Chemistry & Biology
Article

Context-Based Identification of Protein-Protein Interfaces and "Hot-Spot" Residues

Tim Geppert,[1] Benjamin Hoy,[2,3] Silja Wessler,[2] and Gisbert Schneider[1,*]
[1]Swiss Federal Institute of Technology (ETH), Department of Chemistry and Applied Biosciences, 8093 Zürich, Switzerland
[2]University of Salzburg, Department of Molecular Biology, 5020 Salzburg, Austria
[3]Paul-Ehrlich Institute, 63325 Langen, Germany
*Correspondence: gisbert.schneider@pharma.ethz.ch
DOI 10.1016/j.chembiol.2011.01.005

SUMMARY

Reliable determination of protein-protein interaction sites is of critical importance for structure-based design of small molecules modulating protein function through macromolecular interfaces. We present an alignment-free computational method for prediction of protein-protein interface residues. The method ("iPred") is based on a knowledge-based scoring function adapted from the field of protein folding and small molecule docking. Based on a training set of 394 *hetero*-dimeric proteins iPred achieves sustained accuracy on an external unbound test set. Prediction robustness was assessed from more than 1500 diverse complexes containing *homo*- and *hetero*-dimers. The technique does not rely on sequence conservation, so that rapid interface identification is possible even for proteins for which homologs are unknown or lack conserved residue patterns in interface region. Functional "hot-spot" residues are enriched among the predicted interface residues, rendering the method predestined for macromolecular binding site identification and drug design studies aiming at modulating protein-protein interaction that might influence protein function. For a comparative structural model of peptidase HtrA from *Helicobacter pylori*, we performed mutation studies for predicted hot-spot residues, which were confirmed as functionally relevant for HtrA activity or oligomerization.

INTRODUCTION

Most protein functions result from interactions between multiple partners (Gong et al., 2005; Keskin et al., 2008), including signal transduction and modulation, and a whole class of enzymatic reactions that can be enhanced or inhibited by protein interaction partners (Aranda et al., 2010). Often only transient macromolecular complexes are formed, but permanent complexes have also been observed (Keskin et al., 2008). Interaction modes can be experimentally determined by nuclear magnetic resonance (NMR) or X-ray crystallography, and sometimes by electron microscopy (Shoemaker and Panchenko, 2007a). If only epitope information is needed, e.g., to guide mutation studies, mass spectrometric analyses or alanine scanning may be utilized (Lefevre et al., 1997; Juszczyk et al., 2009). As a downside, such methods are often time-consuming and cost-intensive. X-ray crystallography of protein-protein complexes is particularly challenging because of the requirement for native complex formation, and severe problems arise for transient complexes (Bravo and Aloy, 2006).

Robust computational methods might help identify interface residues and predict protein-protein interaction sites. A sound prediction can guide mutation studies and reduce the actual number of necessary mutations (Kortemme et al., 2004). It is also possible to gain information about ligand binding sites, which can then be targeted by small molecular inhibitors arising from drug design studies (Shoemaker and Panchenko, 2007b). Another field that will evidently profit from interface prediction is protein-protein docking (Halperin et al., 2002). Predicted interface residues can dramatically reduce the search space for protein docking algorithms, thereby enabling the use of time-intensive docking methods (Huang and Schroeder, 2008).

A wide variety of protein-protein complexes is available from the Protein Data Bank (PDB) (Berman et al., 2000) for the development of knowledge-based scoring functions for automated assessment of docking poses and simulated protein folding. The general assumption of knowledge-based scoring is that favored partners are found more often in a complex than nonpreferred interaction partners (Keskin et al., 2008). Here, we introduce and experimentally validate such a heuristic for protein interface prediction. Related techniques are based on the common assumption that the interface region can be separated from the rest of the surface by (1) discriminative potentials, such as electrostatic or hydrophobic potentials (Neuvirth et al., 2004), and (2) geometric features like planarity or solvent accessibility (Liang et al., 2006). In fact, there are slight differences in the amino acid composition and their local neighborhoods between interface and other surface regions in protein-protein complexes (Neuvirth et al., 2004; Liang et al., 2006). Based on these observations we propose to consider the local structural environment (the local context) of each surface residue for interface prediction, with the aim to enhance the discrimination between interface and noninterface residues by knowledge-based scoring.

RESULTS

During feature extraction for knowledge-based scoring one usually counts occurrences of atoms or groups that are involved

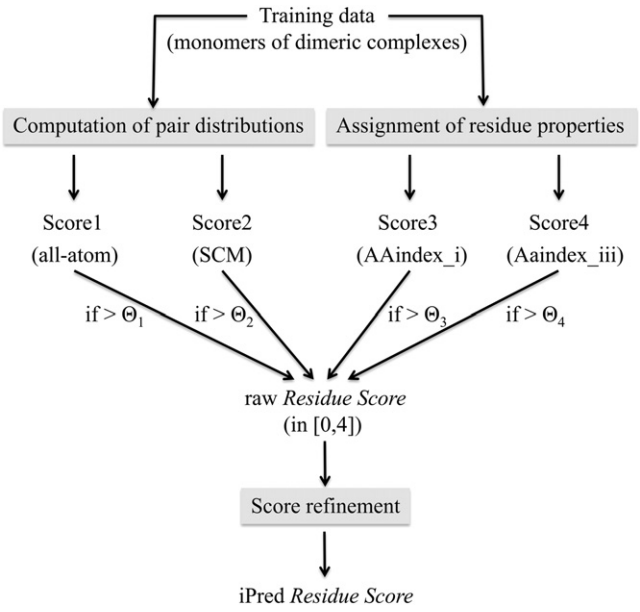

Figure 1. Overview of the iPred Scoring Workflow

Computation of the pair distribution and residue property assignments was based on the training data set. *Score1* is based on an all-atom description, and *Score2*-*Score4* on the side-chain center of mass (SCM) description of the protein structure. Each score that exceeds a threshold Q value increments the *Residue Score* by one. Score refinement results in the final iPred *Residue Score*. See also Table S1 and Table S7.

in an intermolecular interaction. While for protein-protein or protein-ligand docking one typically considers interaction partners from both molecules (Gohlke et al., 2000; Zhang et al., 2005), for interface prediction we measure the dissimilarity between the interface and all other surface residues from only one protein (one partner of the complex) (Melo et al., 2002). The goal is to find relevant differences in the local structural environment of these two residue types. We developed such a technique for protein interface prediction (iPred) without the need for sequence alignment or homology information and demonstrate its practical applicability in a case study using a peptidase (HtrA) from the human pathogen *Helicobacter pylori* as an example.

Algorithm Development and Testing

To extract the information for the discrimination of the "noninterface" and the "interface" area we devised four different scoring functions (Figure 1). One of the scoring functions (*Score1*) is based on atom types (Huang and Zou, 2008), the remaining three scoring functions are based on the protein side-chain center of mass (SCM, AAindex_i, AAindex_ii) representation of the 20 standard amino acids (Kocher et al., 1994). Each scoring function considers two pair potentials, namely, the interface and noninterface potential, which were extracted from the interface and noninterface area, respectively. The individual predictive score value for each amino acid residue is calculated from the difference between these two potentials. Each of the four scoring functions was individually evaluated on the training data (see Table S1 available online). *Score1* yielded the highest precision (59%), while *Score3* achieved the best coverage (16%). *Score1* and *Score4* had the highest success rate (57%), i.e., the largest predicted interface patch lies in the true interface area.

