Università degli Studi di Salerno

Facoltà di Scienze Matematiche Fisiche e Naturali

IN SILICO STUDY OF PROTEIN-PROTEIN INTERACTIONS

Philosophiae Doctor Thesis
in
Scienza e Tecnologie dell'Industria Chimica, Alimentare e Farmaceutica – Indirizzo Chimico
Ciclo XI

Supervisor

Prof. Luigi Cavallo

Contro-relatori

Prof. Gianluca Sbardella

Prof. Daniele Sblattero
(Università Piemonte Orientale)

Co-supervisor

Dr. Romina Oliva
(Università Parthenope)

Coordinatore

Prof. Gaetano Guerra

PhD candidate

Anna Vangone

2009-2012

INDEX

ABSTRACT

Protein-protein interactions are at the basis of many of the most important molecular processes in the cell, which explains the constantly growing interest within the scientific community for the structural characterization of protein complexes.[1] However, experimental knowledge of the 3D structure of the great majority of such complexes is missing, and this spurred their accurate prediction through molecular docking simulations, one of the major challenges in the field of structural computational biology and bioinformatics.[2,3]

My PhD work aims to contribute to the field, by providing novel computational instruments and giving useful insight on specific case studies in the field. In particular, in the first part of my PhD thesis, I present novel methods I developed: i) for analysing and comparing the 3D structure of protein complexes, to immediately extract useful information on the interaction based on a contact map visualization (COCOMAPS[4] web tool, Chapter 2), and ii) for analysing a set of multiple docking solutions, to single out the key inter-residue contacts and to distinguish native-like solutions from the incorrect ones (CONS-COCOMAPS[5] web tool and CONS-RANK program, Chapter 3 and 4, respectively).

In the second part of the thesis, these methods have been applied, in combination with classical state-of-art computational biology techniques, to predict and analyse the binding mode in real biological systems, related to particular diseases. This part of the work has been afforded in collaboration with experimental groups, to take advantage of specific biological information on the systems under study. In particular, the interaction between proteins involved in the autoimmune response in celiac disease[6,7] (Chapters 5 and 6) has been studied in collaboration with the group directed by Prof. Sblattero, University of Piemonte Orientale (Italy) and the group directed by Prof. Esposito, University of Salerno (Italy). In addition, recognition properties of the FXa enzymatic system[8] has been studied through dynamic characterization of a FXa pathogenic mutant that causes problems in the blood coagulation cascade (Chapter 7). This study has been performed in collaboration with the group directed by Prof. De Cristofaro, Catholic University School of Medicine, Rome (Italy) and the group

directed by Prof. Peyvandi, Ospedale Maggiore Policlinico and Università degli Studi di Milano (Italy).

Finally, during my PhD I spent seven months in the groups of Prof. Charlotte Deane, Department of Statistics, University of Oxford (UK). During this period I studied the geometrical features of the proteins' regions most recurrent in the protein-protein interaction, the loops, clarifying some structural aspects of them in one of the most important and huge class of proteins: the membrane proteins (Appendix 1).

Web tools and programs:

COCOMAPS[4] web tool freely available at:
https://www.molnac.unisa.it/BioTools/cocomaps/

CONS-COCOMAPS[5] web tool freely available at:
https://www.molnac.unisa.it/BioTools/conscocomaps/

CONS-RANK program available upon request from the autors.

CHAPTER 1 - Protein-protein interactions and molecular docking

1.1 - Introduction to the protein-protein interaction

Biological complexes: preliminary remarks

The thousands of proteins expressed in the cells perform many of their functions through interactions with other proteins. The protein-protein interactions are intrinsic to every cellular process; in fact, protein complexes have been implicated as an essential component in the major research topics in biology and medicine, such as DNA replication, transcription, translation, splicing, secretion, cell cycle control, signal transduction, and intermediary metabolism.[1,9] Therefore, the analysis at a molecular level of proteins in complexes is a matter of interest for biochemists, but also geneticists, cell biologists, developmental biologists, molecular biologists and biophysicists.[10]

Protein-protein interactions play diverse roles and differ based on the composition, affinity, lifetime and nature of the association. In the permanent/obligate complexes the interactions are usually very stable and the interacting proteins are not found as stable structures on their own *in vivo*, while in the transient/non-obligate complexes there are transient interactions that associate and dissociate *in vivo* and the interacting proteins can also exist in the unbound form. Obligate complexes can be further divided into homodimers, i.e. interactions occurring between identical chains, heterodimers and multimers.[11] It has been observed that different classes of association exhibit different physical and chemical properties in their interaction sites and different functions.[12-14] So, for example, interactions in intracellular signaling are expected to be transient, since their function requires a ready association and dissociation, while an antigen-antibody interaction is generally permanent. Anyway, it is important to note that many protein-protein interactions do not fall into distinct types. Rather, a continuum exists between non-obligate and obligate interactions, and the stability of all complexes very much depends on the physiological conditions and the environment.[11]

In the last years, experimental and theoretical work has been devoted to unravel the

principles of protein-protein interactions.[15-20]

The formation of biological complexes is driven by the free energy of the complex (determined by physicochemical and geometrical interface properties) and the concentration of the protein components.[11] The association of two proteins, in fact, relies on an encounter of the interacting surfaces, requiring co-localization in time and space. Generally a protein resides in a crowded environment with many potential binding partners with different surface properties; therefore, during the evolution the surfaces presumably evolve to optimize the interacting efficacy.[21] When proteins collide, they do not diffuse away immediately (kinetic experimental evidence from Northup *et al.*[22] and Wells[23]); instead, they are held loosely, rolling on one another and thereby sampling considerably more surface area than would be the case for a single elastic collision; this allows them time to become reorientated and repositionated on the surface or to adjust their shape to fit together more tightly (Figure 1).[24] Recent studies are beginning to describe the dynamic of the assembly processes and to show that these non specific collisions producing transient 'encounter complexes' play an important role in macromolecular associaction.[25] The role of long-range forces in bringing molecules together has been studied from both experimental and theoretical viewpoints,[26,27] suggesting the electrostatic interactions to be predominant.[25]

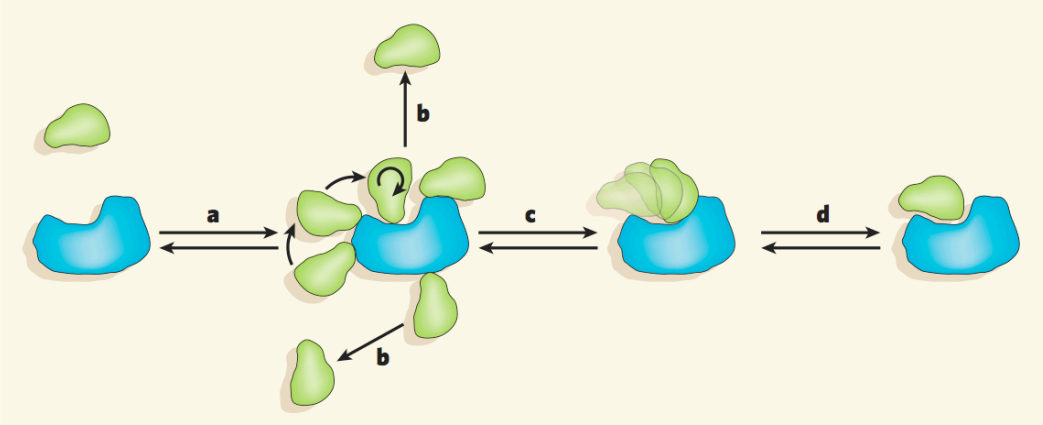


Figure 1. Protein-protein interactions
Equilibrium steps in a possible mechanism for protein–protein association. a) Formation of transient encounter complexes by nonspecific collisions, guided mostly by electrostatic interactions. b) Many encounter complexes separate rapidly. c) Some productive encounter complexes reorientate and come closer to the final, specific orientation, guided mostly by desolvation, as water molecules move away from the protein surfaces. d) Formation of the specific complex, with final fitting of interacting surfaces.[24]

In this scenario, it is extremely valuable to obtain structural information for a complete understanding of both the biochemical nature of the process for which the components come together, and the facilitated design of compounds that might influence it. In particular, the structural characterization of a protein-protein interface includes the identification of interatomic hydrogen bonds, of salt bridges, of hydrophobic interactions, determination to the interaction surface area and possibly the presence of bridging water molecules[28,29] The combination of all this information about the network of interactions defines the nature of the binding site and makes it possible to point out the residue-residue contacts with a key role in the interaction.

Here below it is reported an example of a protein-protein interface characterization for the complex between the hemagglutinin (HA) and its antibody HC45.[30] This antigen-antibody complex has a fundamental role in one of the most common world diseases: the influenza. Hemagglutinin, in fact, is the influenza virus glycoprotein that interacts with infectivity-neutralizing antibodies. It has a primary role in influenza infection mediating the binding of the virus to its cellular receptor. Over the years, amino acids substitution that arise by mutations in the genes for HA lead to escape of immune surveillance and recurrent epidemics - this process is called antigenic drift. So, the structural study of the complexes between HA and its antibodies is fundamental to understand the mechanism of the infection and to ensure the development vaccines of variants closely related to the circulating virus. Fleury *at al.*[30] reported the structure of the X31 HA-HC45 Fab complex (PDB entry: 1QFU; resolution 2.8 Å), describing the atomic characteristics of their interactions (Figure 2). Upon complex formation, a surface area of 1.840 $Å^2$ is buried; 36 amino acids participate in the intermolecular contacts, and 10 hydrogen bonds are established, involving antigen's residues such as Asp36 and Arg94. The HC45 epitope, i.e. the antigen binding site, comprises in total 17 residues. It was also proved that the mutation Asp63Asn (Figure 2, *right*) leads to escape from neutralization by HC45, underlining the importance of this residue in the interaction.

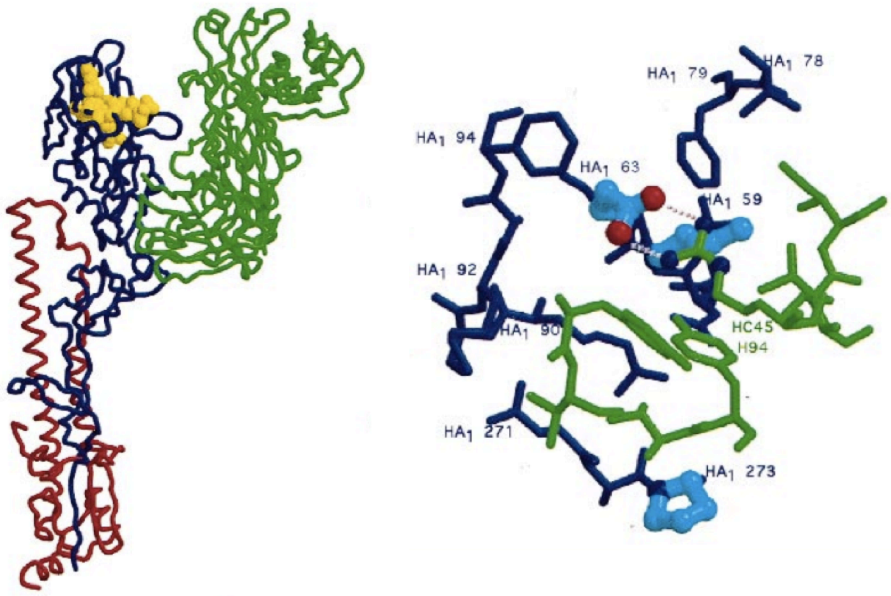


Figure 2
The X31 HA–HC45 Fab complex. *Left:* Ribbon diagram of the complex showing one HA monomer (the two domains HA1 and HA2 in blue and res, respectively) and the HC45 Fab (in green); the receptor binding site is shown in yellow. *Right:* Stick view of the HC45–HA interface (HA in blue, Fab in green). Of the 17 amino acids in the epitope, 12 are in the four polypeptide stretches of the HA1 chain (residues 59–63, 78–79, 90–94 and 271–273) and are represented here. HA residues substituted in mutants with decreased affinity for the HC45 antibody (Asp63 and Arg94) are highlighted in cyan; their nitrogen and oxygen atoms are colored in cyan and red, respectively. Hydrogen bonds involving atoms of these HA residues are shown as dotted lines.[30]

Structure of protein complexes

As shown in the example, the structural characterization of biological complexes has a supreme significance in the study of the system and in all the possible pharmaceutical and medicinal applications,[31] and although experimental methods for protein-structure determination have improved over the past decade, the number of structures for protein complex determined is still very little. Protein structures have been mainly achieved by two methods so far: X-ray crystallography and nuclear magnetic resonance (NMR). X-ray and NMR encounter difficulties to prepare complexes suitable for structural studies: by X-ray, the dynamics of the complex formation makes the crystallization difficult, while complexes of high molecular weight are difficult to deal with NMR.[18,32,33]

Due to the greater difficulty in obtaining suitable protein-protein complexes for the experimental determination, there is relatively little structural information available

about them compared to the proteins that exist as single chains or form permanent oligomers.[33] Hence, experimental studies are faced with outstanding technical difficulties and the number of solved complexes deposited in the Protein Data Bank[34] (PDB: www.rcsb.org/pdb) is still orders of magnitude smaller than structures of individual proteins, as show in Figure 3.[18,31]

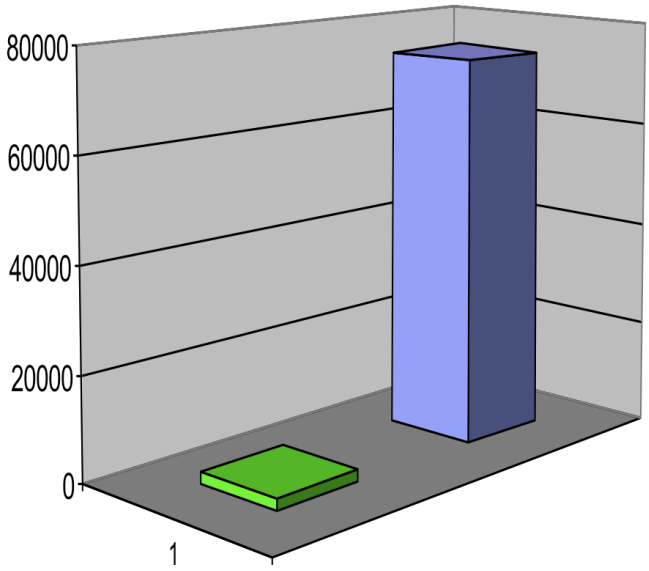


Figure 3
Number of X-ray structures of protein-protein complexes (in green) and single chain proteins (in blues) deposited in the wwPDB[34] within October 2011.

Despite this disproportion, the growing number of available experimental structures for protein-protein complexes in the years has allowed a statistical study of the properties and the chemical-physical forces that regulate protein-protein interactions (hydrophobicity, hydrogen bonding, electrostatic interactions, van der Waals interactions, and so on), that are useful information in the development of computational strategies helping in the structural prediction and characterization.[35] In fact, notwithstanding the practical difficulties, for a better understanding of the biological function of a protein, knowledge of its three-dimensional structure is fundamental. Therefore, it would be quite rewarding to have efficient and reliable computational algorithms available to predict correctly conformations of protein complexes based on the structures of the free molecules. Indeed, in the past two decades there was an emergence of a large variety of theoretical algorithms designed to predict the structures of protein-protein and protein-ligand complexes: a procedure

named molecular docking.[36]

Interest in protein docking is growing within the scientific community, and many interdisciplinary approaches are being applied to model, predict, and understand protein-protein interactions, one of the major challenge in the field of structural bioinformatics. [37]

1.2 - Approaches to the docking problem

The docking technique has the task of assembling two separate protein components (as the ones seen in Figure 4a and Figure 4b) into their biologically relevant complex structure (Figure 4c), giving a model of the way the two proteins bind each other.[38,39] Computational docking, if accurate and reliable, can therefore play an important role, both to infer functional properties and to guide new experiments. So, due to its potential applications in generating models of molecular complexes, it has attracted a vast deal of attention.[40]

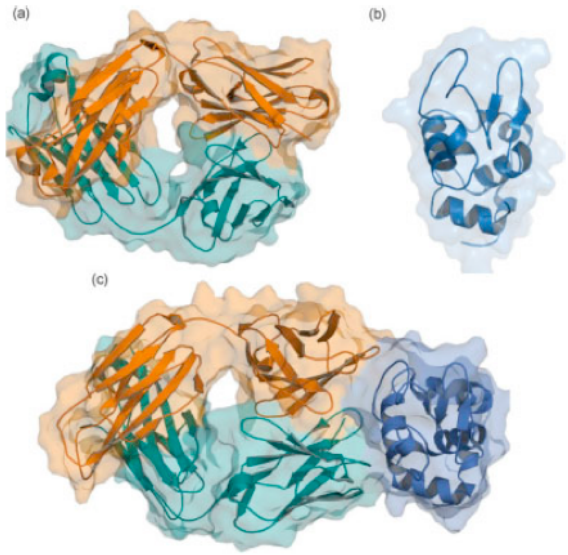


Figure 4. Schematic representation of the protein-protein docking technique
X-ray structure of (a) FAB Hyhel63 antibody (PDBID: 1DQQ), (b) HEW lysozyme (PDBID: 3LZT) and (c) the biological complex formed between the two (PDBID: 1DQJ).

The docking in general, and the protein-protein docking in particular, is not a simple problem. The objective of it is to predict the three-dimensional arrangement of a protein-protein complex from the coordinates of its component molecules, hopefully pointing out most of the residue-residue contacts involved in the interaction.[41-47]

There are no general rules to predict a binding interface. Basically, all docking approaches assume that the native complex is near the global minimum of the energy landscape. In fact, based on thermodynamic hypothesis, at fixed temperature and pressure the Gibbs free energy of the macromolecule-solvent system reaches its global minimum at the native state of the complex.[48] It has been established over the last two decades that the energy landscape of a foldable protein resembles a many-dimensional funnel with a free energy gradient toward the native structure (Figure 5).[21,49,50] A number of studies suggest that the landscape theory also applies to protein-protein association.[51-54] This theory states that the assembly of two proteins is initiated by the formation of nonspecific encounter complexes,[24] followed by rearrangements of them driven by stronger and more specific interactions. Taking into account that it is the structural features that determine if two proteins interact,[55] then such hypothesis implies that not only the 'final' binding but also other parts of the surface contain information for interacting with the partner. The size of the funnel will be determined by the length scales of the long-range electrostatic and hydrophobic interactions and the geometry of the proteins, and hence the funnel is restricted to a neighborhood of the native complex.[56] There is a free energy gradient toward the native state, but the funnel is rough, giving rise to many local minima.[21,57]

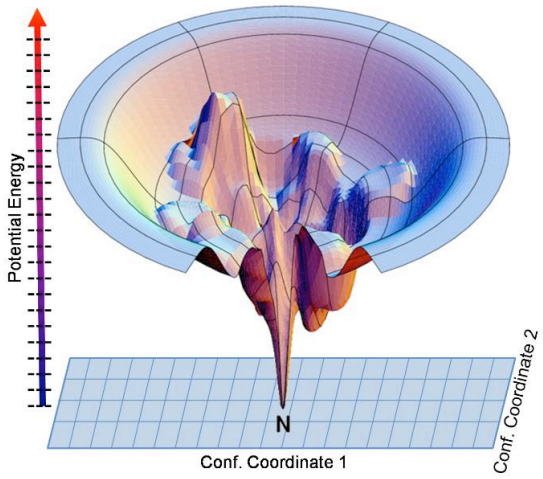


Figure 5. Protein-protein complex energy landscape
The many-dimensional funnel representing the energy landscape of a protein-protein complex. With "N" the native conformation is indicated.

Therefore, all the current docking methods are based on the optimization of a function approximating the free energy of the complex.
In all the docking algorithms, there are two crucial steps to generate possible models of the three-dimensional arrangement of a complex:

1. *Searching (low-resolution search)*, consisting in the generation of thousands of alternative poses (decoys) to sample the rotational/translational space;

2. *scoring and ranking (high-resolution refinement)*, consisting in scoring these poses using a 'pseudo-energy' function in order to rank the poses and so to identify the native-like solutions.

A simple docking algorithm may fail predicting the native complex. Anyway, a recent work[58] shows that the docking technique is able to distinguish between binding and non-binding partners, based on their score distributions. This may indicate that although protein surface morphology is not enough to find the native interface, it at least contains sufficient information to identify a 'bona fide' interactor.[58]
Anyway, it has been shown in CAPRI that, whereas approximately correct solutions are generated by the first step of the docking, scoring functions unfortunately often fail to correctly rank them.[58,59]

Step 1: sampling the conformational space
The searching step involves an exhaustive search of the conformational space of one protein with respect to the other, resulting in a six-dimensional search (6D). The search of through the entire conformational space of the complex geometry makes the calculation expensive, so it is necessary to simplify the system preserving the geometrical and physicochemical properties of the atoms, using mathematical models, such as geometrical shape descriptors or a grid.[42] Once having the easier representation of the system, almost all the docking programs use the same approach for the searching step: one protein is fixed in space (usually the bigger one) and the second one is rotated and translated around the first one. To minimize the degrees of freedom, both molecules are treated as rigid bodies, but still a simple systematic search is usually impracticable because the searching algorithm entails evaluating in the order of billions (10^9) distinct possibilities.[60]

Although geometric complementarity of the protein surface is the filtering criterion most commonly used to eliminate a large number of solutions with poor surface matching,[47] the docking problem is not simply matching two irregular shapes, but there are also other geometric, electrostatic or hydrophobic factors to take into account.[61]

So, there are a lots of possible search methods that have been used in protein-protein docking programs. Most methods that perform well in CAPRI are based only on three approaches. Some programs use grid-based spatial searches that are sped up with a Fast Fourier Transform (FFT), a method first applied in 1992 by Katchalski-Katzir and co-workers.[62] The other approaches for docking searches include instead Monte Carlo based searching[63,64] and geometric hashing.[65]

Step2: scoring and ranking docking decoys

The initial stage, which treats proteins as rigid bodies and generates many prediction (10.000 or more), is followed by the refinement stage, which performs any combination of detailed scoring, energy minimization, side chain optimization to the aim of valuate the energies of protein-protein docking poses in order to identify the one with the lowest energy as the predicted binding mode.[47]

A fundamental point of any docking method is to be computationally efficient, having a scoring scheme able to evaluate a huge number of solutions and discriminate the native-like binding modes from the wrong decoy complex structures in a reasonable computational time.[42]

The free energy of binding, $\Delta G_{binding}$, is not easily accessible but other and faster scoring functions that model $\Delta G_{binding}$ as accurately as possible, i.e. provide good correlations with experimental binding affinities, can be used.[60] Considering the energy function as a funnel-like function, as described above, the original free energy function is extremely rugged with huge number of local minima even in a small region of conformational space. Yet its approximated scoring function is much smoother and still capture the overall funnel-like landscape, which provides an easier free energy minimization (see Figure 6).[66] Further, according to the general idea of the funnel-shaped binding energy, there are an ensemble of encounter complexes from which the binding process initiates and precedes, that follow different pathways to converge in native state defined by the global minimum. So, there are many

possible routes for downhill in the binding funnel, and these are determined by transient interactions in the encounter complexes, which carry a track of the native interactions.[24]

Figure 6
The schematic representation of a funnel-like function (dark line) and an approximated scoring function (dotted line), still catching some of the local minima (indicated as small squares).

Whether this ensemble of orientations reflects the true binding-energy landscape will depend on the accuracy of the energy description and the efficiency of the sampling method. Most of the docking algorithms developed so far use the extent of geometric complementarity of the protein surfaces because it is a fast filter to eliminate a large number of solutions with poor surface matching. It is, however, usually recognized that a criterion based exclusively on geometric complementarity is far from being enough to distinguish among native and non-native docked geometries, except for a very a small number of cases.[67] Numerous criteria have been implemented with different levels of success: steric complementarity of the shapes of the interaction sites, electrostatic interactions, hydrogen bonding, van der Waals, pair potential, desolvation, rotamer probabilities, contact pair potential and knowledge-based potentials. Different docking programs can use different combinations of these terms in a weighted sum. Furthermore, exclusion of the solvent from the interface and the associated solvent entropy change play an important role in the stabilization of protein interactions, and can be estimated from empirical potentials or database derived functions.[18,68]

Finally, the scoring part is generally followed by a final post-processing stage, in which a large number of low energy conformations (usually 2000 to 20000) are

retained and ranked. A common way to rank the retained decoys is clustering them using pairwise root mean square deviation (RMSD, a number that quantifies the structural diversity between two structures) as the distance measure, and then ranks the clusters according to their size, i.e., identifying conformations that have large numbers of neighbors.[56,57] The method is based on the observation that, in the free energy landscapes of partially solvated receptor-ligand complexes, the free energy attractor at the binding site generally has the greatest breadth among all local minima.[69] Hence, following the uniform sampling of the conformational space defined by translations and rotations of the ligand, the docked conformations that are below an energy threshold are expected to form the largest cluster around the native complex.

Biological information

Although important progresses, protein-protein docking remains a quite difficult procedure, due to the complex nature of the problem it tries to solve. One of the most useful approach to improve the quality of the docking simulations is the use of biological information about the complex interface to confine the search of allowed configurations or filter out wrong solutions.[42,70] Biological information available from experiments or from computational methods on the regions or residues likely involved in the interaction are one of the key points for the improvement of a docking simulation. Almost all the docking programs have a section in which it is possible to exclude regions not involved in the interaction, or driving the docking towards the ones involved (for example, the software HADDOCK[32] dedicate a section to express the NMR data such as chemical shift perturbation and residual dipolar couplings in terms of ambiguous interactions restrains). If experimental data are not available for the protein-protein system that is simulated, it is also very helpful to carry out structural comparisons of the same protein family.[42,70] Fox example, the binding crevice centered on the catalytic triad of serine proteases (His, Asp, Ser)[71,72] (see Chapter 7), as well as the complementarity defining regions of immunoglobulins (CDRs), which are part of the biological surface involved in the interaction with protein interactors (see Chapter 5 and Chapter 6), are both well characterized; although in general, a protease-inihibitor interface is more static and consequently more easily predicted than an antibody-antigen interface.[73]

1.3 - The CAPRI experiment: what is the state of protein-protein docking

As described above, protein-protein docking procedure is a very helpful method to model biological complexes and to guide biochemical experiments. A general docking algorithm can be briefly described as an initial searching step yields a long list of candidate structures; the following step requires some forms of post-processing, which may include: i) scoring or re-scoring of the docked conformations using a more accurate energy function, or ii) refining the conformations followed by re-scoring.[74] These treatments usually improve the number of near-native conformations among the 10 to 100 lowest energy structures, but in most cases are unable to eliminate all false positives (steps showed in Figure 7).

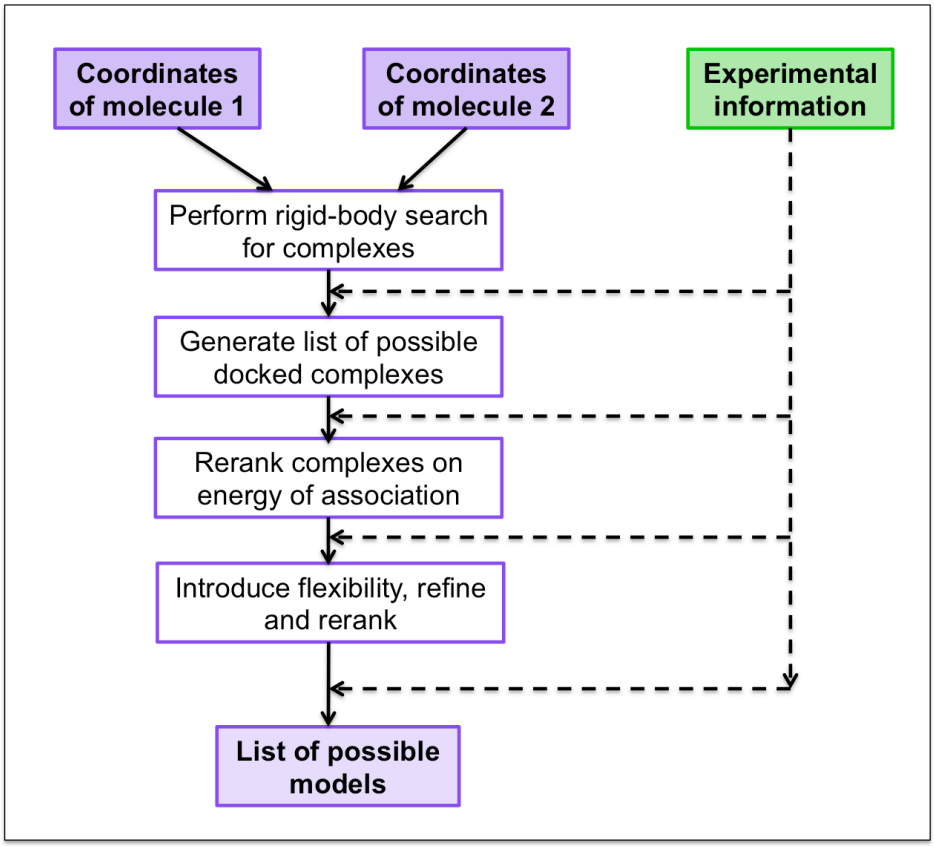


Figure 7
The stages of protein-protein docking.

A variety of approaches have been used in docking programs that mostly differ in the stages of the algorithms, showing different performances depending on the approach and the nature of the biological system. In this scenario, the comparison of different

docking programs to establish their relative performances is very important. Indeed, it is required an objective valuation of the model quality. To this aim, the international Critical Assessment of Prediction of Interactions (CAPRI) experiment was designed, precisely to evaluate current computational approaches of protein–protein docking.[75] The CAPRI is a community-wide experiment designed according to the model of the Critical Assessment of Techniques for Protein Structure Prediction (CASP).[76] It was designed in June 2001 at the Conference on Modeling Protein Interactions in Genomes organized in Charleston, SC, by Ilya Vakser (Medical University of South Carolina) and Sandor Vajda (Boston University). CAPRI targets are protein–protein complexes and it is data-driven, meaning that it can start whenever an experimentalist offers an adequate target and ends 6–8 weeks later with the submission of predicted structures.[76-78] Computational researchers are given the three-dimensional coordinates of the unbound structures for a given target before the experimental structure of the complex is published. The researchers are then given a few weeks to dock the two structures together, possibly using biological information and literature searches. Therefore, CAPRI challenge provides the docking community with a unique blind setting of simultaneously assessing of all docking algorithms, and has led to significant advances in the field.[79,80]

From the analysis of CAPRI results, it can be noted that there are some docking programs that give globally better predictions, such as ICM,[81] ZDOCK,[79] HADDOCK,[32] RosettaDock,[64,82] ClusPro[56] and Camacho group's Smooth-Dock.[75] Furthermore, in Figure 8, the number of citations per year of the most common docking programs joining to CAPRI is plotted (references took from ISI Web). From the plot it is possible to observe that only after 2003 there was an increase of the number of citations of the protein–protein docking software. Since their publication the most cited programs are HADDOCK,[32] RosettaDock,[64] three-dimensional-Dock,[83] BIGGER,[67] and Dot.[53] It is possible to observe an increase of the number of citations per year of the Patch-Dock,[44,45] ClusPro,[56] HADDOCK,[32] RosettaDock[64] and ZDOCK.[79] When considering only papers that apply the different software to specific biological problems (represented in Figure 8b) HADDOCK results to be the most popular one, followed by ClusPro, PatchDock and RosettaDock.

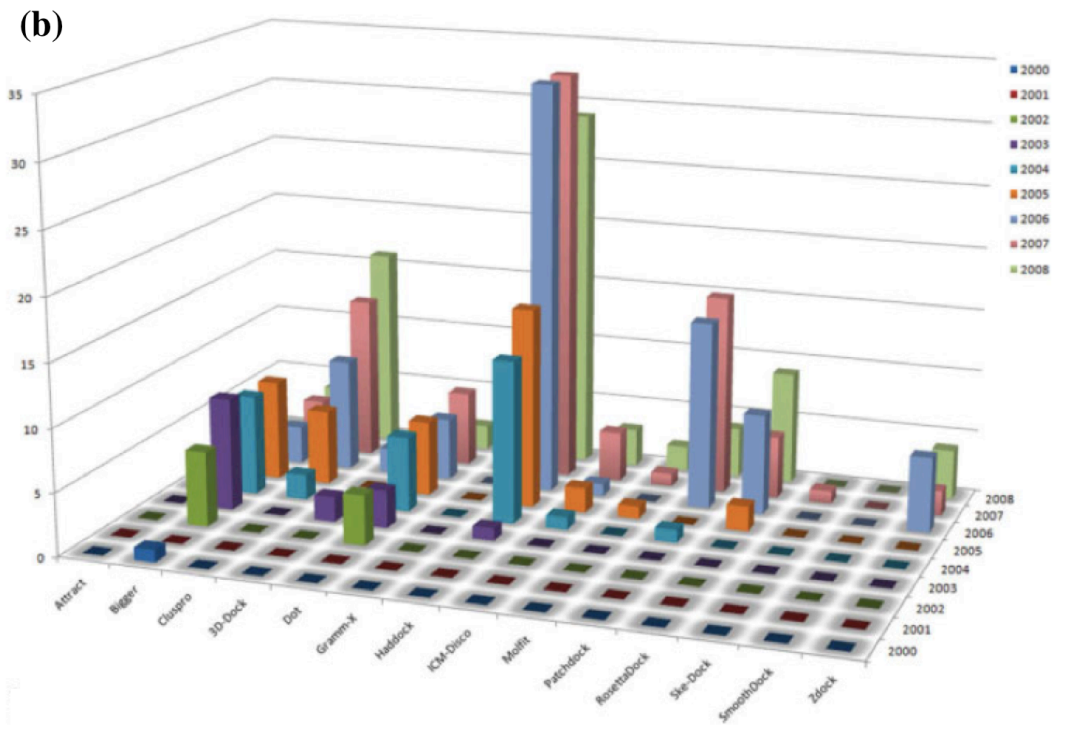


Figure 8
Number of citations per year of the docking programs described earlier. Data taken from ISI Web of Science (February of 2007).; only the articles with experimental predictions were considered.

Four of the most common docking programs are RosettaDock,[64] ZDOCK,[79] HADDOCK[32] and ClusPro.[56] The advantage of RosettaDock compared with the other three programs is the close correspondence of the lowest free energy structures with the X-ray complex, the disadvantage is that using a Monte Carlo technique in the searching step and a detailed energy function, it is quite slower than the others. Instead, ZDOCK is a FFT based algorithm, so it is faster but it does not perform well in the cases of complexes with large conformational change. HADDOCK seems combine the rapidity with the fact that the both side chains and backbone are allowed to move, and this increase the accuracy of the scoring if compared with classical rigid body docking programs. The big disadvantage in HADDOCK is that it is data-driven, so its performance closly depends on the availability and the level of confidence of experimental information. Compared with the other programs, ClusPro has the advantage to be a fully automated algorithm that rapidly docks, filters and ranks potential models within a short amount of time, using only the structures of the component proteins, and eventually adding experimental data if available.
A more detailed description of these methods is reported in the Appendix 2.

1.4 - The PhD project

My PhD work has been focused on the study of protein-protein interactions, taking advantage of computational techniques. The study has been devoted to two main aspects: i) the development of new methods to analyse and rank docking solutions (Chapters 2,3,4), and ii) the application of these methods, in combination with classical state-of-art computational biology simulations, to predict and analyse the binding mode in real biological systems, which are related to particular diseases (Chapters 5,6,7). As availability of biological information is guarantee of a better success rate in the docking simulations, we afforded the latter part of the work in collaboration with experimental groups. In particular, interaction between proteins involved in the autoimmune response in celiac disease has been studied in collaboration with the group directed by Prof. Daniele Sblattero, University of Piemonte Orientale (Italy) and the group directed by Prof. Carla Esposito, University of Salerno (Italy). In addition, recognition properties of the FXa enzymatic system has been studied through dynamic characterization of a FXa pathogenic mutant that causes problem in the process of blood coagulation. This study has been performed in collaboration with the group directed by Prof. Raimondo De Cristofaro, Catholic University School of Medicine, Rome (Italy) and the group directed by Prof. Flora Peyvandi, Ospedale Maggiore Policlinico and Università degli Studi di Milano (Italy).

Finally, during my PhD I spent seven months in the groups of the Prof. Charlotte Deane, Department of Statistics, University of Oxford (UK). In that period, I studied the geometrical features of the proteins' regions most recurrent in the protein-protein interaction, the loops, clarifying some structural aspects of them in one of the most important and huge class of proteins: the membrane proteins (Appendix 1).

CHAPTER 2 - COCOMAPS: a web tool for analyzing, visualizing and comparing the interface in protein-protein and protein-nucleic acid complexes

2.1 - Introduction

Interaction between biomolecules is at the basis of many of the most important molecular processes in the cell. As described in Chapter 1, protein-protein interactions underlie for instance signaling, regulation, immunogenic recognition, whereas protein-nucleic acid interactions under- lie processes such as DNA transcription, repair, replication, as well as post-transcriptional events, including RNA splicing and editing.

Availability of a 3D structure for a complex allows detailed analysis of the interaction at atomic level between the molecular partners, which is a fundamental step for possible biomedical and biotechnological applications. Moreover, the recent development of well performing docking software (see Chapter 1 and Appendix 2) to predict the 3D structure of macromolecular complexes requires, in the analysis step, the accurate and tedious screening of all the best solutions. It is indeed well accepted that the correct solution, if any, can be found within the 10-20 best ranked ones (e.g. the CAPRI assessment accepts 10 different models per target from each predictor).

It is therefore of timely interest, both for bioinformaticians and wet biologists, to have programs and tools able to automatically analyse features of a complex interface, and to easily and intuitively discriminate between similar and different binding solutions

Several valuable web tools have been made available for the analysis of the interface in biomolecular complexes.[84-92] However, no available web tool has been implemented to provide interactive contact maps from the 3D structure of a biomolecular complex.

Introduced to provide a reduced representation of a protein structure, contact maps have been successfully exploited for describing similarity between protein structures. Analogously, an intermolecular contact map between two or more interacting molecules could identify uniquely and intuitively the surface of interaction, representing a sort of fingerprint of the complex and reporting the crucial information

in a ready-to-read form. Interesting work has in fact been done to demonstrate the advantages of using contact map representations for the alignment of protein-protein interfaces.[93,94]

For this reason, during my PhD study my groups and I have implemented COCOMAPS (bioCOmplex Contact MAPS).[4] It is a novel web tool to easily and effectively analyse and visualize the interface in biological complexes, such as protein-protein, protein-DNA and protein-RNA complexes, by making use of intermolecular contact maps.

2.2 - Methods

All the programs under COCOMAPS have been written in python, taking advantage of python libraries such as SciPy and Matplotlid. We made it available at the URL: http://www.molnac.unisa.it/BioTools/cocomaps.

2.3 - Results and Discussion

Description of the tool

The tool takes in input the PDB type file of the complex, that contains the Cartesian coordinates of the complex. Usually, the two interacting parts of the complex are distinguished by different names of the chains, indicate by a single letter. In fact, a user-friendly interface of the tool allows to download input files directly from the data bank wwPDB[95] (for the experimental structures) or to upload locally stored PDB formatted files. The user is requested to specify the chain identifiers for the molecules involved in the interaction to be analyzed. More chains can be selected for each interacting partner, which overcomes a limitation of the other available tools that either work on all the chains present in a PDB file, or on one pair of them at a time. Therefore, COCOMAPS can be used to analyze the interface between two molecules, between one molecule and an ensemble (made by two or more molecular chains) or between two ensemble, depending on how many chains are specified.

COCOMAPS outputs are displayed on the results HTML page for one month and archived as downloadable compressed files. A link to the online resource is also emailed to the user, if requested.

COCOMAPS provides three graphical contact maps defining the interface of the complex:

1. *Black and white* contact map;
2. *Distange range* contact map;
3. *Properties* contact map.

The first one is a classical intermolecular contact map (Figure 9, a) where a black dot is present at the crossover of residues i and j, belonging to molecule/assembly 1 and molecule/assembly 2, respectively, if any pair of atoms belonging to the two residues is closer than a cut-off distance chosen by the user (default value being 8 Å). The second map (Figure 9, b), named "*distance range contact map*", reports in different colors inter-residues contacts at increasing distances. Red, yellow, green and blue indicate contacts within 7 Å, 10 Å, 13 Å and 16 Å, respectively. The third contact map (Figure 9, c), named "*properties contact map*", is similar to the first one, but each contact is colored according to the physico-chemical nature of the two interacting residues: hydrophobic-hydrophobic in green, hydrophilic-hydrophilic in violet and hydrophobic-hydrophilic in yellow.
By mousing over the maps, it is possible to visualize the identity of the residues pairs corresponding to the dots.

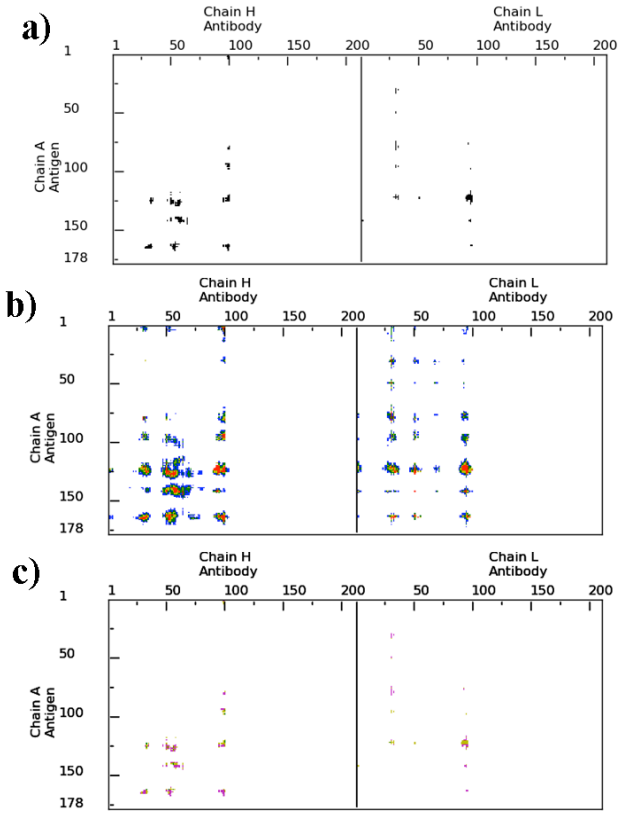


Figure 9.
A sample of COCOMAPS contact maps for the complex Ibalizumab antibody (chains L and H) with the CD4 antigen (chain A), PDBcode: 3O2D. a) *Black and white* contact map; b*) Distance range* contact map; c) *Properties* contact map.

