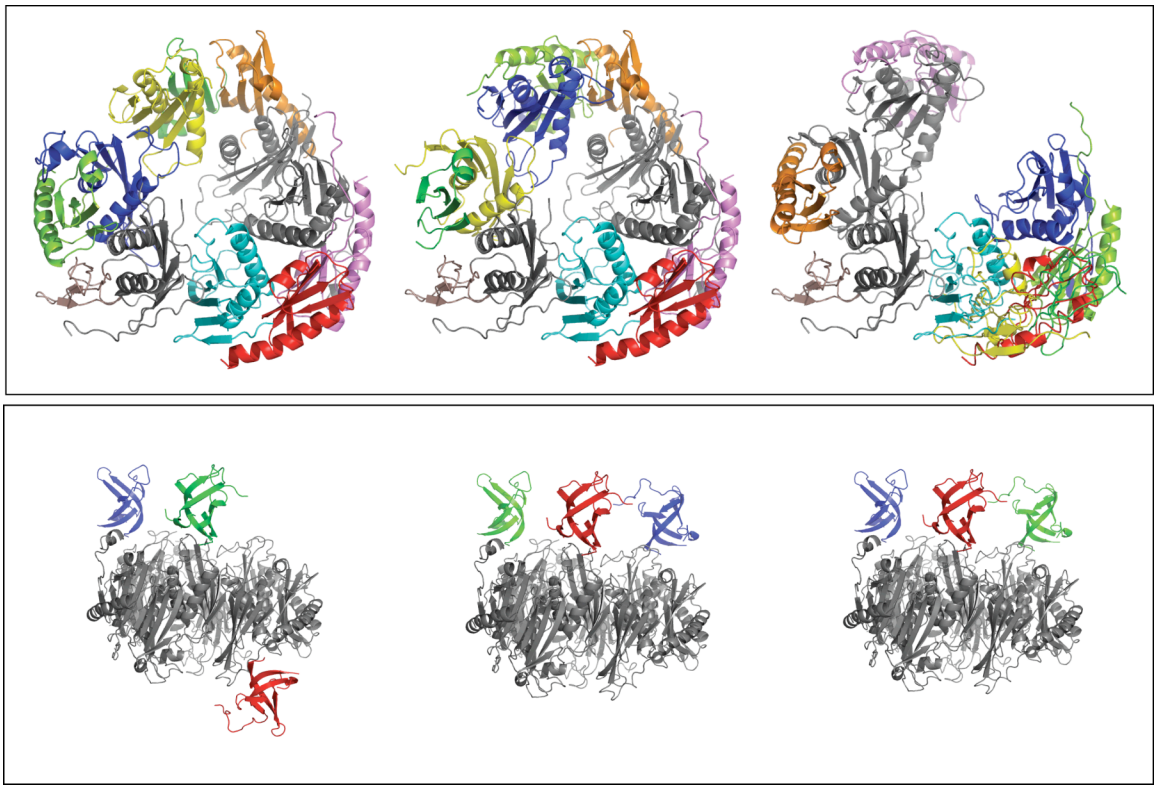

Figure 18: Reconstruction of the Eukaryotic exosome
Top: three models (out of 32) obtained using the seed search. Each model was obtained using two interaction templates multiple times.

Bottom: three models (out of 120) obtained when adding of Rrp4, Csl4 and Rrp40 to the ring of the exosome. Each protein can be placed on top or at the bottom of the ring given the current interaction templates.

From the study of the exosome, we have shown that, when using binary interaction templates, we could still predict structures with higher-level of organization, that the program can achieve fewer, more reliable predictions when using a protocol to search for symmetrical assemblies and that the use of predicted models instead of the structure for each domain decreases the quality of the complex model.

2. Applications

2.1. Estimation of the applicability of the method at different time points

We considered 615 complexes derived from high-throughput TAP-purification/ mass spectrometry experiments, along with their core components, their modules and attachments (as defined in (Gavin *et al.* 2006)). We estimated the fraction of the complex that can be assembled in the best-case scenario (*i.e.* when all interaction templates predicted are effectively suitable to model the interactions). We focused only on the interactions that are between inter-chain domains. The coverage at different periods was computed to appreciate how it changes as shown on Figure 19.

With the data contained in the latest version of SCOP, it is possible to orientate 30% of chains of the complexes. The portion of inter-chain interactions that can be modeled in the complexes varies greatly: for one fourth of the complexes, no interaction is modelable at all (*c.f.* value of the first quartile – see Methods for more details about the representation), whereas for more than half of the complexes, we can in principle predict more than 30% of the required interactions.

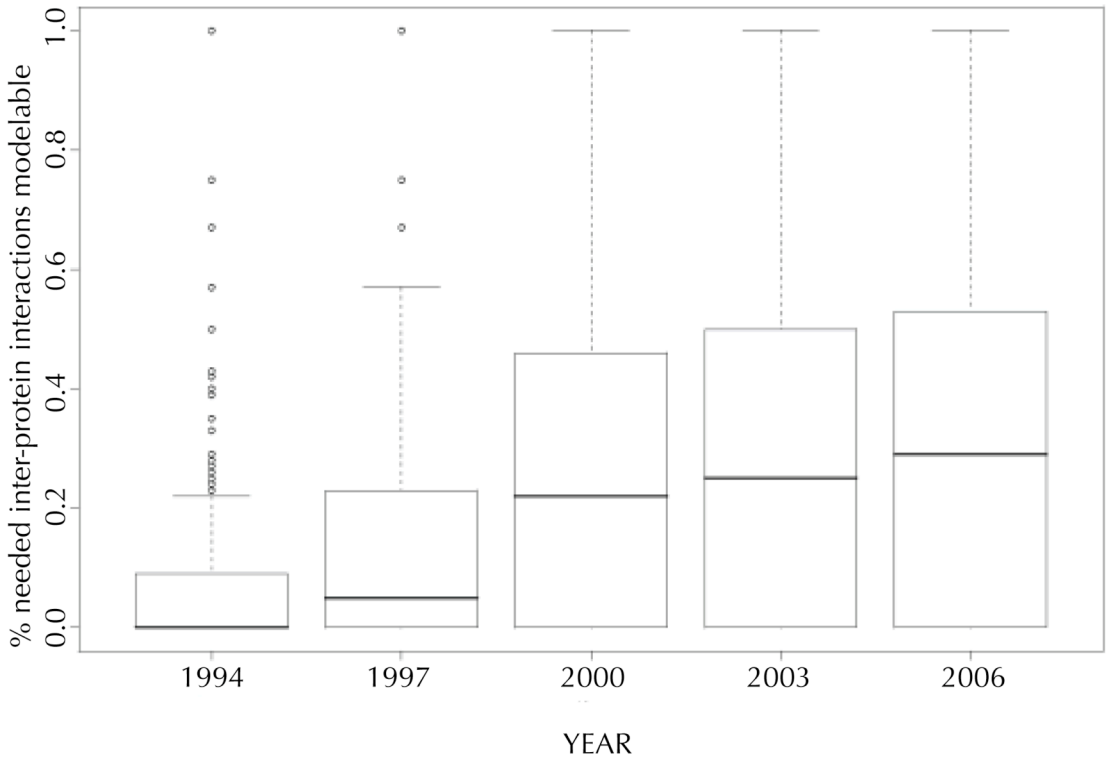

Figure 19: Box-plot representing the potential application of the method at different time points
The box-plot indicates the distribution of the percentages of crucial interactions that can be modeled for each complex (Gavin *et al.* 2006) using interaction templates available at different times

Our ability to predict interactions improves with each release of SCOP. However, after the rapid increase of in 2000, the trend seems to slow down. It suggests that even if the number of structures released steadily increases, the knowledge of the interaction structures increases at a slower pace.

2.2. Predictions

2.2.1 RNA polymerase

RNA polymerases are essential enzymes involved in the transcription of genes into RNA and are found in all organisms and many viruses. While in bacteria, only one type of RNA polymerase is found, three variations of RNA polymerases co-exist in

eukaryotic cells. RNA polymerase I synthesizes a precursor ribosomal RNA (rRNA 45S) that matures into three major RNA sections of the ribosome (Russell and Zomerdijk 2006). RNA polymerase II synthesizes precursors of RNAs and most snRNAs and miRNAs (Sims *et al*. 2004). RNA polymerase III synthesize**s** tRNAs, another ribosomal RNA (rRNA 5S) and small RNAs found in the nucleus and cytosol (Haeusler and Engelke 2006). The structures of both, bacterial RNA polymerase and RNA polymerase II, were determined and they share great structural similarity ((Woychik and Hampsey 2002; Borukhov and Nudler 2003) and (Chen and Hahn 2003; Chung *et al*. 2003; Bushnell *et al*. 2004) respectively). However, eukaryotic RNA polymerase II differs in that it misses domains to initiate transcription by itself and recruits general transcription factors. Moreover, in eukaryotic cells, RNA polymerase must deal with the typical DNA packing that does not exist to the same extent in bacterial RNA polymerase.

SCOP	**I core**	**I nocore**	**II core**	**II nocore**	**III core**	**III nocore**	**PDB entry: 1i50**
a.114.1.1				SPT5			
a.143.1.2		RPAB2		RPAB2		RPAB2	F
a.177.1.1		RPC7					
a.4.11.1		RPAB5		RPAB5		RPAB5	J
a.4.5.15			T2FB				
a.60.8.2			RPB4			RPC9	
a.8.3.1		MAN1					
b.15.1.1		HSP42					
b.30.5.6		MAN1					
b.40.4.5	RPA43		RPB7		RPC8		
b.40.4.8		RPAB3		RPAB3		RPAB3	H
b.43.4.1		MET10					
b.65.1.1			T2FB				
b.69.4.1		SNI1					
c.25.1.4		MET10					

c.36.1.8		MET10					
c.37.1.19		DHH1 (x2)					
c.45.1.1		CDC14 (x2)					
c.48.1.3		MET10					
c.52.3.1		RPAB1		RPAB1		RPAB1	E:1-143
c.6.2.1		MAN1					
c.64.1.1		MET10					
d.181.1.1	RPAC1		RPB3		RPAC1		C:42-172
d.230.1.1	RPA43		RPB7		RPC8		
d.74.3.1	RPAC1		RPB3		RPAC1		C:3-41, C:173-268
d.74.3.2			RPB11			RPAC2	K
d.78.1.1		RPAB1		RPAB1		RPAB1	E:144-215
e.29.1.1		RPA2	RPB2		RPC2		B
e.29.1.2	RPA1			RPB1	RPC1		A
g.41.3.1	RPA12		RPB9 (x2)			RPC10 (x2)	I:1-49 I:50-122
g.41.9.2							L

Table 3: Comparison of SCOP domains from RNA polymerases I, II and III
The three RNA polymerase complexes are defined as in Gavin *et al.* (Gavin *et al.* 2006). The structure found in the Protein Data Bank under code 1i50 corresponds to one instance of the RNA polymerase II complex

Each of the RNA polymerases I, II and III proteins were aligned to SCOP domains (Table 3). Domain assignments for each of these proteins were compared in order to obtain a domain map across RNA polymerases. The domains from the structure of RNA polymerase II were added (PDB code: 1i50) in order to show which part of the RNA polymerase structure is known. Strikingly, most domain types are present in the three RNA polymerases. Thus, whenever it is possible we will directly position the domains onto their equivalent in the known structure. However, the classification of complex proteins in core and not-core components seems difficult.

