PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Great interactions: How binding incorrect partners can teach us about protein recognition and function

Lydie Vamparys,[1] Benoist Laurent,[1] Alessandra Carbone,[2,3] and Sophie Sacquin-Mora[1]*

[1] Laboratoire De Biochimie Théorique, CNRS UPR 9080, Institut De Biologie Physico-Chimique, 13 Rue Pierre Et Marie Curie, Paris 75005, France

[2] Sorbonne Universités, UPMC Univ-Paris 6, CNRS UMR7238, Laboratoire De Biologie Computationnelle Et Quantitative, 15 Rue De L'Ecole De Médecine, Paris 75006, France

[3] Institut Universitaire De France, Paris 75005, France

ABSTRACT

Protein–protein interactions play a key part in most biological processes and understanding their mechanism is a fundamental problem leading to numerous practical applications. The prediction of protein binding sites in particular is of paramount importance since proteins now represent a major class of therapeutic targets. Amongst others methods, docking simulations between two proteins known to interact can be a useful tool for the prediction of likely binding patches on a protein surface. From the analysis of the protein interfaces generated by a massive cross-docking experiment using the 168 proteins of the Docking Benchmark 2.0, where all possible protein pairs, and not only experimental ones, have been docked together, we show that it is also possible to predict a protein's binding residues without having any prior knowledge regarding its potential interaction partners. Evaluating the performance of cross-docking predictions using the area under the specificity-sensitivity ROC curve (AUC) leads to an AUC value of 0.77 for the complete benchmark (compared to the 0.5 AUC value obtained for random predictions). Furthermore, a new clustering analysis performed on the binding patches that are scattered on the protein surface show that their distribution and growth will depend on the protein's functional group. Finally, in several cases, the binding-site predictions resulting from the cross-docking simulations will lead to the identification of an alternate interface, which corresponds to the interaction with a biomolecular partner that is not included in the original benchmark.

Proteins 2016; 84:1408–1421.

Key words: protein–protein interaction; docking; protein–protein interfaces; binding sites prediction; coarse grain models.

INTRODUCTION

Because of their essential role in performing, coordinating and regulating the majority of cell activities, proteins are undeniably amongst the most fascinating and complex macromolecules in living systems. Over recent decades, numerous studies have been devoted to the molecular properties and functions of individual proteins. However, proteins often fulfil their roles through interactions, and are capable of forming large edifices that can act as complex molecular machines.[1] Transient interactions also form a complex protein network that controls these machines as well as a host of other cellular processes. As a consequence, protein–protein interactions (PPI) play a central role in biological systems,[2–4] defining the *interactome* of an organism, or, as elegantly expressed by Robinson *et al.* the *molecular sociology of the cell*.[5]

Many experimental approaches are used to investigate PPI, including yeast two-hybrid,[6,7] tandem affinity purification[8,9] and mass spectroscopy[10] (see ref. 11 for a general review). These approaches have enabled PPI maps to be established for various organisms, including

Additional Supporting Information may be found in the online version of this article.

Abbreviations: AUC, area under the curve; CC-D, complete cross-docking; NPV, negative predicted value; PDB, Protein Data Bank; PPI, protein-protein interactions; ROC, receiver operating characteristics

*Correspondence to: Sophie Sacquin-Mora, Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique,13 rue Pierre et Marie Curie, 75005 Paris, France. E-mail: sacquin@ibpc.fr

Received 29 January 2016; Revised 1 June 2016; Accepted 2 June 2016
Published online 10 June 2016 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.25086

yeast,[12] Escherichia coli,[13] *Drosophila melanogaster*,[14] *Caenorhabditis elegans*[15] and humans.[16] Experimental mapping of interactomes however suffers from several drawbacks. It involves expensive experiments and, despite continued progress, it still suffers from inaccuracies and generates significant numbers of false positives and negatives.[17–19] As a complementary approach to *in vitro* methods, several *in silico* methods have also been developed for predicting binary protein interactions. Many of them are based on protein sequence information, using gene clustering or phylogenetic profiling.[20–26] However, although these methods can predict interactions they do not provide any atomic-level information on the conformation of the complex or on the origins of its formation and stability. Another approach has been to develop so-called template methods, which predict interactions between pairs of proteins that are homologous (either globally, or at the interface region) to pairs of proteins within known binary complexes. These methods have achieved good results, but are naturally limited by the quality and coverage of the available template database.[27–29] Still, important aspects of PPI, such as the influence of the crowded cellular environment,[30–32] or the time-dependence of these networks,[33–35] cannot be addressed via any of these approaches.

Molecular modeling potentially offers an alternative route for identifying protein interactions, while at the same time providing structural models for the corresponding complexes and insight into the physical principles behind complex formation. In particular, the characterization on the molecular level of protein interfaces represents a key issue from a therapeutic point of view, since these PPI sites are potential targets for drugs designed to modulate or mimic their effects.[36] As a consequence, numerous interface prediction methods have been developed over the last years, which combine evolutionary and structural, and sometimes experimental, information.[37–40] In this perspective, docking methods, which were originally developed to predict the structure of a complex starting from the structures of two proteins that are known to interact,[41] are of specific interest. The collection of docking poses between two protein partners can be used to derive a consensus of predicted interface residues. Following the NIP (Normalized Interface Propensity) approach that was originally developed by Fernandez-Recio *et al.*,[42] docking calculations have been used for binding sites predictions both in simple docking studies,[43–47] i. e. studies involving only protein partners that are already known to interact, and complete cross-docking (CC-D) studies, which involve performing docking calculations on all possible protein pairs within a given dataset.[48,49]

After an early CC-D study on a reduced test set,[48] the computational power of the public World Community Grid, has allowed us to carry out CC-D calculations on the complete Mintseris Docking Benchmark 2.0,[50] which comprises 168 proteins forming 84 known binary complexes. A first analysis of the PPI energies and interfaces resulting from this large scale study (14,196 potential binary interactions) showed how the combination of docking and evolutionary information could improve partner identification within the benchmark.[51] This work focuses more specifically on the information that can be obtained using the *PIP* (Protein Interface Propensity) value regarding the protein binding sites, which does not necessitate any prior knowledge regarding a protein's interaction partner. In addition, we developed a clustering algorithm for residues with a high interface propensity, that is, residues that are the most commonly found in protein interfaces resulting from docking calculations. This new approach shows how the distribution of potential binding patches on a protein surface will depend on the functional category this protein belongs to. Finally, we highlight several cases where cross-docking calculations will lead to the identification of alternate protein binding sites corresponding to interaction partners that are not included in the original dataset.

MATERIALS AND METHODS

Cross-docking calculations

In this section, we describe the MAXDo (Molecular Association via Cross Docking) algorithm that was developed for CC-D studies.[48] Since CC-D involves a much larger number of calculations than simple docking, we chose a rigid-body docking approach using a reduced protein model in order to make rapid conformational searches.

Protein dataset

All simulations were performed using the unbound conformations of the proteins from the Docking Benchmark 2.0 of Mintseris *et al.*[50] with the exception of 12 antibodies for which the unbound structure is unavailable and the bound structure was used instead. Any further reference to these proteins uses their name, or the Protein Data Bank (PDB) code[52] of the experimental complex they belong to with the *r* or *l* extension denoting a receptor or a ligand protein respectively. For example, 1AY7_r and 1AY7_l refer to barnase (receptor) and barstar (ligand) in the barnase-barstar complex 1AY7. The coordinates for the bound and unbound structures of both receptor and ligand proteins are available in the PDB and can also be found at the following address: http://zlab.bu.edu/zdock/benchmark.shtml. The 84 binary complexes listed in the Docking Benchmark 2.0 cover three broad biochemical categories and three difficulty categories related to the degree of conformational change in the protein-protein interface upon complex formation. They are classified as Enzyme-Inhibitors (E, 23 complexes, enzymes with a *r* extension and inhibitors with a

