1

Computational Methods for Prediction of Protein-Protein Interaction Sites

Aleksey Porollo and Jaroslaw Meller
University of Cincinnati
USA

1. Introduction

Studies of protein-protein interactions play a central role in understanding protein function in biological systems, closing the gap between large-scale sequencing efforts and medically relevant outcomes. Increasingly, protein interaction interfaces that mediate communication between proteins are becoming targets for therapeutics, offering a possibility to disrupt critical interactions and specifically attenuate function (Fletcher and Hamilton, 2007; Fry, 2006).

Efforts to catalog, characterize, and link protein interactions with disease states and other phenotypes are ongoing, building on improvements in experimental techniques, such as high throughput two-hybrid assays or chip-based proteomics. Significant progress has also been achieved in structural genomics, providing detailed information for a growing number of macromolecular complexes and interaction interfaces by means of X-ray crystallography, NMR spectroscopy and other methods.(Aloy et al., 2005; Slabinski et al., 2007)

Despite impressive progress, existing experimental methods for mapping protein interactions suffer from many limitations. High throughput methods, such as two-hybrid or chip-based essays, are characterized by high rates of false positives and false negatives (Bader and Chant, 2006; Han et al., 2005), requiring further validation and detailed characterization of individual interactions. Obtaining detailed high-resolution information about protein interaction interfaces can also be challenging in many instances.

For example, some complexes may not crystallize, or crystallize in a different than biologically relevant conformation. X-ray crystallography may also fail when multiple and incompletely mapped interactions or membrane domains are involved.(Lacapere et al., 2007) This is exacerbated by the fact that each protein has been estimated to have around 9 distinct interacting partners (and some are estimated to have hundreds interactants), with majority of the implied complexes unlikely to be resolved experimentally in the foreseeable future.(Aloy and Russell, 2004; Ritchie, 2008)

Limitations of experimental techniques and attempts to circumvent the problem by focusing directly on protein interactions create an opportunity for computational approaches to complement and facilitate experimental efforts in that regard. In particular,

statistical and machine learning-based approaches are being increasingly used to facilitate identification of protein interfaces. There are a growing number of methods for protein interaction sites prediction that vary in terms of principles of the recognition of interaction interfaces, descriptors used to identify interacting sites (feature space) and learning algorithms used.

From the point of view of a representation used to capture characteristics of interaction interfaces, one may distinguish two main groups of methods. The first group attempts to predict interaction sites using sequence information only.(Gallet et al., 2000; Ofran and Rost, 2007) The second group of methods, takes available structural information into account (Fariselli et al., 2002; Lichtarge et al., 1996), typically involving the identification of sites on the surface of a monomeric structure that are either evolutionarily conserved (as for example in the pioneering evolutionary trace method by Lichtarge and colleagues (Lichtarge et al., 1996)), or have a propensity for interaction interfaces (see, e.g., (Jones and Thornton, 1997)).

Although evolutionary trace methods are relatively insensitive to structural detail and can identify conserved "hot spots", their overall accuracy is limited.(Caffrey et al., 2004; Porollo and Meller, 2007) On the other hand, detailed structural information can be used to characterize patches on the surface of a protein in terms of their geometric and other properties (see, e.g., (Bordner and Abagyan, 2005; Koike and Takagi, 2004; Neuvirth et al., 2004)). Structural conservation can also be taken into account when multiple structures within families are available.(Chung et al., 2006; Ma et al., 2003)

While structural information improves prediction accuracies (with the risk of increasing the sensitivity to the choice of a specific structure), challenges remain and new insights are required to improve state-of-the-art in the field.(de Vries and Bonvin, 2008; Zhou and Qin, 2007) Further progress also requires continued systematic evaluation of new methods. In this regard, the lack of standard definitions and consistent evaluation criteria adds to the challenge and often makes direct comparison of existing methods impossible.

One problem that contributes to the difficulty of fair evaluation and objective comparison of different methods is related to the uncertainty concerning the definition of the negative class. The assignment to the "non-interacting" class is at best tentative, given the incompleteness of information regarding all possible interactions and interacting partners. Despite the growing number of resolved structures of protein-protein complexes, another challenge is the relative paucity of carefully curated and properly stratified (to represent different types of complexes) benchmarks.

This chapter reviews computational methods for the prediction of protein interaction sites, with a primary focus on structure-based approaches. The goal is to help the reader better understand the underlying concepts and limitations pertaining to current methods in the field. A number of methodological issues related to the training and validation of such methods are discussed as well. The benchmarks and assessment included in this chapter should also help making an informed decision as to when computational predictions can be regarded as sufficiently confident for a particular system of interest to warrant further experimental validation.

2. Definition of protein-protein interaction site

The recognition of protein-protein interaction sites can be cast as a classification problem, i.e., each amino acid residue is assigned to one of the two classes: interacting or non-interacting residues. Consequently, the problem may be solved using statistical and machine learning techniques, such as neural networks (Ofran and Rost, 2003b; Zhou and Shan, 2001) or Support Vector Machines (Bock and Gough, 2001; Yan et al., 2004).

A clear definition of interacting residues is obviously required in order to predict whether a given amino acid residue is involved in protein-protein interactions. However, many alternative definitions are being used in the field. As the definition of an interaction site varies from one prediction method to another, it becomes difficult to directly compare their performance.

2.1 Commonly used definitions

If available, high resolution structural data readily provides a basis for atom or residue based definition of interaction sites. In fact, prediction methods discussed in this chapter primarily use information from resolved protein complexes to define the positive ("interacting") and negative ("non-interacting") classes. Protein quaternary structures are typically resolved by X-ray crystallography, and less frequently by NMR-spectroscopy or other techniques (Protein Data Bank, PDB – http://www.pdb.org/). While providing a high resolution structure, crystallographic data often remains inconclusive regarding the nature of the observed intermolecular contacts between protein chains. In particular, some of the observed contacts (and the resulting putative interaction interfaces) may be the result of crystal packing, rather than representing biologically relevant interactions.

A number of methods have been introduced to facilitate the process of filtering out crystal packing artefacts. Here, we used the approach adopted by the PISA server (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html). PISA discriminates crystal packing contacts from the functional protein–protein interaction using the size of solvent exposed area buried during association, as well as the number of residues constituting the interface, the number of salt and disulphide bridges at the interface, and the difference in approximate solvation energy upon complex formation.(Henrick and Thornton, 1998; Krissinel and Henrick, 2007)

Two different approaches are commonly used to define an interaction site based on 3D structural data: (i) interatomic distance and (ii) change in accessible surface area (ASA) upon complex formation. Following the first approach, interaction sites can be defined based on the distance between non-hydrogen atoms of different protein chains. For example, distance cutoffs of 4Å (Bordner and Abagyan, 2005); 4.5Å (Hamer et al., 2010); 5Å (Chen and Zhou, 2005); or 6Å (Ofran and Rost, 2003b) are used. This way of defining interaction sites is likely to miss some interchain contacts when water molecules are involved. A polar solvent, such as water, may bridge the interaction between two charged groups of amino acids that are too far apart to form a direct hydrogen bond.(Janin, 1999) In this regard, Neuvirth *et al.* introduced the Connolly interface index (CII) that is computed for circles of radius 10 Å around anchoring dots on the surface of monomeric structures. Atoms with CII above certain threshold are assigned to be interaction sites.(Neuvirth et al., 2004)

The second approach defines an interaction interface by using the concept of solvent accessibility or ASA. Specifically, ASA or the solvent accessibility of an amino acid residue in an unbound protein chain is contrasted with the corresponding ASA value for the same residue in a complex. Residues with a significant difference in ASA between the isolated chain and complex structures are then classified as "interacting". The following cutoffs for ASA were used: the loss of > 99% ASA for a given atom (Bradford and Westhead, 2005); a residue loses > 1Å^2 ASA in the complex (Chen and Jeong, 2009; Jones and Thornton, 1995; Liang et al., 2006); a residue ASA change by more than 20Å^2 (Kufareva et al., 2007); relative solvent accessibility (RSA) of a given residue decreased by more than 4% and its ASA decreased greater than 5Å^2 (Porollo and Meller, 2007). The latter definition uses relative ASA to address the considerable difference in size of amino acids, e.g. between glycine and tryptophan.

