CHEMICAL REVIEWS

Review

pubs.acs.org/CR

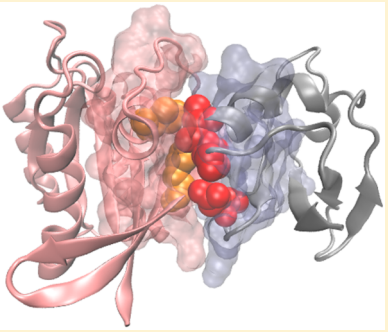
Predicting Protein−Protein Interactions from the Molecular to the Proteome Level

Ozlem Keskin,[*,†,‡] Nurcan Tuncbag,[*,§] and Attila Gursoy[*,‡,∥]

[†]Chemical and Biological Engineering, College of Engineering, [‡]Center for Computational Biology and Bioinformatics, and [∥]Computer Engineering, College of Engineering, Koc University, 34450 Istanbul, Turkey

[§]Graduate School of Informatics, Department of Health Informatics, Middle East Technical University, 06800 Ankara, Turkey

ABSTRACT: Identification of protein−protein interactions (PPIs) is at the center of molecular biology considering the unquestionable role of proteins in cells. Combinatorial interactions result in a repertoire of multiple functions; hence, knowledge of PPI and binding regions naturally serve to functional proteomics and drug discovery. Given experimental limitations to find all interactions in a proteome, computational prediction/modeling of protein interactions is a prerequisite to proceed on the way to complete interactions at the proteome level. This review aims to provide a background on PPIs and their types. Computational methods for PPI predictions can use a variety of biological data including sequence-, evolution-, expression-, and structure-based data. Physical and statistical modeling are commonly used to integrate these data and infer PPI predictions. We review and list the state-of-the-art methods, servers, databases, and tools for protein−protein interaction prediction.



CONTENTS

1. INTRODUCTION

Completion of the genome sequencing for more than 200 organisms in addition to human genome uncovered that the phenotypical complexity cannot be explained by the number of genes of the organism. This finding revolutionized the systems biology era, and the postgenomic events took extra attention toward explaining the phenotypical complexity. One of the mechanisms amplifying the complexity is alternative splicing.[1] More than 90% of all human genes are estimated to generate alternatively spliced mRNA isoforms.[1] Despite ∼20 000 protein coding genes in the human genome, 196 345 different transcripts have been released from these genes in Ensembl database (GRCh38, version 77)[2] that contribute to the diversity of the human proteome. Very recently, two proteome map drafts have been released as a complement to the available genome and transcriptome data and confirmed that the protein translation of more than 90% of the human genes exists.[3,4] Further, post-translational modifications (such as phosphor-

Received: November 24, 2015
Published: April 13, 2016

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

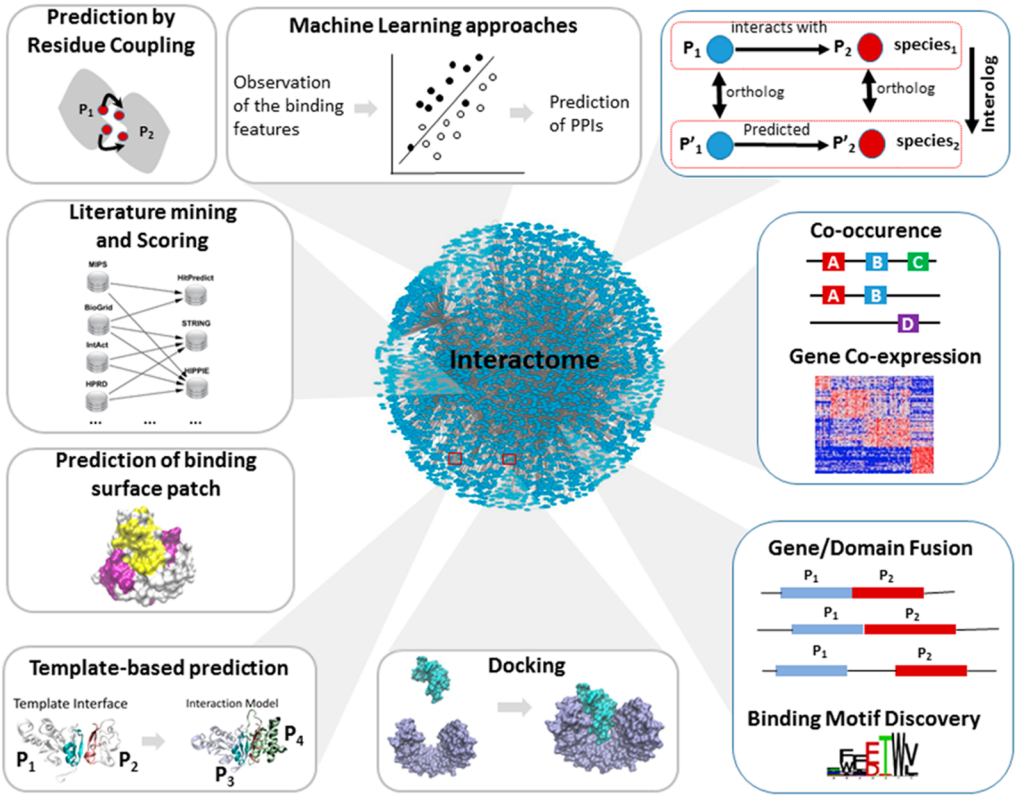


Figure 1. (A) Conceptual view on PPI prediction methods. PPI prediction approaches can be divided into three groups at the top level. (i) Pairwise PPI prediction methods to learn which proteins interact. Learning-based approaches, literature mining and scoring, interolog search, gene or domain fusion, gene coexpression, and co-occurrence belong to this class of PPI prediction. (ii) Binding site prediction methods to learn which region on a protein surface is used for binding. Binding patch and motif search belongs to this class. (iii) Protein assembly prediction to find out how proteins interact and form a complex. Docking and template-based prediction methods belong to this class. Eventually, each of these methods contributes to constructing a more complete interactome.

ylation, acetylation), tissue specificity, and cellular localization majorly contribute to the complexity.

Another contributor is the communication between proteins. Proteins do not act in isolation, and more than 80% of all proteins in the cell interact with other molecules to get functional.[5] A broad range of cellular processes including signal transduction, cell-to-cell communication, transcription, replication, and membrane transport are achieved by protein interactions. Protein interactions tell us how proteins come together to construct metabolic and signaling pathways in order to fulfill their functions. Dysfunction or malfunction of pathways and alterations in protein interactions have shown to be the cause of diseases, such as neurodegenerative diseases or cancer. Although solving the proteome is much harder than the genome, having a complete map of protein interactions is even more difficult because of the temporal and spatial heterogeneity in protein interactions. Some protein interactions are transient where protein partners associate and dissociate temporally.[6] In addition, proteins may need to be chemically modified such as phosphorylated to interact with their partners. Another constraint is the localization and transportability of the proteins. Proteins that are expressed in completely different locations and not transported to other locations may never associate although their interaction is possible physicochemically. Also, the expression levels of the proteins vary across different tissues as the proteome maps illustrated; therefore, protein interactomes are not the same in all cell types. Further, proteins are prone to changes in their three-dimensional structure, which directly changes their binding preferences.

The complete map of protein interactions that can occur in an organism is called the interactome. As of 2006, available PPI data was estimated to represent only 10% of all PPIs in human,[7] which is believed to be a small portion of all PPIs and may not be a good representative set of the complete interactome.[8] Still the known part of the interactome is valuable for explaining biological processes such as the molecular level links between diseases and proteins.[9] While the exact number of human PPIs is unknown, estimates range from hundred thousands to around a million.[10,11] Further, protein−protein interactions are dynamic. They might change depending on the condition and state of the cell as explained above. This heterogeneity leads to difficulty in PPI detection techniques and thus in definition of an interactome. Pairwise protein interactions can be detected with high-throughput or low-throughput experimental techniques. Recently, Rolland et al.[12] tested potential interactions between proteins of ∼13 000 genes. They reported 13 944 high-quality human binary protein−protein interactions between 4303 proteins. Actually, the number of possible binary interactions among the products of ∼200 000 transcripts might be much larger.

It is still an open question whether a complete interactome will ever be found by experimental techniques. Predictive methods became increasingly popular in the systems biology era to reveal the interaction principles at multiple scales, to detect new interactions, and to construct structural assemblies