Table I
Interface Residues Prediction

Protein dataset	AUC	Err_{min}	PIP_{min}	Cov.	Sen.	Spec.	Prec.	Rand. Prec.
Complete benchmark (48161 residues)	0.77	0.29	0.09	32%	71%	71%	17%	11%
Enzymes (6757 residues)	0.73	0.33	0.12	28%	60%	76%	20%	13%
Inhibitors (2589 residues)	0.77	0.28	0.18	37%	73%	70%	35%	22%
Antigens (6669 residues)	0.72	0.33	0.10	28%	61%	74%	15%	9%
Antibodies (3873 residues)	0.90	0.16	0.11	18%	82%	87%	30%	8%
Bound-antibodies (4442 residues)	0.85	0.20	0.11	20%	78%	84%	24%	8%
Others (23828 residues)	0.75	0.30	0.08	35%	72%	68%	15%	10%

Results of the interface residues prediction using the *PIP* index for the complete benchmark, or depending on the protein's biochemical type. All values in the *Cov.*, *Sen.*, *Spec.* and *Prec.* columns are obtained with the optimal PIP_{min} value (column 4) which corresponds to the minimum error in column 3. The *Random Precision* column on the far right gives the ratio of experimental interface residues over the surface residues.

l extension after the PDB code), Antibody-Antigen (A, 10 complexes, antibodies with a *r* extension and antigens with a *l* extension after the PDB code), Bound Antibody-Antigen (AB, 12 complexes, bound-antibodies with a *r* extension and antigens with a *l* extension after the PDB code) and Others (O, 36 complexes). Note that for three cases in the AB category, namely 1IR9, 1KXQ and 2HMI, there was an inversion in the pdb files names, the antigen protein has a *r* extension and the bound antibody has a *l* extension after the PDB code.

Reduced protein representation

We use a coarse-grain protein model developed by Zacharias,[53] where each amino acid is represented by one pseudoatom located at the Cα position, and either one or two pseudoatoms representing the side-chain (with the exception of Gly). Ala, Ser, Thr, Val, Leu, Ile, Asn, Asp, and Cys have a single pseudoatom located at the geometrical center of the side-chain heavy atoms. For the remaining amino acids, a first pseudoatom is located midway between the Cβ and Cɣ atoms, while the second is placed at the geometrical center of the remaining side-chain heavy atoms. This description, which allows different amino acids to be distinguished from one another, has already proved useful both in protein–protein docking[53–55] and protein mechanics studies.[56–58]

Interactions between the pseudoatoms of the Zacharias representation are treated using a soft LJ-type potential with appropriately adjusted parameters for each type of side-chain, see Table I in Ref. 53. In the case of charged side-chains, electrostatic interactions between net point charges located on the second side-chain pseudoatom were calculated by using a distance-dependent dielectric constant $\varepsilon = 15r$, leading to the following equation for the interaction energy of the pseudoatom pair *i,j* at distance r_{ij}:

$$E_{ij} = \left(\frac{B_{ij}}{r\,ij^8} - \frac{C_{ij}}{r\,ij^6}\right) + \frac{q_i q_j}{15r\,2_{ij}}$$

where B_{ij} and C_{ij} are the repulsive and attractive LJ-type parameters respectively, and q_i and q_j are the charges of the pseudoatoms *i* and *j*.

Systematic docking simulations

Our systematic docking algorithm (see Supporting Information Fig. S1) is derived from the ATTRACT protocol[53] and uses a multiple energy minimization scheme. For each pair of proteins, the first molecule (called the receptor) is fixed in space, while the second (termed the ligand) is used as a probe and placed at multiple positions on the surface of the receptor. The initial distance of the probe from the receptor is chosen so that no pair of probe-receptor pseudoatoms comes closer than 6 Å. Starting probe positions are randomly created around the receptor surface with a density of one position per 10 $Å^2$, and for each starting position, 210 different ligand orientations are generated, resulting in a total number of start configurations ranging from roughly 100,000 to 450,000 depending on the size of the receptor.

During each energy minimization, the ligand protein is kept at a given location over the surface of the receptor protein, using a harmonic restraint to maintain its center of mass on a vector passing through the center of mass of the receptor protein. The direction of this vector is defined by two Euler angles θ and φ, (where $\theta = \varphi = 0°$ was chosen to pass through the center of the binding interface of the receptor protein) as shown in Supporting Information Figure S1. By using a Korobov grid[59] and varying the Euler angles from 0°→360° and 0°→180° respectively, it is possible to uniformly sample interactions over the complete surface of the receptor and to represent its binding potential using 2D energy maps (each point corresponding to the best ligand orientation for the chosen θ/φ pair). These maps where developed during the first phase of this project for validating the docking algorithm.[48]

Computational implementation

Each energy minimization for a pair of interacting proteins typically takes 5 s on a single 2 GHz processor. As noted above, approximately 1,00,000 to 4,50,000 minimizations are needed to probe all possible interaction conformations, as a function of the size of the interacting proteins. Therefore, a CC-D search on the benchmark, namely 168 × 168 = 28,224 receptor/ligand pairs, would

require several thousand years of computation on a single processor. However, since each minimization is independent of the others, this problem belongs to the "embarrassingly parallel" category and is well adapted to multiprocessor machines, and particularly to grid-computing systems. In the present case, our calculations have been carried out using the public *World Community Grid* (WCG, www.worldcommunitygrid.org) during the first phase of the Help Cure Muscular Dystrophy (HCMD) project. It took approximately six months to perform CC-D calculations on the complete dataset of 168 proteins. More technical details regarding the execution of the program on WCG can be found in Ref. [60]. The resulting CC-D data is available for download at the following address: http://www.lcqb.upmc.fr/CCDMintseris/

DATA ANALYSIS

Definition of surface and interface residues

Surface residues have a relative solvent accessible surface area larger than 5%. The accessibility is calculated with the NACCESS program,[61] using a 1.4 Å probe. Interface residues present at least a 10% decrease of their accessible surface area in the protein bound structure compared to the unbound form.

Interface propensity of the surface residues

In order to see whether cross-docking simulations can give us information regarding protein interaction sites, we use the interaction propensity approach initially developed by Fernandez-Recio *et al.*[42] That is, we count the number of docking hits for each surface residue r_i in protein P_1, that is, the number of times each surface residue belongs to a docked interface between P_1 and all its interaction partners in the benchmark. In earlier works,[48,51] we used a Boltzmann weighting factor which would favor docked interfaces with low energies. As a consequence, for a given protein pair P_1P_2, all interfaces with a 2.7 kcal.mol^{-1} or more energy difference from the lowest energy docked interface would have a Boltzmann weight lower than 1% (see ref. 51 for more details). Here, in order to limit the number of docked interfaces that would have to be reconstructed for determining the interface residues, which is the time consuming part of the analysis process, we chose to calculate the residues *PIP* (Protein Interface Propensity) values using only the lowest energy docking poses within this 2.7 kcal.mol^{-1} energy criterion, therefore we have

$$PIP_{P_1P_2}(i) = \frac{N_{\mathrm{int},P_1P_2}(i)}{N_{\mathrm{pos},P_1P_2}}$$

where $N_{pos,P1P2}$ is the number of retained docking poses of P_1 and P_2 (which will vary with protein P_2) satisfying the energy criterion, and $N_{int,P1P2}(i)$ is the number of these conformations where residue i belongs to the binding interface. Finally, the *PIP* value for a given residue i belonging to protein P_1 taking into account the CC-D calculations within the whole benchmark will simply be the average *PIP* of this residue over all the possible partner proteins P_2, that is