Both approaches require high resolution structural data. However, the interatomic distance based approach seems to be more sensitive to problems with missing atoms or atoms with multiple occupancies. Table 1 illustrates the difference in the protein interface recognition resulting from alternative definitions. As can be seen from the table, the same protein quaternary structure may yield different subsets of residues deemed to be interaction sites, therefore leading to different prediction models and their reported performances.

In what follows, we will refer to protein interfaces derived using our own ASA-based definition, dRSA > 4% and dASA > 5Å^2 (Porollo and Meller, 2007), unless stated otherwise. This definition takes into account both relative and absolute change in ASA, and it attempts to filter out noise related to variation in RSA observed in structures resolved under different conditions, or for closely related homologs.

Definition	Chain	Residues at the interface	Interface ASA, Å^2
dASA > 1Å^2	I	Y35 T41 C42 H57 C58 D60 R61 N95 T96 D97 D98 V99 A99A L143 L151 W172 T175 C191 Q192 G193 S195 T213 S214 F215 V216 S217 R217A L218 K224	830
	E	I18 I19 L20 I21 R22 C23 A24 M25 L26 N27 P29 R31 E46 G47 S48 C49 A52 C53 F54	994
dRSA > 4% and dASA > 5Å^2	I	Y35 T41 H57 D60 R61 T96 D97 V99 A99A L143 L151 W172 T175 C191 Q192 G193 S195 S214 F215 V216 S217 R217A L218	810
	E	I18 I19 L20 I21 R22 C23 A24 M25 L26 N27 P29 R31 E46 G47 S48 C49 A52 F54	989
dASA > 20Å^2	I	Y35 H57 R61 T96 D97 V99 W172 Q192 S195 F215 V216 R217A L218	692
	E	I19 L20 I21 R22 C23 A24 M25 L26 N27 P29 R31 E46 S48 C49	938

Table 1. The effects of using alternative definitions of protein interaction interfaces for a specific hetero-dimeric complex (PDB ID 1fle); dASA is the total loss of ASA for a given protein chain upon complex formation.

It should be noted that information on protein interaction sites may be also derived from the alanine scanning mutagenesis (ASM). Systematic replacement of the residues at the protein interface with alanine enables the evaluation of individual contribution of each interaction site to the binding energy. In this regard, the Alanine Scanning Energetics database (ASEdb, http://www.asedb.org/) provides ASM data on a number of protein-protein, as well as on some protein-DNA and protein-ligand interactions (Thorn and Bogan, 2001)

However, ASM approach is very costly and laborious, thus considerably limiting the number of comprehensively studied proteins. A protein interface needs to be approximately defined beforehand to limit the number of alanine mutants to evaluate. Results of ASM may not necessarily indicate the contribution to the binding energy, as some alanine mutants may cause an adverse protein conformational change and therefore indirectly decrease the efficacy of the protein-protein binding. Moreover, some protein-protein interactions are allosterically regulated, and ASM may not reflect the actual driving forces for a given protein complex. Nevertheless, such data is of great value and may be used as an additional validation of prediction methods. For example, it was used to evaluate ability of the methods ISIS (Ofran and Rost, 2007) and APIS (Xia et al., 2010) to identify hot spots.

2.2 Mapping interaction sites

Methods that do not require information about the interacting partner(s) are the primary focus of this chapter. These methods aim at the recognition of either individual residues, surface patches, or whole interaction interfaces using only sequence, structure and other information about an individual target protein, assuming that it is involved in some sufficiently stable interactions.

In light of the above, an important part of defining the residues as interaction sites is to retrieve as much information as possible on physical interactions for a given protein. Published studies on methods for the prediction of protein-protein interaction quite often ignore the fact that most proteins have multiple interaction partners that are mediated by alternative or overlapping interfaces. Therefore, using just one particular complex to identify the interaction interface and to derive the corresponding definition of the positive class, while ignoring all other complexes and interactions involving the same target protein chain (or its close homolog), may result in highly biased estimates of both false positive and false negative rates.

With the significant growth of structural data, the problem can be addressed by taking into account interaction sites from alternative complexes that contain the same protein chain or its close homologs. Interaction sites identified in such homologs can be mapped to a representative sequence in order to enable more sensitive prediction and perform its fair accuracy evaluation. Figure 1 illustrates this issue for two proteins resolved in complexes with different partners.

The protein shown in the left panel, caspase-9, utilizes overlapping interfaces for homo-oligomerization (PDB ID 1jxq), and for its interaction with ecotin (PDB ID 1nw9). However, the former protein-protein interaction involves many more residues than the latter interaction (affected ASA 1954Å^2 and 1019Å^2, respectively). If the definition of the positive ("interacting") class in caspase-9 were to be derived from the complex with ecotin (1nw9),

the accuracy of any method predicting correctly also the more extensive interface would have been wrongly underestimated. This problem can be addressed by mapping the interface from the homooligomer into the target structure, leading to the union of homo-dimerization and caspase-9/ecotin interfaces to be taken as the true positive class.

The second example on the right illustrates the mapping of the known interfaces into the beta subunit of *E. coli* DNA polymerase III. In addition to homodimerization interface (PDB ID 2pol), physical interactions with the delta subunit of the gamma complex (PDB IDs 1jqj, 1jql) and DNA polymerase Pol IV (PDB ID 1unn) are mapped. Again, without this additional mapping step, prediction of these alternative interfaces would be considered as false positives during the evaluation process.

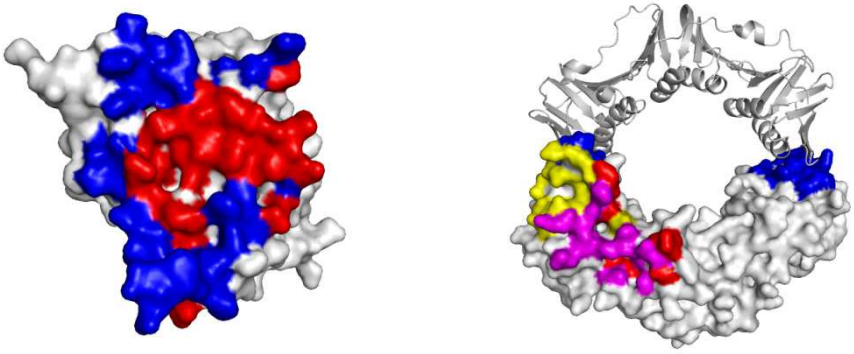


Fig. 1. Mapping interfaces from alternative protein complexes: A. Interaction interfaces in caspase-9, derived from the complex with ecotin (PDB ID 1nw9, chains B-A, shown in red) and caspase-9 homooligomer (PDB ID 1jxq, chains A-B), which includes both red and blue patches; B. Interaction interfaces mapped into DNA Pol III from the homodimer of the beta subunit of DNA Pol III (PDB ID 2pol, blue), delta subunit (PDB IDs 1jqj and 1jql, red), and DNA Pol IV (PDB ID 1unn, yellow), with the overlap of the latter two shown in magenta. Interfaces identified by using the SPPIDER server (http://sppider.cchmc.org/) and mapped into the target structure by using POLYVIEW-3D (http://polyview.cchmc.org/polyview3d.html).

The mapping, though, needs to be performed carefully, keeping in mind some important caveats. Sequence homology-based approach assumes that similar protein sequences adopt the same 3D fold and carry the same function, which is not always true. For example, paralogs may evolve to have distinct interaction partners and therefore perform different functions while having high sequence homology. Mapping interaction sites from such homologs might then result in incorrect expansion of the positive class to include patches utilized by other proteins with sequence similarity but distinct functions. In this context, one should comment that many methods for the prediction of interaction sites incorporate information about evolutionary profiles of protein families (e.g., obtained using PSI-BLAST to generate PSSM (Altschul et al., 1997)). Therefore, at least in some cases such methods arguably identify sites with a propensity to interact within the whole family, rather than just for the target protein.