Our tool also provided detailed information, organized in table (Figure 10, **a** and **b**), about:

1) interacting residues, defined on the basis of a cut-off distance that can be customized by the user;
2) residues at the interface, defined on the basis of the buried surface upon complex formation;
3) intermolecular H-bonds, with specification of the acceptor and donor atoms.

A 3D visualization of the complex in JMol (http://www.jmol.org) (Figure 10, **c**) is also provided online, with the interacting residues highlighted. Finally, a ready-to-run Pymol[96] script, which generates a visualization of the interface in the corresponding 3D-structure, is downloadable. Accessible surfaces and H-bonds are calculated by NACCESS[97] and HBPLUS,[98] respectively.

All the programs under the COCOMAPS web tool have been written in python, taking advantage of python libraries such as SciPy and Matplotlib.

a)

Molecule1				Molecule2				Distance	aaProperty	
Res1	N° res1	Atom1	Chain1	Res2	N° res2	Atom2	Chain2	Dist(Å)	property1	property2
LYS	1	O	A	TYR	99	CD1	H	7.32	phil	phil
LYS	2	C	A	TYR	99	CD1	H	7.53	phil	phil
VAL	3	CG2	A	THR	100	CG2	H	6.87	phob	phil
VAL	3	CG2	A	ASN	98	OD1	H	6.26	phob	phil
VAL	3	CG1	A	TYR	99	CD2	H	3.76	phob	phil
VAL	4	N	A	TYR	99	CD2	H	6.95	phob	phil
LEU	5	CD2	A	TYR	99	CD2	H	5.16	phob	phil
ASN	30	ND2	A	THR	30	O	L	4.92	phil	phil
SER	31	OG	A	TYR	32	OH	L	7.96	phil	phil
SER	31	OG	A	ASN	30	CB	L	3.61	phil	phil
ASN	32	CB	A	THR	30	O	L	4.01	phil	phil

b)

Title	Value
Buried area upon the complex formation (Å²)	2271.3
Buried area upon the complex formation (%)	7.63
Interface area (Å²)	1135.65
Interface area MOL1 (%)	11.33
Interface area MOL2 (%)	5.75
POLAR Buried area upon the complex formation (Å²)	1228.9
POLAR Interface (%)	54.11
POLAR Interface area (Å²)	614.45
NON POLAR Buried area upon the complex formation (Å²)	1042.4
NON POLAR Interface (%)	45.89
NON POLAR Interface area (Å²)	521.2
Residues at the interface_TOT (n)	64
Residues at the interface_Mol1	32
Residues at the interface_Mol2	32

c)

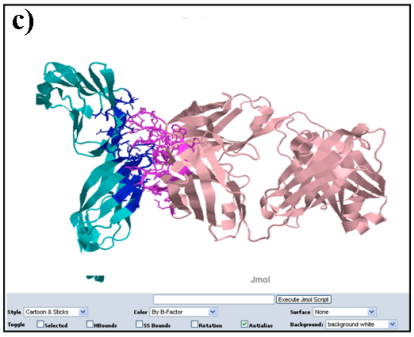


Figure 10.
Sample COCOMAPS outputs for the complex Ibalizumab antibody (chains L and H) with the CD4 antigen (chain A), PDBcode: 3O2D. a) First part of the table of *interacting residues*, defined on the basis of the cut-off distance; b) Overview table of the interaction properties ; c) 3D visualization of the complex in Jmol

The example

Although COCOMAPS provides a complete characterization of the interfaces in biological complexes, the real novelty that we have introduced is the generation of intermolecular contact maps. Contact maps give an immediate view of which regions of the two partners are in contact. From the properties map, it is also possible to immediately appreciate the physico-chemical nature of the interaction.

As an example, in Figure 11 properties contact maps are reported for the biological complexes of the antigen *hen egg lysozyme* (HEL) with two different antibodies, namely *D1.3* (PDBcode: 1VFB)[99] and *F10.6.6* (PDBcode: 1P2C)[100], together with the corresponding Pymol 3D representation of the complexes, as generated by COCOMAPS.

The 2D contact-maps of the HEL-antibody complexes reported in Figure 11 show in a glance that the two binding solutions are completely alternative, and the

corresponding epitopes present no overlap. In addition, contact maps specify which regions of the antibodies and of the antigen are in contact.

As expected, both the antibodies contact HEL with their six hypervariable loops (L1, L2, L3, H1, H2 and H3, also labeled in the figure, for the sake of clarity). As for the HEL antigen, it contacts the *D1.3* antibody with about 30 N- and 30 C-terminal residues and the *F10.6.6* antibody with its central region (residues 40-85). The same information could of course be extracted either from lists of interacting residues or from the 3D view of the complexes (such as that in Figure 11). However, differently from the contact-map view, which is immediate, in both of the above cases, manual intervention by the user would be required to extract the needed information. Further, the contact maps in Figure 11 immediately indicate that the H3 loop of the *D1.3* antibody is more involved in the interaction with HEL than the *F10.6.6* H3 loop, and that it mostly gives hydrophilic-hydrophilic contacts (magenta dots). This is a consequence of the *D1.3* H3 loop amino-acids sequence, (one code amino-acids sequence: ERDYRLDY), which is longer than the *F10.6.6* one (one code amino-acids sequence: GDGFYVY), and much more hydrophilic, presenting five charged residues.

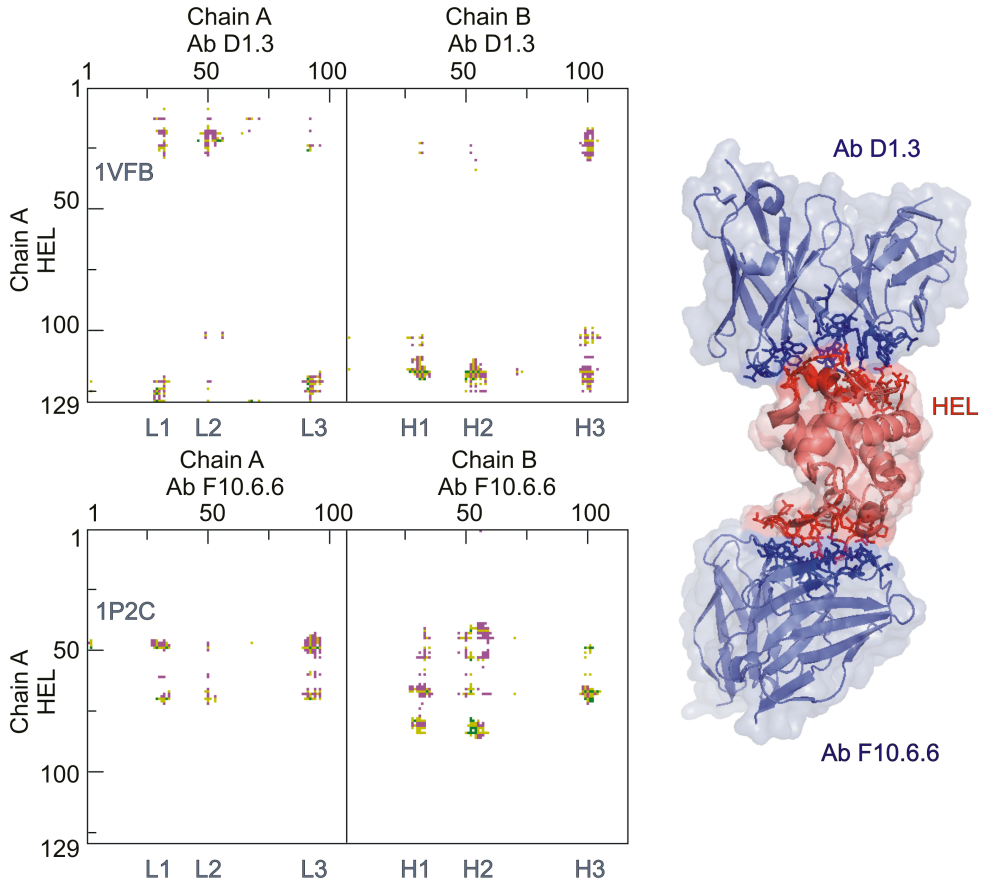


Figure 11.
Comparison of the complexes of HEL with two different antibodies: *D1.3* (PDBcode: 1VFB) and *F10.6.6* (PDBcode: 1P2C). *Left*: COCOMAPS "properties contact maps". Labels have been added for the antibody hypervariable loops L1-L3 and H1-H3. Magenta, green and yellow dots indicate hydrophilic-hydrophilic, hydrophobic-hydrophobic and hydrophobic-hydrophilic contacts, respectively. The cut-off distance is set to 10 Å. *Right*: A Pymol visualization of the complexes based on the automatic COCOMAPS script .pml; residues at the interface are shown as "sticks".

2.4 - Conclusion

In conclusion, this first study has been focused on the development of a tools able to automatically analyze, visualize and compare the interfaces both in experimental and predicted 3D structures of protein-protein and protein-nucleic acids complexes. COCOMAPS combines in a single tool the traditional analysis and 3D visualization of interfaces in biocomplexes with the effectiveness of the contact map view. It can straightforwardly be applied to the analysis of interfaces both in experimental and predicted 3D structures of biological complexes.

CHAPTER 3 - CONS-COCOMAPS: a novel web tool to measure and visualize the conservation of inter-residue contact in multiple docking solutions

3.1 - Introduction

As described in Chapter 1, most important molecular processes in the cell rely on the interaction between biomolecules. Understanding the molecular basis of the recognition in a functional biological complex is thus a fundamental step for possible biomedical and biotechnological applications. However, the 3D structure of a significant fraction of biomolecular complexes is difficult to solve experimentally. In this scenario, the development of accurate protein-protein docking programs is making this kind of simulations an effective tool to predict the 3D structure and the surface of interaction between the molecular partners in macromolecular complexes.[101] Unfortunately, correctly scoring the obtained solutions to extract native-like ones is still an open problem [95,102], which is recently also object of assessment in CAPRI (Critical Assessment of PRedicted Interactions), a community-wide blind docking experiment [59]. As a consequence, the confidence to have a near-native solution among the ten best ranked ones is still an unreached task [102]. This requires the accurate and tedious screening of many docking models in the analysis step.

Typically, as described in Chapter 1 and Appendix 2, the first step of a docking simulation generates a large number, around 10^5-10^6, of 3D models (decoys). Such decoys are then clusterized on the basis of RMSD values, usually calculated on the atoms of the smaller molecular partner (or "ligand") [56,64,103]. The different solutions are ranked according to the cluster population: the most populated the cluster, the higher the rank. However, RMSD has two major limitations: i) its statistical significance is length dependent and ii) it is a global metric, that may not be able to characterize local similarities. As a consequence, solutions belonging to different RMSD-based clusters may share a notable number of intermolecular contacts, pointing essentially to the same interface. Therefore, as already reported [50,102,104,105], RMSD cannot be the only descriptor for the similarity of multiple docking solutions.

Indeed, in the CAPRI experiment the correctness of a prediction, i.e. its similarity to the native structure, is assessed not only by means of RMSD based criteria, but also from the conservation of ligand-receptor contacts, as compared to the native structure [50]. Alternative scores have also been proposed to evaluate the correctness of a docking prediction, based on the geometric distance between the interfaces, and the residue-residue contact similarity [104].

However, the normal case in real-life research is having many different docking solutions to analyse and obviously no native structure to compare them to. Therefore, it would be of great utility both for bioinformaticians and wet biologists to have programs and tools to easily and effectively analyse and compare multiple docking solutions, based on criteria other than 'simple' RMSD. Most of all, it would be useful to visualize the consensus of multiple docking solutions, in order to appreciate at a glance which is the conservation rate of the predicted interface and which are the residues most often predicted as interacting.

As a matter of fact, if different docking solutions, especially from a series of well recognized programs, point to the same interacting regions, it is likely that the prediction can be better trusted. Consequently, it will be reasonable to focus attention, as for instance in site-directed mutagenesis experiments, on the residues most frequently predicted to be involved in the interaction. The concept of "consensus" has indeed been widely demonstrated to improve the performance of bioinformatics tools in many fields, including the prediction of protein and RNA secondary structure [106-112], of membrane protein topology [113], of protein retention in bacterial membrane [114], of docking small ligands to proteins [115,116], etc. Recently, consensus interface prediction has also been used to improve the performance of macromolecular docking simulations [117-119].

However, although many valuable tools have been made available to analyse the interface in biomolecular complexes [4,84-88,90-92], no tool has been developed to the aim of measuring and visualizing the consensus of multiple docking solutions. In Chapter 2 there is the description of COCOMAPS (bioCOmplexes COntact MAPS, available at the URL [90]), a comprehensive tool that my group and I developed to analyse and visualize the interface in biological complexes, by making use of intermolecular contact maps [4]. We have shown that intermolecular contact maps can be very effective in providing an immediate 2D-view of the interaction, allowing to easily

discriminate between similar and different binding solutions. They represent a sort of fingerprint of the complex, providing the crucial information in a ready-to-read form. Then, we used intermolecular contact maps to develop the second novel tool, CONS-COCOMAPS (CONSensus-COCOMAPS), to measure and visualize the conservation of inter-residue contacts in multiple docking solutions. CONS-COCOMAPS provides both numerical values of the contacts conservation and a graphical representation in the form of a "consensus map". To show its performance, here we applied CONS-COCOMAPS to the analysis and visualization of a few test cases taken from recent CAPRI rounds.

3.2 - Methods

Given an ensemble of N models of the same biomolecular complex, the pairwise contacts conservation score, C_{pair}^{ij}, between models i and j is calculated as in Eq. 1.

$$C_{pair}^{ij} = \frac{nc_{ij}}{(nc_i + nc_j)/2} \quad (1);$$

where nc_i and nc_j are the total number of inter-residue contacts in models i and j, respectively, and nc_{ij} is the total number of inter-residue contacts common to models i and j. Following this definition, the average pairwise contacts conservation score C_{pair}^{av} simply is the value of C_{pair}^{ij} averaged over all the possible pairs of models in the considered ensemble, see Eq. 2.

$$C_{pair}^{av} = \sum_{i,j>i}^{N} \frac{C_{pair}^{ij}}{N(N-1)/2} \quad (2).$$

However, Eq 1. can be generalized to a conservation score defined over all the N models in the considered ensemble, as in Eq.3.

$$C_{100} = \frac{nc_{100}}{\sum_{i}^{N} \frac{nc_i}{N}} \quad (3);$$

where nc_{100} is the total number of inter-residue contacts common to all (100%) the models in the ensemble. The contacts conservation score of Eq. 3 can be extended to measure any amount of inter-residue contacts common to a given percentage of analysed models. For instance, C_{70} is calculated as in Eq. 4, where nc_{70} is the total number of inter-residue contacts conserved in 70 % of the analysed models.

$$C_{70} = \frac{nc_{70}}{\sum_{i}^{N} \frac{nc_i}{N}} \quad (4).$$

The total number of inter-residue contacts in an ensemble of N models, Nt, is calculated as in Eq. 5.

$$Nt = \sum_{i}^{N} nc_i \,. \quad (5)$$

Finally, on a residue level we define the conservation rate, CR_{kl}, of Eq. 6, where nc_{kl} is the total number of models where residues k and l are in contact.

$$CR_{kl} = \frac{nc_{kl}}{N} \,. \quad (6)$$

Within this work, two residues are defined in contact if any pair of atoms belonging to the two residues is closer than a cut-off distance of 5 Å, which is the threshold distance adopted in the assessment of CAPRI predictions to define native residue-residue contacts [50]. Conservation rates can be plotted in the form of consensus contact maps, which are depicted in a grey scale. The highest conservation corresponds to a

black dot, absence of conservation corresponds to white, and contacts at increasing conservation appear in darker grey.

All the programs under CONS-COCOMAPS have been written in python, taking advantage of python libraries such as SciPy and Matplotlib. It is freely available as a web tool at the URL [92]).

3.3. CAPRI models

The docking models for recent CAPRI targets were downloaded from the official web site (at the URL [88]). We selected seven recent protein-protein targets (T24-T26, T28-T29, T32, T36) for which the docking models were made available to the public. Four of them, T25, T26, T29 and T32, have at least one medium quality prediction and are more extensively discussed in the text. A total of 2130 CAPRI models have been analysed, 300 for target T24, round 9, 300 for target 25, round 9, 310 for target 26, round 10, 320 for target 28, round 12, 350 for target 29, round 13, 350 for target 32, round 15, and 200 for target 36, round 15 (see Table 1). Note that targets T24 and T25 refer to the same native complex. The quality score (Q-score) for each Predictor was calculated by summing 0, 1, 2 and 3 for each incorrect, acceptable, medium quality and high quality solution, respectively, as assessed in CAPRI [59]. Predictors which submitted less than the ten allowed models and those who submitted models with a ligand and/or receptor sequence not corresponding to the target were excluded from the analysis. L_rmsd is the pair-wise RMSD calculated on all the heavy atoms of the ligand after a LSQ RMS fit of the receptor invariant residues backbone, as in the CAPRI assessment [50].

Target	CAPRI Round	Incorrect	Acceptable	Medium quality	High quality	All
T24	R 09	296	4	0	0	300
T25	R 09	268	19	12	1	300
T26	R 10	276	19	15	0	310
T28	R 12	320	0	0	0	320
T29	R 13	333	8	9	0	350
T32	R 15	316	6	13	15	350
T36	R 15	199	1	0	0	200

Table 1. Analysed models

3.4 - Results and Discussion

Given a number of multiple docking solutions, we calculated the conservation score of the inter-residue contacts at different percentages, from 0 to 100%. For instance, C_{70} gives the amount of inter-residue contacts which are conserved in 70% of the compared models. When only two models are compared, the pair-wise conservation score, C^{ij}_{pair}, is calculated. CONS-COCOMAPS then plots the inter-residue contacts conservation to an intermolecular contact map, that we call "consensus map".

The conservation of inter-residue contacts has been here measured and visualized with CONS-COCOMAPS for a total of 2130 models submitted to CAPRI for seven different targets: T24, T25, T26, T28, T29, T32 and T36 (see Table 1). The percentage of correct solutions among those submitted is 10-11% for T25, T26 and T32 and 5% for T29. For the remaining targets, T24, T28 and T36, it is instead much lower: 1% and 0% and 0.5%, respectively (see Table 1).

Inter-residue conservation versus L_rmsd

The pair-wise conservation score, C^{ij}_{pair}, between all the models within each of the CAPRI targets T25, T26, T29 and T32 have been plotted versus the corresponding L_rmsd values in Figure 12. As expected, C^{ij}_{pair} rapidly decreases as the L_rmsd increases, with C^{ij}_{pair} approaching to zero at L_rmsd higher than 30-40 Å. The C^{ij}_{pair} distribution is significantly spread out, even at C^{ij}_{pair} values around 0.5 (which means that one out of two contacts at the interface is conserved in the two considered models), and several outliers are indeed observed that contemporarily show either low C^{ij}_{pair} and low L_rmsd values or high C^{ij}_{pair} and high L_rmsd values. As an example, the 3D representation of the models M03 and M07 submitted by the P86 predictor for T26, responsible for the point outlined by the arrows, is shown in the same Figure. The L_rmsd for their superimposition is as high as 19.6 Å, notwithstanding a pair-wise conservation score C^{ij}_{pair} of 0.47 is calculated. This is due to a significant conformational change undergone by both the receptor and the ligand in the two models (RMSD for the best superposition of the two receptors and

the two ligands is 4.8 Å and 2.8 Å, respectively), which causes a remarkably different orientation of the ligand. Nevertheless, regions involved in the interaction are substantially the same, because the ligand somehow "follows" the receptor in its conformational change. This case and many others demonstrate once more that the RMSD cannot be selected as the only descriptors for the similarity of two docking solutions and that descriptors directly describing the property of interest, in this case the interface, should be used [50,102,104,105].

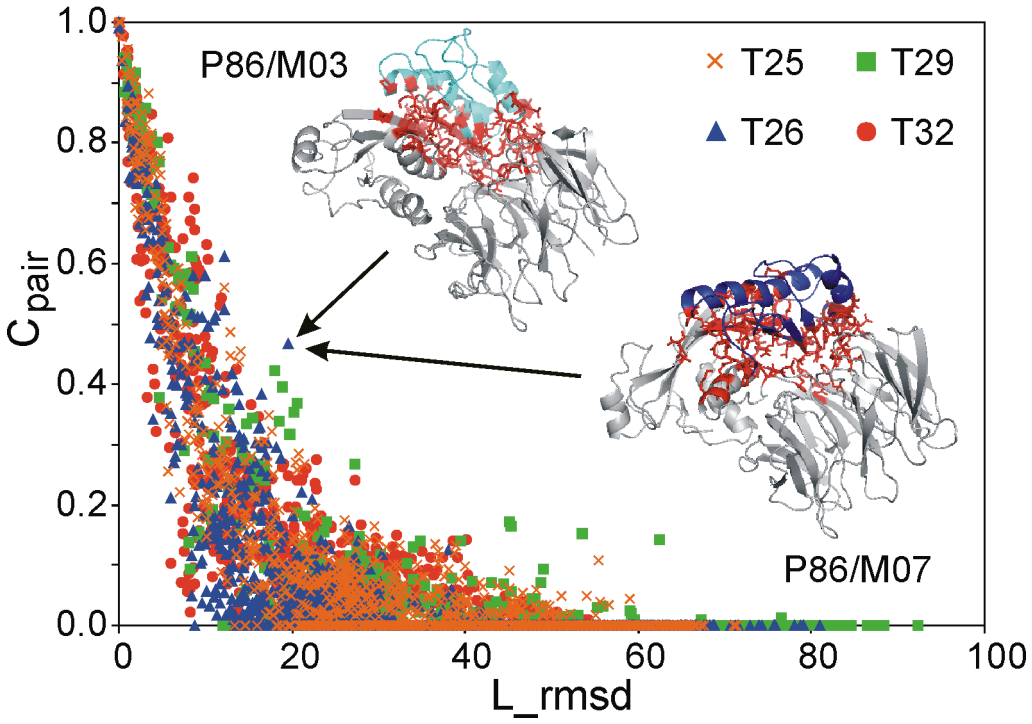


Figure 12. C^{ij}_{pair} **versus L_rmsd**

Chart of the C^{ij}_{pair} values versus L_rmsd values for targets T25, T26, T29 and T32. A comparison of the M03 and M07 models submitted by the P86 predictor for T26 and corresponding to the point indicated by the arrows is also shown with the ligand coloured in cyan and blue, respectively; residues involved in the contacts common to the two models are shown as red sticks.

Conservation and Consensus maps for the multiple solutions submitted by each predictor

Conservation scores have also been calculated for each set of ten models submitted for each CAPRI target by the same predictor. C_{30}, C_{50} and C_{70} (data not showed). They correspond to amount of inter-residue contacts which are conserved in 30%,

50% and 70% of the models, respectively. The average C_{pair}^{av} and the quality score, Q-score, for each predictor, obtained on the basis of the CAPRI assessment, are also reported.

As expected, the inter-residue conservation rate within each set of multiple solutions submitted by each predictor is very variable. As an illustrative example, in Figure 13a-b, the graphical CONS-COCOMAPS outputs (consensus maps) are shown for the set of ten predictions submitted by predictors P04 and P49 for target T32. For comparison, the intermolecular contact map for the native structure (PDB code 3BX1[120]) is also reported (Figure 13c). The calculated C_{pair}^{av} values are 0.003 and 0.400 for predictors P04 and P49, respectively. Visual inspection of Figure 13a-b immediately indicates that the solutions proposed by predictor P49 are very conservative as concerns the predicted inter-residue contacts, whereas the predicted inter-residue contacts in the solutions proposed by predictor P04 are extremely diverse and spread out all over the map. Further, the maps of Figure 13b-c also immediately show that the consensus contact map of predictor P49 is extremely similar to the contact map of the native complex structure. In fact, predictor P49 performed very well in this test case, having one acceptable, two medium quality and five high quality predictions. On the contrary, predictor P04 had only incorrect predictions.

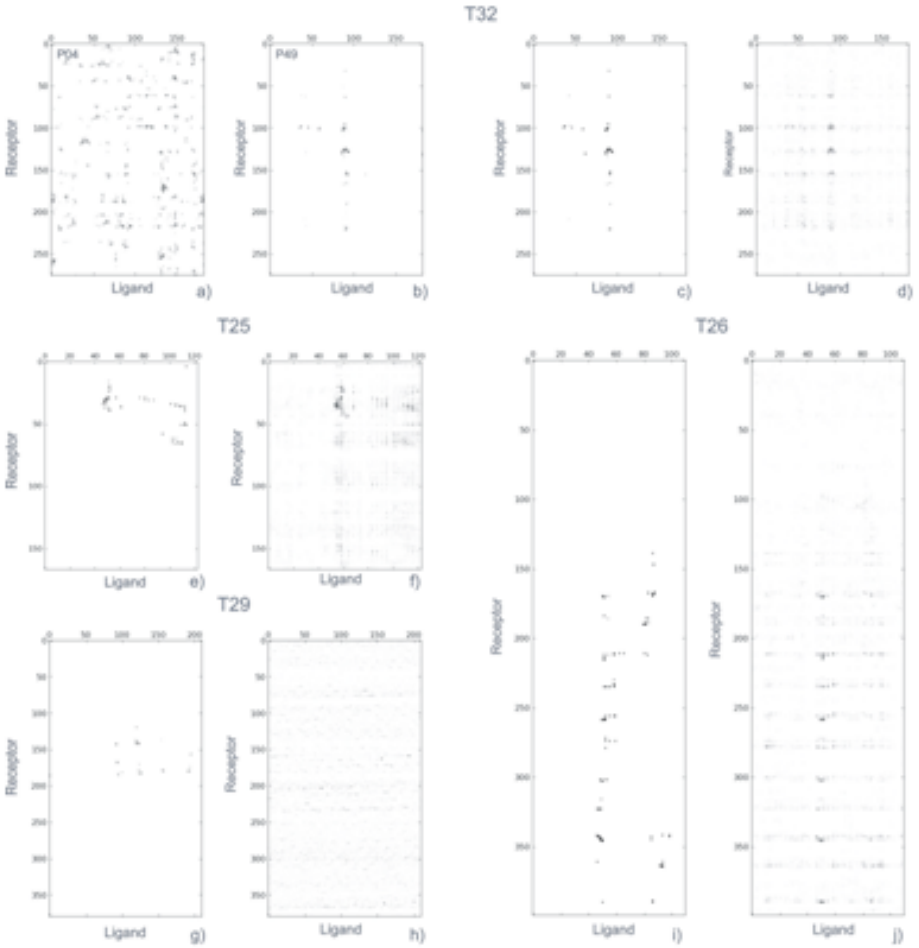


Figure 13. Consensus maps
a-b) CONS-COCOMAPS consensus maps obtained from the 10 models submitted for the CAPRI target T32 by the P04 and P49 predictors. **c-j)** Comparison between the CONS-COCOMAPS consensus maps (**d,f,h,j**) obtained from all the 300, 310, 350 and 350 models submitted to CAPRI for the targets T25, T26, T29 and T32, respectively, and the intermolecular contact maps (**c,e,g,i**) of the corresponding native structures (PDB codes: 2J59, 2HQS, 2VDU and 3BX1).

We noted that there is indeed a nice correlation, especially for targets T26 and T32, between the success of the predictor and a high conservation of the inter-residue contacts. However, it is worth to remark that the opposite does not hold true, i.e. we also observed cases where a predictor submitted very similar predictions in terms of inter-residue contacts but they were far away from the native structure. For instance, the ten predictions submitted by predictor P89 for target T25 share an average C_{pair}^{av} as high as 0.772, notwithstanding all the predictions have been assessed as incorrect.

The corresponding consensus map is shown and compared with the native structure contact map in the Figure 14.

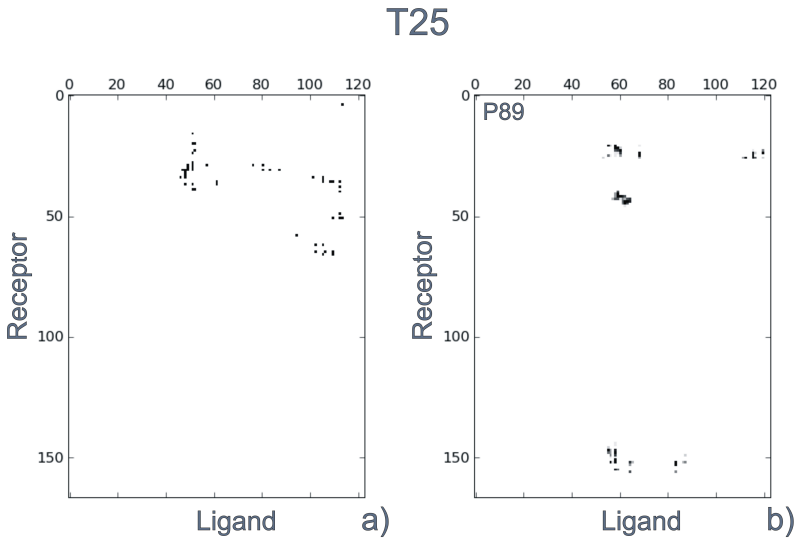


Figure 14. Consensus map from the P89 predictor for T25.
Comparison between the CONS-COCOMAPS consensus map (b) obtained from the 10 models submitted for the CAPRI target T25 by the P89 predictor, and the intermolecular contact map (a) of the corresponding native structure (PDB code: 2J59).

Consensus maps for the multiple solutions submitted by all the predictors

Overall conservation scores of the inter-residue contacts in all the models submitted for the analysed targets are quite low. Conservation scores at 5, 10, 15 and 20 % are reported in Table 2, both for all the docking models and for only the incorrect solutions. They correspond to the number of inter-residue contacts which are conserved in 5, 10, 15 and 20 models out of 100, divided by the average number of contacts per model. From Table 2 it is apparent that the conservation of inter-residue contacts in T24, T28, T29 and T36 is particularly low. The conservation score of contacts common to the 5% of all the models, including the correct ones, is indeed below 0.7 (0.398, 0.056, 0.176 and 0.643, respectively). At higher percentages the conservation scores for these targets are zero, with the only exception of T36, whose C10 value is 0.016.

On the contrary, C5 assumes higher and similar values for the other three targets, from 2.274 for target T32 to 2.455 for target T25. These values are remarkably lower when the correct predictions are excluded from the analysis. C10 values are also quite similar and range from the 0.420 for target T32 to 0.576 for target T26. C15 values

are more variable, ranging from 0.078 for target T25 to 0.183 for target T26. Exclusion of the correct predictions causes a dramatic decrease of the C15 values, which approach to zero. At percentages of 20% or more, the conservation score is not higher than 0.027 for any of the analysed targets.

Target	**Nt**	**C5**	**C10**	**C15**	**C20**
T24	15818	0.398	0.000	0.000	0.000
T24-incorrect	15618	0.322	0.000	0.000	0.000
T25	15399	2.455	0.448	0.078	0.000
T25-incorrect	13613	1.477	0.020	0.000	0.000
T26	22063	2.318	0.576	0.183	0.020
T26-incorrect	19825	2.019	0.125	0.014	0.000
T28	29360	0.056	0.000	0.000	0.000
T29	23890	0.176	0.000	0.000	0.000
T29-incorrect	22923	0.000	0.000	0.000	0.000
T32	25859	2.274	0.420	0.081	0.027
T32-incorrect	23420	1.754	0.202	0.027	0.000
T36	12750	0.643	0.016	0.000	0.000
T36-incorrect	12673	0.628	0.016	0.000	0.000

[a] Calculations performed upon excluding all the correct predictions.

Table 2.
Inter-residue conservation scores at different percentages for all the models submitted for each target.

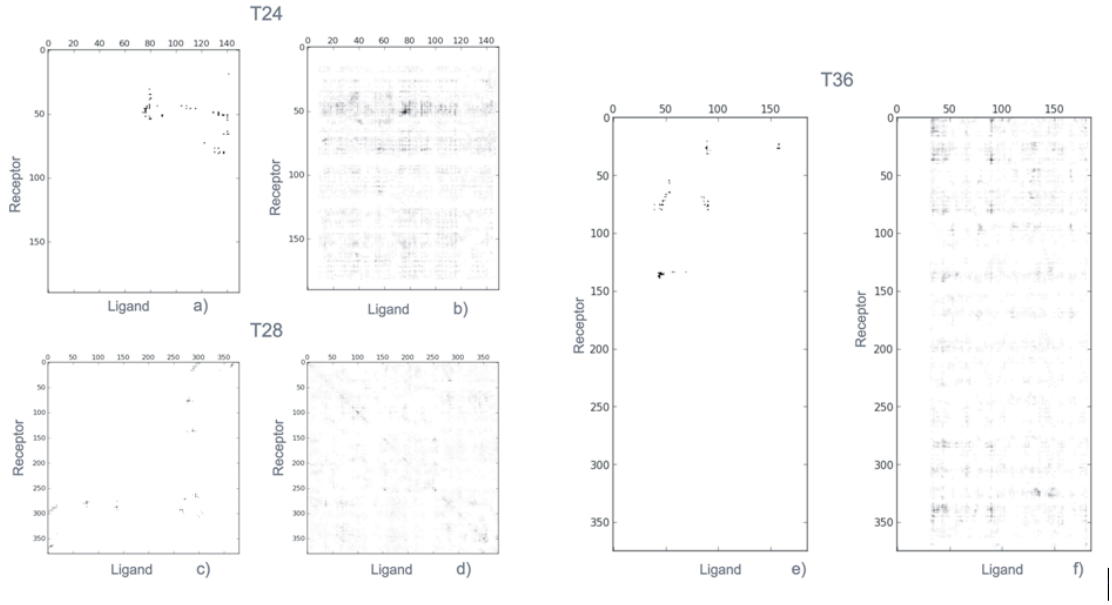


Figure 15.
Comparison between the CONS-COCOMAPS consensus maps (b,d,f) obtained from all the 300, 320 and 200 models submitted to CAPRI for the targets T24, T28 and T36, respectively, and the intermolecular contact maps (a,c,e) of the corresponding native structures (PDBcodes: 2J59, 2ONI and 2W5F).

Conservation rates at the residue level have been plotted in consensus maps and are reported in Figure 13 for T25, T26, T29 and T32 and in the Figure 15 for T24, T28 and T36, together with the intermolecular contact map of the corresponding native structures (PDB codes: 2J59 [121], 2HQS [122], 2ONI, 2VDU [123], 3BX1 [120] and 2W5F [124] for T24/T25, T26, T28, T29, T32 and T36, respectively). The consensus maps reported in Figure 13d-f-h-j and Figure 14b-d-f therefore represent the consensus emerging from the analysis of 200 to 350 different solutions, for each target, submitted by different predictors and obtained and selected on the basis of different methods and criteria.

As a consequence of their very low conservation scores, the consensus maps of T24, T28, T29 and T36 are quite spread out and only for T24 a week signal emerges from the background noise (Figure 13h and Figure 14b-d-f). On the contrary, in case of targets T25, T26 and T32, some darker hot spots, due to the best conserved inter-residue contacts in the multiple solutions, clearly emerge (Figure 13b-d-f).

	CR_{kl}	Receptor		Ligand		Distance (Å)
T25						
	0,173	TYR	35	TYR	999	3,48
	0,167	PHE	51	ASP	996	**5,82**
	0,163	PHE	51	ILE	1053	4,00
	0,150	ASN	52	ASP	996	3,84
	0,147	THR	44	TYR	999	2,60
	0,140	ASN	52	TYR	999	4,20
	0,140	ILE	46	ILE	997	3,65
	0,137	THR	45	TYR	999	3,49
	0,133	ILE	49	GLN	1035	**6,09**
	0,130	ILE	49	ILE	995	**5,29**
T26						
	0,232	GLU	293	GLU	116	3,62
	0,210	GLU	293	THR	114	2,66
	0,197	PHE	424	PRO	115	3,43
	0,190	ALA	249	GLU	116	2,92
	0,187	SER	205	GLU	116	2,66
	0,174	PHE	424	GLU	116	**5,55**
	0,174	HIS	246	GLU	116	2,79
	0,168	MET	204	GLU	116	3,75
	0,158	GLN	336	THR	114	2,94
	0,158	GLY	248	GLU	116	3,94
T29						
	0,069	TRP	236	PHE	165	**7,67**
	0,063	HIS	221	PHE	165	3,65
	0,063	VAL	195	ARG	195	**6,53**
	0,060	TRP	236	GLU	204	3,03
	0,057	PHE	231	PRO	236	3,88
	0,057	LYS	223	THR	200	**5,73**
	0,054	VAL	195	PHE	165	**7,28**
	0,051	PHE	231	LEU	237	3,35
	0,051	TRP	236	TYR	207	3,67

	0,051	VAL	233	THR	200	**6,82**
T32						
	0,223	LEU	126	TYR	87	3,71
	0,200	GLY	127	TYR	87	3,74
	0,183	SER	125	TYR	87	**7,68**
	0,169	GLY	100	TYR	87	4,03
	0,160	ASN	62	TYR	87	**9,91**
	0,157	SER	128	TYR	87	3,49
	0,146	ASN	62	THR	89	4,65
	0,143	ASN	155	THR	89	4,56
	0,140	LEU	96	TYR	87	3,52
	0,137	GLY	127	LEU	91	3,51

Table 3. Ten most conserved inter-residue contacts.
The ten most conserved inter-residue contacts are reported for targets T25, T26, T29 and T32, together with corresponding distances in the native structures [120-123]. Distances above 5 Å are outlined in bold.

Interestingly, analysis of the CONS-COCOMAPS outputs indicates that among the ten inter-residue contacts with highest conservation rates, reported in Table 3, several correspond to native inter-residue contacts. Indeed, for targets T25, T26 and T32, seven, nine and eight of the ten most conserved contacts correspond to distances within 5 Å in the native structure [120-123] (see again Table 3). Considering that only ~10% of the CAPRI models for the three targets was assessed to be correct (Table 1), this indicates that focusing on the consensus of predicted inter-residue contacts, rather than on the correctness of the entire models, can significantly increase the success rate of the prediction. Importantly, hot spots of the interactions are highlighted by this approach, such as for instance residue Tyr87 of the T32 ligand (the barley a-amylase/subtilisin inhibitor), whose mutation to alanine has been experimentally shown to dramatically decrease the ligand-receptor affinity [120]. A useful consensus, five correct contacts among the ten most conserved contacts, also emerges for T29, for which only 5% of the models was assessed to be correct (Table 3). Further, when drawing the consensus maps for targets T25, T26 and T32 using only the incorrect solutions, some inter-residue contacts corresponding to the native ones still emerge, and are clearly distinguishable from the noise (Figure 16). In particular, considering

only the incorrect models submitted for T25, T26 and T32, two, seven and four contacts, respectively, correspond to native ones (data not shown). Surprisingly, even T24, having no medium/high quality prediction, presents three native contacts among the ten most conserved ones (Table 4). Quite strikingly, these findings indicate that the consensus of many solutions, even incorrect according to the CAPRI definition, may point to the correct inter-residue contacts. If confirmed, this result could be of great interest and utility in applications such as mutagenesis experiments design, considering that the main aim of bioinformaticians and wet biologists, when performing macromolecular docking simulations, is often to predict the residues at the interface, more than the fine details of the biomolecular complex.

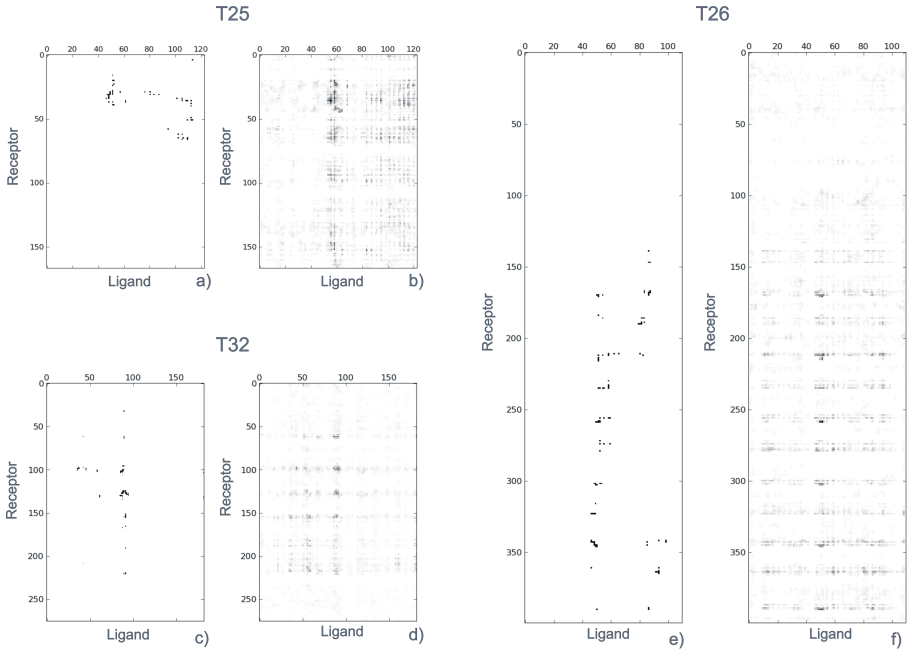


Figure 16.
Comparison between the CONS-COCOMAPS consensus maps (**b,d,f**) obtained from the 268, 276 and 316 incorrect models submitted to CAPRI for the targets T25, T26 and T32, respectively, and the intermolecular contact maps (**a,c,e**) of the corresponding native structures (PDBcodes: 2J59, 2HQS and 3BX1).

	CR_{kl}	Receptor		Ligand		Distance (Å)
T24						
	0,093	PHE	51	ASP	996	5,82
	0,083	PHE	51	ILE	997	8,11
	0,080	PHE	51	LEU	994	6,47
	0,073	PHE	51	ILE	995	9,16
	0,073	ILE	49	TYR	999	9,50
	0,070	ILE	49	ILE	997	8,59
	0,067	GLY	50	ASP	996	6,17
	0,060	ASN	52	ASP	996	3,84
	0,057	ASN	52	TYR	999	4,2
	0,057	ILE	49	ASP	996	4,92

Table 4
Ten most conserved inter-residue contacts for the target T24 and corresponding distances in the native structure [120-123].