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

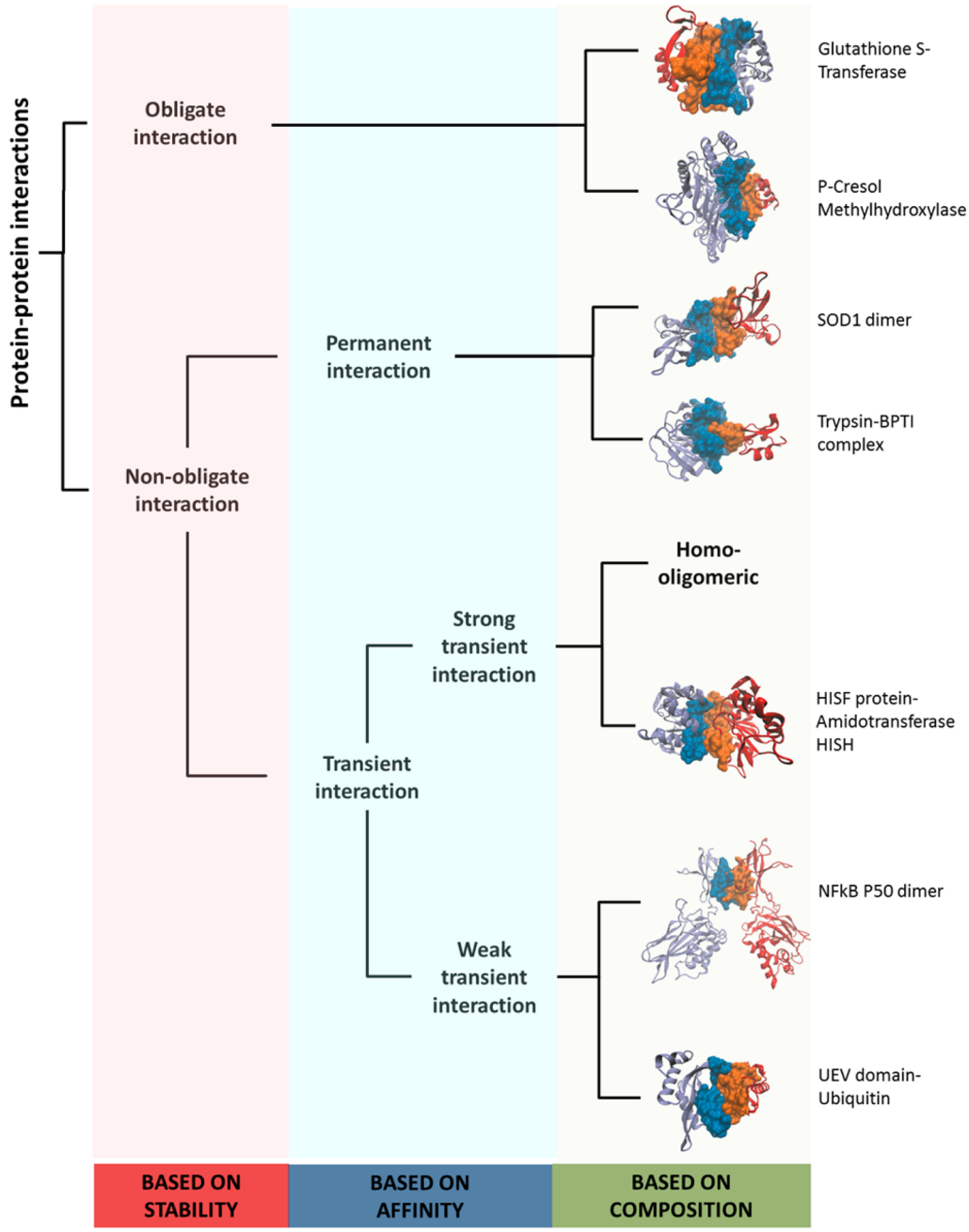


Figure 2. Classification of PPI types. On the basis of the stability of complex, interaction can be obligate or nonobligate. On the basis of the lifetime of the complex, interaction can be permanent or transient. Affinity of the interaction implies if the interaction is weak or strong. On the basis of composition, a complex can be a homo- (upper panel dimers) or heterodimer (lower panel cases) in the last column.

and networks of proteins. However, one should keep in mind that computational methods will only be better and cover more interactions with the help of reliable experimental data. We believe that a collective effort between the experiments and computations can make it possible to have a near complete set of interactions.

This review provides available methods to predict interaction of protein pairs, binding regions, and structural assemblies of proteins. Given the incomplete knowledge of PPIs and the interactome, in silico approaches emerge to fill out the gap and assist in reconstructing pathway maps. As illustrated in Figure 1, there are several types of methods aiming to solve the prediction of protein interaction problem. These methods can be classified into four groups based on the question they are addressing: (i) Which proteins interact with which others? (ii) What are the types of interactions and their importance? (iii) At which region of a protein does the binding occur? (iv) How strong is the interaction between two proteins which implies the importance of binding energies?

2. EXPERIMENTAL DETECTION OF PROTEIN INTERACTIONS

Pairwise protein interactions can be detected with high-throughput or low-throughput experimental techniques. Most detection methods are based on genetic or biochemical techniques; for a comprehensive review of the methods and their applications, please see refs 13 and 14. The most popular experimental methods for protein interactions are the yeast-

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

two-hybrid (Y2H) system, affinity purification followed by mass spectrometry (AP-MS), and literature-derived low-throughput experiments. The Y2H system is a genetic, high-throughput technique to detect direct interactions between proteins in vivo. The main challenge in the Y2H system is the failure of considering the dynamic aspects of protein interactions. For example, mammalian proteins expressed in completely different cellular compartments and at different time points can be detected as interacting by the Y2H system which produces false positives. Also, detection of interactions occurring after post-translational modifications is not possible with the Y2H system, which leads to false negatives. This leads to decreased overlap between different protein interaction data sets. For example, the overlap between yeast interactomes obtained by Ito el al. and Uetz et al. is only 20%, although they use the same 6000 open reading frames (ORFs) in their experimental setup.[15,16] Although the false positives included in the Y2H system are estimated to be 25−45% of the identified interactions,[17] the quality of Y2H data improved over time.[18] The other drawback of the Y2H system is that it can only identify binary interactions. The other technique AP-MS is a proteomic high-throughput approach and performs well in characterization of stable protein complexes or molecular machines under "native" conditions. Compared to the Y2H system, transient interactions are less represented in the AP-MS approach.[19] An additional drawback is that indirect (non-physical) interactions can be included with AP-MS, because co-purification does not always imply a physical interaction, and to identify the type of the interaction, detailed information about the protein complex is necessary.[20] However, with recently developed cross-linking and quantitative proteomic technologies, direct and dynamic interactions can also be elucidated by AP-MS.[20−23] A list of major AP-MS studies is provided in ref 24. Several computational methods are developed to accurately assess the reliability of interactions obtained by AP-MS experiments. Usually these methods are constructed based on a single AP-MS experiment. In a recent work, the performance of these scoring methods has been assessed across multiple interaction data sets.[25] The results show that the overlap between the high-confidence interactions found by each scoring method is very low, and they are still biased toward highly abundant proteins. In addition, they are not successful in discriminating the specific interactions.

Another high-throughput method is luminescence-based mammalian interactome mapping (LUMIER).[26,27] This strategy fuses the luciferase enzyme to a bait protein expressed in a mammalian cell along with candidate protein partners tagged with a polypeptide called Flag. Flag-tagged preys are detected when light is emitted. Other techniques to identify PPIs include co-immunoprecipitation, protein microarrays, fluorescence spectroscopy, resonance-energy transfer systems, mammalian-two-hybrid, mammalian protein−protein interaction trap (MAPPIT), phage display, surface plasmon resonance, protein-fragment complementation assay, and isothermal titration calorimetry (ITC) (extensively reviewed in refs 13 and 14).

3. PPI TYPES AND CHARACTERISTICS

Protein−protein interaction types are diverse ranging from transient or permanent nonobligate interactions to obligate interactions.[6,28−30] Different types of complexes with specific functions can be observed. Large macromolecular complexes, such as the small and large subunits of ribosomes or rings of GroELs, are highly stable and permanent.[31,32] Dynamic and transient interactions are key components in signaling and regulatory networks such as the interactions of Ras protein with its effectors (i.e., Raf, PI3K, etc.) where Ras acts as a switch in signaling.[33] Therefore, it is of great interest to classify PPI types. However, usually a continuum exists and it is not straightforward to separate PPIs into one of the classes.[6]

3.1. Homo-Oligomeric and Hetero-Oligomeric Complexes

This type of classification is straightforward. If the proteins in a complex are identical (interactions occurring between identical protein chains), they form a homo-oligomer, whereas if the PPI takes place among nonidentical chains then it forms a hetero-oligomer. Homo-oligomers are mostly symmetric and provide a good scaffold for stable macromolecules. The stability of hetero-oligomers, on the other hand, varies. Most of the homodimers are only observed in the oligomeric form (i.e., section 3.2), and it is often impossible to separate them into independently stable folded monomers. As an example, F-type ATP synthases are large multisubunit complexes. They convert the energy stored in electrochemical gradients of H^+ for the synthesis of ATP. The C-subunit assembly (the C-ring) is the key element that transduces the electrochemical energy into mechanical rotation and vice versa. The cyclic structure of ATP sythase C-ring is an example of the homooligomeric, highly symmetric, and stable protein complexes (i.e., PDB ID: 2xqt).[34]

3.2. Obligate and Nonobligate Complexes

In order to classify interactions as obligate/nonobligate, one needs to know the affinity and stability of the proteins in the complex and monomeric states, see section 6 (Figure 2). If the proteins (monomers) of a complex are unstable on their own in vivo then this is an obligate interaction, whereas the components of nonobligate interactions can exist independently.[35,36] Obligate interactions are named as two-state folders. Protein components fold and bind at the same time to form stable complexes. The individual proteins cannot exist as stable, folded structures, but they are stable in the complex form. The components of the nonobligate interactions are three-state folders; they first fold and then come together to form the complex. Most of the stable machineries in the cell are examples of obligate complexes.

3.3. Transient and Permanent Complexes

Protein interactions can be classified based on the lifetime of the complex. This classification is relevant only to non-obligatory interactions. Permanent interactions are usually very stable; once two proteins interact they permanently stay as a complex.[35] Transient interactions associate and dissociate temporarily in vivo (Figure 2). Binding between hormone−receptors, signal transduction, inhibition of proteases, and chaperone-assisted protein folding are examples of transient interactions. These types of interactions dominate signaling and regulatory pathways as they provide a mechanism for the cell to quickly respond to extracellular stimuli and relay the signals when needed.

It is hard to differentiate between transient/permanent and obligate/nonobligate complexes. Usually these classifications are intermixing such that obligate interactions are permanent whereas nonobligate interactions can be either transient or permanent. Antibody−antigen interactions are examples of permanent, nonobligate interactions. Large stable supramolecular systems should be strong, and they are examples of obligate and permanent interactions. Usually binding free

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

energy is used to determine the stability and affinity of complexes. The interacting proteins of permanent complexes are more likely to be coexpressed and colocalized than proteins in transient complexes.