Interactions specific to only some (or even only one) family members may require the identification of distinct interaction patches, rather than considering the problem of predicting the union of alternative interaction interfaces. Thus, mapping interaction interfaces might not be appropriate for evaluation of methods that attempt to predict such individual interaction patches. On the other hand, if ANY interaction patch that corresponds to a stable protein complex is to be found, then the union of all known interfaces constitutes the best approximation of the positive class and should be used for evaluation of the overall accuracy. As indicated above, this issue is often ignored altogether, even though it highlights the difficulty with a proper definition of a classification problem that best captures biologically relevant information while providing sufficiently "accurate" predictions.

Conversely, some protein domains with conserved 3D structure and specific function may be very divergent in terms of amino acid sequence, and only structure alignment might be able to detect such distant similarity. For example, PB1 domain displays low sequence homology between proteins, but it has a highly conserved secondary structure pattern and the overall 3D fold.(Lamark et al., 2003) While having just a few conserved residues playing a role of hot spots, this domain is widely utilized in various biological systems for interactions between the PB1-containing proteins to conduct cell signaling.(Moscat et al., 2006)

A PDB-wide structure alignment remains a computationally challenging task when it comes to a large protein set compiled for training or benchmarking a method for protein-protein interaction prediction. However, some current efforts, including for example the Dali database (http://ekhidna.biocenter.helsinki.fi/dali/start) (Holm et al., 2008), provide valuable resources in this regard. There have been also a number of studies published on the structure-based mapping of interaction sites, utilizing different schemes of hit weighting and homology recognition.(Albou et al., 2011; Oldfield, 2002; Park et al., 2001; Xu and Dunbrack, 2011)

However, it remains to be seen how structure-based mapping methods can deal with situations when a protein undergoes a significant conformational change upon complex formation (e.g., in case of calmodulin), and a structure alignment is likely to fail to identify similarity between apo- and holo-forms. Most likely, the future methods will utilize a balanced combination of sequence- and structure-based homology in order to more accurately map interaction sites from the known physical interactions. In this work, in order to test the effects of mapping interaction sites from multiple resolved complexes, we used a sequence homology-based mapping with conservative thresholds for homology hits: 70 or 90% of sequence identity. The interaction sites mapping process was automated through the SCORPPION web-server (http://scorppion.cchmc.org/).

3. Types of protein complexes

Biological diversity is very well represented at molecular level, in particular showing broad versatility in protein-protein interactions. Protein complexes can be classified into a number of broad categories, for example as homo- and hetero-oligomers; transient and obligatory (permanent), rigid and flexible complexes. Homo-oligomers are complexes consisting of two or more protein chains with identical amino acid sequence. Accordingly, assemblies of chains with different sequences are hetero-oligomeres. The number of chains participating in the assembly dictates the distinction on dimers, trimers, tetramers, and so forth.

Obligatory complexes (sometimes called obligomers) are considered to be protein assemblies that perform function only in the coupled state, whereas transient complexes are formed by proteins that were found to exists as monomers and to function separately as well. Rigid complexes may be considered as products of interaction between stable rigid-body domains. Flexible complexes, on the other hand, are formed when one or more constituting proteins undergo significant conformational changes.

Systematic analysis of the known protein complexes by several studies resulted in a number of observations that have significantly influenced the field of protein-protein interaction sites prediction. Ofran and Rost suggested that there are at least 6 types of contacts in proteins that display distinct amino acids compositions and contact preferences.(Ofran and Rost, 2003a) Thus, methods utilizing statistical contact propensities in their prediction models have to take into account different types of interactions. Another study found that even within a single interface the composition of amino acids varies depending on where the interacting amino acids are located, in the core of the interface or at its rim.(Chakrabarti and Janin, 2002)

A closer look at transient complexes was presented in (Nooren and Thornton, 2003). The study distinguished "weak" and "strong" homodimers, and it found that weak transient homodimers demonstrate smaller, more planar and polar interfaces compared to permanent homodimers, whereas strong transient homodimers undergo large conformational changes upon complex formation, and demonstrate larger, less planar, and more hydrophobic interfaces. Interestingly, only weak transient homodimers were found to have residues at interfaces more conserved than other surface residues, whereas other proteins with different oligomeric states showed no pronounced amino acid conservation.

These findings were further supported by the study on a larger set of protein complexes.(Caffrey et al., 2004) Comparing the conservation scores derived from multiple sequence alignments to orthologs vs. paralogs, the study demonstrated that residues at the interfaces are rarely more conserved than other residues on the protein surface. This observation implies that prediction models solely based on evolutionary profiles are likely to have limited overall accuracy.

Another large scale study has recently reported the results of PDB-wide analysis of protein-protein interactions. Both sequence and structure based characteristics of protein interfaces were characterized, with special focus on proteins with multiple interaction partners.(Kim et al., 2006) This analysis showed that, while there are ancient interfaces conserved across archea, bacteria, and eukaryotes (attributed primarily to symmetric homodimers), by and large interfaces are not conserved and vary in shape and amino acid composition due to broad diversity of interactions and interaction partners. The suggested classification introduced as many as 6000 different types of interfaces that are available for search and matching from the SCOPPI database (http://www.scoppi.org/).

4. Benchmarks of protein complexes

Benchmarks specifically designed for the training and evaluation of methods for the recognition of protein-protein interaction sites are critical for further progress in the field. Such benchmarks should allow an unbiased and fair evaluation of prediction methods. Consequently, benchmark sets used for comparison of different methods should comprise a

diverse representative set of protein-protein interactions and contain no redundancy to the training sets used by individual methods.

The uncertainty of the negative class assignment further complicates the choice of appropriate benchmarks. Designing a dataset that includes only carefully curated and well-studied proteins, or their domains, with all known physical interactions mapped, may result in a very limited number of data points for training and validation. As a more feasible alternative one could consider assembling several diverse and non-redundant training and validation data sets that include complexes of different type and are characterized by some level of completeness of information regarding interactions and interaction sites.

As a result of these difficulties, there is no established gold standard in the field. Most of the published methods refer to their own compilation of protein complexes derived from PDB. Here, we consider three protein sets used in the literature. The first compilation of protein complexes is a benchmark set for protein-protein docking, current version 3.(Hwang et al., 2008) For this set, proteins in bound and unbound state were retrieved from PDB in a semi-automated manner. Current version contains the total of 124 test cases; among those 88 are rigid-body cases, 19 of medium difficulty, and 17 difficult cases, which are classified by the degree of conformational change at the interface upon complex formation.

While the primary purpose of Hwang *et al*. benchmark was to evaluate the protein docking methods, many protein interface prediction methods used it for their own and comparative evaluation.(de Vries and Bonvin, 2011; de Vries et al., 2006; Fiorucci and Zacharias, 2010; Guharoy and Chakrabarti, 2010; Li et al., 2008; Liu and Zhou, 2009; Qin and Zhou, 2007; Zhou and Qin, 2007) However, a thorough analysis of this benchmark set led us to conclusion that it is not suitable for evaluation of the methods predicting protein-protein interaction sites. For example, it contains 25 antibody-antigen cases (PDB IDs: 1fc2, 1ahw, 1bvk, 1dqj, 1e6j, 1jps, 1mlc, 1vfb, 1wej, 2fd6, 2i25, 2vis, 1bj1, 1fsk, 1i9r, 1iqd, 1k4c, 1kxq, 1nca, 1nsn, 1qfw, 2jel, 1bgx, 1e4k, 2hmi), which are asymmetrical functional protein-protein interactions, i.e. while one partner (in general: antibody, protease, or major histocompatibility complex) is evolved to bind its substrate, the second partner is not (except for the protease inhibitors).