$$PIP_{P_1}(i) = \langle PIP_{P_1P_2}(i) \rangle_{p_2}$$

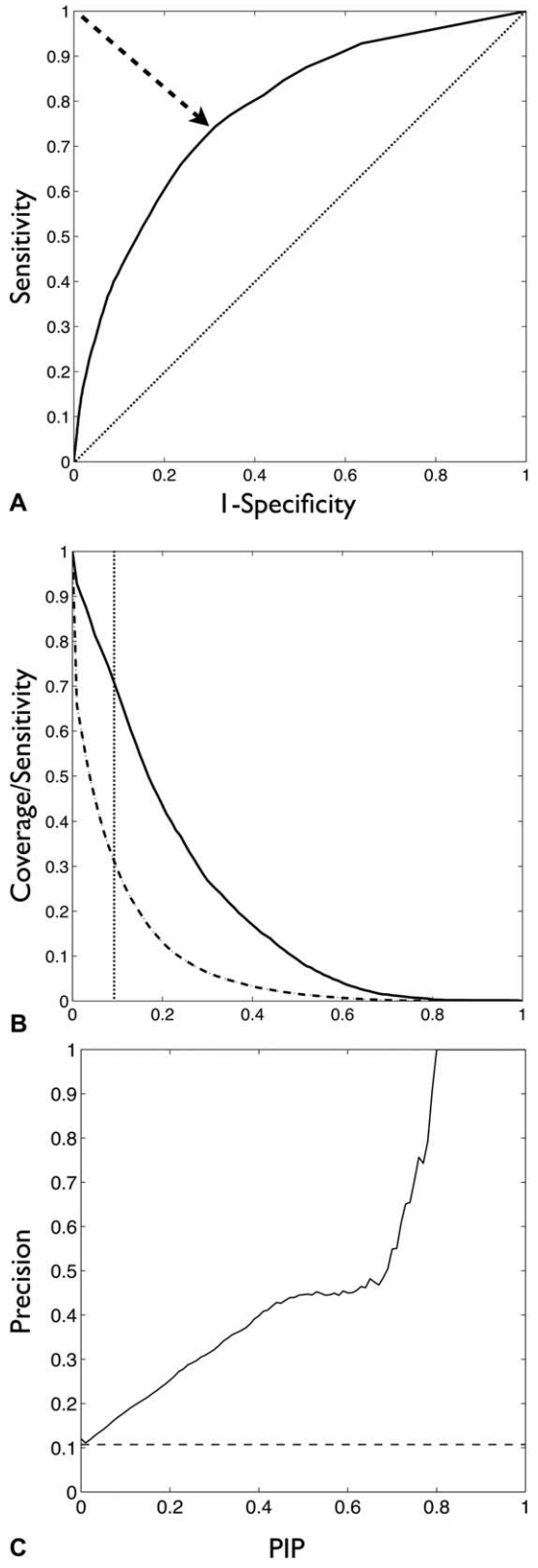
PIP values are comprised between 0 (the residue does not appear in any docked interface) and 1 (the residue is present in every single docked interface involving protein P_1) and will be used for the prediction of binding sites. For each protein pair in the benchmark, between 1 and 215 docking poses were kept using the 2.7 kcal.mol^{-1} energy criterion, with an average of 11 docking poses (see Supporting Information Fig. S2a in the supplementary material for the distribution of the number of conserved poses for each protein pair), and for 60% of the protein pairs, five docking poses or less are kept after filtering on the interaction energy. These low statistics on each individual protein pair are compensated by the fact that every protein was docked with 168 different partners. Eventually, for each protein in the benchmark, between 900 and 4000 docking poses were used for to calculate the residues PI*P* values (see Supporting Information Fig. S2b).

Evaluation of the binding sites prediction

Considering the *PIP* values results for all the residues, we define as predicted interface residues, residues whose *PIP* value lies above a chosen cutoff, and we can use the classical notions of sensitivity, specificity and the error function to evaluate their efficiency for the identification of protein interaction sites. Sensitivity (Sen.) is defined as the number of surface residues that are correctly predicted as interface residues (true positives, TP) divided by the total number of experimentally identified interface residues in the set (T). Specificity (Spe.) is defined as the fraction of surface residues that do not belong to an experimental protein interface and that are predicted as such (true negatives, TN). Additional useful notions that are commonly used include the positive predicted value (PPV, also called precision, Prec.), which is the fraction of predicted interface residues that are indeed experimental interface residues (TP/P), and the negative predicted value (NPV), which is the fraction of residues that are not predicted to be in the interface and which do not belong to an experimental interface (TN/N).

An optimal prediction tool would have all notions (Sen., Spe., Prec. and NPV) equal to unity. If this cannot be achieved, a compromise can obtained by minimizing a normalized error function based on the sensitivity and specificity values, which is comprised between 0 and 1 and defined as:

$$Norm.Err. = \sqrt{(1-Sen.)^2 + (1-Spe.)^2}/\sqrt{2}$$



On a classic receiver operating characteristics (ROC) curve (with the sensitivity plotted as a function of 1-specificity) the minimum error corresponds to the point on the curve that is the farthest away from the diagonal (which corresponds to random prediction).

Clustering surface residues

In order to visualize how binding patches composed of residues with the highest *PIP* values can form on a protein surface, we use the following clustering algorithm: for a given protein, its surface residues are ordered following their *PIP* value. Starting with the residue with the largest *PIP*, each surface residue leads to the creation of a new isolated cluster, or, if any of its heavy atoms is <5 Å away from the heavy atoms of a residue already included in a cluster, is added to an already existing cluster. This process is implemented as long as the average *PIP* value of every cluster is larger than a given threshold named PIP_{clust}. PIP_{clust} is a relative value, and is expressed as a percentage of the maximum *PIP* value found on a single protein surface. For $PIP_{clust} = 100\%$ only the residue with the largest *PIP* value on this specific surface is selected, while for $PIP_{clust} = 0\%$ all surface residues are selected and form a single cluster covering the whole protein's surface. We define a protein's clustering profile as the curve showing the number of clusters on its surface as a function of the PIP_{clust} criterion.

RESULTS

We must recall that, since the point of this work is to investigate the general binding behavior of protein surfaces with no prior knowledge of the binding partners, and not the correct docking of experimentally known partners, which can be achieved via other more effective but much more computationally demanding methods,[62] we did not evaluate the quality of the best structural predictions for the 84 docked complexes. However, in an earlier work,[48] where we performed cross-docking simulations on a limited test-set involving 12 proteins (using their bound structures), our method was able to predict correctly the position of the ligand protein with respect to its receptor with an rsmd of the Cα

Figure 1

(**a**) ROC curve of the *PIP* prediction. The diagonal dotted line corresponds to random predictions. The dashed arrow indicates the lowest error point. (**b**) Enrichment of the interface residues from the 168 proteins in the Benchmark 2.0[50] using the *PIP* index shown by comparing the fraction of true interface residues detected (sensitivity, solid line) with the total fraction of residues detected (coverage, dashed line) as a function of the *PIP* cutoff. The vertical dotted line corresponds to the position of the optimal *PIP* cutoff leading to the minimal error function shown in Figure 1a. (**c**) Precision as a function of the *PIP* cutoff. The dashed horizontal line corresponds to random predictions.

pseudoatoms below 3 Å; thus validating the quality of the force-field used in the MAXDo program. Furthermore, this force-field, which was originally developed by Zacharias for protein-protein docking,[53] has been successfully used on numerous occasions for the prediction of protein complex structures, especially during the CAPRI contest where the unbound structures of the protein partners are used.[54,63,64]

Identification of protein interaction sites

Figure 1(a) gives us a quantitative view of the results using the sensitivity and specificity notions defined in the Methods section. The ROC curve for the complete dataset (which contains >48,000 protein residues) with the variation of the sensitivity and specificity of the predictions is plotted in Figure 1(a), while Figure 1(b) shows the selection of residues potentially belonging to a protein interface as a function of a *PIP* cutoff comprised between 0 and 1. In Figure 1(a), the dotted diagonal corresponds to a random sampling of surface residues and divides the ROC space into areas of correct (above the diagonal) and incorrect (below the diagonal) classification. The greatest distance between the ROC curve and the diagonal yields the lowest error estimate, which is indicated by a dashed arrow in Figure 1(a). With a minimum error of 0.29 [see Supporting Information Fig. S3(a)], the optimal *PIP* cutoff [0.09, vertical dotted line in Fig. 1(b)] enables us to select 32% of the residues with a sensitivity of 71% and a specificity of 71% for the interface residues. These results are comparable to those that were obtained in our previous work cross-docking the unbound structures of 12 proteins from a reduced dataset and which led to 30% coverage and 70% and 75% sensitivity and specificity, respectively. To measure the efficiency of our predictions independently of the cutoff value, we can also use the area under the curve (AUC) value, which is 0.77 for the black line in Figure 1(a) (while random predictions would yield an AUC value of 0.50). Finally, Figure 1(c) shows the variation of the precision with the *PIP* cutoff. Experimental interface residues represent only 11% of all the surface residues from our data set, which corresponds to the random precision value that would be obtained with a *PIP* cutoff set to 0. Increasing the *PIP* cutoff will increase the precision and the specificity of the prediction while decreasing the sensitivity. For example, setting the *PIP* cutoff to 0.4 would lead to a 40% precision, 98% specificity, but only 17% sensitivity.