We built models for the components of RNA polymerase I and III and then sought interactions in each complex (core + attachment + module) that could in principle be predicted based on domain types. In order to limit superposition problems due to bad model predictions, we first tried to get a prediction using the SCOP domains reference for each domain.

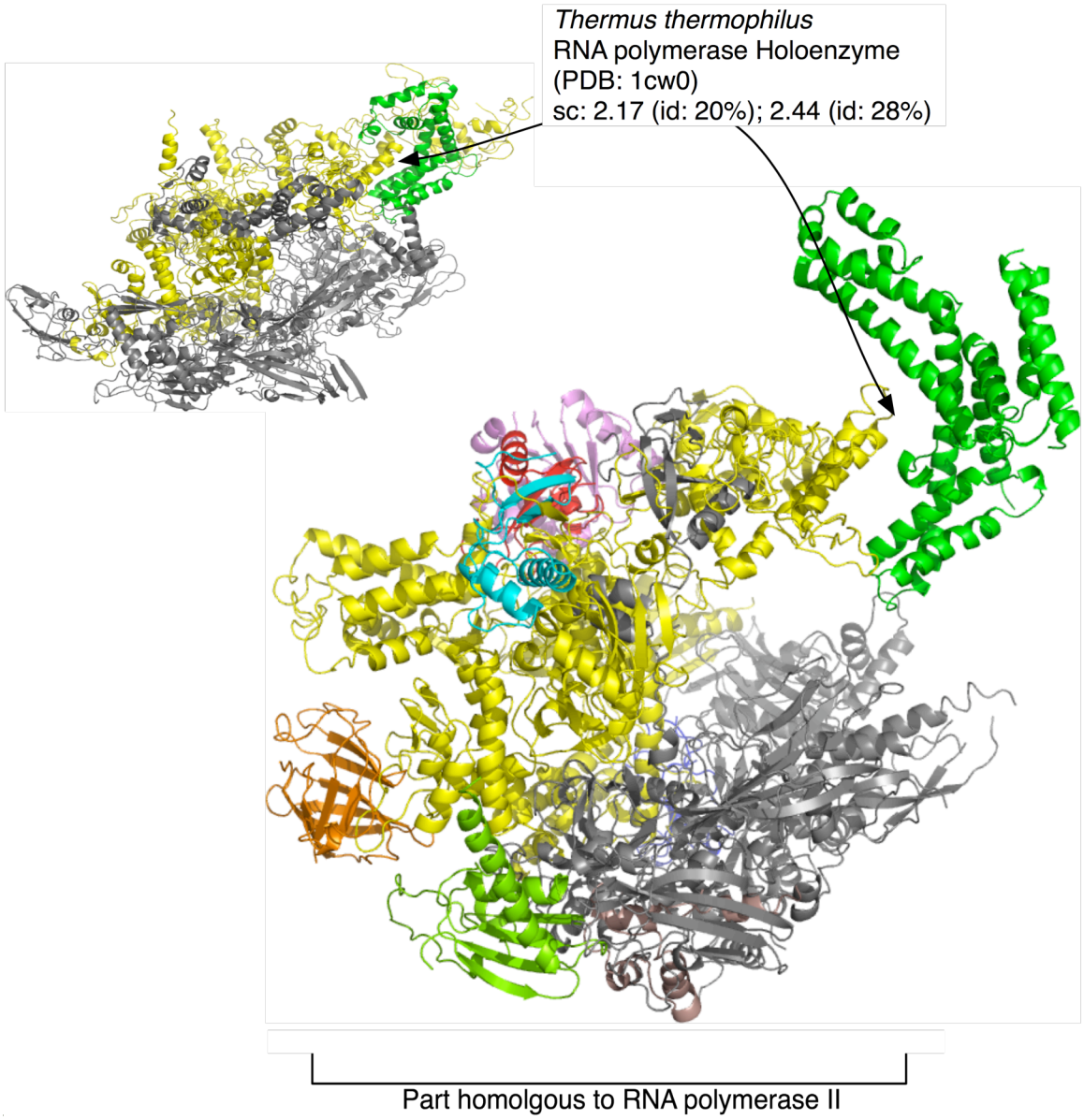

Figure 20: RNA polymerase I derived from RNA polymerase II with the addition of RPC7

Unfortunately, only one domain of RNA polymerase I could be added to the assembly derived trivially from RNA polymerase II structure (Figure 20). The domain corresponds to RPC7 and could be assembled onto the RNA polymerase beta-prime domain using an interaction template from *Thermus thermophilus* RNA polymerase holoenzyme (PDB code: 2cw0). However, the poor quality of the superposition of RPC7 on its template makes the prediction dubious (sc score: 2.44 and 2.17) and several means were used to evaluate the quality of the prediction. InterPreTS was used to evaluate the likelihood of the interaction built and scored poorly when compared with the likelihood of the interaction in the template (-36.56). We used Consurf (Armon *et al.* 2001) to map the conservation of residues onto the structures of the two interacting domains (Figure 21). The number and type of atomic interactions created was assessed using a derivative of Ligplot for domain-domain interaction (Wallace *et al.* 1995). When we compared the interaction pattern, obtained within the interaction we have predicted, to the original pattern, the difference is striking and we cannot have much confidence in our prediction.

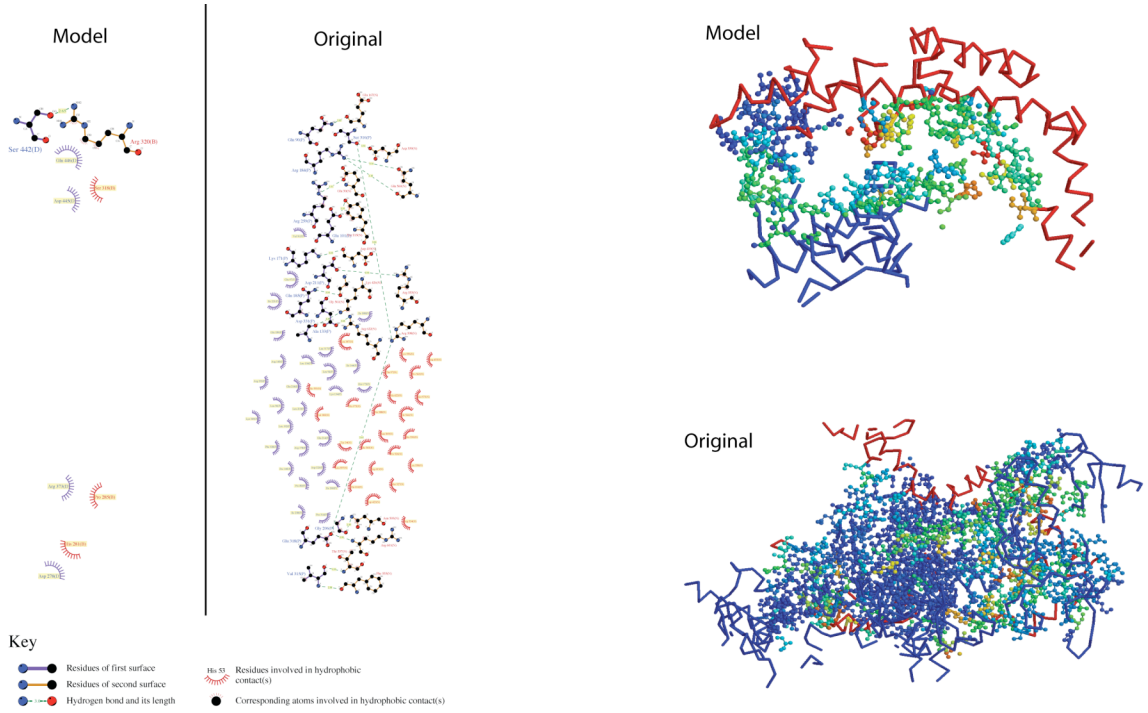

Figure 21: Study of the interface predicted between RNA polymerase I and RPC7
Left: Ligplot (Wallace *et al.* 1995) 2D-representation of the interface

Right: the interfaces displayed with evolutionary information ((Armon *et al.* 2001))

in the predicted structure and in the correponding interaction within the interaction template. The backbone in blue corresponds to RNA polymerase, the backbone in red to RPC7 and its homolog. Residues in red are more conserved and residues in blue are not.

2.2.2 Cdc 48/Npl4/Ufd1 complex

One complex identified as a potential modeling candidate in the above screen is the yeast Cdc48/Ufd1/Npl4 complex, consisting of 5 components: Ufd1, Npl4, Cdc48, YDR049W and Shp1. The AAA (ATPase associated with various cellular activities) ATPase p97, the well-studied vertebrate homolog of Cdc48 is involved in a broad variety of cellular activities, amongst which are ubiquitin-dependent protein degradation (Hetzer *et al.* 2001), spindle disassembly (Cao *et al.* 2003), Golgi preassembly, centromere targeting (Vong *et al.* 2005) and post-mitosis nuclear envelope reassembly (Johnson *et al.* 1995). Different proteins adapt to Cdc48 to achieve their functions. For instance, it requires Ufd1 and Npl4 to

participate in spindle disassembly at the end of mitosis (Cao *et al.* 2003). p97 forms a homo-hexamer (Peters *et al.* 1992; Zhang *et al.* 2000; Beuron *et al.* 2003; DeLaBarre and Brunger 2003; Huyton *et al.* 2003) on which the different adaptors bind.

Protein	#AA	Sequence	SCOP	Hit	E-value	Pfam
Ufd1	361	118-207	d.31.1.1	1e32A:118-207	0.037	UFD1
		90-164	b.52.2.3	1e32A:15-84	0.028	UFD1
Npl4	580	1-84	d.15.1.1	1v2yA:8-104	1.60E-05	Pfam-B
Cdc48/p95	835	211-468	c.37.1.20	1e32A:201-458	0	AAA
		117-207	d.31.1.1	1e32A:107-200	3.00E-07	CDC48 2
		31-116	b.52.2.3	1e32A:21-106	3.30E-14	CDC48 N
		481-757	c.37.1.20	1r7r:471-735	0	AAA
UBX1/Shp1	423	355-421	d.15.1.2	1i42A:1-87	2.50E-29	UBX
		226-299	d.245.1.1	1vazA:3-76	3.40E-22	SEP
		1-45	a.5.2.3	1v92a:1-46	2.30E-16	Pfam-B
YDR049W	632	338-562	d.211.1.1	1s70B:20-288	0	

Table 4: Modeling of the domains for each protein from the Cdc48/Npl4/Ufd1 complex
From left to right, protein name, number of amino-acids, part of the sequence that matches a SCOP domain, SCOP category of the match, description of the SCOP domain hit, E-value, corresponding Pfam classification.

Each protein was assigned plausible SCOP domains using the described sequence-based protocol (Material and Methods) with corresponding structural models (Table 4). Interestingly, Ufd1 sequence ambiguously hits two SCOP domains found in Cdc48 (Golbik *et al.* 1999). TAP/MS experiments indicate that 5 proteins interact tightly: Shp1 (Ubiquitin Regulatory X), YDR049W (an hypothetical protein), Ufd1 (Ubiquitin fusion degradation 1), Npl4 (Nuclear protein localization 4) and Cdc48 (homolog of p97 in yeast) via 4 main interactions (Cdc48-Shp1, Cdc48-Ufd1, Cdc48-Npl4 and Ufd1-Npl4 (Figure 22)). With the sole study of these interaction

data we can consider two independent organizations: one with Cdc48-Ufd1-Npl4 and the other with Cdc48-Shp1.