Therefore, all antibody-antigen complexes were removed from the set. In addition, protein chains no longer available in PDB (PDBID_ChainID: 1cd8_B, 1ml0_B, 2pab_C, 2pab_D, 2viu_C, 2viu_E, 1aly_B, 1aly_C, 1jb1_B, 1jb1_C), difficult to interpret in terms of protein chains (1hia_A, 1hia_B, 1n8o_B, 1n8o_C) or too short (1n8o_A, 1k74_B, 1mzn_B, 1zgy_B) were removed. Finally, before using this benchmark set for evaluation of protein interface prediction methods, redundant chains were also removed.

The second benchmark set represents 85 cases of proteins found in PDB both in bound and unbound state.(Albou et al., 2009) No complexes with asymmetrical function are included, such as antibody-antigen cases and others listed above. This set represents diverse protein-protein interactions and allows the evaluators to estimate the role of conformational change on the accuracy of the methods, when predictions using bound structures *versus* unbound are compared. However, the set contains two cases, when only α-carbon coordinates are available (PDBID_ChainID: 3dpa_A and 2tld_I). These cases may be challenging to prediction methods that rely on high resolution data with all atoms resolved.

The last benchmark set to be used in this work is the control set of the SPPIDER method.(Porollo and Meller, 2007) It was compiled based on the protein complexes

deposited in PDB after the compilation of the training set for the same prediction method. This manually curated and non-redundant (to the training set and within itself) set includes 149 protein chains, deemed to be sufficiently diverse and representative enough to be used for cross-validation studies. The only update to the set involved replacing the chain 1r72_A by 1xcb_A, as the PDB entry 1xcb now supersedes 1r72. In what follows, this set is referred to as SPPIDER149.

Table 2 and Figure 2 summarize the three datasets described above, after removing problematic cases from the first set, and redundant proteins from the first two sets. Redundancy was defined in terms of sequence homology: BLAST e-value < 0.001 when the alignment covers at least 70% of the query sequence (derived from the ATOM section of a PDB file). 150 chains derived from complexes in the first set and 78 chains in the second set were found non-redundant, and these (sub-) sets will be referred to as Hwang150B and Albou78B, respectively. The corresponding sets of chains that were retrieved from their unbound structures will be referred to as Hwang150U and Albou78U, respectively.

Dataset	Total chains	Families	Domains
Hwang150B	150	42	107
Albou78B	78	16	44
SPPIDER149	149	76	75

Table 2. Protein families and domains represented in non-redundant chains of the three benchmark sets used in this work. Families and domains defined according to the Pfam database (http://pfam.sanger.ac.uk/) (Finn et al., 2008) and mapped using sequence based search as implemented in SCORPPION (http://scorppion.cchmc.org/).

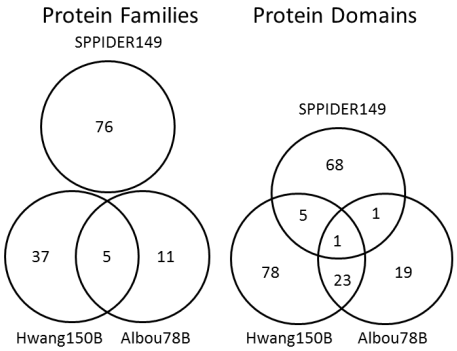


Fig. 2. Overlap between protein families (left) and domains (right) identified within the three benchmark sets used here.

Low to no overlap between the datasets discussed here is observed in terms of protein families and domains, suggesting a broad coverage of protein-protein interactions. This bodes well for estimates of the performance on different types of protein interfaces. On the other hand, the training sets for tested methods might partially overlap with the benchmark sets used here, leading to potentially overestimated accuracy.

Mapping of known interaction interfaces from alternative complexes was performed for each set using different approaches discussed in Section 2.2. Table 3 shows the number and

fraction of interacting residues for each protein set. Interaction sites were derived from (i) asymmetric units defined in the original PDB files, (ii) biological units (BUs) as defined by Protein Quaternary Structure (PQS) database, and (iii) BUs as defined by the PISA database. In addition, interaction sites were mapped from the PISA-based BUs of their close homologs using sequence identity 90 and 70% as a cutoff (Table 4). The estimates of accuracy for methods compared here were overall quite similar, and only the results for the latter threshold are reported in the following sections of the chapter.

PDB also provides its own definition of biological units that differs from PISA.(Xu and Dunbrack, 2011) PDB defines biological units as separate models in the same PDB file. In addition, both PISA and PDB may rename chain labels starting from 'A' within each BU. This all makes it difficult sometimes to trace back the chains from the asymmetrical unit in automated manner. To be consistent, we will map interaction sites from BUs as defined by PISA. However, when no information can be mapped for a given chain, due to technical difficulties or inconsistency in BU definition, we will use a PDB-based asymmetric unit for the mapping of interaction sites.

Dataset	Total residues / On the surface	PDB-based, %	PQS-based, %	PISA-based, %
Hwang150B	31208 / 24687	19	21	19
Albou78B	17412 / 13375	16	16	15
SPPIDER149	25883 / 20885	29	28	28

Table 3. Summary of the benchmarks used in this work with regards to the total number of residues, residues on the surface, and percentage of the surface residues found to be at protein interfaces derived from the asymmetric unit (PDB-based), and biological units (PQS-based and PISA-based), respectively.

Dataset	Total / Surface residues	SI70	SI90
Hwang150B	31208 / 24687	10011	9674
Hwang150U	32471 / 24595	10201	9661
Albou78B	17412 / 13375	5819	5506
Albou78U	16838 / 12342	5572	5294
SPPIDER149	25883 / 20885	7863	7668

Table 4. Summary of the benchmarks used in this work with regards to the total number of residues, residues on the surface, and interacting residues on the surface mapped to representative protein chains using BUs derived from the PISA database and 70 or 90% sequence identity cutoffs (SI70 and SI90), respectively.

5. Prediction methods

All prediction methods can be broadly classified by the type of data they use as an input. Sequence-based methods rely on some combination of the following protein features: amino acid hydrophobicity, evolutionary profile (e.g., similarity scores or Shannon entropy), amino acid composition or propensity to be at the interface, predicted structural features (e.g., secondary structure, solvent accessibility, order/disorder region, *etc.*), or their derivatives like mean or weighted average over a sequence window.

The structure-based methods, on the other hand, also utilize features derived from a 3D protein structure, such as solvent accessibility and secondary structure states, local topology (e.g., protrusions and cavities), hydrophobic and polar surface patches, temperature or B-factors (for X-ray based structures), etc. In addition, there are a number of methods built using a consensus of the individual predictors with reportedly improved accuracy.(de Vries and Bonvin, 2011; Huang and Schroeder, 2008; Qin and Zhou, 2007) However, consensus-based methods are not discussed here in detail, as the goal is to evaluate the discriminating power of the underlying principal features for each representative method.

Described below are selected structure-based methods with at least somewhat orthogonal feature spaces that were available as web-servers at the time of data preparation for this work. Methods are listed in the order of the publication year of the original work.

Evolutionary trace (ET) method (Lichtarge et al., 1996) identifies evolutionary conserved residues and maps them onto a protein 3D structure. Conserved residues in the core of a protein are deemed to be structurally important, whereas those on the surface are assumed to be functionally important. The method starts from constructing a multiple sequence alignments, and partitions the aligned sequences into groups by using their mutual sequence similarity. For each group, a consensus sequence is defined highlighting the positions with invariant amino acids. Consensus sequences are further aligned to identify (i) conserved residues across the entire protein family; (ii) class-specific residues that are invariant in some groups; and (iii) neutral residues that are not preserved in any single sequence group. Conserved and class-specific residues are then mapped onto 3D structure. Clusters of such residues on the surface of a protein structure are predicted to be functional. The ET method is available at http://mammoth.bcm.tmc.edu/ETserver.html

ConSurf (Glaser et al., 2003) follows a similar approach by mapping the evolutionary conserved residues on 3D protein structure. The difference lies in computing the conservation scores that are relative with respect to other residues in a given protein. In addition, the outcome of the method is sensitive to the quality of multiple sequence alignment and to the overall length of a query sequence. For example, two 3D structures of the same protein, but with different sequence length representing its resolved part, may result in different location of the most conserved residues. The ConSurf method is available at http://consurf.tau.ac.il/, whereas its pre-computed results for the PDB deposited proteins are available from the ConSurfDB database (http://consurfdb.tau.ac.il/).