3.4. Disordered-to-Ordered Complexes

A not so well-defined type of PPIs is the one formed by disordered proteins. Intrinsically disordered proteins have regions that are unstructured whose amino acid compositions cannot provide a stable folded structure.[37] Disordered proteins are especially abundant in eukaryotic proteins including the tails of histone proteins and proteins that control the cell division cycle and signaling. These proteins can bind to several different proteins by adapting a conformation compatible with partner proteins. Post-translational modifications on these regions also mediate binding to different partners. A large portion or a small region of the protein might be disordered (intrinsically disordered).[38] Larger disordered segments can fold simultaneously when they bind to their biological targets (coupled folding and binding), whereas shorter flexible disordered linkers might have a role in the assembly of macromolecular complexes.[39,40]

3.5. Biological and Crystal Complexes

Structures found by X-ray crystallography in the Protein Databank (PDB)[41] can contain nonbiological contacts which can be considered as experimental artifacts.[42−45] Although these interactions might sometimes be similar to biological ones,[46] for accurate analysis of protein binding preferences and properties, crystal contacts need to be identified. Toward this aim, there are several efforts to distinguish biological complexes from crystal ones. The most discriminative feature between crystal and biological interfaces is the interface area size. Although interface area size is a good determinant of crystal interfaces, there are also counter examples having a very large interface area but the interface is completely composed of crystal contacts. Another approach to distinguish biological interfaces from nonbiological ones is checking the conservation rate. Combination of interface size and conservation distinguishes biological interfaces with an accuracy of 98.3%.[47] Amino acid compositions of interfaces are another feature to distinguish biological interfaces. If the amino acid composition of the interface is similar to the rest of the protein surface then these interfaces are labeled as nonbiological. Some other interface properties to distinguish biological and crystal interactions are hydrogen bonds and salt bridges across the interface, free energy, and hydrophobicity.

4. PHYSICOCHEMICAL PROPERTIES OF PPI BINDING SITES

Proteins interact through their interfaces. Structural aspects, physicochemical properties, affinity, and specificity of binding are diverse across different protein−protein interfaces.[6,48] In this section, we review characteristics of protein interfaces and available databases and tools about protein interface properties. For the analysis of binding preferences of proteins, interface regions need to be extracted. There are several approaches to find interface regions from 3-dimensional coordinates of a protein complex, such as calculating the accessible surface area (ASA) of the residues or calculating the atomic distances. If the difference between the ASA of a residue in monomeric state and complex state is greater than a threshold (usually 1 Å^2) that residue is labeled to be an interface residue.[6] If the distance between any atoms of two residues each from one chain of a protein complex is less than a threshold (usually 4.5 Å) those residues are labeled as contacting.[49] A list of some available protein interface databases and tools to find interfaces is provided in Table 1.

Table 1. Databases of Known Protein Interfaces

name	web link	interface type	interface clustering
ProtCID[147]	http://dunbrack2.fccc.edu/protcid/	protein−protein	yes
PISITE[148]	http://pisite.hgc.jp/	protein−protein	no
PISA[149]	http://www.ebi.ac.uk/pdbe/pisa/	protein−protein	no
PSIBASE[150]	http://psibase.kobic.re.kr/	protein−protein and domain−domain	no
2P2I Inspector[151]	http://2p2idb.cnrs-mrs.fr/2p2i_inspector.html	protein−protein	no
PiFace[152]	http://prism.ccbb.ku.edu.tr/piface/	protein−protein	yes
PDBSum[153,154]	http://www.ebi.ac.uk/pdbsum/	protein−protein	no
3DID[155]	http://3did.irbbarcelona.org/	domain−domain	yes
iPFAM[156]	http://www.ipfam.org/	domain−domain	no
IBIS[157]	http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi	protein−protein	no
SCOPPI[158]	http://www.scoppi.org/	domain−domain	yes
SCOWLP[159]	http://www.scowlp.org/scowlp/	domain−domain	yes
SPIN-PP	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:SPIN-PP	protein−protein	no
Dockground	http://dockground.compbio.ku.edu	protein−protein	yes

Interface databases are rich resources for characterization of binding surfaces. These databases can contain two types of interfaces: protein−protein interfaces or domain−domain interfaces. Domains are functional modular substructures of proteins. A domain−domain interface can be extracted from intrachain or interchain contacts. Some of the databases listed in Table 1 cluster protein interfaces based on a similarity measure which may use structure, sequence, or topology-based similarities.

Physicochemical properties of protein−protein interfaces include structural and chemical properties. These should be examined to understand the nature of the intermolecular interactions. For example, the surface area that is buried by the interacting molecules and the nonpolar fraction, the hydrogen bonds and the salt bridges across the interface, buried water molecules, the charge distribution and the composition of the interface, residue conservation, the strength of the interaction, flexibility of the interface residues and residues that contribute significantly to the free energy of binding (hot spots), the shape of the binding interface, complementarity of two binding sites, and the types of secondary structures are some of the properties of binding sites.[35,50]

One of the most striking features in protein binding is the energy distribution in the interface region. Hot spots in protein interfaces are energetically critical and contribute more to the binding.[51] These residues can be found experimentally by

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

alanine scanning mutagenesis.[52] If there is a change in binding affinity, usually a variation in binding energy greater than 2 kcal/mol, when a residue is mutated to alanine then this residue is labeled a hot spot. As an example to show an interface and hot spot localizations, Ras/Raf1 complex is illustrated in Figure 3, highlighting predicted hot spots[53] in the interface region.

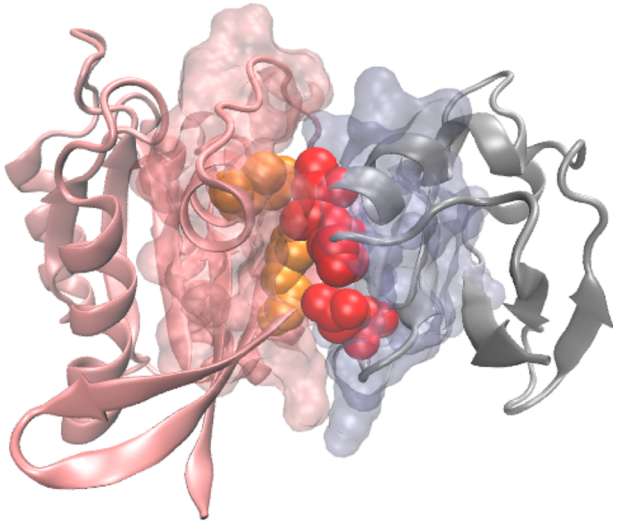


Figure 3. Visualization of the HRAS/RAF complex (4g0n). Overall structure is shown with a cartoon representation where each chain is colored differently. Interface region between HRAS and RAF is in transparent surface representation, the predicted hot spots for this interaction are in opaque van der Waals representation, and hot spots in different chains are colored in red and orange.

Although the alanine scanning experiment is invaluable in hot spot identification, the available data deposited in several databases is limited. Given these limitations, several predictive methods have been developed which successfully distinguish hot spots from nonhot spots in protein interfaces. Computational alanine scanning is one of them.[54] Other methods include learning-based[55−58] and molecular dynamics-based[59,60] approaches. The most discriminative feature in hot spot prediction is the solvent accessibility. Usually hot spots are buried and excluded from solvent and found in close proximity to each other.[61,62] Some computational methods have been listed in Table 2. Hot spots are potential drug targets, and drug molecules have a tendency to bind hot regions in protein interfaces.[63] As an example, IL-2/IL-2RA protein complex and the small molecule FRH ligand targeting IL-2 are provided in Figure 4. FRH binds to the region where IL-2 hot spots in IL-2RA interaction are located.[64]

As to the chemical properties of protein interfaces, aromatic side chains have preference to be in the binding site.[65] Also, the stability and specificity of protein interactions are highly dependent on the presence of hydrogen bonds, electrostatic interactions, salt bridges, and hydrophobic attractions. Although the frequency of seeing disulfide bonds is very low, they contribute to the rigidity and stability of relatively small protein complexes. Protein interfaces can be divided into core and rim regions where the rim region is more exposed to the solvent. Core regions are shown to be more similar to the interior part of the proteins, and rim regions are more similar to the protein surface in terms of residue frequency.[66] Besides, protein binding regions are less flexible than the remaining surface region.[30]

There are differences between interfaces of different types of interactions. For example, permanent complexes are more hydrophobic compared to transient interfaces.[6] While interfaces of the obligate ones are more conserved in sequence than the transient ones,[30] the shape complementarity is less important in transient interactions. Hydrophobic interactions are more preferred in obligate complexes, while salt bridges and hydrogen bonds are more preferred in transient complexes. In globular complexes and receptor−ligand complexes, interfaces are larger than transient and oncogenic interactions.[67] Figure 5 displays two protein−protein interfaces where stefin B/papain complex represents a nonobligate interaction (Figure 5A) while methylmalonyl-CoA mutase complex represents an obligate interaction (Figure 5B). Stefin B/papain has a relatively smaller interface area compared to the interface in methylmalonyl-CoA mutase complex. The gap volume index (GV index) between interacting pairs gives some insight about the interface complementarity which is the gap volume between two protein interfaces normalized with the interface area size. A small GV index corresponds to better complementarity. For example, the interface complementarity of methylmalonyl-CoA mutase complex (GV index = 1.65 Å) is higher than stefin B/papain complex (GV index = 2.12 Å). The nonobligate stefin B/papain interface has 7 hydrogen bonds and 105 nonbonded atomic interactions. In the obligate methylmalonyl-CoA mutase interface, 30 hydrogen bonds, 10 salt bridges, and 649 nonbonded atomic interactions are formed.

Post-translational modifications have critical roles in protein binding. In hetero-oligomers and weak transient homo-oligomers, phosphosites are significantly located in the interface regions.[68] In hetero-oligomers, phosphosites are located preferably at binding site hot spots. Additionally, phosphosites in protein interfaces are more conserved when compared to the rest of the protein complex.[68] Methylation, acethylation, and ubiqination are also important in PPIs.

