Predicting Molecular Interactions in Structural Proteomics

Irina Kufareva and Ruben Abagyan

CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>185</td>
</tr>
<tr>
<td>Characterizing Molecular Interfaces</td>
<td>187</td>
</tr>
<tr>
<td> Comprehensive Sets of Transient Molecular Interactions in 3D</td>
<td>187</td>
</tr>
<tr>
<td> Properties and Flexibility of Transient Molecular Interfaces</td>
<td>188</td>
</tr>
<tr>
<td>Predicting Protein–Protein Interactions</td>
<td>192</td>
</tr>
<tr>
<td> Physicochemical Properties or Evolutionary Patterns?</td>
<td>192</td>
</tr>
<tr>
<td> Protein Interface Prediction with REVCOM</td>
<td>196</td>
</tr>
<tr>
<td> Optimal Docking Area (ODA)</td>
<td>196</td>
</tr>
<tr>
<td> Protein Interface Recognition (PIER)</td>
<td>198</td>
</tr>
<tr>
<td> Predicting Membrane Interfaces: MODA</td>
<td>200</td>
</tr>
<tr>
<td> Protein Docking</td>
<td>200</td>
</tr>
<tr>
<td>Protein Interactions with Small Molecules</td>
<td>202</td>
</tr>
<tr>
<td> Pocketome Based on Gaussian Convolution of the Lennard-Jones Potential</td>
<td>203</td>
</tr>
<tr>
<td> Cross-Docking of a Ligand to a Single Receptor Conformation with the</td>
<td></td>
</tr>
<tr>
<td> ICM Algorithm Overcomes Limited Induced Fit</td>
<td>203</td>
</tr>
<tr>
<td> Advanced Approaches to Induced Fit in Ligand Docking</td>
<td>204</td>
</tr>
<tr>
<td> Ligand Screening and Profiling</td>
<td>204</td>
</tr>
<tr>
<td>Conclusions</td>
<td>205</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>205</td>
</tr>
<tr>
<td>References</td>
<td>205</td>
</tr>
</tbody>
</table>

INTRODUCTION

As the number of files in the Protein Data Bank (PDB) exceeded 50,000 (representing around 10,000 protein domains at 95% level of sequence identity), it is becoming increasingly important to start developing the understanding of the protein function and the next level of subcellular structural organization.\textsuperscript{1,2} This, among other aspects, requires understanding of what other biological molecules or cellular
structures interact with each domain, which residues are involved in this interaction (e.g., References 3 and 4), and what conformational changes accompany the binding. Structure-based computational approaches to these questions invariably face the issue of protein flexibility, which is further complicated by the existence of unstructured, partially structured, or conditionally structured interfaces. While the dream of predictive millisecond-scale molecular dynamics serving as a “computational microscope” persists (K. Schulten, award lecture at the ISQBP meeting in Ascona, 2008; also Reference 6) and may even be getting more tangible as computers become faster, the ability to make reliable predictions on the basis of such trajectory is still lacking.

The task of predicting molecular interactions has three principal aspects:

A. Predicting the interfaces on a given molecule that are involved in intermolecular interactions. As a subtask one may include predicting a class (but hardly the identity) of the interaction partner (say, protein, peptide, membrane, a small substrate). During the last years, computational methods making these kinds of predictions have improved dramatically and may be quite useful.

B. Predicting the spatial arrangement of two interacting molecules given the apostructures of both, aka docking. Existence of homologous interacting pairs with already solved three-dimensional (3D) structures greatly facilitates solving this problem. However, when such template complex structure is not available, obtaining a crystallographic quality model may be exceedingly difficult due to the induced fit.

C. Predicting the identity of molecules (including proteins) involved in direct transient specific interactions with each other. In the most general form, solving this problem requires precise, large-scale prediction of conformational ensembles and Gibbs free binding energies between all possible pairs of biological molecules, which is unrealistic even with the use of the best state-of-the-art computing resources.

From the biological standpoint, the three aspects should be considered in a different order, by increasing attention to details: C (what) to A (where) to B (how). We, however, order them by their computational complexity. For example, in context of protein–protein interactions, task A is tangible and applicable to thousands of proteins constituting entire structural genomes. Task B, in spite of the achieved limited success in protein docking, largely remains an academic exercise. Task C appears to be impossible to solve due to the enormous complexity of biological systems and the imperfections in existing methods of free energy calculations.

While for protein–protein interactions only task A can be solved with reasonable effort and outcome, all three kinds of predictions are approaching widespread practical use in cases when the interacting partner is a small chemical. Recent advances in small molecule docking and related applications led to a number of successful solutions of tasks A and B in this context. Though more difficult than others, task C becomes quite tangible for druglike compounds and is represented by two kinds of screening:
C1. **Ligand screening**, that is, searching for a natural substrate or a new compound to specifically bind to the source protein.

C2. **Ligand specificity profiling**, that is, searching for the proteins in a subclass or even in the entire structural proteome that bind specifically to a given small molecule.