Influence of conformational changes upon binding

Conformational changes upon binding usually define the difficulty level for docking experimentally identified complexes. The Docking Benchmark 2.0 comprises three groups labelled as *rigid* (63 complexes), *medium* (13 complexes) and *difficult* (8 complexes), which present average RMSDs of the Cα atoms from the interface residues of

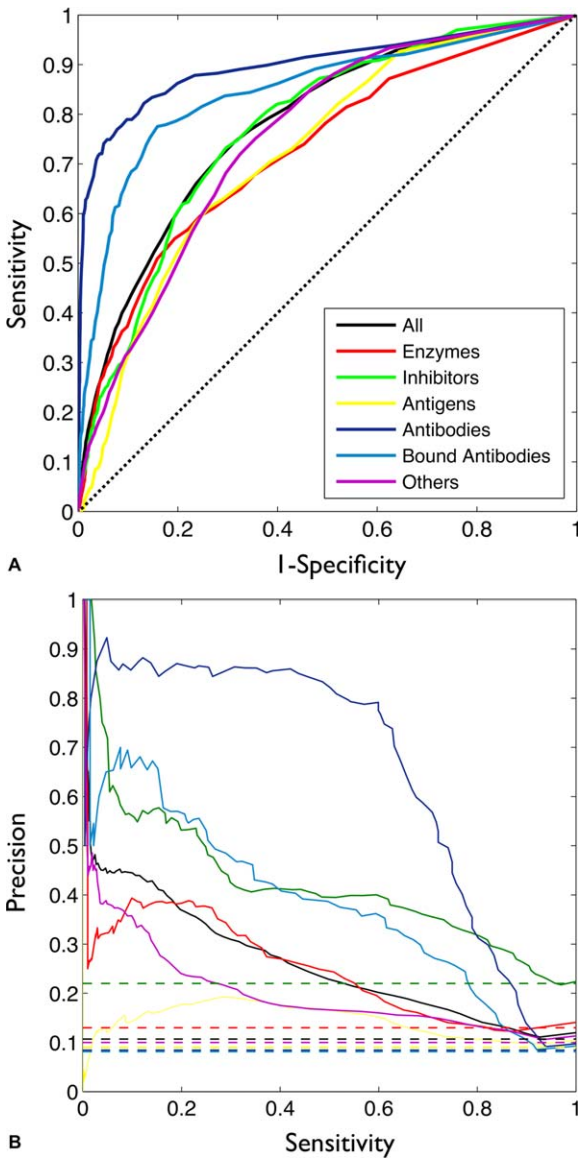


Figure 2

PIP predictions depending on the protein biochemical type with the same color code in both panels. (**a**) ROC curves, the diagonal dotted line corresponds to random predictions. (**b**) Precision/Sensitivity curves, the dashed horizontal lines correspond to random predictions for each biochemical group.

0.82 Å, 1.63 Å and 3.67 Å, respectively. Interestingly, the prediction of binding interfaces using the *PIP* values performs slightly better for the *medium*, than for the *rigid* and *difficult* groups with average AUCs of 0.78 for the *medium* group, and 0.77 for the *rigid* and *difficult* groups respectively (see Supporting Information Table S1).

Influence of the protein biochemical type for protein interface prediction

Using the biochemical categories discussed above (Enzymes, Inhibitors, Antigens, AntiBodies, Bound

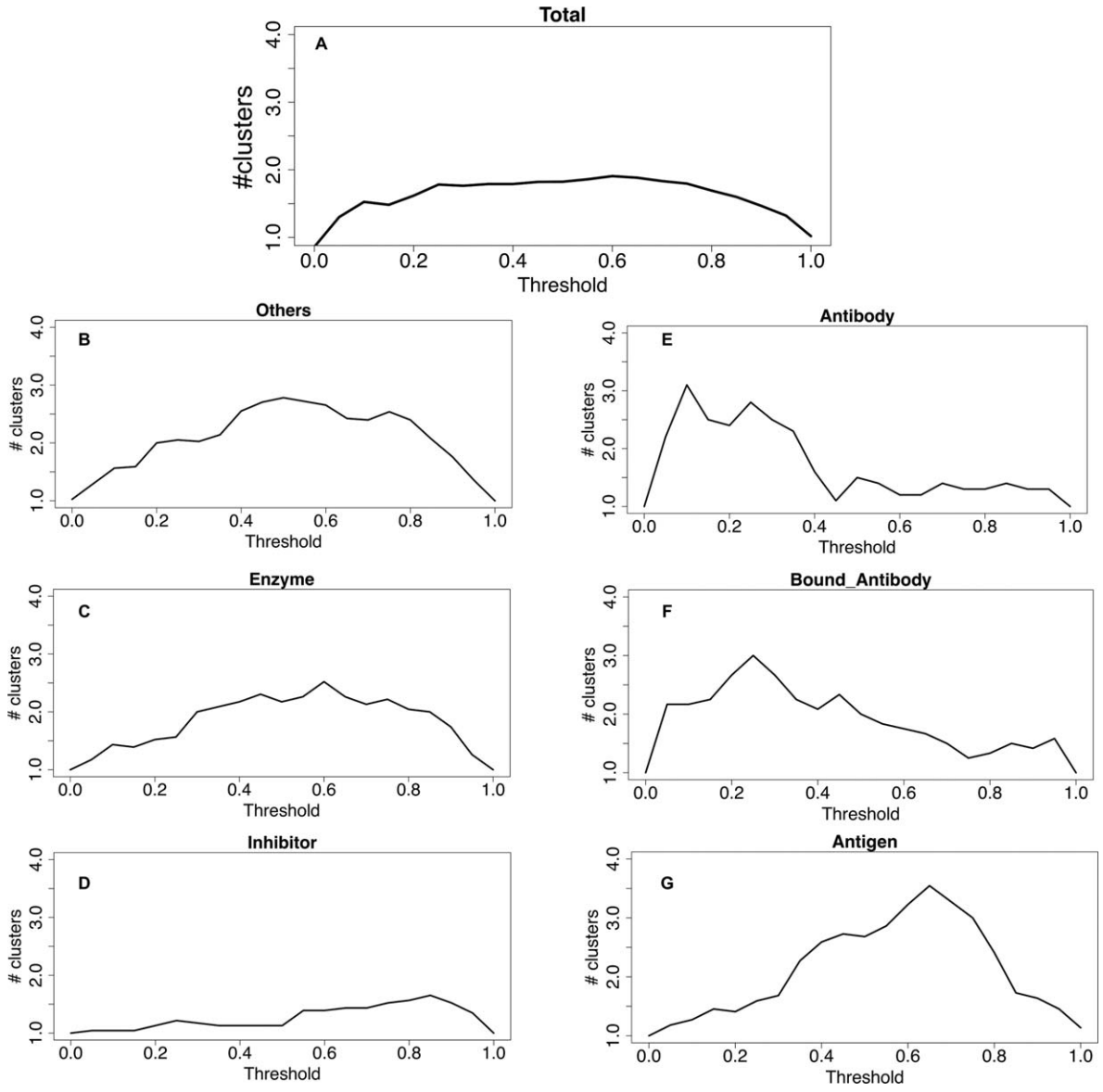


Figure 3

Average number of binding clusters on the protein surface as a function of the PIP_{clust} threshold (a) Complete benchmark, (b) Others, (c) Enzymes, (c) Inhibitors, (e) Antibodies, (f) Bound antibodies, (g) Antigens.