Although all of these properties provide a static view of binding, still they are valuable to understand the nature and strength of PPIs. Table 2 lists the servers that give physicochemical properties of interfaces, such as conservation, binding energies, hot spot residues, and many more.

5. MULTIPARTNER PROTEINS

Proteins can have multiple binding sites and therefore interact with their partners simultaneously. Proteins can use the same binding site repeatedly and bind to multiple partners at different times, forming mutually exclusive interactions; these interfaces show the characteristics of transient interfaces.[69,70] Therefore, adapting multiple binding sites or reutilizing a single site by several partners is crucial for interaction with many different proteins. This adaptation has been illustrated by experimentally determined protein complexes in the PDB, where it has been shown that structural information can change the classical network representation of protein interactions.[71] Hub proteins have been classified into multi-interface hubs and singlish-interface hubs where the former are more conserved and essential than the latter.[71] In order to reconstruct the structural p53 pathway, the multi-interface nature of p53 and Mdm2 proteins has been illustrated including both known and predicted interactions.[72] As another example, the complement cascade pathway in KEGG has been reconstucted with structural information to illustrate the order of events.[73] Structural networks of the human ubiquitination pathway, pathways in breast cancer, the Interleukin 1 initiated signaling

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

Table 2. Databases and Tools To Analyze Protein Interfaces and Calculate Physicochemical Properties of Interface Residues

name	web link	features
ASEdb[160]	http://nic.ucsf.edu/asedb/	collection of experimental alanine scanning data
BID[161]	http://tsailab.chem.pacific.edu/wikiBID/index.php/Main_Page	collection of experimental alanine scanning data
Kinetic Data of Biomolecular Interactions (KDBI)[162]	http://xin.cz3.nus.edu.sg/group/kdbi/kdbi.asp	a database of experimentally determined kinetic data of protein−protein, protein−RNA, protein−DNA, protein−ligand, RNA−ligand, DNA−ligand binding, or reaction events available in the literature
SKEMPI[163]	http://life.bsc.es/pid/mutation_database/	
Protein−Protein Interaction Affinity Database[164]	http://bmm.cancerresearchuk.org/~bmmadmin/Affinity/	a nonredundant set of 144 protein−protein complexes having high-resolution strutures and experimentally determined dissociation constants
PINT[165]	http://www.bioinfodatabase.com/pint/	
Hotpoint[53]	prism.ccbb.ku.edu.tr/hotpoint	empirical formulation for hot spot prediction
Hotsprint[166]	prism.ccbb.ku.edu.tr/hotsprint	
Hotregion[167]	prism.ccbb.ku.edu.tr/hotregion	prediction of hot regions in protein interfaces
KFC, KFC2[55]	http://kfc.mitchell-lab.org/	ML-based hot spot prediction
FoldX[168]	http://foldx.crg.es/	estimates interaction those contribute to the stability of proteins and protein complexes
Robetta[169]	http://www.robetta.org/submit.jsp	in silico alanine scanning
FTMAP[170]	http://ftmap.bu.edu/	identification of druggable hot spots of proteins using Fourier domain correlation techniques.
PCRPi-DB[171]	http://www.bioinsilico.org/PCRPIDB	prediction with Bayesian networks
PredHS[172]	http://www.predhs.org	identification of computational hot spots in protein interfaces using structural neighborhood properties
ANCHOR[173]	http://structure.pitt.edu/anchor/	analyzes protein−protein interfaces for their druggability
DrugScorePPI[174]	http://cpclab.uni-duesseldorf.de/dsppi/	in silico alanine scanning
Consurf[175]	http://bental.tau.ac.il/new_ConSurfDB/	residue conservation
SCORECONS[176]	https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons_server.pl	residue conservation
Galinter[177]	http://galinter.bioinf.mpi-inf.mpg.de/	interface alignment by considering interchain contacts as vectors.
PCAlign[178]	http://brooks.chem.lsa.umich.edu/index.php?page=pcalign&subdir=articles/resources/software	alignment of protein−protein interfaces by considering spatial and physicochemical similarities
IAlign[179]	http://cssb.biology.gatech.edu/iAlign	structural interface alignment
I2ISiteEngine	http://bioinfo3d.cs.tau.ac.il/I2I-SiteEngine/	structural and physicochemical alignment of interfaces
MAPPIS[180]	http://bioinfo3d.cs.tau.ac.il/MAPPIS/	interface alignment by considering structure and chemical conservation
PQS Database	pqs.ebi.ac.uk/	distinguishing biological versus crystal contacts features: interface area size
Conserved Domain Database (CDD)	http://www.ncbi.nlm.nih.gov/cdd/	distinguishing biological versus crystal contacts features: conserved binding modes using structural alignment
DiMoVo[181]	http://fifi.ibbmc.u-psud.fr/DiMoVo	machine-learning-based integrating multiple features, i.e., surface area, conservation, residue propensity
NOXClass[182]	http://noxclass.bioinf.mpi-inf.mpg.de/	machine-learning-based integrating multiple features, i.e., surface area, conservation, residue propensity
PISA[149]	http://www.ebi.ac.uk/pdbe/pisa/	free energy of formation, solvation energy gain, interface area, hydrogen bonds and salt bridges across the interface and hydrophobic specificity
DynaFace[183]	http://safir.prc.boun.edu.tr/dynaface	predicts obligatory and nonobligatory interactions using the Gaussing Network Model (GNM)

pathway, and ERKs in the MAPK pathway have also clearly illustrated the multipartner character of proteins in functional pathways.[74−77]

The partners cGMP 3′,5′-cyclic phosphodiesterase, regulator of G-protein signaling (RGS) 14, KB752 peptide, and KB-1753 phage display peptide bind to the same region on guanine nucleotide binding protein. Hence, their interactions are not simultaneously possible. Distinct from this region, there are 3 more binding regions on guanine nucleotide binding proteins where RGS4, RGS8, and G protein beta subunit bind, respectively. Although the binding regions of RGS4, G protein beta subunit, and cGMP 3′,5′-cyclic phosphodiesterase are not overlapping, the rest of the partner proteins interpenetrate each other. Hence, their simultaneous interactions are not possible. Interactions of G protein alpha subunit with itself, beta subunit, and RGS4 are shown in Figure 6.

The serine protease subtilisin BPN′ is another multipartner protein that uses the same region to interact with different proteins (see Figure 7). Streptomyces subtilisin inhibitor (3sic:I), eglin C (1sib:I), and chymotrypsin inhibitor 2 (1y34:I) are the binding partners of subtilisin. Although their overall structures are dissimilar, their interface regions are structurally very similar. Computational hot spots in the interface regions of subtilisin remain unchanged while interacting with different partners.

6. AFFINITY OF PPIS

The types of PPI depend on different conditions, most importantly pH, protein concentration, concentration of other components in the cell, temperature, etc. The term binding affinity is important to define the strength of the interaction between two proteins that bind to each other physically. Hence, the affinity of proteins determines whether the interaction will actually take place under a given condition. For the binary interaction A + B ↔ AB, the binding affinity is a function of the interaction force of attraction between the A and the B proteins and therefore describes the strength of the complex, while the underlying forward (k_{on}) and reverse (k_{off}) rates determine the

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

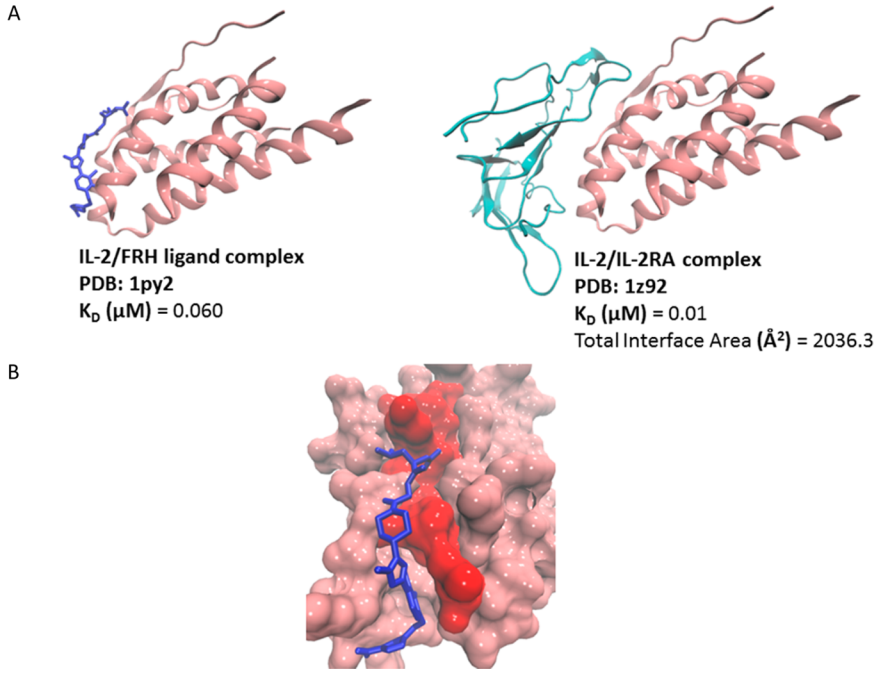


Figure 4. (a) Interaction between FRH ligand and IL-2 protein (PDB: 1py2) where K_D = 0.060 μM. FRH binds to the same region where IL-2RA protein binds to IL-2. Interaction between IL-2 and IL-2RA (PDB: 1z92). Dissociation constant (K_D) of this interaction is 0.01 μM, and total interface area is 2036.3 Å^2. (b) Binding pocket in IL-2 with FRH ligand has been zoomed where predicted hot spots on IL-2 in IL-2RA binding are colored red and FRH ligand is colored blue.