Figure 22: TAP-MS characterization of the Cdc48/Npl4/Ufd1 complex
Red lines describe tight links, pink and yellow line looser interactions.

The monomer structure and the ring formed by 6 copies of protein Cdc48 were modeled on the structure of *murine p97/VCP* (PDB code: 1r7r). Then, one copy of proteins Npl4 and Ufd1 was attached to the structure of the ring. Only one copy of Npl4 and Ufd1 was added to reflect the stoichiometry of the complex (Pye *et al.* 2007). To complete the study, we searched possible means to bind Shp1 or YDR049W to the structure.

We found 3 distinct means to bind Npl4 and Ufd1 to the structure of the Cdc48 ring (Figure 23), none of which could accommodate Shp1 or YDR049W. The difference between the 3 predictions lies in the orientation of the binding of Ufd1 on Cdc48. The two first predictions were almost identical provided the 6-fold symmetry of the Cdc48 ring and in both predictions the interaction template for the interaction between Ufd1 and Cdc48 was found in the structure of the Cdc48 homologue. The other prediction used a template from the amino-terminal domain of N-ethylmaleimide-sensitive fusion protein to model the interface between Cdc48 and Ufd1 (PDB code: 1qdn). The interaction between Ufd1 and Shp1 was based again on the structure of p97 in complex with p47 (PDB code: 1s3s).

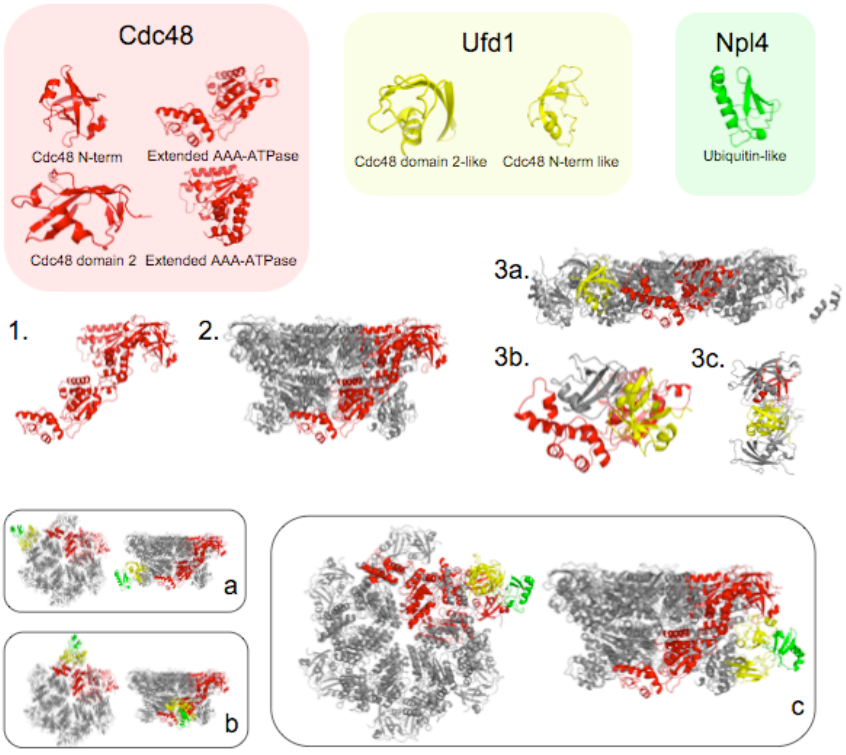

Figure 23: The assembly process leading to the three predictions for Cdc48/Npl4/Ufd1
First row: collected structures for the domains of Cdc48, predicted structures of Ufd1 and Npl4; second row: 1: assembly of Cdc48 domains; 2: assembly of the ring of six Cdc48; 3a, 3b, 3c the three templates used for the predictions m1, m2, m3

The last model compares well with the negative stain EM image of the complex (Pye *et al.* 2007) (Figure 24). As shown on the picture of the negative stain EM, the Ufd1-Npl4 heterodimer binds on the side of Cdc48, not on top.

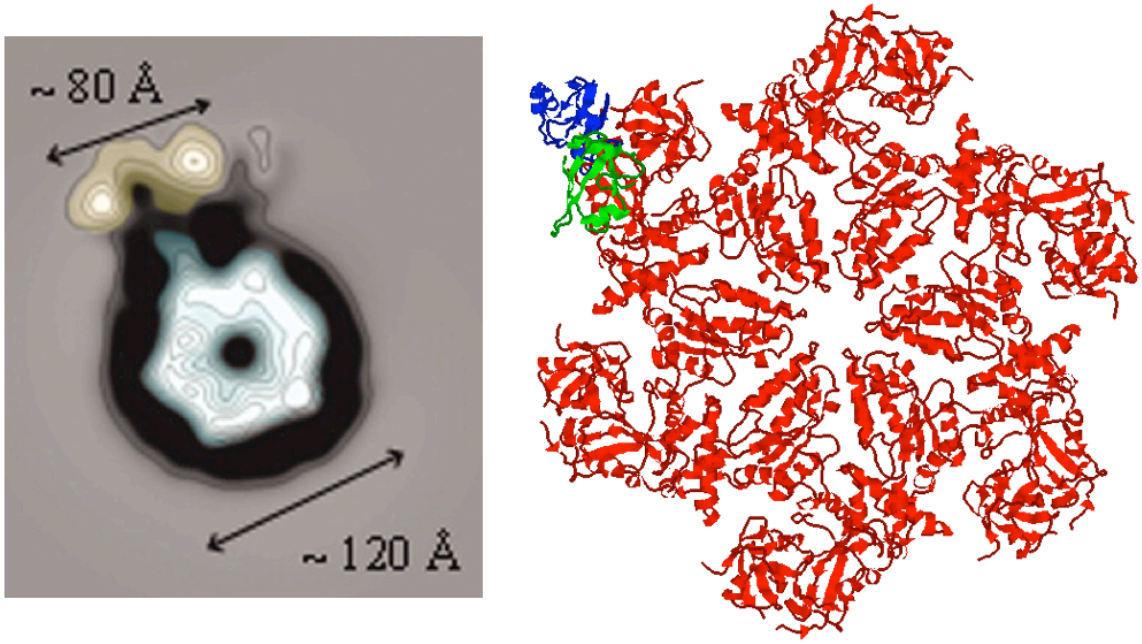

Figure 24: Comparison of the first model with a negative stain EM from Cdc48/Npl4/Ufd1

The structure of Ufd1 was solved recently (Park *et al.* 2005). In our prediction, the orientation of the 2 domains of Ufd1 is close to the orientation they obtained by NMR structure despite the ambiguity in the domain assignment.

Finally, we evaluated the models using purely bioinformatics methods. We assessed the likelihood of the interactions created with InterPreTS first and then by a combination of Ligplot (Wallace *et al.* 1995) and conservation study.

In the two first models, the anchoring of Ufd1 on Cdc48 is made via an interaction that has a bad InterPreTS score. In the third, the interaction scores better. In order to confirm this evaluation, we first drew all the interactions (*i.e.* hydrogen bonds and hydrophobic contacts) using Ligplot (a tool to represent interactions between structures in two dimensions) (Figure 25). The bound between Ufd1 and Npl4 being the same in the three predictions, we display only the interaction patterns between the Cdc48 ring and the Ufd1/Npl4 sub-complex. The third prediction

contains more connections between the ring and Ufd1/Npl4. Surprisingly, the interaction used to anchor the Cdc48 ring and Ufd1/Npl4 (between one Cdc48 AAA ATPase domain from Cdc48 and the other of Ufd1) is not the one that generates the most interactions. The interaction between the other Cdc48 AAA ATPase domain from Cdc48 and Ufd1's Cdc48 2-like domain is an interaction that is indirectly predicted and accounts for most of the interaction between the ring and the Ufd1/Npl4 subunit. We performed sequence alignments to locate on the interactions those that involve conserved residues (Table 5). The interactions in model 3 involve residues that are better conserved.

Finally, even if it is difficult to assess the accuracy of these models, we think that the three models are worth further investigation: the two first models are built using interaction templates directly derived from the structures of the Cdc48 ring to bind the adaptors to the ring of Cdc48, whereas the last prediction is based on a remote template but supported by interaction and conservation studies.

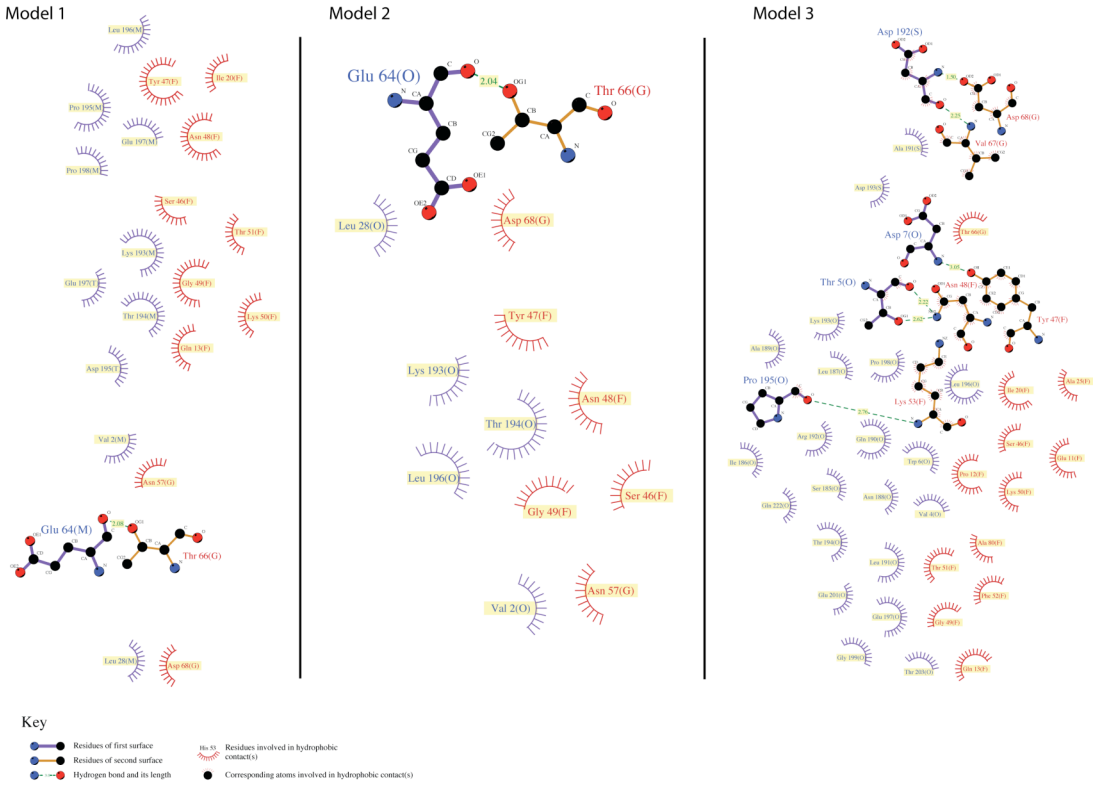

Figure 25: 2D representations of the three interfaces predicted between the Cdc48 ring and the Ufd1/Npl4 complex

a. data for model 1

H Bond	Donor	Res	Conservation	Acceptor	Res	Conservation
	THR	G66	3	GLU	M64	0

Hydrophobic contact	Atom1	Res	Conservation	Atom2	Res	Conservation
	ILE	F20	6	LEU	M196	0
	TYR	F47	6	THR	M194	0
	TYR	F47	6	PRO	M195	0
	TYR	F47	6	LEU	M196	0
	TYR	F47	6	GLU	M197	0
	TYR	F47	6	PRO	M198	0
	GLN	F13	5	ASP	T195	0
	GLN	F13	5	GLU	T197	0
	ASP	G68	3	LEU	M28	0
	ASN	G57	2	VAL	M2	5