AntiBodies and Others) (see Table I), we obtained the ROC curves of the *PIP* predictions for each of these subgroups [see Fig. 2(a)]. We can observe noticeable variations in the quality of the predictions depending on the protein type. While the curves for proteins from the Others group (purple line) and inhibitors (green line) stay close to the overall curve (black line), the predictions for enzymes and antigens interfaces appear to be much less effective, with ROC curves (in red and yellow respectively) below the overall curve, an increased minimum error function (0.33 for both groups, compared to 0.29) and decreased AUCs of 0.73 and 0.72 respectively. In contrast, interfaces from antibodies (both in the unbound, AB, and bound, BAB, groups), are better predicted than average, with ROC curves (dark and light blue lines) above the overall one, reduced minimum errors of 0.16 and 0.20 respectively, and increased AUCs of 0.90 and 0.85 respectively.

Instead of using the classic *Sen./(1-Spe.)* ROC curve, one can also evaluate the performance of the *PIP* predictions with the *Precision/Sensitivity* (Recall) curve like what has been done by Hwang *et al.* for the prediction of binding interfaces from simple docking calculations.[47] In that case the random AUC is no longer a constant and will depend on the proportion of experimental interface residues in each dataset, which corresponds to the

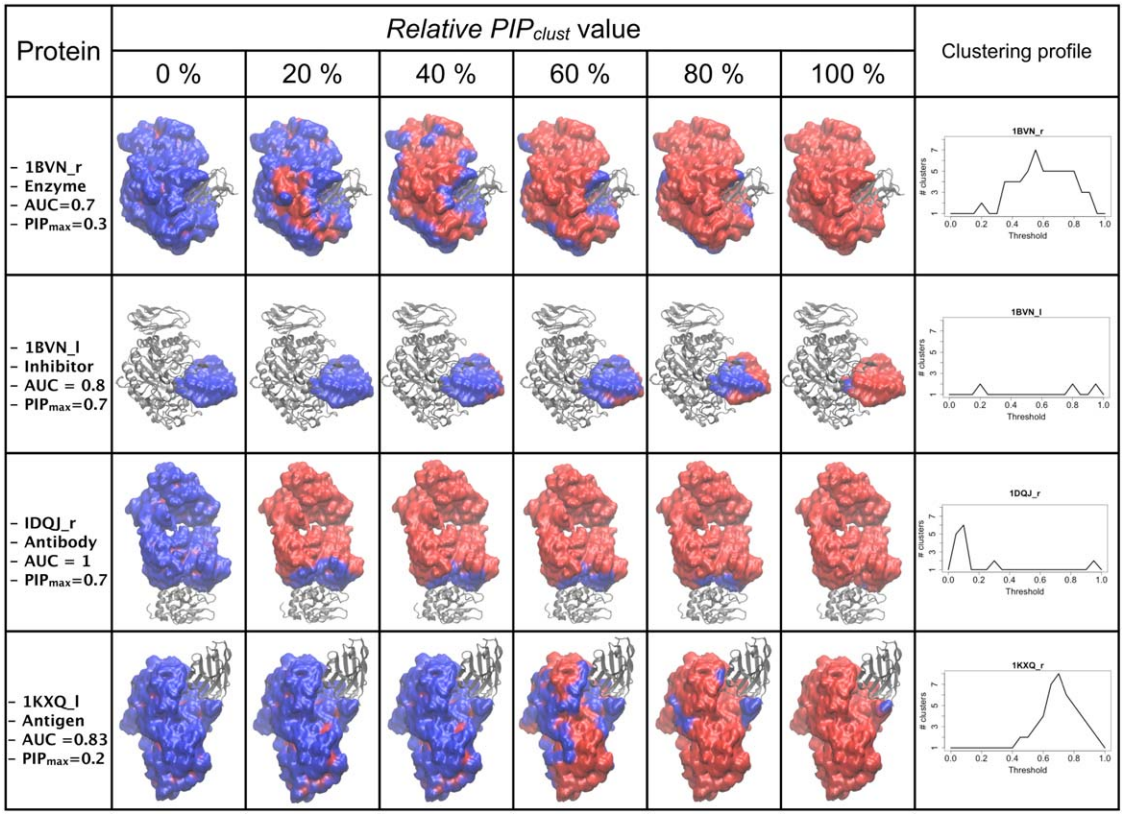


Figure 4

Evolution of the binding clusters (shown in blue) growth on a protein surface (red background) as a function of the PIP_{clust} threshold for representative cases of different functional categories. The column on the far right shows the individual clustering profile for each protein.

Random Precision column on the far right of Table I. Figure 2(b) shows the *Precision/Sensitivity* curves for the complete benchmark (in black) and all the functional subgroups. Once again we can observe a striking difference between the antigens (yellow line), with predictions that are barely above the random line, and the antibodies (dark blue), which present a curve that is markedly above all the others.

Clustering binding patches on the protein surface

Using the clustering algorithm presented in the Material an Methods section, we can plot an average clustering profile for the complete benchmark, see Figure 3(a). As expected, for a maximum threshold $PIP_{clust} = 1$, we find a single cluster formed by the residue with the largest *PIP* value on the protein surface. As the threshold decreases, the number of clusters will either increase, when high-*PIP* residues are scattered on different parts of the surface, or decrease, since adding a residue that belongs to two disjoint clusters will lead to their merging and the formation of a single cluster. The average number of clusters on the ensemble of protein surfaces will pass through a maximum value $\langle N_{clust} \rangle_{max} = 2$ for $PIP_{clust} = 60\%$, before decreasing and reaching a final value $\langle N_{clust} \rangle = 1$ for $PIP_{clust} = 0\%$, where all surface residues are selected. Figure 4 presents the progressive binding clusters growth on the protein surface for representative proteins from the enzyme, inhibitor, antibody and antigen categories. For example, reading the first line on Figure 4 from right to left shows how binding clusters (in blue), will increase in size and number on the surface of the α-amylase enzyme (1BVN_r) upon decreasing the PIP_{clust} threshold value. As could be expected, there are important variations in the proteins individual cluster profiles, in particular regarding the maximum number of clusters that can be found on the protein surface, which is comprised between 1 and 21 and appears to be strongly correlated with the proteins size (see Supporting Information Fig. S4). The position of the PIP_{clust} threshold corresponding to this maximum will also vary widely from one protein to the other.

Influence of the protein biochemical type on the clustering profile

Figure 3(b–g) show the average clustering profiles for the six biochemical categories that are represented in the