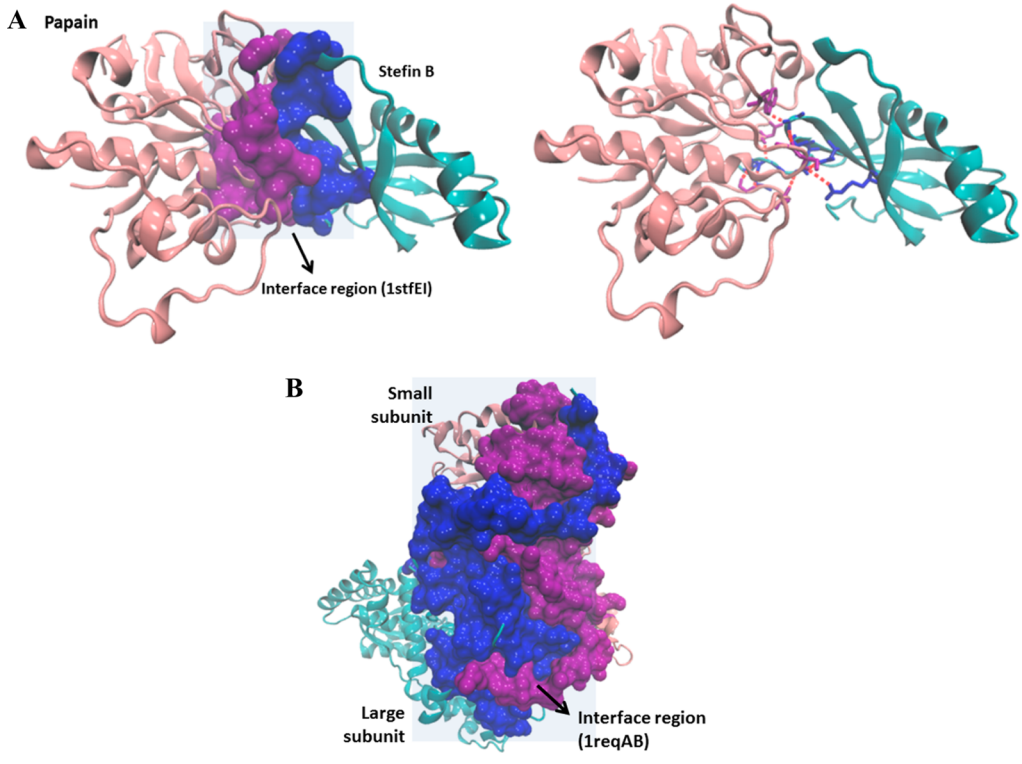


Figure 5. Visual representation of (a) the nonobligate complex formed between stefin B and papain and (b) the obligate complex formed between methylmalonyl-CoA mutase small and large subunits. Purple and dark blue balls represent the interface residues. For the stefin B/papain complex structure, two different representations have been provided. Structure of the complex and interface region is illustrated in the left part, and some atomic details have been provided in the right part. Hydrogen bonds within stefin B/papain interfaces have been drawn with red dashed lines.

time scales of association and dissociation, respectively. k_{on} and k_{off} can be used to find the dissociation equilibrium constant (K_d) with $K_d = C_A C_B / C_{AB} = k_{off}/k_{on}$, where C's represent the concentrations of the proteins. Fast association may enhance binding affinity with little A and B in the environment. High affinity can also be achieved through slow dissociation, i.e., once

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

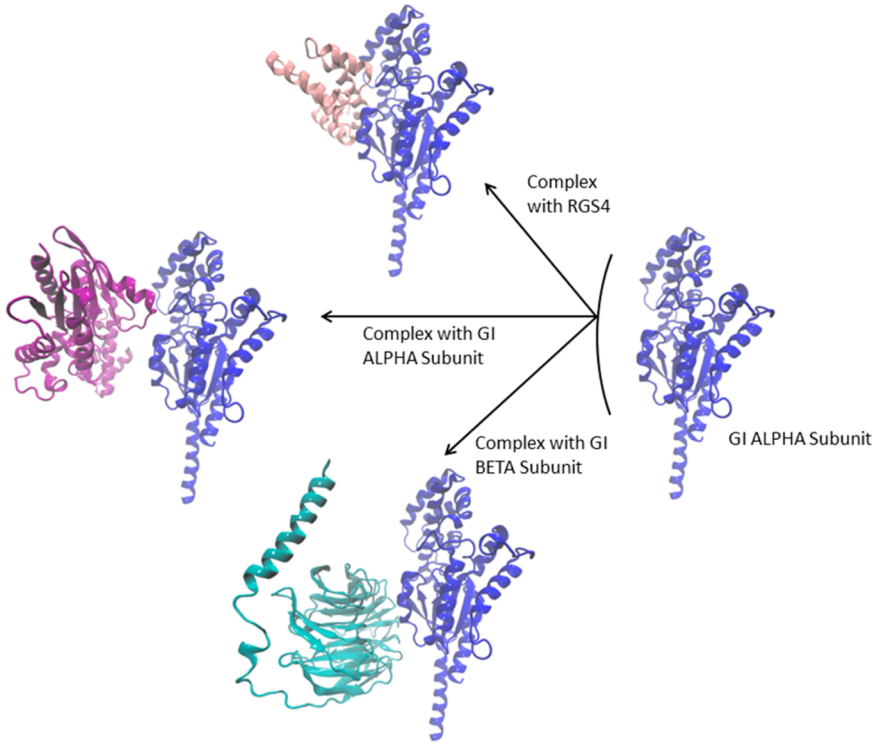


Figure 6. Partners of guanine nucleotide binding protein alpha subunit(GI). The first binding site on G protein is used by itself to form a homodimer and by Regulator of G-Protein Signaling (RGS) 4. The second region is used by G protein beta subunit.

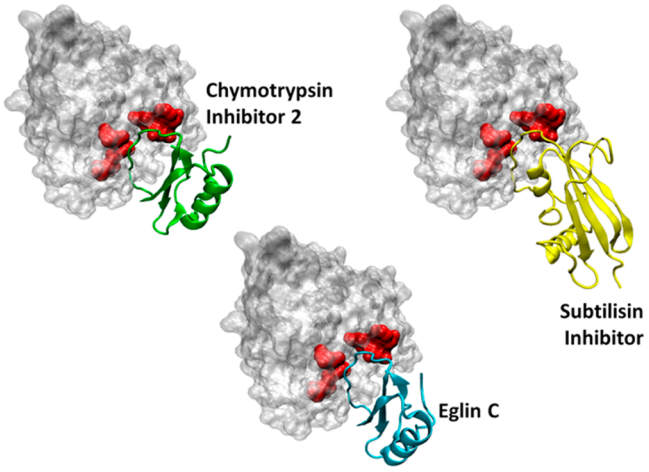


Figure 7. Subtilisin and its inhibitor molecules chymotrypsin inhibitor-2, subtilisin inhibitor, and eglin C. Subtilisin is shown in surface representation colored white. Predicted hot regions are colored red.

AB is formed, it does not dissociate back to A and B fast. However, since slow dissociation results in a long-lasting complex (AB), they are not observed in signaling where proteins need to respond fast and effectively to stimuli. The equilibrium constant, K_d, can be empirically used to find the binding free energy of the reaction with the relation $\Delta G = -RT \ln K_d$ and therefore can be used to score the binding processes.[78] Dissociation constant values (K_d) can be used to differentiate strong and weak transient interactions. Dissociation constants of strongly permanent complexes are typically in the nanomolar range, whereas transient complexes typically are in the micromolar range or higher. In Figure 8, three protein complexes are presented with their dissociation constants and binding energies as examples. The UEV ubiquitin complex is an example of weak transient interaction, while the complex formed between uracyl DNA glycosylase/glycosylase inhibitor complex and trypsin/BTPI complex are examples of permanent interactions.

There are structure-based methods to predict binding affinity and therefore score PPIs. These methods use empirical scoring functions, physics-based, knowledge-based methods, or quantitative structure−activity relationships. Sequence-based methods can also be used which mainly use amino acid properties of interacting proteins. This area is still open, and there are many ongoing studies to predict the binding affinity of proteins.[79−81]

7. COMPUTATIONAL METHODS FOR PREDICTION OF PPIS

Limitations of experimental methods necessitate computational prediction of protein interactions. Various computational approaches exist to predict PPIs. The majority of them can be grouped as simulation-based and statistical/machine-learning-based approaches. The simulation-based methods model the forces governing interactions of proteins, usually at atomistic level, and compute the strength of the interactions. These methods include molecular dynamics simulations and docking and are mostly used either studying the dynamics of interactions or determining strength of interactions rather than finding which proteins interact with which others due to high computational cost. Statistical and machine-learning-based methods, on the other hand, can be used at large scale. Protein interactions can be inferred by using information from known interacting proteins.[82] With the recent advances in biotechnology, for example, next-generation sequencing, a wealth of protein data are being produced. This trend further necessitates computational approaches to integrate, understand, and extract information from the large and diverse data such as sequence,

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

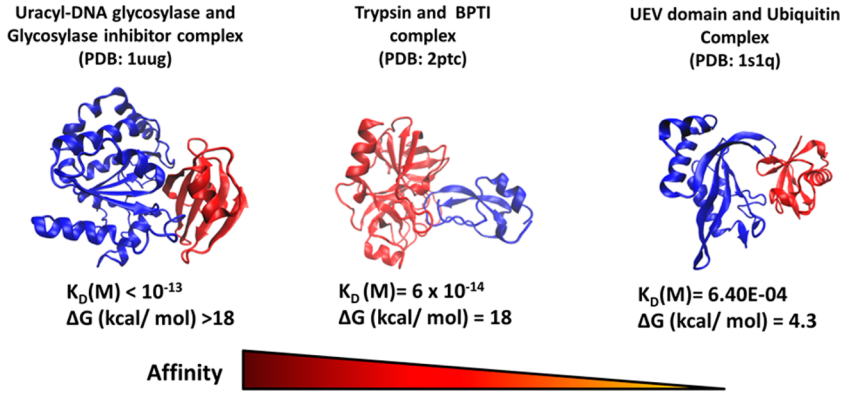


Figure 8. Examples of protein complexes with changing affinities retrieved from the PPI Affinity Database with their dissociation constants (K_D) and binding energies (ΔG). The strongest interaction among these examples is uracyl-DNA glycosylase and glycosylase inhibitor complex, and the weakest interaction is UEV domain and ubiquitin complex.