In this chapter, we present an overview of some methods for predicting the three aspects of molecular interactions. We will focus on targets where a good quality atomic resolution 3D model has either been determined experimentally or can be reliably built by homology (unfortunately, de novo predictors of the three-dimensional structure from the amino acid sequence are still unreliable). We will also focus on the transient, not permanent, interactions. In most cases, permanent binding partners are known in advance, and when this is not the case, they are more easily predictable. We will present an analysis of the induced conformational changes upon binding that create the single biggest challenge for modelers of protein interactions, and describe several methods to overcome this difficulty. Our analysis and the optimization of the prediction methods relied on an ever-growing body of structural data and the improved methods of molecular mechanics with related energy functions.

**CHARACTERIZING MOLECULAR INTERFACES**

**COMPREHENSIVE SETS OF TRANSIENT MOLECULAR INTERACTIONS IN 3D**

Of more than 10,000 unique protein domains found in the 2008 release of the PDB, only about 10% are represented in transient complexes with their biological protein partners. Selection and preparation of a sufficiently large collection of these complexes to be used as a training and validation set is a prerequisite for any study addressing the problem of protein interface prediction. Unfortunately, artificial constructs, crystal packing, and other artifacts present a substantial challenge for both manual and automatic identification of true biological interactions. Although manual intervention during the set collection helps reduce the number of errors, it limits the size of the set and possibilities of timely updates. On the other hand, only a truly large-scale effort can lead to a statistically significant and diverse set without overrepresentation of large families of homologues.

We collected a set of as many as 858 protein domains participating in crystallized transient protein–protein complexes. The entire PDB was organized into families, one family per domain, with each family containing all publicly available good quality structures of the domain with its possible binding partners. To reduce the noise while preserving the automation, we only collected the domains represented by multiple, yet “partner-diverse” structures, and used consistency criteria to achieve the following goals:

- Transient complexes were distinguished from permanent ones based on comparison of PDB complex compositions across the family.
- Each domain was treated in context of its permanent biological unit. In ~20% of the set, the biological unit was found to be different from monomer.
The permanent biological multimers were treated as a whole to avoid potential contamination of the data set with intersubunit (obligate, permanent) interfaces that never get exposed in biological environments.

- Each transient complex was guaranteed to have at least one unbound structure of its receptor domain.
- Multiple protein partners binding to the same or different sites on the protein surface were taken into account.
- Superimposition and structural comparison of the multiple structures provided means for characterization of the induced conformational changes.

For simplicity, we did not include in the set any protein domain that formed permanent heterotrimers or higher multimers, or any domain that was simultaneously bound to more than four distinct protein partners.

The collected set provides a fairly comprehensive representation of transient protein–protein interactions in the PDB. It covered all major classes of biological interactions such as enzyme–inhibitor, hormone–receptor, structural protein, and many types of regulatory interactions. However, antibody–antigen interactions were purposely excluded from the set, as well as all families featuring antibodies as the only type of interacting partner. Epitope prediction must be considered as a standalone task in computational biology. Being different from biological interfaces by both physicochemical properties and (typically) location, epitopes are only recognized by antibodies, naturally selected to target even most noninterface-like patches.

The family size ranged from 2 to 30 (median 6, mean 8.61) structures (Figure 10.1) and was limited by the requirement of using no more than 15 PDB entries and no more than two chains from each entry per protein domain. In a large fraction of cases (361 of 858, 42%), protein domains were found to interact with a variety of protein partners. Such interactions often involved nonoverlapping patches on the protein surface.

Using a similar approach, we also collected a set of ~800 protein domains that have been crystallized apo or in complexes with small molecule ligands. In the following, we present a comparative analysis of the two sets and a comprehensive description of induced fit changes.

**Properties and Flexibility of Transient Molecular Interfaces**

Protein surface patches involved in transient interaction with other proteins or small molecule ligands differ

- from the rest of the surface,
- from permanent multimer interfaces (e.g., Figure 10.2), and
- from each other (small molecule interface vs. protein interface)

by a number of properties. Properties such as relative residue frequencies, physical fields, hydrophobicity, size, charge, evolutionary rates, and so forth have statistically significant differences when compared between classes of protein surface patches (e.g., Reference 16). These properties can be used to predict molecular interfaces (task A).
These properties can be used to predict molecular interfaces
significant differences when compared between classes of protein surface patches
fields, hydrophobicity, size, charge, evolutionary rates, and so forth have statistically
by a number of properties. Properties such as relative residue frequencies, physical
molecule ligands differ
Protein surface patches involved in transient interaction with other proteins or small
properties and flexibility of transient molecular interfaces
description of induced fit changes.
following, we present a comparative analysis of the two sets and a comprehensive
have been crystallized apo or in complexes with
ners. Such interactions often involved nonoverlapping patches on the protein surface.
protein domains were found to interact with a variety of protein part-
than two chains from each entry per protein domain. In a large fraction of cases (361
and was limited by the requirement of using no more than 15 PDB entries and no more
antibodies, naturally selected to target even most noninterface-like patches.
physicochemical properties and (typically) location, epitopes are only recognized by
task in computational biology. Being different from biological interfaces by both
antibody–antigen interactions were pur-
types of regulatory interactions. However, antibody–protein interactions in the PDB. It covered all major classes of biological inter-
bound to more than four distinct protein partners.
permanent heterotrimers or higher multimers, or any domain that was simultaneously
fIgure 10.1
Eight hundred fifty-eight transient protein interfaces represented by two or
more PDB entries. Family size refers to the number of PDB structures representing the same
protein domain in apo form or in complex with transient protein partners. These families
were used to evaluate the induced conformational changes at protein interfaces.