Table 1. Results for the Training and Evaluation Data

		Test data		
	Training	*Hetero*-Dimers	*Homo*-Dimers	Peptide Binding
Precision	60%	45%	47%	31%
Coverage	21%	13%	14%	14%
Background	34%	25%	24%	14%
No. of Proteins	394	526	656	265
Success	67% (265)	51% (269)	53% (346)	29% (78)

Predictions with more than 50% of the actual interface residues were considered successful. "Background" describes the background precision of the respective data set. See also Tables S2–S5.

After extraction of pair potentials from the training data, we evaluated the iPred method using independent test data. Results are summarized in Table 1. Individual results for each protein are given in Table S2. In total, 1442 proteins were used for evaluation, split into 530 *hetero*-dimeric structures and 616 monomers from *homo*-dimeric structures. We observed nearly identical performance of the iPred method on both data sets although training was based on *hetero*-dimeric structures only. As a hard test, we assessed the performance of iPred on a data set containing 265 examples of protein-peptide interactions (Table 1). Such interfaces are likely to be transient (Petsalaki et al., 2009). Examples of protein-peptide interactions were absent from the training data. Total success (correct interface prediction) was observed in 29%. It is noteworthy to mention that the percentage of interface residues within the protein-peptide data set drops by 10% in comparison to the *hetero*- and *homo*-dimeric structures (Table 2). A complete overview of the method's performance for each protein is provided in Tables S3–S5.

Since training was done with bound data, it is important to additionally measure performance using unbound data. For comparability to other interface prediction methods, we used an unbound data set compiled by Qin and Zhou (2007) from the DOCKGROUND database (Liu et al., 2008). It contains 31 structures, 18 of which have a sequence identity below 30% to our training data. Four different methods were compared with

Table 2. Comparison of iPred to the Performance of Four Other Methods Based on an Unbound Set of 31 Protein Structures

	cons-PPISP[a]	Promate[a]	PINUP[a]	meta-PPISP[a]	iPred
Precision	38%	47%	41%	42%	51% (51%)
Coverage	39%	17%	35%	46%	14% (12%)
Success	–	–	–	47%	50% (61%)
Background	–	–	–	–	23% (23%)

The iPred column gives the results for the complete data set and for a subset of 18 structures (in brackets), which are nonhomologous to the training data set.

[a] Taken from the study by Qin and Zhou (2007).

Table 3. Comparison of Hot-Spot Residue Prediction for Residues with a DDG >1 kJ/mol

Method	iPred	SVM	Robetta
Precision	60%	64%	69%
Recall	56%	79%	65%
F-Measure	0.53	0.71	0.67

Data from Lise et al. (2009). iPred scores were *z*-transformed and all residues yielding a final score >0.7 were used as positive predictions.

iPred: meta-PPISP (Qin and Zhou, 2007), cons-PPISP (Chen and Zhou, 2005), PINUP (Liang et al., 2006), and ProMate (Neuvirth et al., 2004). Table 2 and Table S6 present the results of the different methods as determined by Qin and Zhou (2007) together with the performance of iPred, which was evaluated on the whole data set set as well as the nonhomologous part of the data set. Overall, the precision of the iPred method is higher while the coverage is lower than for the other techniques. The Promate method performs similar to our method. This outcome demonstrates that our knowledge-based approach is able to competitively predict interface residues for unbound protein monomers, and the precision of the predictions is well above background.

Several studies have shown that not all interface residues contribute equally to binding and there are so-called hot-spot residues (Moreira et al., 2007; Bogan and Thorn, 1998). We tested whether iPred can also be used for hot-spot identification on a data set compiled by Lise et al. (2009) providing ΔΔG values for several interface residues. Table 3 presents the performance measures for the hot-spot residues among the predicted interface residues. As expected, overall performance is lower but still in the range of the results presented in the study by Lise et al. (2009), which is based on computational alanine scanning. This outcome confirms that iPred not only correctly identifies interface residues, but also retrieves the more important hot-spot residues within the predicted interface region.

Prediction of Potential "Hot-Spot" Residues

Inspired by the promising results of hot-spot analysis we performed a retrospective evaluation of the iPred method. The first case study uses an interferon alpha 2a structure (PDB-ID: 1itf, NMR model 1; Klaus et al., 1997) for which alanine-scanning mutagenesis data for its interaction with interferon-alpha receptor (IFNAR) was available (Piehler and Schreiber, 1999; Piehler et al., 2000). There were no interferon structures within our training set. iPred predicts a clustered region of potential interface residues and two separated residues (Figure 2A). The clustered residue patch perfectly lies within the IFNAR contact region as determined by experimental alanine scanning (Piehler and Schreiber, 1999; Piehler et al., 2000). Evidently, the predictions are in agreement with the mutagenesis studies. Most interestingly, in the first mutagenesis study by Piehler and Schreiber (1999) Arg149, which resides on loop E of interferon-alpha 2a, was not mutated because IFNAR interaction was assumed to be restricted to the AB region of the interferon. Only in a later study (Piehler et al., 2000), Arg149 was shown to be important for the interaction. iPred correctly determined both the correct interface region as well as the separate hot-spot residue.

In the second case study, we focused on the interface prediction of the amyloid-beta (Aβ) fibril (PDB-ID: 2beg, NMR model 1; Lührs et al., 2005). Aβ is known form plaques in brain tissue, which are hold responsible for neurodegenerative illnesses like Alzheimer's disease (Shankar et al., 2008). The prediction of possible interaction residues on the Aβ structure is particularly challenging due to the brevity of the Aβ monomer (42 residues). Training samples used for parameterization of iPred had an average length of 166 residues. Thus, we used the fibril structure consisting of five identical Aβ subunits. Prediction results are presented in Figure 2B. The maximal prediction score is

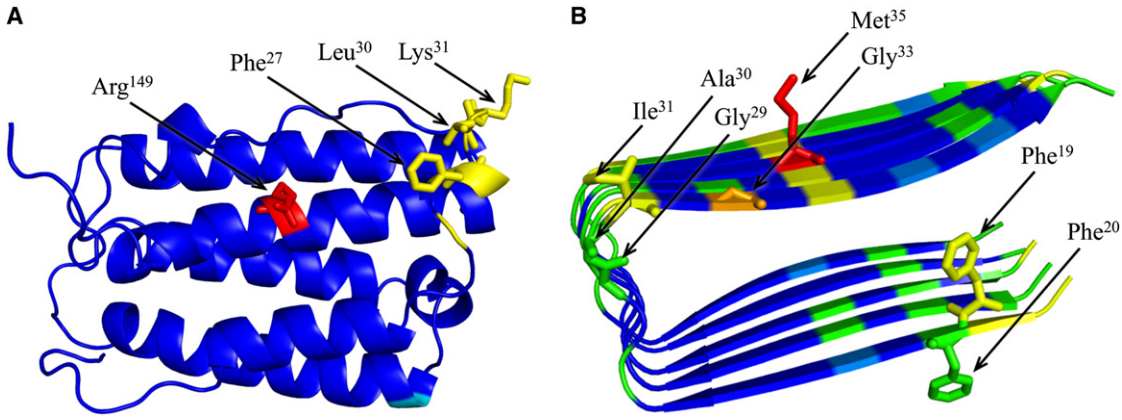

Figure 2. iPred-Based Predictions of Interferon alpha as well as the Amyloid-beta Fibril Interaction Area