3.5 - Conclusions

Here I described the second computational method I developed during my PhD work to easily measure and visualize the consensus in multiple docking solutions. Our novel tool CONS-COCOMAPS uses the conservation of inter-residue contacts as an estimate of the similarity between different docking solutions. The conservation of ligand-receptor contacts is indeed used as one of the fundamental criteria in CAPRI for assessing the similarity of a predicted complex to the native structure, and recently it has been emphasized that it can be the most useful descriptor when looking at the biological significance of the prediction, i.e. the individuation of the interface area [102]. To visualize the conservation, CONS-COCOMAPS uses intermolecular contact maps, that we recently showed to be a very effective way to visualize a biomolecular complex interface [4]. There is virtually no limit on the number of models that can be compared by CONS-COCOMAPS. This novel tool is freely available to the scientific community (at the URL [92]) and can straightforwardly be applied to the analysis of the outputs of one or more docking programs.

The application of CONS-COCOMAPS to some test-cases taken from recent CAPRI rounds shows that it is efficient in highlighting even a very weak consensus. Interestingly, in three out of the seven analysed cases, T25, T26 and T32, consensus maps clearly point to the native contacts (Figure 13 and Table 3). In other two cases, T24 and T29, although the consensus is less visually apparent from the maps (Figure 13 and Figure 15), three and five native contacts, respectively, are included among the ten most conserved inter-residue contacts (Table 3 and Table 4). Importantly, in none of the analysed cases a false-positive consensus emerged. This opens the road to further studies to test and prove whether the consensus of a large number of docking solutions may be used to successfully predict residue-residue contacts in biomolecular complexes.

CHAPTER 4 - CONS-RANK: a novel tool to rank multiple docking solutions based on the conservation of inter-residue contacts

4.1 - Introduction

Although most proteins fulfil their functions through interaction with other proteins, a dramatic disproportion still exists between the number of experimental structures solved for protein complexes and the number of structures available for single proteins.[125] In this scenario, molecular docking, i.e. the prediction of a protein complex structure starting from the two separate components, is the method of choice for investigating the molecular basis of the recognition in many functional biological systems (see Chapter 1 and Appendix 2 for details about the docking technique). In a docking process, a large number of possible conformations (docking decoys) are sampled, from which native-like solutions, i.e. solutions close to the native structure, should be extracted. Unfortunately, correctly scoring the obtained solutions to rank native-like conformations before the incorrect ones is still an open problem, which is also object of assessment in CAPRI (Critical Assessment of PRedicted Interactions), a community-wide blind docking experiment.[59] In the last CAPRI edition it was shown that, although signs of progress are evidenced, correctly ranking models to single out the best ones from a decoys ensemble remains a challenge.[126]

During my PhD work, my group and I developed CONS-RANK (CONSensus-RANKing), a novel method to rank multiple docking solutions. CONS-RANK deeply differs from other valuable algorithms developed to the aim,[95,127-141] as it uses neither knowledge-based nor physics-based energy functions. Instead, it relies on the conservation of inter-residue contacts in the analysed decoys ensemble.

The importance of inter-residue contacts when analysing docking decoys is well established. In the CAPRI experiment, for instance, the correctness of a prediction, i.e. its similarity to the native structure, is assessed based on a combination of RMSD criteria and of conservation of inter-residue contacts, as compared to the native structure.[50] Interestingly, the fraction of common inter-residue contacts among a set of

docking decoys has been recently shown by Bonvin and colleagues[142] to successfully apply to their clustering, and a similar concept, i.e. the atom contact frequency in a set of predictions, has also been recently added to the ZRANK docking pipeline.[143]

Here we introduce the use of the conservation of inter-residue contacts to the task of ranking multiple docking solutions. The basic idea behind our approach is to move away from ranking methods based on the analysis of the single model *per se*. Rather, we first decompose the whole ensemble of decoys into an inter-residue contacts matrix (that can be visualized as a contact map, see below). Contacts that occur more frequently can be seen as "hot spots" for the interaction, and the decoys in the ensemble are ranked according to their ability to match the more frequently observed contacts. We had this idea when analysing several CAPRI targets to extract the most conserved inter-residue contacts for visualization in a "consensus map" (i.e. an intermolecular contact map where absence of conservation corresponds to white and contacts at increasing conservation appear in darker grey).[5] Quite strikingly, we observed that even if a small fraction of native-like solutions was present in the decoys ensemble, a clear native-like consensus in terms of inter-residue contacts emerged from the background noise and, more importantly, a significant fraction of native contacts was included within the ten contacts with highest conservation rate. This finding clearly indicates that also incorrect solutions may point to some correct inter-residue contacts and is in line with results of the analysis that Lensink and Wodak performed on 20 CAPRI targets to the aim of evaluating the ability of docking calculations in predicting the interface in protein-protein complexes.[144] Lensink and Wodak interestingly showed that about one quarter of the interfaces in incorrect docking models are in fact correctly predicted and that 70% of all the submissions with correct interface predictions are contributed by incorrect models. On the other hand, analogously to regular or irregular arrays of atoms when scattering X-ray beams, correct contacts, which can also be present in incorrect solutions, add constructively towards the native consensus, whereas incorrect contacts are expected to be wrong in a different way (unless the underlying docking algorithm is biased towards a specific wrong interface) and thus to give destructive interference, cancelling one another and not contributing to a false consensus.

With these considerations on mind, we developed a simple and fast algorithm to rank docking decoys according to their ability to match the most conserved inter-residue contacts in the analysed decoys ensemble. In the following, we illustrate the algorithm

and demonstrate its performance on over 100 targets from three benchmarks: RosettaDock,[64] DOCKGROUND[145] and CAPRI.[146,147]

4.2 - Methods

The algorithm we have implemented is split into two sections. In the first, we analyse the decoys ensemble to find the most conserved inter-residue contacts. In the second, we rank the decoys in the ensemble according to their ability to match the most conserved contacts.

Given an ensemble of N models of the same biomolecular complex, to find the most conserved contacts, we define the conservation rate, CR_{kl}, of each inter-residue contact as in Eq. 1,

$$CR_{kl} = nc_{kl}/N, \tag{1}$$

where nc_{kl} is the total number of models where residues *k* and *l* are in contact.

To rank the models in the ensemble according to their ability to match the most conserved inter-residue contacts according to their conservation rate, we first calculate a score per each model *i*, as in Eq. 2:

$$S_i = \sum_1^{M_i} CR_{kl}, \tag{2}$$

where M_i is the total number of contacts in model *i*. Then, we calculate a normalized score, $\bar{S}_i$, as in Eq. 3:

$$\bar{S}_i = S_i/M_i. \tag{3}$$

Note that the normalized score $\bar{S}_i$ of Eq. 3 coincides with the average conservation of the inter-residue contacts in each model. Models are ranked according to their $\bar{S}_i$

value. Within this work, two residues are defined in contact if any pair of atoms belonging to the two residues is closer than a cut-off distance of 5 Å. Conservation rates were plotted in the form of consensus contact maps as in the CONS-COCOMAPS program.[5] Contact maps for the corresponding native structures were obtained by COCOMAPS.[4]

All the programs under CONS-RANK have been written in python, taking advantage of python libraries such as SciPy and Matplotlib. The program is freely available upon request from the authors. ROC curves were obtained by plotting the fraction of true positives (FTP) against the fraction of false positives (FFP).

RosettaDock benchmark

A total of 6270 decoys for the 35 targets of the Global-Unbound RosettaDock benchmark having at least one native-like solution[64] (available at http://graylab.jhu.edu/docking/decoys/) have been downloaded and analysed. Models having a ligand RMSD (Lrmsd) <= 5 Å were classified as high/medium quality (HM). All models having a ligand RMSD (Lrmsd) <= 10 Å, i.e. high/medium quality plus acceptable ones, were classified as native-like (NL). On average, each target presented 179 decoys, including 6 high/medium quality models and 16 native-like models.

DOCKGROUND benchmark

A total of 6605 decoys for the 61 targets of the DOCKGROUND benchmark[145] (available at http://dockground.bioinformatics.ku.edu/) have been downloaded and analysed. For the decoys classification into high/medium quality and native-like, see the above section. Each target presented on average 108 decoys, including about 8 high/medium quality models and 10 native-like models.

CAPRI models

The docking models for recent CAPRI targets[146,147] were downloaded from the official web site (available at: ftp://ftp.ebi.ac.uk/pub/databases/msd/capri/). We analysed all the 6 recent protein-protein targets having at least one acceptable quality prediction (T24, T25, T26, T29, T32 and T36, round 9 on), for which the docking models were made available to the public. A total of 1810 CAPRI models have been analysed, 300 for target 24, round 9, 300 for target 25, round 9, 310 for target 26,

round 10, 350 for target 29, round 13, 350 for target 32, round 15, and 200 for target 36, round 15. Models were classified as incorrect, acceptable, medium quality or high quality, according to the CAPRI assessment.[50,59]

4.3 - Results and Discussion

We tested CONS-RANK on three different benchmarks: RosettaDock (global-unbound), DOCKGROUND and CAPRI. We remind that decoys in the RosettaDock benchmark were obtained by Rosetta global docking searches,[64] those in DOCKGROUND were generated by the GRAMM-X docking procedure,[145] whereas the CAPRI decoys were submitted by different predictors using different programs and procedures. A total of 14685 models, corresponding to 102 targets, were downloaded (6270 from RosettaDock, 6605 from DOCKGROUND and 1810 from CAPRI) and analysed.

Given an ensemble of multiple docking solutions for a specific target, CONS-RANK first calculates the conservation rate, CR_{kl}, of each observed inter-residue contact in the ensemble (see Methods). Then, it calculates the average inter-residue contact conservation rate, or normalized score, $\overline{S_i}$, for each model. Models are ranked according to their $\overline{S_i}$ values: the higher the $\overline{S_i}$, the better the rank. A consensus map[5] is also obtained for the ensemble of decoys provided in each benchmark for a given target, to possibly compare with the intermolecular contact map of the corresponding native structure.

After ranking the models, the number of high/medium quality and native-like (acceptable or better) solutions ranked within the top five, ten and twenty positions was counted. To further investigate the performance of the method, we also calculated the Receiver Operating Characteristics (ROCs) curve for each target by plotting the fraction of true positives *vs.* the fraction of false positives. The Area Under the ROC Curve (AUC), compared to the 0.5 value for a random function, was used to assess the overall performance of the method.

Ranking of decoys in the Global-Unbound RosettaDock benchmark

A total of 6270 models from the RosettaDock global-unbound benchmark were analysed, corresponding to 35 targets, which included 20 enzyme-inhibitor, 10 antibody-antigen and 5 other complexes. Results of the ranking of RosettaDock decoys are summarized in Table 5.

CONS-RANK proved to be effective in correctly ranking the docking solutions in the benchmark. It was indeed able to rank 16.8% of all the native-like solutions among the top ten positions, and 31.6% of them among the top twenty positions (Table 5), which is a striking difference from the random fraction of native-like solutions in the top ten positions, 1.6%. Furthermore, CONS-RANK proved to be able to specifically single out the high/medium quality solutions. In fact, as it is apparent from Table 5 (columns 5-10), the percentage of high/medium quality solutions ranked among the top five, ten and twenty positions is consistently larger for high/medium quality solutions than for the total native-like ones (13.5 vs. 8.2 %, 26.5 vs. 16.8 % and 46.2 vs. 31.6 %, in the top five, ten and twenty positions, respectively). Remarkably, almost half of the high/medium quality solutions are ranked within the top twenty positions.

For 17 out of the 35 analysed targets (shadowed in the table), the performance of our ranking method is excellent (AUC values above 0.9). Except for the 1PPE target, having almost all correct solutions (150 out of the total 179 ones), these targets, including examples of enzyme-inhibitor, antibody-antigen and other complexes, presented a total of native-like solutions ranging from 9 to 78, corresponding to 5.3% and 41% of the total solutions, respectively. In two cases, targets 1ATN and 1UGH, all the native-like solutions were correctly ranked before any incorrect solution, dealing to an AUC value of 1. In these two cases the correct solutions were 13 and 65, corresponding to 7.0 and 36 % of the total decoys, respectively.

Two factors may concur in explaining such a good performance of our ranking method when applied to the above targets: i) the incorrect solutions are instead near-native, at least in terms of inter-molecular contacts (i.e. even the incorrect solutions point to some native contacts); ii) the provided solutions are really unbiased, i.e. all the wrong solutions are wrong in a different way, thus not pointing to a false consensus and making the native consensus to easily emerge.

We have examples where either one or the other of the two factors clearly prevails. As it can be seen from Figure 17, the consensus map obtained using only the 162

incorrect solutions for the 1ATN target is completely spread, not highlighting any consensus. Therefore, the few (13) native-like solutions are sufficient to highlight a consensus towards the native contacts. On the contrary, in the case of the 2SIC target (AUC value 0.965), a consensus roughly corresponding to the native contacts is also observed in the consensus map drawn by using the 162 incorrect solutions, which means that a significant fraction of native contacts are also found in the solutions classified as incorrect (Figure 17f).

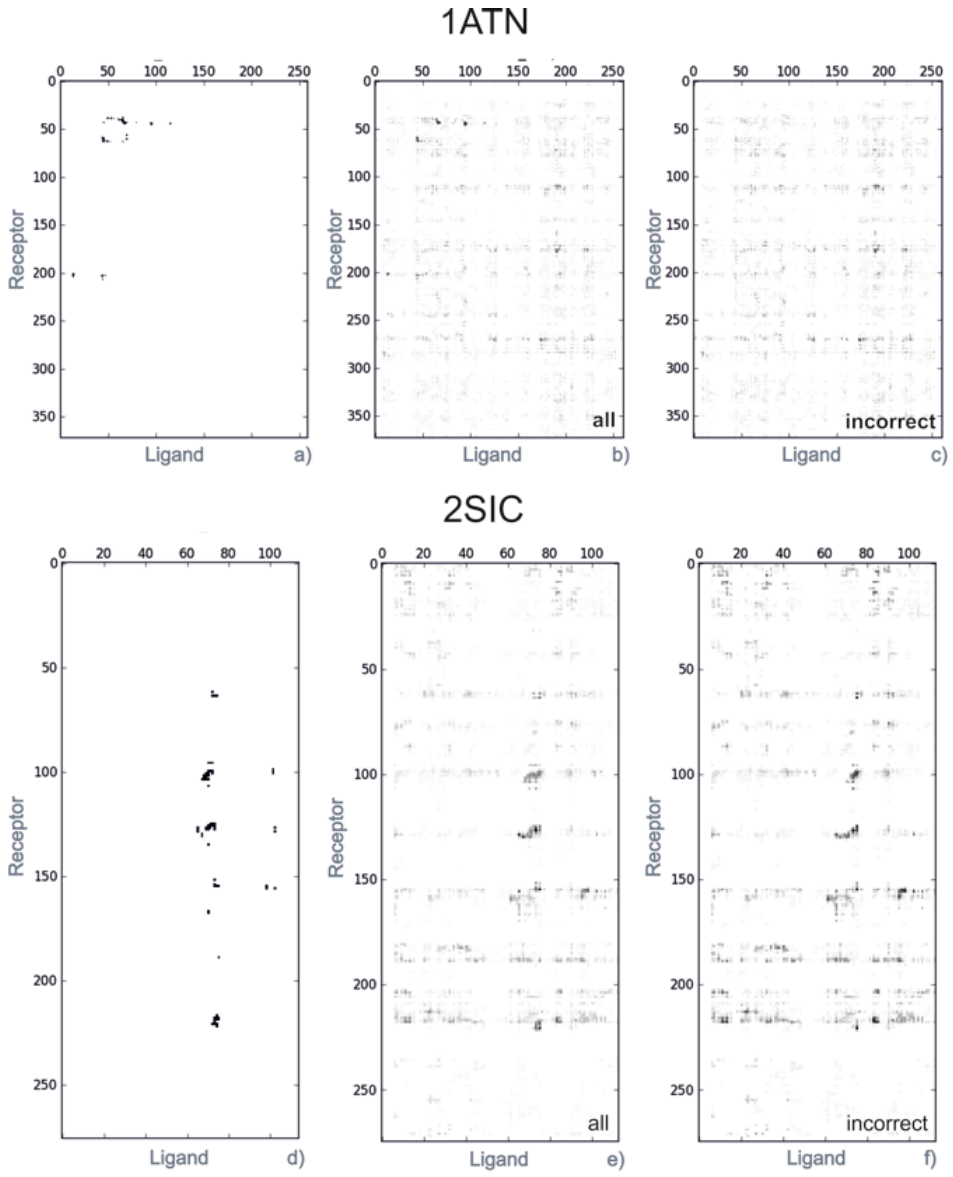


Figure 17. Consensus maps for the RosettaDock 1ATN and 2SIC targets
Comparison between the COCOMAPS[4] intermolecular contact maps of the 1ATN and 2SIC native structures (a,d), the consensus maps obtained from all the 185 and 179 models in the RosettaDock Global-Unbound benchmark (b,e) and the consensus maps obtained by only the 172 and 162, respectively, incorrect models (c,f).

The average AUC value over the 35 RosettaDock targets is 0.758, with 22 of them having an AUC value higher than 0.8. The average AUC value is significantly improved (to 0.799), when excluding from the analysis the 1AVZ and 1DQJ targets, having only two native-like solutions. For these targets our method ranked the two correct solutions at positions 76^{th} - 107^{th} and 88^{th} - 169^{th}, respectively, resulting in the particularly low AUC values of 0.174 and 0.015. A bad AUC value (0.114) was also obtained for target 1FBI, having only 3 native-like solutions.

The method performed badly in only four additional cases. In particular, AUC values worse that random (i.e. below 0.5) were obtained for the 1ACB and 2PTC targets, whereas AUC values around 0.5 were obtained for the 1CSE and 1MLC targets. Of these, 1ACB and 1MLC/2PTC present 5 and 7 native-like solutions, corresponding to 2.8 and 3.7/3.6 % of the total solutions. In Figure 18, a comparison between the consensus map for the 2PTC target (7 native-like solutions, AUC value of 0.202) and the native contact map is reported, from which it is apparent that the available solutions are biased because they point to a consensus that does not correspond to the native contacts. In particular, wrong regions of the ligand are docked to the receptor binding site. See for instance the four dark spots at the crossover of residues 30 to 45 of the ligand, with the receptor residues around the positions 60, 150, 190 and 220, which are absent in the native structure contact map. However, we note that for other targets having a comparable or even lower fraction of native-like solutions, for instance 1QFU, 1FSS or 1WQ1, significantly better results are obtained. For instance the AUC value for the 1QFU target, having only 3 native-like solutions, is as high as 0.886.

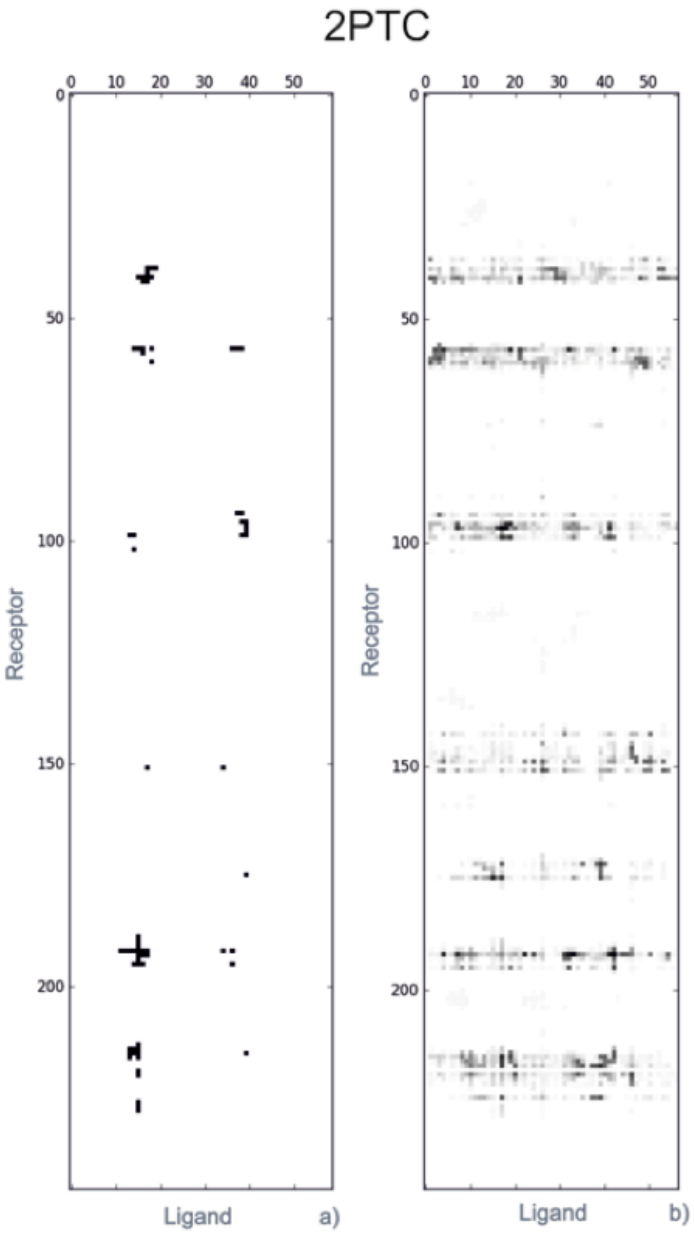


Figure 18. Consensus map for the RosettaDock 2PTC target.
Comparison between the COCOMAPS[4] intermolecular contact map of the 2PTC native structure (a) and the consensus map obtained from all the 192 models in the RosettaDock Global-Unbound benchmark (b).

Finally, the 1CSE target deserves a special mention. As said above, it presents the disappointingly low AUC value of 0.492, although having 19 native-like solutions (10% of the total). However, beside the 19 native-like models, with a Lrmsd < 10 Å, there are other 28 models with a Lrmsd < 12 Å. As a matter of fact, many of the

solutions classified as incorrect are instead "near native", both in terms of Lrmsd and of inter-molecular contacts. Therefore, it is not surprising that CONS-RANK ranks several of these "near native" solutions in the top positions, thus decreasing the AUC value. To clarify the concept, in Figure 19, the center of mass of all the ligand heavy atoms in contact with the receptor is shown both for the native structure[148] (gold) and for the native-like (hot pink) and non native-like (light pink) docking solutions. From the figure it clearly appears that, apart from two dozens of outliers, the ligand interface of most of the 171 'incorrect' solutions is indeed correctly centered on the receptor binding site.

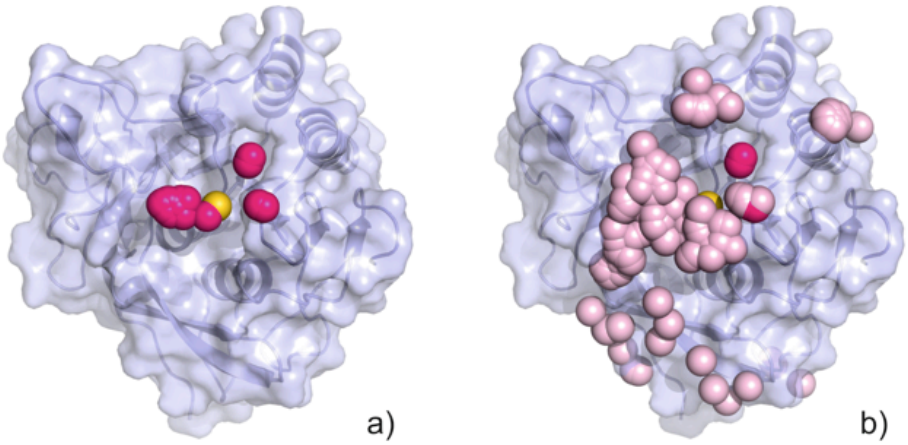


Figure 19. 3D representation of the native structure and docking decoys for the RosettaDock 1CSE target.
Receptor of the 1CSE target is shown in a light blue ribbon and surface representation. The center of mass of the ligand heavy atoms in contact with the receptor is shown as a gold sphere for the native structure (a,b), as a hot-pink sphere for the 19 native-like solutions (a,b) and a light-pink sphere for the 171 non native-like solutions (b). The figure was prepared with Pymol (www.pymol.org).

Ranking of decoys in the DOCKGROUND benchmark

A total of 6605 decoys from the DOCKGROUND benchmark were analysed, corresponding to a total of 61 targets. Results of the ranking of DOCKGROUND decoys are summarized in Table 6

The overall performance of the method is quite good, since it is able to rank 9.2% of all the native-like solutions within the top five positions, and 17.8% and 32.9% of them, respectively, within the top ten and twenty positions (Table 6). As in this

benchmark the 505 high/medium quality solutions represent almost the totality of all the 589 native-like ones, comparable results were obtained when considering only the high/medium quality solutions (Table 6 , columns 5-10)

The method performs in an excellent way, AUC values above 0.9, on 11 targets (shadowed in the table), and very well for additional 10 targets, AUC values between 0.8 and 0.9. Except the 1T6G target, having more than half native-like solutions (57 out of the total 110 ones), these targets, presented a number of native-like solutions ranging from 10 to 18, corresponding to 9% to 16% of all the solutions.

Also for this benchmark, in three cases, targets 1FM9, 1GPW and 1UGH, all the native-like solutions were correctly ranked before any incorrect solution, dealing to an AUC value of 1. In these three cases the native-like solutions were 13, 18 and 12, corresponding to 12, 16 and 11% of the total decoys, respectively. Analogously to results obtained on the RosettaDock benchmark, in some of the above cases, like for target 1GPW, the native consensus disappears from the consensus map when native-like solutions are excluded from the analysis, whereas in other cases, like target 1PPF, spots corresponding to native contacts are also observed in the consensus map obtained using only the incorrect solutions (see Figure 20).

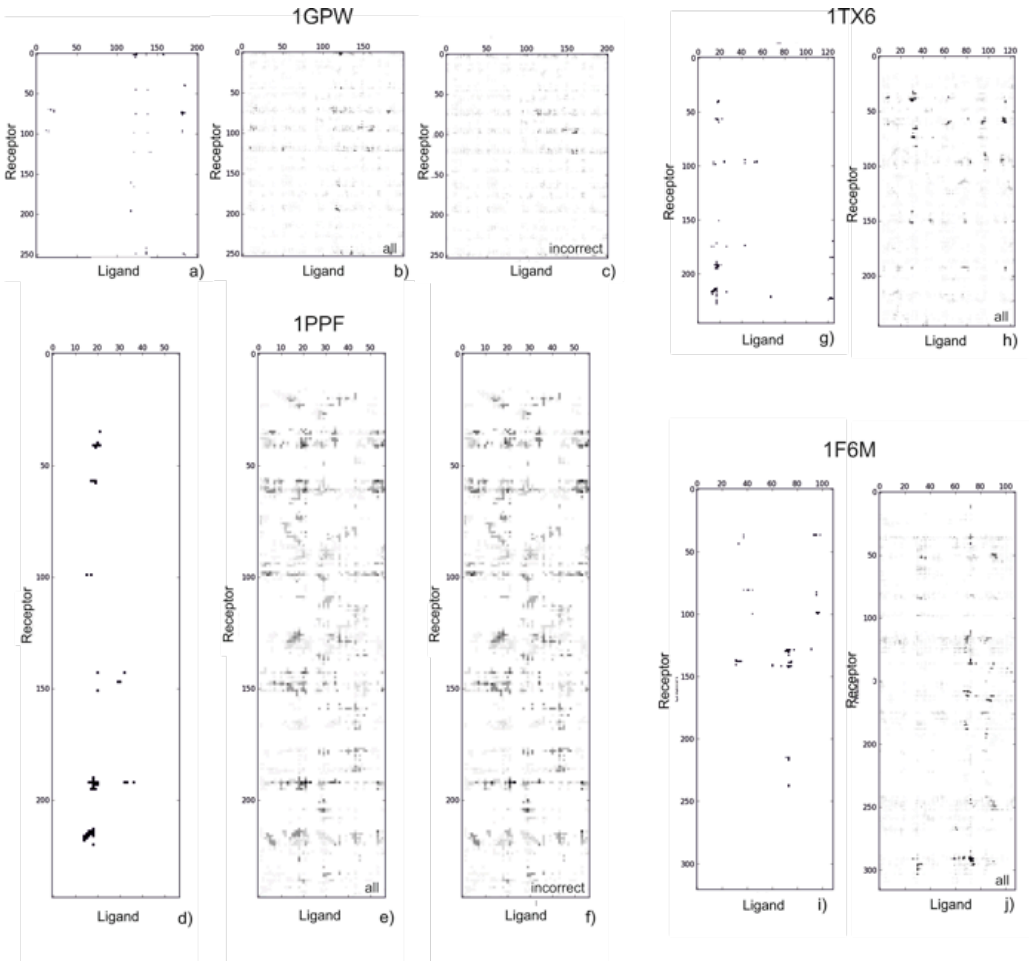


Figure 20. Consensus maps for the DOCKGROUND 1GPW,1PPF, 1TX6 and 1F6M targets.
Comparison between the intermolecular contact maps of the 1GPW, 1PPF, 1TX6 and 1F6M native structures (a,d,g,i) and the consensus maps obtained from all the 110 models in the DOCKGROUND benchmark (b,e,h,j). Consensus maps obtained by only the 92 and 100, respectively, incorrect models are also shown for 1GPW and 1PPF (c,f).

The average AUC value over the 61 targets is 0.654 and rises to 0.743 when excluding from the analysis the nine targets having only one or two native-like solutions. Analogously to results on the RosettaDock benchmark, the method performs badly on targets having only 1-2 native-like solutions (maximum AUC value 0.295). AUC values around 0.5 (ranging from 0.390 to 0.552) were also obtained for the 1OOK, 1P7Q and 1S6V targets, having only 4 native-like solutions (3.8% of the total solutions). Bad AUC values were obtained in four additional cases, in particular for the 1EZU, 1F6M, 1G6V and 1TX6 targets, having 8/10 native-like solutions out of 108/110. In all these cases, the decoys in the benchmark are biased

toward a wrong solution. This can be easily seen for the 1TX6 and 1F6M targets from Figure 20, where corresponding consensus and native contact maps are reported.

Ranking of CAPRI targets

A total of 1688 models for 6 CAPRI targets were ranked by CONS-RANK. Results are summarized in Table 7

Performance of CONS-RANK on the CAPRI targets is strikingly good. It ranks 32.0% of all the native-like solutions in the top ten positions, and 59.0% of them in the top twenty positions (Table 7). Like for the RosettaDock benchmark, the method specifically singles out the high/medium quality solutions. Analysis of the data in Table 7 (columns 5-10) indicates that the percentage of high/medium quality solutions ranked among the top five, ten and twenty positions is consistently larger for high/medium quality solutions than for the total native-like ones (20.0 vs. 18.0 %, 43.1 vs. 32.0 % and 72.3 vs. 59.0 %, in the top five, ten and twenty positions, respectively). Therefore, about three quarters of all the high/medium quality solutions are ranked within the top twenty positions.

The average AUC value is 0.870, and only for target T36, having one native-like solution out of 199 (0.5%), the performance of the method is not better than random (AUC value of 0.490). Instead AUC values approximate to 1 for the T25, T26, T29 and T32 targets, having a number of native-like solutions ranging from 17 to 34 (from 5% to 12% of the total solutions). For targets T25, T26 and T32, it is pretty clear that also incorrect solutions point to native contacts, as can be easily seen from the corresponding consensus maps (see Figure 21). In the case of target T29, the map is pretty spread and it is not easy to visually distinguish the native contacts from the background noise. However, we have previously shown that five out of the ten best conserved inter-residue contacts are native, i.e. correspond to distances within 5 Å in the native structure (while the remaining five are within a maximum distance of 7.7 Å).[5]

Target T24, having only four native-like solutions (1.3 % of the total) also has an AUC value as high as 0.818 (the four native-like solutions are ranked at positions 40, 57, 62 and 66 out of 296). Also in this case, we have previously shown that the ten best conserved contacts among the available models correspond to an average

distance in the native structure below 7 Å. [5] Therefore, these and other contacts with high conservation rate can correctly drive the ranking of the docking decoys.[5]

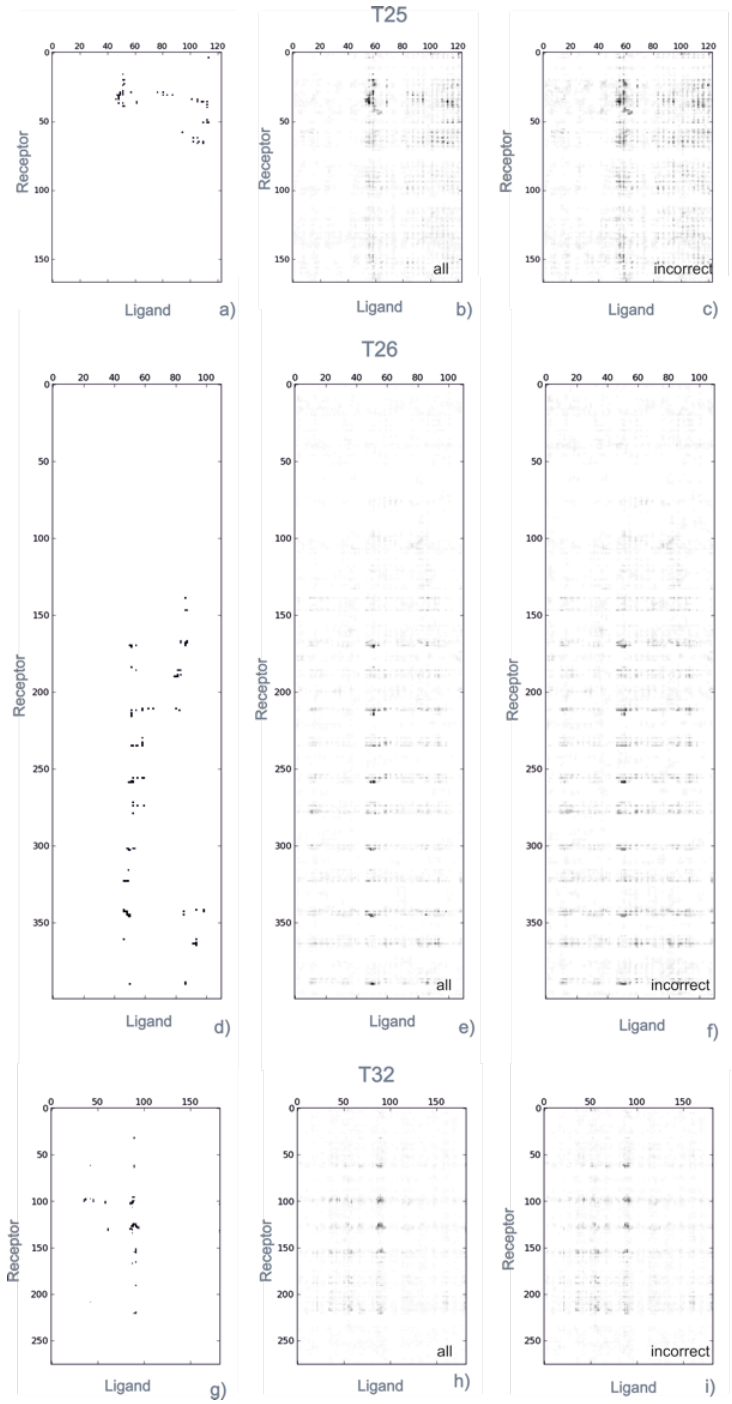


Figure 21. Consensus maps for the CAPRI T25, T26 and T32 targets.
Comparison between the consensus maps (b,e,h) obtained from all the 300, 310, and 350 models submitted to CAPRI for the targets T25, T26 and T32, respectively, the consensus maps (c,f,i) obtained from only the incorrect models and the intermolecular contact maps (a,d,g) of the corresponding native structures (PDBcodes: 2J59, 2HQS and 3BX1).

Dependence of the method performance on the percentage of native-like solutions

In Figure 22a, the obtained AUC values are reported vs. the percentage of native-like solutions available for each analysed target. As expected, AUC values are low for those targets having a very low percentage of native-like solutions and significantly increase for targets having a higher percentage of correct solutions. As a general rule, a percentage of 10% or better is guarantee of a performance better than random. However, AUC values approaching to 1 have also been found for many targets, especially in the RosettaDock and CAPRI benchmarks, having a percentage of native-like solutions as low as 5% or below. It is worth noticing that for the CAPRI targets a percentage of native solutions of 1.3 % or more leads to AUC values above 0.80 (and higher than 0.96 when considering targets with a percentage of correct solutions above 5%).

We also tried to correlate the maximum score obtained for each target, i.e. the $\overline{S_i}$ score of the top ranked decoy (ranging from 0.04 to 0.66), with the percentage of native-like solutions available. As it can be seen from Figure 22b, however, a linear correlation seems to emerge only for $\overline{S_i}$ values above 0.35 and percentages of native-like solutions higher than 40%. At lower values, instead, no clear correlation emerges and $\overline{S_i}$ values of 0.2 or 0.3 may correspond to a range of native-like percentages from 1 to 40%. Therefore, unless assuming very high values (above 0.35), the $\overline{S_i}$ absolute value alone is not sufficient to recognize decoy ensembles containing a significant faction of correct solutions.

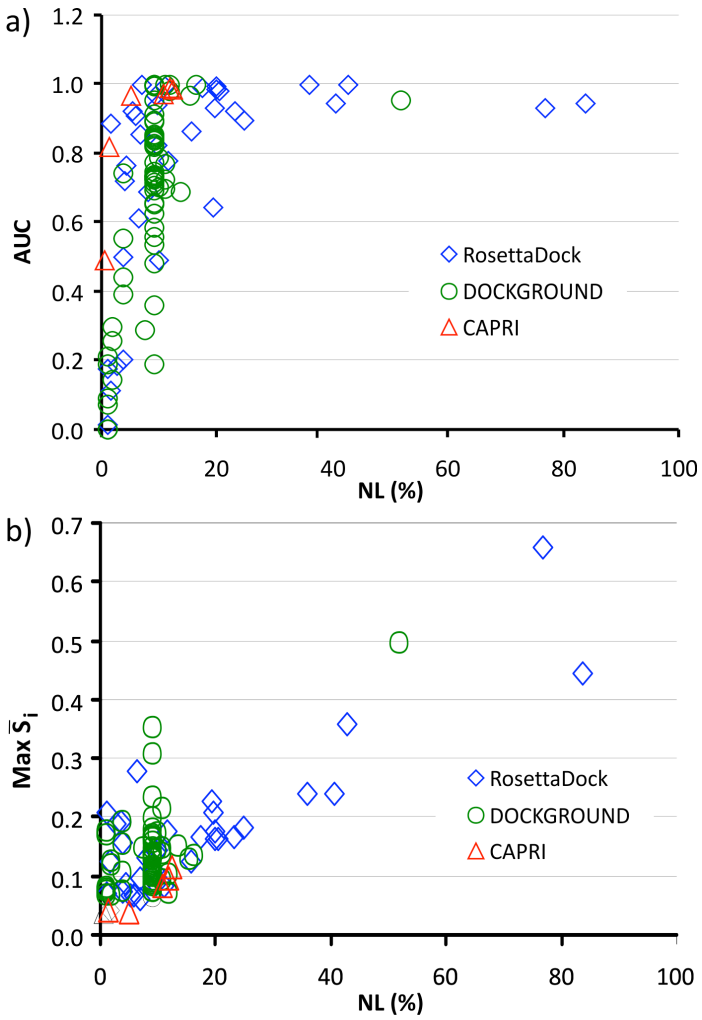


Figure 22. AUC value and Maximum score vs. the number of native-like solutions (%)

Charts of: a) the AUC-ROC value and b) the calculated Maximum score ($\bar{S}_i$) versus the percentage of native-like solutions for the analyzed targets in the RosettaDock, DOCKGROUND and CAPRI benchmarks.

Analysis of merged decoys from RosettaDock and DOCKGROUND

The previous analysis clearly indicated that our method outperforms on the CAPRI targets as compared to the RosettaDock and DOCKGROUND ones. It is reasonable to think that this depends on the fact that the CAPRI decoys have been obtained by several docking algorithms, whereas decoys in the other two analysed benchmarks came from single docking programs. In case this hypothesis is correct, merging decoys from different programs should improve the performance of the method. Luckily, this hypothesis could be tested as the DOCKGROUND and RosettaDock benchmarks have six common targets. For two of them, 1CHO and WQ1, AUC

values below 0.8 were obtained when using decoys from the single benchmarks, and no native-like solution was ranked within the top twenty positions. Therefore, we collected all the available decoys for these two targets and analyzed the augmented number of decoys (285 for target 1CHO and 296 for the target 1WQ1). Results summarized in Table 8, clearly show a significant improvement over results obtained when the single benchmarks were analyzed (Table 5 and Table 6). In particular, the twenty available native-like solutions for 1WQ1 were ranked between position 21 and 70, leading to an AUC value of 0.862 (it was 0.766 and 0.697 for RosettaDock and DOCKGROUND, respectively). For 1CHO the prediction power of the method improves even more, with 18 native-like solutions, out of the total 49, ranked in the top 20 positions and an AUC value as high as 0.898 (it was 0.644 and 0.688 for RosettaDock and DOCKGROUND, respectively).

A comparison of the native 1CHO intermolecular map with consensus maps obtained from the single RosettaDock and DOCKGROUND benchmarks and from the merged decoys is reported in Figure 23. It is pretty clear that in the RosettaDock and DOCKGROUND maps false contacts emerge, whose conservation competes with that of native-like ones. As hypothesized, such false contacts are different for the two benchmarks, and their conservation is consequently weakened when the decoys are analysed all together, allowing the native consensus to more easily emerge.