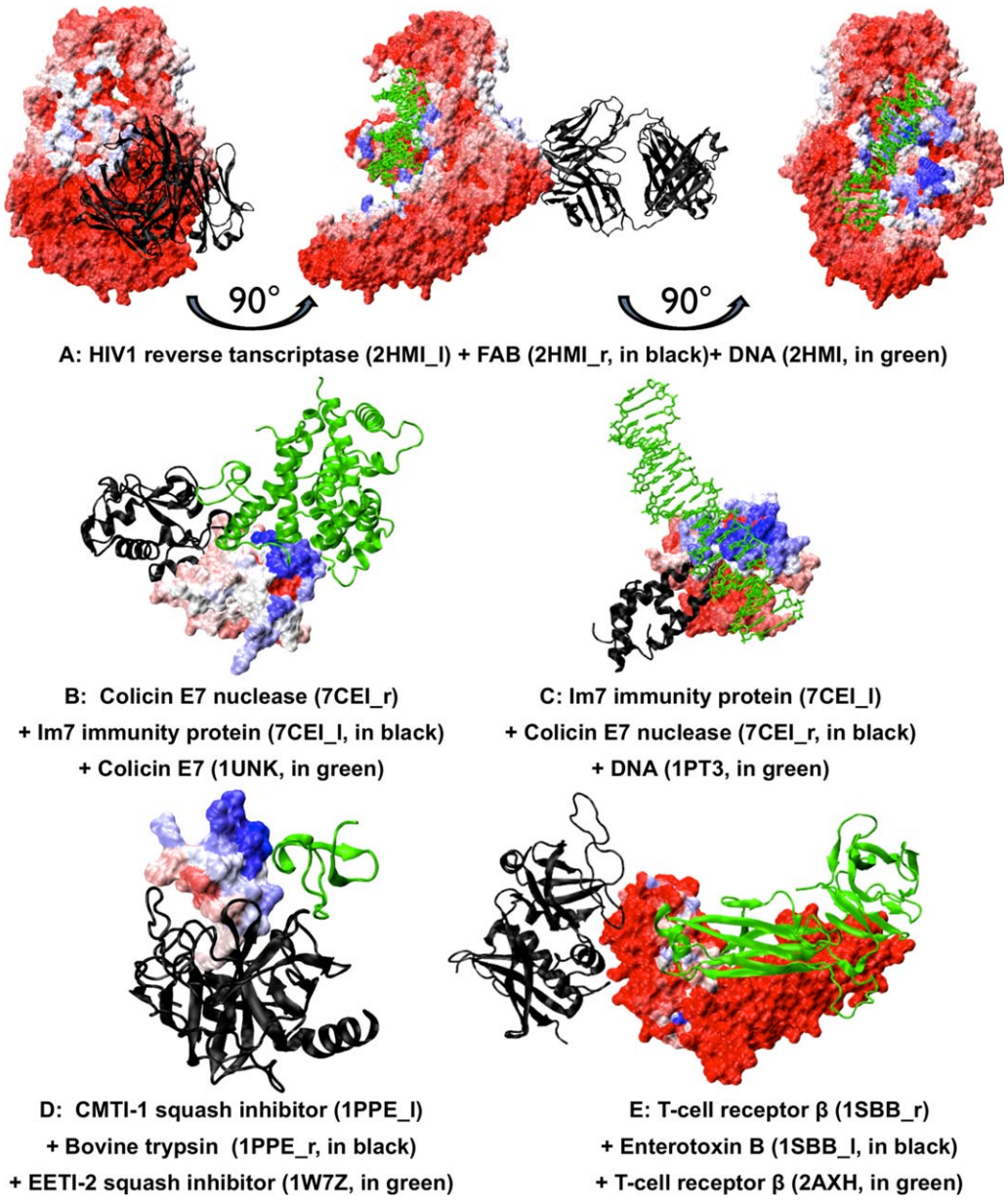


Figure 5

Mapping the *PIP* values on a protein's surface, high *PIP* residues are shown in blue and low *PIP* residues are shown in red. The reference experimental partner is shown in a black cartoon representation, while the alternate partner is shown in green. (**a**) HIV1 reverse transcriptase (2HMI_l), with FAB28 (2HMI_r) and DNA (2HMI). (**b**) Colicin E7 nuclease (7CEI_r), with Im7 immunity protein (7CEI_l) and Colicin E7 (1UNK). (**c**) Im7 immunity protein (7CEI_l), with Colicin E7 nuclease (7CEI_r) and DNA (1PT3). (d) CMTI-1 squash inhibitor (1PPE_l), with Bovine trypsin (1PPE_r) and EETI-2 squash inhibitor (1W7Z). (e) T-cell receptor β (1SBB_r), with Enterotoxin B (1SBB_l) and T-cell receptor β (2AXH).

benchmark. As seen in Figure 3(b,c), proteins from the others and enzymes categories present an average profile that is similar to the general cluster profile from Figure 3(a), while inhibitors [Fig. 3(d)] have a somewhat flatter profile, with a maximum cluster number under 2. This is probably due to the fact that, since inhibitors represent the smallest proteins on the benchmark, their surface is only large enough for two clusters at most, which will rapidly merge into a single one as the PIP_{clust} criterion decreases. On the second line of Figure 4 we can see for example how the tendamistat inhibitor (1BVN_l) only presents one or two binding clusters, which will cover most of its surface at an early stage (PIP_{clust} around 60% and under).

On the other hand, the average profiles for proteins from the antibody [bound, Figure 3(e), or unbound, Fig. 3(f)] and antigen [Fig. 3(g)] categories are very specific. Antibodies usually present a single dominant cluster that

will grow on the protein surface, and alternative binding patches will only appear at a later stage for low values of the PIP_{clust} criterion (under 40%). As an illustration, the third line of Figure 4 shows the binding clusters growth in the case of the FAB Hyhel63 antibody (1DQJ_r), with a very late coverage of the protein surface. On the opposite, antigens will often present various disjoint binding clusters on their surface for high values of the PIP_{clust} criterion (around 70%), which will afterwards merge in a single binding patch as the PIP_{clust} criterion keeps on decreasing, see for example the fourth line in Figure 4 with the case of the pancreatic α-amylase antigen (1KXQ_l). The individual clustering profiles for all the proteins in the dataset are available in the Supporting Information Figure S5.

When false positives turn out to be true positives: the case of proteins with multiple interfaces

Promiscuity, or multispecificity, is a frequent feature in the protein world[65,66] and a key issue when trying to predict proteic interfaces, since many proteins can actually present multiple binding sites, even when interacting with a single partner.[67] This is also the case for proteins from the docking benchmark 2.0, and the apparent failure of cross-docking for predicting experimental binding sites can result from the detection of alternative interfaces formed by the protein with other biomolecular partners. For example, Figure 5(a) presents the case of HIV1 reverse transcriptase (2HMI_r). The prediction of the binding site between this antigen and the FAB 28 antibody (2HMI_l) leads to a AUC value of 0.71, which is slightly smaller than the average AUC obtained for the prediction of interface residues in antigens (see Table I). Coloring the antigen's surface using the PI*P* values resulting from cross-docking, we can see that we only have a weak binding signal [Fig. 5(a), left panel, in white] for the antigen/antibody binding site. However, the 2HMI complex also comprises a DNA double strand that binds HIV1 reverse transcriptase on the opposite side of the antibody binding site [Fig. 5(a), center panel]. As we can see in the right panel from Figure 5(a), the antigen surface residues surrounding the DNA double strand present high PI*P* values (in blue), which means that our procedure can lead to the prediction of binding sites with a molecular partner (in that case DNA) that do not belong to the benchmark used for the cross-docking calculations. We used the list of alternate interfaces and partners that has been established by Martin and Lavery[49] for proteins from the Docking Benchmark 4.0[68] (which comprises many complexes from the 2.0 version) and found out several other cases of proteins with a low AUC value where the binding patches formed by surface residues with high *PIP*-values correspond to experimental interfaces with alternate biomolecular partners. Figure 5(b,c) show how cross-docking lead to the prediction of an alternate protein binding site for then enzyme colicin E7 nuclease (7CEI_r) and a DNA binding-site for its inhibitor, the Im7 immunity protein (7CEI_l). More alternate interfaces are shown in Supporting Information Figure S6 with the references partner and the alternate partner systematically shown in black and green, respectively. Using the PiQSi webserver,[69] we also searched for alternate experimental interfaces for the two proteins in the benchmark which presented the poorest binding site predictions (as expressed by their AUC value, see Supporting Information Fig. S7), namely the CMTI-1 squash inhibitor (1PPE_l, AUC = 0.21, green dot surrounded by a red circle in Supporting Information Fig. S7) and the T-cell receptor β (1SBB_r, AUC = 0.33, purple dot surrounded by a red circle in Supporting Information Fig. S7). Interestingly, in both cases the alternate binding sites predicted via the *PIP*-values matches with the interfaces formed in homodimeric structures as can be seen on Figure 5(d,e).

DISCUSSION

Conservation of the interacting interfaces throughout cross-docking calculations

We have shown that the *PIP* index provides useful information on residues more likely to be involved in protein-protein interfaces, although its performance depends on the protein family (see Fig. 2). Antigen surfaces in particular, seem to be more difficult to locate than the other interaction sites, including those of antibodies. Interestingly, this result agrees with the observations of Kowalsman and Eisenstein.[70] In their work analysing protein-protein docking models, they showed that interaction interfaces are enriched with high interface propensity residues, with the exception of antigenic sites. Both enzymes and inhibitors yield similar results regarding the interface propensity of their interface core residues (that is, residues that are totally buried in the experimental protein interface). In contrast, while antibody binding sites seem very well defined, including an important proportion of core residues with high interface propensity, antigen interfaces on the contrary are enriched with residues that have low interface propensities. Distinctions between enzyme/inhibitor and antigen/antibody interfaces have also been noted in other studies,[71–75] reflecting the fact that, while both surfaces evolve to promote binding in the former case, only the antibody surface evolves in the latter.