(A) Interface prediction for interferon alpha 2a (PDB-ID: 1ITF, model 1; Klaus et al., 1997). Predicted interface residues are highlighted in red and yellow according to the iPred score. Arg^{149} was assigned the highest prediction score. Residues that were also found by alanine scanning mutagenesis are shown in stick representation (Piehler and Schreiber, 1999; Piehler et al., 2000). The change in free energy of binding for the four residues is as follows: Phe^{27Ala} ΔΔG = 2.0 kJ/mol; Leu^{30Ala} ΔΔG ~15 kJ/mol; Lys^{31Ala} ΔΔG = 2.1 kJ/mol; Arg^{149Ala} ΔΔG = 13.9 kJ/mol.

(B) Interface prediction for the amyloid-beta fibril (PDB-ID: 2BEG, model 1 (Lührs et al., 2005)). Predicted noninterface residues are colored blue. Residues in stick representation are known to interact.

See also Figure S1.

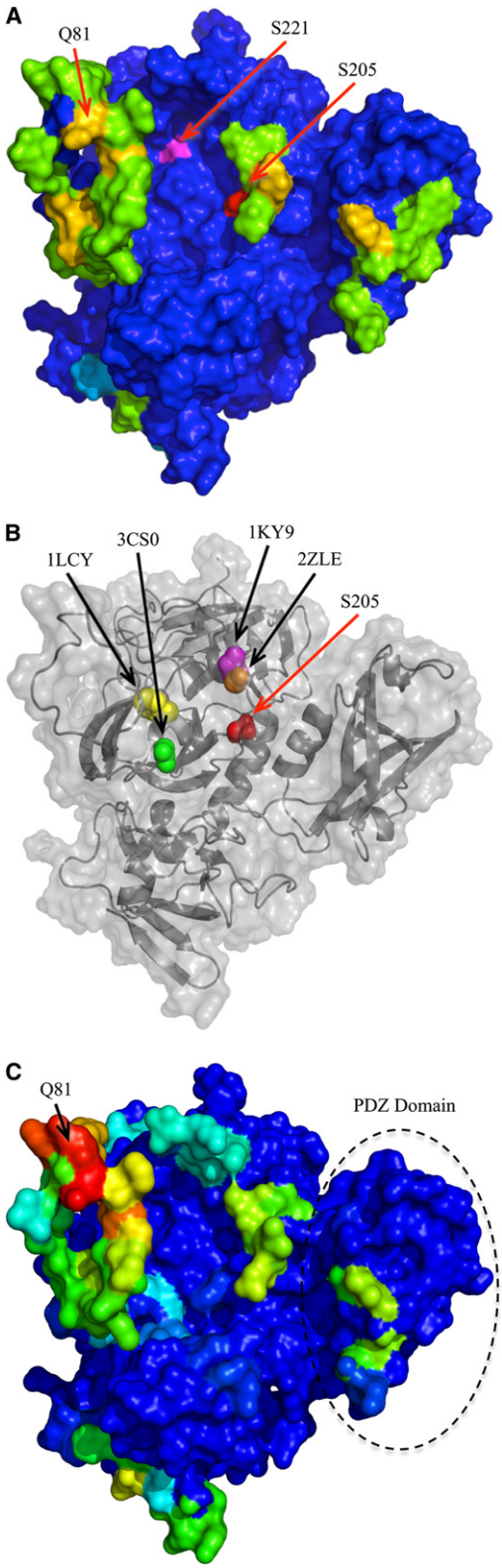

Figure 3. Interface Analysis of the HtrA Homology Model and Homolog Structures

(A) iPred interface prediction based on the homology model of HtrA. Predicted interface residues are colored in green, yellow and red according to increasing prediction score. The highest scored reside is Ser^{205}. The catalytic Ser^{221} is colored in magenta.

(B) Location of the highest scoring residue for four different structures, which are sequence homolog to HtrA. The location of Ser^{205} is highlighted in red. All predicted residues are on the same face of the protein. (PDB-ID 1lcy: serine protease HtrA2, Li et al., 2002; 3cs0: DegP24, Krojer et al., 2008; 1ky9: DegP, Krojer et al., 2002; 2zle: DegP12/OMP, Li et al., 2002).

(C) Average iPred interface prediction based on an ensemble of 2000 protein conformations resulting from a 20 ns MD simulation. Coloring is based on the mean predicted residue score using the color-coding scheme from panel A. Highest scored residue is Glu^{81}. A dotted circle indicates the hypothetical PDZ domain of HtrA.

assigned to the central Met^{35}, which it is known to crosslink with residue four (Lührs et al., 2005). Phe^{19} and Phe^{20} are correctly identified as hot spots, as they interact with human cystatin C (Juszczyk et al., 2009). The third area of interest is the region around residues 29–33, which is important for fibril oligomerization (Juszczyk et al., 2009). Notably, a helix-binding GxxxG motif is located within this region (Russ and Engelman, 2000). It is important to mention that motifs were not explicitly considered for predictor development.

Prospective Application

Encouraged by the consistency of predicted interface residues and actual hot spots, we performed a prospective study. The iPred prediction was carried out for a protein homology model of secreted *H. pylori* peptidase HtrA, which had been identified by us as an important virulence factor of these gastric bacteria (Hoy et al., 2010; Löwer et al., 2008), and four additional homologous structures.

Among the potential interface residues, Ser^{205} of *H. pylori* HtrA received the highest prediction score of all residues (Figure 3A). This residue is located in proximity to the potential catalytic Ser^{221} of the active site (Jiang et al., 2008). In Figure 3B, the positions of the top-scoring predicted interface residues are highlighted in four close HtrA homologs. Without exception, these are located near Ser^{205} on the same face of the protein. These results suggest that HtrA binds to a peptide or protein with the surface patch around Ser^{205} being part of the interface.

A second patch of potential hot-spot residues was predicted around Gln^{81}. To test the validity of this second surface area, which is located in a structurally poorly resolved and potentially flexible loop, we evaluated iPred prediction sustainability for an ensemble of HtrA conformations that were generated by a 20 ns molecular dynamics simulation. The conformer ensemble consisted of 2000 snapshots sampled for each 0.01 ns time step. iPred scores were computed for all 2000 protein conformations. For each residue the average score was visualized by color-coding on the initial starting conformation of HtrA (Figure 3C). Apparently, the overall number of potential protein-protein interaction sites remained constant over the MD simulation. In the ensemble average, the surface patch around Gln^{81} receives most favorable scores. Ser^{205}, which was top scoring in the starting structure of HtrA prior to MD simulation (Figure 3A), was no longer considered as interface residue in the ensemble average. This points to some structural flexibility in the presumable active site of HtrA.