THR	F51	2	LYS	M193	0
SER	F46	1	LYS	M193	0
ASN	F48	0	LYS	M193	0
ASN	F48	0	THR	M194	0
ASN	F48	0	PRO	M195	0
ASN	F48	0	LEU	M196	0
GLY	F49	0	LYS	M193	0
GLY	F49	0	THR	M194	0
GLY	F49	0	PRO	M195	0
LYS	F50	0	LYS	M193	0
LYS	F50	0	THR	M194	0

b. data for model 2

H Bond	**Donor**	**Res**	**Conservation**	**Acceptor**	**Res**	**Conservation**
	THR	OG1	3	GLU	O64	0

Hydrophobic contact	**Atom1**	**Res**	**Conservation**	**Atom2**	**Res**	**Conservation**
	TYR	F47	6	LEU	O196	0
	ASP	G68	3	LEU	O28	0
	ASN	G57	2	VAL	O2	5
	SER	F46	1	LYS	O193	0
	ASN	F48	0	THR	O194	0
	ASN	F48	0	LEU	O196	0
	GLY	F49	0	THR	O194	0

c. data for model 3

H Bond	**Donor**	**Res**	**Conservation**	**Acceptor**	**Res**	**Conservation**
	LYS	F53	3	PRO	O195	0
	ASP	O7	2	TYR	F47	6
	VAL	G67	2	ASP	S192	0
	ASP	S192	0	ASP	G68	3
	ASN	F48	0	THR	O5	1
	ASN	F48	0	THR	O5	1

Hydrophobic contact	Atom1	Res	Conservation	Atom2	Res	Conservation
	PHE	F52	9	ARG	O192	0
	PHE	F52	9	LYS	O193	0
	PHE	F52	9	PRO	O198	0
	PRO	F12	7	ARG	O192	0
	PRO	F12	7	GLU	O197	0
	PRO	F12	7	PRO	O198	0
	PRO	F12	7	GLY	O199	0
	ILE	F20	6	TRP	O6	6
	ALA	F25	6	TRP	O6	6
	TYR	F47	6	TRP	O6	6
	GLU	F11	6	PRO	O198	0
	GLU	F11	6	GLY	O199	0
	ILE	F20	6	SER	O185	0
	ILE	F20	6	ILE	O186	0
	ILE	F20	6	ALA	O189	0
	ILE	F20	6	THR	O203	0
	TYR	F47	6	GLN	O190	0
	TYR	F47	6	VAL	O4	0
	GLN	F13	5	PRO	O195	0
	GLN	F13	5	LEU	O196	0
	GLN	F13	5	GLU	O197	0
	GLN	F13	5	PRO	O198	0
	GLN	F13	5	GLU	O201	0
	ASP	G68	3	ALA	S191	7
	LYS	F53	3	THR	O194	0
	LYS	F53	3	PRO	O195	0
	LYS	F53	3	LEU	O196	0
	LYS	F53	3	PRO	O198	0
	ALA	F80	3	GLY	O199	0
	THR	G66	3	ASP	S192	0
	THR	G66	3	ASP	S192	0
	THR	G66	3	ASP	S193	0
	ASP	G68	3	ASP	S192	0
	THR	F51	2	ARG	O192	0
	THR	F51	2	LYS	O193	0
	THR	F51	2	THR	O194	0
	VAL	G67	2	ASP	S192	0
	SER	F46	1	LYS	O193	0
	ASN	F48	0	TRP	O6	6
	ASN	F48	0	LEU	O187	1

ASN	F48	0	LEU	O187	1
ASN	F48	0	THR	O5	1
ASN	F48	0	ILE	O186	0
ASN	F48	0	ALA	O189	0
ASN	F48	0	GLN	O190	0
ASN	F48	0	VAL	O4	0
GLY	F49	0	GLN	O190	0
GLY	F49	0	LEU	O191	0
LYS	F50	0	ASN	O188	0
LYS	F50	0	ALA	O189	0
LYS	F50	0	GLN	O190	0
LYS	F50	0	LEU	O191	0
LYS	F50	0	ARG	O192	0
LYS	F50	0	LYS	O193	0
LYS	F50	0	GLN	O222	0

Table 5: Conservation of residues at the interface in the three models predicted for Cdc48/Ufd1/Npl4

Consevation of residues at the interface in models "1", "2" and "3" (corresponding tables a, b and c). The columns "Atom" and "Res" describe the residue, the column "Conservation" describes the level of conservation (0: poor). For each model, the first residue described belongs to the Cdc48 ring, and the second residue is from Ufd1/Npl4. Conserved interactions are shown in red.

Discussion

1. Summary of the results

1.1. Results

During this work, I developed a method to predict the structure of protein complexes. The problem differs from the prediction of the structure of protein-protein interactions because the number of interfaces available for a set of subunits is limited and not all arrangements are sterically possible.

The procedure searches for interaction templates from solved protein structures that can be used to predict the orientation of any pair of subunits of the complex. These pairwise orientations are represented in a graph and combinations of orientations are searched to form protein assemblies. The number of possible combinations grows at least exponentially with the number of subunits and thus it is rarely possible to generate all the models. To limit the predictions to the most significant, several complementary methods have been implemented: (i) subunits that have matches in a known structure are directly oriented, (ii) subunits known to be in direct contact constrain predictions and (iii) multimeric assemblies of multimers are searched directly.

The method was benchmarked on 425 elementary complexes consisting of three domains in interaction. The influence of three parameters (sequence identity, superposition and InterPreTS score) on the quality of predicted assemblies was tested. Sequence identity and superposition score significantly improved the specificity and the ratio of 'true positives' predicted. Nine structures from a benchmark set that could be constructed using non-trivial templates from other structures were used to illustrate the performance of the method. Moreover, the

exosome, a circular assembly of 6 proteins arranged in a trimer of dimers was used to illustrate how the method handles such multimer of multimers assemblies.

Yeast complexes on which the method could be applied were found. As more structures are solved, the number of interactions that can be predicted in complexes increases. In the specific example of RNA polymerase I and III, the method was successful in using the structure of RNA polymerase II as a template. However, no additional subunits could be fitted in a satisfying manner. Finally, three assemblies were predicted for Cdc48 with Ufd1 and Npl4. The predictions were compared to low-resolution structure and evaluated by a detailed study of the interactions formed.

1.2. Application

The database of interactions can be used to locate potential interfaces, find alternative modes of binding for two protein domains and in general detect the potential interacting proteins in a complex.

The assembly method can be applied to a large number of protein assemblies when some structural information (determined or predicted) is available for the subunits. The ability to make a prediction depends on the interactions seen in solved structures but it can be extended with structures of protein-protein interactions determined by any other technique. The more information about the protein complex that is available, the better the prediction will be.

1.3. Comments

1.3.1 Domain issues

In this work and the study of Inbar *et al*, protein domains are the basic structural units that are combined to form a prediction. Using domains is relevant, as they are elementary structural 'blocks' that constitute protein structures (Murzin *et al*. 1995;

Orengo *et al*. 1997). However, manipulating domains is not trivial. For instance, domain boundaries are sometimes difficult to determine, yet they are critical to the success of structure superposition attempts, or sometimes a single part of the protein is similar to several domains of different types, which makes the assignment ambiguous.

Because of these limitations, one could consider that domains are not appropriate structural units for the prediction of such an assembly and that alternative structural units should be considered, for instance, complete structures. However, as most interactions involve protein domains, as domain structures re-occur more than complete protein structures and as domains are well characterized and classified, we consider that they are the most appropriate units to use.

The SCOP classification of protein domains (Murzin *et al*. 1995) is central to this work: it is used for the assignment of a reference structure to each domain from the query, for building the database of interaction templates and for assigning domains before comparing two multi-domain structures. In our perspective, the SCOP classification suffers from two limitations: the database is not updated frequently (only once per year usually) and each update does not account for the most recent structures (in the worst case, structures solved during the last year and a half may be omitted, even in the automatically-determined SCOP pre-release). In addition, structural variations can be significant inside a single SCOP family (Suhrer *et al*. 2007) and it is difficult to know what degree of structural similarities can be expected from two domains with the same SCOP classification. Thus, working with these categories contributes dramatically to the scarcity and obsolescence of interaction templates that can be used. Furthermore, assigning a SCOP family to a domain from sequence or structure is sometimes ambiguous and yet it is critical for the success of the approach, as the program will search amongst interaction templates involving domains of similar types only.

SCOP was used, as it is believed to be one of the most accurate classifications of domains. However, now that the program is prototyped, we can define better requirements for the classification of protein domains needed. The problem is not to use the most precise classification of protein domains, but a classification that fits best our need: it must be up-to-date in order to cover as many structures as possible, and it must provide good insights into the structural similarities between domains. It would be useful as well to be able to assign classifications directly from sequence.

An FSSP-like approach seems the most convenient. Structures are automatically classified in a hierarchy that reflects their structural similarity and the content of the database is controlled and updated regularly. Expert-curated databases are too refined for our framework and delay the extraction of interaction templates from the most recent structures. For our method, good coverage of the domain space and correct estimates of the similarity between protein domains are more critical. Indeed, the accuracy of a template directly depends on the structural similarity and sequence identity between its domains and those in the query.

Finally, assigning a single class to a protein domain remains difficult and prone to ambiguity. When modelling protein assemblies from the structure of subunits, there are two ways to circumvent such a problem: when domain types are assigned, one may bias the assignment to domains for which interaction templates are available, or, alternatively, one domain may be assigned several domains when its classification is ambiguous. Either of the two approaches would ensure that for a set of domains, all possible interaction templates are tried and that domain classification does not limit the set of orientations tried.

1.3.2 Predicting the structure of complexes from binary interactions

1.3.2.1 Prediction of binary interactions

As seen above, it is very difficult to select a set of relevant interaction templates for the prediction of the orientation of two structures. In this work, three different approaches were assessed: (i) selection by comparing the sequences of the domains and the template domains; (ii) selection by similarity of domain structures; (iii) selection by InterPreTS score for resulting interface. Sequence identity and structural similarity performed similarly and helped to select good interaction templates. It could also be argued that both are quite similar, since there is a clear relationship between them (Lesk and Chothia 1980). However, InterPreTS did not perform well for the scoring of interfaces and was not used further.

Despite the fact that comparing domain features from the query and the interaction template achieves good performance, an accurate method for the recognition of protein-protein interfaces seems intuitively more adapted to the problem. Several techniques can be used to achieve such an evaluation (Guerois *et al.* 2002; Verkhivker *et al.* 2002; Kortemme *et al.* 2004; Schymkowitz *et al.* 2005; Li *et al.* 2006) and will be tested in the next version of the software. If such a method is found, it could be used to evaluate the quality of inter-chain interactions and help the selection of the best interaction templates; for the structure of intra-chain orientation, using domain-centered evaluation method prevails as the interfaces do not follow the same rules.

However, the structures of the different components of a complex are rarely known and predicted structures often have to be used. With a high level of homology and therefore a good model for the components, the interface is likely to be preserved (Aloy *et al.* 2003) whereas for more difficult predictions the conformation of

residues at the interface is uncertain. Thus, “scoring” interfaces is even more difficult when the two interacting structures are not perfectly determined as in the case of protein models. This phenomenon has to be considered when choosing a method to evaluate interaction templates: the method used to evaluate interfaces has to be very robust to small changes in the orientation of residues at the interface.