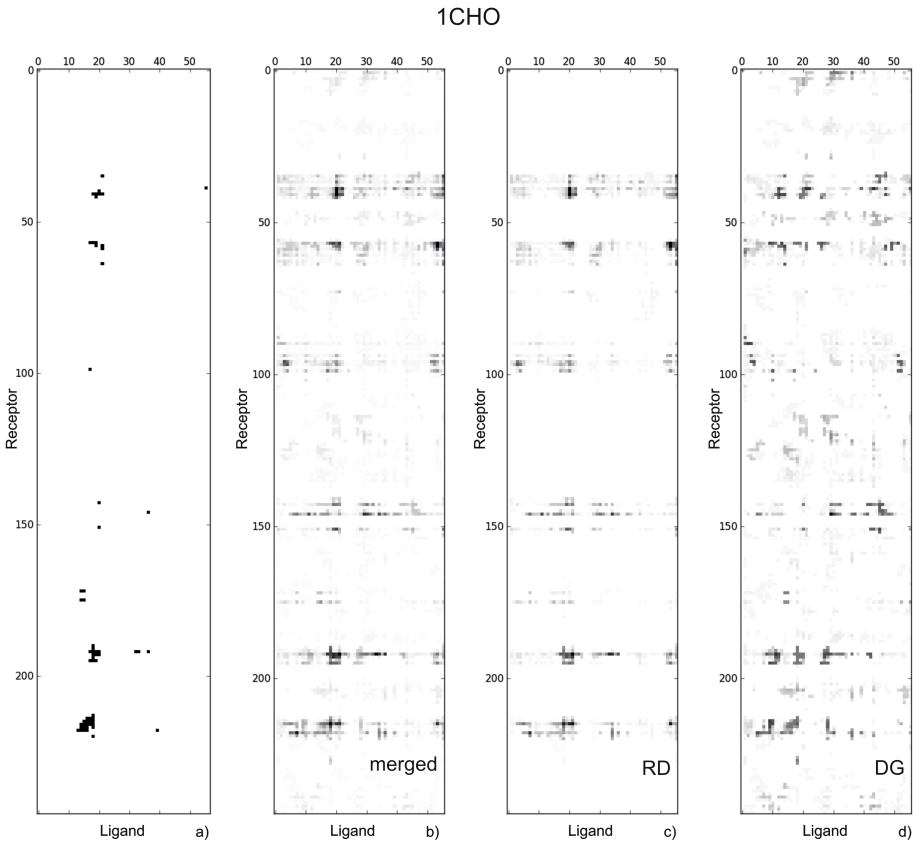


Figure 23. Consensus maps for the RosettaDock and DOCKGROUND 1CHO target.
Comparison between the COCOMAPS[4] intermolecular contact map of the 1CHO native structure (a), the consensus map obtained from all the 175 models in the RosettaDock Global-Unbound benchmark (c), the consensus map obtained from the 110 models in the DOCKGROUND benchmark (d) and the consensus map obtained from the 285 merged decoys from the two benchmarks (b).

4. Conclusions

In this chapter I described CONS-RANK, a simple and effective method to rank multiple docking solutions that I developed during my PhD project. The novelty and strength of the method is that it is based on the conservation of contacts at the complex interface: decoys are ranked according to their ability to match the most conserved contacts. We applied CONS-RANK to 102 targets from three different benchmarks, finding it to perform consistently well. CONS-RANK also proved able to specifically single out the high/medium quality solutions from the docking decoys ensemble. Remarkably, 46.2% and 72.3% of the total high/medium quality

predictions available for the RosettaDock and CAPRI targets, respectively, were ranked within the top twenty positions. Due to its philosophy, CONS-RANK performs particularly well when applied to decoys coming from different docking programs and procedures, as in the case of CAPRI targets, since the noise of the incorrect solutions from the specific docking procedures cancels, whereas the signal of the correct contacts gets stronger. For instance, an AUC value of 0.818 was obtained for the CAPRI target T24, having only 1.3% of acceptable solutions. We proved this concept on the 1WQ1 and 1CHO targets, which are common to the RosettaDock and DOCKGROUND benchmarks. Analysis of the merged decoys from the two benchmarks (AUC of 0.862 and 0.989 for 1WQ1 and 1CHO) indeed offers a clear improvement over analysis of the single benchmarks (AUC of 0.766 and 0.697 for 1WQ1 and of 0.644 and 0.688 for 1CHO), due to an increased signal to noise ratio in the analysis of the conserved contacts.

The main drawback of CONS-RANK is probably that it depends by its nature on the presence of correct solutions in the decoys ensemble. However, this seems to be a common feature to scoring algorithms, as evidenced in the last CAPRI edition, where it was shown that the success rate of scoring algorithms strongly depends on the percentage of available models of acceptable or better quality. We remind that a significant enrichment of native-like solutions by scoring algorithms was observed only in few cases, among those having a percentage of correct solutions of 5% or higher.[126]

Nevertheless, the approach to the ranking of docking solutions we have presented is very well performing and robust, thus offering a valid alternative to the ranking methods already available. Our approach can be particularly useful to analyse docking solutions collected from different docking procedures. Analysis is extremely fast, and hundreds of docking decoys can be reliably ranked in minutes on a standard PC.

Table 5
Summary of the ranking of RosettaDock targets. N-decoys is the total number of decoys; N-HM is the number of high or medium quality models; N-NL is the number of native-like, i.e. acceptable or better, models; R5-HM and R5-NL are the number of HM and NL models ranked in the top 5 positions, respectively.; R10-HM and R10-NL are the number of HM and NL models ranked in the top 10 positions, respectively.; R20-HM and R20-NL are the number of HM and NL models ranked in the top 20 positions, respectively.

Target	N decoys	N HM	N NL	R5 HM	R5 NL	R10 HM	R10 NL	R20 HM	R20 NL	AUC
1A0O	184	1	36	0	4	0	9	0	16	0.931
1ACB	181	1	5	0	0	0	0	0	0	0.186
1AHW	171	0	9	0	0	0	3	0	5	0.923
1ATN	185	9	13	5	5	9	10	9	13	1.000
1AVW	177	1	12	0	0	0	0	0	1	0.856
1AVZ	177	0	2	0	0	0	0	0	0	0.174
1BQL	178	18	31	4	5	9	10	17	20	0.991
1BRS	179	2	28	1	2	2	4	2	8	0.864
1BVK	99	0	76	0	5	0	10	0	20	0.930
1CGI	182	18	37	4	5	8	10	13	20	0.979
1CHO	175	5	34	0	0	0	0	0	0	0.644
1CSE	190	0	19	0	0	0	0	0	0	0.492

1DQJ	197	0	2	0	0	0	0	0	0	0.015
1FBI	190	3	3	0	0	0	0	0	0	0.114
1FSS	179	6	7	0	0	0	0	0	0	0.718
1JHL	186	0	12	0	0	0	0	0	0	0.609
1MAH	169	9	10	0	0	2	2	4	5	0.906
1MEL	181	9	36	2	5	5	10	8	18	0.984
1MLC	187	5	7	0	0	0	0	0	0	0.501
1PPE	179	43	150	1	5	2	10	8	20	0.945
1QFU	176	2	3	0	0	0	0	2	2	0.886
1SPB	174	8	14	0	0	0	0	0	0	0.686
1STF	184	16	18	5	5	10	10	14	15	0.938
1TAB	185	25	43	5	5	10	10	20	20	0.920
1TGS	185	12	46	2	5	4	10	7	20	0.895
1UDI	163	10	18	4	5	7	10	10	16	0.994
1UGH	181	33	65	3	5	6	10	11	20	1.000
1WQ1	186	0	8	0	0	0	0	0	0	0.766
2JEL	192	31	78	4	5	6	10	13	20	0.94

2KAI	180	56	77	5	5	10	10	19	20	5 0.996
2PTC	192	4	7	0	0	0	0	0	0	0.202
2SIC	179	14	17	5	5	7	7	10	11	0.965
2SNI	184	5	18	0	1	1	4	1	6	0.822
2TEC	183	15	21	0	0	0	0	0	0	0.776
4HTC	180	9	36	0	5	0	9	3	19	0.994
TOT	6270	370	998	50	82	98	168	171	315	
%		5.9[a]	15.9[a]	13.5[b]	8.2[c]	26.5[b]	16.8[c]	46.2[b]	31.6[c]	
average	179.1	10.6	28.5	1.4	2.3	2.8	4.8	4.9	9.0	0.758 0.799[d]
ST-DEV	15.5	13.1	30.4	2.0	2.4	3.7	4.7	6.3	8.8	0.292 0.247[d]

[a] Compared to N-decoys. [b] Compared to N-HM. [c] Compared to N-NL. [d] Values obtained by excluding the 2 targets having only two native-like solutions.

Table 6
Summary of the ranking of the DOCKGROUND targets. N-decoys is the total number of decoys; N-HM is the number of high or medium quality models; N-NL is the number of native-like, i.e. acceptable or better, models; R5-HM and R5-NL are the number of HM and NL models ranked in the top 5 positions, respectively.; R10-HM and R10-NL are the number of HM and NL models ranked in the top 10 positions, respectively.; R20-HM and R20-NL are the number of HM and NL models ranked in the top 20 positions, respectively.

Target	N decoys	N HM	N NL	R5 HM	R5 NL	R10 HM	R10 NL	R20 HM	R20 NL	AUC
1A2K	102	2	2	0	0	0	0	0	0	0.295
1A2Y	110	10	10	0	0	0	0	0	0	0.536
1AKJ	110	10	10	0	0	1	1	9	9	0.894
1AVW	110	10	10	0	0	0	0	0	0	0.656
1BTH	101	1	1	0	0	0	0	0	0	0.190
1BUI[a]	110	10	10	2	2	4	4	5	5	0.854
1BUI[b]	110	10	10	0	0	1	1	4	4	0.822
1BVN	110	10	12	0	0	0	0	0	0	0.770
1CHO	110	10	15	0	0	0	0	0	0	0.688
1DFJ	109	9	10	0	0	0	0	1	1	0.774
1E96	110	10	10	0	0	0	0	5	5	0.839
1EWY	110	10	10	0	0	0	0	0	0	0.624
1EZU	110	10	10	0	0	0	0	0	0	0.190
1F51	110	10	10	0	0	0	0	2	2	0.735
1F6M	110	10	10	0	0	0	0	0	0	0.360
1FM9	110	10	13	5	5	9	10	10	13	1.000
1G20	110	10	10	0	0	0	0	0	0	0.746
1G6V	108	8	8	0	0	0	0	0	0	0.286
1GPQ	110	10	10	2	2	4	4	7	7	0.911
1GPW	110	10	18	3	5	7	10	10	18	1.000
1HE1	110	10	13	5	5	9	9	10	12	0.979
1HE8	101	1	1	0	0	0	0	0	0	0.000
1HXY	102	2	2	0	0	0	0	0	0	0.145

1JPS	110	10	10	0	0	0	0	4	4	0.829
1KU6	110	10	10	0	0	0	0	0	0	0.706
1L9B	110	10	10	0	0	0	0	0	0	0.557
1MA9	110	10	10	0	0	0	0	6	6	0.849
1NBF	110	10	10	0	0	0	0	0	0	0.713
1OOK	104	4	4	0	0	0	0	0	0	0.552
1OPH	110	10	10	0	0	1	1	7	7	0.888
1P7Q	104	4	4	0	0	0	0	0	0	0.440
1PPF	110	10	10	5	5	9	9	10	10	0.992
1R0R	110	10	12	0	0	0	0	0	0	0.722
1R4M	101	1	1	0	0	0	0	0	0	0.070
1S6V	104	4	4	0	0	0	0	0	0	0.390
1T6G	110	10	57	1	5	1	10	1	18	0.951
1TMQ	110	10	10	0	0	0	0	0	0	0.586
1TX6	110	10	10	0	0	0	0	0	0	0.482
1U7F	110	10	10	0	0	2	2	8	8	0.847
1UEX	101	1	1	0	0	0	0	0	0	0.210
1UGH	110	10	12	5	5	9	10	10	12	1.000
1W1I	104	4	4	0	0	0	0	0	0	0.742
1WEJ	110	10	10	0	0	0	0	0	0	0.729
1WQ1	110	10	12	0	0	0	0	0	0	0.697
1XD3	110	10	10	0	0	0	0	0	0	0.722
1XX9	102	2	2	0	0	0	0	0	0	0.255
1YVB	110	10	10	0	0	0	0	0	0	0.650
1ZY8[c]	110	10	10	5	5	9	9	10	10	0.999
1ZY8[d]	110	10	10	5	5	5	5	8	8	0.955
2A5T	101	1	1	0	0	0	0	0	0	0.090
2BKR	110	10	11	0	0	0	0	2	2	0.788
2BNQ	101	1	1	0	0	0	0	0	0	0.000
2BTF	110	10	10	5	5	9	9	10	10	0.998
2CKH	110	10	10	0	0	0	0	0	0	0.722
2FI4	110	10	10	0	0	1	1	3	3	0.837
2GOO	110	10	10	0	0	0	0	1	1	0.734

2KAI	110	10	11	0	0	0	0	0	0	0.700
2SNI	110	10	10	0	0	0	0	0	0	0.692
3FAP	110	10	10	0	0	0	0	3	3	0.734
3PRO	110	10	17	4	5	6	10	8	14	0.967
3SIC	110	10	10	0	0	0	0	2	2	0.817
TOT	6605	505	589	47	54	87	105	156	194	
%		7.6[e]	8.9[e]	9.3[f]	9.2[g]	17.2[f]	17.8[g]	30.9[f]	32.9[g]	
average	108.3	8.3	9.7	0.8	0.9	1.4	1.7	2.6	3.2	0.654 0.743[h]
ST-DEV	3.3	3.3	7.3	1.7	1.9	2.9	3.4	3.7	4.9	0.280 0.191[h]

[a] 1BUI_A:C. [b] 1BUI_B:C. [c] 1ZY8_AB:K1. [d] 1ZY8_AB:K2. [e] Compared to N-decoys. [f] Compared to N-HM. [g] Compared to N-NL. [h] Values obtained by excluding the 9 targets having only one or two native-like solutions.

Table 7
Summary of the ranking of the CAPRI targets. N-decoys is the total number of decoys; N-HM is the number of high or medium quality models; N-NL is the number of native-like, i.e. acceptable or better, models; R5-HM and R5-NL are the number of HM and NL models ranked in the top 5 positions, respectively.; R10-HM and R10-NL are the number of HM and NL models ranked in the top 10 positions, respectively.; R20-HM and R20-NL are the number of HM and NL models ranked in the top 20 positions, respectively.

Target	N decoys	N HM	N NL	R5 HM	R5 NL	R10 HM	R10 NL	R20 HM	R20 NL	AUC
T24	296	0	4	0	0	0	0	0	0	0.818
T25	268	13	32	2	5	5	10	9	19	0.990
T26	276	15	34	3	8	10	10	13	20	0.986
T29	333	9	17	4	5	5	10	9	16	0.965
T32	316	28	34	4	4	8	9	16	17	0.969
T36	199	0	1	0	0	0	0	0	0	0.490
TOT	1688	65	122	13	22	28	39	47	72	1688
%		3.9[a]	7.2 %[a]	20.0 [b]	18.0[c]	43.1 [b]	2.0[c]	72.3 [b]	59.0[c]	
Average	281.3	10.8	20.3	2.2	3.7	4.7	6.5	7.8	12.0	0.870
ST-DEV	47.1	10.5	15.2	1.8	3.1	4.1	5.0	6.6	9.4	0.197

[a] Compared to N-decoys. [b] Compared to N-HM. [c] Compared to N-NL.

Table 8
Summary of the ranking of merged decoys from RosettaDock and DOCKGROUND for the 1CHO and 1WQ1 targets. N-decoys is the total number of decoys; N-HM is the number of high or medium quality models; N-NL is the number of native-like, i.e. acceptable or better, models; R5-HM and R5-NL are the number of HM and NL models ranked in the top 5 positions, respectively.; R10-HM and R10-NL are the number of HM and NL models ranked in the top 10 positions, respectively.; R20-HM and R20-NL are the number of HM and NL models ranked in the top 20 positions, respectively.

Target	N decoys	N HM	N NL	R5 HM	R5 NL	R10 HM	R10 NL	R20 HM	R20 NL	AUC
1CHO	285	15	49	1	5	2	9	6	18	0.898
%		5.3[a]	17[a]							
1WQ1	296	10	20	0	0	0	0	0	0	0.862
%		3.4[a]	6.8[a]							

[a] Compared to N-decoys.

CHAPTER 5 - Study of the interaction between celiac auto-antibodies and the auto-antigen Tissue Transglutaminase (TG2)

5.1 - Introduction

The immune system

One of the most important and fascinating example of protein-protein interaction is the complex made by an antigen and its antibody.

The antibodies, or immunoglobulins, are a class of protein at the basis of the immune system. The immune system, in fact, provides a defense mechanism against foreign parasites such as virus and bacteria. Foreign invaders, the antigens, are recognized through specific binding of the antibodies. The site on foreign molecules that are specifically recognized by the antibody is called antigenic determinant or epitope.

Structurally, an antibody is a "Y"-shaped protein composed by a light (L) and a heavy (H) chain linked together by disulfide bonds (see Figure 24). There are two different classes, or isotypes, of light chains, λ and κ, but there is no known functional distinction between them. Heavy chains, by contrast, have five different isotypes that divide the antibodies into different functional classes: IgG, IgM, IgA, IgD and IgE, each with different effector properties in the elimination of the antigen. Each class of heavy chain can combine with either of the two different classes of the light chain. The IgG class is the major type of immunoglobulin in normal human serum, and it has the simplest structure.

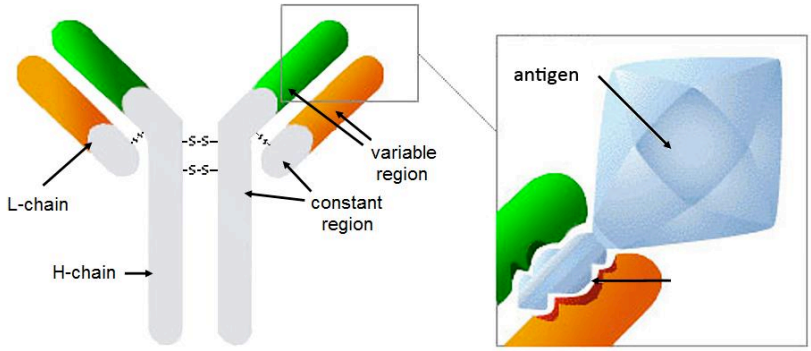


Figure 24. Schematic representation of an antibody.
On the left: The heavy (in green and gray) and light (in orange and gray) chains are represented, connetected by disulphide bonds (in black). The variable region is highlight in green and orange for the H and L chains, respectively. *On the right*: the detail of the CDR region, interacting with the antige, is represented.

Both light and heavy chains are built up from one amino-terminal variable domain (V_L and V_H, respectively) (the two "arms" of the Y) and one carboxy-terminal constant domain (C_L and C_{H1}, C_{H2} and C_{H3}) (see Figure 24) . The variable domains are not uniformly variable throughout their length; in particular, three small regions for both L and H chains show much more variability than the rest of it: they are called complementary determining regions (CDRs), made by six hypervariable loops, as showed in Figure 25. They are indicated with L1, L2, L3, H1, H2, H3 depending on the chain, and represent the binding site for the antibody, presenting a specific and complementary structure for the antigen recognition.

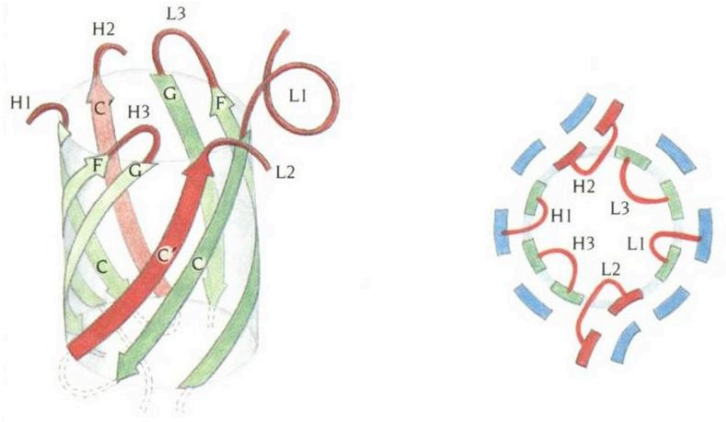


Figure 25. Schematic representation of the six hypervariable loops
Side (on the left) and top (on the right) view of a schematic representation of arrangement of the CDR and the six hypervariable loop. Labels to each loop have been addes.

This six loops show a variability in the length and in the amino acids composition, being specific for the antigen. In particular, the loop H3 is the most variable in sequence and structure, having a key role in the antigen recognition.
Therefore, the study of a binding between and an antibody and its can be focused on the analysis of the six hypervariable loops.[149]

The autoimmunity and celiac disease

The antibody-antigen recognition is an high precise and specific interaction, but sometimes the immune system mistakes parts of the body as a pathogen, attacking its own cells and proteins. When the body arise an immune response against a self antigen, there is the development of a peculiar class of disease, termed autoimmune disease. This is a quite common pathologies (involving more than the 5% of the population), and their gravity can vary on the basis of the organ and tissue that is erroneously recognize by auto-antibodies, i.e. antibodies direct against a self antigen.
Multiple sclerosis, Mellitus diabetes, some kind of allergies are all examples of autoimmune disease, and are that characterized by the presence of auto-antibodies.[150,151]

One of the most common food intolerance in Europe has been classified as auto-immune disease, being characterized by the presence of auto-antibodies: the celiac disease.[150]Celiac disease is a multifactorial disorder affecting approximately 1 in 100 individual in the European population.[152] It is a long-life food intolerance affecting susceptible individuals and caused by the exposure to the gluten, the constituent protein in wheat and cereal.[153] The presence of gluten causes an abnormal immune response not only against the gluten's proteins, but also againt the self antigen Tissue Transglutaminase (or type 2 Transglutaminase TG2).[154] The clinical consequence is an intestinal mucosal injury and malabsorption; the absence of typical symptoms makes the celiac disease not easy to diagnose. [155]
Therefore, the disease is characterized by the presence of specific antibodies recognizing gliadins (the food proteins comes from the gluten digestion) and the autoantigen TG2. TG2 is a member of a family of seven isoforms of enzymes involved in protein cross-linking. It is a Ca^{2+} -dependent ubiquitous intracellular enzyme that catalyzes the covalent and irreversible formation of gamma glutamyl-

lysine bonds. Furthermore, TG2 plays a role in the transduction of extracellular signals, mediated by its additional GTP-hydrolyzing activity.[156]

Human TG2 consists of four domains:

- the N-terminal domain, with a β -sandwich structure;
- the enzyme Core domain, formed by a series of α-helices;
- two C-terminal domains, β1 and β2, containing β-structures arranged in barrel-like conformations.

The catalytic site, the so called triad,[157] formed by Cys 277, His 335 and Asp 358, as well as the Ca^{2+} and the GTP binding sites, are located in the Core domain and the nearby first β-barrel domain. [158]

Experimental x-ray structures of several transglutaminase have been crystallized,[158-161] proving that the four TG2's domains can organize themselves in two different ways. In one case, named closed conformation, the two C-terminal domains are folded on the Core domain, hiding the active site; in the other one, named open conformation, the four domains are straight on an axis, exposing the active site (see Figure 26). From the superimposition of the two TG2's conformations is so apparent that the conformational change is not changing the 3D orientation of the first two domains between the closed and open conformation. To crystallize and so to solve the structure of the two conformations of the TG2, it was required the stabilization of the two transient state. Then, the human TG2 open conformation has been crystallized in a GDP-bound state, while the TG2 closed conformation has been crystallized in an inhibitor-bound state.

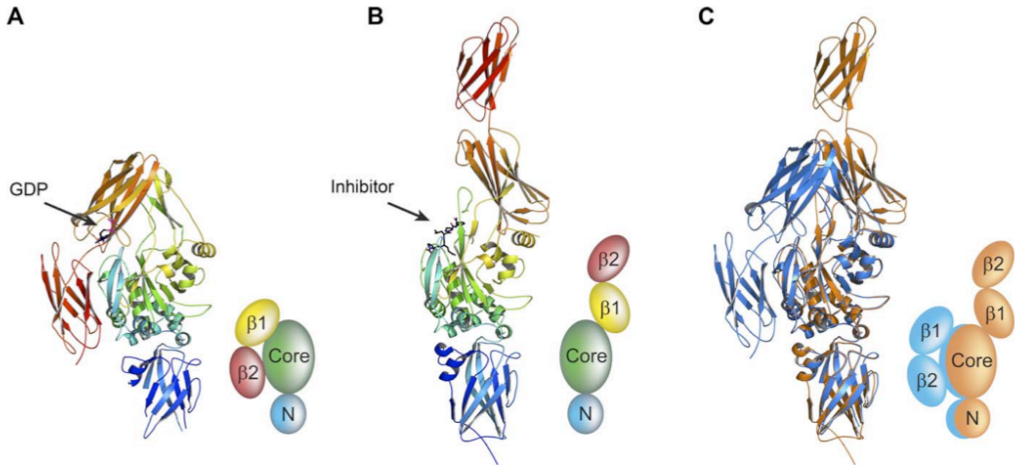


Figure 26. Overall structures of CDP-bound and inhibitor-bound TG2
The crystal structures are shown as ribbons, and simplified cartoons are included for clarity.(A and B) The N-terminal β-sandwich is shown in blue (N), the catalytic domain (Core) in green, and the C-terminal β -barrels (β1 and β2) in yellow and red, respectively. (A) GDP-bound TG2 (PDB ID: 1KV3[158]). (B) TG2 inhibited with the active-site inhibitor Ac-P(DON)LPF-NH2 (PDB ID: 2Q3Z[162]).(C) The N-terminal β-sandwich and catalytic domains of the two structures are superimposed, highlighting the conformational change. The GDP-bound structure (named "TG2-closed") is shown in blue and the inhibitor-bound structure (named "TG2-open") in gold.

Studies conducted on celiac patients have demonstrated the presence of the auto-antibodies against the TG2 in the blood of the patients as a peculiarity of this illness.

The TG2/auto-antibody interaction has so a fundamental role in the study of the disease, founding a substantial role in the diagnosis.[154,163] In the previous year, in fact, the only way to diagnose this illness was through an intestinal biopsy, to check an eventual tissue damage. Nowadays, a blood test is enough for a first diagnosis of it, checking the presence of auto-antibody specific for the TG2 through an ELISA test.[154,163]

Due to the key role that TG2 seems to have in the pathogenesis of celiac disease, and the fundamental role in the diagnosis strategy, it is of great importance the characterization at atomic levels of the TG2's epitope and, consequently, the binging mode of the complex TG2/auto-antibody.

Experimental studies

For this reason, in the last decades, some experimental studies have been performed. Sblattero *et al*. isolated and characterized in sequence some anti-transglutaminase

antibodies from blood of celiac patients,[7] and then, using transglutaminase gene fragments, they identified a region of TG2 recognized by these antibodies (in particular three, called clone 2.8, clone 4.1 and clone 3.7) as being conformational and located in the *Core* domain of the enzyme; in particular, they proved the epitope belong to the fragment 140-376, showed in Figure 27 in red and in blue for the TG2-closed and TG2-open, respectively.[155]

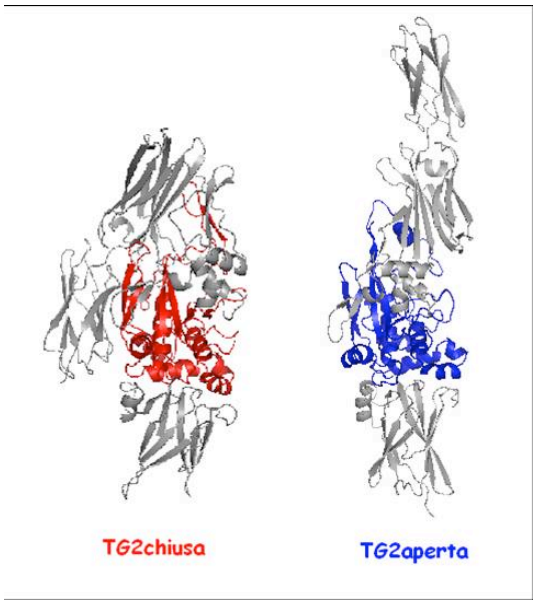


Figure 27. Representation of the Core domain of TG2
The closed (*on the left*) and open (*on the right*) conformation of TG2 are showed in cartoon. The fragment 140-376 in which the epitope is located is highlighted in red (for the TG2-closed) and in blue (for the TG2-open).

Due to the importance of anti-TG2 antibodies in diagnosis and pathogenesis of the celiac disease, we believe that the characterization at atomic level of the interaction between the TG2 (both closed and open conformation) with clone 2.8, clone 3.7 and clone 4.1 could be of great interest. For this reason, we performed docking simulation between the autoantibody the two TG2's conformations and the three Abs from CD patients, taking in account the experimental data.

5.2 - Methods

Abs and TG2 strucutres

The structures of both TG2's closed and open conformation were experimental solved and deposited in the Protein Data Bank, with the code 1KV3 and 2Q3Z, respectively. The three anti-TG2 antibodies clone 2.8, clone 3.7 and clone 4.1 were isolated from celiac patients blood and the sequences were characterized by the group of the prof. Sblattero, University of Piemonte Orientale (Italy). Then, the variable domain structures of clone 2.8, clone 3.7 and clone 4.1 were modeled by the RosettaAntibody Fv homology modeling server,[164] using the full refinement protocol option.

Docking simulations

The TG2s crystal structures and the obtained Abs models were then used for protein-protein docking simulations, performed by the ClusPro 2.0 server (see Chapter 1 and Appendix 2 for details about docking technique).[56] By default, ClusPro server docks the two proteins using PIPER rigid-body docking algorithm. The top 2000 complexes generated by PIPER are then filtered according to electrostatic and desolvation energies and retained for further processing. The retained 2000 conformations are then clustered according to interface RMSD values and the top 10 docked models, following a short Charmm10 energy minimization, are made available for download. In all the simulations, all the Abs' residues that do not fall into the Complementary Determining Region (CDR) were masked (ClusPro Antibody Mode). Differently, for the two TG2 conformations all the residues were considered on an equal basis.

Analysis

The representative structures of the top clusters for each simulation were analyzed. To analyze the docking results we used script on the basis of COCOMAPS web tool.

5.3 - Results and Discussion

We performed docking simulations to obtain the structure of the complex between the TG2 and the antibodies anti-TG2 isolated from CD patients. For both closed (TG2closed) and open (TG2open) conformation of TG2, the experimental structures from Protein Data Bank were used, while the Abs structure of clone 2.8, clone 3.7 and

clone 4.1 were reliably predicted by homology modeling using Rosetta Antibody. The docking simulations were performed by ClusPro 2.0.

In the first moment, we ran a set of six docking simulation, testing all the combination between the three Abs (clone 2.8, clone 3.7 and clone 4.1) with the two possible experimental conformation of the antigen TG2 (closed and open conformation).
However, as experimental data showed that TG2 interacts with these Abs using only the *Core* domain, indicating that the epitope should be in that region, we additionally ran another set of docking simulations in according with the experimental data, for a total of twelve simulations. In fact, for the same systems, we repeated the docking using only the 1-376 fragment for the two TG2open and TG2closed structures.

FIRST SET OF SIMULATION	**SECOND SET OF SIMULATION**
TG2closed + clone 2.8	TG2closed (fragment 1-376)+ clone 2.8
TG2closed + clone 4.1	TG2closed (fragment 1-376)+ clone 4.1
TG2closed + clone 3.7	TG2closed (fragment 1-376)+ clone 3.7
TG2open + clone 2.8	TG2open (fragment 1-376)+ clone 2.8
TG2open + clone 4.1	TG2open (fragment 1-376)+ clone 4.1
TG2open + clone 3.7	TG2open (fragment 1-376)+ clone 3.7

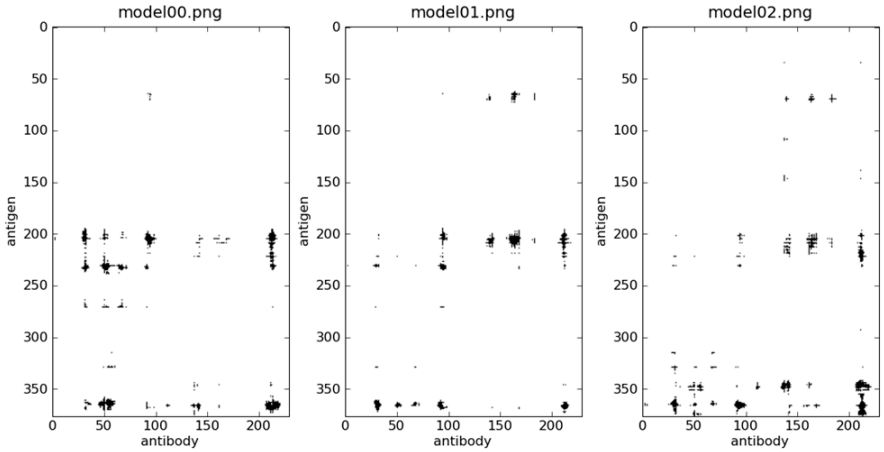
Table 9
List of the TG2-Ab systems used in the first (*on the left*) and second (*on the right*) set of simulations.

In our ClusPro simulations, as generally done for antibodies, the Abs were fixed and all their residues not falling into the CDR were masked, while all the antigen's residues were considered unmasked, and therefore available for the interaction.
For each of the twelve simulations, we then analyzed the representative structure of the top three clusters, that should capture most of the important rigid-body binding geometries, and we compared the results.
The docking results of the first set of simulations (using the whole TG2s' residues) corresponded to the ones obtained from the second set (using only the 1-376 fragment of the two TG2 conformation - Figure 27), therefore the data and analysis show below concern the simulations using only the second set of them (see Table 9).

To easily and intuitively compare the solutions came from the various simulations, we made intermolecular contact maps for the best three clusters for each of the six systems (see Table 9), reported in Figure 28.

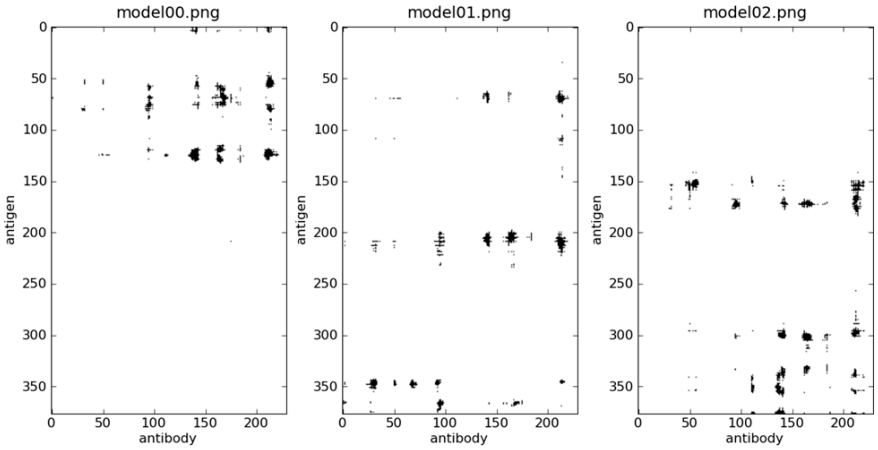
CLONE 2.8 + TG2-OPEN

Popolazione:266 | Popolazione:150 | Popolazione:47



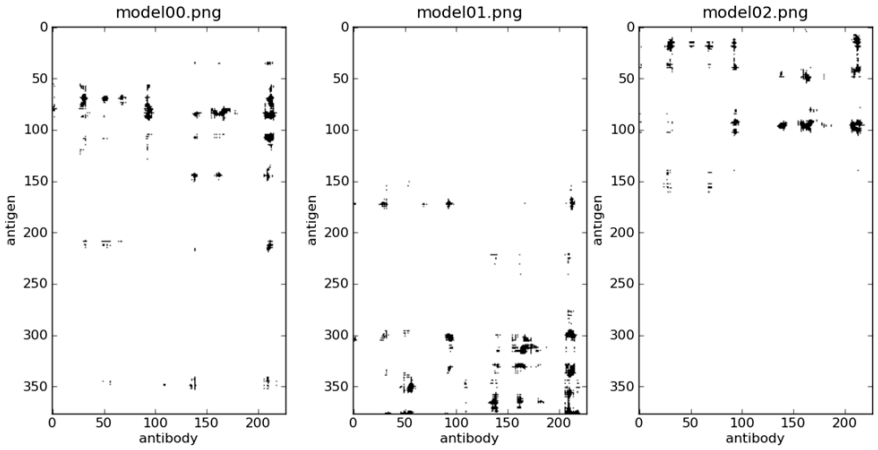
CLONE 4.1 + TG2-OPEN

Popolazione:139 | Popolazione:96 | Popolazione:73



CLONE 3.7 + TG2-OPEN

Popolazione:111 | Popolazione:57 | Popolazione:50



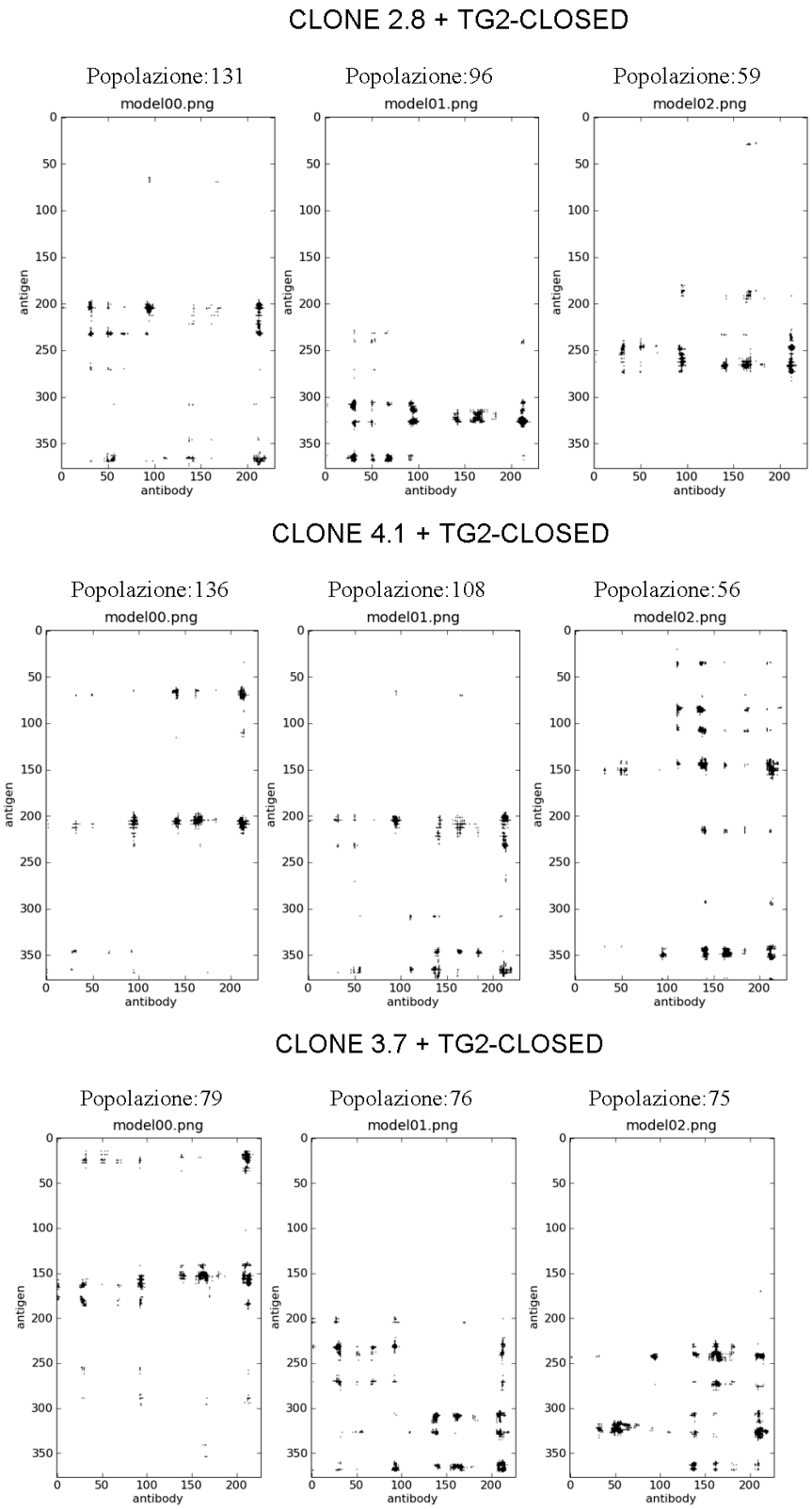


Figure 28. Contact maps

Contact maps for the best three clusters between the Abs (clone 2.8, clone 4.1 and clone 3.7) and the TG2 1-376 fragment for both the closed and the open conformation. Labels and titles have been added of each contact map. The population of each cluster is also indicated on the top of the map.

Abs/TG2 open systems

In the simulations involving the open conformation of TG2, the analysis of clone 2.8/TG2open system showed that:

- the first and second top solutions are very populated;
- the top three solutions present more o less the same epitope, made up of three interacting zones: the first zone is around the amino acids 200-210, the second one is around the amino acid 230, and the third one is around the amino acid 365. We defined this epitope as "EP1".

Summing the population of these three clusters, we obtained more than 500/2000, increases our confidence in the EP1 solution as a close-native one.

Looking at the results for clone 4.1/TG2open complex, the first best solution (population: 111) clearly pointed in the N-terminal domain of TG2, so we excluded the this solution as disagreement with the experimental data. The second best solution (population: 96) presented EP1 as epitope, while the third one pointed in the *Core*, but in a region different from EP1.

Finally, in the clone 3.7/TG2 open solutions, the second one is the only one involving the *Core* domain in the interaction, but is a different region from EP1.

Abs/TG2 closed systems

In the simulations involving the closed conformation of TG2, the first top solution for clone 2.8/TG2closed showed again EP1 as epitope, even though the preference (in terms of population) is less if compared with the clone 2.8/TG2 open.

The clone 4.1/TG2closed presented similar results, again with a preference for EP1, while the clone 3.7/TG2closed system seed to prefer other interaction regions.

In conclusion, the analysis of the contact maps showed that the complexes involving clone 2.8 and clone 4.1 preferred to interact through EP1, and this solution is preferential in particular for the clone 2.8/TG2open system, in clear accord with the experimental data. Therefore, this increases our confidence in proposing EP1 as possible epitope for TG2.