Table 3. Comparison of PPI Prediction Methods

method	input	better in predicting (transient/permanent)	enegy (scoring) available
interolog search	complete genome sequences of several organisms	usually for obligate but not transient	yes
gene/domain fusion	complete genome sequences of several organisms/ protein domain association data	not good for promiscuous domain interactions but for physical permanent interactions	yes
gene cluster and gene neighborhood	genome sequences (not necessarily complete) of several organisms	not physical interaction but rather functional interaction.	
residue coupling and coevolution	amino acid sequnces of interacting proteins	good for direct and physical interactions	yes
phylogenetic similarity	genome sequences (not necessarily complete) of several organisms	good for functional and transient interactions	no
network topology	binary protein−protein interaction data	for functional interactions for permanent and transient interactions	no
gene coexpression profiles	gene expression and ORF data	mostly for functional interactions	no
docking	protein structure	mostly for direct interactions	yes
machine learning and text mining	all sorts data including sequence, structure, gene expression, literature	can be used for both functional and direct interactions and for transient and permanent interactions	yes

structure, gene expression, binding affinity, and many other types of data to complement experimental detection methods.

Classification, or supervised learning, is a commonly used machine-learning (ML) technique employed for this purpose. Supervised machine learning builds a predictor using a training data, which is a set of labeled data for interacting and non-interacting proteins. The prediction model then can be applied to a new pair of proteins to infer interactions. A typical ML prediction study requires (a) deciding the features to represent proteins, such as sequence, physicochemical characteristic of amino acids, affinity, or any other available data, and (b) choosing a learning algorithm. Support vector machines (SVM), neural networks, decision trees, and random forests are examples of commonly used learning algorithms differing in certain strengths and weaknesses. Excellent reviews are available covering ML techniques and application to PPI in the literature.[82−84]

In general, a combination of the features representing PPIs and selecting an appropriate learning algorithm leads to a reliable PPI prediction method. Most of the learning-based computational methods require a list of positive examples (interacting pairs of proteins) and negative examples (pairs of noninteracting proteins) for training the classifiers. Construction of gold standard positive and negative sets is crucial to develop reliable ML-based prediction algorithms. For positive interaction set, there are several resources which contain verified PPIs (discussed in the PPI databases section). For negative interaction set, this problem is more complicated. A recent work is Negatome v2, which can be used as a data set for negative interactions.[85]

Besides the ML-based methods, computational prediction of pairwise PPIs and their analysis can be done using interolog mapping, gene/domain fusion events, learning-based prediction using sequence information, domain co-occurrence, and gene coexpression. A summary of these prediction methods with their aspects such as their performance in predicting transient or obligate interactions and availability of a scoring scheme is listed in Table 3. Below, we discuss some of the methods in detail.

7.1. Gene/Domain Fusion-Based Methods

Different genes can fuse into a single open reading frame and be translated to multidomain protein sequences. Gene (and consequently domain) fusion events are useful in detecting functional relation of proteins. In a pioneering study, Enright et al. stated that there must be selective pressure for certain genes to be fused over the course of evolution.[86] With this hypothesis, they predicted functional associations of proteins. They observed that 215 genes or proteins in the complete genomes of *E. coli*, *H. influenzae*, and *M. jannaschii* were involved in 64 unique fusion events.

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

Table 4. Pairwise PPI Prediction

name	web link	prediction method
hPRINT	www.print-db.org	ML-based pairwise physical interaction prediction
I2D	http://ophid.utoronto.ca/ophidv2.204/	orthologs
POINT	http://point.bioinformatics.tw/	orthologs
CODA	ftp://ftp.biochem.ucl.ac.uk/pub/gene3d_data/v6.1.0/CODA/	domain fusion
HPID	http://www.hpid.org	prediction by protein superfamily similarity and orthology
Bio::Homology::InterologWalk	http://search.cpan.org/~ggallone/Bio-Homology-InterologWalk/	orthologs
PRIME	N/A	text mining
EP-PPI	http://www.bioinf.ebc.ee/EP/EP/PPI/	gene coexpression
LocFuse	http://lbb.ut.ac.ir/Download/LBBsoft/LocFuse/	localization
PPInterFinder	http://www.biomining-bu.in/ppinterfinder/	text mining
PIANA	http://sbi.imim.es/piana	structual similarity + finding interologs
PPI-Search	http://gemdock.life.nctu.edu.tw/ppisearch/	interologs
BIPS[95]	http://sbi.imim.es/web/index.php/research/servers/bips	interologs
FpClass	http://www.cs.utoronto.ca/~juris/data/fpclass/	
EVComplex	http://evfold.org/	residue coupling

Gene fusion algorithms use nucleotide sequences, whereas domain fusion algorithms use domain-based annotations (e.g., CATH, SCOP, or PFAM) where a fusion event is between two proteins which contain distinct domains that are found fused together in another genome. Domain or gene fusion approaches may sometimes not be robust due to the complex structure of genomes. In this type of approach, promiscuous domains are challenging, because they are present in many proteins and fused to their partner domains, thus increasing the false positive rates in prediction of functional associations.

The advantage of this method is its high accuracy. One of the disadvantages is that gene-fusion events do not occur to a large extent, particularly in simple organisms. Thus, this method is not useful to detect interactions of all proteins. For example, Marsh et al. showed that only 3.7% of the nonredundant subunit pairs in their data set are associated with evolutionary fusion events.[87] Additionally, they mapped gene fusion events identified from fully sequenced genomes onto protein complex assembly orders and demonstrated evolutionary selection for conservation of assembly order. They further used structural and high-throughput interaction data and showed that gene fusion tends to optimize protein complex assembly by simplifying protein complex topologies.[87] Gene fusion events are more related to physical interaction between proteins than to other weaker functional relationships such as participation in a common biological pathway.[88]

7.2. Gene Cluster- and Gene Neighborhood-Based Methods

Gene clusters are defined as a set of genes within an intergenic distance of a threshold number of nucleotide bases (e.g., 1000). The size of gene clusters can vary significantly, from a few genes to several hundred genes. Genes with similar (or related) functions encoding potentially interacting proteins are often transcribed as a single unit, an operon, in bacteria and are assumed to be coregulated in eukaryotes. It is observed that in eukaryotes genes involved in the same biological process or pathway are frequently situated in close genomic proximity.[89] Gene cluster-based methods calculate co-occurrence probability of orthologs of query proteins encoded from the same gene clusters. This method is also named domain/gene co-occurrence. If two proteins' genes are not close by in the genome; then this method cannot reliably predict an interaction between these two genes.

The gene clusters do not point to a physical interaction between proteins but rather a functional interaction. A number of excellent resources exists that allow one to determine whether two proteins may interact using this approach. String, for example, can provide the co-occurrence data for the query protein.[90] Muley and Ranjan reported the analyses of 14 biological pathways of *E. coli* catalogued in the KEGG database using five protein−protein functional linkage prediction methods: phylogenetic profiling, gene neighborhood, co-occurrence of orthologous genes in the same gene clusters, a mirror tree variant, and expression similarity.[91] They showed that metabolic pathways are best predicted by using neighborhood of orthologous genes. They also showed that the effective use of a particular prediction method depends on the pathway under investigation. Although gene clusters and gene neighborhood-based methods are fairly simple to adapt, their reliability depends on the number of the genomes used.

7.3. Interolog Search Methods

The basic assumption in these methods is the conservation of interactions among species. Protein−protein interactions can be transferred between different species. PPIs within an organism or between two organisms can therefore be predicted based on the known interactions of the orthologous genes of other organisms (interologs). Usually model organisms are used to find interologs of higher eukaryotes. However, the prediction is also possible from higher to lower eukaryotes. Automated and efficient methods that map ortholog interactions are of great interest. Several recent papers describe the success of such methods: Folador et al. described a protocol to map interologs by using public databases and freely available tools.[92] Sahu et al. used interolog and domain mapping to predict the interactions between *Arabidopsis* and a pathogen at genome scale.[93] This study therefore provides the PPI between a host−pathogen system. Around 11 000 interactions were found by both methods where interolog and domain-based methods gave 0.8 million and 86 000interactions, respectively.[94] The BIPS-BIANA (Biologic Interactions and Network Analysis) Interolog Prediction Server[95] offers a web-based tool to facilitate PPI predictions based on interolog information by utilizing integrated multiple interaction. GO functional annotations have been used for ranking predicted interologs.[95] The specificity of the server is between 72% and 98%, whereas sensitivity varies between 1% and 59%, depending on the sequence identity cutoff used to calculate similarities between

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

sequences. Table 4 lists several methods for mapping conserved interactions across different species to find interologs. In interolog search, the first step is searching for the conservation of ortholog proteins in species. Then if two interacting proteins in one species are conserved in the second species, they possibly interact in the second species too. Because some proteins interact through only one domain and that domain can just cover a portion of the whole protein sequence, BLASTing the sequences can fail in ortholog search. Therefore, there is a need for postanalysis for those proteins.

Interolog search reveals the conserved protein interactions. Obligate interactions are evolutionarily more conserved across different species than the transient ones. These results suggest that interolog search is not appropriate for transient protein interaction prediction. The quality of the interologs can be assessed by the correlation in GO molecular functions. Another assessment approach is analyzing domain pair conservation and functional conservation. Additionally, gene coexpression levels can also be used for the assessment of predicted interologs. The advantages of interolog search include its reliable results, although the method is not applicable to large-scale proteomes.