To predict complex geometries (task B) a different question gains primary impor-
tance. Since induced fit presents the major challenge for all docking algorithms, one
needs a clear understanding of the nature and the degree of changes that can happen
upon binding of a protein to a protein or a small molecule partner. Such studies were
previously performed only for small sets of proteins.17,18

To collect the induced fit data, we used the sets of transient protein interactions
in 3D described earlier. Given a family of complexes formed by a particular protein
domain, we compared each complex with all other complexes of the same com-
position (same protein partner in case of protein interactions, same small molecule
for protein–ligand interactions), complexes of other compositions, and unbound
structures. The unbound structures were also compared to one another to assess the
degree of changes stemming from natural protein flexibility rather than induced by
binding partners.

For protein–protein interactions, the obtained data for 858 protein ensembles are
presented in Figure 10.3. In the majority of the cases (77%), comparison of a bound
form of a protein to its unbound form or a complex of different compositions shows
a strong deviation (>1.5 Å) of at least one interface residue. On average, about one-
fourth interface residue backbones deviate above that threshold. Moreover, at least
one interface side chain is displaced by more than 1.5 Å almost always (99%), and
more than one-half of side chains strongly deviate on average. The corresponding

All: Is A used to
represent angstrom?
FIGURE 10.2 Desolvation properties of obligate and transient interfaces in the collected set of 858 protein domains involved in crystallized transient protein complexes in PDB. (a) Buried solvent accessible surface area, (b) desolvation energy. As shown, the transient interfaces are smaller in size, are associated with smaller desolvation penalty, and, therefore, more difficult to predict.
Predicting Molecular Interactions in Structural Proteomics

**FIGURE 10.3** Flexibility of protein–protein interfaces and induced fit. On average, more than one-half of interface side chains are displaced by more than 1.5 Å when compared between different complex compositions (bound vs. unbound/bound to a different protein partner), giving an average interface side chain RMSD of ~4 Å. In contrast, when compared between complexes of the same composition (bound vs. bound to the same protein partner), the expected fraction of strongly deviating side chains is less than one-fourth, and the average interface side chain RMSD is below 2 Å. At least one interface residue backbone deviates by more than 1.5 Å in 78% of the cases, and by at least one side chain in 99% of the cases.
values observed between complexes of the same composition due to natural protein flexibility (white bars), or even between unbound structures (gray bars), are significantly lower. Some induced changes involved large-scale domain, termini, or loop movements and extended as far as 25–50 Å (Figure 10.4).

In contrast, ligand-binding interfaces appear more stable. Being a little smaller in size (the number of residues involved in small-molecule binding is about two-thirds of an average protein interface size; Figure 10.5) they are usually more buried, which restricts potential movements of the interface side chains. Only about 4% of the interface residue backbones deviate above the threshold of 1.5 Å, and about 18% of the side chains (1 to 2 side chains per interface; Figure 10.6).

In short, our analysis proved that in spite of comparable sizes on the interface, small molecules induce less conformational changes upon binding to their receptors than possible protein partners. Even though some exceptions to this rule exist (e.g., activation loop transitions in protein kinases), the task of predicting interfaces, binding geometry, and even identifying the small molecule ligands appears more tangible compared with protein–protein interaction predictions.

**PREDICTING PROTEIN–PROTEIN INTERACTIONS**

**PHYSICOCHEMICAL PROPERTIES OR EVOLUTIONARY PATTERNS?**

Computational methods for protein interface prediction can be divided into two major classes: (1) methods incorporating evolutionary conservation information derived from multiple sequence alignments (MSA) and projected on a protein surface, and (2) those based solely on geometrical and physicochemical properties of the surface.

Methods on the first class rely on the broad evidence of interface residues mutating at slower rates than the rest of protein surface. In general, functionally important surface residues are expected to be conserved. Since interior residues responsible for efficient folding and stability also fall under this category, a strong conservation signal from protein interfaces is only observed when the residue conservation of the interface is compared with that of the surface. Modern multiple sequence alignment methods incorporating residue substitution matrices and phylogenetic trees allow detection of even weak conservation signal. It was argued, however, that conservation score alone is not sufficient for accurate discrimination and can be misleading in several ways. The high variability of alignment composition and extent, unbalanced subfamily representations, and local alignment errors need to be taken into account. The prediction greatly depends on the algorithm of deriving scores from the alignment. Even the most sophisticated algorithms break down on the proteins with no or few orthologs. Most important, many protein interfaces are not expected to be better conserved at all, either because of their function (e.g., the adaptable binding surfaces of the immune system proteins) or because they were formed late in evolution.