Our cross-docking results also show how the different evolutionary histories of antigens and antibody proteins impact their surface binding properties. In Figure 2(a,b), we can see that enzymes and inhibitors have comparable ROC curves, while the *PIP* index performs much better for the determination of antibody interfaces than for

antigens. In Figure 2(b), it is quite striking that unbound antibodies are the only functional group for which it is possible to make precisions that combine a high precision (above 50%) and high sensitivity (80%). Another interesting feature of Figure 2 is that unbound antibodies (dark blue curve) perform better that bound antibodies (light blue curve) for the prediction of interface residues. This somewhat puzzling result may explained by the fact that, in the case of bound antibodies, the protein interface underwent specific conformational changes so as to specifically fit a given antigen, leading to a decrease of the quality of binding with every other potential partner. In contrast, for unbound antibodies, the protein interface may be less adapted to the specific antigen, but performs better with all the other proteins. This effect also appears when looking at the average AUC for the three classes denoting conformational changes upon binding in the benchmark, with cross-docking on proteins from the *medium* category leading to slightly better binding site predictions than for protein from the *rigid* and *difficult* groups. Again, the unbound structure of proteins in the *rigid* group might be already adapted to their specific partner and less suitable for arbitrary docking, while a random partner will dock more easily on the unbound, and therefore unadapted, structure of a protein from the *medium* group. On the other hand, the conformational change undergone upon binding for proteins from the *difficult* group might this time be too important, thus making the binding site predictions more complicated for these systems. Interestingly, we would observe the opposite effect when looking at the prediction of interaction partners within the benchmark,[51] that is, proteins from the *rigid* group would naturally perform better than those from the *medium* and *difficult* groups, since their unbound structure is already adapted to the specific partner we were looking for, thus highlighting the fact that partner and binding site prediction for protein are two different issues that each have to be dealt with via specific tools. In particular, while efficient binding site prediction tools, using for example evolutionary data,[25] are nowadays available, binding partner prediction seems to be a much more complex problem which is still difficult to address.[48,51]

Proteins from the antibody and antigen groups show specific PIP clustering profiles

Looking at the average clustering profile for proteins in different biochemical groups, we can see a specific binding behavior emerge for the surfaces of antibodies and antigens. Antibodies usually present a single binding cluster that will be located on a very specific part of the protein surface, while antigens have several binding clusters that are scattered all over the protein's surface. These trends are in agreement with the different performances that are observed for the prediction of the experimental interfaces in these two groups using the *PIP* values, and again reflect the different evolutionary pressures undergone by the antigen and the antibody functional groups.

Cross-docking can lead to the prediction of alternate interfaces

An interesting feature of binding sites prediction via cross-docking simulations, is that for some cases it will lead to the prediction of an alternate interface that corresponds to the interaction of the target protein with biomolecular partners that are not present in the original protein benchmark. This is the case for several proteins in the enzyme, inhibitor, antigen and others categories, for which, at first sight, the binding site predictions via cross-docking seemed to be working badly, as shown from their weak AUC values compared to the average obtained for their functional category. In particular, it is quite striking that for two cases (the antigen HIV1 reverse transcriptase and the inhibitor Im7 immunity protein), this alternate binding site is actually a DNA binding site while the cross-docking experiment was exclusively performed on PPI. Alternate predicted interfaces can also correspond to the interfaces found in homodimeric structures of the protein under study. In their work on arbitrary protein-protein docking,[49] Martin and Lavery could list 59 cases of protein with multiple interfaces. For over one half (31 proteins) of these cases, arbitrary docking would target the alternate experimental interface instead of the reference experimental interface corresponding to the binding partner present in the docking benchmark. In our case, for those nine proteins (shown in Fig. 5 and Supporting Information Fig. S6) where alternate experimental interfaces were found, taking into account all the experimental interfaces residues (both from the reference and the alternate interface) for the evaluation of the binding site predictions via cross-docking leads to a slight increase of the AUC (from 0.74 to 0.75) and a noticeable increase of the method's precision as can be seen on Supporting Information Figure S8.

CONCLUSIONS

In this work, we use high-throughput cross-docking calculations on a dataset of 168 protein unbound structures to develop a *PIP* index for the prediction of interface residues on the protein surface. The *PIP* index does not require any prior knowledge of a protein's potential interaction partners, and its predictive power depends on the protein functional group and is poorer for antigens. We also develop a clustering algorithm which permits to group together surface residues with high *PIP*-values. The resulting clustering profiles for the various biochemical categories show remarkable differences, especially between the antibody and antigen groups, thus

suggesting that the binding ability of the surface residues actually holds some information regarding a protein's potential function. To the best of our knowledge, this is the first time such a clustering analysis has been performed on proteins surfaces, and that a close connection between a protein's function and its binding behaviour has been established. Finally, the *PIP* index can also lead to the prediction of alternate interface that are not present in our original benchmark, but are still biologically significant. This means that the evaluation of its predictive power, that is based only on the prediction of a restricted number of proteic interfaces already present in the docking benchmark, is probably well underestimated. The regular finding of alternate experimental interfaces which concur with our cross-docking predictions also highlight the fact that defining «true negatives» when dealing with PPI is actually a difficult matter.

In future work, we plan to apply the analysis tools that have been developed within the frame of this first experiment to the results from the cross-docking calculations subsequently performed on a much-larger protein database (made of 2246 proteins) that has been used in the second stage of the HCMD project. This database comprises a subset of >200 proteins that are known to be mutated and over-expressed in neuromuscular disorder. Combining data resulting from cross-docking simulations and evolutionary approaches, we aim to retrieve key information regarding the function, binding sites and potential binding partners of these target proteins.

Globally it is quite remarkable how a simple coarse-grain rigid approach can still bring us many information on *protein sociology*. Just like we cannot expect every single individual in a group to behave the same way, a protein will not always behave and bind the way we expect it to, no matter how exact or detailed our protein model might be. Because proteins, like people, are complex systems, driven by a profusion of parameters that we cannot hope to identify in their entirety. However, if we set aside individual cases, that cannot, and should not, be trusted to tell us exactly how a protein behaves, and now turn our attention to the statistical data emerging from high-throughput calculations, we can still retrieve valuable information regarding proteins general binding behaviour and function.

ACKNOWLEDGMENTS

This work was originally carried out in the framework of the DECRYPTHON Project, set up by the CNRS (Centre National de la Recherche Scientifique), the AFM (French Muscular Distrophy Association) and IBM. The cross-docking calculations were carried out on the *World Community Grid* (www.worldcommunitygrid.org) in phase 1 of the Help Cure Muscular Distrophy project. The authors wish to thank all the members of the WCG team for adapting our docking program for use on this grid, and also WCG volunteers for donating the computing time that made this work possible. S. S.-M., A. C. and L. V. thank the Mapping Project (Investissement d'Avenir) ANR-11-BINF-003 for postdoctoral funding.

REFERENCES

1. Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 1998; 92:291–294.
2. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996; 93:13–20.
3. Wodak SJ, Pu S, Vlasblom J, Seraphin B. Challenges and rewards of interaction proteomics. Mol Cell Proteomics 2009; 8:3–18.
4. Wass MN, David A, Sternberg MJE. Challenges for the prediction of macromolecular interactions. Curr Opin Struct Biol 2011; 21:382–390.
5. Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. Nature 2007; 450:973–982.
6. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li SM, Albala JS, Lim JH, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005; 437:1173–1178.
7. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005; 122:957–968.
8. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. Nature 2006; 440:631–636.
9. Krogan NJ, Cagney G, Yu HY, Zhong GQ, Guo XH, Ignatchenko A, Li J, Pu SY, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MHY, Butland G, Altaf-Ui AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 2006; 440:637–643.
10. Zhou M, Robinson CV. When proteomics meets structural biology. Trend Biochem Sci 2010; 35:522–529.
11. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Comput Biol 2007; 3:337–344.
12. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002; 415:141–147.