7.4. Phylogenetic Similarity (Profile) and Conservation-Based Methods

The assumption in the phylogenetic profile method is that if two nonhomologous proteins are functionally related (i.e., involved in the same pathway or biological process or the subunits of a macromolecule, etc.) then they may potentially interact and coevolve. One needs a phylogenetic profile of a protein across many organisms.[96] A phylogenetic profile of a protein can be found as a vector with entries indicating whether the protein is present or absent in an organism. Then proteins with similar profiles are clustered. Proteins in a cluster are hypothesized to be functionally related or interacting. The Clusters of Orthologous Groups (COGs) contains large numbers of profiles. One needs complete genomes of as many as possible organisms to get reliable results. Another effort is based on the rationale that if two nonhomologous genes (proteins) are present and absent together across multiple genomes (proteomes) then these pairs of proteins are likely to be functionally associated. Predicted functional associations can be scored with the probability of observing co-occurences. It is also possible to apply this method to individual domains rather than proteins.[97] This application gives information on domain associations and functional relation of domains. A phylogenetic tree provides an evolutionary link between protein sequences. Interacting proteins tend to have topologically similar phylogenetic trees, and this has been used by some methods to predict interaction partners of proteins (known as mirror tree method).[98] The mirror tree method is based on the hypothesis that proteins that interact coevolve and have orthologs in different organisms. The advantage of the mirror-tree method over the phylogenetic profiles is that it does not require having the fully sequenced genomes, since this method is based on the trees of protein families of interest.

A disadvantage is that phylogenetic profiling method does not provide information for housekeeping or essential proteins, since these proteins will always be present in all organisms. The opposite is also true: the proteins specific to an organism will not be detected with this method either.

Functionally or stability-wise important residues are often evolutionarily conserved. Therefore, finding conserved motifs can help in identifying the binding sites and thus binding partners. The evolutionary trace method is one example which uses not only absolute conservations of key residues of a protein but also divergences in the phylogenetic tree of that sequence family.[99] Residues that are conserved among the widely divergent branches in the evolution tree are expected to have a larger functional impact than other residues that vary among closely related species.

7.5. Gene Coexpression-Based Methods

Recent advances in technology make it possible to simultaneously measure the expression levels of all genes in a genome rapidly. Gene coexpression data can be used to identify proteins that are likely to interact.[100] Raw and normalized expression data can be obtained from several sources (i.e., Gene Expression Omnibus (GEO) of NCBI). By applying clustering algorithms, genes with similar expression profiles can be grouped together according to their expression levels. Proteins whose genes exhibit similar patterns of expression across multiple time points or states may then be considered candidates for functional association and possibly direct physical interaction.[101] Gene coexpression data can be combined with other data to increase the accuracy of methods. One should always keep in mind that expression data is a high-throughput data and can be noisy. Using gene coexpression is an indirect way to infer protein interactions which introduces a caveat: protein levels do not correlate perfectly with gene expression levels; therefore, it may be misleading to deduce interaction knowledge from gene expression data.

7.6. Network Topology-Based Methods

Protein interaction networks can be represented as a graph where each node represents a protein and each edge represents an association between two proteins. In this way, many topological properties such as the number of direct or indirect neighbors of a protein and the shortest paths between proteins can be calculated. In general, a small number of proteins called "hubs" in this graph has many interaction partners. Hub proteins are essential for the functionality and integrity of biological processes in the cell. The mathematical representation (usually in the form of adjacency matrix) of a PPI network is useful for identification of functional relations between proteins and prediction of novel interactions as well. In principle, if two proteins have many common partners in the network they tend to function in similar biological processes.[102,103] Further, if there are many shared partners, the sequence, structure, and biochemical properties of these proteins are assumed to be also similar and they are more likely to interact with each other. Here, the advantage is that the method predicts PPI just by topological properties independent from the sequence or structure properties of proteins.[104] To improve the performance of topology-based approaches, Phan and Sternberg[105] developed a new approach which integrates protein sequence, function, and network topology information and predicts protein complexes and function. Alignment of human and yeast interaction networks has been used, and conserved subnetworks have been detected.

7.7. Residue Coupling- and Coevolution-Based Methods

In principle, coordinated changes across proteins in the residue level helps in predicting protein−protein interactions. Co-evolution-based approaches use multiple sequence alignment of a protein family. It has been demonstrated that correlated mutations are important in maintaining protein stability, allosteric pathways, protein function, and folding. Later, this

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

Table 5. Binding Region Prediction

name	web link	feature set	predictor or scoring function
PROFISIS[84]	https://rostlab.org/owiki/index.php/PROFisis	sequence features	neural networks
metaPPI	http://projects.biotec.tu-dresden.de/metappi/	interface prediction by combining results from five prediction servers PPI−Pred, PPISP, PINUP, Promate, and SPPIDER	
PresCont	http://www-bioinf.uni-regensburg.de/	solvent-accessible surface area, hydrophobicity, conservation, and the local environment of each amino acid on the protein surface	SVM
Cons-PPISP	http://pipe.scs.fsu.edu/ppisp.html	position-specific sequence profiles and solvent accessibilities of each residue and its spatial neighbors from the 3D structure	neural network
PINUP	http://sparks.informatics.iupui.edu/PINUP/	side-chain energy score, conservation, and propensity	empirical scoring function
SPPIDER	http://sppider.cchmc.org/	relative surface accessibility fingerprints	SVM and neural networks
InterProSurf[184]	http://curie.utmb.edu/	solvent-accessible surface area and residue propensities	clustering
Meta-PPISP	http://pipe.scs.fsu.edu/meta-ppisp.html	raw scores from cons-PPISP, PINUP, and Promate web servers	linear regression
Patch Finder Plus[185]	http://pfp.technion.ac.il/	computes electrostatic potential and finds the largest positive patch	scoring function
PPI-Pred	http://bioinformatics.leeds.ac.uk/ppi-pred		
ProMate[186]	http://bioportal.weizmann.ac.il/promate/	biophysical properties	calculation of mutual information and clustering
PredUs[187]	https://bhapp.c2b2.columbia.edu/PredUs/	contacting frequencies and solvent-accessible surface areas of the residues and their 14 spatially nearest surface residues	SVM
SHARP2[188]	http://www.bioinformatics.sussex.ac.uk/	solvation potential, hydrophobicity, accessible surface area, residue interface propensity, planarity. and protrusion	scoring function
WHISCY[189]	http://nmr.chem.uu.nl/Software/whiscy/index.html	conservation and residue propensities	scoring function

approach has been extended to identify correlated mutations between protein partners (in silico two-hybrid (I2H)). This method is based on the assumption that interacting proteins should undergo coevolution in order to keep the proteins functional.[106] Because the coevolution of residues is searched for, in addition to predicting interaction partners, residues contributing to binding are also implicitly predicted. Alignments within the same protein families and after concatenation of different protein families are used to prepare the position-specific matrix to find correlated residues. The main limitation in the evolutionary coupling approach is the number of available homologues for multiple sequence alignment. Another challenge is the presence of indirect correlations such as mutations located far from the binding interface but having allosteric effect. Previously, direct and indirect correlations were not distinguishable. In a statistical physics-based method this problem has been solved and direct and indirect correlations between residues were successfully distinguished. Here, mutual information between every combination of positions each from one protein chain is evaluated and a score is calculated. Residue pairs are ranked based on the calculated score. A global inference approach has been applied to distinguish direct and indirect correlation. In this way, the specificity of predicted contact pairs has been improved.[107] In a recent work, 50 *E. coli* protein complexes with unknown structures have been analyzed with the coevolution approach and possible contacts have been identified. Using the evolutionary couplings, 3D models of the complexes have been built.[108]

7.8. Sequence-Only-Based Methods

Some methods use only amino acid sequence information on the putative target proteins of particular interest to predict whether they interact or not. One of the most remarkable works using sequence was performed by Shen et al. (2007).[82] Amino acids were grouped into seven classes based on the dipoles and volumes of side chains. Following that the "conjoint triplet method" was used in order to create input vectors for SVM. The positive interactions used in a training set were human PPIs taken from HPRD. The negative interaction set was also created. There were 16 000 positive and 16 000 negative interactions classified via SVM with a 5-fold cross-validation technique. The positive precision was reported as 84%.

The advantage of sequence-only-based methods is that obtaining sequences is straightforward and there is no need for additional data like phylogenetic profiles or structures of proteins; thus, sequence-based methods are more universal and applicable for large-scale predictions. The disadvantage of the methods is that the accuracy might be lower compared to the other methods.