We then performed biochemical experiments, and first investigated the activity of wild-type (WT) protein ($HtrA^{wt}$), a Ser^{221}Ala

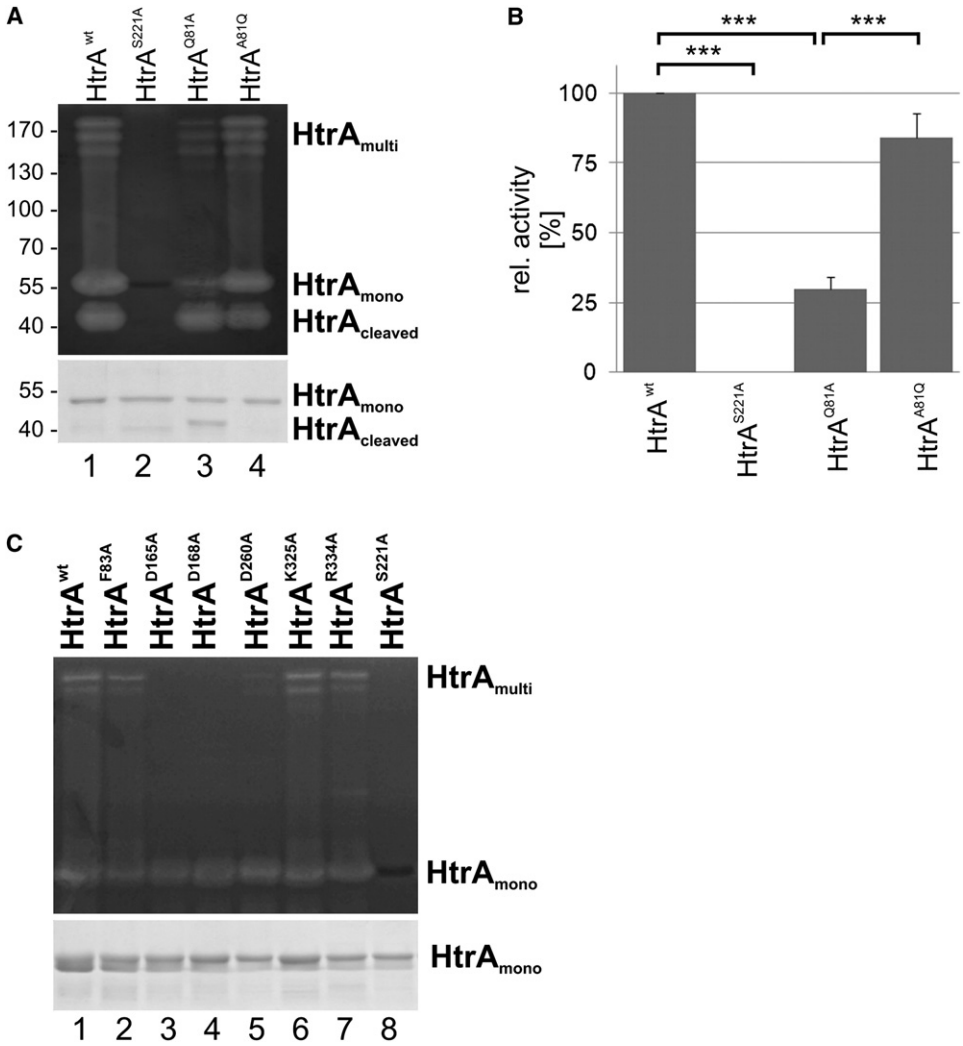

Figure 4. Mutation Study of HtrA Interface Resides Based on iPred Predictions

(A) One μg of purified HtrA^{wt}, HtrA^{S221A}, HtrA^{Q81A}, and HtrA^{A81Q} were separated in a casein zymogram to analyze the proteolytic activity of multimeric (HtrA$_{multi}$), monomeric (HtrA$_{mono}$) and processed HtrA (HtrA$_{cleaved}$) (upper panel). Equal protein amounts were shown in Coomassie-stained SDS PAGEs (lower panel).

(B) HtrA activity was quantified from five independent experiments and is expressed as relative activity compared to the activity of HtrA^{wt}. Asterisks indicate statistically significant differences (***$p < 0.001$).

(C) Two micrograms of HtrA^{wt} or the isogenic mutants were analyzed by casein zymography to visualize the activity of HtrA monomers and multimers (upper panel). Equal protein amounts were used and visualized in Coomassie-stained SDS PAGEs (lower panel).

mutant (HtrA^{S221A}), and a Gln^{81}Ala mutant (HtrA^{Q81A}). The Gln^{81} residue was selected as representative for the high-scoring surface patch according to the iPred prediction. We could not directly probe protein-protein association due to unknown binding partners of HtrA. Therefore, we decided to test for effects of the mutations on substrate processing by HtrA. Residue Ser^{221} has recently been shown to be important for catalytic activity (Hoy et al., 2010; Löwer et al., 2008), and served as reference for the present study. Casein is a commonly used substrate for bacterial HtrA, which induces large complexes upon binding to DegP (Jiang et al., 2008). In casein zymography, we observed strong proteolytic activity of HtrA^{wt}, which was inactivated by the Ser^{221}Ala mutation in the postulated active site (HtrA^{S221A}). Importantly, mutation of Gln^{81} (HtrA^{Q81A}) also resulted in significant inhibition of the proteolytic function of HtrA by approximately 70% (Figure 4B). The loss of proteolytic activity of HtrA^{Q81A} was restored after re-mutation from alanine to WT-glutamine (HtrA^{A81Q}, Figures 4A and 4B).

In addition to these top-predicted residues, we also probed the effect of additional mutations on HtrA activity. We selected Phe^{83}Ala which is located within the predicted interface patch of Gln^{81} as well as various charged residues at the HtrA surface: Asp^{165}Ala, Asp^{168}Ala, Asp^{260}Ala, Lys^{325}Ala, and Arg^{334}Ala. Lys^{325} as well as Arg^{334} are located in a hypothetical PDZ domain (residues 274–361, Figure 3C) of the HtrA protein (Kennedy, 1995). While iPred actually predicts an interface patch within the hypothetical PDZ domain (around Ile^{300}), none of the mutated residues were predicted as hot spots (score = 0, blue colored

residues in Figure 3). In agreement with these predictions we did not observe any effect of the mutations on the catalytic activity of the HtrA monomer (Figure 4C). Notably, activity of the HtrA homolog DegP from *Escherichia coli* is controlled by oligomer conversion induced upon substrate binding (Jiang et al., 2008). In fact, two of our HtrA mutants ($Asp^{165}Ala$, $Asp^{168}Ala$) exhibit strongly reduced oligomer activity or formation (Figure 4C, lanes 3 and 4). HtrA mutants $Asp^{260}Ala$ and $Gln^{81}Ala$ also exhibit reduced oligomer activity (Figure 4C, lane 5, and Figure 4A), while $Phe^{83}Ala$, $Lys^{325}Ala$, and $Arg^{334}Ala$ mutations have no effects (Figure 4C, lanes 2, 6, and 7).