Alignment-independent prediction methods rely on an assumption that protein interfaces are different from the rest of the surface by their physicochemical and geometrical properties. Although it was demonstrated that the composition of
FIGURE 10.4  Some examples of large conformational changes at protein–protein interfaces. Whereas the interaction surfaces could be predicted with the methods described in this chapter, the induced rearrangements and the docking pose could not.
protein interface patches had statistically significant biases, the attempts of using the differences for patch discrimination have encountered several difficulties. The physical properties of the interfaces are highly diverse and vary between protein families and complex types. Even within a single interface the binding energy is not distributed evenly among residues; instead, there are so-called hot spots, which contribute most of the interaction energy, while the other interface residues are of relatively minor importance. Finally, the extent and shape of a protein patch in which the small local biases accumulate into a statistically significant signal is not known in advance.

Despite the described difficulties, both approaches have been successfully applied to prediction of protein interfaces on isolated protein structures. The decision about choosing one of the two approaches in each particular case depends on the nature of the protein of interest and available resources. It should be taken into account that in realistic situation, the absence of knowledge about the interaction patch shape (which depends strongly on the partner), and the ambiguity of interface definition make 100% success rates unachievable. On the other hand, all methods provide a statistically significant prediction with high likelihood for the predicted interface residues to be really involved in protein interactions.

In the following we present three methods for prediction of protein interfaces on the surface of isolated proteins with available 3D structures. The first method, REVCOM, belongs to the class of alignment-dependent methods, and the other two, ODA and PIER, to that of the alignment-independent methods.

**FIGURE 10.5** Protein–protein and protein–ligand interface sizes in residues. The average size of a ligand-binding interface constitutes about two-thirds of a protein-binding interface.
Predicting Molecular Interactions in Structural Proteomics

**Figure 10.6** Flexibility of small molecule binding interfaces and induced fit. Less than one-fifth of interface side chains are displaced by more than 1.5 Å when compared between different complex compositions. At least one interface residue backbone deviates by more than 1.5 Å in only 33% of the cases, and by at least one side chain in 77% of the cases. The median number of strongly deviating side chains is one.
**Protein Interface Prediction with REVCOM**

Many residue conservation calculation algorithms only work well if a positionally accurate and compositionally balanced sequence alignment is used, and provide highly variable results otherwise. This variability, along with limited performance of simple conservation measures on weakly conserved interfaces, urged the development of a Bayesian method, robust evolutionary conservation measure (REVCOM), that employs phylogenetic trees to calculate evolutionary rates. REVCOM improves the conservation prediction algorithm by making it more robust and less sensitive to (1) local alignment errors, (2) overrepresentation of sequences in some branches, and (3) occasional presence of unrelated sequences. The method was evaluated and compared with an entropy-based conservation measure on a set of 1494 protein interfaces. By REVCOM conservation measures, 62% of the analyzed protein interfaces were found to be more conserved than the remaining surface at the 5% significance level. A consistent method to incorporate alignment reliability was proposed and demonstrated to reduce arbitrary variation of calculated rates upon inclusion of distantly related or unrelated sequences into the alignment.

REVCOM measures were combined with residue-type distributions in a support vector machine (SVM)-based method for predicting protein interfaces on the structure of an isolated protein. The models were trained and cross-validated on a carefully selected set of biologically relevant protein–protein interfaces. Data for noninterface residues was not removed from the data set. Removing this data reduces the number of false positives in the cross-validation, which provides a biased measure of accuracy since the identity of noninterface residues is not known beforehand for an actual prediction.

The recall and precision achieved by the model on cross-validation were, respectively, 35% higher and 24% higher than expected from a random assignment. Ninety-seven percent of the predicted interface patches overlapped with the actual interface, even though on average only 22% of the surface residues were included in the predicted patch. The receiver operating characteristic (ROC) curve for the fivefold cross-validation on the complete dimer set is shown in Figure 10.7. This curve shows the tradeoff between sensitivity and specificity for the prediction.

**Optimal Docking Area (ODA)**

Extensive experimental and theoretical kinetic studies (reviewed in Reference 46) indicate that specific protein association is often preceded by formation of the encounter complex, which is primarily driven by electrostatics and desolvation. More specific interactions, such as hydrogen bonding and salt bridges, form later and account for the specificity of the final orientation. The important role of desolvation was discussed and used in several applications, for example, to discriminate between docking solutions.

The relative contributions of electrostatic and hydrophobic forces to complex formation vary widely among different complexes. In particular, statistical analyses of known protein–protein complex structures have clearly shown the hydrophobic character of protein interfaces in obligate complexes. Moreover, it has been
demonstrated that large hydrophobic patches correlate with obligate protein interfaces.\textsuperscript{50} However, in transient complexes, the residue preferences at the interface are less pronounced and are not strong enough to unambiguously predict the interface location.

The optimal docking area (ODA)\textsuperscript{11} method represented an attempt to characterize the desolvation properties of protein surfaces and further analyze their role in transient complex formation.