It should be noted that the two methods described above were not designed to identify specifically protein-protein interaction sites, but rather to reveal any functional residues, e.g. involved in protein-DNA or protein-ligand interactions. However, since the authors of these methods refer to identification of protein interfaces as examples in their original publications, we chose these methods to serve as a separate group of predictors that rely primarily on evolutionary information, and can be contrasted with structure-based methods.

PROMATE (Neuvirth et al., 2004) considers residues on the surface of a protein structure within 10Å circles around a given point. Spatially neighboring residues provide the following descriptors: (i) statistically derived chemical composition of binding sites, such as

propensity of individual amino acids, atom types, pairs of amino acids, and collective chemical properties (positively and negatively charged, polar, hydrophobic, and aromatic residues); (ii) evolutionary conservation in terms of diagonal elements of the PSI-BLAST-derived position specific scoring matrix (PSSM); (iii) distance in the sequence between residues in the circle; (iv) secondary structure states, including extent of the loops. Additionally, temperature factors (B-factors) and bound waters are incorporated into the model whenever available. These descriptors are combined to yield a cumulative score that allows the circles to be classified as Interface, Non-interface, or Boundary. The neighboring circles are further clustered to define predicted interface patches. PROMATE is available at http://bioinfo.weizmann.ac.il/promate/

Cons-PPISP (Chen and Zhou, 2005) employs a consensus of neural networks trained on (i) the position specific similarity scores derived from the PSI-BLAST multiple sequence alignment and (ii) observed (in the target structure provided as input) solvent accessibility for spatially neighboring residues. In addition to validation on crystal structures, cons-PPISP was shown to provide accurate prediction of protein interfaces for a set of 8 NMR-derived complexes, non-redundant to its training set. The web-server is available at http://pipe.scs.fsu.edu/ppisp.html

WHISCY (de Vries et al., 2006) introduces prediction scores that are based on evolutionary and structural information. Conservation of residues on the surface is computed as the corrected sum of similarity scores between amino acids at a given position by pairwise comparison of a query sequence and sequences from a multiple alignment. Similarity scores are taken from the Dayhoff mutation matrix. ASA is the only structural information used. WHISCY is available at http://nmr.chem.uu.nl/Software/whiscy/index.html

PIER (Kufareva et al., 2007) combines (i) statistically derived interatomic contact potentials, (ii) physical descriptors, such as observed solvent accessibility for separate atomic groups within amino acids, and (iii) sequence alignment based features, in particular, three different conservation scores (frequency-based, similarity matrix-based, and entropy-based). The surface of a protein structure is divided on individual patches. Using the descriptors listed above, all patches obtain a set of cumulative scores that further fed to a partial least squares (PLS) based regression model to predict protein interfaces. Since the PIER scoring heavily relies on atomic resolution, it may have difficulties with incomplete or of low resolution crystal structures. The corresponding prediction server is available at http://abagyan.ucsd.edu/PIER/

SPPIDER (Porollo and Meller, 2007) is a neural network-based method that uses the difference between predicted from sequence and observed in an unbound structure RSA of amino acid residue as a novel and highly informative signal of interaction sites. Solvent accessibility prediction methods tend to predict residues at protein interfaces as buried, which is consistent with the fact that they are indeed getting buried upon complex formation, even though they are exposed in an unbound structure. The SABLE (Adamczak et al., 2004) method for RSA prediction was used to generate the input for SPPIDER. Additional features include averaged over spatially neighboring residues of (i) RSA predicted by SABLE; (ii) evolutionary conservation (in terms of Shannon entropy) of amino acid type, charge, hydrophobicity, and side chain size; (iii) amino acid contact numbers and hydropathy constants. The server is available at http://sppider.cchmc.org/

6. Evaluation

6.1 Accuracy measures

Prediction of protein interaction sites is typically cast as a classification problem. Therefore, a number of commonly used measures for two class classification problems can be employed to evaluate the accuracy. These measures include the two-class classification accuracy (Q_2), recall or sensitivity (R), and precision or specificity (P), all expressed as percentage.

$$Q_2 = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100\% \quad (1)$$

$$R = \frac{TP}{TP+FN} \cdot 100\% \quad (2)$$

$$P = \frac{TP}{TP+FP} \cdot 100\%, \quad (3)$$

where TP are true positives, TN – true negatives, FP – false positives, and FN – false negatives.

However, since the number of interaction sites can be much smaller than the number of non-interacting residues, the classification problem at hand may be highly unbalanced. As a result, the measures listed above may be difficult to interpret and compare for different benchmarks. For example, with 90% of data points assigned to the negative class, a baseline classifier that predicts all residues as non-interacting achieves numerically high 90% classification accuracy. To provide a measure that balances sensitivity and specificity of predictions, the Matthews correlation coefficient (MCC) is often used (4) together with other measures. MCC ranges from -1, indicating an inverse prediction, through 0, which corresponds to a random classifier, to +1 for perfect prediction.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \quad (4)$$

Other measures that can be used to assess and compare classification methods are area under the receiver operating characteristic (ROC) curve and F-measure.

6.2 Performance of selected methods

The performance of several representative methods discussed in the previous section is assessed here in order to compare more systematically individual methods, and to quantify the effects of mapping additional interaction interfaces and using truly unbound structures. Different aspects of the performance are evaluated using benchmark datasets described in section 4 (SPPIDER149, Hwang150B/U, and Albou78B/U).

For all evaluations, only residues with RSA of at least 5% were considered, thus excluding all fully buried residues in a given protein conformation. For methods providing a real valued score, multiple thresholds were tested as a basis for projection into two classes. The results for the best performing threshold in terms of MCC are reported in Tables 5 through 9. The following values were found to be optimal for each method: ET with residues being ranked 1 (out of top 1, 5, and 10 rankings evaluated), ConSurf with evolutionary rank ≥ 5 (5,

7, 9 evaluated), WHISCY with threshold ≥ 0 (0, 0.18 evaluated), PIER with threshold ≥ 15 (0, 15, 30 evaluated), and SPPIDER with threshold ≥ 0.3 (0.3, 0.5, 0.7 evaluated).

Method	SPPIDER149	Hwang150B	Albou78B
ET	0.08	0.04	0.01
ConSurf	0.12	0.07	0.02
PROMATE	0.10	0.10	0.09
Cons-PPISP	0.30	0.22	0.17
WHISCY	0.19	0.11	0.08
PIER	0.37	0.27	0.22
SPPIDER	0.41	0.28	0.20

Table 5. The performance of representative methods measured using MCC on three different sets, with only the original PDB complexes used to define the positive class.

As can be seen from Table 5, the overall accuracy of the methods evaluated here is rather limited. The two best performing methods, i.e., PIER and SPPIDER achieve MCC of about 0.4 for SPPIDER149 set, 0.3 for Hwang150B, and 0.2 for Albou78B, respectively. Similar relative drop in accuracy is also observed for other methods, indicating that Hwang150B and Albou78B sets are more difficult to classify. This can be explained in part due to a larger imbalance between positive and negative classes in these benchmarks, especially in the Albou78B dataset (see Table 3).

Method	SPPIDER149		Hwang150B		Albou78B	
	R, %	P, %	R, %	P, %	R, %	P, %
ET	7.03	43.92	3.99	28.18	2.84	17.55
	6.39	51.73	3.44	48.89	3.57	60.60
ConSurf	65.27	32.87	61.42	22.18	55.17	16.40
	63.00	40.97	55.91	41.07	53.19	41.66
PROMATE	3.91	60.71	4.06	48.98	3.69	43.43
	3.22	64.29	2.56	63.78	1.85	58.29
Cons-PPISP	33.40	60.59	26.25	42.42	22.46	34.80
	29.39	69.12	19.35	67.62	15.33	64.40
WHISCY	29.38	45.42	21.15	29.77	20.49	21.83
	26.66	54.32	17.21	51.71	16.53	48.38
PIER	61.10	52.62	49.66	37.46	45.43	30.61
	54.38	60.31	38.64	60.86	31.20	56.99
SPPIDER	80.36	48.47	63.15	34.11	56.22	26.49
	73.14	56.81	53.04	59.82	43.48	55.52

Table 6. The effect of mapping interaction sites from homologous protein complexes on recall (R) and precision (P): the first line in each row shows R and P using original PDB complexes, whereas the second line indicates accuracy derived after mapping interaction sites using PISA BUs and homologous chains with 70% sequence identity.