DISCUSSION

The iPred method for prediction of protein interface residues is based on the concept of residue environment conservation. As an alternative to considering only residue frequencies for the distinction between interface and noninterface positions, we have introduced the concept of context-dependent scoring. Based on an independent evaluation set, we could demonstrate that a combination of four scoring functions captures relevant differences between interface and noninterface residues. This scheme resulted in a robust interface classifier with competitive performance, as determined for a diverse data set of more than 1500 proteins including *homo*- and *hetero*-dimers. Notably, iPred also showed to be applicable to recognizing peptide-binding regions on protein surfaces. Examples of protein-peptide interaction were absent from the training data set. Although prediction accuracy is not perfect, we are able to achieve for all tested data sets a performance that is more than twice as good as the background probability for interface residues without relying on information about the binding peptide (Petsalaki et al., 2009). A comparison to four other interface prediction methods based on unbound data showed that the performance of our method is unaffected by differences between unbound and bound data. Prediction precision of 51% was achieved, and based on the performance measure iPred is comparable to the ProMate method, while relying on a novel scoring concept.

We suggest that by using both property types and distance information for scoring one captures physicochemical as well as structural information of interface regions. Such features might be locally conserved in protein-protein interfaces. Importantly, iPred does not use sequence conservation for prediction, as the other compared methods do. Consequently, as a unique feature, it may be applied to protein structures for which close homologous proteins are unknown.

Using an ensemble of HtrA structures from a 20 ns molecular dynamics simulation, we were able to show the robustness of the iPred method. Structural fluctuations within the molecular dynamic trajectory did not affect the predicted interface areas. Due to prediction times of only 20 s *per* HtrA structural model (454 residues) iPred qualifies itself as part of an ensemble prediction workflow.

Testing the conclusiveness of iPred prediction, we analyzed the *H. pylori* serine protease HtrA in our study. Recently, *H. pylori* HtrA has been identified as a novel virulence factor that is secreted by *H. pylori* into the extracellular space to cleave E-cadherin and fibronectin on epithelial host cells; obviously an important step in bacterial pathogenesis (Weydig et al., 2007; Hoy et al., 2010). Although less is known about how *H. pylori* HtrA activity is controlled, it has become apparent that *E. coli* DegP activity is regulated via formation of large cage-like oligomers upon substrate binding (Jiang et al., 2008). Based on a DegP-derived homology model, residues located around Ser^{221} in the presumable active center of *H. pylori* HtrA received high scores indicating that this surface area is actually involved in substrate binding. A second functional hot-spot site was experimentally confirmed by mutating Gln^{81} to Ala^{81} located at the tip of an HtrA protuberance that might also interact with substrate or other yet unknown protein(s). Back-mutation to Gln^{81} fully restored HtrA activity, which supports the validity of both the predictions and the assay system used. In addition, we probed mutations of several charged surface residues that were not predicted by iPred as potential interface residues. None of the mutants exhibited modified enzymatic activity of the HtrA monomer, which again corroborates the predictive model.

Overall, the mutation study for HtrA demonstrates that iPred has the potential to detect functionally relevant surface patches. This is somewhat surprising, as we did not explicitly consider protein *function* during the development of iPred. Although our biochemical study does not provide direct proof for the ability of the method to predict protein-protein interfaces, its predictive power supports the notion that some functional effects might result from modified protein-protein association. The HtrA family contains one or more C-terminal PDZ domains mediating specific protein-protein interactions and binding preferentially to the C-terminal part of target proteins (Jiang et al., 2008). We identified Asp^{165}, Asp^{168}, and Asp^{260} as important amino acids for HtrA oligomer activity and possible oligomer formation. Asp^{260} is located between the first hypothetical PDZ domain and loop L3, which stabilizes the active state of *E. coli* DegP. Hence, mutation of Asp^{260} might influence the protein interaction through the PDZ domain or the activity of the HtrA multimer (Krojer et al., 2010). The roles and function of Asp^{165} and Asp^{168} in HtrA from *H. pylori* are currently being investigated by us.

Evidently, several structural aspects that are crucial for protein-protein interface formation are not considered by iPred. For example, in its present implementation, we neglect solvent effects resulting from water molecules in the interface, and do not explicitly account for the thermodynamics of the interaction process. Possibly, iPred will not be able to recognize entropy-driven interactions leading to the formation of large flat interface patches. Also, in this study we restricted iPred application to monomer structures. Prediction accuracy might be improved by considering both binding partners and their surface flexibility. Recently, Krüger and Gohlke developed a specialized method for hot-spot prediction that explicitly considers both interaction partners (Krüger and Gohlke, 2010). This overall concept is similar to the Robetta approach that had been developed earlier by Baker and coworkers (Kim et al., 2004). We see a particular complementarity between our monomer-focused iPred concept and such dimer-specific techniques. It should not be expected that monomer-based prediction outperforms dimer approaches, but rather provide first indications about potential protein-protein contacts when only one binding partner is structurally known, such as for the specific case of HtrA investigated here.

The conceptual framework presented here is not limited to interface prediction, but could also be used for protein-protein docking, which would profit from a reduced search space based on predicted interface areas. Although our method was not specifically designed to predict hot-spot resides, we could show that concept of "residue environment conservation" actually enriches hot-spot residues among the predicted interface residues. In other words, the spatial context of surface residues contains essential information that cannot be extracted when looking at individual residues only, without explicit consideration of the local neighborhood. Conservation of hot spots and their structural environment might be greater than for other interface residues, making hot-spot residues easier to distinguish for our method. Reliable computer-based identification of hot spots might also facilitate the design of small molecule inhibitors of protein-protein interactions. Such compounds are urgently sought for as protein-protein interaction networks represent a promising target area in drug discovery (Fry, 2006; Keskin et al., 2007).

SIGNIFICANCE

We have developed an alignment-free method for the prediction of protein interface residues. It is the first prediction method that incorporates the complete structural environment information of a residue to determine its interface propensity. We could show that predicted interface residues are enriched in hot-spot residues. The method also provides a new framework for the prediction of important interface residues for proteins without close homologs. We predicted and tested two residues of the *H. pylori* serine protease HtrA and were able to show that mutation of each residue lead to an activity reduction of the virulence factor HtrA. The prediction not only confirmed our retrospective studies but also showed the applicability of the iPred method for a comparative structural model.

EXPERIMENTAL PROCEDURES

Cloning, Mutation, and Purification of HtrA

Cloning of $HtrA^{wt}$ and $HtrA^{S221A}$ was described previously (Löwer et al., 2008; Hoy et al., 2010). The generation of the HtrA mutants ($HtrA^{Q81A}$, $HtrA^{A81Q}$, $HtrA^{F83A}$, $HtrA^{D165A}$, $HtrA^{D168A}$, $HtrA^{D260A}$, $HtrA^{K325A}$, and $HtrA^{R334A}$) was performed using the QuikChange Lightning Site-Directed Mutagenesis Kit (Stratagene) according to the manufacturer's instructions. Additionally, $HtrA^{Q81A}$ mutation was restored by remutation of alanine 81 to glutamine ($HtrA^{A81Q}$). For heterologous overexpression and purification of proteins, transformed *E. coli* was grown in 500 ml TB medium to an OD550 of 0.6 and the expression was induced by the addition of 0.1 mM isopropylthiogalactoside (IPTG). The bacterial culture was pelleted at 4,000 × g for 30 min and lysed in 25 ml PBS by sonification. The lysate was cleared by centrifugation and the supernatant was incubated with glutathione sepharose (GE Healthcare Life Sciences) at 4°C over night. The fusion proteins were cleaved with 180 U Prescission Protease for 16 hr at 4°C (GE Healthcare Life Sciences).