Different subsets of adjacent surface residues were evaluated in order to map low desolvation areas on protein surfaces. In contrast to the common approach that involves dividing the protein surface into equal-area patches, the ODA method generated a series of patches of increasing size with a common center and searched for the ODA, that is, the patch with the lowest desolvation energy. The surface points that generated the ODAs with significant low-energy values were used to define a region over the protein surface most likely to be involved in interaction with other proteins. Instead of accounting for trivial hydrophobicity, the patch surface energy was evaluated based on atomic solvation parameters previously derived from octanol/water transfer experiments and adjusted for protein–protein binding.\textsuperscript{51} The method was applied and shown to successfully identify \textit{nonobligate} protein interaction sites.

\textbf{FIGURE 10.7} Receiver operator characteristic (ROC) curve for fivefold cross-validation of the REVCOM-based protein interface prediction method on the set of 632 dimer interfaces. The area under the curve is 0.79. False positive rate = (false positives)/(false positives + true negatives) and true positive rate = (true positives)/(true positives + false negatives). The default SVM operating point with a decision value cutoff is marked.
Protein Interface Recognition (PIER)

Inspired by the success of ODA, the protein interface recognition (PIER)\textsuperscript{52} method applied machine learning techniques for further optimization of atom desolvation parameters in context of protein–protein associations. This led to an alignment-independent method for protein interface identification with improved reliability, accuracy, and speed.

Interface prediction by PIER starts with generation of surface patches in the spirit of ODA, however, for each protein the patch generation radius was fixed and given by the formula

$$d = \sqrt{27.235 + 0.018 \times ASA_{1000}^{400}},$$

where $ASA_{1000}^{400}$ is the accessible surface area (ASA) of the isolated molecule, trimmed to fit in the range of $[400,1000]$ Å\(^2\). For proteins whose total ASA exceeded 10000 Å\(^2\), using this equation produced the distance of 14 Å, and resulted in surface patches with average ASA between 900 and 1000 Å\(^2\). The deviations in the values of ASA between different patches reflected the curvature and packing of the surface atoms within the patch. For each patch ($P$), 12 patch descriptors were calculated. These descriptors simply represented total ASA of 12 subresidue atomic groups, whose representation was previously found significantly different between interface and noninterface patches. The PIER value for the patch was calculated as a linear combination of the obtained descriptors and further transferred to individual residues within the patch. Based on the per-residue PIER decision value, the residues are predicted to be either interface or noninterface.

The proposed alignment-independent method demonstrated improved performance over the previously published methods. On a diverse benchmark of 748 proteins known to be involved in homo- and heterodimeric interactions, permanent as well as transient, the overall precision at the residue level was 60% at the recall threshold of 50%. The method was also tested on other benchmarks. Using the method, we identified potential new interfaces and corrected mislabeled oligomeric states. Several predictions with PIER are presented in Figure 10.8.

A cross-validated partial least squares (PLS) regression algorithm\textsuperscript{53} provides a natural environment to incorporate and evaluate the relative contribution of new surface descriptors including those derived from sequence alignments. In particular, we found that when added to the set of PIER descriptors, the evolutionary signal contributed as little as 7%–10%, with the rest 90%–93% being provided by atomic group composition descriptors. Adding evolutionary signal only marginally influenced the prediction performance; moreover, for certain classes of proteins, using conservation scores actually resulted in deteriorated prediction. This exercise demonstrated that while both alignment-dependent and alignment-independent approaches maybe successful, combining them does not improve the success rate significantly.
Predicting Molecular Interactions in Structural Proteomics

Inspired by the success of ODA, the protein interface recognition (PIER) method applied machine learning techniques for further optimization of atom desolvation parameters in context of protein–protein associations. This led to an alignment-independent method for protein interface identification with improved reliability, accuracy, and speed.

Interface prediction by PIER starts with generation of surface patches in the spirit of ODA, however, for each protein the patch generation radius was fixed and given by the formula

\[ r = \frac{ASA}{k} \]

where \( ASA \) is the accessible surface area (ASA) of the isolated molecule, trimmed to fit in the range of \([400,1000]\) Å\(^2\). For proteins whose total ASA exceeded 10000 Å\(^2\), using this equation produced the distance of 14 Å, and resulted in surface patches with average ASA between 900 and 1000 Å\(^2\). The deviations in the values of ASA between different patches reflected the curvature and packing of the surface atoms within the patch. For each patch, 12 patch descriptors were calculated. These descriptors simply represented total ASA of 12 subresidue atomic groups, whose representation was previously found significantly different between interface and noninterface patches. The PIER value for the patch was calculated as a linear combination of the obtained descriptors and further transferred to individual residues within the patch. Based on the per-residue PIER decision value, the residues are predicted to be either interface or noninterface.

The proposed alignment-independent method demonstrated improved performance over the previously published methods. On a diverse benchmark of 748 proteins known to be involved in homo- and heterodimeric interactions, permanent as well as transient, the overall precision at the residue level was 60% at the recall threshold of 50%. The method was also tested on other benchmarks. Using the method, we identified potential new interfaces and corrected mislabeled oligomeric states. Several predictions with PIER are presented in Figure 10.8.