It should be noted that due to a sufficiently large number of data points (surface residues, see Table 3) included in each benchmarks, each of the correlation coefficients reported above

is statistically significantly different from 0 with a p-value < 0.05. Nevertheless, practical applicability of methods that achieve correlations of 0.2 and lower has to be judged using also other criteria and specific examples. In particular, evolutionary methods achieve very limited accuracy in this test, even though they may provide biologically valuable insights, as discussed later.

The effects of mapping interaction residues from alternative complexes are illustrated in Table 6 using measures of sensitivity and specificity. The accuracy using the assignment of the positive class (interaction sites) derived from the original complexes is compared to the accuracy obtained re-labeling the "non-interacting" residues in mapped interfaces as "interacting" sites. Due to largely canceling effects of decreased rates of false positives and increased rates of false negatives, the mapping of interaction sites from PISA biological units does not affect significantly the performance of the prediction methods in terms of MCC, although a systematic small drop in accuracy is observed in most cases (data not shown).

However, as can be seen from Table 6, all methods show a drop in recall while precision improves when mapping is applied. These results also allow one to trace how the trade-off between sensitivity and specificity was optimized for different methods. One striking example is ConSurf vs. ET comparison. On the other hand, most structure-based methods provide fairly well balanced predictions. In particular, precision improves considerably, with only a relatively limited drop in recall for the best performing SPPIDER method, followed by PIER and Cons-PPISP. The observed ranking could reflect the fact that SPPIDER was trained (although on a different set without homology to SPPIDER149 set) using mapping from alternative complexes to reduce the noise in learning from data and to provide a more balanced classification problem.

Method	Hwang150B SI70	Hwang150U SI70	Albou78B SI70	Albou78U SI70
ET	0.03	0.00	0.06	0.08
ConSurf	0.03	0.05	0.00	0.00
PROMATE	0.06	0.05	0.04	0.01
Cons-PPISP	0.20	0.18	0.14	0.13
WHISCY	0.09	0.16	0.06	0.08
PIER	0.24	0.23	0.15	0.11
SPPIDER	0.29	0.29	0.17	0.14

Table 7. The effect of the bound versus unbound state of the protein structures used as an input in terms of MCC. In all cases, interacting residues were mapped using homology to PISA BUs with 70% sequence identity.

The impact of conformational change and the use of structures in bound as opposed to unbound state as an input is assessed in Table 7. For that purpose, the overall accuracy in terms of MCC is compared using two pairs of sets of bound (taken from a complex by simply ignoring other chains) and truly unbound structures: Hwang150B *vs.* Hwang150U and Albou78B *vs.* Albou78U, respectively. Slight decrease in performance is observed for all but one structure-based method, the exception being WHISCY. The latter method starts from a low level, though. In addition, the WHISCY server did not generate results for a number of more difficult cases, suggesting that this trend might not hold on other data sets.

While the drop in accuracy is limited for other methods tested, it should be emphasized that benchmarks included here sample relatively small conformational changes due to induced fit. Therefore, further systematic studies will be required to better delineate the range of applicability of structure-based method for the recognition of protein interaction sites.

Table 8 demonstrates how the performance estimates can be inflated when accuracy measures are computed based on all residues as opposed to computing the accuracy for each protein and then averaging over all proteins. Per protein averages, together with measures of variance (here we report standard deviations), allow one to assess better the range of expected accuracies for individual proteins. As can be seen from Table 8, the observed large standard deviations suggest large protein to protein variation and indicate that all tested methods fail dramatically for at least some proteins. It should be also noted that using per protein measures PIER is the top performing method, followed by SPPIDER and Cons-PPISP.

Method	MCC	Q_2, %	R, %	P, %
ET	0.06±0.12 0.08	65.64±17.83 71.21	9.60±16.07 7.03	29.35±35.01 43.92
ConSurf	0.12±0.15 0.12	54.44±8.16 52.80	64.54±14.06 65.27	39.61±22.69 32.87
PROMATE	0.07±0.13 0.10	64.01±19.63 71.16	5.72±8.93 3.91	28.30±39.31 60.71
Cons-PPISP	0.23±0.23 0.30	69.52±13.23 74.15	37.50±22.11 33.40	58.99±29.71 60.59
WHISCY	0.14±0.20 0.19	67.39±13.14 71.03	26.58±19.79 29.38	42.64±28.00 45.42
PIER	0.30±0.23 0.37	71.18±11.47 72.54	58.73±24.80 61.10	55.22±27.09 52.62
SPPIDER	0.29±0.20 0.41	66.94±13.82 69.39	79.16±24.79 80.36	49.19±21.69 48.47

Table 8. Comparison of the accuracy measures calculated per residue by merging data from all chains (the bottom line in each row) and per protein averages and standard deviations (the top line in each row), using the SPPIDER149 set (similar effect is observed on other benchmarks).

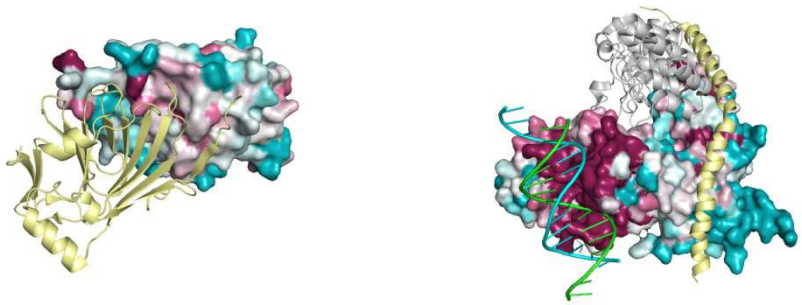
Not all web-based implementations of the methods are reliable. While requesting and retrieving predictions from the evaluated servers, we faced multiple failures. Table 9 illustrates the reliability of the corresponding servers from the user`s point of view by presenting the numbers of proteins failed to be processes within each benchmark set. The most reliable web-servers appear to be PIER and SPPIDER, whereas ET, ConSurf, and WHISCY are quite unreliable, which makes it more difficult to evaluate servers on a large scale.

Prediction methods that seemingly perform poorly according to some evaluation criteria can still greatly facilitate further experimental and computational studies on protein interactions. One might argue that predicting possible interaction interfaces should be directed at the recognition of the sites that contribute most to the binding energy. Such hot

spots also represent the most natural target for further validation, e.g., using mutagenesis, or as targets for therapeutics.

Method	SPPIDER149	Hwang150B	Hwang150U	Albou78B	Albou78U
ET	14	14	28	8	8
ConSurf	21	12	13	4	4
PROMATE	1	3	8	3	12
Cons-PPISP	7	3	1	0	4
WHISCY	34	15	17	8	9
PIER	0	0	0	0	0
SPPIDER	0	0	0	0	0

Table 9. The number of proteins not included in each benchmark due to problems with the retrieval of the results as an indicator of the reliability of web-servers tested.



A **B**

Fig. 3. Examples of protein interaction sites predicted by ConSurf: **A.** A successful identification of the protein interface for the homodimer of phosphoglucose isomerase (PDB ID 1qxr, chain A); **B.** A multi-interface protein (CSL transcription factor) illustrates possible confusion with DNA binding sites that are the most slowly evolving residues at the surface of the protein in this case (PDB ID 2fo1, chain A). Residues in magenta are the most conserved, whereas variable sites are colored using cyan (see the ConSurf documentation).