Zymography

Proteins (1.0–2.0 μg) were loaded onto 8% SDS-PAGE containing 0.1% casein (Invitrogen, Germany) and separated by electrophoresis under non-reducing conditions. After separation, the gel was renaturated in 2.5% Triton X-100 solution at room temperature for 60 min with gentle agitation, equilibrated in developing buffer (50 mM Tris-HCl [pH 7.4]; 200 mM NaCl, 5 mM $CaCl^2$, 0.02% Brij35) at room temperature for 30 min with gentle agitation, and incubated overnight at 37°C in fresh developing buffer. Transparent bands of caseinolytic activity were visualized by staining with 0.5% Coomassie blue R250 and quantified by the Fusion FX-7 image analyzer and Bio-1D analysis software (Vilber Lourmat Deutschland GmbH). To show equal protein amount, 1.0 μg of the proteins were separated by SDS PAGE followed by Coomassie staining. The experiments were repeated four or five times.

Source Data

For the present study we used five different protein sets. Protein structures for the different data sets were downloaded from the PDB (Tables S1–S4). To derive the parameters of our scoring function, we relied on a nonredundant data set ("training set") containing 197 hetero-dimeric structures compiled by Huang and Zou (2008). For evaluation purposes we compiled a data set ("evaluation set") of 263 *hetero*-dimeric structures with no more than 30% pairwise sequence identity, and 656 *homo*-dimeric structures from the same source as the *hetero*-dimeric data set (Huang and Zou, 2008). Of the *homo*-dimeric data set we used only one monomer per dimeric structure. We excluded structures with fewer than 30 residues and accepted only resolutions below 2.0 Å. We also assessed the performance of the iPred method on a data set that had been compiled by Petsalaki et al. (2009) containing 265 protein-peptide complexes and predicted the interfaces for the larger partner of the complex. To compare the performance of iPred with other available methods we used the recently published data by Qin and Zhou (2007) collected from the DOCKGROUND database. This data set contains only unbound structures. A data set compiled by Lise et al. was used to assess whether a predicted interface residue is a hot spot. For all structures of this data set the relative change in binding affinity (ΔΔG) had been measured for selected interface residues by alanine scanning (Lise et al., 2009).

Data Preparation

To determine the sequence identity within the different data sets, we used CLUSTALW (Larkin et al., 2007). For all proteins, we computed the percent identity to all other proteins whose sequence length differed by not more than 30% from the query. Unless otherwise stated, we then removed all proteins from the respective evaluation data set with a sequence identity of greater than 30% to the training set. Calculation of surface accessibility was done using DSSP (Kabsch and Sander, 1983). As in previous studies by Qin and Zhou (2007) and Liang and Neuvirth (Neuvirth et al., 2004; Liang et al., 2006), we flagged all protein residues with a surface accessibility exceeding 10% as "surface residues." Surface residues undergoing an accessibility change of more than 5% upon complex formation were marked as "interface residues." Interface residues are a subset of the surface residues. For scoring, we reimplemented the surface protrusion method introduced by Pintar et al. (2002).

Molecular Dynamics Simulation

MD simulation was performed using the CUDA accelerated NAMD 2.7b3 (Phillips et al., 2005) based on the CHARMM 27 (Brooks et al., 1983) force field. We divided the simulation into four consecutive parts. In the first phase, the system was minimized for 2000 steps with constrained backbone and than minimized for 2000 steps with harmonically restrained backbone. In the second phase, the system was heated to 310 K over 12 ps. The system was then equilibrated for 5 ns. Following the equilibration phase we executed the actual production run for 20 ns. The system was set up within a cubic water box and neutralized using chlorine counterions in 0.1 mM concentration. We used periodic boundary conditions as well as the particle mesh Ewald method for long-range electrostatic interaction (Ewald, 1921). The mean energy of the whole simulation system during the 20 ns of dynamics simulation was –241,716 $kcal \cdot mol^{-1}$ with a standard deviation of 328 $kcal \cdot mol^{-1}$.<<COMP: middots needed between kcal and mol in this sentence x 2>>.

Residue Scoring

During the training phase four different scoring functions were developed (Figure 1). All calculations were done on monomeric structures extracted from the dimeric structures within the training data set. Of the monomeric structures only the surface residues were used for the computations. The final iPred predictor consists of a total of four individual functions: *Score1* is based on the all-atom pair distributions, and *Score2* on the amino acid side-chain center

of mass (SCM) pair distributions computed from the training data set. *Score1* and *Score2* were calculated following an adaption of Huang's method to calculate interaction potentials for the scoring of docked protein complex poses (Huang and Zou, 2008). *Score3* and *Score4* are based on potentials from the AAindex database (Kawashima et al., 2008). The AAindex database contains (1) 544 residue properties, (2) 88 mutation matrices, and (3) 47 interaction potentials. For iPred only AAindex_i (*Score3*) and AAindex_iii (*Score4*) were used.

Score1

Score1 is based on the atom types (all-atom) (Huang and Zou, 2008). The pair potential $\Phi_{ij}(r)$ was extracted in two different ways, namely an interface and a noninterface pair potential. For the interface pair potential, pairings of each interface residue i with all structurally adjacent surface residues j were counted. Pairings of each noninterface residue with all remaining surface residues outside the interface were counted for the noninterface distribution. This scheme results in an interface pair potential $\Phi_{ij}^{interface}(r)$ as well as a noninterface $\Phi_{ij}^{non-interface}(r)$ pair potential for atom type i with atom type j from the shell $r-b/2$ (lower bound) and $r+b/2$ (upper bound). Here b denotes the shell thickness, which was set to 0.2 Å, r equals the distance between i and j, where $r < r_{max}$, which was set to 10 Å. The calculation of pair potentials was is done according to Equation (1) for the interface as well as the noninterface case, respectively.

$$\phi_{ij}(r) = -\ln\frac{g_{ij}(r)}{f_{ij}(r)}, \tag{1}$$

where $g_{ij}(r)$ represents the normalized pair distribution and $f_{ij}(r)$ defines a volume correction factor that accounts for size differences of the different residue types used (Huang and Zou, 2008). If one of the two functions equals zero the pair potential $\Phi_{ij}(r)$ is defined as zero, too. The normalized pair distribution $g_{ij}(r)$ was calculated using Equation (2).

$$g_{ij}(r) = \frac{p_{ij}(r)}{p_{ij}^{bulk}}, \tag{2}$$

where $p_{ij}(r)$ gives the pair distribution of the two residue types i and j in the shell defined by distance r [Equation (3)]. $p_{ij}^{bulk}(r)$ is equal to the distribution of a pairing ij within the whole distance range [Equation (4)]. The distance's upper bound r_{max} was set to 10 Å.