A cross-validated partial least squares (PLS) regression algorithm provides a natural environment to incorporate and evaluate the relative contribution of new surface descriptors including those derived from sequence alignments. In particular, we found that when added to the set of PIER descriptors, the evolutionary signal contributed as little as 7%–10%, with the rest 90%–93% being provided by atomic group composition descriptors. Adding evolutionary signal only marginally influenced the prediction performance; moreover, for certain classes of proteins, using conservation scores actually resulted in deteriorated prediction. This exercise demonstrated that while both alignment-dependent and alignment-independent approaches maybe successful, combining them does not improve the success rate significantly.

**FIGURE 10.8** Precise identification of multiple interfaces (white patches) on the surface of an isolated protein with PIER.
Predicting Membrane Interfaces: MODA

Reversible recruitment of soluble proteins to cellular membranes represents another class of important biological intermolecular interactions and underlines many cellular processes and events. In many cases, this recruitment is not mediated by explicit covalently attached membrane anchors and is based solely on the surface properties. A number of computational methods predicting peripheral membrane interactions are based on homology with known membrane-targeting modules. They, however, fail to encompass the entire biological space and yield many false positives, since conservation of the structural motif does not always mean conservation of the function (i.e., PH domains and C2 domains). By combining the PIER approach with the analysis of protein surface curvature and local electrostatic potential, we developed a fast computational method, which, given a 3D structure of a protein, predicts its membrane propensity and identifies the membrane contact elements on its surface. The method was named MODA, or membrane ODA. MODA prediction of the protein–membrane interfaces on known peripheral proteins correlated well with the experimental data; it achieved the precision of 64% at 50% recall at the residue level (7% precision expected from random prediction). Moreover, using the method, we were able to identify several novel potential peripheral proteins, which were subsequently validated using NMR spectroscopy and micelle titration (Figure 10.9).

Protein Docking

Predicting the geometry of association of two proteins known to bind each other is also a legitimate, albeit more rare and much more difficult task than predicting protein interaction patches. Its applicability is limited to the cases where both the identity and three-dimensional structure of the second partner are known. Since the beginning of the Critical Assessment of PRedicted Interactions (CAPRI) competitions, two main problems make the protein docking task unreliable: (1) an unpredictable degree of induced rearrangements or restructuring upon complex formation, and (2) poorly predictable specifics of conformational changes even for small degrees of induced fit (e.g., Reference 59).

The large degree of induced rearrangements may include (1) conditional structuring of the previously unstructured loop or region (e.g., the activation loop of protein kinases, or loop 6 of TIM α/β barrels); (2) domain swapping (e.g., References 62 and 63); (3) large displacements of secondary structure elements, especially N- or C-terminal (e.g., helix 12 in nuclear receptors, opening of two parallel β-strands in serpins); and (4) large relative movements of protein domains, for example, the “swivelling” in pyruvate phosphate dikinase, Ca-dependent domain movements in calmodulin, and integrin domains. The changes may have some functional context, for example, being associated with protein activation like in serpins or kinases. Sadly, no computational method so far claimed any success in predicting those changes without prior knowledge of the answer.

For cases where the restructuring of domains, parts, and the interface atoms is minor, the protein docking has been a half-solved problem (e.g., References 8 and 64). If the changes are limited to minor backbone shifts and rearrangements of a
Predicting Molecular Interactions in Structural Proteomics

Reversible recruitment of soluble proteins to cellular membranes represents another class of important biological intermolecular interactions and underlines many cellular processes and events. In many cases, this recruitment is not mediated by explicit covalently attached membrane anchors and is based solely on the surface properties. A number of computational methods predicting peripheral membrane interactions are based on homology with known membrane-targeting modules. They, however, fail to encompass the entire biological space and yield many false positives, since conservation of the structural motif does not always mean conservation of the function (i.e., PH domains and C2 domains). By combining the PIER approach with the analysis of protein surface curvature and local electrostatic potential, we developed a fast computational method, which, given a 3D structure of a protein, predicts its membrane propensity and identifies the membrane contact elements on its surface. The method was named MODA, or membrane ODA. MODA prediction of the protein–membrane interfaces on known peripheral proteins correlated well with the experimental data; it achieved the precision of 64% at 50% recall at the residue level (7% precision expected from random prediction). Moreover, using the method, we were able to identify several novel potential peripheral proteins, which were subsequently validated using NMR spectroscopy and micelle titration (Figure 10.9).

Predicting the geometry of association of two proteins known to bind each other is also a legitimate, albeit more rare and much more difficult task than predicting protein interaction patches. Its applicability is limited to the cases where both the identity and three-dimensional structure of the second partner are known. Since the beginning of the Critical Assessment of PRedicted Interactions (CAPRI) competitions, two main problems make the protein docking task unreliable: (1) an unpredictable degree of induced rearrangements or restructuring upon complex formation, and (2) poorly predictable specifics of conformational changes even for small degrees of induced fit (e.g., Reference 59). The large degree of induced rearrangements may include (1) conditional structuring of the previously unstructured loop or region (e.g., the activation loop of protein kinases, or loop 6 of TIM α/β barrels); (2) domain swapping (e.g., References 62 and 63); (3) large displacements of secondary structure elements, especially N- or C-terminal (e.g., helix 12 in nuclear receptors, opening of two parallel β-strands in serpins); and (4) large relative movements of protein domains, for example, the “swivelling” in pyruvate phosphate dikinase, Ca-dependent domain movements in calmodulin, and integrin domains. The changes may have some functional context, for example, being associated with protein activation like in serpins or kinases. Sadly, no computational method so far claimed any success in predicting those changes without prior knowledge of the answer.