Moreover, predicting the structure of interactions by docking of two structures requires that the interfaces presented by each structure are accurately determined and that the complementarities between them can be evaluated. Consequently, the knowledge of the interface is critical in docking, whereas in homology modeling of interactions, the global structural features of the two domains are used for the prediction of the interaction and not the interface only. So the overall shape of a protein is sufficient for homology modeling of interactions but not for docking. This issue also illustrates the pertinence of using sequence identity and superposition score for the evaluation of the interaction template, because they accurately estimate the similarity between a domain and a possible template while not assuming that the interfaces are perfectly determined.

Currently, the database of interaction templates is made of the collection of distinct interactions in each protein structure. Redundancies were purged within each structure but no comparisons or selections were made across structures. To increase the speed of the method, the set of interaction templates could be restricted. However, it must be done carefully to ensure that no information is lost in the process and from our point of view, it is not trivial.
For instance, one could decide to compare all interactions of the same kind across structures using iRMSD and keep one representative for the whole set. The consequence would be that only one pair of interacting proteins is left to represent the whole set of similar orientations. Then when computing sequence identity or the structural similarity between a pair of structure from the query and the interaction template, these values will reflect only the match to the representative

selected and better matches with other similar structures will be discarded. Thus, comparing interaction templates against each other is not sufficient and more refined clusters of interaction templates have to be searched.

1.3.2.2 Relevance of the final assembly

Some drawbacks are inherent to the bottom-up strategy employed. When building protein assemblies by combining the orientations predicted for binary interactions with no record of previous orientations, the risk is to lose the higher-level structural features of the final assembly, *i.e.* structural features that encompass more than two domains. In this study, we illustrated this point by studying the ring structure of the exosome and the dimer of trimers structure of the CDK-cyclin complex (PDB code: 1g3n). In these two cases, the naive application of the bottom-up approach consisting of the addition of structures with no memory of the orientations used previously in the construction is time consuming and generates many poor predictions.

To tackle this re-occurring problem, we propose to use seeds consisting of sets of transformations used to predict the structure of several binary interactions in the structure. This approach favors the formation of higher levels of structural organization, as we illustrated in the case of the exosome where we could form the hexameric ring in an accurate fashion and generate a restricted set of predictions with all the possible variations of the ring, or in the case of the inhibited CDK-cyclin complex for which we obtained few symmetrical predictions. The seed approach is easy to implement in the homology-based approach as interaction templates are picked from a finite set of orientations (the database of interaction templates) and pairs of domains where the same interaction templates can be used are easy to find. The method could still apply to docking approaches but needs some adaptation because the space of candidate interaction template is bigger and could be quasi infinite.

Finally, when the assembly process is complete, the prediction has to be evaluated. Currently, we evaluate the prediction by the score obtained for each prediction of interaction and ensure that all the subunits are connected and do not overlap. We count the number of interactions between elements in the prediction (to check how many indirect interactions were created) and optionally score the interfaces built using InterPreTS. The next versions of InterPreTS are indeed more accurate (Russell, personal communication).

Other means to estimate the quality of the prediction can be considered: the interactions formed indirectly during construction can be scored and compared to the interactions from the database, or the energy of the overall structures can be computed and compared to the energies of the constituents separately.

At the moment, the scoring of the complete predictions is imperfect: it remains difficult to score good assemblies better and to discriminate realistic structures from artifacts. We compensate for this lack by providing tools to efficiently explore large sets of predictions.

1.3.2.3 Performance of the method

As time goes by, more protein structures are solved and more interaction templates become available. At the same time, the quality of all the methods to predict the structure of protein-protein interactions improves and complexes will be known in more detail. Thus, we expect potential applications of the method to increase in the coming years and constraints to increase the success of the method to be more numerous.

At the moment, the potential applications are quite limited and achieve mixed results. The task is complicated by the fact that we are integrating results from several error-prone studies. In the construction of the model for Cdc48/Ufd1/Npl4, for instance, we combined data from TAP-purification experiments and from

protein modeling. TAP-purification happened to reveal the superposition of two complexes (namely Cdc48-Npl4-Ufd1 and Cdc48-Shp1) and the modeling of Ufd1 was hampered by difficulties in the alignment of target and template sequences. Obviously, combining results from different methods may increase the potential error, or the drawbacks can compensate each other and it may be possible to discard some errors from previous studies. In the case of Cdc48-Npl4-Ufd1, the structural study of the complex shows that the binding of Shp1 occurs at the same location and that the binding of Npl4-Ufd1 and Shp1 must be exclusive.

2. Comparison with combinatorial docking

It is interesting to compare the method that we have developed to the other method for the prediction of assemblies developed by Inbar *et al*. The main difference lies in the prediction of the relative orientation of two structures: Inbar *et al*. use docking and we have used homology (Table 6).

	Docking-based	**Homology-based**
Orientations explored	Infinite	Finite – Limited to interaction of homologues
Structuring element	Interface	Complete structure
Comparison to native	RMSD	iRMSD across interactions
Search method	Heuristic search	Kruskal adaptation

Table 6: Differences between the docking-based approach and the homology-based approach

Docking enables one to search all possible orientations between two structures, which is an appealing feature. However, the search is computationally demanding and generates mostly false-positives. The correct answer, if present, might be lost in the noise. A compromise has to be found between the search space and the computing time required. Moreover, docking cannot be used currently to distinguish real interactions from artifacts and usually orientations are found for any

pair of proteins. In contrast, finding interacting homologues for a pair of domains happens seldom but the information provided is based on an existing interaction.

Inbar *et al.* used the canonical root-mean-square deviation (RMSD) to assess structural differences between predicted assemblies and originals. RMSD is the measure used when comparing a real protein structure to a model (*c.f.* evaluation of structures predicted in CASP). We chose to evaluate the quality of a prediction by computing iRMSD between all the interacting pairs in the two structures and keep the highest value for a score. Basically, their scoring is based on the overall similarity of the structures, where ours compares the similarity of each interaction in the two structures.

We think that an interaction-centric score is better than RMSD for the comparison of multi-domain assemblies predicted by arrangement of binary interactions (especially when comparing the model to the original structure during benchmarking). It does not seem to overreact to wrong interactions predicted: if one interaction is poorly predicted amongst many others, this interaction and this interaction only impacts the iRMSD score, whereas it may have a large effect on RMSD. This is because all the subunits bound to the domains involved in the interaction contribute to the final RMSD score (Figure 26). Still, when comparing two domains only, the two measures should be equivalent.

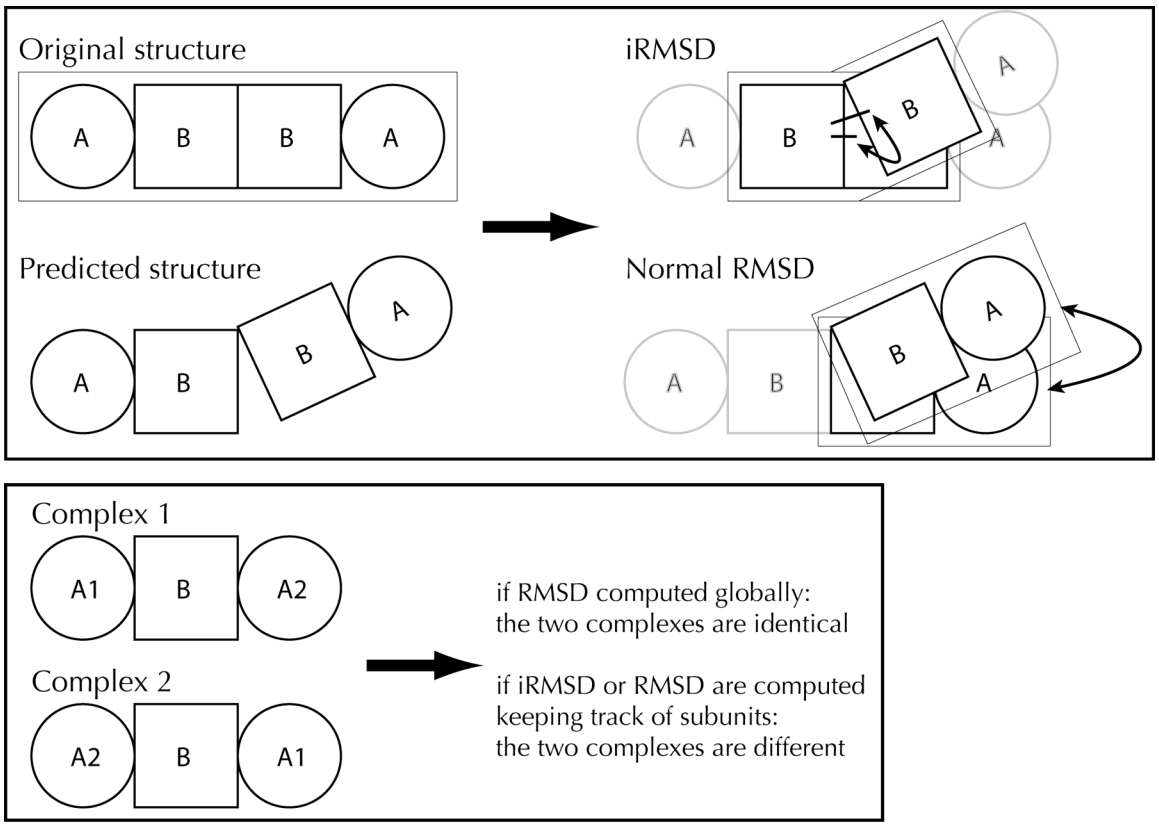

Figure 26: Comparison of RMSD and an iRMSD-based score as means to evaluate the difference between two assemblies
First box: the effect of a bad prediction on RMSD and the iRMSD-based score. The former is penalized by the two displaced domains when only the bad interaction is detected by the latter. Second box: direct RMSD evaluation cannot account for swapped domains.

In the case where a structure contains several distinct copies of the same subunits, computing RMSD of the two structures will not account for misplacement of the similar subunits, and they will all be treated the same. In contrast, our interaction-based score carefully considers all the possible equivalences of domains from one structure to those in the other before it gives a final score.

Thus, RMSD remains a good evaluation of the shape of the overall assembly. It seems more adapted to the method employed by Inbar *et al*, since their method is based on docking and since an obvious follow-up may use flexible docking. With flexible docking they will need an accurate estimation of the interactions predicted

between subunits and an estimate of the changes induced in the backbone of each subunit. RMSD measures both the differences in the interaction and the structural difference of each constituent. In our homology-based method, the flexibility of the subunits is not considered and the estimation of the quality of the interactions predicted is more important, thus using the highest value of iRMSD across interactions seems more appropriate here.

3. Other potential uses of protein interactions

In this work, the structure of protein assemblies is predicted by combining several homology-based predictions of binary interactions, as we believe much can be learned from the structures already determined. Here, we present different benefits that can be derived from the knowledge of protein-protein interactions.

3.1. Prediction of interfaces

Usually, an experimentalist would compare his/her two proteins of interest to similar proteins that directly interact and for which the structure has been solved. Then the probable location of the interface can be derived by homology (Bork 1989). Similarly, our method can contribute to the prediction of interfaces when there is no structure for a direct interaction between homologues and when it is possible to combine interactions to form new interactions. By considering complexes formed by several proteins, one multiplies the chances to be able to predict the conformation of an interaction.