13. Butland G, Peregrin-Alvarez JM, Li J, Yang WH, Yang XC, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A. Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature 2005; 433:531–537.
14. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. A protein interaction map of Drosophila melanogaster. Science 2003; 302:1727–1736.
15. Li SM, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JDJ, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li QR, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu HY, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, van den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE. Vidal M. A map of the interactome network of the metazoan C-elegans. Science 2004; 303:540–543.
16. Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, Taylor R, Dharsee M, Ho Y, Heilbut A, Moore L, Zhang S, Ornatsky O, Bukhman YV, Ethier M, Sheng Y, Vasilescu J, Abu-Farha M, Lambert JP, Duewel HS, Stewart II, Kuehl B, Hogue K, Colwill K, Gladwish K, Muskat B, Kinach R, Adams SL, Moran MF, Morin GB, Topaloglou T, Figeys D. Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol 2007; 3:17
17. Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, Hogue CWV. BIND—The biomolecular interaction network database. Nucleic Acids Res 2001; 29:242–245.
18. Kuchaiev O, Rasajski M, Higham DJ, Przulj N. Geometric denoising of protein-protein interaction networks. PLoS Comput Biol 2009; 5:10
19. Tyagi M, Hashimoto K, Shoemaker BA, Wuchty S, Panchenko AR. Large-scale mapping of human protein interactome using structural complexes. EMBO Rep 2012;13:266–271.
20. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002; 12:368–373.
21. Sprinzak E, Altuvia Y, Margalit H. Characterization and prediction of protein-protein interactions within and between complexes. Proc Nat Acad of Sci USA 2006; 103:14718–14723.
22. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol 2007; 3:595–601.
23. Juan D, Pazos F, Valencia A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Nat Acad of Sci USA 2008; 105:934–939.
24. Engelen S, Trojan LA, Sacquin-Mora S, Lavery R, Carbone A. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. PLoS Comput Biol 2009; 5:e1000267
25. Laine E, Carbone A. Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein-protein interactions. PLoS Comput Biol 2015; 11:e1004580
26. Andreani J, Guerois R. Evolution of protein interactions: from interactomes to interfaces. Arch Biochem Biophys 2014; 554:65–75.
27. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. PRISM: protein interactions by structural matching. Nucleic Acids Res 2005; 33:W331–W336.
28. Aloy P, Russell RB. Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 2006; 7:188–197.
29. Launay G, Simonson T. Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. BMC Bioinformatics 2008; 9:427
30. Minton AP. Implications of macromolecular crowding for protein assembly. Curr Opin Struct Biol 2000; 10:34–39.
31. Ellis RJ, Minton AP. Cell biology—join the crowd. Nature 2003; 425:27–28.
32. Deeds EJ, Ashenberg O, Gerardin J, Shakhnovich EI. Robust protein-protein interactions in crowded cellular environments. Proc Natl Acad Sci USA 2007; 104:14952–14957.
33. Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 2004; 430:88–93.
34. de Lichtenberg U, Jensen LJ, Brunak S, Bork P. Dynamic complex formation during the yeast cell cycle. Science 2005; 307:724–727.
35. Tuncbag N, Kar G, Gursoy A, Keskin O, Nussinov R. Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example. Mol BioSyst 2009; 5:1770–1778.
36. Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. Cell 2011; 144:986–998.
37. Fernandez-Recio J. Prediction of protein binding sites and hot spots. Wiley Interdisciplinary Rev Comput Mol Sci 2011; 1:680–698.
38. Grosdidier S, Fernandez-Recio J. Protein-protein docking and hot-spot prediction for drug discovery. Curr Pharm Des 2012; 18:4607–4618.
39. Rodrigues JP, Bonvin AM. Integrative computational modeling of protein interactions. FEBS J 2014; 281:1988–2003.
40. Aumentado-Armstrong TT, Istrate B, Murgita RA. Algorithmic approaches to protein-protein interaction site prediction. Algorithms Mol Biol AMB 2015; 10:7
41. Gray JJ. High-resolution protein-protein docking. Curr Opin Struct Biol 2006; 16:183–193.
42. Fernandez-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. J Mol Biol 2004; 335:843–865.
43. Grosdidier S, Fernandez-Recio J. Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinformatics 2008; 9:447
44. Lensink MF, Wodak SJ. Blind predictions of protein interfaces by docking calculations in CAPRI. Proteins 2010; 78:3085–3095.
45. Hwang H, Vreven T, Pierce BG, Hung JH, Weng Z. Performance of ZDOCK and ZRANK in CAPRI rounds 13-19. Proteins 2010; 78: 3104–3110.
46. de Vries SJ, Bonvin AM. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PLoS One 2011; 6:e17695
47. Hwang H, Vreven T, Weng Z. Binding interface prediction by combining protein-protein docking results. Proteins 2014; 82:57–66.
48. Sacquin-Mora S, Carbone A, Lavery R. Identification of protein interaction partners and protein-protein interaction sites. J Mol Biol 2008; 382:1276–1289.
49. Martin J, Lavery R. Arbitrary protein-protein docking targets biologically relevant interfaces. BMC Biophys 2012; 5:7
50. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng ZP. Protein-protein docking benchmark 2.0: an update. Proteins 2005; 60:214–216.
51. Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A. Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. PLoS Comput Biol 2013; 9:e1003369
52. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The protein data bank. Acta Crystallogr, Sect D: Biol Crystallogr 2002; 58:899–907.

53. Zacharias M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci 2003; 12: 1271–1282.
54. Zacharias M. ATTRAACT: protein-protein docking in CAPRI using a reduced protein model. Proteins 2005; 60:252–256.
55. Schneider S, Saladin A, Fiorucci S, Prevost C, Zacharias M. ATTRACT and PTools: open source programs for protein-protein docking. Methods Mol Biol 2012; 819:221–232.
56. Lavery R, Sacquin-Mora S. Protein mechanics: a route from structure to function. J Biosci 2007; 32:891–898.
57. Sacquin-Mora S, Delalande O, Baaden M. Functional modes and residue flexibility control the anisotropic response of guanylate kinase to mechanical stress. Biophys J 2010; 99:3412–3419.
58. Sacquin-Mora S. Motions and mechanics: investigating conformational transitions in multi-domain proteins with coarse-grain simulations. Mol Simul 2014; 40:229–236.
59. Korobov NM. Approximate evaluation of repeated integrals. Doklady Akademii Nauk Sssr 1959; 124:1207–1210.
60. Bertis V, Bolze R, Desprez F, Reed K. From dedicated grid to volunteer grid: large scale execution of a bioinformatics application. J Grid Computing 2009; 7:463–478.
61. Hubbard SJ. ACCESS: a program for calculating accessibilities. Department of Biochemistry and Molecular Biology. University College of London; 1992.
62. Ritchie DW. Recent progress and future directions in protein-protein docking. Curr Protein Peptide Sci 2008; 9:1–15.
63. Bastard K, Prevost C, Zacharias M. Accounting for loop flexibility during protein-protein docking. Proteins 2006; 62:956–969.
64. May A, Zacharias M. Protein-protein docking in CAPRI using ATTRACT to account for global and local flexibility. Protein Struct Funct Bioinform 2007; 69:774–780.
65. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nat Biotechnol 2009; 27:157–167.
66. Erijman A, Aizner Y, Shifman JM. Multispecific recognition: mechanism, evolution, and design. Biochemistry 2011; 50:602–611.
67. Hamp T, Rost B. Alternative protein-protein interfaces are frequent exceptions. PLoS Comput Biol 2012; 8:e1002623
68. Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. Proteins 2010; 78:3111–3114.
69. Levy ED. PiQSi: protein quaternary structure investigation. Structure 2007; 15:1364–1367.
70. Kowalsman N, Eisenstein M. Combining interface core and whole interface descriptors in postscan processing of protein-protein docking models. Proteins 2009; 77:297–318.
71. Janin J, Chothia C. The structure of protein-protein recognition sites. J Biol Chem 1990; 265:16027–16030.
72. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999; 285:2177–2198.
73. Jackson RM. Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. Protein Sci 1999; 8:603–613.
74. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein - protein interfaces. J Mol Biol 2004; 336:943–955.
75. Vajda S. Classification of protein complexes based on docking difficulty. Proteins 2005; 60:176–180.