In this context, a special note needs to be made on the performance of evolutionary methods, such as ET and ConSurf. As we mentioned before, these methods were not designed specifically to predict protein-protein interaction sites, but rather to identify evolutionary conserved residues. Therefore, these methods may not able to discriminate between protein-protein, protein-ligand (e.g., co-factor or substrate), and protein-DNA/RNA binding sites. An example of such a case is shown in Figure 3.

On the other hand, highly conserved residues that are exposed on the surface of a protein are very likely functionally relevant, irrespective of the actual involvement in interaction. Despite all the limitations, evolutionary methods for the prediction of interaction sites have significantly contributed to the mapping of protein interactions and other functional

annotations, see e.g., (Kniazeff et al., 2002; Shenoy et al., 2006) and (He et al., 2003; Lietha et al., 2007), for ET and ConSurf, respectively.

7. Discussion and conclusions

Protein-protein interactions are essential for enzymatic functions, signal transduction, cell cycle regulation and other fundamental biological processes. In addition to addressing the fundamental questions of molecular biology, identification of residues involved in protein-protein interactions has important medical relevance. Combined with recent advances in genome sequencing it facilitates delineating natural functional variants from pathological mutants, and conducting 'molecular diagnostics' as part of personalized medicine.(Su et al., 2011) Detailed structural information on thousands of protein complexes also stimulates growth in the field of rational drug design by providing a new class of targets that include known protein interaction interfaces.(White et al., 2008)

However, experimental identification and validation of a protein interface remains a challenging task, both in terms of labor and cost. Therefore, efforts to map and characterize protein interactions can considerably benefit from computational biology and structural bioinformatics. In particular, methods that integrate sequence and structure information achieved accuracies that are useful in selecting and prioritizing targets for mutagenesis and other experimental studies.

In this chapter, we reviewed state-of-the-art in the field of computational prediction of protein-protein interaction sites. We evaluated some representative methods using several published benchmarks of protein complexes. The overall accuracy of existing methods, in accord with other recent evaluations, was found to be limited (the Matthews correlation coefficient between the predicted and true class assignment of up to 0.4). Therefore, further concerted efforts will be required to improve state-of-the-art in the field. To that end, we discussed the need for standard definition of protein interaction sites, developing more comprehensive benchmark protein sets, and appropriate ways of measuring/reporting the accuracy of predictions.

We quantified the effects of taking into account multiple interaction interfaces and using as an input unbound structures that were resolved without interacting partners. Both of these issues are often ignored when evaluating the performance of interaction sites prediction methods. Yet, they are shown to impact significantly the estimates of performance. These two issues also highlight more fundamental difficulties with the definition of the negative class and current attempts to cast the problem in a computationally feasible way.

Casting the prediction of interaction sites in terms of a two-class classification problem requires that examples of the negative ("non-interacting") class be used for the training. With data points representing both "interacting" and "non-interacting" residues, a decision boundary separating the two classes can be optimized. These negative examples are defined in most cases by simply taking the complement of the positive class, i.e., all other (surface exposed) residues that are not known to be involved in interactions.

Consequently, without mapping known interfaces alternative complexes, residues within such interfaces are incorrectly regarded as "non-interacting". This could introduce problems in training, as misclassified vectors from the negative class may coincide with the bulk of the

density for the positive class. One strategy to address this issue is to filter out such difficult cases. As an alternative, one could also consider one-class approaches, in which only the positive class examples are used to learn a predictor. On the other hand, if residues from multiple complexes are systematically mapped, as advocated here, the negative class assignment as a source of noise should be gradually reduced with the progress in experimental mapping of interaction sites.

Conformational changes upon complex formation pose another problem for the methods considered here. Protein flexibility and the induced fit effects upon complex formation are assumed to be limited. Obviously, this assumption does not hold in many instances of protein-protein interactions (and sometimes it breaks spectacularly, e.g., when the co-folding of otherwise disordered interacting domains occurs). Therefore, methods presented here are of limited applicability when large conformational changes or flexible domains are involved.

It should be also stressed that even a limited induced fit can pose significant challenges for structure-based methods. Simply ignoring all but one chain in a protein complex, and thus taking a *de facto* bound conformation as input, may lead to spurious effects in training and overly optimistic estimates of accuracy. For example, low B-factors of surface residues, which can be "locked" in a specific conformation by interactions with a co-factor, may not be a true signal of interaction sites (in many cases the opposite can actually be observed). Features that are capable of identifying interaction sites starting from a truly unbound structure should be emphasized.

Reliable identification of residues that participate in binding to other proteins can help direct and streamline mutagenesis and other experimental studies, and to facilitate efforts to map entire interactomes. It can also reduce the levels of false positives (by assessing compatibility between predicted interfaces), and false negatives (by helping identify novel interactions) observed for experimental approaches that are used to map protein interactions. Another promising application is protein docking, in which predicted interfaces can be used for evaluating and ranking potential complex structures (de Vries and Bonvin, 2011), in analogy to docking methods that utilize limited NMR data. (Dominguez et al., 2003; Kohlbache et al., 2001)

Further progress in the field will require new insights to overcome current limitations, as well as careful assessment of the accuracy in order to address possible biases in training and validation. Constant improvements in experimental techniques and a growing number of resolved macromolecular complexes, from which to learn better predictors, bode well for future efforts in this regard.

8. References

Adamczak, R., Porollo, A., and Meller, J. (2004). Accurate prediction of solvent accessibility using neural networks-based regression. Proteins *56*, 753-767.

Albou, L. P., Poch, O., and Moras, D. (2011). M-ORBIS: mapping of molecular binding sites and surfaces. Nucleic Acids Res *39*, 30-43.

Albou, L. P., Schwarz, B., Poch, O., Wurtz, J. M., and Moras, D. (2009). Defining and characterizing protein surface using alpha shapes. Proteins *76*, 1-12.

Aloy, P., Pichaud, M., and Russell, R. B. (2005). Protein complexes: structure prediction challenges for the 21st century. Curr Opin Struct Biol *15*, 15-22.

Aloy, P., and Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. Nat Biotechnol *22*, 1317-1321.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Bader, J. S., and Chant, J. (2006). Systems biology. When proteomes collide. Science *311*, 187-188.

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.

Bock, J. R., and Gough, D. A. (2001). Predicting protein--protein interactions from primary structure. Bioinformatics *17*, 455-460.

Bordner, A. J., and Abagyan, R. (2005). Statistical analysis and prediction of protein-protein interfaces. Proteins *60*, 353-366.

Bradford, J. R., and Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics *21*, 1487-1494.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., and Huang, E. S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci *13*, 190-202.

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

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

Chen, X. W., and Jeong, J. C. (2009). Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics *25*, 585-591.

Chung, J. L., Wang, W., and Bourne, P. E. (2006). Exploiting sequence and structure homologs to identify protein-protein binding sites. Proteins *62*, 630-640.

de Vries, S. J., and Bonvin, A. M. (2011). CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PLoS One *6*, e17695.

de Vries, S. J., and Bonvin, A. M. J. J. (2008). How proteins get in touch: Interface prediction in the study of biomolecular complexes. Curr Protein Pept Sc *9*, 394-406.

de Vries, S. J., van Dijk, A. D., and Bonvin, A. M. (2006). WHISCY: what information does surface conservation yield? Application to data-driven docking. Proteins *63*, 479-489.

Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc *125*, 1731-1737.

Fariselli, P., Pazos, F., Valencia, A., and Casadio, R. (2002). Prediction of protein--protein interaction sites in heterocomplexes with neural networks. Eur J Biochem *269*, 1356-1361.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2008). The Pfam protein families database. Nucleic Acids Res *36*, D281-288.

Fiorucci, S., and Zacharias, M. (2010). Prediction of protein-protein interaction sites using electrostatic desolvation profiles. Biophys J *98*, 1921-1930.

Fletcher, S., and Hamilton, A. D. (2007). Protein-protein interaction inhibitors: small molecules from screening techniques. Curr Top Med Chem *7*, 922-927.

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

Gallet, X., Charloteaux, B., Thomas, A., and Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. J Mol Biol *302*, 917-926.

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics *19*, 163-164.