Finding the key-residues for the interaction

To characterize the interaction interface and have a clue about the most common TG2 residues at the interface, we performed a statistical analysis counting how many times

a TG2 residue is at the interface with the Ab in the best top twenty solutions for each simulation. In particular, a TG2 residue is at the interface if any atom is closer than the cut-off distance of 5 Å from any Ab's residue. In Table 10 are reported the residues that are at the interface at least in the 20% of the models analyzed.

The analysis of Table 10, such as the contact maps in Figure 28, shows that the TG2's regions involved in the interaction are the same in both the complexes TG2/clone 2.8 and TG2/clone 4.1, in according with the experimental data. In particular, it is possible to identify three specific regions:

- region 200-230;
- region around the residue 345;
- region around the residue 365

Therefore, the statistical analysis of the top twenty clusters for the systems involving clone 2.8 and clone 4.1, interacting with both open and closed conformation of TG2, converges towards EP1 (defined of the basis of the best solution for the system clone2.8/TG2-open), with the only addition of the region around the residue 345. This result increases our confidence in EP1 as native-like epitope.

aa TG2		Clone2.8	Clone4.1	Clone3.7
TRP	142	2.5	0.0	22.5
TYR	149	2.5	7.5	27.5
SER	152	7.5	15.0	27.5
GLU	153	7.5	5.0	40.0
GLU	154	7.5	17.5	35.0
GLU	155	7.5	15.0	22.5
ARG	156	2.5	2.5	35.0
GLN	157	2.5	10.0	32.5
GLU	158	2.5	12.5	27.5
TYR	159	5.0	7.5	20.0
THR	162	2.5	17.5	20.0
GLN	166	0.0	10.0	20.0
LYS	173	10.0	20.0	15.0
PRO	201	20.0	12.5	5.0
LYS	202	35.0	20.0	5.0
LYS	205	40.0	30.0	7.5
ASN	206	30.0	30.0	2.5
ARG	209	35.0	35.0	5.0
ARG	213	25.0	25.0	0.0
TYR	219	27.5	15.0	0.0
ARG	222	27.5	20.0	2.5
ASN	231	27.5	20.0	7.5
ASP	232	32.5	17.5	15.0
ASP	233	25.0	17.5	7.5
ARG	296	5.0	10.0	20.0
ASN	308	20.0	2.5	17.5
SER	309	20.0	5.0	10.0
GLU	314	20.0	5.0	10.0
TYR	315	20.0	10.0	7.5
PHE	316	22.5	2.5	17.5
SER	328	20.0	2.5	17.5
GLU	329	22.5	12.5	17.5
ARG	344	17.5	22.5	2.5
PRO	345	22.5	22.5	0.0
ASP	346	20.0	30.0	0.0
LEU	347	22.5	25.0	0.0
GLN	348	25.0	25.0	10.0
PRO	349	20.0	20.0	10.0
TRP	354	5.0	12.5	22.5
GLU	363	35.0	15.0	20.0
LYS	364	35.0	17.5	17.5
SER	365	45.0	22.5	20.0
GLU	366	52.5	25.0	15.0
GLY	367	45.0	22.5	17.5
THR	368	37.5	17.5	17.5
TYR	369	22.5	15.0	7.5

Table 10

List of TG2's amino acids numbering and typology at the interface with the Abs at least in the 20% of all the models analyzed (columns 1-2), with the corresponding percentage (column 3-5).

Simulations on the mutants

A valid experimental approach to identify and verify the nature of a binding site is testing the affinity of mutants of the protein. A mutant has the same structure of the wild-type protein itself, with exception of the amino acids that are considered involved in the binding site; in fact, these residue are mutated in a different amino acid with different chemical-physical properties. If the mutated residues are involve in the interactions, their mutation will compromise the binding with the molecule partner.

Then, to test the reliability of our result (EP1 as binging site), we performed docking simulations testing mutants of the TG2. The simulations were so performed on the system that showed the highest selectivity for EP1: clone2.8/TG2-open. The method is based on the idea that if the mutations are able to decrease a so strong affinity *in silico*, probably they can give the same result in experimental tests too. So, we should be able to identity the most promising mutants.

Considering the residues in Table 10 with a high frequence at the interface, we designed the list of mutants reported in Table 11, in which all the key residues were muted in alanine, and the region 360-369 (corresponding to a loop) was deleted.

Mutant	Mutated residues
M1	202+205 mutati in A
M2	209+213 mutati in A
M3	232+366 mutati in A
M4	202+205+232 mutati in A
M5	202+205+209+213 mutati in A
M6	202+205+232+366 mutati in A
M7	202+205+209+213+222 mutati in A
M8	360-369 deleted
M9	202+205 mutati in A e loop 360-369 deleted
M10	202+205+209+232 mutati in A e loop 360-369 deleted
M11	202+205+206+209+232+346 mutati in A e loop 360-369 deleted
M12	202+205+206+209+213+231+232+233+346+347+348 mutati in A e loop
M13	365+366+368 mutati in A
M14	365+366 mutati in A 368 mutato in V
M15	202+205+232+365+366+368 mutati in A
M16	202+205+232+365+366 mutati in A 368 mutato in V
M17	202+205+232+365+366+368+346
M18	202+205+209+232+365+366+368
M19	202+205+209+232+365+366+368 mutato in V
M20	202+205+209+232+346+348+365+366+368
M21	202+205+209+232+346+348+ loop(360-369) deleted

Table 11
List of the mutants of the 1-376 TG2-open's fragment used in the docking simulations with clone 2.8

In Figure 29 are reported the results of the docking simulations between clone 2.8 and the TG2's mutants (listed in Table 11) compared with the wild-type (WT) TG2 result. In particular, we looked at the population of the cluster reporting EP1 as representative solution, due to the fact that in the docking technique the cluster's population is generally a signal about how preferential is the solution. From FIGX it is apparent that the mutants involving only one of the three regions characterizing EP1, i.e. the one around 200-210 (M1 and M2 mutants) or around 360-369 (M8 mutant), show again EP1 as preferential binding mode. In fact, in these cases EP1 is

the most populated cluster if compared with the other solution, showing only a little reduction in the affinity (the EP1's cluster population in these cases is less than the EP1's cluster population in the WT). Again, the mutants M5 and M7 have only one key-region of EP1 changed in alanine but a major number of residues compared with M1, M2 and M8. In fact, the mutations have a stronger negative effect on the binding affinity, as showed by the fall of EP1's cluster position and population (becoming the third/fourth cluster in order of population).

We can observe a similar effect also for the mutant M3, that is characterized by one mutation in two of the EP1's key-regions (see Table 11).

Therefore, looking at the results of the TG2 mutants' simulations, it is apparent that the most strong effect on the TG2/clone 2.8 binding is performed by the mutants that have mutations in all the three key-regions of EP1. In fact, the mutant M6, M15, M16 and in particular M18, M19, M20 and M21 present a disappearance of the solution EP1 between the top 30 models.

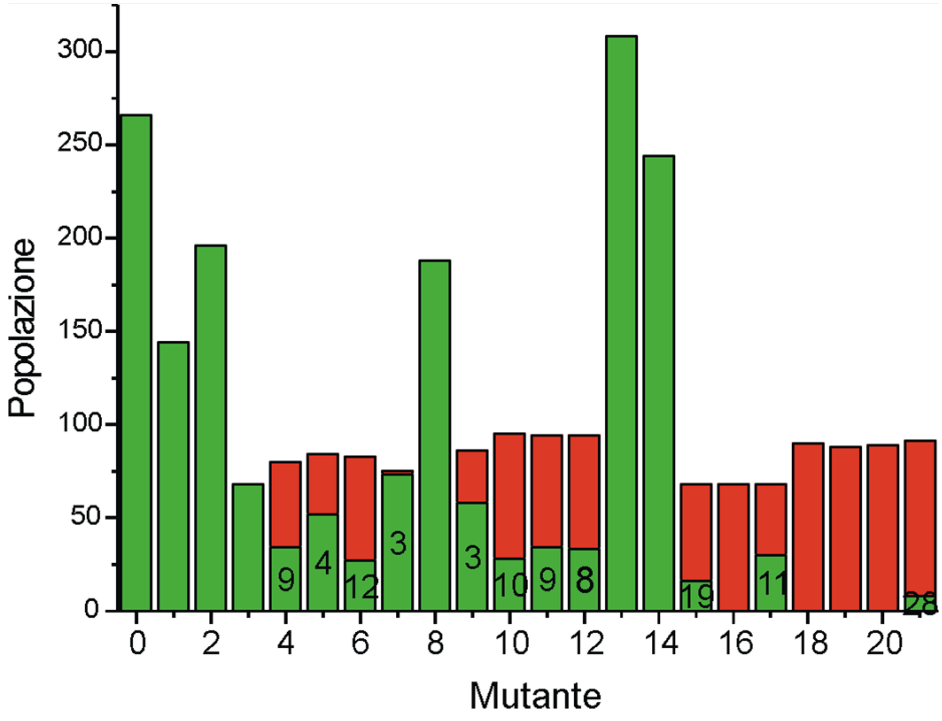


Figure 29

The histrogram's bars report the cluster population corresponding to EP1 epitope (in green) and the population of the most populated cluster (in red) if the solution does not correspond to EP1. In this case, the number reported on the green bar indicates the rank of the cluster presenting EP1 as solution. For example, in the case of M6 mutant, the most populated cluster has a population of 83 (as indicated by the high of the red bar), while the cluster presenting EP1 as binding mode has a population of 27 (as indicated by the high of the green bar) and it is th 12^{th} cluster ranked in order of

population. The WT is indicated as 0, and the mutants are labels with the numeration reported in Table 11.

On the basis of the results and the effect of the mutations, the idea is that TG2 binds the antibody through epitope EP1, characterized by three anchor regions: the region 200-230, the region around 345 and the one around the loop 365. Due to the topology of the binding site, made by three anchor sites, the mutation of only one site decrease the affinity between the two proteins but it is not enough to prevent the binding. Only the mutation of all the three site is able to compromise the binding, destroying the interacting network of the three anchor points (Figure 29).
The mutants that *in silico* showed the most promising performance were suggested to prof. Daniele Sblattero, Department of Medical Science, University of Piemonte Orientale (Italy), for experimental tests.

5.4 - Conclusion

In this chapter I described the docking study I performed to obtain a molecular model of the complex between the celiac autoantibodies (clone 2.8, clone 4.1 and clone 4.1) and the auto-antigen TG2. The present investigation provides better picture and gives useful insight into the orientation and characterization of the complex's binding site, showing that the interaction involve the TG2 epitope made of three anchor sites, all of them fundamental for the binding.
The improvement achieved in recent years by methods for predicting structures and protein-protein interaction give us the confidence in the results of our computational approach. Most importantly, the model is validated by its ability to explain the experimental data, its coherence resulting by different docking simulations (with clone 2.8, clone 4.1 and clone 3.7) and by the comparison with the simulations performed on 21 mutants of TG2.
Due to the crucial involvement of the complex TG2-antoantibody in the celiac disease, the diagnosis application and the promising therapeutic applications, the proposed model could help rationalizing the experiments as crucial step for the study of the celiac disease mechanism, the improvement of diagnosis strategies and the rational design of molecules for pharmacological and therapeutic purposes.

CHAPTER 6 - Prediction and analysis of an idiotype - anti-idiotype antibody complex associated to celiac disease

6.1 - Introduction

The idiotypic network

The antigenic determinants of antibodies is named "idiotype" and it is located in the variable region of the antibodies.[165,166] The 1984 Nobel laureate Jerne proposed an idiotypic network theory,[167] predicting that the idiotypic determinants of each antibody are recognized by those of another antibody, thus creating an "idiotypic network" through which immunoglobulins expression might be controlled. In fact, under physiological conditions, each antigenic stimulation (due to an antigen Ag) leads to the production of idiotype antibodies (termed Ab1) against Ag and then the unique structure of its antigen-binding site triggers the immune system to produce a series of anti-idiotypes directed against the Ab1's antigenic-determinant, termed Ab2 (Figure 30). Finally, anti- anti-idiotypes antibodies (Ab3) are induced by the presence of Ab2, which may have binding capabilities similar to those of Ab1, recognizing the original Ag.[168,169]

This idiotypic Ab1-Ab2-Ab3 network has a crucial role in the regulation of immune response to external and self antigens. For example, in a healthy subject and after eradication of the invading organism (the Ag), anti-idiotype antibodies Ab2 are useful to decrease the idiotype Ab1 titers to lower levels,[168] and their presence can maintain B cell memory during the absence of antigen in the system, helping in the maintenance of immunological memory.[170]

In some cases Ab2 can also act inhibiting the binding of Ab1 to the original antigen. On the basis of this "inhibiting" property, the Ab2 are classified as follows (Figure 30):

- *"Ab2-alpha"* (Ab2α are directed against idiotypes which are distinct from the antigen-binding site on Ab1. In this case, the idiotype/anti-idiotype interaction does not inhibit the Ab1-Ag binding.
- *"Ab2-beta"* (Ab2 β) binds exactly the antigen-binding site of the Ab1 antibody, inhibiting the Ab1-Ag binding. In fact, this class of anti-idiotypes interacts with Ab1 through structures that resemble the epitope of the antigen, carrying a so defined "internal image" of it.
- *"Ab2-gamma"* (Ab2γ) refers to antibodies directed against idiotypes close to, rather than within, the antigen-binding site. So, they can sterically inhibit the Ab1-Ag binding such as Ab2β, but they do not carry an internal image of the antigen. [165,168]

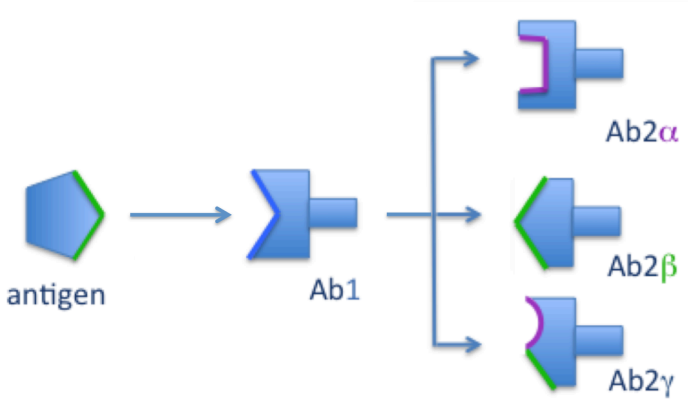


Figure 30. The idiotypic network
An antigen Ag is recognized by its antibody Ab1. The Ab1 becomes itself an antigen eliciting the production of anti-antibodies Ab2. This response can be divided into: i) an antigen-noninhibitable group (Ab2α), ii) an antigen-inhibitable group bringing an internal image of the Ag (Ab2β), and iii) an antigen-inhibitable group due to steric hindrance with the antigen binding-site (Ab2γ).

The theory of the "internal image" was then experimentally proved, showing that an anti-idiotypic antibody can provide an approximate topological and binding-group mimicry of an external antigen in different ways. Ban *et al.*[171] determined the crystal structure of an anti-idiotypic antibody 409.5.3 raised against the antibody that

neutralizes E2 peplomer, a large glycoprotein of feline infectious peritonitis virus FIPV (wwPDB code: 1IAI), describing also the mimicry of the Ab2 409.5.3 for the original antigen E2. The experimental data, in fact, showed that Ab2, when injected back into mice, elicited the production of Ab3s that had FIPV neutralizing properties. A comparison of the sequence of the Ab2's CDR loops with the antigen showed sequence homology in two regions of about six amino acids, both regions providing important contacts with Ab1. This evidence was also consistent with experimental tests, showing that Ab1 recognizes an epitope on the E2 peplomer even when the E2 protein is completely denaturated.[172] This suggests that Ab1 may be specific not strictly for a structurally unique epitope but for a sequence-unique epitope on the antigen.

However, other cases of Ab2 mimicry for the Ag was proved even though there is no sequence homology between the anti-idiotype and the antigen. Bradford *et al.*[173] determined the crystal structure of an anti-hen-egg-white lysozyme antibody (D1.3) complexed with an anti-idiotypic antibody (E5.2) (wwPDB[34] code: 1DVF) and they discussed the molecular mimicry of E5.2 for the original antigen HEL, showing that the mimicry did not depend on amino acid sequence homologies between the Ag and Ab2. In fact, they compared the structure of E5.2 and HEL both in complex with Ab1-D1.3. After superimposition of D1.3, they did not found a similar topology between the E5.2 and HEL, but they did find similar size of the solvated cavities, almost the same number of van der Waals contacts, the same patterns in the hydrophilic interaction, and six of the 14 interface hydrogen bonds in D1.3-E5.2 conserved also in D1.3-HEL complex. In particular, they found that much of the mimicry of E5.2 for HEL resides in the similar interaction made by the CDR loop H3 of the anti-idiotype and two particular residues of the Ag. Finally, also previous experimental data about CDR side-chain mutations of D1.3[174] are consistent with the structural/functional mimicry of E5.2 for HEL.

Even though internal image on anti-idiotypes are three-dimensional amino acid constructs, they can mimic also peptides epitopes. This is the case of the complex between an antibody and the angiotensin II peptide. The elicited Ab3 recognizes the original antigen with high affinity. Backbone atoms of this peptides closely resemble a CDR loop belongin to the canonical structure of the CDR loop L3.[175] This suggest

that the Ab2 has a CDR-L3 that resembles angiotensis II, mimicking the peptide with this single hypervariable loop.[176]

Applications of anti-idiotypic antibodies in medicine

Due to the key role of the immune system in the defence against diseases, the idiotypic network can be harnessed to develop new therapeutic strategies in many possible ways.

First of all, the anti-idiotype Ab2 offers an elegant concept for developing vaccines not based on the conventional approach of using nominal antigens. In particular, because of the Ab2β anti-idiotype brings the internal image of the Ag, it can induce specific immune responses similar to response induced by nominal Ag and it can be used to surrogate Ag. In fact, immunization with Ab2β can lead to the generation of anti - anti-Id antibodies Ab3 that recognize the corresponding original Ag identified by the Ab1. These so called anti-idiotype vaccines (anti-Id vaccines) have many advantages over conventional vaccines. For example, they contain neither nominal Ag nor its fragments. This exclude the possibility that anti-Id vaccines would have the same undesired effects which are sometimes associated with conventional antigen vaccines. There are also practical, economical and biological advantages.[177] In fact, anti-Id vaccines do not depend on the availability of large amounts of pure Ag, which often is a limiting economical factor in vaccine production and, by virtue of their being proteins, they can be easily manipulated.[178]

Anti-idiotypic antibody for cancer immunotherapy

Vaccination can become a decisive factor in situation where the responding immune system is immature or suppressed, such as in cancer patients, who may be immunodeficient or tolerant against their own tumour.

In particular, active immunotherapy is a really attractive therapeutic approach for tumours because it harnesses the body's immune potential to attack malignant cells in an antigen-specific manner and have immunological memory, but normally cancer patients are immunodeficient or tolerant against their own tumour. A common explanation for the absence of anti-tumour immunity is that the immune system has been tolerized by the tumour antigen. An effective method of breaking tolerance is to present the critical epitope in a different molecular environment to the tolerized host.

This is impossible to do with most tumour antigens because they are chemically elaborate and difficult to purify.[179] In this contest anti-Id Ab2β, mimicking the three-dimensional shapes of antigens, can present the antigen in a different molecular environment and it can be considered a powerful approach to generate targeted antigen immunity.[178]

This strategy has been used for the last decades[180]. In particular, anti-idiotype antibodies have been usefully used to implement active immunotherapy in patients with breast cancer,[181,182] colorectal carcinoma,[183] melanoma and ovarian lymphoma.[184,185] Nowadays, induction of adaptive tumour-antigen-specific immune responses continues to hold great promise for cancer prevention and therapy.

Ab2 anti-idiotypes are having great application in the development of new therapeutic approaches in cancer treatment, but they seem promising also in a lot of other important applications, such us the design of future anti-HIV strategies against AIDS, one of the deeply challenged in the case of persistent infections,[186,187] or as potent anticoagulant as an ideal antidote in restoring normal hemostasis.[188]

Although the Ab2b class of anti-idiotype has been receiving the most attention in the development of new vaccines, also the Ab2γ seem promising in the induction of an anti-viral response. In fact, Bryson *et al.* 2008[186] reported the crystal structure of the complex between the mouse antibody anti-HIV-1 Ab1/2F5 and its anti-idiotype Ab2/3H6 (wwPDB code: 3BQU), in which they showed that Ab2/2F5 does not resemble the structure of the original Ab1/2F5's antigen, but still interferes with the Ag-Ab1 binding[189], classifying the Ab2 as γ. Anyway, it is still uncertain if this information can concretely help the attempts at creating a vaccine targeting human immunodeficiency virus (HIV).

The role of the anti-idiotypic antibodies in autoimmune diseases

The idiotypic network has also a fundamental role in the autoimmune diseases. The regulation of the autoimmune response is still an intriguing and largely explored area. The factors leading to the onset of the autoimmune response remain obscure, but the idiotypic dysregulation is now recognized as a major mechanism for autoimmunity. In fact, in subjects susceptible to autoimmune diseases as a result of genetic predisposition and environmental factors, the immune response to a particular self

antigen involves an uncontrolled production of idiotypic antibodies (autoantibodies) that recognize pathogenic epitopes. Deficient idiotypic regulation of autoantibodies has been considered responsible for a number of autoimmune diseases[190] such as systemic lupus erythematosus (SLE)[191], autoimmune thyroiditis,[192] systemic vasculitis[193] and the Guillain-Barrè syndrome[194]. Furthermore, it has been demonstrated that autoimmune patients show a large ratio of autoantibody to anti-idiotype concentration whereas this ratio is small in healthy controls.[195]

Therefore, the use of anti-idiotype could be very promising in the study and the treatment of auto-immune disease. It is still very hard to find a definitive cure for this kind of diseases, due to the fact that the pathogen triggering the immune response is a self antigen, but *in vivo* study have indicated that anti-idiotypic antibodies might be able to downregulate the autoantibodies. In type 1 diabetes, for example, it was shown that anti-idiotypes may play a protective role in the immune response, by preventing that the auto-antibody binds its antigen.[196]

Another important application is creating animal models to study autoimmune diseases, by inducing them in animals through the usage of pathogenic idiotypes of autoantibodies. Following immunization with Ab1 and production of Ab2, the animals develop Ab3 having original autoantibodies properties and are associated with the respective serological and clinical manifestations of the disease.[168,197]

The celiac diseases

One of the most common disease with autoimmune features that suffers from a lack of animal models is celiac disease (CD). It is a disorder affecting approximately 1 in 100 individuals in the European population[150] occurring as a result of the interplay between genetic and environmental factors.[198] It is a long-life inflammatory condition characterized by flattering of the intestinal mucosa and malabsorption. The pathogenesis involves dietary exposure to gliadins, specific antigenic determinants found in gluten. The disease is characterized by presence of specific antibodies recognizing gliadins, food proteins and an endomysial autoantigen identified as tissue transglutaminase 2 (TG2).[154] The antibody level against gliadins and TG2 increase upon exposure to gluten, and decrease during the course of a gluten-free diet. Although considerable scientific progress has been made in understanding celiac disease and in preventing or curing its manifestations, a strict gluten-free diet is the only treatment for celiac disease to date.

Recently, celiac antibodies recognizing TG2 have also been shown to elicit the production of anti-idiotypes antibodies in mouse.[6] To the aim of developing an animal model in order to verify the possibility of expressing human anti-TG2 antibodies *in vivo*, Di Niro *et all.*[6] injected in mice non predisposed to gluten-intolerance two anti-TG2 antibodies isolated from patients with CD, both recognizing human TG2, but only one autoantibody (clone 2.8) is cross-reactive to mouse TG2. Since the other antibody (clone 3.7) does not recognize murine TG2, and so presumably had no *in vivo* immunological effect, it was used as a negative control.

What they found was a clear proof of an anti-idiotypic response in mice treated with clone2.8 (the only one able to recognize mouse TG2), in which the production of anti-TG2 antibodies was counterbalanced by the production of anti-Id antibodies. In particular, the results showed that all the anti-Id antibodies competed strongly with TG2 for clone 2.8 binding, indicating that both interact with the antigen-binding site.

To the aim of better understand the role and the characteristics of the interaction between the autoantibody anti-TG2 and its anti-idiotype, the sequence of the anti-idiotype antibody (AIT2) from mouse was characterized.

Aim of the work

Due to the growing importance of anti-idiotype antibodies in the development of new strategies for the treatment and the study of celiac disease (and diseases in general), we believe that the characterization at atomic level of the interaction between the autoantibody anti-TG2 clone 2.8 (Ab1) and its anti-idiotype (Ab2) could be of great interest. For this reason, we performed docking simulation between the autoantibody clone 2.8 and its anti-idiotype AIT2 isolated from mouse. We also compared the obtained solutions with available experimental Ab1-Ab2 structures.[171,186,188,199-201] Finally, we searched for local structural similarities between the Ab2 and the original Ab1's antigen TG2.

6.2 - Methods

Abs modeling

The variable domain structures of clone 2.8 and Ab2-mouse AIT2 were modeled by the RosettaAntibody Fv homology modeling server,[164] using the full refinement protocol option.

The PDB codes of the templates for Ab2 AIT2 are as follows: 1MH5 for the heavy-chain framework (97,01%) and 1AY1 for the light chain (96,77%); 1AY1 for L1 (100,00%), 1SEQ for L2 (100,00%), 1AY1 for L3 (77,78%); 1IQW for H1 (90,00%), 1IQW for L2 (100%) and 1A2Y for H3 (same length, no identity).

Docking

The obtained models were then used for Ab1/Ab2 protein-protein docking simulations, performed by the ClusPro 2.0 server.[56] By default, ClusPro server docks the receptor (Ab2) and the ligand (clone2.8) structures using DOT rigid-body docking algorithm. The top 20.000 complexes generated by DOT are then filtered according to electrostatic and desolvation energies, and then the top 2000 complexes are retained for further processing. The retained 2000 conformations are then clustered according to interface RMSD values and the top 10 docked models, following a short Charrm10 energy minimization, are made available for download. In all the simulations, all the Ab2's residues that do not fall into the Complementary Determining Region (CDR) were masked (ClusPro Antibody Mode). Differently, for clone2.8 two situations were explored. In the former, indicated as 'blind' docking, all the clone 2.8 residues were considered on an equal basis; in the latter, indicated as 'active' docking, all but CDR residues were marked, to have only CDR interacting.

Analysis

The representative structures of the ten best clusters for each simulation were analysed. To analyze the docking results we used the CONS-COCOMAPS[5] web tool, that uses the conservation of inter-residue contacts as an estimate of the similarity between different docking solution. Then, the visualization and comparison of the interface in the docking models and crystallographic complexes were performed with the COCOMAPS[4] web tool, through intermolecular contact maps. Finally, a local

structural similarity between the Ab2 and the original clone 2.8's antigen TG2 was perfomed by RASMOT 3D PRO[202] and ProBis 2012[203] web tools.

6.3 - Results and Discussion

We performed docking simulations to obtain the structure of the complex between the idiotype clone 2.8 and its anti-idiotype AIT2, both isolated from CD patients. The structures of both Ab1 and Ab2 were reliably predicted by homology modeling using Rosetta Antibody,[164] while protein-protein docking simulations were performed by ClusPro 2.0.[56]

In our ClusPro Ab1-Ab2 simulations, Ab2-AIT2 acts as the antibody (i.e the recognizing molecule), while Ab1-clone2.8 acts as the antigen (i.e. the recognized molecule). Therefore, as generally done for antibodies, Ab2-AIT2 was fixed and all its residues not falling into the CDR were masked. As for the Ab1-clone 2.8, in a first 'blind' docking approach, all residues were considered unmasked, and therefore available for the interaction. However, as experimental data showed that AIT2 strongly competes with the original TG2 antigen for the clone 2.8 binding, indicating that it also binds to the clone 2.8 CDR region, we additionally ran 'active' simulations, where the non-CDR regions of clone 2.8 were masked. For each docking approach, we then analyzed the representative structure of the ten best clusters, that should capture most of the important rigid-body binding geometries, providing good starting structures for further analyses.

'Blind docking'

The 'blind' docking simulations gave solutions with relative low population (86/2000 for the first cluster). However, although no constrains were applied, all the ten models, with the only exception of model 6, pointed to the CDR region of clone 2.8 (in particular to its light chain), (see Figure 28a and Figure 28c).

(a) 'BLIND' DOCKING

	Population	Score
Model 1	86	-300.4
Model 2	82	-250.3
Model 3	72	-244.0
Model 4	69	-250.3
Model 5	44	-269.1
Model 6	43	-257.9
Model 7	41	-245.7
Model 8	37	-241.9
Model 9	34	-249.9
Model 10	33	-243.4

(b) 'ACTIVE' DOCKING

	Population	Score
Model 1	152	-237.9
Model 2	118	-255.8
Model 3	67	-234.4
Model 4	65	-237.9
Model 5	58	-216.5
Model 6	38	-230.3
Model 7	38	-217.4
Model 8	32	-221.5
Model 9	32	-228.3
Model 10	32	-256.6

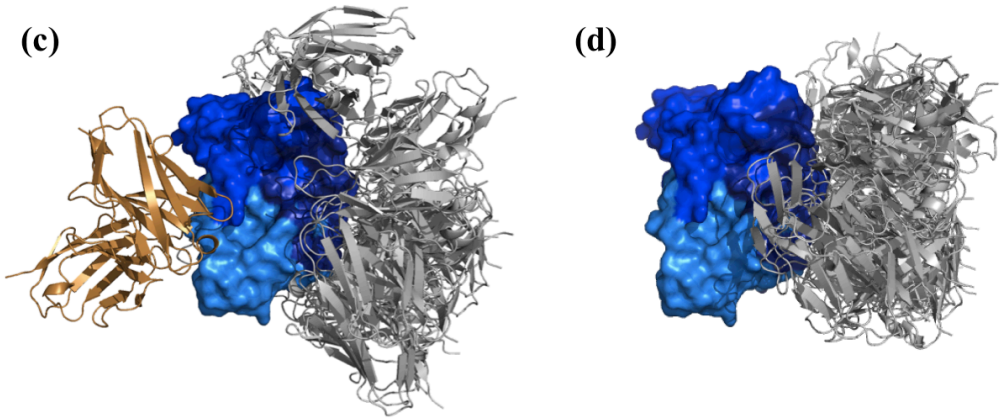
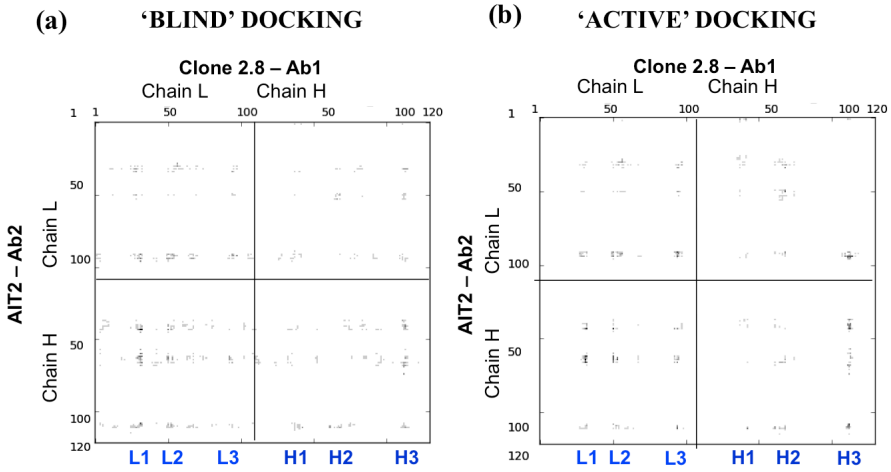


Figure 31. 'Blind' and 'active' docking clusters population and 3D visualization
(a) and (b): Table reporting the population of the top ten clusters, and the score of the representative model of each cluster; the values are reported for both 'blind' (a) and 'active' docking simulations (b). (c) and (d): Pymol[204] visualization of the representative models of the top ten clusters in both 'blind' (c) and the 'active' docking simulations (d), after superimposition of clone 2.8. The clone 2.8's light chain is colored in cyan, its heavy chain is colored in blue and its CDR loops are highlighted in dark blue. All the models involving the clone 2.8's CDR region for the interaction have AIT2 colored in silver, while the only one pointing in the other direction has it colored in copper.

Furthermore, running CONS-COCOMAPS[5] on the top ten solutions, a significant consensus was found in terms of intermolecular contacts. In particular, CONS-COCOMAPS[5] gives in output: i) a 'consensus' map, i.e. a 2D map were inter-molecular contacts are shown in a scale of grays were the more conserved the contact, the darker the spot, and ii) a list of the most conserved contacts.

The consensus map (Figure 32a) showed at a glance the similarity among these ten best docking solutions; in fact, the map was not spread but the dark spots converged in well defined and conserved regions, most of them located at the crossover of the CDR loops of both clone 2.8 and AIT2. Please note that involvement in the

interaction of the clone2.8 CDR loops was not obvious as a result of the 'blind' simulations, where all its residues were treated on an equal basis. This consensus was also quantified by the CONS-COCOMAPS' table that reports the list of the most conserved inter-residue contacts, in which it was evident that all the clone 2.8's CDR-loops were involved in the interaction in more that one model (reaching a maximum of seven models on ten having the L1 and L2 loops at the interface, and a minimum of three models on ten having the H1 at the interface, see Figure 32).



(c)

Clone 2.8 CDR-loops	N° models
CDR- Loop L1	7/10
CDR- Loop L2	7/10
CDR- Loop L3	5/10
CDR- Loop H1	3/10
CDR- Loop H2	4/10
CDR- Loop H3	8/10

(d)

Clone 2.8 CDR-loops	N° models
CDR- Loop L1	9/10
CDR- Loop L2	9/11
CDR- Loop L3	9/12
CDR- Loop H1	3/10
CDR- Loop H2	4/10
CDR- Loop H3	8/10

Figure 32. Consensus maps and consensus lists
(a) and (b): The CONS-COCOMAPS consensus map between clone 2.8 and AIT2 for the ten best 'blind' (a) and the 'active' (b) docking. Labels have been added for clone 2.8 CDR loops L1-L3 and H1-H3. (c) and (d): List of the number of models presenting the clone 2.8's CDR loop at the interface of interaction, in the case of the 'blind' and the active 'docking'.

So, overall the 'blind' simulation clearly show that the region recognized by the anti-idiotype antibody AIT2 involved the clone 2.8's CDR loops, in agreement with the experimental data.[6]

'Active' docking

We therefore ran the 'active' docking simulation, where the CDR residues of clone 2.8 were the only ones not masked in the docking. The results do not vary greatly. Comparing the 'blind' and 'active' consensus maps (see Figure 32a and b) it is apparent that the 'active' solutions represent a subset of the 'blind' ones. In fact, almost all the spots showed in the 'active' consensus map are included in the spots showed in the 'blind' consensus map, and darker in same cases. The first cluster coincides with the first most populated 'blind' solution and its population is about doubled (Figure 28a and Figure 28b).

Analyzing and characterizing the interaction interface for these docking models by the COCOMAPS[4] web tool, one preferred solution clearly emerges finding a significant consensus among the most populated 'blind' and 'active' solutions (and also the ones of lowest score), involving about 20% of all the solutions. In fact, by the comparison of the COCOMAPS[4] contact maps it is apparent at a glance the overlap of the epitopes (corresponding in an overlap of the spots in the contact maps), in particular between model 1 of the 'blind' docking and model 1 and model 2 of the 'active' one (cluster population of 86, 152 and 118, respectively) (Figure 33). Also the resemblance of the accessible surface area in the three complexes, the lists of residues at the interaction interface and of the intermolecular H-bonds given by COCOMAPS[4] confirmed the similarity between them (*data not shown*).

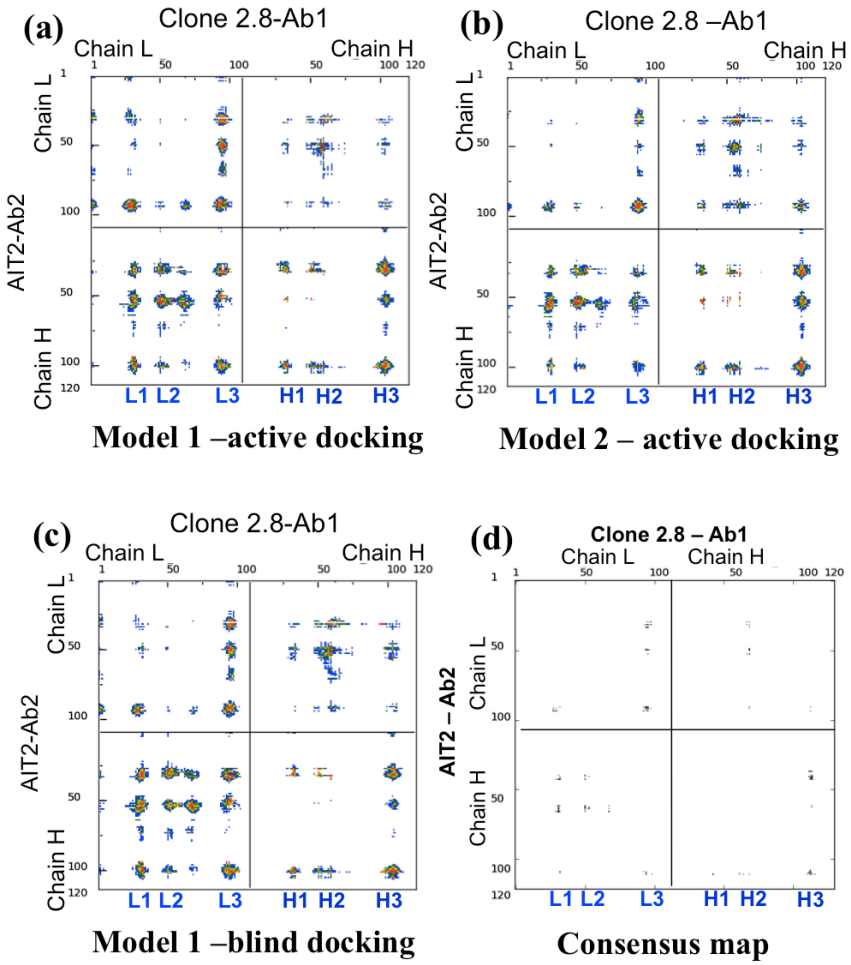


Figure 33. Comparison of the COCOMAPS's property contact maps and the CONS-COCOMAPS's consensus map.
(a), (b) and (c): the distance range contact maps by COCOMAPS[4], calculated for the model 1 of the 'blind' docking (c) and model 1 (a) and model 2 (b) of the 'active' one. Labels have been added for clone 2.8 CDR loops L1-L3 and H1-H3. The dots at the crossover of two residues are colored in red, yellow, green and blue if any pair of atom is closer than 7, 10, 13 and 16 Å, respectively. (d) Consensus map calculated on the blind simulation's model 1, active simulation's model 1 and model 2. Also, labels have been added for clone 2.8 CDR loops L1-L3 and H1-H3.

6.4 - Study of experimental cases from literature: comparison with other Ab1-Ab2 X-ray structures

We than decided to analyze the features of the complex interface in all the available experimental structures of Ab1-Ab2 complexes (in the wwPDB: 1CIC,[201] 1IAI,[171] 1DVF,[199] 1PG7,[188] 3BQU[186]; experimental structure of Ab1 YsT9.1 in complex with the Ab2 T91AJ5 determinated by Evans *et al.*,[200]) and to compared these with the AIT2-clone 2.8 'consensus' complex we selected upon the docking simulations. We

found that there is one recurrent binding solution in the experimental structures, that interestingly closely resembles the one we found for clone2.8 and AIT2. This increases our confidence in the proposed orientation of the molecules in the complex.

First, the RMSD between the experimental case and the proposed model is of only 3.2 Å. In Figure 34 the contact maps of both AIT2/clone 2.8 'consensus' model and the crystallographic structure between an anti-hen-egg-white lysozyme antibody (Ab1-D1.3) and an anti-idiotypic antibody (Ab2-E5.2) (wwPDB code: 1DVF)[199] are reported. Comparing the two contact maps, it is evident that the AIT2/clone 2.8 model provides similar binding interaction to the x-ray structure. In fact, both complexes seems stabilized preferentially by contacts between:

- Ab1's loops L1, L3 and H2 interacting with the light chain of Ab2;
- Ab1's loops L1, L2, L3 and H3 interacting with the heavy chain of Ab2.

The chemical-physical nature of the interaction residue involved in the interaction seems similar among the two complexes, showing in both cases a major involvement of hydrophilic residues. In fact, using the COCOMAPS default cutoff value to define the interaction residues (i.e. 8 Å), the clone 2.8/AIT2 model present 68,6% of hydrophilic/hydrophilic interaction, similar to the 60,9% of D1.3/E5.2 x-ray structure. On the contrary, the percentage of hydrophobic/hydrophobic interaction is of 2,2% for the model and 4,2% for the experimental structure we are considering, in both cases centered on the Ab1's H3 – Ab2's H3 interactions. Also the number of intermolecular H-bonds results the same in both cases.

Finally, we found similarity also in the values of the interface area, being of 816.4 $Å^2$ for clone 2.8-AIT2 complex and of 839.6 $Å^2$ for D1.3-E5.2 complex.