$$p_{ij}(r) = \frac{1}{|M|}\sum_{m}^{M}\frac{n_{ij}^{m}(r)}{4\pi r^2 b}. \tag{3}$$

$$p_{ij}^{bulk}(r) = \frac{1}{|M|}\sum_{m}^{M}\frac{N_{ij}^{m}}{4\pi r_{max}^3/3}. \tag{4}$$

$n_{ij}^{m}(r)$ provides the relative frequency of pairs i and j found within the shell defined by distance r and thickness b = 0.2 Å, as occurring in molecule m. N_{ij}^{m} is the sum of all n_{ij}^{m} from all shells. M is the number of monomers in the training set. The volume correction factor $f_{ij}(r)$ was calculated using Equation 5.

$$f_{ij}(r) = \frac{1}{2}\left(\frac{\sum_i p_{ij}(r)}{\sum_i p_{ij}^{bulk}(r)} + \frac{\sum_j p_{ij}(r)}{\sum_j p_{ij}^{bulk}(r)}\right), \tag{5}$$

where $p_{ij}(r)$ and p_{ij}^{bulk} are computed as given in Equations 3 and 4. Following the computation of the pair potential $\Phi_{ij}(r)$, each residue on the protein surface was scored with the noninterface as well as the "interface" pair potential. The score of each residue is calculated according to Equation 6.

$$\text{Score1}_i = \sum_{\substack{j\neq i \\ r>r_{max}}} \Phi_{ij}^{interface}(r) - \sum_{\substack{j\neq i \\ r>r_{max}}} \Phi_{ij}^{non-interface}(r). \tag{6}$$

In Equation 6, the two variables $\Phi_{ij}^{interface}(r)$ and $\Phi_{ij}^{non-interface}(r)$ represent the pair potentials extracted from the interface and noninterface area as defined previously. r_{max} was set to 10 Å.

Score2

Computation of *Score2* is similar to *Score1*. While *Score1* is based on atom types, *Score2* is based on the side-chain center of mass (SCM) representation of the 20 amino acids.

Score3

Score3 is based on residue properties from the AAindex database part (1) (Kawashima et al., 2008) (Equation 7).

$$\text{Score3}_i^k = \frac{1}{|M_{ij}^k|}\sum_{a_j^k \in M_{ij}^k}\frac{a_j^k}{D_2(\mathbf{a}_i^k, \mathbf{a}_j^k)}, \tag{7}$$

where $M_{ij}^k = \{a_j^k | D_2(\mathbf{a}_i^k, \mathbf{a}_j^k) < r_{max}\}$, and r_{max} = 10 Å. a_j^k is the k^{th} property value of residue j from molecule M, and $D_2(\mathbf{a}_i^k, \mathbf{a}_j^k)$ the Euclidian distance between $\mathbf{a}_i^k$ and $\mathbf{a}_j^k$. The calculation results in k = 544 different scores corresponding to the 544 different residue properties of the AAindex_i database.

Score4

To compute *Score4* the residue pair potentials of the AAindex database part (3) were used. Equally to *Score3*, *Score4* is based on the side-chain center of mass representation of the protein surface Equation (8).

$$\text{Score4}_i^k = \frac{1}{|M_{ij}^k|}\sum_{a_j^k \in M_{ij}^k}\frac{a_i^k \cdot a_j^k}{D_2(\mathbf{a}_i^k, \mathbf{a}_j^k)}. \tag{8}$$

For $Score4_i^k$, the type of residue a_i^k as well as the partner residue a_j^k is used because AAindex_iii contains pair potentials. To ensure enhanced discrimination between interface and non-interface surface patches we focused the four scoring functions on residue types resulting in a precision of at least 50% (Table S7).

Score Thresholds

Threshold values (Θ_1, Θ_2, Θ_3, Θ_4) for all four scoring functions were established to classify interface and noninterface residues. All surface residues from the training data set were scored and thresholds defined that resulted in a precision of at least 50% and a minimum of 100 correct predicted interface residues. This criterion was chosen to guarantee that the scoring functions consider multiple proteins rather than focusing on individual training examples.

Residue Score

The four scores were calculated for each surface residue. When a score exceeds the defined cutoff value the residue's vote counter is incremented by one. This results in a raw residue score in the range between zero to four for each surface residue.

Score Refinement

Following the calculation of the raw residue score, residues receive additional votes according to the following rules:

(1) The two nearest neighbors of a residue with two votes, and the three nearest neighbors of a residue with three or four votes receive one additional vote (contrast enhancement).
(2) The protrusion of each predicted interface residue is calculated (Pintar et al., 2002). If a residue is below the mean of the protrusion of all predicted residues of this protein it loses one vote, otherwise its vote counter is incremented by one.
(3) The one predicted interface residue with the largest number of predicted neighbors within a radius of 10 Å receives the largest vote count of all predicted residues. The vote counters of all residues that are within 10 Å of this residue are incremented by one.
(4) All predicted interface residues without any predicted interface residue within a radius of 6 Å were omitted to remove solitary interface residues from the final result.

Performance Analysis

A prediction was successful if more than 50% of all predicted residues are true interface residues (Neuvirth et al., 2004). For quantification, we used *Precision*, *Coverage* (Neuvirth et al., 2004;Liang et al., 2006; Qin and Zhou, 2007) and the F-measure (van Rijsbergen, 1979).

Programming Environment

Software development and evaluation were done using JAVA version 1.6 (Sun Microsystems).

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven tables and one figure and can be found with this article online at doi:10.1016/j.chembiol.2011.01.005.

ACKNOWLEDGMENTS

We thank Gert Carra for excellent technical assistance. We are grateful to Felix Reisen for discussions about protein-small molecule interaction sites.

Received: October 8, 2010
Revised: December 3, 2010
Accepted: January 5, 2011
Published: March 24, 2011

REFERENCES

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, T., Kerrien, S., Khadake, J., et al. (2010). The IntAct molecular interaction database in 2010. Nucleic Acids Res. *38*, D525–D531.

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

Bogan, A., and Thorn, K. (1998). Anatomy of hot-spots in protein interfaces. J. Mol. Biol. *280*, 9–10.

Bravo, J., and Aloy, P. (2006). Target selection for complex structural genomics. Curr. Opin. Struct. Biol. *16*, 385–392.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, B.J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A Programm for Macromolecular Energy, Minimization, and Dynamics Calculations. J. Comput. Chem. *4*, 187–217.

Chen, H., and Zhou, H. (2005). Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins *61*, 21–35.

Ewald, P. (1921). Die Berechnung optischer und elektrostatischer Gitterpotentiale. Ann. Phys. *369*, 253–287.

Fry, D.C. (2006). Protein-protein interactions as targets for small molecule drug discovery. Biopolymers *84*, 535–552.

Gohlke, H., Hendlich, M., and Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. J. Mol. Biol. *295*, 337–356.

Gong, S., Yoon, G., Jang, I., Bolser, D., and Dafas, P. (2005). PSIbase: a database of Protein Structural Interactome map (PSIMAP). Bioinformatics *21*, 2541–2543.

Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002). Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins *47*, 409–443.