Guharoy, M., and Chakrabarti, P. (2010). Conserved residue clusters at protein-protein interfaces and their use in binding site identification. BMC Bioinformatics *11*, 286.

Hamer, R., Luo, Q., Armitage, J. P., Reinert, G., and Deane, C. M. (2010). i-Patch: interprotein contact prediction using local network information. Proteins *78*, 2781-2797.

Han, J. D., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol *23*, 839-844.

He, X. L., Bazan, J. F., McDermott, G., Park, J. B., Wang, K., Tessier-Lavigne, M., He, Z., and Garcia, K. C. (2003). Structure of the Nogo receptor ectodomain: a recognition module implicated in myelin inhibition. Neuron *38*, 177-185.

Henrick, K., and Thornton, J. M. (1998). PQS: a protein quaternary structure file server. Trends Biochem Sci *23*, 358-361.

Holm, L., Kaariainen, S., Rosenstrom, P., and Schenkel, A. (2008). Searching protein structure databases with DaliLite v.3. Bioinformatics *24*, 2780-2781.

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

Hwang, H., Pierce, B., Mintseris, J., Janin, J., and Weng, Z. (2008). Protein-protein docking benchmark version 3.0. Proteins *73*, 705-709.

Janin, J. (1999). Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. Structure *7*, R277-279.

Jones, S., and Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol *63*, 31-65.

Jones, S., and Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. J Mol Biol *272*, 121-132.

Kim, W. K., Henschel, A., Winter, C., and Schroeder, M. (2006). The many faces of protein-protein interactions: A compendium of interface geometry. PLoS Comput Biol *2*, e124.

Kniazeff, J., Galvez, T., Labesse, G., and Pin, J. P. (2002). No ligand binding in the GB2 subunit of the GABA(B) receptor is required for activation and allosteric interaction between the subunits. J Neurosci *22*, 7352-7361.

Kohlbache, O., Burchardt, A., Moll, A., Hildebrandt, A., Bayer, P., and Lenhof, H. P. (2001). Structure prediction of protein complexes by an NMR-based protein docking algorithm. J Biomol NMR *20*, 15-21.

Koike, A., and Takagi, T. (2004). Prediction of protein-protein interaction sites using support vector machines. Protein Eng Des Sel *17*, 165-173.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J Mol Biol *372*, 774-797.

Kufareva, I., Budagyan, L., Raush, E., Totrov, M., and Abagyan, R. (2007). PIER: protein interface recognition for structural proteomics. Proteins *67*, 400-417.

Lacapere, J. J., Pebay-Peyroula, E., Neumann, J. M., and Etchebest, C. (2007). Determining membrane protein structures: still a challenge! Trends Biochem Sci *32*, 259-270.

Lamark, T., Perander, M., Outzen, H., Kristiansen, K., Overvatn, A., Michaelsen, E., Bjorkoy, G., and Johansen, T. (2003). Interaction codes within the family of mammalian Phox and Bem1p domain-containing proteins. J Biol Chem *278*, 34568-34581.

Li, N., Sun, Z., and Jiang, F. (2008). Prediction of protein-protein binding site by using core interface residue and support vector machine. BMC Bioinformatics *9*, 553.

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.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol *257*, 342-358.

Lietha, D., Cai, X., Ceccarelli, D. F., Li, Y., Schaller, M. D., and Eck, M. J. (2007). Structural basis for the autoinhibition of focal adhesion kinase. Cell *129*, 1177-1187.

Liu, R., and Zhou, Y. (2009). Using support vector machine combined with post-processing procedure to improve prediction of interface residues in transient complexes. Protein J *28*, 369-374.

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci U S A *100*, 5772-5777.

Moscat, J., Diaz-Meco, M. T., Albert, A., and Campuzano, S. (2006). Cell signaling and function organized by PB1 domain interactions. Mol Cell *23*, 631-640.

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.

Nooren, I. M., and Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol *325*, 991-1018.

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

Ofran, Y., and Rost, B. (2003b). Predicted protein-protein interaction sites from local sequence information. FEBS Lett *544*, 236-239.

Ofran, Y., and Rost, B. (2007). ISIS: interaction sites identified from sequence. Bioinformatics *23*, e13-16.

Oldfield, T. J. (2002). Data mining the protein data bank: residue interactions. Proteins *49*, 510-528.

Park, J., Lappe, M., and Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol *307*, 929-938.

Porollo, A., and Meller, J. (2007). Prediction-based fingerprints of protein-protein interactions. Proteins *66*, 630-645.

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

Ritchie, D. W. (2008). Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci *9*, 1-15.

Shenoy, S. K., Drake, M. T., Nelson, C. D., Houtz, D. A., Xiao, K., Madabushi, S., Reiter, E., Premont, R. T., Lichtarge, O., and Lefkowitz, R. J. (2006). beta-arrestin-dependent, G protein-independent ERK1/2 activation by the beta2 adrenergic receptor. J Biol Chem *281*, 1261-1273.

Slabinski, L., Jaroszewski, L., Rodrigues, A. P., Rychlewski, L., Wilson, I. A., Lesley, S. A., and Godzik, A. (2007). The challenge of protein structure determination--lessons from structural genomics. Protein Sci *16*, 2472-2482.

Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W., and Shi, L. (2011). Next-generation sequencing and its applications in molecular diagnostics. Expert Rev Mol Diagn *11*, 333-343.

Thorn, K. S., and Bogan, A. A. (2001). ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics *17*, 284-285.

White, A. W., Westwell, A. D., and Brahemi, G. (2008). Protein-protein interactions as targets for small-molecule therapeutics in cancer. Expert Rev Mol Med *10*, e8.

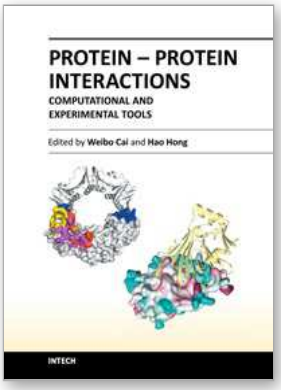
Xia, J. F., Zhao, X. M., Song, J., and Huang, D. S. (2010). APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics *11*, 174.

Xu, Q., and Dunbrack, R. L., Jr. (2011). The protein common interface database (ProtCID)--a comprehensive database of interactions of homologous proteins in multiple crystal forms. Nucleic Acids Res *39*, D761-770.

Yan, C., Honavar, V., and Dobbs, D. (2004). Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach. Neural Comput Appl *13*, 123-129.

Zhou, H. X., and Qin, S. (2007). Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics *23*, 2203-2209.

Zhou, H. X., and Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins *44*, 336-343.



Protein-Protein Interactions - Computational and Experimental Tools

Edited by Dr. Weibo Cai

ISBN 978-953-51-0397-4
Hard cover, 472 pages
Publisher InTech
Published online 30, March, 2012
Published in print edition March, 2012

Proteins are indispensable players in virtually all biological events. The functions of proteins are coordinated through intricate regulatory networks of transient protein-protein interactions (PPIs). To predict and/or study PPIs, a wide variety of techniques have been developed over the last several decades. Many in vitro and in vivo assays have been implemented to explore the mechanism of these ubiquitous interactions. However, despite significant advances in these experimental approaches, many limitations exist such as false-positives/false-negatives, difficulty in obtaining crystal structures of proteins, challenges in the detection of transient PPI, among others. To overcome these limitations, many computational approaches have been developed which are becoming increasingly widely used to facilitate the investigation of PPIs. This book has gathered an ensemble of experts in the field, in 22 chapters, which have been broadly categorized into Computational Approaches, Experimental Approaches, and Others.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jarek Meller and Alexey Porollo (2012). Computational Methods for Prediction of Protein-Protein Interaction Sites, Protein-Protein Interactions - Computational and Experimental Tools, Dr. Weibo Cai (Ed.), ISBN: 978-953-51-0397-4, InTech, Available from: http://www.intechopen.com/books/protein-protein-interactions-computational-and-experimental-tools/computational-methods-for-prediction-of-protein-protein-interaction-sites

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821