3.2. Limiting the number of structural determinations required for predicting assemblies

When determining the structure of large protein assemblies, the program can be used to break the problem into pieces and to find subunits predicted to be in contact that lack structural information. If good homology to some other structures is detected, the structural biologist can focus on novel interactions and combine them with the known interactions using this program.

3.3. Spatial constraints

As shown in this study, the spatial arrangement of protein structures is very constrained: not every structural arrangement of a protein chain is possible, the same is true for the structures of protein-protein interactions and possibly for protein assemblies. Knowing that structures are constrained provides a wealth of information. For instance, homology modeling, fold recognition and fragment-based methods are based on the re-occurrence of some structural features (protein/domain structures in the first two cases, fragment structures in the third). Here, two possible applications of the re-occurrence of protein-protein interaction features are considered that can contribute to the determination of stoichiometry (the number of copies of each constituent of a complex).

3.3.1 Repeating a pattern to form loops or helices

Some structures contain several copies of the same subunit and in such cases, the study of sub-complexes is enough to obtain the complete structure. As we have seen, the assembly of symmetrical structures (multimers of multimers and protein rings) can be achieved with few data. The exosome for instance, consists of six chains and can be assembled with only two interaction templates (and appropriate constraints) when five interaction templates are theoretically needed. Amongst symmetrical assemblies, protein rings and macro-helices are the most repetitive; the difference being that rings close while helices can extend endlessly.

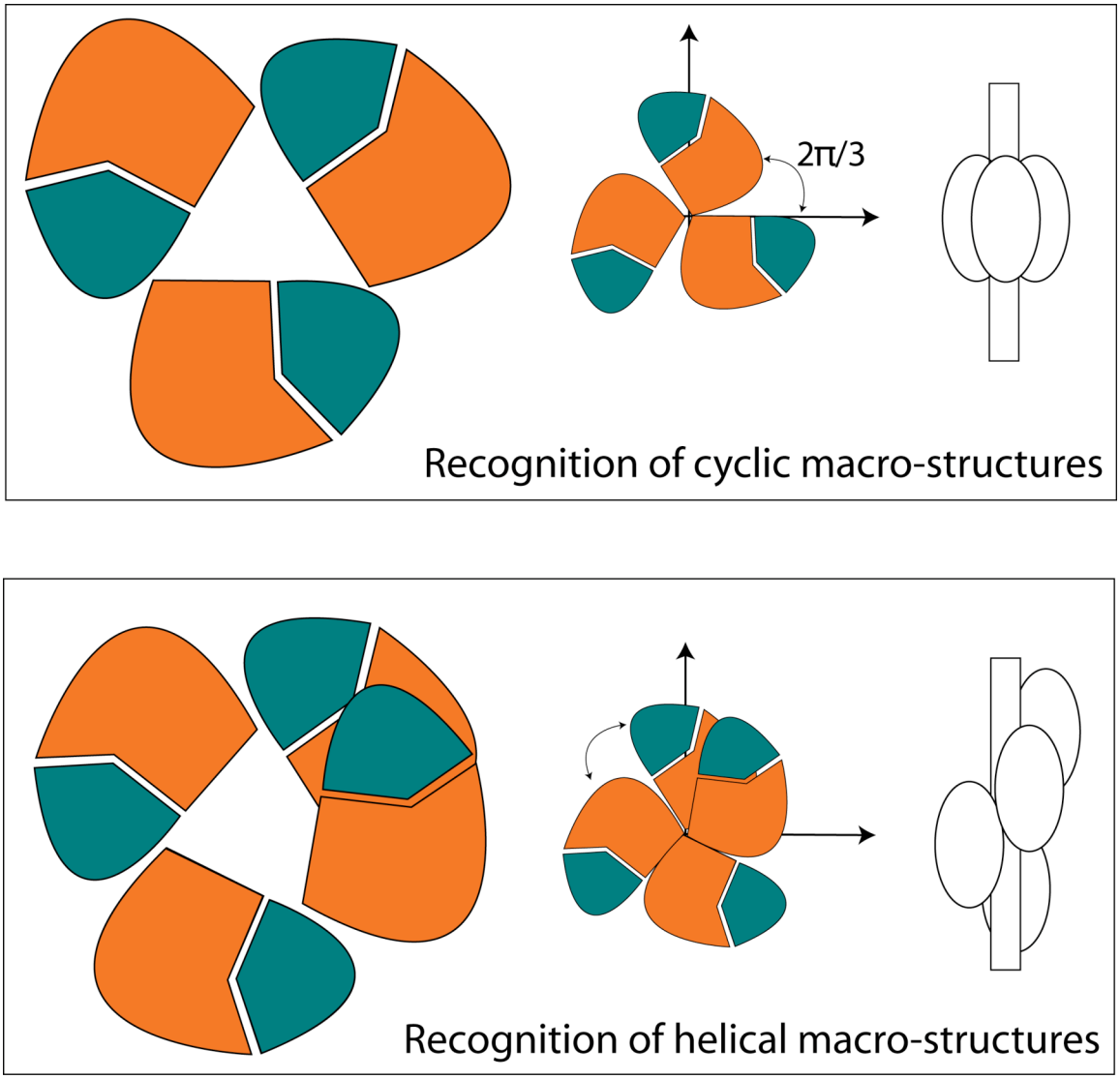

Figure 27: Prediction of the structure of circular or helical assemblies from binary interactions
On top: three subunits arranges in a circle. The angle of the rotation from one subunit to the next is a fraction of 2π and all subunits are in the same plan.

At the bottom: the angle of the rotation is not a multiple of 2π and the subunits cannot be arranged in a circle. However, it may be possible to arrange the subunits in a helical structures.

3.3.2 Rings

Every interaction does not form a circular assembly when repeated. For a ring to close, it is necessary (but not sufficient) that the rotation between two successive

subunits of the ring is a fraction of 2π (the rationale being that there is a finite number k of subunits placed evenly on the circle oriented in a regular manner). When these conditions are met (Figure 27), the subunit is moved back to its original orientation after applying the same transformation a finite number of times. In this case, the number of elements in the ring is a multiple of 2π. For instance, in the exosome, the subunit consists of two proteins (the exosome assembles in a trimer of dimers) and the rotation from one subunit to the next has an angle of $2\pi/3$. Then the number of subunits in the ring is a multiple of 3 and we can try to repeat the transformation 3+1 times to see if the subunit is transformed back to its original position. Hence, the knowledge of the structure of the subunit and the transformation from one subunit to the next are sufficient to indicate the possibility to form a ring and the number of times the subunit is repeated in the ring. As a rough pre-study, we searched amongst all interactions between domains of the same type those that could form a ring and proved the interest of the approach by detecting accurately subunits that can be arranged in circles. This idea can obviously be applied to larger subunits to determine if they could form rings.

3.3.3 Macro-helices

To test whether multiple copies of a protein can arrange in a helix, one must know the structure of the subunit and compute the angle of the rotation from one subunit to the next one (all transformations that conserve distances in 3D-space can be decomposed in a rotation and a translation). Then, it is necessary to find how many repetitions k of the subunit are needed to cover more than 2π (Figure 27). If the assembly built by repeating the transformation 2k times is valid, one detects that it is possible to arrange two turns of the helix in a valid manner, suggesting that the helix is valid.

Thus, with the help of some basic geometry principles, the structure of one subunit and the interaction between two subunits (only the interface matters) suffices to

recognize cyclic and helical macro-structures with obvious consequences on the stoichiometry of those assemblies.

3.4. A glimpse at the stoichiometry of any complex

The determination of protein complexes is usually limited to the list of its components. The stoichiometry is rarely known before the structure is solved. It is obvious that proteins are limited in the nature and number of contacts they can make: the number of interfaces per structure is limited and those interfaces are specific to few compatible binding partners. Basically, it means that given a list of proteins, not just any quantity of these proteins can form a structure: interfaces will be occupied and multiple copies of certain proteins can be required to achieve interactions.

The same phenomenon occurs in chemistry: a molecule of water is composed of atoms of oxygen and hydrogen. This is the composition of a water molecule without information about the stoichiometry of atoms. In this case, oxygen has two interfaces available for hydrogen and hydrogen has one interface available for oxygen. Thus, the combination that best fills the interface consists of two hydrogens for one oxygen. A similar reasoning can be applied to protein structures: in this study we listed domains that form interactions (like the H-O bound in chemistry for example) and know the interactions that are seen simultaneously in a complex, which provides information about protein-protein interfaces (oxygen has two slots for an interaction with hydrogen in chemistry). This information can be used to predict the number of copies of each domain needed to saturate the interfaces and thus determine the stoichiometry of the complex.

The knowledge of protein-protein interactions is crucial. As we showed in this work, it can be used for the prediction of the structure of protein assemblies and moreover it may apply to many other fields of research, in particular in the

determination of contacts between proteins or stoichiometry of proteins within complexes.

4. Conclusion

In the present work, the structures of protein complexes were predicted by combining pairwise orientations of subunits predicted by homology. Finding the most accurate assembly from the mass of possible predictions is a difficult task, but some parameters can efficiently evaluate the quality of the predictions. Moreover, when the method is combined with interaction and structural data, the predictions are limited to those that are the most relevant. This approach is still at an early stage and many improvements are possible that will undoubtedly make the approach even more reliable.

References

Alexandrov N, Shindyalov I (2003) PDP: protein domain parser. Bioinformatics 19(3): 429-430.

Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33(Database issue): D418-424.

Aloy P, Russell RB (2002a) The third dimension for protein interactions and complexes. Trends Biochem Sci 27(12): 633-638.

Aloy P, Russell RB (2002b) Potential artefacts in protein-interaction networks. FEBS Lett 530(1-3): 253-254.

Aloy P, Russell RB (2003) InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics 19(1): 161-162.

Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. Nat Biotechnol 22(10): 1317-1321.

Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol In Press.

Aloy P, Querol E, Aviles FX, Sternberg MJE (2001) Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 311: 395-408.

Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. J Mol Biol 332(5): 989-998.

Aloy P, Ciccarelli FD, Leutwein C, Gavin AC, Superti-Furga G et al. (2002) A complex prediction: three-dimensional model of the yeast exosome. EMBO Reports 3: 628-635.

Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C et al. (2004) Structure-based assembly of protein complexes in yeast. Science 303(5666): 2026-2029.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17): 3389-3402.

Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 310(2): 311-325.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29(1): 37-40.

Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307(1): 447-463.

Arnau V, Mars S, Marin I (2005) Iterative cluster analysis of protein interaction data. Bioinformatics 21(3): 364-378.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1): 25-29.

Bader GD, Hogue CW (2002) Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol 20(10): 991-997.

Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.

Baker ML, Jiang W, Wedemeyer WJ, Rixon FJ, Baker D et al. (2006) Ab initio modeling of the herpesvirus VP26 core domain assessed by CryoEM density. PLoS Comput Biol 2(10): e146.

Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289(5481): 905-920.

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439): 509-512.

Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ et al. (2007) Still stratus not altocumulus: further evidence against the date/party hub distinction. PLoS Biol 5(6): e154.

Beck M, Forster F, Ecke M, Plitzko JM, Melchior F et al. (2004) Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. Science 306(5700): 1387-1390.

Beuron F, Flynn TC, Ma J, Kondo H, Zhang X et al. (2003) Motions and negative cooperativity between p97 domains revealed by cryo-electron microscopy and quantised elastic deformational model. J Mol Biol 327(3): 619-629.

Blundell TL, Johnson MS (1993) Catching a common fold. Protein Sci 2(6): 877-883.

Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 280(1): 1-9.

Bork P (1989) Recognition of functional regions in primary structures using a set of property patterns. FEBS Lett 257(1): 191-195.

Bork P (1991) Shuffled domains in extracellular proteins. FEBS Lett 286(1-2): 47-54.

Bork P, Dandekar T, Eisenhaber F, Huynen M (1998) Characterization of targeting domains by sequence analysis: glycogen-binding domains in protein phosphatases. J Mol Med 76(2): 77-79.

Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J, 3rd (2005) The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci 62(4): 435-445.

Borukhov S, Nudler E (2003) RNA polymerase holoenzyme: structure, function and biological implications. Curr Opin Microbiol 6(2): 93-100.

Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7: 488.

Bushnell DA, Westover KD, Davis RE, Kornberg RD (2004) Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. Science 303(5660): 983-988.

Cao K, Nakajima R, Meyer HH, Zheng Y (2003) The AAA-ATPase Cdc48/p97 regulates spindle disassembly at the end of mitosis. Cell 115(3): 355-367.
Ceulemans H, Russell RB (2004) Fast fitting of atomic structures to low resolution electron density maps by surface overlap maximization. J Mol Biol In Press.
Chacon P, Wriggers W (2002) Multi-resolution contour-based fitting of macromolecular structures. J Mol Biol 317(3): 375-384.
Chen HT, Hahn S (2003) Binding of TFIIB to RNA polymerase II: Mapping the binding site for the TFIIB zinc ribbon domain within the preinitiation complex. Mol Cell 12(2): 437-447.
Cho S, Park SG, Lee DH, Park BC (2004) Protein-protein interaction networks: from interactions to networks. J Biochem Mol Biol 37(1): 45-52.
Chothia C (1992) Proteins. One thousand families for the molecular biologist. Nature 357(6379): 543-544.
Chung WH, Craighead JL, Chang WH, Ezeokonkwo C, Bareket-Samish A et al. (2003) RNA polymerase II/TFIIF structure and conserved organization of the initiation complex. Mol Cell 12(4): 1003-1013.
Clackson T, Wells JA (1995) A hot spot of binding energy in a hormone-receptor interface. Science 267(5196): 383-386.
Copley RR, Doerks T, Letunic I, Bork P (2002) Protein domain analysis in the era of complete genomes. FEBS Lett 513(1): 129-134.
Corpet F, Gouzy J, Kahn D (1998) The ProDom database of protein domain families. Nucleic Acids Res 26(1): 323-326.
Cramer P, Bushnell DA, Kornberg RD (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science 292(5523): 1863-1876.
Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23(9): 324-328.
Darnell SJ, Page D, Mitchell JC (2007) An automated decision-tree approach to predicting protein interaction hot spots. Proteins 68(4): 813-823.
Davis FP, Sali A (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics 21(9): 1901-1907.
del Sol Mesa A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues. J Mol Biol 326(4): 1289-1302.
DeLaBarre B, Brunger AT (2003) Complete structure of p97/valosin-containing protein reveals communication between nucleotide domains. Nat Struct Biol 10(10): 856-863.
Devos D, Russell RB (2007) A more complete, complexed and structured interactome. Curr Opin Struct Biol 17(3): 370-377.
Dezso Z, Oltvai ZN, Barabasi AL (2003) Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. Genome Res 13(11): 2450-2454.
Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA (2007) The protein folding problem: when will it be solved? Curr Opin Struct Biol 17(3): 342-346.
Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9): 755-763.

Ekman D, Light S, Bjorklund AK, Elofsson A (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biol 7(6): R45.
Elcock AH (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 312(4): 885-896.
Enright AJ, Iliopoulos, I. L, Kyrpides, N. C., Ouzounis, C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 25-26.
Fiaux J, Bertelsen EB, Horwich AL, Wuthrich K (2002) NMR analysis of a 900K GroEL GroES complex. Nature 418(6894): 207-211.
Finn RD, Marshall M, Bateman A (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 21(3): 410-412.
Gaasterland T (1998) Structural genomics: bioinformatics in the driver's seat. Nat Biotechnol 16(7): 625-627.
Gagneur J, David L, Steinmetz LM (2006) Capturing cellular machines by systematic screens of protein complexes. Trends Microbiol 14(8): 336-339.
Gao Y, Wang R, Lai L (2004) Structure-based method for analyzing protein-protein interfaces. J Mol Model 10(1): 44-54.
Gavin AC, Aloy P, Grandi P, Krause R, Boesche M et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature In press.
Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868): 141-147.
Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B et al. (2003) A protein interaction map of Drosophila melanogaster. Science 302(5651): 1727-1736.
Golbik R, Lupas AN, Koretke KK, Baumeister W, Peters J (1999) The Janus face of the archaeal Cdc48/p97 homologue VAT: protein folding versus unfolding. Biol Chem 380(9): 1049-1062.
Goll J, Uetz P (2006) The elusive yeast interactome. Genome Biol 7(6): 223.
Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 28(7): 849-857.
Govindarajan S, Recabarren R, Goldstein RA (1999) Estimating the total number of protein folds. Proteins 35(4): 408-414.
Gray JJ (2006) High-resolution protein-protein docking. Curr Opin Struct Biol 16(2): 183-193.
Greene LH, Lewis TE, Addou S, Cuff A, Dallman T et al. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35(Database issue): D291-297.
Groll M, Ditzel L, Lowe J, Stock D, Bochtler M et al. (1997) Structure of 20S proteasome from yeast at 2.4 A resolution. Nature 386(6624): 463-471.

Grunewald K, Desai P, Winkler DC, Heymann JB, Belnap DM et al. (2003) Three-dimensional structure of herpes simplex virus from cryo-electron tomography. Science 302(5649): 1396-1398.
Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320(2): 369-387.
Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A et al. (2006) MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 34(Database issue): D436-441.
Guo JT, Xu D, Kim D, Xu Y (2003) Improving the performance of DomainParser for structural domain partition using neural network. Nucleic Acids Res 31(3): 944-952.
Haeusler RA, Engelke DR (2006) Spatial organization of transcription by RNA polymerase III. Nucleic Acids Res 34(17): 4826-4836.
Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M (2005) Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol 23(7): 839-844.
Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF et al. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995): 88-93.
Henrick K, Thornton JM (1998) PQS: a protein quaternary structure file server. Trends Biochem Sci 23(9): 358-361.
Hernandez H, Dziembowski A, Taverner T, Seraphin B, Robinson CV (2006) Subunit architecture of multimeric complexes isolated directly from cells. EMBO Rep 7(6): 605-610.
Hetzer M, Meyer HH, Walther TC, Bilbao-Cortes D, Warren G et al. (2001) Distinct AAA-ATPase p97 complexes function in discrete steps of nuclear assembly. Nat Cell Biol 3(12): 1086-1091.
Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868): 180-183.
Holland TA, Veretnik S, Shindyalov IN, Bourne PE (2006) Partitioning protein structures into domains: why is it so difficult? J Mol Biol 361(3): 562-590.
Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233(1): 123-138.
Holm L, Sander C (1994) The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 22(17): 3600-3609.
Holm L, Sander C (1997) Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 25(1): 231-234.
Holm L, Sander C (1998) Dictionary of recurrent domains in protein structures. Proteins 33(1): 88-96.
Hoog JL, Schwartz C, Noon AT, O'Toole ET, Mastronarde DN et al. (2007) Organization of interphase microtubules in fission yeast analyzed by electron tomography. Dev Cell 12(3): 349-361.

Hubbard SJ, Gross KH, Argos P (1994) Intramolecular cavities in globular proteins. Protein Eng 7(5): 613-626.
Huyton T, Pye VE, Briggs LC, Flynn TC, Beuron F et al. (2003) The crystal structure of murine p97/VCP at 3.6A. J Struct Biol 144(3): 337-348.
Inbar Y, Benyamini H, Nussinov R, Wolfson HJ (2005) Prediction of multimolecular assemblies by multiple docking. J Mol Biol 349(2): 435-447.
Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98(8): 4569-4574.
Janin J, Chothia C (1990) The structure of protein-protein recognition sites. J Biol Chem 265(27): 16027-16030.
Jefferson ER, Walsh TP, Roberts TJ, Barton GJ (2007) SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. Nucleic Acids Res 35(Database issue): D580-589.
Jeffrey PD, Tong L, Pavletich NP (2000) Structural basis of inhibition of CDK-cyclin complexes by INK4 inhibitors. Genes Dev 14(24): 3115-3125.
Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. Nature 443(7111): 594-597.
Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411(6833): 41-42.
Johnson ES, Ma PC, Ota IM, Varshavsky A (1995) A proteolytic pathway that recognizes ubiquitin as a degradation signal. J Biol Chem 270(29): 17442-17456.
Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93: 13-20.
Jones S, Marin A, Thornton JM (2000) Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Eng 13(2): 77-82.
Kann MG, Jothi R, Cherukuri PF, Przytycka TM (2007) Predicting protein domain interactions from coevolution of conserved regions. Proteins 67(4): 811-820.
Kawashima T, Berthet-Colominas C, Wulff M, Cusack S, Leberman R (1996) The structure of the Escherichia coli EF-Tu.EF-Ts complex at 2.5 A resolution. Nature 379(6565): 511-518.
Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A et al. (2007) IntAct--open source resource for molecular interaction data. Nucleic Acids Res 35(Database issue): D561-565.
Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314(5807): 1938-1941.
Kim WK, Ison JC (2005) Survey of the geometric association of domain-domain interfaces. Proteins 61(4): 1075-1088.
King AD, Przulj N, Jurisica I (2004) Protein complex prediction via cost-based clustering. Bioinformatics 20(17): 3013-3020.

Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420(6912): 218-223.
Korkin D, Davis FP, Alber F, Luong T, Shen MY et al. (2006) Structural modeling of protein interactions by analogy: application to PSD-95. PLoS Comput Biol 2(11): e153.
Kortemme T, Kim DE, Baker D (2004) Computational alanine scanning of protein-protein interfaces. Sci STKE 2004(219): pl2.
Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60(Pt 12 Pt 1): 2256-2268.
Krogan NJ, Cagney G, Yu H, Zhong G, Guo X et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440(7084): 637-643.
Krueger JK, Gallagher SC, Wang CA, Trewhella J (2000) Calmodulin remains extended upon binding to smooth muscle caldesmon: a combined small-angle scattering and fourier transform infrared spectroscopy study. Biochemistry 39(14): 3979-3987.
Kumar A, Snyder M (2002) Protein complexes take the bait. Nature 415(6868): 123-124.
Kundrotas PJ, Alexov E (2007) PROTCOM: searchable database of protein complexes enhanced with domain-domain structures. Nucleic Acids Res 35(Database issue): D575-579.
Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E et al. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33(Web Server issue): W299-302.
Landgraf R, Xenarios I, Eisenberg D (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 307(5): 1487-1502.
Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33(Web Server issue): W89-93.
Lesk AM, Chothia C (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 136(3): 225-270.
Li L, Zhao B, Cui Z, Gan J, Sakharkar MK et al. (2006) Identification of hot spot residues at protein-protein interface. Bioinformation 1(4): 121-126.
Li S, Armstrong CM, Bertin N, Ge H, Milstein S et al. (2004) A map of the interactome network of the metazoan C. elegans. Science 303(5657): 540-543.
Liu Q, Greimann JC, Lima CD (2006) Reconstitution, activities, and structure of the eukaryotic RNA exosome. Cell 127(6): 1223-1237.
Lorentzen E, Walter P, Fribourg S, Evguenieva-Hackenberg E, Klug G et al. (2005) The archaeal exosome core is a hexameric ring structure with three catalytic subunits. Nat Struct Mol Biol 12(7): 575-581.

Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins 49(3): 350-364.

Ludtke SJ, Jakana J, Song JL, Chuang DT, Chiu W (2001) A 11.5 A single particle reconstruction of GroEL using EMAN. J Mol Biol 314(2): 253-262.

Malhotra A, Harvey SC (1994) A quantitative model of the Escherichia coli 16 S RNA in the 30 S ribosomal subunit. J Mol Biol 240(4): 308-340.

Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY et al. (2005) CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res 33(Database issue): D192-196.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285(5428): 751-753.

Matadeen R, Patwardhan A, Gowen B, Orlova EV, Pape T et al. (1999) The Escherichia coli large ribosomal subunit at 7.5 A resolution. Structure 7(12): 1575-1583.

Medalia O, Weber I, Frangakis AS, Nicastro D, Gerisch G et al. (2002) Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. Science 298(5596): 1209-1213.

Mendez R, Leplae R, Lensink MF, Wodak SJ (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. Proteins 60(2): 150-169.

Mewes HW, Frishman D, Gruber C, Geier B, Haase D et al. (2000) MIPS: a database for genomes and protein sequences. Nucleic Acids Res 28(1): 37-40.

Milgram S (1967) The small-world problem. Psychology Today 1: 61-67.

Miranker A, Karplus M (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. Proteins 11(1): 29-34.

Morgan DH, Kristensen DM, Mittelman D, Lichtarge O (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. Bioinformatics 22(16): 2049-2050.

Morgunova E, Tuuttila A, Bergmann U, Isupov M, Lindqvist Y et al. (1999) Structure of human pro-matrix metalloproteinase-2: activation mechanism revealed. Science 284(5420): 1667-1670.

Mrowka R, Patzak A, Herzel H (2001) Is there a bias in proteome research? Genome Res 11(12): 1971-1973.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31(1): 315-318.

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4): 536-540.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB et al. (1997) CATH--a hierarchic classification of protein domain structures. Structure 5(8): 1093-1108.

Orengo CA, Bray JE, Buchan DW, Harrison A, Lee D et al. (2002) The CATH protein family database: a resource for structural and functional annotation of genomes. Proteomics 2(1): 11-21.

Orlowski J, Kaczanowski S, Zielenkiewicz P (2007) Overrepresentation of interactions between homologous proteins in interactomes. FEBS Lett 581(1): 52-56.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96(6): 2896-2901.

Park S, Isaacson R, Kim HT, Silver PA, Wagner G (2005) Ufd1 exhibits the AAA-ATPase fold with two distinct ubiquitin interaction sites. Structure 13(7): 995-1005.

Park SY, Beel BD, Simon MI, Bilwes AM, Crane BR (2004) In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved. Proc Natl Acad Sci U S A 101(32): 11646-11651.

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271(4): 511-523.

Pazos F, Ranea JA, Juan D, Sternberg MJ (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol 352(4): 1002-1015.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96(8): 4285-4288.

Pereira-Leal JB, Audit B, Peregrin-Alvarez JM, Ouzounis CA (2005) An exponential core in the heart of the yeast protein interaction network. Mol Biol Evol 22(3): 421-425.

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13(10): 2363-2371.

Peters JM, Harris JR, Lustig A, Muller S, Engel A et al. (1992) Ubiquitous soluble Mg(2+)-ATPase complex. A structural study. J Mol Biol 223(2): 557-571.

Petoukhov MV, Svergun DI (2005) Global rigid body modeling of macromolecular complexes against small-angle scattering data. Biophys J 89(2): 1237-1250.

Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP et al. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 32(Database issue): D217-222.

Prabu MM, Suguna K, Vijayan M (1999) Variability in quaternary association of proteins with the same tertiary fold: a case study and rationalization involving legume lectins. Proteins 35(1): 58-69.

Przulj N, Corneil DG, Jurisica I (2004) Modeling interactome: scale-free or geometric? Bioinformatics 20(18): 3508-3515.
Pye VE, Beuron F, Keetch CA, McKeown C, Robinson CV et al. (2007) Structural insights into the p97-Ufd1-Npl4 complex. Proc Natl Acad Sci U S A 104(2): 467-472.
Raijmakers R, Egberts WV, van Venrooij WJ, Pruijn GJ (2002) Protein-protein interactions between human exosome components support the assembly of RNase PH-type subunits into a six-membered PNPase-like ring. J Mol Biol 323(4): 653-663.
Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C et al. (2001) The protein-protein interaction map of Helicobacter pylori. Nature 409(6817): 211-215.
Ranson NA, Farr GW, Roseman AM, Gowen B, Fenton WA et al. (2001) ATP-bound states of GroEL captured by cryo-electron microscopy. Cell 107(7): 869-879.
Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586): 1551-1555.
Richardson JS (1981) The Anatomy and taxonomy of protein structure. Advances in Protein Chemistry 34: 167-338.
Ringe D (1995) What makes a binding site a binding site? Curr Opin Struct Biol 5(6): 825-829.
Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383: 66-93.
Rossmann MG (2000) Fitting atomic models into electron-microscopy maps. Acta Crystallogr D Biol Crystallogr 56 (Pt 10): 1341-1349.
Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437(7062): 1173-1178.
Russell J, Zomerdijk JC (2006) The RNA polymerase I transcription machinery. Biochem Soc Symp(73): 203-216.
Russell RB, Barton GJ (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 14(2): 309-323.
Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3): 779-815.
Sali A, Glaeser R, Earnest T, Baumeister W (2003) From words to literature in structural proteomics. Nature 422(6928): 216-225.
Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32(Database issue): D449-451.
Sato T, Yamanishi Y, Kanehisa M, Toh H (2005) The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21(17): 3482-3489.

Schultz J, Copley RR, Doerks T, Ponting CP, Bork P (2000) SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res 28(1): 231-234.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F et al. (2005) The FoldX web server: an online force field. Nucleic Acids Res 33(Web Server issue): W382-388.

Sharon M, Witt S, Glasmacher E, Baumeister W, Robinson CV (2007) Mass spectrometry reveals the missing links in the assembly pathway of the bacterial 20 S proteasome. J Biol Chem 282(25): 18448-18457.

Sherr CJ, Roberts JM (1999) CDK inhibitors: positive and negative regulators of G1-phase progression. Genes Dev 13(12): 1501-1512.

Silberstein M, Dennis S, Brown L, Kortvelyesi T, Clodfelter K et al. (2003) Identification of substrate binding sites in enzymes by computational solvent mapping. J Mol Biol 332(5): 1095-1113.

Sims RJ, 3rd, Belotserkovskaya R, Reinberg D (2004) Elongation by RNA polymerase II: the short and long of it. Genes Dev 18(20): 2437-2468.

Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28(18): 3442-3444.

Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7): 951-960.

Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28(3): 405-420.

Spahn CM, Beckmann R, Eswar N, Penczek PA, Sali A et al. (2001) Structure of the 80S ribosome from Saccharomyces cerevisiae--tRNA- ribosome and subunit-subunit interactions. Cell 107(3): 373-386.

Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A 100(21): 12123-12128.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A et al. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34(Database issue): D535-539.

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

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122(6): 957-968.

Sternberg MJ, Bates PA, Kelley LA, MacCallum RM (1999) Progress in protein structure prediction: assessment of CASP3. Curr Opin Struct Biol 9(3): 368-373.

Stoffler D, Feja B, Fahrenkrog B, Walz J, Typke D et al. (2003) Cryo-electron tomography provides novel insights into nuclear pore architecture: implications for nucleocytoplasmic transport. J Mol Biol 328(1): 119-130.

Suhrer SJ, Wiederstein M, Sippl MJ (2007) QSCOP--SCOP quantified by structural relationships. Bioinformatics 23(4): 513-514.

Sun Z, Reid KB, Perkins SJ (2004) The dimeric and trimeric solution structures of the multidomain complement protein properdin by X-ray scattering, analytical ultracentrifugation and constrained modelling. J Mol Biol 343(5): 1327-1343.

Svergun DI, Koch MH (2002) Advances in structure analysis using small-angle scattering in solution. Curr Opin Struct Biol 12(5): 654-660.

Topf M, Baker ML, John B, Chiu W, Sali A (2005) Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. J Struct Biol 149(2): 191-203.

Topf M, Baker ML, Marti-Renom MA, Chiu W, Sali A (2006) Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. J Mol Biol 357(5): 1655-1668.

Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A (2005) Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 61 Suppl 7: 27-45.

Tung CH, Yang JM (2007) fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies. Nucleic Acids Res 35(Web Server issue): W438-443.

Tung CS, Walsh DA, Trewhella J (2002) A structural model of the catalytic subunit-regulatory subunit dimeric complex of the cAMP-dependent protein kinase. J Biol Chem 277(14): 12423-12431.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770): 623-627.

Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST et al. (2002) Monte Carlo simulations of the peptide recognition at the consensus binding site of the constant fragment of human immunoglobulin G: the energy landscape analysis of a hot spot at the intermolecular interface. Proteins 48(3): 539-557.

Vincent JJ, Tai CH, Sathyanarayana BK, Lee B (2005) Assessment of CASP6 predictions for new and nearly new fold targets. Proteins 61 Suppl 7: 67-83.

Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA (2004) Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 14(2): 208-216.

Volkmann N, Hanein D (1999) Quantitative fitting of atomic models into observed densities derived by electron microscopy. J Struct Biol 125(2-3): 176-184.

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P et al. (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 31(1): 258-261.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887): 399-403.

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T et al. (2007) STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35(Database issue): D358-362.

Vong QP, Cao K, Li HY, Iglesias PA, Zheng Y (2005) Chromosome alignment and segregation regulated by ubiquitination of survivin. Science 310(5753): 1499-1504.

Wall ME, Gallagher SC, Trewhella J (2000) Large-scale shape changes in proteins and macromolecular complexes. Annu Rev Phys Chem 51: 355-380.

Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 8(2): 127-134.

Watts DJ (1999) Small Worlds: The Dynamics of Networks Between Order and Randomness.

Wetlaufer DB (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci U S A 70(3): 697-701.

Winter C, Henschel A, Kim WK, Schroeder M (2006) SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Res 34(Database issue): D310-314.

Woychik NA, Hampsey M (2002) The RNA polymerase II machinery: structure illuminates function. Cell 108(4): 453-463.

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM et al. (2000) DIP: the database of interacting proteins. Nucleic Acids Res 28(1): 289-291.

Yook SH, Oltvai ZN, Barabasi AL (2004) Functional and topological characterization of protein interaction networks. Proteomics 4(4): 928-942.

Yu H (1999) Extending the size limit of protein nuclear magnetic resonance. Proc Natl Acad Sci U S A 96(2): 332-334.

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M et al. (2002) MINT: a Molecular INTeraction database. FEBS Lett 513(1): 135-140.

Zhang X, Shaw A, Bates PA, Newman RH, Gowen B et al. (2000) Structure of the AAA ATPase p97. Mol Cell 6(6): 1473-1484.

Zhang Y, Arakaki AK, Skolnick J (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 61 Suppl 7: 91-98.

Zhao J, Hoye E, Boylan S, Walsh DA, Trewhella J (1998) Quaternary structures of a catalytic subunit-regulatory subunit dimeric complex and the holoenzyme of the cAMP-dependent protein kinase by neutron contrast variation. J Biol Chem 273(46): 30448-30459.