Hoy, B., Löwer, M., Weydig, C., Carra, G., Tegtmeyer, N., Geppert, T., Schröder, P., Sewald, N., Backert, S., Schneider, G., and Wessler, S. (2010). Helicobacter pylori HtrA is a novel secreted virulence factor which cleaves E-cadherin to disrupt intercellular adhesion. EMBO Rep. *11*, 798–804.

Huang, B., and Schroeder, M. (2008). Using protein binding site prediction to improve protein docking. Gene *422*, 14–21.

Huang, S., and Zou, X. (2008). An iterative knowledge-based scoring function for protein-protein recognition. Protein Struct. Funct. Genet *72*, 557–579.

Jiang, J., Zhang, X., Chen, Y., Wu, Y., Zhou, Z.H., Chang, Z., and Sui, S.F. (2008). Activation of DegP chaperone-protease via formation of large cage-like oligomers upon binding to substrate proteins. Proc. Natl. Acad. Sci. USA *105*, 11939–11944.

Juszczyk, P., Szymanska, A., Kolodziejczyk, A.S., Rodziewicz-Motosidlo, S., Grzonky, Z., and Przybylski, M. (2009). Binding epitopes and interaction structure of the neuroprotective protease inhibitor cystatin C with β-amyloid revealed by proteolytic excision mass spectrometry and molecular docking simulation. J. Med. Chem. *52*, 2420–2428.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. *36*, D202–D205.

Kennedy, M.B. (1995). Origin of PDZ (DHR,GLGF) domains. Trends Biochem. Sci. *20*, 350.

Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. (2007). Towards drugs targeting multiple proteins in a systems biology approach. Curr. Top. Med. Chem. *7*, 943–951.

Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. (2008). Principles of protein–protein interactions: What are the preferred ways for proteins to interact? Chem. Rev. *108*, 1225–1244.

Kim, D.E., Chivian, D., and Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids. Res. *32*, W526–W531.

Klaus, W., Gsell, B., Labhardt, A.M., Wipf, B., and Senn, H. (1997). The three-dimensional high resolution structure of human interferon alpha-2a determined by heteronuclear NMR spectroscopy in solution. J. Mol. Biol. *274*, 661–675.

Kocher, J.-P.A., Rooman, M., and Wodak, S. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J. Mol. Biol. *235*, 1598–1613.

Kortemme, T., Kim, D.E., and Baker, D. (2004). Computational alanine scanning of protein-protein interfaces. Sci. Signal. *219*, pl2.

Krojer, T., Garrido-Franco, M., Huber, R., Ehrmann, M., and Clausen, T. (2002). Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. Nature *416*, 455–459.

Krojer, T., Sawa, J., Schäfer, E., Saibil, H.R., Ehrmann, M., and Clausen, T. (2008). Structural basis for the regulated protease and chaperone function of DegP. Nature *453*, 885–890.

Krojer, T., Sawa, J., Huber, R., and Clausen, T. (2010). HtrA proteases have a conserved activation mechanism that can be triggered by distinct molecular cues. Nat. Struct. Mol. Biol. *17*, 844–852.

Krüger, D.M., and Gohlke, H. (2010). DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. Nucleic Acids. Res. *38*, W480–W486.

Larkin, M.A., Blackshilds, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWiiliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). ClustalW and ClustalX version 2. Bioinformatics *23*, 2947–2948.

Lefevre, F., Remy, M., and Masson, J. (1997). Alanine-stretch scanning mutagenesis: a simple and efficient method to probe protein structure and function. Nucleic Acids Res. *25*, 447–448.

Li, W., Srinivasula, S.M., Chai, J., Li, P., Wu, J.W., Zhang, Z., Alnemri, E.S., and Shi, Y. (2002). Structural insights into the pro-apoptotic function of mitochondrial serine protease HtrA2/Omi. Nat. Struct. Mol. Biol. *9*, 436–441.

Liang, S., Zhang, C., Liu, S., and Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function. Nucleic Acids Res. *34*, 3698–3707.

Lise, S., Archambeau, C., Pontil, M., and Jones, D. (2009). Prediction of hot-spot residues at protein-protein interfaces by combining machine learning and energy-based methods. BMC Bioinformatics *10*, 365.

Liu, S., Gao, Y., and Vakser, I.A. (2008). DOCKGROUND protein-protein docking decoy set. Bioinformatics *24*, 2634–2635.

Löwer, M., Weydig, C., Metzler, D., Reuter, A., Starzinski-Powitz, A., Wessler, S., and Schneider, G. (2008). Prediction of extracellular proteases of the human pathogen Helicobacter pylori reveals proteolytic activity of the Hp1018/19 protein HtrA. PLoS ONE *3*, e3510.

Lührs, T., Ritter, C., Adrian, M., Riek-Loher, D., Bohrmann, B., Döbeli, H., Schubert, D., and Riek, R. (2005). 3D structure of Alzheimer's amyloid-beta (1-42) fibrils. Proc. Natl. Acad. Sci. USA *102*, 17342–17347.

Melo, F., Sanchez, R., and Sali, A. (2002). Statistical potentials for fold assessment. Protein Sci. *430*, 430–448.

Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2007). Hot-spots - A review of the protein-protein interface determinant amino-acid residues. Proteins *68*, 803–812.

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

Petsalaki, E., Stark, A., García-Urdiales, E., and Russell, R.B. (2009). Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput. Biol. *5*, e1000335.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., and Schulte, K. (2005). Scalable molecular dynamics with NAMD. J. Comput. Chem. *26*, 1781–1802.

Piehler, J., and Schreiber, G. (1999). Mutational and structural analysis of the binding interface between type I interferons and their receptor Ifnar2. J. Mol. Biol. *294*, 223–237.

Piehler, J., Roisman, L.C., and Schreiber, G. (2000). New structural and functional aspects of the type I interferon-receptor interaction revealed by comprehensive mutational analysis of the binding interface. J. Biol. Chem. *275*, 40425–40433.

Pintar, A., Carugo, O., and Pongor, S. (2002). CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics *18*, 980–984.

Qin, S., and Zhou, H. (2007). meta-PPISP: A meta web server for protein-protein interaction site prediction. Bioinformatics *23*, 3386–3387.

Russ, W.P., and Engelman, D.M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. J. Mol. Biol. *296*, 911–919.

Shankar, G.M., et al. (2008). Amyloid-beta protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory. Nat. Med. *14*, 837–842.

Shoemaker, B.A., and Panchenko, A.R. (2007a). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Comput. Biol. *3*, e42.

Shoemaker, B.A., and Panchenko, A.R. (2007b). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput. Biol. *3*, e43.

van Rijsbergen, C.J. (1979). Information Retrieval, Second Edition (London: Butterworths).

Weydig, C., Starzinski-Powitz, A., Carra, G., Loewer, J., and Wessler, S. (2007). CagA-independent disruption of adherence junction complexes involves E-cadherin shedding and implies multiple steps in Helicobacter pylori pathogenicity. Exp. Cell Res. *313*, 3459–3471.

Zhang, C., Liu, S., Zhu, Q., and Zhou, Y. (2005). A knowledge-based energy function for protein−ligand, protein−protein, and protein−DNA complexes. J. Med. Chem. *48*, 2325–2335.