For cases where the restructuring of domains, parts, and the interface atoms is minor, the protein docking has been a half-solved problem (e.g., References 8 and 64). If the changes are limited to minor backbone shifts and rearrangements of a

(a) Von Willebrand factor, A3 domain
(b) ARF1 GTPase
(c) AT1G acetyltransferase
(d) Methionine sulfoxide reductase

(e-h) Predicted membrane-docking positions confirmed by NMR spectroscopy

> FIGURE 10.9 Discovery of novel peripheral membrane proteins and identification of membrane interfaces with MODA. MODA-positive residues are shown in sticks; backbone amides with chemical shifts as CPK.
fraction of the interface side chains, in 40% to 80% of the cases (depending on a problem set) the top scoring pose will have most of the residue–residue contacts correctly predicted and the interface patch can be around 2 Å RMSD from the crystallographic solution. Of course the caveats include the fact that the correct solution can be either completely missed because of a large rearrangement, or not ranked as a top solution due to inability of the refinement to find all smaller changes necessary for the native interface and scoring function to recognize the near-native solution.65

The first successful combination ab initio of all-atom protein–protein docking and interface refinement that led to a single solution was published in Nature Structural Biology in 1994.66 The uncomplexed lysozyme was docked to the HyHel5 antibody without any prior knowledge about the epitope location. The procedure used all-atom representation and worked in two natural phases. The initial positional sampling was performed by searching from multiple starting positions with the pseudo-Brownian jumps followed by minor relaxations of torsional angles. At the refinement phase, the 30 lowest-energy conformations from all the searches were then refined by a global optimization of the interface side chains and positional variables67 with relaxed backbone using an extended set of energy terms with desolvation and side-chain entropy. The near-native solution found was surprisingly close to the crystallographic structure (root mean square deviation [RMSD] of 1.57 Å for all backbone atoms of lysozyme) and had a considerably lower energy (by 20 kcal/mol) than any other solution after refinement.

Although the first phase of the docking procedure can be performed on various resolution models using a variety of search methods (most notably, multidimensional fast Fourier transform, or FFT) and scoring functions68 (also reviewed in References 47, 64, 69, and 70), it became clear after 1994 that the refinement with a detailed atomic model and detailed energy function is necessary to both improve the ability to recognize the native solution and improve the quality of the pose. The 1994 refinement protocol using internal coordinate sampling was essentially preserved in the later modifications of the Internal Coordinate Mechanics–based protocols leading to the high accuracy predictions in the first two CAPRI meetings71,72 and was later “rediscovered.”73 Refinement procedures using molecular dynamics were also proven to be successful.74–76

Overall, the unpredictable and unexpected rearrangements of protein segments and domains, their partially unstructured character,3 as well as a relative paucity of the available high quality and complete (i.e., “dockable”) three-dimensional models of the interacting components prevent protein docking from becoming a routine practical tool of molecular biology comparable to x-ray crystallography.

PROTEIN INTERACTIONS WITH SMALL MOLECULES

Small molecules are much easier to deal with. First, their interfaces with proteins are guaranteed to be small in contrast to proteins where a large variation of interface area is observed. Consequently, specific ligands cannot afford to avoid a tight fit to one of the ligand’s conformers, which, in turn, makes the prediction of ligand binding sites for nanomolar ligands a feasible task (see References 77 and 78 for a list).
Furthermore, for any new ligand it is relatively easy to predict how it docks into a “preformed” pocket because the number of degrees of freedom including the positional ones is small enough for a reliable sampling and structure prediction.

**Pocketome Based on Gaussian Convolution of the Lennard-Jones Potential**

Since strong, small ligands require a tight fit whether the binding has polar, charge, hydrophobic, or mixed nature, the most general form of interatomic interaction potential, van der Waals potential, may be predictive of a small-molecule interface on a protein. Straightforwardly computed with the Lennard-Jones equation, it is not predictive; however, its aggregate value allows to clearly distinguish between an area occupied by a typical small molecule and the rest of the protein surface.\(^7^9\) The cumulative value can be mathematically defined as a convolution of the potential with a Gaussian kernel. We tried different radii and found that the radius of 2.6Å results in the best performance of the method.

The generated potential can be contoured at a certain level to give an envelope that resembles the envelope of strong binders in most cases. The predicted envelope is particularly valuable for open and extended pockets with many binding surfaces and does not require anything but a set of coordinates. We used this method to generate the bounding box for unbiased ligand–receptor cross-docking.\(^8^0\)

The method is fast and can be applied to a currently known structural proteome of an organism. For example, we applied the Gaussian convolution pocket prediction method to a small set of proteins from the malarial parasite, *Plasmodium falciparum*, but it can be scaled up to much larger sets of proteins.\(^8^1\) The site prediction methods can also be adjusted to a particular task and involve other principles, from geometrical to evolutionary (see Reference 82 for a review).