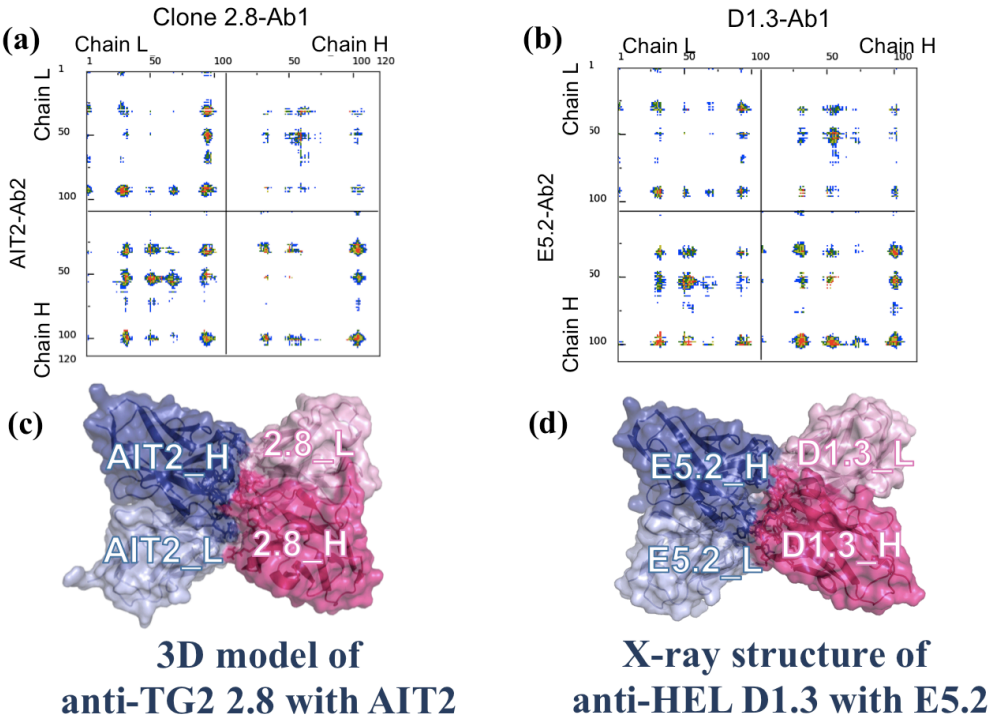


Figure 34. Comparison with x-ray
(a) and (b): the distance range contact maps by COCOMAPS[4], calculated for the clone 2.8/AIT2 model (a) and the experimental structure of E5.2/D1.3 complex (PDB code: 1DVF) (b). The dots at the crossover of two residues are colored in red, yellow, green and blue if any pair of atom is closer than 7, 10, 13 and 16 Å, respectively. (c) and (d): Pymol[204] visualization of clone 2.8/AIT2 model (c) and the experimental structure E5.2/D1.3 (d). The color code is the same in both figures: the Ab1 light and heavy chains are colored in light and dark blues, respectively; the Ab1 light and heavy chains are colored in light and dark pink, respectively. Labels have been added for the Ab1's and Ab2's light and heavy chains.

6.5 - Searching for structural similarities between Ab2 and Ag

The last step in the characterization of a Ab1-Ab2 complex is to identify possible structural similarities between the anti-idiotype Ab2 and the original antigen TG2. Unfortunately an experimental structure for the TG2-clone2.8 complex (the corresponding Ag-Ab1 complex) is missing, therefore we could just search for a possible local structural similarity between Ab2 and the original Ab1's antigen TG2. However the two web tools used to the aim, RASMOT 3D PRO[202] and ProBis 2012,[203] were unable to detect any significant similarity between them.

6.6 - Conclusion

Here we report a molecular model of the complex between the mouse anti-idiotype antibody Ab2-AIT2 elicited against the celiac autoantibody Ab1-clone 2.8. The present investigation provides better picture and gives useful insight into the orientation and characterization of the complex's binding site, showing that the interaction involves the Ab2-clone 2.8's binding site specific for the original antigen TG2, in according with the experimental data. Unfortunately, the experimental structure for the corresponding Ag-Ab1 is missing, and the only searching for structural similarity between Ab2 and original Ab1's antigen did not detect significant similarity. So, on the basis of this property, Ab2 can be classify the Ab2β or Ab2γ.

The improvement achieved in recent years by methods for predicting structures and protein-protein interaction give us the confidence in the results of our computational approach. Most importantly, the model is validated by its ability to explain the experimental data, its coherence resulting by different docking simulations and by the comparison with experimental structure complexes of the same typology, resembling the most recurrent binding mode of the experimental Ab1-Ab2 complexes.

Due to the crucial involvement of the idiotypic network in the autoimmune diseases and the promising therapeutic applications, the proposed model could help rationalizing the experiments as crucial step for the study of the celiac disease and the development of new possible therapeutic strategies.

CHAPTER 7 - Dynamic properties of a pathogenic mutant of the blood coagulation Factor X activated (FXa) and their effect on the substrate recognition and the catalytic efficiency

7.1 - Introduction

As described in Chapter 1, protein-protein interactions are intrinsic to every cellular process. Protein complexes underline for instance signaling, regulation, immunogenic recognition, as well as post-transcriptional events.[58] Apart from the antibody-antigen interactions described in Chapter 5 and 6, another fascinating biological complex typology is the one between an enzyme and its substrate. The enzymes are large proteins responsible for thousands of chemical interconversions occurring in the cells. An enzyme, in fact, acts as a highly selective biological catalyst, increasing the velocity and the rate of the reaction. Most enzymes act specifically with one reactant (called substrate) to produce products. On the basis of the function and the reaction that the enzymes catalyze, they are divided in families. A family of enzyme that is an interesting case in the field of the protein-protein interaction study is the serine protease one. These enzymes catalyze the cleavage of the peptide bonds in proteins and they are involved in a lot of fundamental processes, such as blood coagulations, digestion, immune response and reproduction. There are many experimental data available about the structure, the nature of the binding site (made of the catalytic triad His, Asp and Ser) and the function of the serine protease that helps to carry out structural study. In this scenario, my group and I focused the attention on a study about the recognizing properties of the serine protease factor X and its pathogenic mutant that causes problems in the blood coagulation cascade, taking advance of the molecular dynamics technique.

Factor X

The factor X (FX) is a vitamin K-dependent glycoprotein synthesized in the liver as a precursor molecule.[71,72] FX plays a pivotal role in the coagulation cascade being the point of convergence between intrinsic and extrinsic pathway of blood coagulation.

The FX activation results from the cleavage of the peptide bond Arg194-Ile195 (Arg15-Ile16 in the chymotrypsinogen numbering) that transforms the inactive zymogen of FX in a fully active enzyme (FXa). Upon its activation, FXa assembles into the prothrombinase complex to convert prothrombin (its substrate) to thrombin in the final stage of the blood coagulation cascade.[205,206]

FX circulates in plasma, at a concentration of 8-10 μg/mL, as a two chain protein: a light chain of 17 kDa linked with a disulphide bond to a 45 kDa heavy chain. FX shares extensive amino acid sequence identity to other vitamin-K-dependent serine proteases such as prothrombin, FVII, FIX, protein C and protein S.[207] In particular, the catalytic sites of the hemostatic proteinases share the same fold of the trypsin-like serine proteinases.[208] This allowed the use of the chymotrypsinogen numbering system for residues of the catalytic domains, which facilitate comparison of the various factors. The FXa catalytic domain is composed of two six- strand β–barrels and four short helices. The three serine protease catalytic residues His57, Asp102 and Ser195 (chymotrypsinogen numbering) are located at the crevice of the two β–barrels (Figure 35). The catalytic Ser195, together with the adjacent Gly193, forms the "oxyanion hole", helping to stabilize the tetrahedral intermediate during catalysis.

Inherited FX deficiency is a rare (1:1,000,000) coagulopathy with severe bleeding symptoms presenting early in life in homozygous patients.[209,210] About 105 causative mutations have been described in the FX gene so far, the majority of which are missense. The study of naturally occurring mutants in FX offers considerable insight into the structure and function of FX molecule. However, only few FX mutants have been expressed and characterized so far, and among them, only four located in the catalytic domain, were analyzed, namely Val342Ala, Arg347His, Gly366Ser and Gly381Asp.[8,211-213] Expression studies showed that all recombinant proteins were normally synthesized and secreted, but further functional characterization revealed that all of them had reduced coagulant activity, even those having a rate of activation similar to the wild type protein. All these studies confirmed the existence of a strict correlation between the localization of the mutation and protein function.[214]

In the present study, my group and I focused our attention on the molecular characterization of a recurrent p.Gly262Asp mutation (Gly43Asp in the chymotrypsinogen nomenclature; to better localize the amino acid residues in the

molecular modeling of the mutant FX, the chymotrypsin numbering system was used throughout the text) in the catalytic domain of mature FX protein, performing molecular dynamic simulations. We performed the present investigation to the aim of provide an explanation about the influence of the mutation on the structure and function of the protein ant its consequence on the interaction with the substrate. The study was conducted in collaboration with the experimental groups of the Prof. De Cristofaro (Hemostasis Research Centre, Institute of Internal Medicine and Geriatrics, Catholic University School of Medicine, Rome, Italy) and Prof. Peyvandi (Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Università degli Studi di Milano and Luigi Villa Foundation, Milan, Italy), that performed in vitro expression analyses and steady state kinetic studies of the wild type FXa and of its Gly43Asp mutant. The current name of the mutation refers to initiating methionine numbered as +1, but it was originally reported as Gly222Asp.[215] Gly43 is a buried residue located in the β3-strand (residues 40-46) of the N-terminal β-barrel, under the oxyanion hole formed by the Ser195 and Gly193 residues. It also gives an H-bond with the Ser195 backbone. Its spatial position relative to the catalytic triad is pretty fixed, as a disulphide bridge connects the immediately upstream Cys42 to Cys58 that follows the catalytic His57 (Figure 35).

Therefore, to investigate the effect of the naturally occurring Gly43Asp mutation on the FXa structure and dynamics, in vitro expression analyses, steady state kinetic studies and molecular dynamics (MD) simulations of the wild type FXa and of its Gly43Asp mutant were performed. This is the first report of the cellular fate characterization of a mutation located in the so-called loop-40 of the FXa, a conserved region of serine proteases including the catalytic His57.[216]

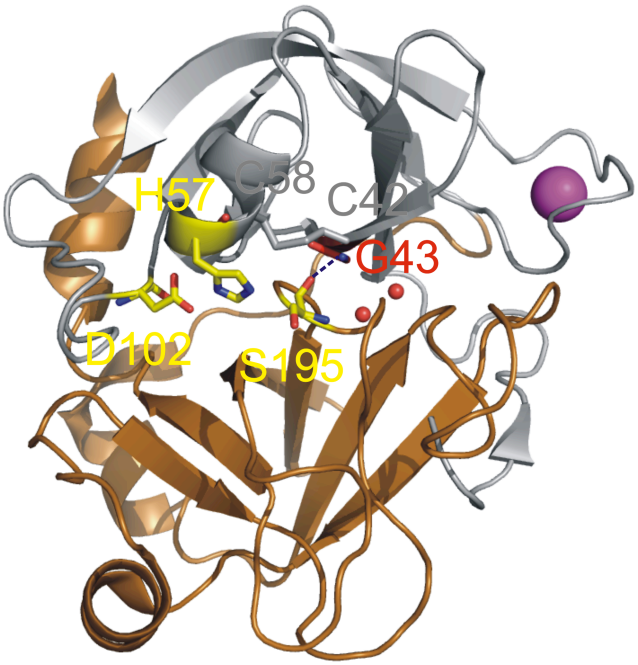


Figure 35

Cartoon representation of the crystallographic structure of the catalytic domain of human FXa (PDB code: 2BOH). In this orientation, the N-terminal β-barrel is up (colored gray) and the C-terminal β-barrel is down (colored copper). Gly43 and the catalytic residues are shown as red and yellow sticks, respectively. The bound calcium ion and oxygens of the two crystallographic waters close to Gly43 are shown as large magenta and small red spheres, respectively. The H-bond between Ser195(O) and Gly43(N) is also shown and a stick representation of the Cys42-Cys58 disulfide bridge is given.

7.2 - Methods

Molecular dynamics simulations and electrostatic potential calculations

For the MD simulations of FXa, the 2.2 Å resolution crystal structure corresponding to the PDB code 2BOH[217] was selected. For the MD simulations of the heavy chain of FXa, the 2.2 Å resolution crystal structure corresponding to the PDB code 2BOH was selected.[218] Since missing crystallographic waters may lead to artifacts in the FXa dynamics,[219] we selected a starting structure with a large set of well- determined water molecules,[217] which were included in the simulations. Interestingly, in the X-ray structure two buried water molecules are located close to the Gly43 residue. The starting structure for the FXa43Asp mutant was generated by the Mutagenesis PyMol module.[96] Insertion of the quite bulky aspartate side chain did not cause any dramatic

clash with the rest of the molecule, and the oxygens of the carboxylate group replaced the two aforementioned crystallographic waters.

All the MD simulations were performed using GROMACS ver. 4.5.4 [220] that adopts the AMBER99SB[221] force-field for energy minimization and molecular dynamics simulations. Both the WT and MT structures were solvated in a periodic cubic box with about 12500 TIP3P[222] water molecules, and at least 10 Å between the protein and the box sides. Electroneutrality was achieved by random replacement of water molecules with enough counter ions. The Particle-Mesh Ewald algorithm was applied to treat electrostatic interactions. The systems were first energy minimized, then a short 100 ps NVT MD simulation at 300 K was run to equilibrate them. These structures were used as the reference in the analysis of the MD trajectories. For better sampling, four different 60 ns long NPT MD simulations for each protein were performed assigning different initial velocities. A Berendsen thermostat with a time constant of 0.1 ps was used to control temperature of protein and of solvent. Pressure was controlled with a Parrinello-Rahman barostat with a time constant of 2 ps. The time step of the simulations was set to 2 fs, coordinates were saved every 10 ps. Analysis was performed on the last 10 ns.
Structural properties, such as root mean-square deviation (RMSD), root-mean square fluctuation (RMSF) and hydrogen bond interactions, were calculated with the built-in functions of GROMACS. Essential dynamics analysis was based on the diagonalization of the covariance matrix of the protein alpha-carbon atomic fluctuations. To calculate the H-bonds occupancy, the cut-offs on the Donor-Acceptor heavy atoms distance and on the Hydrogen-Donor-Acceptor angle were set at 3.5 Å and 30°, respectively. A representative structure for each system was also extracted, to be used for visualization and electrostatic potential calculations, by selecting the nearest frame to the average coordinates during the last 10-ns. Electrostatic potentials were calculated by solving the Poisson-Boltzmann equation using the APBS program [223] and visualized with PyMOL [224]. Calculations were carried out on a grid spacing between 0.31 and 0.37 Å, with a temperature of 298.15 K. The dielectric constrants were set to 4 for the protein and to 78 for the solvent. The solvent probe radius used was 1.4 Å.

7.3 - Results

To investigate the effect of the naturally occurring Gly43Asp mutation on the FXa structure and dynamics, we performed molecular dynamics (MD) simulations of the wild type FXa (WT) and of its Gly43Asp mutant (MT). We selected a starting structure with a large set of well-determined water molecules [217], which were included in the simulations. Interestingly, in the X-ray structure two buried water molecules are located close to the Gly43 residue. When we modeled the Gly43Asp mutant, the insertion of the quite bulky aspartate side chain did not cause any dramatic crash with the rest of the molecule, and the oxygens of the carboxylate group replaced the two above crystallographic waters.

RMSD and RMSF analysis

Four 60-ns-long MD simulations have been performed both for the WT and MT FXa with different initial velocities, to ensure better sampling. The overall stability of the proteins throughout the simulations were monitored through the RMSD of Ca atoms from the appropriate starting structures for each of the eight simulations. The systems remain stable during the 60-ns both in WT and MT (see Figure 36), and reach the equilibrated structures after 30-ns of simulation time based on the pleateuing of the RMSD curves, both for the WT and MT simulations.

To investigate how the mutation affects the overall flexibility of the protein, we also calculated the RMSF of the C-alpha atoms during the last 10-ns of the simulations (Figure 36). The Cα-RMSF values reveal the same trend among the four different simulations for each system. The trend and overlap of the RMSD curves and also of the RMSF profiles are clear indicators of the similarity between the spaces sampled by the four simulations for both the systems. Due to convergence of these simulations, in the following the results of one simulation per system will be discussed in detail.

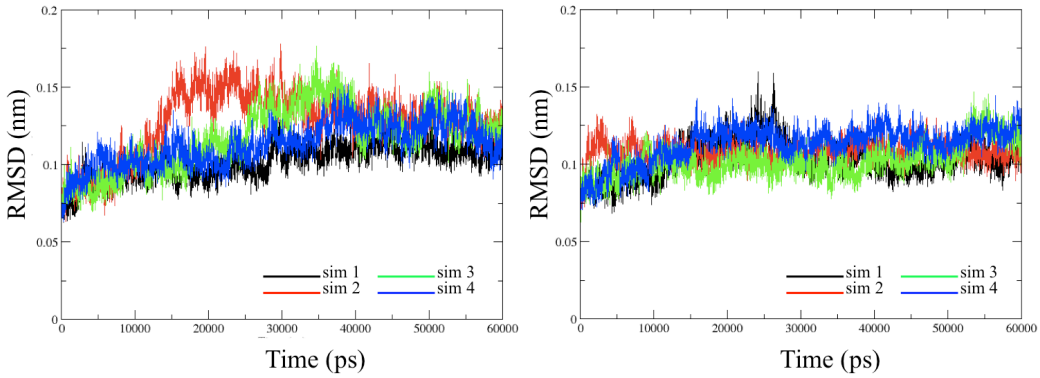


Figure 36.
The time dependence of RMSDs for the Cα of WT (on the left) and MT FXa (on the right) in the 60-ns MD simulations. The four simulations are shown in different colors in both the WT and MT systems.

From the RMSD values and the RMSF trends (Figure 37c), it is apparent that the Gly43Asp mutation does not dramatically affect the FXa structure, neither globally nor locally. This is also confirmed by the conservation of the secondary structure in the two systems, monitored during the last 10-ns of MD simulation (data not shown). It is worth noting that the fluctuation of residue 43, both in WT and MT is particularly low (see Figure 37c). As already said, this is related to the presence of a disulphide bridge connecting Cys42 and Cys58, both in the WT and MT FXa.

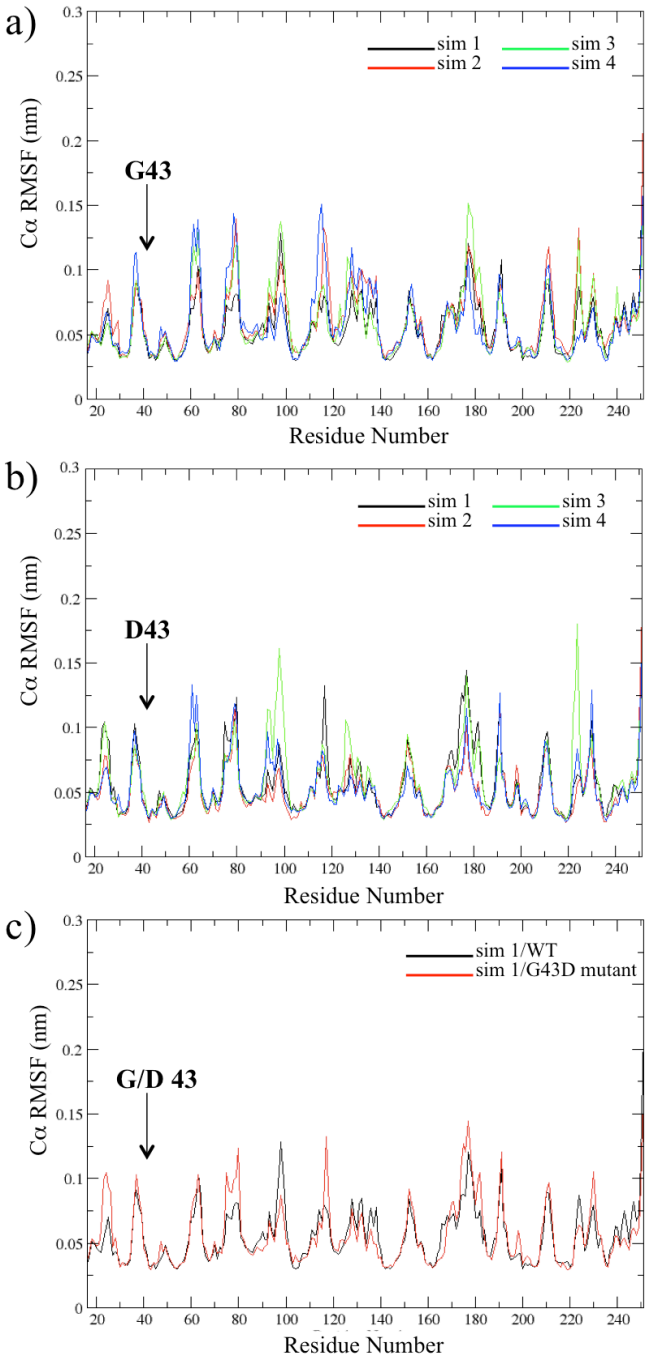


Figure 37
The calculated RMSF of Cα atoms vs protein residue number during the last 10-ns of simulation for : **a)** WT and **b)** MT FXa. The four simulations are shown in different colors in both the WT and MT systems. **c)** A comparison between the RMSF plot for simulation 1 of WT (in black) and MT FXa (in red).

Catalytic hydrogen bonds

A notable difference between the two systems is however observed, when looking at the protein H-bonds. Interestingly, on average four more H-bonds are found in the MT as compared to the WT (total number 167,4 *vs.* 163,6). To investigate whether this involve the catalytic site, all the H-bonds involving the catalytic triad and/or the Gly/Asp43 residue were monitored during the last 10-ns of the MD simulations. In Table 12, the occupancy of these "catalytic H-bonds" is reported for WT and MT; in Figure 38a view of the "catalytic H-bond network" is also given for the two systems.

Acceptor	**Donor**	**WT**	**G43D mutant**
43D/G(O)	T54(OG1)	97,9	0,0
43D(OD1)	Q30(NE2)	-	100,0
43D/G(OD2)	F141(N)	-	98,6
S195(O)	43D/G(N)	98,7	66,2
D102(OD1/OD2)	H57(ND1)	86,5	95,7
D102(OD1/OD2)	H57(N)	99,0	96,9
H57(O)	Y60(N)	73,2	70,3
H57(NE)	S195(OG)	0,3	11,8
D102(OG)	T229(OG)	97,2	1,7
D102(OD1)	S214(OG)	99,4	99,7
D102(O)	A56(N)	2,3	81,1
S195(OG)	G193(N)	12,8	19,3
S214(O)	S195(OG)	0,8	29,2
I227(O)	S214(N)	81,1	78,9
A104(O)	T54(OG1)	0,0	69,3

Table 12
The occupancy percentage (%) of WT and MT FXa hydrogen bond interactions involving mutated and/or catalytic residues, during the last 10-ns of simulation.

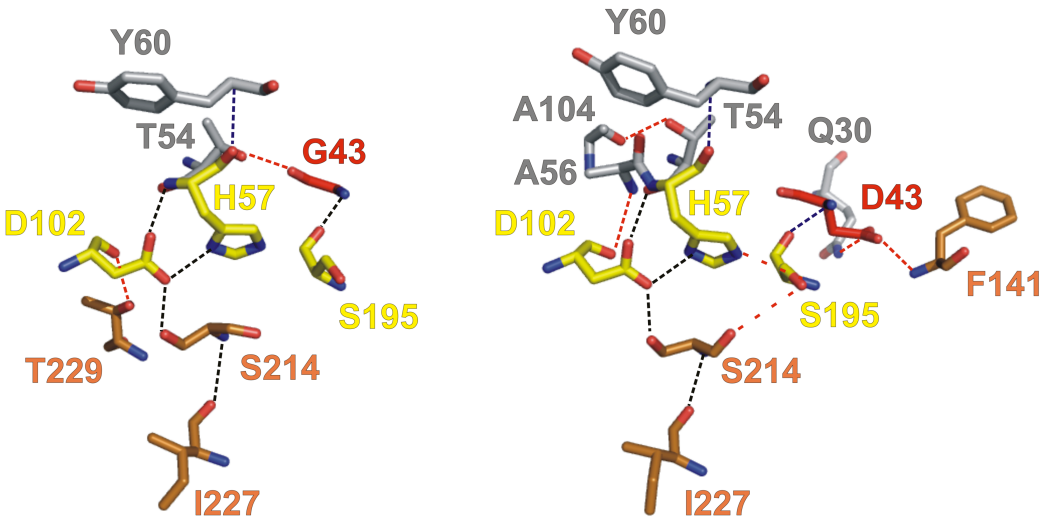


Figure 38
The H-bonds network around the catalytic site, as calculated during the last 10-ns of simulation for WT (on the left) and MT FXa (on the right). H-bonds which are remarkably different between the two systems are shown in red, those conserved are instead shown in blue. The mutated residue (Gly/Asp43) is colored red, the catalytic triad is colored yellow. Remaining residues are colored gray if belonging to the N-terminal and copper if belonging to the C-terminal β-barrel. Note that for the sake of simplicity, the Gly193(N)-Ser195(O) H-bond, whose occupancy is comparable between the two systems, is not reported.

Both from Table 12 and Figure 38, it can be easily seen that the mutation modifies the H-bonds network around the catalytic residues, by thickening it in the MT, as compared to the WT. In particular, in the WT the Gly43 backbone acts as acceptor in a high-occupancy H-bond with the side-chain of Thr54. Such H-bond is lost in the MT, however it is here compensated by a H-bond between the Thr54 side chain and the Ala104 backbone (see Figure 38). Importantly, MT Asp43 uses its additional H-bond acceptors (the oxygens of the carboxylate group) to give two novel H-bonds, as compared to the WT, with the side chain of Gln30 and with the backbone amide proton of Phe141 (from the opposite C-terminal β-barrel), respectively. It is interesting that one of these H-bonds substitutes a WT H-bond involving the Phe141 backbone and one of the above crystallographic waters.

As a possible conseguence, the backbone-backbone H-bond between residue 43 and the catalytic Ser195, which is very stable in the WT (occupancy 98.7%), becomes weaker in the MT (occupancy 66.2%). Other two ”catalytic H-bonds” are also

significantly affected by the mutation: one involves the side chains of His57 and Ser195, while the other involves the Ser195 side chain together with the Ser224 backbone. Both these hydrogen bonds are observed in a significant fraction of MT structures (occupancy 11.8 and 29.2%, respectively), whereas they are absent in the WT (occupancy below 1%). Finally, a backbone-backbone H-bond between the catalytic Asp102 and Ala56 is observed in the MT, instead of the side chain-side chain H-bond between Asp102 and Thr229 of the WT.

Essential dynamics

It is commonly accepted that essential degrees of freedom (or correlated motions) of a protein describe motions relevant for its function,[31-33] with the first several eigenvectors normally representing most of the correlated motions.[31] Therefore, essential dynamics analysis was carried out on the C-alpha atoms of wild type and mutant FXa. For both systems, the last 10 ns trajectories were projected onto their eigenvectors and the RMSF curve of the C-alpha atoms of the protein along the first eigenvector was plotted (Figure 39a). The regions show similar fluctuation for wild type and mutant, except in same peripheral parts of the proteins and in a region involved in the binding of the substrate. Specifically, the loop containing residues 94-97, at the border between the S2 and S4 binding pockets, shows a larger correlated motion in the wild type than in the mutant (Figure 39b)

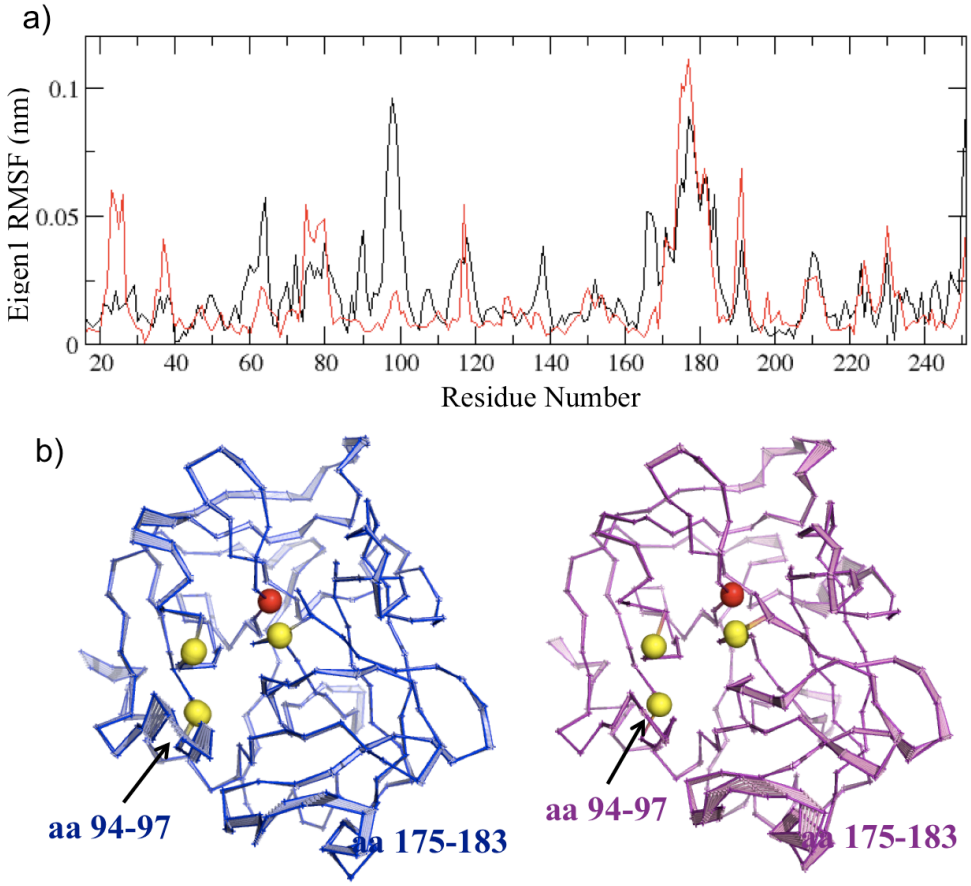


Figure 39
a) Comparison of the RMSF curves for the Cα atoms of WT (black line) and MT FXa (red line) along the first eigenvector. b) Cα ribbon representation of the ensemble structures representing the movement of the first eigenvector for WT (colored blue) and MT. (colored magenta). The Cα atoms of Gly/Asp43 and of the catalytic triad are shown as red and yellow balls, respectively. Binding regions showing a different behavior in terms of correlated motions are also indicated.

Electrostatic potentials

To investigate the possible effect of the negative charge introduced by Asp43 in the MT, in proximity of the catalytic site, we also calculated continuum Poisson-Boltzmann electrostatic potentials for the two systems.

In Figure 40, representative structures are shown for WT and MT FXa, colored according to the calculated electrostatic potentials. The structures are oriented with the N-terminal and C-terminal β-barrels up and down, respectively. The crevices hosting the substrates are in the middle and cross the proteins horizontally. It is apparent that the potential along the crevice is predominantly negative for both the systems. However, in the MT the negative character of the electrostatic potential

around the active site is accentuated and, importantly, in correspondence of the oxyanion hole it is reversed from the positive values of the WT to negative ones.

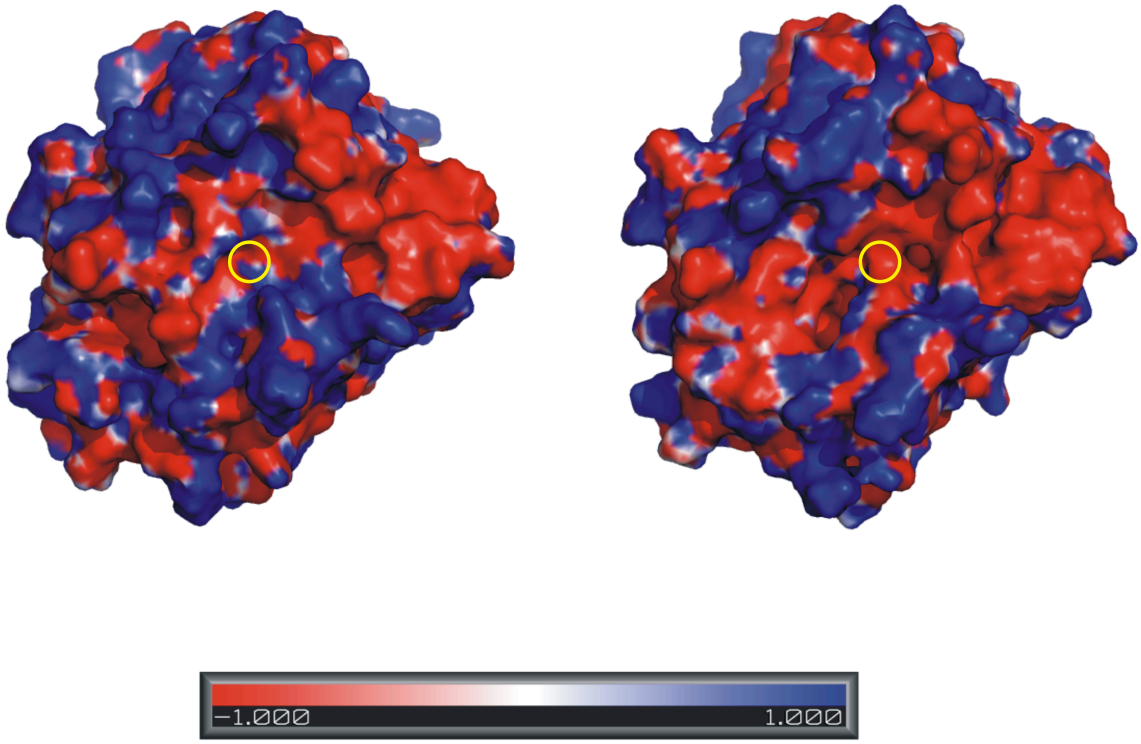


Figure 40
Representative structures are shown for WT and MT FXa, with the Van der Waals surface colored according to the calculated electrostatic potentials. The crevice hosting the substrate crosses the proteins horizontally. The color scale (in kT/e) is also reported. A yellow circle is used to indicate the position of the oxyanion hole.

7.4 - Discussion

Several natural FX variants due to missense mutations were previously reported in the literature.[206,207] These FX mutants, through different molecular mechanisms, are responsible for mild to severe reduction of FX procoagulant activity but only a few naturally occurring variants have been characterized so far. Hence, functional consequences of the majority of them remain largely unexplained. In addition, among naturally occurring FX mutations, only those compatible with normal or reduced biosynthesis and secretion are appropriate tools to investigate the molecular mechanisms underlying the disease, the structure-function relationships and bleeding tendency.

Previous *in vitro* expression studies were performed only on 10 mutations (out of the

82 missense mutations hitherto identified); each of these studies elucidated the effect of the mutation, providing information on the involved molecular mechanisms. In particular, those located in the catalytic domain, Val342Ala [Val160], Gly366Ser [Gly183] and Gly381Asp [Gly197] mutations, led to normally synthesized proteins with variable reduction of the activity due to the disruption of the native conformational structure of the catalytic domain with rearrangements of the molecule,[8,211,213] except for Arg347His [Arg165] that provokes the attenuation of FVa binding.[212] All these mutations but Val342Ala [Val160] are predictive of a clinically severe FX deficiency characterized by absence of FX procoagulant activity.

This work provides information about one of these naturally-occurring recurrent FX variants, the Gly43Asp mutation, that was firstly identified in two patients from Iran,17 and subsequently in 13 other patients from Turkey (unpublished data). All of them had a severe clinical phenotype, with such symptoms as hematomas and hemarthroses, hematuria and epistaxis, associated to a coagulant activity <1%.[209]

The Gly43 is a buried residue at the core of a completely inaccessible β-strand in the mature FX protein. The Gly43Asp substitution occurs at the highly conserved 42-58 residues region (loop-40) shared among all trypsinogen-like proteins (Table 13). In addition, Krawczak et al., through a homology modeling study of the catalytic domain of the ancestors of the present-day serine proteinases, showed that sequence and structure of this region is maintained despite the processes of duplication and divergence that the genes coding for coagulation proteins have undergone during their evolution.[225] These findings emphasize the importance of the loop-40 in the context of the functional architecture of blood coagulation serine proteinase.

Protein	**Aminoacids sequences**
Thrombin	**CG**ASLISDRWVLT**AAHC**
FXa	**CG**GTILSEFYILT**AAHC**
FIXa	**CG**GSIVNEKWIVT**AAHC**
FVIIa	**CG**GTLINTIWVVS**AAHC**
Protein C	**CG**AVLIHPSWVLT**AAHC**

Table 13
Aminoacids sequences from residue 42 to 58 (chymotrypsin numbering) of thrombin, FXa, FIXa and FVIIa and anticoagulant protein C.

Mutations occurring at this region have been shown to be causative of severe type I prothrombin, FVII, FIX or protein C deficiency.[226-229] The mechanism responsible for the severe FX deficiency caused by the Gly43Asp substitution was previously investigated showing a partial defect in secreting FX43Asp. Then, the experimental part of the present study on FX43Asp confirmed a secretion defect due to an alteration in the secretion efficiency of the mutant recombinant protein.

It is clear that in the process of secreted proteins is orchestrated by a group of molecules with a quality control function. The proteins involved in the folding system are lectins such as calreticulin, calnexin and Erp57.[230] These molecules (chaperons) facilitate protein folding[231] and ensure that correctly folded, assembled and modified proteins are transported along the secretory pathway. Therefore, the partial defect in secreting FX43Asp probably occurs due to the introduction of a charged hydrophilic Asp residue instead of a neutral Gly into the protein core, altering the recognition site involved in the intracellular trafficking.

In order to understand why the reduced amount of secreted mutant protein did not conserve its procoagulant activity, additional kinetic studies were performed in this study. In fact, the Gly43Asp substitution seemed to affect the amidolityc activity of the FX protein. These findings suggested that the Gly43Asp substitution does not totally disrupt the architecture and thus the catalytic function of the enzyme [210] but causes a more discrete change of the active site conformation. The Gly43 residue is sunk in the core of the protein, and is far from binding sites known to directly interact with substrates and/or cofactors to facilitate the specific assembly of the coagulation activation complexes.[232,233] Gly43 is also located at the N-terminus of the region referred to as loop-40, an exosite strongly conserved in the family of serine proteases.

Usually, a single point mutation affecting the activity of an enzyme is expected to compromise its native fold. However, our MD simulations clearly indicate that the Gly43Asp mutation neither disrupts nor destabilizes the FXa native structure. Rather, it makes it someway more rigid, by thickening the H- bonding network around its catalytic site and by affecting the correlated motions involving the substrate binding site. This should not be surprising, considering that the mutant shows a residual catalytic activity that would be incompatible with a completely misfolded protein. Moreover, it is now definitely established that enzymes have evolved under synergistic pressure between structure and dynamics and that their motions underlie catalysis.[234] Experimental evidences for the correlation between the dynamic

flexibility of the active site and its catalytic activity have also been specifically collected for α-chymostrypsin, a prototype serine protease.[235] Therefore, compromising the FXa enzyme flexibility, the mutation may also compromise its catalytic efficiency. The increased rigidity of the mutant FX might also affect the molecular recognition by chaperone proteins inside the cell, thus causing the retention and defective secretion. Further, we have shown that introducing a negative charge (Asp43) spatially close to the FXa catalytic Ser195, results in a negative electrostatic potential around the oxyanion hole, where the negative charge of the tetrahedral intermediate needs to be accommodated and stabilized. This is also expected to dramatically affect the enzyme catalytic efficiency.

7.5 - Conclusion

In conclusion, this work was focused on the study of the structural and functional aspects of a severe FX deficiency due to the frequent Gly43Asp mutation occurring at a highly conserved region (residues 42-58) shared among all trypsinogen-like proteins. This region of the FX protein has never been studied before and the replacement of the Gly43 by an Asp is like to cause a stiffening of the protein due to an altered distribution of H-bonds network as well as to a change of the electrostatic potential around the active site. These structural changes, although not dramatic, lead to the impairment of protein secretion and to a drastic reduction of its coagulant activity, proved by kinetic studies performed by the laboratories of the Prof. De Cristofaro and Prof. Peyvandi.

This study can help the emergence of new therapeutic products for the treatment of coagulation deficiencies.

APPENDIX 1 - Differences between membrane and soluble protein loop structures

During the third year of my, I spent seven months in the group of Prof. Charlotte Deane, Department of Statistics, University of Oxford (UK). In this period I studied the geometrical features of the proteins' regions most recurrent in the protein-protein interaction, the loops, clarifying some structural aspects of them in one of the most important and huge class of proteins: the membrane proteins. Here below there is the description of this work.

Introduction

Membrane proteins (MPs) represent about one third of all known proteins. They regulate the transport of molecules and information into and out of every living cell. Due to their involvement in many medically relevant processes, they comprise over half of current drug targets.[236]

Unlike globular soluble proteins (SPs), whose natural environment is an aqueous solution (such as the cytoplasm), MPs sit inside a lipid bilayer. Thus, a large proportion of a MP's amino acids are in direct contact with the hydrophobic fatty acid tails of the membrane lipids. The presence of the membrane around the protein creates a very different physicochemical environment that has direct effects upon a MP's three-dimensional (3D) structure. Transmembrane (TM) segments are usually one of two structure types: α helices or β strands. These TM segments are connected to each other by stretches of amino acids with irregular structure, known as loops. Especially in helical TM proteins, the geometry of secondary structure elements is often well conserved, with approximately parallel helices being oriented perpendicular to the membrane plane (parallel to the membrane normal) and spanning the entire width of the membrane. The structure of the loop regions connecting the TM segments can vary greatly between homologues.[237] Therefore, loops tend to be the parts of MPs that are the hardest to model.

In MPs loops can interact with the polar head groups of the membrane lipids as well as with water molecules and thus tend to contain many hydrophilic and charged residues. Positively charged amino acids such as Lys and Arg are especially common in loops protruding into the cytosol (the positive inside rule).[238,239] In addition to their

chemical properties, MP loops can also be expected to have characteristic shapes. The typical MP loop connects two roughly parallel TM segments and protrudes from the membrane into a polar environment.
Due to the physical crowding of the membrane, the loop tends not to interact with other parts of the protein, except other loops, but might be found touching the polar head groups of the membrane lipids. Loops in SPs, on the other hand, can interact with sequentially distant residues and often lie on the surface of the protein rather than protruding from it.

Due to the biological and medical importance of membrane proteins, they have become a major focus in structure prediction. Nevertheless, there is a lack of fast and reliable methods that specialize in modeling of membrane protein loops. Often methods designed for soluble protein are directly applied to membrane proteins, but obviously the difference between the structures of membrane and soluble protein loops influences their accuracy.
The group of Prof. Deane has showed an evidence of this difference, and how this can influence the performance of a structure prediction. In fact, using FREAD[240] program for loop modeling they have found that it is possible to predict accurately the structure of membrane protein loops using database of membrane protein fragments, (achieving an accuracy of 0.5-1 Å median RMSD), rather than using fragments of soluble proteins (achieving on accuracy of only 1-4 Å median RMSD). In fact, they found many fragments of soluble proteins with similar shapes to their membrane protein counterparts but with a very different sequence.
The aim of my work in the Prof. Deane's group was of exploring the reasons for the membrane and soluble protein loops difference by analyzing statistical and geometrical properties of both classes of loops. I have identified two features of loop structures that appear to differ between membrane and soluble proteins: the angle between the loop and flanking helices, as well as the contacts between residues and the remainder of the protein.

Methods

To understand how the conformations of membrane and water-soluble loops differ, I performed a series of tests on two sets of loops.

Test set

This study uses two sets of X-ray structures: one containing only water-soluble proteins (SPs), another containing only membrane proteins (MPs). An initial list of potential MPs was culled from PDB_TM[241]. An initial list of potential SPs was created using PISCES server[242] under the below criteria:

- Only X-ray crystallographic determined structure
- Resolution ≤ 3Å
- R-factor ≤ 0.3
- Each chain sharing less than 99% in sequence identity

For both sets, residues annotated by JOY[243] as being anything but helices and sheets were treated as loop residues. For the MPs, only loops within the membrane, or close to it, were considered. Loops close to the membrane were defined as those residues less than 40Å from the central plane of the membrane.[244] For each loop length, loops were clustered by sequence identity and made non-redundant at the 40% identity level. Lengths range from 3 to 15 residues. Only loops connecting two helices were considered.

Loop angle θ

I calculated the loop angle θ, which we define as the angle between the "loop" plane and the "helix" plane. First, the centres of mass (average co-ordinates) of the loop and the helices were calculated (the points labelled as "A" and "B" in Figure 41, respectively). To calculate the centre of mass of the "loop" A, all the Cα atoms of the loop were used, while to calculate the centre of mass of the "helices" B the Cα atoms of the six residues before and after the loop in the sequence were considered. The loop plane was defined as the plane passing through the two anchor residues (points labelled "C" and "D" in Figure 41) and the centre of mass of the loop A. In the same way, the helix plane was defined as the one passing through the two anchor residues and the centre of mass of the helices B. The loop plane's orthogonal vector "a" and

the helix plane's orthogonal vector "b" were calculated as the cross product between the vectors AC◦DA, and CB◦BD, respectively (Figure 41b).

$$a = AC \circ DA$$

$$b = CB \circ BD$$

Finally, the angle θ was calculated using the scalar product between the two orthogonal vectors a and b, which corresponds to the angle between the helix plane and the loop plane.

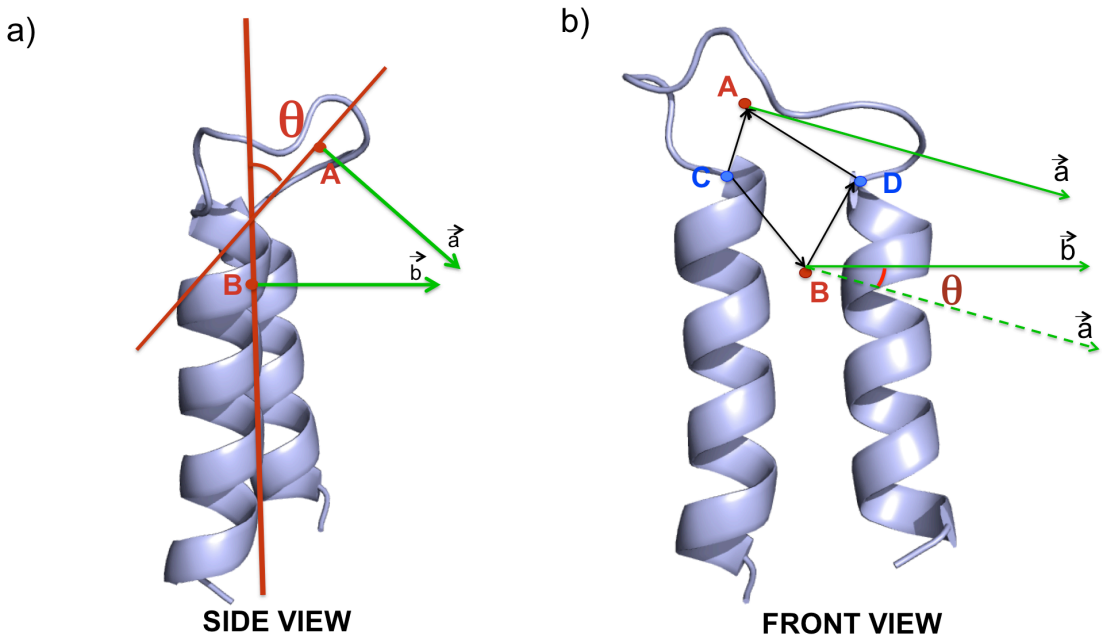


Figure 41
Geometrical representation of a loop connecting two helices. A) side view, B) front view. The two red points "A" and "B" correspond to the loop's centre of mass and helices' centre of mass, respectively; the two red lines are the side projection/section of the loop plane and helix plane; the blue points "C" and "D" correspond to the positions of the two anchor residues' Cα atoms; the green vectors "a" and "b" are the orthogonal vectors of the loop plane and helix plane, respectively, that are pointing out of the page plane. The angle θ between the two planes is also indicated.

Contact number $N_{contact}$

For each residue in the loop, the number of residues within a cut-off distance of 4Å is calculated (only backbone atoms N, Ca, C, O are considered). Potentially interacting residues include any residues that are outside the loop itself *and* are further than two residues from the beginning and end of the loop along the protein sequence. $N_{contact}$ is defined as the number of contacts between the loop and its surroundings (where any

particular pair of residues counts only once), divided be the total number of loop residue is in contact. $N_{contact}$ can be interpreted as the average number of residue than an interacting loop residue is in contact with. A higher $N_{contact}$ would indicate tighter contact, i.e. any single interacting loop residue in closely surrounded by many sequentially distant residues.

Results and Discussion

The hypothesis developed in the Prof. Deane's group was that the shapes of membrane loops tend to be biased, due to the presence of nearly parallel TM segments and the crowded environment of the membrane lipids. We propose that MP loops will favour a straight conformation, sticking out of the membrane, away from the remainder of the protein's transmembrane domain. In contrast, while some SP loops might have similar shapes, they are not confined by the membrane and will be more often able to "lie down" on the surface of the protein, in contact with sequentially distant residues, thus forming a more globular shape.

In order to test this hypothesis, I performed several tests on datasets of membrane and soluble protein loops. All loops, in both datasets, connected two helices. The first measure to assess this hypothesis is the loop angle θ, which is calculated as the angle between the plane of the loop and the plane of the two adjacent helices (Figure 41). We expected SPs to have a wide variety of θ angles, perhaps biased more towards a "lying down" conformation (larger θ values), we expected MPs to be biased towards a "straight" conformation (smaller θ values). We did indeed observe such a bias in loops up to 6 residues in length (Figure 42a), although only length-6 loops achieved a significant P value ($p < 0.05$) in a Kolmogorov-Smirnov test. MP loops have smaller θ angles than SP loops, indicating that they tend to "stand straighter", away from the protein, than the average soluble loop. Loops with lengths above 6 residues showed no clear difference. One concern was that this might be due to the way we calculate the angle θ. Given that our definition of θ includes the center of mass of all loop C-alpha atoms, longer loops might produce unpredictable behaviours. I thus repeated the calculations while calculating θ by only considering the first and last loop residue. The results obtained were virtually identical to the previous test (Figure 42b). While this validates our results, it also raises an interesting point: the difference observed in

the θ angles for short MP and SP loops (≤6 residues) is entirely due to changes in the conformation of the first and last loop residues.

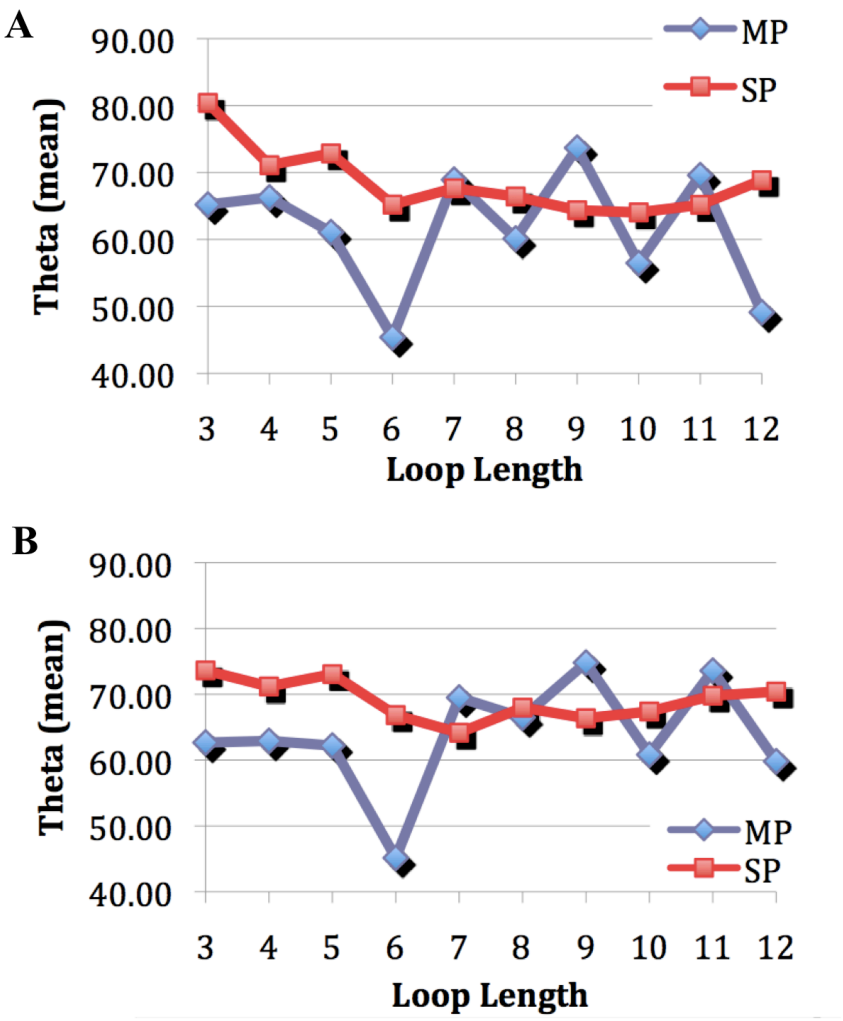


Figure 42.
Loop angle θ vs loop length. A) Angle θ calculated using the centre of mass of the entire loop; B) Angle θ calculated using the centre of mass of only the first and last loop residue.

I also investigated a second feature of membrane and soluble protein loops, namely their contacts with the rest of the protein. For this purpose I defined the contact number $N_{contact}$ (see Methods). A high contact number indicates "tight" contacts, where a single loop residue is in contact with many residues in the rest of the protein, a contact number close to 1 indicates "loose" contacts, where a single loop residue is in contact with only a single residue in the rest of the protein. We would expect MP to have lower average contact numbers, since their loops tend to stick out of the membrane, away from the bulk of the transmembrane domain. SP loops are expected

to have higher contact number, as they are not constrained by the membrane and can bend to be in closer contact with the bulk of the globular protein. On average, we do observe a small difference in $N_{contact}$ between MP ($N_{contact}$ = 1.53) and SP loops ($N_{contact}$ = 1.60). As Figure 43 shows, $N_{contact}$ tends to be smaller for MP loops when compared to SP loops of the same length, although some loop lengths show identical behaviour in the two datasets. Given the low numbers of example in the case of MP loops it is unlikely that this fluctuation is meaningful. We assume that, as more MP structure become known, this curve will smooth out to resemble that of SP loops, but shifted towards lower values of $N_{contact}$.

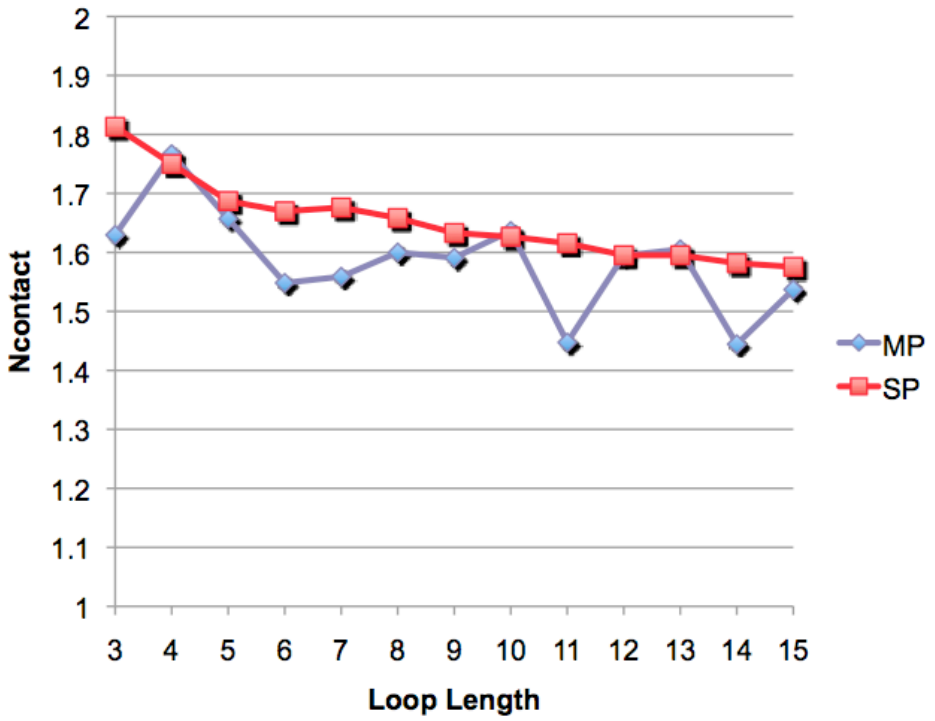


Figure 43
Contact number Ncontact vs loop length in membrane proteins (MP) and soluble proteins (SP).

I performed further investigations into the differences between the shapes of MP and SP loops. My results indicate a difference in the average conformations of the first and last loop residues resulting in a difference in loop angle θ. Short MP loops of up to 6 residues in length tend to have a lower θ, meaning they "stick out" away from the rest of the protein, rather than lying flat against it. We also observed a difference in the contacts between MP or SP loops and the remainder of the protein (Figure 43).

The results suggest that MP loops are more loosely connected to the rest of the protein structure than their soluble counterparts.

Conclusion

It thus seems that the loop sequence, shape of the first and last residues, as well as the loop's contacts with the remainder of the protein all contribute to the difference between MP and SP loops. It should be possible to engineer a statistical scoring scheme that utilizes measures similar to those defined here to identify fragments of soluble proteins that can be used to model the shape of membrane protein loops.

This is the first case in which a parameter to define, describe and quantify the difference between membrane and soluble protein loops is reported.

APPENDIX 2 - Docking technique: details

Docking tecnique

Due to the importance of protein-protein interactions in nature and the difficulty to obtain experimental data about the 3D structure of such complexes, the interest in protein docking is growing within the scientific community, and is currently one of the major challenges in the field of structural computational biology and bioinformatics.[36]

As described in Chapter 1, the docking technique has the task of assembling two separate protein components into their biologically relevant complex structure giving a model of the way the two proteins bind each other.[38,39]
Computational docking, if accurate and reliable, can therefore play an important role, both to infer functional properties and to guide new experiments. So, to its potential applications generating models of molecular complexes, although being a demanding problem, has attracted a vast deal of attention.[40]
There are no general rules to predict a binding interface. Basically, all docking approaches assume that the native complex is near the global minimum of the energy landscape. Therefore, all the current docking methods are based on optimization and attempt to find the global minimum of a function approximating the free energy of the complex. For details see Chapter 1.

Docking steps
In all the docking algorithms, there are two crucial steps to generate possible models of the three-dimensional arrangement of a complex:

3. *Searching (low-resolution search)*, consisting in the generation of thousands of alternative poses (decoys) to sample the rotational/translational space;

4. *scoring and ranking (high-resolution refinement)*, consisting in scoring these poses using a 'pseudo-energy' function in order to rank the poses and so to identify the native-like solutions.

Sampling the conformational space

The searching step involves an exhaustive search of the conformational space of one protein with respect to the other, resulting in a six-dimensional search (6D). The search of through the entire conformational space of the complex geometry makes the calculation expensive, so it is necessary to simplify the system.

First of all, the protein surface is represented in an easier way than the full-atoms representation, preserving the geometrical and physicochemical properties of the atoms. The basic description of the protein surface is the atomic representation of exposed residues, which can be usually achieved by mathematical models, such as geometrical shape descriptors or a grid.[42] The geometrical shape descriptors are widely used, in which each amino acid is represented by a sphere. As an alternative, a grid representation may be used, in which the points are calculated on the basis of force field potentials for van der Waals and electrostatic interactions.[245] The protein interior, the surface and the outer space can be differentiated by the use of grid-based molecular representations in combination with Fourier correlation algorithms.

Once having the easier representation of the system, almost all the docking programs use the same approach for the searching step: one protein is fixed in space (usually the bigger one) and the second one is rotated and translated around the first one. To minimize the degrees of freedom, both molecules are treated as rigid bodies, but still a simple systematic search is usually impracticable because the searching algorithm entails evaluating in the order of billions (10^9) distinct possibilities.[60]

Although geometric complementarity of the protein surface is the filtering criterion most commonly used to eliminate a large number of solutions with poor surface matching,[47] the docking problem is not simply matching two irregular shapes, but there are also other geometric, electrostatic or hydrophobic factors to take into account.[61] So, there are a lots of possible search methods that have been used in protein-protein docking programs. Most methods that perform well in CAPRI (see Chapter 1 and paragraph below) are based only on two approaches. These approaches are rigid body exhaustive search, involving Fast Fourier Transforms (FFT) and Monte Carlo Minimization.

Fast Fourier Transform

The Fast Fourier Transform (FFT) correlation approach, introduced in 1992 by Katchalski-Katzir and co-workers,[62] revolutionized rigid body protein-protein docking. The basic idea of the method is to represent one of the proteins (which will be identified as the receptor) on a fixed grid, the second protein (which will be referred to as the ligand) on a movable grid, and consider an interaction energy written in the form of a correlation function (or as a sum of a few correlation functions).[246] Since such energy functions can be efficiently calculated via Fast Fourier Transforms, it is possible to exhaustively sample the conformational space of protein-protein complexes evaluating the energies for billions of conformations on the grids, and thus to dock proteins without any *a priori* information on the expected structure.[50,247] The original scoring function, introduced by Katchalski-Katzir et al.,[62] accounted only for shape complementarity, but was later extended to include additional terms representing electrostatic interactions,[53,248] or both electrostatic and solvation contributions.[79] Since the FFT correlation method performs exhaustive sampling on a dense grid, it necessarily samples near-native conformations, independently of the shape of the energy surface. Anyway, as said before, even if correct solutions are generated, scoring functions often fail to rank them properly, so the structures that are close to the native conformation do not necessarily have the lowest energies.

Monte Carlo method

While the rigid body search based on FFT is global but has to rely on simplified energy functions defined on a grid, the Monte Carlo analyses involve more detailed energy functions and more thorough searches, possibly accounting for side chain flexibility. However, Monte Carlo is a statistical method, and due to the improved energy evaluation, the Monte Carlo based algorithms (such as the ones at the basis of RosettaDock[64] and ICM[81] docking programs) requires extensive calculations, and the simulations can explore only limited regions of the conformational space on reasonable time scales. Although the Monte Carlo minimization (MCM) trajectories can move ”uphill” and thus cross energy barriers, there is no guarantee that the search converges to the global minimum. In fact, the Monte Carlo based docking methods include a first stage that uses simplified protein models and energy functions to explore the conformational space, and only then switch to simulations that involve

models with more detailed geometry and more accurate energy functions. Thus, results provided by Monte Carlo minimization may heavily depend on the initial points of the simulations. For this reason, quite often this method is preceded by a perturbation of the ligand's position by random translations and rotations.

Scoring and ranking docking decoys

After a first low-resolution search step, a high-refinement scoring step is required to evaluate the energies of protein-protein docking poses in order to identify the one with the lowest energy as the predicted binding mode.[47] Docking algorithms can be classified on the basis of the position of the scoring phase in the algorithm flow, into two groups: integrated and edge functions. In integrated algorithms, scoring is integrated into the search stage and is thus used to filter emerging solutions. In edge algorithms, scoring is applied at the end of the search stage. The major difference is therefore that the scoring function is part of the design of the binding solutions in integrated algorithms, but not in edge algorithms.[42]

The fundamental point of any docking method is to be computationally efficient, having a scoring scheme able to evaluate a huge number of solutions and discriminate the native-like binding modes from the wrong decoy complex structures in a reasonable computational time.[42] Most of the docking algorithms developed so far use the extent of geometric complementarity of the protein surfaces as an initial filter to eliminate a large number of solutions with poor surface matching. It is, however, usually recognized that a criterion based exclusively on geometric complementarity is far from being enough to distinguish among native and non-native docked geometries, except for a very a small number of cases.[67] Numerous criteria have been implemented with different levels of success: steric complementarity of the shapes of the interaction sites, electrostatic interactions, hydrogen bonding, van der Waals, pair potential, desolvation, rotamer probabilities, contact pair potential and knowledge-based potentials. Different docking programs can use different combination of this terms in a weighted sum. Furthermore, exclusion of the solvent from the interface and the associated solvent entropy change play an important role in the stabilization of protein interactions, and can be estimated from empirical potentials or database derived functions.[18,68]

Resuming the docking procedure and scoring, the initial searching step yields a long list of candidate structures; the following step requires some forms of post-processing,

which may include: i) scoring or re-scoring of the docked conformations using a more accurate energy function, or ii) refining the conformations followed by re-scoring and, eventually, clustering.[74] These treatments usually improve the number of near-native conformations among the 10 to 100 lowest energy structures, but in most cases are unable to eliminate all false positives.

The flexibility problem

One of the most important difficulties in protein docking is that the interface residues of both interacting molecules may undergo a conformational change on complex formation. Although often the conformational change is limited to side-chains, a comparison of bound and unbound structures from PDB[34] reveals significant changes also in backbone conformation upon binding.[249] In protein-protein docking, because of the large number of atoms and degrees of conformational freedom involved, it would be impracticable to treat molecular flexibility in an explicit way with the current available computers, so flexibility is still the major challenge in protein-protein docking in terms of computational time.[250]

Since it is infeasible to explore all possible conformations, protein flexibility is introduced into docking protocols only in some steps, and in a variety of ways. As it is not feasible to execute extensive conformational searches during docking, unless the binding site is known, it has been generally adopted the two-stage approach. Initially the interacting molecules are treated as rigid bodies and a fully exploration of the six-dimensional rotational and translational space is made. At a second stage, a much smaller number of structures acquired in the initial stage are refined and re-ranked by more scrupulous energy functions that include small backbone and side-chain movements as well as rigid-body adjustments to take into account conformational changes.[79] Quite often among docking programs, both backbone and side-chain flexibility are being introduced using molecular dynamics (MD) in combination with some form of rigid-body docking, either before or after the MD simulations.[33,251]

In addition, the backbone flexibility can be modeled implicitly as a pregenerated ensemble of rigid structures generated from the unbound structure. The ensembles can be achieved by using different solved experimental structures from X-ray or NMR studies of diverse conformations of the same protein. If the experimental ensembles structures are not available, MD and Monte Carlo simulations have been used to

generate full protein ensembles and so to incorporate protein flexibility in docking (but only for small-scale movements).[46,252]

In alternative, normal-mode analysis can also be used to calculate the normal modes that are related with the flexibility of the protein and therefore may be used to model large global motion.[253] Is this way, only backbone motions are studied because the model used to calculate the normal modes considers C-alpha atoms only.[254]

In addition to inducing flexibility in the backbone, also the side-chain flexibility has a fundamental role and it can permit an efficient docking if some interfacial residues are in incorrect conformation.[39] In 1994 Totrov *et al.* published one of the first successful *ab initio* predictions of a complex that combined pseudo Brownian Monte Carlo minimization with a biased-probability global side-chain placement procedure. They showed that side-chain optimization was fundamental for discrimination of near-native conformations from false positives.[255]

The majority of the docking methods adjust side chain conformations explicitly during a refinement stage following the rigid-body search, which is characteristically performed only for a selected set of protein side chains close to the putative binding site and side chain conformations are represented as a discrete set of rotamers from libraries. These libraries are derived from statistical analysis of side-chain conformations in known high-resolution protein structures.[256] In fact, the 20 amino acids do not show the same degree of freedom. Amongst the protein complexes, arginine, lysine, glutamate and methionine present the highest frequency and amplitude of movements between the structures of free and co-crystallized proteins. [257] In contrast, many of the smaller polar or charged residues, such as asparagine, aspartate and histidine, and the large aromatics, phenylalanine, tyrosine and tryptophan, are markedly inflexible. So, for example, the lysine side chains flex 25 times more often than do phenylalanine side chains.[253,258-260]

Critical Assessment of Prediction of Interactions (CAPRI)

A variety of approaches have been used in docking programs have that mostly differ in the stages of the algorithms, showing different performances depending on the approach and the nature of the biological system. In this scenario, the comparison of different docking programs to establish their relative performances is very important. Indeed, it is required an objective valuation of the model quality. To this aim, the international Critical Assessment of Prediction of Interactions (CAPRI) experiment

was designed, precisely to evaluate current computational approaches of protein–protein docking (details in Chapter 1).[75]

To assess the quality of the models in CAPRI, after a least-square superimposition of the receptor in the model and target, three aspects are analyzed:

1. the RMSD distance L_{rms} between Ca atoms of the ligand (L) in the model and target;
2. the interface RMSD distance I_{rms}, calculated with the Ca's of the epitopes only;
3. the fraction of native contacts $f_{nc} = nc/Nc$, where Nc is the number of residue pairs in contact in the target, and nc the number of those native contacts that are present in the model.

These parameters I_{rms}, L_{rms}, and f_{nc} are then combined to classify and rank the models in correct and incorrect ones. In models of the ''high-quality'' and ''medium'' categories, f_{nc} is higher than 0.3, I_{rms} is lower than 2 Å and L_{rms} is lower than 5.0 Å. Models with 10–30% of the native contact pairs and I_{rms} between 2 Å and 4 Å, are placed in the ''acceptable'' category. Although their geometry is poor, they should still be useful for site-directed mutagenesis and other experiments, because a large part of the epitopes must be correctly identified to yield f_{nc} >= 0.1.Table 14 summarizes the criteria available for ranking the CAPRI predictions.[68]

Rank	f_{nc}	L_{rms}	Or I_{rms}
High	≥0.5	≤1.0	Or ≤1.0
Medium	≥0.3	1.0–5.0	Or 1.0–2.0
Acceptable	≥0.1	5.0–10.0	Or 2.0–4.0
Incorrect	<0.1		

Table 14. Criteria for ranking CAPRI predictions.

Four of the most common docking programs are RosettaDock,[64] ZDOCK,[79] HADDOCK[32] and ClusPro.[56] The present different advantages and disadvantages described in Chapter 1. Here a brief description of these methods is reported.

Docking programs

RosettaDock

RosettaDock[64] starts with a step in which the position of the ligand is perturbed by random translation and rotations. Next, a fast Monte Carlo minimization optimize the complex orientation with respect to features that do not depend on the explicit conformations of the side-chains (e.g. amino acid propensity at the interface, amino acid pair preferences, etc;[64]). After this step, explicit side chains are added back using a backbone-dependent rotamer packing algorithm and an all-atom optimization locates the local minimum energy conformation. No filters are applied to filter out promising models; in fact, the sampling problem is attacked creating a very large numbers of poses (decoys), which are then discriminated using a detailed scoring function including van der Waals and solvation interactions, hydrogen bonding, desolvation energy, residue-residue pair statistics, rotamer probabilities and a simple electrostatic term across the interface.[261] While the weights of most of the terms in the scoring function are of the same order of magnitude, the dominant contributions to discrimination are the van der Waals (packing) interactions, followed by solvation.[64] In this procedure, no backbone flexibility is allowed. Decoys are then ranked and clustered. To select final models, the decoy with the highest score is selected for each of the top ten largest clusters.

Predictions are usually performed without including any *a priori* biological information, being the energy of a model the primary criterion for the selection of the possible models. However, in some cases, biological information constrains can be used. [64]

ZDOCK

ZDOCK is a rigid body Fast Fourier Transform (FFT) based algorithm. It exhaustively samples the rigid body mutual orientations of the docking partners[79] and this stage could be filtered introducing biological structural information. In fact, ZDCOK procedure allows the definition of blocking residues (which in contrast with interfacial residues would be given zero desolation energy). The scoring function of ZDOCK is a weighted sum of energy terms representing shape complementarity, van

der Waals energy, Coulombic electrostatics and a simplified implementation of the atomic contact potential score, which essentially measure the solvation/desolvation function contributions to the binding free energy. A protocol on CHARMm removes possible clashes, optimize the polar interface and optimize the charge interaction. Finally, a cluster of the top predictions is performed to reduce structural redundancy.

HADDOCK

The searching step in HADDOCK[32] algorithm starts with a randomization of the orientation of the two interacting molecules, followed by a rigid body docking and energy minimization. After this step, in which the two proteins are treated as rigid bodies, there is a semirigid simulated annealing in torsion angle space, and a final refinement in Cartesian space with explicit solvent. During the last two steps, the amino acids at the interface (both side chains and backbone) are allowed to move to optimize the interface packing.[32] The filtering applied in the searching stage take in account experimental information. In fact, an HADDOCK's peculiar approach is the possibility to use biochemical and/or biophysical interaction data, such as chemical shift perturbation data resulting from NMR experiments or mutagenesis data, to reduce the conformational search space and filter the solutions. In particular, the most fundamental differences in comparison with other algorithms is that HADDOCK translates information about the interface into highly ambiguous inter-molecular distance restraints used to directly drive the docking process.[32,262]

Flexibility is introduced at several levels in the algorithm: in the searching stage, it is introduced by docking from ensembles of structures (coming from experimental data or short MD simulation in explicit solvent[263]) and taking all possible pairwise combinations and by introduction of flexibility in the side chain at the interface; instead, at the final refinement stage the algorithm allows both side chains and backbone flexibility by simulated annealing MD and steepest descent minimization.

The final structures are clustered using the pairwise backbone RMSD at the interface and they are scored as a sum of electrostatic, van der Waals, electrostatic, buried surface area, desolvation energy and Ambiguous Interaction Restraints (AIR) that are derived from any kind of experimental information available concerning the residues involved in the inter-molecular interaction.[32] Recently, explicit inclusion of interfacial water was incorporated in the docking protocol and incorporated in CAPRI predictions,[264] observing a improvement in the f_{nat}.[103]

ClusPro

The ClusPro[56] is a fully-automated docking program that includes three main steps. First, it runs PIPER, a rigid body docking program based on the Fast Fourier Transform (FFT) correlation approach. The major advantage of PIPER is the inclusion of pairwise interaction potentials.[246] The top 1000 structures are retained from PIPER to the second step consisting in clustering.[57] The clustering of the retained conformations is based on the pairwise RSMD of ligand structures, calculated for the atoms that are within 10 Å of any atom of the fixed receptor. It uses a simple greedy algorithm to find the structures with the largest number of neighbors within a clustering radius *RC. T*he choice of *RC* depends on a clustering parameter $0 \leq \Delta \leq 1$, which is based on the histogram of pairwise RMSD values, and measures the depth of the separation between clusters. Once a clustering radius *RC* is selected (default value of 9 Å), the structure with the highest number of neighbors within *RC* is considered as the center of the first cluster and is the representative structure for the cluster. The members of this cluster are removed, and the algorithm selects the next structure with the highest number of neighbors from the remaining ligands until the set is exhausted, thereby generating 10 to 30 rank ordered clusters.[57] The 30 largest cluster centers are then subjected to a straightforward (300 step and fixed backbone) van der Waals minimization using CHARMm[265,266] to remove potential side chain clashes.

Conclusive notes

The success of docking algorithms has consistently improved over the last years, as measuring by the CAPRI blind docking experiment. Due to such efforts, on one hand the applicability of in silico created complexes is becoming widly accepted, and on other hand the various available docking programs can be objective compared.

CONCLUSIONS

The aim of my PhD work has been to provide novel computational instruments and to give useful insight into one of the most crucial topics in nature: the protein-protein interaction (Chapter1).

In particular, my research has been devoted to two main aspects: i) the development of new methods to analyse protein complexes, and to compare and rank multiple docking solutions (Chapters 2, 3 and 4), and ii) the application of these methods, in combination with classical state-of-art computational biology techniques, to predict and analyse the binding mode in real biological systems, which are related to particular diseases. The second part of the work has been afforded in collaboration with experimental groups (Chapters 5, 6 and 7), in order to take advantage of specific biological information on the systems under study.

Part 1: development of new methodologies

Due to the importance of protein-protein interactions, the interest in their structural characterization is constantly growing within the scientific community.[1] However, due to the difficulty to obtain experimental 3D structures for protein-protein complexes, their accurate prediction through molecular docking simulations has become one of the major challenges in the field of structural computational biology and bioinformatics.[2,3] Unfortunately, although success in docking algorithms has consistently improved over the last years,[59] correctly ranking predicted models to single out the best ones from a decoys ensemble remains an open challenge.

In this scenario, it is of timely interest, both for bioinformaticians and wet biologists, to have programs and tools able to: i) automatically analyze features of a complex interface, and to easily and intuitively discriminate between similar and different binding solutions, ii) compare multiple docking solutions, in order to appreciate at a glance which are the residues most often predicted as interacting and iii) accurately rank hundreds of docking solutions to distinguish native-like from incorrect ones..

On this basis, in my PhD work I developed three web tools to automatically analyze biological complex structures, COCOMAPS[4], and to compare and rank multiple docking solutions, CONS-COCOMAPS[5] and CONS-RANK. The web tool COCOMAPS (available at https://www.molnac.unisa.it/BioTools/cocomaps/, details

in Chapter 2) analyzes the interfaces of protein-protein and protein-nucleic acids complexes, combining in a single tool the traditional analysis and 3D visualization of biocomplexes with the effectiveness of the contact map view.

The web tool CONS-COCOMAPS (available at https://www.molnac.unisa.it/BioTools/conscocomaps/, details in Chapter 3), instead, easily measures and visualizes the consensus in multiple docking solutions. This novel tool uses the conservation of inter-residue contacts as an estimate of the similarity between different docking solutions.

CONS-RANK (available upon request from the authors, details in Chapter 4) is a simple and effective method to rank multiple docking solutions; it is well performing and robust, thus offering a valid alternative to the ranking methods already available.

Part 2: study of protein-protein interactions in real biological systems

Firstly, I studied two cases of biological complexes involved in the celiac disease; both studies were afforded in collaboration with the group directed by Prof. Daniele Sblattero, University of Piemonte Orinteale (Italy) and the group directed by Prof. Carla Esposito, University of Salerno (Italy).

In the first study (Chapter 5), I performed docking simulation to obtain the molecular model for a biological complex involved in the celiac disease, made up by celiac autoantibodies isolated from celiac patients, and its auto-antigen Tissue Transglutaminase type 2 (TG2).[7,154]

In the second study (Chapter 6), instead, I performed docking simulation and the following analysis to the complex between the celiac autoantibody Ab1-clone 2.8 and the mouse anti-idiotype antibody Ab2-AIT2 elicited against Ab1. These investigations provided useful insight into orientation and characterization of the complexes' binding site. [6]

Due to the crucial involvement of the complex TG2-antoantibody and the idiotypic network in the celiac disease causes, the diagnosis applications and the promising therapeutic applications, the proposed models could help rationalizing the experiments as crucial step for the study of the celiac disease mechanism, the improvement of diagnosis strategies and the rational design of molecules for pharmacological and therapeutic purposes.[267]

In addition, I worked on a project regarding a pathogenic mutant of the enzymatic system FXa, that causes problem in the process of blood coagulation, taking

advantage of the computational molecular dynamic technique.[8] The study was afforded in collaboration with the Prof. De Cristofaro's group, Catholic University School of Medicine, Rome (Italy) and the group directed by Prof. Peyvandi, Ospedale Maggiore Policlinico and Università degli Studi di Milano (Italy).
This study can help the emergence of new therapeutic products for the treatment of coagulation deficiencies (for details see Chapter 7)

Visiting PhD at University of Oxford

Finally, during my PhD I spent seven months in the groups of the Prof. Charlotte Deane, Department of Statistics, University of Oxford (UK). In that period, I studied the geometrical features of the proteins' regions most recurrent in the protein-protein interaction, the loops, clarifying some structural aspects of them in one of the most important and huge class of proteins: the membrane proteins (details in Appendix 1).

REFERENCES

1 Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. Protein function in the post-genomic era. *Nature* **405**, 823-826 (2000).

2 Aloy, P. & Russell, R. B. Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol* **7**, 188-197 (2006).

3 Arkin, M. R. & Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov* **3**, 301-317 (2004).

4 Vangone, A., Spinelli, R., Scarano, V., Cavallo, L. & Oliva, R. COCOMAPS: a web application to analyse and visualize contacts at the interface of biomolecular complexes. *Bioinformatics* **27**, 2915-2916 (2011).

5 Vangone, A., Oliva, R. & Cavallo, L. CONS-COCOMAPS: a novel tool to measure and visualize the conservation of inter-residue contacts in multiple docking solutions. *BMC Bioinformatics* **13 Suppl 4**, S19 (2012).

6 Di Niro, R. *et al.* Anti-idiotypic response in mice expressing human autoantibodies. *Mol Immunol* **45**, 1782-1791 (2008).

7 Marzari, R. *et al.* Molecular dissection of the tissue transglutaminase antoantibody response in celiac disease. *Journal Of Immunology* **166**, 4170-4176 (2001).

8 Pinotti, M. *et al.* Impared prothrombinase activity of factor X Gly381Asp results in severe familial CRM+ FX deficiency. *Thromb Haemost* **89**, 243-248 (2003).

9 Gellman, S. H. Introduction: Molecular Recognition. *Chem Rev* **97**, 1231-1232 (1997).

10 Phizicky, E. M. & Fields, S. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* **59**, 94-123 (1995).

11 Nooren, I. M. & Thornton, J. M. Diversity of protein-protein interactions. *Embo J* **22**, 3486-3492 (2003).

12 Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13-20 (1996).

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

14 Ofran, Y. & Rost, B. Analysing six types of protein-protein interfaces. *J Mol Biol* **325**, 377-387 (2003).

15 Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. Dissecting subunit interfaces in homodimeric proteins. *Proteins* **53**, 708-719 (2003).

16 Chakrabarti, P. & Janin, J. Dissecting protein-protein recognition sites. *Proteins* **47**, 334-343 (2002).

17 de Vries, S. J. & Bonvin, A. M. Intramolecular surface contacts contain information about protein-protein interface regions. *Bioinformatics* **22**, 2094-2098 (2006).

18 Fahmy, A. & Wagner, G. TreeDock: a tool for protein docking based on minimizing van der Waals energies. *J Am Chem Soc* **124**, 1241-1250 (2002).

19 Jones, S. & Thornton, J. M. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* **272**, 133-143 (1997).

20 Lo Conte, L., Chothia, C. & Janin, J. The atomic structure of protein-protein recognition sites. *J Mol Biol* **285**, 2177-2198 (1999).

21 Tsai, C. J., Kumar, S., Ma, B. & Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci* **8**, 1181-1190 (1999).
22 Northrup, S. H. & Erickson, H. P. Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc Natl Acad Sci U S A* **89**, 3338-3342 (1992).
23 Wells, J. A. Binding in the growth hormone receptor complex. *Proc Natl Acad Sci U S A* **93**, 1-6 (1996).
24 Blundell, T. L. & Fernandez-Recio, J. Cell biology: brief encounters bolster contacts. *Nature* **444**, 279-280 (2006).
25 Tang, C., Iwahara, J. & Clore, G. M. Visualization of transient encounter complexes in protein-protein association. *Nature* **444**, 383-386 (2006).
26 Schreiber, G. & Fersht, A. R. Rapid, electrostatically assisted association of proteins. *Nat Struct Biol* **3**, 427-431 (1996).
27 Gabdoulline, R. R. & Wade, R. C. Biomolecular diffusional association. *Curr Opin Struct Biol* **12**, 204-213 (2002).
28 Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* **99**, 14116-14121 (2002).
29 Monecke, P., Borosch, T., Brickmann, J. & Kast, S. M. Determination of the interfacial water content in protein-protein complexes from free energy simulations. *Biophys J* **90**, 841-850 (2006).
30 Fleury, D. *et al.* A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site. *Nat Struct Biol* **6**, 530-534 (1999).
31 Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Protein-protein docking dealing with the unknown. *J Comput Chem* **31**, 317-342 (2009).
32 Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**, 1731-1737 (2003).
33 Smith, G. R., Fitzjohn, P. W., Page, C. S. & Bates, P. A. Incorporation of flexibility into rigid-body docking: applications in rounds 3-5 of CAPRI. *Proteins* **60**, 263-268 (2005).
34 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).
35 Smith, G. R. & Sternberg, M. J. E. Prediction of protein-protein interactions by docking methods. *Current Opinion In Structural Biology* **12**, 28-35 (2002).
36 Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Protein-ligand docking: current status and future challenges. *Proteins* **65**, 15-26 (2006).
37 Fernandez-Recio, J., Abagyan, R. & Totrov, M. Improving CAPRI predictions: optimized desolvation for rigid-body docking. *Proteins* **60**, 308-313 (2005).
38 Gardiner, E. J., Holliday, J. D., O'Dowd, C. & Willett, P. Effectiveness of 2D fingerprints for scaffold hopping. *Future Med Chem* **3**, 405-414 (2011).
39 Zacharias, M. ATTRACT: protein-protein docking in CAPRI using a reduced protein model. *Proteins* **60**, 252-256 (2005).
40 Gray, J. J. *et al.* Protein-protein docking predictions for the CAPRI experiment. *Proteins* **52** (2003).
41 Camacho, C. J. & Vajda, S. Protein-protein association kinetics and protein docking. *Curr Opin Struct Biol* **12**, 36-40 (2002).