**Cross-Docking of a Ligand to a Single Receptor Conformation with the ICM Algorithm Overcomes Limited Induced Fit**

As described earlier, in most cases, ligand binding induces only small conformational changes. This opens possibilities for successful prediction of binding poses of ligands in the pockets that have not been co-crystallized with them (the so-called cross-docking problem). To quantify the extent that rigid-body receptor docking can overcome the extent on the induced fit, we performed cross-docking simulations for 1000 ligands and 300 proteins from different complex structures (Bottegoni, submitted, 2008). It was shown that for about one-half of the cases (46%) such cross-docking with the ICM docking algorithm\(^8^3\) predicts a correct near-native geometry as the top-scoring solution. To overcome the induced fit obstacle for the remaining cases, one of the induced fit protocols need to be applied.

If multiple conformations of a pocket are known in advance, a simple protocol, called MRC (multiple receptor conformation) docking can be used.\(^8^4\) This procedure can also be accelerated using a so-called 4D approach in which the receptor conformer becomes a variable in the docking procedure (Bottegoni, submitted, 2008).
**Advanced Approaches to Induced Fit in Ligand Docking**

Often only one conformer of the receptor is known. It was shown that to dock a ligand correctly to an incorrect pocket, one can simply delete the uncertain parts in a certain way and rely on the rest of the pocket to position the ligand correctly. The idea came from the “omission” modeling proven to be successful for side chain prediction. The uncertain parts can later be refined around the identified ligand position. The scan alanines and refine (SCARE) algorithm replaces pairs of neighboring side chains by alanines and docks the ligand to each gapped version of the pocket. Unlike in Reference 86, the selection of the residue pairs by SCARE is performed in an unbiased systematic fashion, which results in a general algorithm applicable on a scale of complete proteomes. All docked positions are scored, refined with original side chains and flexible backbone, and rescored. In the optimal version of the protocol, pairs of residues were replaced by alanines and only one best scoring conformation was selected from each “gapped” pocket for refinement. The optimal SCARE protocol identifies a near-native conformation (under 2 Å RMSD) as the lowest rank for 80% of pairs if the docking bounding box is defined by the predicted pocket envelope, and for as many as 90% of the pairs if the bounding box is derived from the known answer with a 5 Å margin.

**Ligand Screening and Profiling**

As shown earlier, predicting the correct binding pose of a ligand required some form of treatment of the induced fit. Those relevant receptor models provide necessary starting points for (1) identification of a small-molecule binder to a protein of interest in a large database of chemicals (ligand screening), or (2) identification of possible protein targets for a single, given small molecule (ligand profiling).

The DOLPHIN protocol (Kufareva, submitted, 2008) gives a specific recipe for predicting the changes associated with binding of so-called type II kinase inhibitors to their target kinases. Similar to the SCARE protocol, the type II compatible model of a target kinase is built by omission of the part of the structure with consequent ligand docking and full-atom complex refinement. Using the ICM binding score that has been previously derived from a multireceptor screening benchmark as a compromise between approximated Gibbs free energy of binding and numerical errors, we screened a large database of kinase ligands and showed that the correct type II ligands for the modified kinase are selected in the top 1.5%–3.5% of the database. Further on, based on experimental data, we derived the kinase-specific systematic free energy contributions originating from the different abundance of the relevant conformer and other protein features. By combining these energy contributions with the calculated binding energies, we could identify the kinase specificity profile of individual type II ligands. The ligand specificity profiling approach can be extended to the whole structural interactome as the relevant models and protein offsets become available.
CONCLUSIONS

The structural interactome is the next great challenge for structural proteomics that clearly cannot be solved by crystallography alone and requires computational structural methods because of enormous combinatorics of possible interactions. In this chapter we reviewed (1) the interaction interface prediction methods; (2) the induced conformational changes for different types of interacting partners; and (3) the ability of the available methods to predict the binding pose, its relative score or binding energy, and binding specificity. The recent methods for small-molecule ligands, including pocket prediction, receptor flexible docking (SCARE, 4D, and MRC protocols), and improved scoring, are paving the way to predictive structural chemogenomics.

For protein–protein (or membrane) interactions, the most practical kinds of structure-based predictions include predicting interfaces with other proteins (without the prediction of their identity) and with a membrane. Predicting the geometry, identity, or strength of protein–protein interactions de novo from unbound structures, without evolutionary heuristics, seems to be much more problematic because of a much larger role and scale of unpredictable, induced conformational changes. A comprehensive benchmark described in this review may help to develop and test better methods for predicting large-scale conformational changes upon interaction.

ACKNOWLEDGMENTS

We thank Maxim Totrov and Eugene Raush for valuable discussions and technical help. The project was partially funded by National Institutes of Health/National Institute of General Medical Sciences grants 5-R01-GM071872-02 and 1-R01-GM074832-01A1.

REFERENCES