42 Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409-443 (2002).
43 Halperin, I., Wolfson, H. & Nussinov, R. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure* **12**, 1027-1038 (2004).
44 Lee, K., Sim, J. & Lee, J. Study of protein-protein interaction using conformational space annealing. *Proteins* **60**, 257-262 (2005).
45 Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* **33**, W363-367 (2005).
46 Smith, G. R., Sternberg, M. J. & Bates, P. A. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J Mol Biol* **347**, 1077-1101 (2005).
47 Yu, Y. H., Lu, B. Z., Han, J. G. & Zhang, P. F. Scoring protein-protein docked structures based on the balance and tightness of binding. *J Comput Aided Mol Des* **18**, 251-260 (2004).
48 Vajda, S. & Camacho, C. J. Protein-protein docking: is the glass half-full or half-empty? *Trends Biotechnol* **22**, 110-116 (2004).
49 McCammon, J. A. Theory of biomolecular recognition. *Curr Opin Struct Biol* **8**, 245-249 (1998).
50 Mendez, R., Leplae, R., De Maria, L. & Wodak, S. J. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins* **52**, 51-67 (2003).
51 Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167-195 (1995).
52 Leopold, P. E., Montal, M. & Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc Natl Acad Sci U S A* **89**, 8721-8725 (1992).
53 Mandell, J. G. *et al.* Protein docking using continuum electrostatics and geometric fit. *Protein Eng* **14**, 105-113 (2001).
54 Verkhivker, G. M., Rejto, P. A., Gehlhaar, D. K. & Freer, S. T. Exploring the energy landscapes of molecular recognition by a genetic algorithm: analysis of the requirements for robust docking of HIV-1 protease and FKBP-12 complexes. *Proteins* **25**, 342-353 (1996).
55 Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556-560 (2012).
56 Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **20**, 45-50 (2004).
57 Kozakov, D., Clodfelter, K. H., Vajda, S. & Camacho, C. J. Optimal clustering for detecting near-native conformations in protein docking. *Biophys J* **89**, 867-875 (2005).
58 Wass, M. N., Fuentes, G., Pons, C., Pazos, F. & Valencia, A. Towards the prediction of protein interaction partners using physical docking. *Mol Syst Biol* **7**, 469 (2011).
59 Lensink, M. F., Mendez, R. & Wodak, S. J. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins* **69**, 704-718 (2007).

60 Li, C. H., Ma, X. H., Chen, W. Z. & Wang, C. X. A protein-protein docking algorithm dependent on the type of complexes. *Protein Eng* **16**, 265-269 (2003).
61 Kozakov, D., Schueler-Furman, O. & Vajda, S. Discrimination of near-native structures in protein-protein docking by testing the stability of local minima. *Proteins* **72**, 993-1004 (2008).
62 Katchalski-Katzir, E. *et al.* Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A* **89**, 2195-2199 (1992).
63 Abagyan, R. & Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* **235**, 983-1002 (1994).
64 Gray, J. J. *et al.* Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* **331**, 281-299, (2003).
65 Fischer, D., Lin, S. L., Wolfson, H. L. & Nussinov, R. A geometry-based suite of molecular docking processes. *J Mol Biol* **248**, 459-477 (1995).
66 Shen, Y., Paschalidis, I., Vakili, P. & Vajda, S. Protein docking by the underestimation of free energy funnels in the space of encounter complexes. *PLoS Comput Biol* **4**, e1000191 (2008).
67 Palma, P. N., Krippahl, L., Wampler, J. E. & Moura, J. J. BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins* **39**, 372-384 (2000).
68 Janin, J. *et al.* CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **52**, 2-9 (2003).
69 Camacho, C. J., Weng, Z., Vajda, S. & DeLisi, C. Free energy landscapes of encounter complexes in protein-protein association. *Biophys J* **76**, 1166-1178 (1999).
70 Lise, S., Walker-Taylor, A. & Jones, D. T. Docking protein domains in contact space. *BMC Bioinformatics* **7**, 310 (2006).
71 Bajaj, S. P. & Mann, K. G. Simultaneous purification of bovine prothrombin and factor X. Activation of prothrombin by trypsin-activated factor X. *J Biol Chem* **248**, 7729-7741 (1973).
72 Stanton, C. & Wallin, R. Processing and trafficking of clotting factor X in the secretory pathway. Effects of warfarin. *Biochem J* **284 (Pt 1)**, 25-31 (1992).
73 Jackson, R. M. Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. *Protein Sci* **8**, 603-613 (1999).
74 Li, L., Chen, R. & Weng, Z. RDOCK: refinement of rigid-body protein docking predictions. *Proteins* **53**, 693-707 (2003).
75 Camacho, C. J. & Gatchell, D. W. Successful discrimination of protein interactions. *Proteins* **52**, 92-97 (2003).
76 Janin, J. Welcome to CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins-Structure Function And Genetics* **47**, 257-257 (2002).
77 Wodak, S. J. & Janin, J. Structural basis of macromolecular recognition. *Adv Protein Chem* **61**, 9-73 (2002).
78 Wodak, S. J. & Mendez, R. Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Curr Opin Struct Biol* **14**, 242-249 (2004).

79 Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80-87 (2003).
80 Comeau, S. R., Vajda, S. & Camacho, C. J. Performance of the first protein docking server ClusPro in CAPRI rounds 3-5. *Proteins* **60**, 239-244 (2005).
81 Fernandez-Recio, J., Totrov, M. & Abagyan, R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins* **52**, 113-117 (2003).
82 Wang, C., Schueler-Furman, O. & Baker, D. Improved side-chain modeling for protein-protein docking. *Protein Sci* **14**, 1328-1339 (2005).
83 Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. 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 (2001).
84 Cavallo, L., Kleinjung, J. & Fraternali, F. POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res* **31**, 3364-3366 (2003).
85 Fischer, T. B. *et al.* Assessing methods for identifying pair-wise atomic contacts across binding interfaces. *J Struct Biol* **153**, 103-112 (2006).
86 Gabdoulline, R. R., Wade, R. C. & Walther, D. MolSurfer: A macromolecular interface navigator. *Nucleic Acids Res* **31**, 3349-3351 (2003).
87 Kleinjung, J. & Fraternali, F. POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic Acids Res* **33**, W342-346 (2005).
88 Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774-797 (2007).
89 Negi, S. S., Schein, C. H., Oezguen, N., Power, T. D. & Braun, W. InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics* **23**, 3397-3399 (2007).
90 Reynolds, C., Damerell, D. & Jones, S. ProtorP: a protein-protein interaction analysis server. *Bioinformatics* **25**, 413-414 (2009).
91 Salerno, W. J., Seaver, S. M., Armstrong, B. R. & Radhakrishnan, I. MONSTER: inferring non-covalent interactions in macromolecular structures from atomic coordinate data. *Nucleic Acids Res* **32**, W566-568 (2004).
92 Tina, K. G., Bhadra, R. & Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic Acids Res* **35**, W473-476 (2007).
93 Holm, L. & Sander, C. Mapping the protein universe. *Science* **273**, 595-603 (1996).
94 Pulim, V., Berger, B. & Bienkowska, J. Optimal contact map alignment of protein-protein interfaces. *Bioinformatics* **24**, 2324-2328 (2008).
95 Bernauer, J., Aze, J., Janin, J. & Poupon, A. A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics* **23**, 555-562 (2007).
96 DeLano Scientific, L. *http://www.pymol.org* (2002).
97 Hubbard, S. J. & Thornton, J. M. 'NACCESS' Computer Program. (1993).
98 McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**, 777-793 (1994).
99 Bhat, T. N. *et al.* Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc Natl Acad Sci U S A* **91**, 1089-1093 (1994).

100 Cauerhff, A., Goldbaum, F. A. & Braden, B. C. Structural mechanism for affinity maturation of an anti-lysozyme antibody. *Proc Natl Acad Sci U S A* **101**, 3539-3544 (2004).
101 Janin, J. Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst* **6**, 2351-2362 (2010).
102 Bourquard, T., Bernauer, J., Aze, J. & Poupon, A. A collaborative filtering approach for protein-protein docking scoring functions. *PLoS One* **6**, e18541.
103 de Vries, S. J. *et al.* HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* **69**, 726-733 (2007).
104 Gao, M. & Skolnick, J. New benchmark metrics for protein-protein docking methods. *Proteins* **79**, 1623-1634.
105 Dunbrack, R. L. & Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science* **6**, 1661-1681 (1997).
106 Pollastri, G., Martin, A. J., Mooney, C. & Vullo, A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics* **8**, 201 (2007).
107 Albrecht, M., Tosatto, S. C., Lengauer, T. & Valle, G. Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng* **16**, 459-462 (2003).
108 Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. & Mornon, J. P. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng* **6**, 377-382 (1993).
109 Konings, D. A. & Hogeweg, P. Pattern analysis of RNA secondary structure similarity and consensus of minimal-energy folding. *J Mol Biol* **207**, 597-614 (1989).
110 Kiryu, H., Kin, T. & Asai, K. Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics* **23**, 434-441 (2007).
111 Witwer, C., Hofacker, I. L. & Stadler, P. F. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans Comput Biol Bioinform* **1**, 66-77 (2004).
112 Anwar, M., Nguyen, T. & Turcotte, M. Identification of consensus RNA secondary structures using suffix arrays. *BMC Bioinformatics* **7**, 244 (2006).
113 Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res* **37**, W465-468 (2009).
114 Tjalsma, H. & van Dijl, J. M. Proteomics-based consensus prediction of protein retention in a bacterial membrane. *Proteomics* **5**, 4472-4482 (2005).
115 Ginalski, K. & Rychlewski, L. Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins* **53 Suppl 6**, 410-417 (2003).
116 Plewczynski, D., Lazniewski, M., von Grotthuss, M., Rychlewski, L. & Ginalski, K. VoteDock: consensus docking method for prediction of protein-ligand interactions. *J Comput Chem* **32**, 568-581.
117 de Vries, S. J. & Bonvin, A. M. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One* **6**, e17695.

118 Huang, B. & Schroeder, M. Using protein binding site prediction to improve protein docking. *Gene* **422**, 14-21 (2008).
119 Qin, S. & Zhou, H. X. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics* **23**, 3386-3387 (2007).
120 Micheelsen, P. O. *et al.* Structural and mutational analyses of the interaction between the barley alpha-amylase/subtilisin inhibitor and the subtilisin savinase reveal a novel mode of inhibition. *J Mol Biol* **380**, 681-690 (2008).
121 Menetrey, J. *et al.* Structural basis for ARF1-mediated recruitment of ARHGAP21 to Golgi membranes. *Embo J* **26**, 1953-1962 (2007).
122 Bonsor, D. A., Grishkovskaya, I., Dodson, E. J. & Kleanthous, C. Molecular mimicry enables competitive recruitment by a natively disordered protein. *J Am Chem Soc* **129**, 4800-4807 (2007).
123 Leulliot, N. *et al.* Structure of the yeast tRNA m7G methylation complex. *Structure* **16**, 52-61 (2008).
124 Najmudin, S. *et al.* Putting an N-terminal end to the Clostridium thermocellum xylanase Xyn10B story: crystal structure of the CBM22-1-GH10 modules complexed with xylohexaose. *J Struct Biol* **172**, 353-362.
125 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
126 Lensink, M. F. & Wodak, S. J. Docking and scoring protein interactions: CAPRI 2009. *Proteins* **78**, 3073-3084 (2010).
127 Camacho, C. J., Gatchell, D. W., Kimura, S. R. & Vajda, S. Scoring docked conformations generated by rigid-body protein-protein docking. *Proteins* **40**, 525-537 (2000).
128 Tress, M. *et al.* Scoring docking models with evolutionary information. *Proteins* **60**, 275-280 (2005).
129 Cheng, T. M., Blundell, T. L. & Fernandez-Recio, J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins* **68**, 503-515 (2007).
130 Fink, F., Hochrein, J., Wolowski, V., Merkl, R. & Gronwald, W. PROCOS: computational analysis of protein-protein complexes. *J Comput Chem* **32**, 2575-2586 (2011).
131 Andrusier, N., Nussinov, R. & Wolfson, H. J. FireDock: fast interaction refinement in molecular docking. *Proteins* **69**, 139-159 (2007).
132 Pierce, B. & Weng, Z. ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins* **67**, 1078-1086 (2007).
133 Huang, S. Y. & Zou, X. An iterative knowledge-based scoring function for protein-protein recognition. *Proteins* **72**, 557-579 (2008).
134 Martin, O. & Schomburg, D. Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. *Proteins* **70**, 1367-1378 (2008).
135 Liang, S., Meroueh, S. O., Wang, G., Qiu, C. & Zhou, Y. Consensus scoring for enriching near-native structures from protein-protein docking decoys. *Proteins* **75**, 397-403 (2009).
136 Bourquard, T., Bernauer, J., Aze, J. & Poupon, A. A collaborative filtering approach for protein-protein docking scoring functions. *PLoS One* **6**, e18541 (2011).
137 Liu, S. & Vakser, I. A. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC Bioinformatics* **12**, 280 (2011).

138 Mitra, P. & Pal, D. Using correlated parameters for improved ranking of protein-protein docking decoys. *J Comput Chem* **32**, 787-796 (2011).
139 Khashan, R., Zheng, W. & Tropsha, A. Scoring protein interaction decoys using exposed residues (SPIDER): A novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. *Proteins* **80**, 2207-2217 (2012).
140 Gong, X. *et al.* Protein-protein docking with binding site patch prediction and network-based terms enhanced combinatorial scoring. *Proteins* **78**, 3150-3155 (2010).
141 Zimmermann, M. T., Leelananda, S. P., Kloczkowski, A. & Jernigan, R. L. Combining statistical potentials with dynamics-based entropies improves selection from protein decoys and docking poses. *The journal of physical chemistry. B* **116**, 6725-6731 (2012).
142 Rodrigues, J. P. *et al.* Clustering biomolecular complexes by residue contacts similarity. *Proteins* **80**, 1810-1817 (2012).
143 Hwang, H., Vreven, T., Pierce, B. G., Hung, J. H. & Weng, Z. Performance of ZDOCK and ZRANK in CAPRI rounds 13-19. *Proteins* **78**, 3104-3110 (2010).
144 Lensink, M. F. & Wodak, S. J. Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins* **78**, 3085-3095 (2010).
145 Liu, S., Gao, Y. & Vakser, I. A. DOCKGROUND protein-protein docking decoy set. *Bioinformatics* **24**, 2634-2635 (2008).
146 Janin, J. & Wodak, S. The third CAPRI assessment meeting Toronto, Canada, April 20-21, 2007. *Structure* **15**, 755-759 (2007).
147 Janin, J. The targets of CAPRI Rounds 13-19. *Proteins* **78**, 3067-3072 (2010).
148 Bode, W., Papamokos, E. & Musil, D. The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech Hirudo medicinalis. Structural analysis, subtilisin structure and interface geometry. *European journal of biochemistry / FEBS* **166**, 673-692 (1987).
149 Collin, P., Kaukinen, K. & Maki, M. Clinical features of celiac disease today. *Digestive Diseases* **17**, 100-106 (1999).
150 Sollid, L. M. Molecular basis of celiac disease. *Annual Review Of Immunology* **18**, 53-81 (2000).
151 Sollid, L. M., Molberg, O., McAdam, S. & Lundin, K. E. A. Autoantibodies in coeliac disease: tissue transglutaminase - guilt by association? *Gut* **41**, 851-852 (1997).
152 Sollid, L. M. Coeliac disease: dissecting a complex inflammatory disorder. *Nat Rev Immunol* **2**, 647-655 (2002).
153 Wieser, H. Relation between gliadin structure and coeliac toxicity. *Acta Paediatrica* **85**, 3-9 (1996).
154 Dieterich, W. *et al.* Identification of tissue transglutaminase as the autoantigen of celiac disease. *Nature Medicine* **3**, 797-801 (1997).
155 Sblattero, D. *et al.* The analysis of the fine specificity of celiac disease antibodies using tissue transglutaminase fragments. *European Journal Of Biochemistry* **269**, 5175-5181 (2002).
156 Nakaoka, H. *et al.* Gh: a GTP-binding protein with transglutaminase activity and receptor signaling function. *Science* **264**, 1593-1596 (1994).
157 Pedersen, L. C. *et al.* Transglutaminase factor XIII uses proteinase-like catalytic triad to crosslink macromolecules. *Protein Sci* **3**, 1131-1135 (1994).

158 Liu, S. P., Cerione, R. A. & Clardy, J. Structural basis for the guanine nucleotide-binding activity of tissue transglutaminase and its regulation of transamidation activity. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **99**, 2743-2747 (2002).
159 Ahvazi, B., Kim, H. C., Kee, S. H., Nemes, Z. & Steinert, P. M. Three-dimensional structure of the human transglutaminase 3 enzyme: binding of calcium ions changes structure for activation. *Embo J* **21**, 2055-2067 (2002).
160 Noguchi, K. *et al.* Crystal structure of red sea bream transglutaminase. *J Biol Chem* **276**, 12055-12059 (2001).
161 Yee, V. C. *et al.* Three-dimensional structure of a transglutaminase: human blood coagulation factor XIII. *Proc Natl Acad Sci U S A* **91**, 7296-7300 (1994).
162 Pinkas, D. M., Strop, P., Brunger, A. T. & Khosla, C. Transglutaminase 2 undergoes a large conformational change upon activation. *Plos Biology* **5**, 2788-2796 (2007).
163 Caputo, I., Barone, M., Martucciello, S., Lepretti, M. & Esposito, C. Tissue transglutaminase in celiac disease: role of autoantibodies. *Amino Acids* **36**, 693-699 (2009).
164 Sircar, A., Kim, E. T. & Gray, J. J. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res* **37**, W474-479 (2009).
165 Dalgleish, A. G. & Kennedy, R. C. ANTI-IDIOTYPIC ANTIBODIES AS IMMUNOGENS - IDIOTYPE-BASED VACCINES. *Vaccine* **6**, 215-220 (1988).
166 Oudin, J. & Michel, M. [A new allotype form of rabbit serum gamma-globulins, apparently associated with antibody function and specificity]. *C R Hebd Seances Acad Sci* **257**, 805-808 (1963).
167 Jerne, N. K. Towards a network theory of the immune system. *Annales d'immunologie* **125C**, 373-389 (1974).
168 Abu-Shakra M, S. Y. in *Autoantibodies* Vol. 2nd Ch. 10, 69-76 (2007).
169 Pan, Y., Yuhasz, S. C. & Amzel, L. M. ANTIIDIOTYPIC ANTIBODIES - BIOLOGICAL FUNCTION AND STRUCTURAL STUDIES. *Faseb Journal* **9**, 43-49 (1995).
170 Vani, J., Chatterjee, J., Shaila, M. S., Nayak, R. & Chandra, N. R. Structural basis for the function of anti-idiotypic antibody in immune memory. *Mol Immunol* **46**, 1250-1255 (2009).
171 Ban, N. *et al.* CRYSTAL-STRUCTURE OF AN IDIOTYPE ANTIIDIOTYPE FAB COMPLEX. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **91**, 1604-1608 (1994).
172 Escobar, J. C., Kochik, S. A., Skaletsky, E., Rosenberg, J. S. & Beardsley, T. R. Immunization of cats against feline infectious peritonitis with anti-idiotypic antibodies. *Viral Immunol* **5**, 71-79 (1992).
173 Braden, B. C. *et al.* Crystal structure of an Fv-Fv idiotope - Anti-idiotope complex at 1.9 angstrom resolution. *Journal Of Molecular Biology* **264**, 137-151 (1996).
174 Dall'Acqua, W., Goldman, E. R., Eisenstein, E. & Mariuzza, R. A. A mutational analysis of the binding of two different proteins to the same antibody. *Biochemistry* **35**, 9667-9676 (1996).
175 AlLazikani, B., Lesk, A. M. & Chothia, C. Standard conformations for the canonical structures of immunoglobulins. *Journal Of Molecular Biology* **273**, 927-948 (1997).

176 Garcia, K. C., Ronco, P. M., Verroust, P. J., Brunger, A. T. & Amzel, L. M. Three-dimensional structure of an angiotensin II-Fab complex at 3 A: hormone recognition by an anti-idiotypic antibody. *Science* **257**, 502-507 (1992).
177 Wilks, D. & Dalgleish, A. G. Anti-idiotypic therapeutic strategies in HIV infection. *Molecular and cell biology of human diseases series* **1**, 283-308 (1992).
178 Bhattacharya-Chatterjee, M., Chatterjee, S. K. & Foon, K. a. Anti-idiotype vaccine against cancer. *Immunology letters* **74**, 51-58 (2000).
179 Weigle, W. O. The immune response of rabbits tolerant to bovine serum albumin to the injection of other heterologous serum albumins. *The Journal of experimental medicine* **114**, 111-125 (1961).
180 Warncke, M. *et al.* Control of the specificity of T cell-mediated anti-idiotype immunity by natural regulatory T cells. *Cancer Immunol Immunother* **60**, 49-60 (2011).
181 Ladjemi, M. Z. *et al.* Vaccination with human anti-trastuzumab anti-idiotype scFv reverses HER2 immunological tolerance and induces tumor immunity in MMTV.f.huHER2(Fo5) mice. *Breast Cancer Res* **13**, R17 (2011).
182 Mohanty, K. *et al.* Anti-tumor immunity induced by an anti-idiotype antibody mimicking human Her-2/neu. *Breast Cancer Res Treat* **104**, 1-11 (2007).
183 Chong, G. *et al.* Phase III trial of 5-fluorouracil and leucovorin plus either 3H1 anti-idiotype monoclonal antibody or placebo in patients with advanced colorectal cancer. *Ann Oncol* **17**, 437-442 (2006).
184 Park, H. J. & Neelapu, S. S. Developing idiotype vaccines for lymphoma: from preclinical studies to phase III clinical trials. *British journal of haematology* **142**, 179-191 (2008).
185 Weiner, L. M., Dhodapkar, M. V. & Ferrone, S. Monoclonal antibodies for cancer immunotherapy. *Lancet* **373**, 1033-1040 (2009).
186 Bryson, S. *et al.* Crystal structure of the complex between the F(ab)' fragment of the cross-neutralizing anti-HIV-1 antibody 2F5 and the F(ab) fragment of its anti-idiotypic antibody 3H6. *J Mol Biol* **382**, 910-919 (2008).
187 Burioni, R. *et al.* Anti-HIV-1 response elicited in rabbits by anti-idiotype monoclonal antibodies mimicking the CD4-binding site. *PLoS One* **3**, e3423 (2008).
188 Eigenbrot, C. *et al.* Structural insight into how an anti-idiotypic antibody against D3H44 (anti-tissue factor antibody) restores normal coagulation. *J Mol Biol* **331**, 433-446 (2003).
189 Kunert, R. E., Weik, R., Ferko, B., Stiegler, G. & Katinger, H. Anti-idiotypic antibody Ab2/3H6 mimics the epitope of the neutralizing anti-HIV-1 monoclonal antibody 2F5. *AIDS* **16**, 667-668 (2002).
190 Zanetti, M. Idiotypic regulation of autoantibody production. *Crit Rev Immunol* **6**, 151-183 (1986).
191 Jost, C. R. *et al.* Mammalian expression and secretion of functional single-chain Fv molecules. *J Biol Chem* **269**, 26267-26273 (1994).
192 Dietrich, G., Varela, F. J., Hurez, V., Bouanani, M. & Kazatchkine, M. D. Selection of the expressed B cell repertoire by infusion of normal immunoglobulin G in a patient with autoimmune thyroiditis. *Eur J Immunol* **23**, 2945-2950 (1993).

193 Jayne, D. R., Esnault, V. L. & Lockwood, C. M. Anti-idiotype antibodies to anti-myeloperoxidase autoantibodies in patients with systemic vasculitis. *J Autoimmun* **6**, 221-226 (1993).
194 Lundkvist, I. *et al.* Regulation of autoantibodies in inflammatory demyelinating polyneuropathy: spontaneous and therapeutic. *Immunol Rev* **110**, 105-117 (1989).
195 Lundkvist, I., van Doorn, P. A., Vermeulen, M. & Brand, A. Spontaneous recovery from the Guillain-Barre syndrome is associated with anti-idiotypic antibodies recognizing a cross-reactive idiotype on anti-neuroblastoma cell line antibodies. *Clin Immunol Immunopathol* **67**, 192-198 (1993).
196 Tzioufas, A. G. & Routsias, J. G. Idiotype, anti-idiotype network of autoantibodies: pathogenetic considerations and clinical application. *Autoimmun Rev* **9**, 631-633 (2010).
197 Shoenfeld, Y. Common infections, idiotypic dysregulation, autoantibody spread and induction of autoimmune diseases. *J Autoimmun* **9**, 235-239 (1996).
198 Shaoul, R. & Lerner, A. Associated autoantibodies in celiac disease. *Autoimmun Rev* **6**, 559-565 (2007).
199 Braden, B. C. *et al.* Crystal structure of an Fv-Fv idiotope-anti-idiotope complex at 1.9 A resolution. *J Mol Biol* **264**, 137-151 (1996).
200 Evans, S. V., Rose, D. R., To, R., Young, N. M. & Bundle, D. R. Exploring the mimicry of polysaccharide antigens by anti-idiotypic antibodies. The crystallization, molecular replacement, and refinement to 2.8 A resolution of an idiotope-anti-idiotope Fab complex and of the unliganded anti-idiotope Fab. *J Mol Biol* **241**, 691-705 (1994).
201 Bentley, G. A., Boulot, G., Riottot, M. M. & Poljak, R. J. 3-DIMENSIONAL STRUCTURE OF AN IDIOTOPE ANTI-IDIOTOPE COMPLEX. *Nature* **348**, 254-257 (1990).
202 Debret, G., Martel, A. & Cuniasse, P. RASMOT-3D PRO: a 3D motif search webserver. *Nucleic Acids Res* **37**, W459-464 (2009).
203 Konc, J. & Janezic, D. ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. *Nucleic Acids Res* **40**, W214-221 (2012).
204 L, D. S. Pymol. http://www.pymol.org (2002).
205 Jackson, C. M. Factor X. *Prog Hemost Thromb* **7**, 55-109 (1984).
206 Davie, E. W., Fujikawa, K. & Kisiel, W. The coagulation cascade: initiation, maintenance, and regulation. *Biochemistry* **30**, 10363-10370 (1991).
207 Greer, J. Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins* **7**, 317-334 (1990).
208 Bode, W., Brandstetter, H., Mather, T. & Stubbs, M. T. Comparative analysis of haemostatic proteinases: structural aspects of thrombin, factor Xa, factor IXa and protein C. *Thromb Haemost* **78**, 501-511 (1997).
209 Peyvandi, F. *et al.* Congenital factor X deficiency: spectrum of bleeding symptoms in 32 Iranian patients. *Br J Haematol* **102**, 626-628 (1998).
210 Herrmann, F. H. *et al.* Factor X deficiency: clinical manifestation of 102 subjects from Europe and Latin America with mutations in the factor 10 gene. *Haemophilia* **12**, 479-489 (2006).
211 Pinotti, M., Monti, M., Baroni, M., Marchetti, G. & Bernardi, F. Molecular characterization of factor X deficiency associated with borderline plasma factor X level. *Haematologica* **89**, 501-502 (2004).

212 Wang, W. B. *et al.* Molecular characterization of two novel mutations causing factor X deficiency in a Chinese pedigree. *Haemophilia* **11**, 31-37 (2005).
213 Isshiki, I. *et al.* Genetic analysis of hereditary factor X deficiency in a French patient of Sri Lankan ancestry: in vitro expression study identified Gly366Ser substitution as the molecular basis of the dysfunctional factor X. *Blood Coagul Fibrinolysis* **16**, 9-16 (2005).
214 Rao, L. V. & Rapaport, S. I. Activation of factor VII bound to tissue factor: a key early step in the tissue factor pathway of blood coagulation. *Proc Natl Acad Sci U S A* **85**, 6687-6691 (1988).
215 Peyvandi, F. *et al.* Gene mutations and three-dimensional structural analysis in 13 families with severe factor X deficiency. *Br J Haematol* **117**, 685-692 (2002).
216 Furie, B. *et al.* Computer-generated models of blood coagulation factor Xa, factor IXa, and thrombin based upon structural homology with other serine proteases. *J Biol Chem* **257**, 3875-3882 (1982).
217 Nazare, M. *et al.* Probing the subpockets of factor Xa reveals two binding modes for inhibitors based on a 2-carboxyindole scaffold: A study combining structure-activity relationship and X-ray crystallography. *Journal of Medicinal Chemistry* **48**, 4511-4525 (2005).
218 Nazare, M. *et al.* Probing the subpockets of factor Xa reveals two binding modes for inhibitors based on a 2-carboxyindole scaffold: a study combining structure-activity relationship and X-ray crystallography. *Journal of Medicinal Chemistry* **48**, 4511-4525 (2005).
219 Wallnoefer, H. G., Handschuh, S., Liedl, K. R. & Fox, T. Stabilizing of a globular protein by a highly complex water network: a molecular dynamics simulation study on factor Xa. *J Phys Chem B* **114**, 7405-7412 (2010).
220 Van der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **26**, 1701-1718 (2005).
221 Hornak, V. *et al.* Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712-725 (2006).
222 Jorgensen, W. L., Duffy, E. M. & Tiradorives, J. COMPUTATIONAL INVESTIGATIONS OF PROTEIN DENATURATION - APOMYOGLOBIN AND CHAOTROPE-ARENE INTERACTIONS. *Philosophical Transactions of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences* **345**, 87-96 (1993).
223 Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10037-10041 (2001).
224 DeLano, L. & Scientific. Pymol. http://www.pymol.org (2002).
225 Krawczak, M., Wacey, A. & Cooper, D. N. Molecular reconstruction and homology modelling of the catalytic domain of the common ancestor of the haemostatic vitamin-K-dependent serine proteinases. *Hum Genet* **98**, 351-370 (1996).
226 Akhavan, S. *et al.* Identification and three-dimensional structural analysis of nine novel mutations in patients with prothrombin deficiency. *Thromb Haemost* **84**, 989-997 (2000).
227 Millar, D. S. *et al.* Molecular analysis of the genotype-phenotype relationship in factor VII deficiency. *Hum Genet* **107**, 327-342 (2000).

228 Giannelli, F. *et al.* Haemophilia B: database of point mutations and short additions and deletions, fifth edition, 1994. *Nucleic Acids Res* **22**, 3534-3546 (1994).
229 Gandrille, S. & Aiach, M. Identification of mutations in 90 of 121 consecutive symptomatic French patients with a type I protein C deficiency. The French INSERM Network on Molecular Abnormalities Responsible for Protein C and Protein S deficiencies. *Blood* **86**, 2598-2605 (1995).
230 Ruddon, R. W. & Bedows, E. Assisted protein folding. *J Biol Chem* **272**, 3125-3128 (1997).
231 Sidrauski, C., Chapman, R. & Walter, P. The unfolded protein response: an intracellular signalling pathway with many surprising features. *Trends Cell Biol* **8**, 245-249 (1998).
232 Bianchini, E. P., Pike, R. N. & Le Bonniec, B. F. The elusive role of the potential factor X cation-binding exosite-1 in substrate and inhibitor interactions. *J Biol Chem* **279**, 3671-3679 (2004).
233 Chen, L., Manithody, C., Yang, L. & Rezaie, A. R. Zymogenic and enzymatic properties of the 70-80 loop mutants of factor X/Xa. *Protein Sci* **13**, 431-442 (2004).
234 Eisenmesser, E. Z. *et al.* Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**, 117-121 (2005).
235 Banerjee, D. & Pal, S. K. Conformational dynamics at the active site of alpha-chymotrypsin and enzymatic activity. *Langmuir* **24**, 8163-8168 (2008).
236 von Heijne, G. The membrane protein universe: what's out there and why bother? *J Intern Med* **261**, 543-557 (2007).
237 Forrest, L. R., Tang, C. L. & Honig, B. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys J* **91**, 508-517 (2006).
238 Heijne, G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *Embo J* **5**, 3021-3027 (1986).
239 von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225**, 487-494 (1992).
240 Choi, Y. & Deane, C. M. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins* **78**, 1431-1440 (2010).
241 Tusnady, G. E., Dosztanyi, Z. & Simon, I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res* **33**, D275-278 (2005).
242 Wang, G. & Dunbrack, R. L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591 (2003).
243 Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S. & Overington, J. P. JOY: protein sequence-structure representation and analysis. *Bioinformatics* **14**, 617-623 (1998).
244 Scott, K. A. *et al.* Coarse-grained MD simulations of membrane protein-bilayer self-assembly. *Structure* **16**, 621-630 (2008).
245 Sotriffer, C. A. *et al.* Automated docking of ligands to antibodies: methods and applications. *Methods* **20**, 280-291 (2000).
246 Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins* **65**, 392-406 (2006).

247 Mendez, R., Leplae, R., Lensink, M. F. & Wodak, S. J. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins* **60**, 150-169 (2005).
248 Gabb, H. A., Jackson, R. M. & Sternberg, M. J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* **272**, 106-120 (1997).
249 Chen, R., Mintseris, J., Janin, J. & Weng, Z. A protein-protein docking benchmark. *Proteins* **52**, 88-91 (2003).
250 Zacharias, M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci* **12**, 1271-1282 (2003).
251 Brenk, R., Vetter, S. W., Boyce, S. E., Goodin, D. B. & Shoichet, B. K. Probing molecular docking in a charged model binding site. *J Mol Biol* **357**, 1449-1470 (2006).
252 Grunberg, R., Leckner, J. & Nilges, M. Complementarity of structure ensembles in protein-protein binding. *Structure* **12**, 2125-2136 (2004).
253 Chaudhury, S. & Gray, J. J. Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. *J Mol Biol* **381**, 1068-1087 (2008).
254 Dobbins, S. E., Lesk, V. I. & Sternberg, M. J. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci U S A* **105**, 10390-10395 (2008).
255 Totrov, M. & Abagyan, R. Detailed ab initio prediction of lysozyme-antibody complex with 1.6 A accuracy. *Nat Struct Biol* **1**, 259-263 (1994).
256 Andrusier, N., Mashiach, E., Nussinov, R. & Wolfson, H. J. Principles of flexible protein-protein docking. *Proteins* **73**, 271-289 (2008).
257 Krippahl, L., Moura, J. J. & Palma, P. N. Modeling protein complexes with BiGGER. *Proteins* **52**, 19-23 (2003).
258 Heifetz, A. & Eisenstein, M. Effect of local shape modifications of molecular surfaces on rigid-body protein-protein docking. *Protein Eng* **16**, 179-185 (2003).
259 Najmanovich, R., Kuttner, J., Sobolev, V. & Edelman, M. Side-chain flexibility in proteins upon ligand binding. *Proteins* **39**, 261-268 (2000).
260 Segal, D. & Eisenstein, M. The effect of resolution-dependent global shape modifications on rigid-body protein-protein docking. *Proteins* **59**, 580-591 (2005).
261 Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93 (2004).
262 de Vries, S. J., van Dijk, A. D. & Bonvin, A. M. WHISCY: what information does surface conservation yield? Application to data-driven docking. *Proteins* **63**, 479-489 (2006).
263 van Dijk, A. D., Boelens, R. & Bonvin, A. M. Data-driven docking for the study of biomolecular complexes. *FEBS J* **272**, 293-312 (2005).
264 van Dijk, A. D. & Bonvin, A. M. Solvated docking: introducing water into the modelling of biomolecular complexes. *Bioinformatics* **22**, 2340-2347 (2006).
265 Brooks, B. R. *et al.* CHARMM: the biomolecular simulation program. *J Comput Chem* **30**, 1545-1614 (2009).
266 Brooks, B. & Karplus, M. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci U S A* **80**, 6571-6575 (1983).

267 Tzioufas, A. G. & Routsias, J. G. Idiotype, anti-idiotype network of autoantibodies: pathogenetic considerations and clinical application. *Autoimmunity reviews* **9**, 631-633 (2010).

LIST OF PUBBLICATIONS AND CONFERENCES

1. Articles on international journals:

Vangone A, Spinelli R, Scarano V, Cavallo L, Oliva R, "COCOMAPS: a web application to analyse and visualize contacts at the inter-face of biomolecular complexes" *Bioinformatics* 2011, 27(20):2915-2916.

Vangone A, Cavallo L, Oliva R, "A novel tool to measure and visualize the conservation of the inter-residue contacts in multiple docking solutions", *BMC Bioinformatics 2012,* Suppl 4:S19.

Kelm S, **Vangone A,** Choi Y, Ebjer JP, Shi J, Deane C, "Fragment-based modelling of membrane protein loops - successes, failures and prospects for the future". *Submitted.*

Menegatti M, **Vangone A,** Palla R, Milano G, Cavallo L, Oliva R, De Cristofaro R, Peyvandi F, "A recurrent GLY43ASP substitution of coagulation Factor X (FX) rigidifies the catalytic pocket and impairs both catalytic activity and intracellular trafficking causing severe bleeding". *Submitted.*

Oliva R, **Vangone A,** Cavallo L, "CONS-RANK: A tool to rank multiple docking solutions based on the conservation of inter-residue contacts". *Submitted*

Vangone A, Sblattero D, Caputo I, Cavallo L, Oliva R, "Prediction and analysis of idiotype - anti-idiotype antibody complex associated to celiac disease". *In preparation.*

Abdel-Azeim S, **Vangone A,** De Cristofaro R, Oliva R and Cavallo L, "The Effect of the Gly43Asp mutation on the catalytic activity of the coagulation Factor Xa from a molecular dynamics perspective". *In preparation.*

Micheloni S, **Vangone A**, Knew KL, Mantione E, Biggins P, Mazzaferro S, Bermudez I, "Agonist binding to β(+)/β(-) interface modulate the function of (α4β2)2β2 nAChR". *In preparation.*

2. Conference proceedings:

Vangone A, Sblattero D, Caputo I, Cavallo L, Oliva R: "Prediction And Analysis Of Idiotype-Anti-Idiotype Antibody Complexes Associated To Celiac Disease", 3D-SIG 2012: The 8th Structural Bioinformatics and Computational Biophysics Meeting, 13th-14th July 2011, Long Beach, California (USA).

Vangone A, Spinelli R, Scarano V, Cavallo L, Oliva R: "COCOMAPS: a novel tool to analyze and visualize contacts at the interface of biomolecular complexes", 3D-SIG 2011: the 7th Structural Bioinformatics and Computational Biophysics Meeting, 15th-16th July 2011, Vienna, Austria.

Vangone A, Oliva R, Cavallo L: "COCO Maps: A Web Application to Visualize Contacts at the Interface of Biomolecular Complexes", BITS2011: the 8th Meeting of the Italian Society of Bioinformatics, 20th-22th June 2011, Pisa, Italia.
ISBN: 978-884673069-5

Vangone A, Oliva R, Cavallo L: "CMPC: A Web Application to Visualize Residues-Residue Contacts at the Interface of Protein Complexes", NETTAB/BBCC 2010, 29th November – 1st December 2010, Napoli, Italia.

3. Poster or oral communications at Conferences:

Abdel-Azeim S, **Vangone A**, Oliva R, De Cristofaro R, Cavallo L: "Gly43Asp Mutation effect on the catalytic activity of the coagulation factor X: Molecular Dynamics and Metadynamics simulation studies", ECCB12: 11th European Conference on Computational Biology, 9-12 September 2012, Basilea, Svizzera. Poster.

Vangone A, Cavallo L, Oliva R: "A novel method to rank protein-protein docking solutions based on the conservation of inter-residue contacts", ECCB12: 11th European Conference on Computational Biology, 9-12 September 2012, Basilea, Svizzera. Poster.

Vangone A: "Novel a web application to analyze and visualize contacts at the interface of biomolecular complexes: COCOMAPS" NanoMeetsBio@Nanomates, 19th June 2012, Salerno, Italy. Oral communication.

Vangone A, Spinelli R, Scarano V, Cavallo L, Oliva R: "COCOMAPS and CONS-COCOMAPS: novel web tools for the analysis of crystallographic complexes and of multiple docking solutions" BCC2011: 6th Bioinformatica e Biologia Computazionale in Campania, 4th November 2011, Avellino, Italy. Oral communication.

Vangone A, Oliva R, Cavallo L: "Studio *in Silico* del Riconoscimento tra Transglutaminasi di Tipo 2 e Anticorpi Anti-Transglutaminasi Caratteristici della Malattia Celiaca" BBCC 2009: 4th Bioinformatics and Computational Biology in Campania", 13rd November 2009, Avellino, Italy. Oral communication.

4. Schools and other conferences:

"KAUST Workshop on Quantitative Biology: a matter of perspective", June 2011 Rome, Italy.

Workshop and Summer School "Optimization, Machine Learning and Bioinformatics", September 2010, Erice, Italy.

International School “CIOB 2010 – Neural Nets on Computational Intelligence Methods for Data Analysis in Oncology Bioinformatics” May 2012, Vietri sul Mare, Italy.

Winter School “9th Biosapiens European School of Bioinformatics” January 2009, Bruxelles, Belgium.

Conference “BBCC 2008 – Bioinformatics and Computational Biology in Campania” November 2008, Avellino, Italy.

ACKNOWLEDGEMENTS

During my PhD I had the great pleasure to work with people who have made useful collaborations to my work and I really would like to say “thank you” to all of them.

In particular, I would like to thank Prof. Vittorio Scarano and Raffaele Spinelli (University of Salerno), for their web graphic support; Prof. Daniele Sblattero (University of Piemonte Orientale), Dr. Ivana Caputo and Prof. Carla Esposito (University of Salerno) for the experimental support given in various projects; Prof. Raimondo De Cristofaro (Catholic University School of Medicine, Rome) for the helpful experimental collaboration; to Dr. Giuseppe Milano, not only for the interesting collaboration, but also for giving me useful suggestions during all of the PhD period.

Thanks to Prof. Charlotte Deane (University of Oxford) for the precious opportunity she gave to me to spend several months in her group, and to the colleagues of Oxford Protein Information Group. All of you made that months a wonderful working and personal experience.

A thank you to the wonderful people of the Molnac group for the daily sharing of the human and professional experiences. Andrea, Francesco, Antonio, Albert, Raffaele, Edita, Nahime, Ying, Laura, it was a great pleasure working with you.

Last, but not least, I would like to thank my two exceptional supervisors, Prof. Luigi Cavallo (University of Salerno) and Dr. Romina Oliva (University “Parthenope”, Naples), for always encouraging me, for the door I always found open in you and for the infinitive opportunity you gave me. The pages in this thesis are all thanks to you!