8. PREDICTION OF BINDING REGIONS

As the knowledge on the characteristics of protein interfaces ascends, it has become clear that not a single property can significantly distinguish the binding region from the rest of the protein surfaces. In the binding region prediction approaches, the principle is to find the set of best discriminating binding site properties of proteins from the rest of the protein surface and then integrate these properties into a predictor or a scoring function. In Table 5, we listed available approaches and web servers for this purpose and added the feature set and the predictor type to the table. These types of approaches only predict binding regions on protein structures, but they do not address the question of which partner proteins are interacting through this region. The most preferred predictors in this type of approaches are support vector machines and neural networks. For the training, the features include conservation, patch shape, residue propensity, sequence profiles, hydrophobicity, electrostatic potential, accessible surface area, and

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

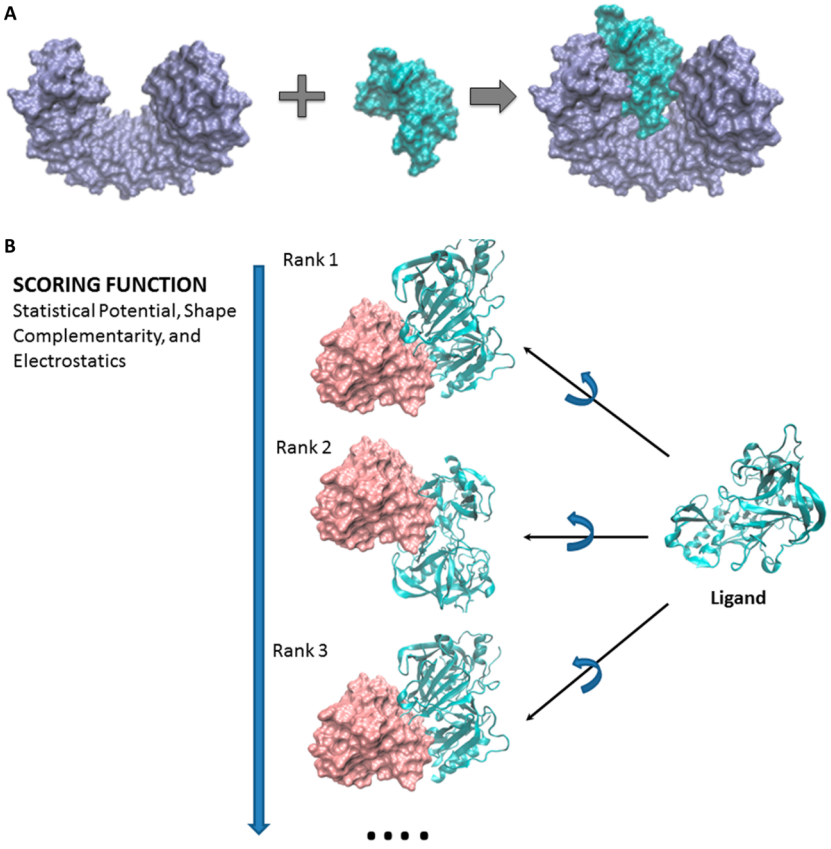


Figure 9. (a) Simple illustration of docking concept where two proteins are docked. (b) Example of docking xylanase and xylanase inhibitor with ZDock[115] approach where different orientations of ligand molecule and their ranking with a scoring function integrating statistical potential, shape complementarity, and electrostatics.

properties of spatial neighbors of each surface residue. For example, residue conservation alone is only marginally discriminative between the interface and the remaining surface region; however, it is an important feature to support other interface properties, for example, with the inclusion of evolutionary divergence.[99] PROFisis is an example of a learning-based approach to predict protein binding sites from sequence by using a neural network.[84] Features for the predictor are learned from 3D structures of transient protein interfaces. Its accuracy reaches 90% in a cross-validation experiment. In another approach, solvation potential, hydrophobicity, accessible surface area, residue interface propensity, planarity, and protrusion features are calculated for patches on the protein surface and combined in a scoring function. A patch with the highest score is predicted to be the binding region. In meta approaches, possible binding patches found by multiple predictors are evaluated if a residue is labeled as binding residue by only one predictor or it is predicted to be in the binding site by all predictors to improve the accuracy.[109] Also, these meta approaches allow incorporating new binding site features and check if the new features are redundant with already known features or they are not discriminating or they are orthogonal to already known significant binding features and improve the prediction performance.

9. STRUCTURE-BASED APPROACHES TO PREDICT PPI

9.1. Docking

Docking is a computational modeling approach for predicting the binding orientation of two protein structures, calculating the binding free energy, and finding the structural assemblies[110] (Figure 9A). The first stage of docking is searching for the possible binding orientations between two proteins. Binding orientations can be found by global and local searches. In a global search, usually one protein is kept static and called the "receptor" and the other protein, called the "ligand" molecule, is rotated around the receptor. The global search finishes when all possible orientations between two proteins in 3D space are sampled. Therefore, a global search is computationally very expensive and requires many translations and rotations. Fast-Fourier transform (FFT) is an example approach to reduce this computational cost in global search strategies.

In the local search approach, surface features are obtained such as the flatness of a patch, pockets, spurs, and solvent-excluded regions on the target surface. Then these local features from each target are matched to obtain a good complementarity. Because the search is local, when two targets are translated based on the local patches, steric clashes between the remaining regions of two proteins are possible. Therefore, there is a need for filtering the predicted complexes if they have atomic clashes.

Another sampling approach is integrating prior knowledge about the interaction. If any biochemical, biophysical, chemical shifting, mutation, or computationally predicted binding patch information is available that can be integrated into the sampling stage. In this way, the search space can be restricted and the accuracy of the prediction can be improved. The sampling stage can be rigid or flexible. In rigid-body sampling, there is no modification in the structure-related features such as bond angles, backbone orientation, and bond lengths. In flexible

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

Table 6. List of Some Docking and Refinement Tools

software name	web site	scoring	type
ZDOCK[115]	http://zdock.umassmed.edu/	shape complementarity, electrostatics, and a pairwise atomic statistical potential developed using contact propensities of transient protein complexes	docking
HEX	http://hexserver.loria.fr/	shape complementarity electrostatic contribution	docking
ClusPro[190]	http://cluspro.bu.edu/	shape complementarity electrostatic and desolvation contributions, pairwise potentials	docking
PatchDock[191]	http://bioinfo3d.cs.tau.ac.il/PatchDock/	geometric fit and atomic desolvation energy	docking
HADDOCK[192]	http://www.nmr.chem.uu.nl/haddock/	ambiguous interaction restraints, buried surface area, electrostatic and desolvation contributions	docking
SwarmDock[193]	http://bmm.cancerresearchuk.org/~SwarmDock/		docking
RosettaDock[194]	http://graylab.jhu.edu/docking/rosetta/	van der Waals, solvation, and hydrogen bond energies	docking
FireDock[195]	http://bioinfo3d.cs.tau.ac.il/FireDock	side chain optimization, energy score	refinement
FiberDock[117]	http://bioinfo3d.cs.tau.ac.il/FiberDock	backbone refinement, side chain optimization, energy score	refinement
ZRANK[196]		electrostatics, van der Waals, and desolvation	refinement
pyDock[197]	http://life.bsc.es/servlet/pydock/home/	electrostatics and desolvation scoring	refinement

sampling, conformational changes in protein structures are taken into account. In rigid-body docking, if a target protein undergoes a significant conformational change during binding, finding the correct orientation is challenging. The sampling stage produces tens to thousands of putative protein complexes. Scoring these putative complexes is crucial to rank them and obtain the best set of solutions. Scoring can be done simultaneously in the sampling stage or right after the sampling stage. Calculated binding free energies by using force fields, shape complementarity measures, and electrostatic complementarity are some of the scoring components to evaluate how tight the predicted binding is.

A docking benchmark, which provides a nonredundant set of protein complexes and their unbound structures, is a valuable source for the systematic assessment of different docking approaches and scoring schemes. The most important aspect is that the docking benchmark has to be as diverse as possible and different types of protein interactions have to be covered. Because each interaction type has different physicochemical aspects, a docking approach has to be able to handle diversity in interactions and be unbiased toward specific interaction types.[111] The latest version of the docking benchmark contains 230 complexes and their unbound structures.[111,112] It contains different types of protein interactions such as enzyme−substrate, antigen−antibody, and other types of interactions. Although the number of complexes and their diversity is limited in the docking benchmark, the benchmark set can be enlarged with the continuous updates in the PDB. The benchmark can be divided into three where the rigid-body set can be used for validation of rigid-body docking approaches. Medium and difficult sets in the docking benchmark comprise conformational changes upon binding. The RMSD of C^{α} atoms of interface residues of the unbound state and bound state implies how difficult the docking is.

The success of the docking predictions on the benchmark set can be assessed by using several criteria. The RMSD of the interface between predicted and native interface regions measures how accurately the interface region has been predicted. Superimposing the receptor molecules in native complex and predicted complex and calculating the RMSD of the ligand molecule measures how accurate the overall binding orientation is. Another measure is the percentage of the predicted interface residues to be in the native interface. The Critical Assessment of PRedicted Interactions (CAPRI) challenge was started to assess the performance of the docking algorithms and scoring approaches which has led to significant improvements in available approaches and development of novel docking approaches.[113] CAPRI is a community-wide blind protein−protein prediction experiment designed to assess the performance of different docking methods. A set of unbound protein structures is released, and many groups averaging about 40 per round join in the challenge to predict the bound states of protein complexes and to evaluate their approaches. Then each group is ranked based on the number of predictions: those are high, medium, acceptable or not acceptable.[114] We should also note that structurally modeling proteome-scale PPIs is not easy since not all of the structures in any proteome are available as X-ray or NMR structures. Therefore, model structures are sometimes needed. Models are less accurate than the X-ray structures; thus, special attention is needed in the development of a methodology for modeling of their complexes.

ZDock is one of the most popular docking algorithms which uses FFT-based global search.[115] In Figure 9B, the top three possible orientations of the receptor molecule (Xylanase) and the ligand molecule (Xylanase Inhibitor) sampled by ZDock are illustrated. The scoring function to rank these predictions is composed of statistical potential, shape complementarity, and electrostatic. Another docking tool is RosettaDock, which uses the Monte Carlo search method.[116] The algorithm starts with either a random orientation of two targets or a user-defined initial pose. This step is coarse grained, and each side chain is represented with a centroid. After a 500 step Monte Carlo search, the lowest energy structure is selected and passed to the high-resolution refinement stage. At this stage, centroids are converted to their initial unbound side chain orientation and the structure undergoes a minimization step. Finally, a score is calculated by an all-atom energy function.[116]

Given the limitations of docking techniques, refinement approaches have been developed for improving the final set of solutions and their rankings.[117] Refinement techniques consider the hits obtained by the docking approaches, refine them by considering flexibility, scoring, and biological information if available, and can optimize the final putative complex

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

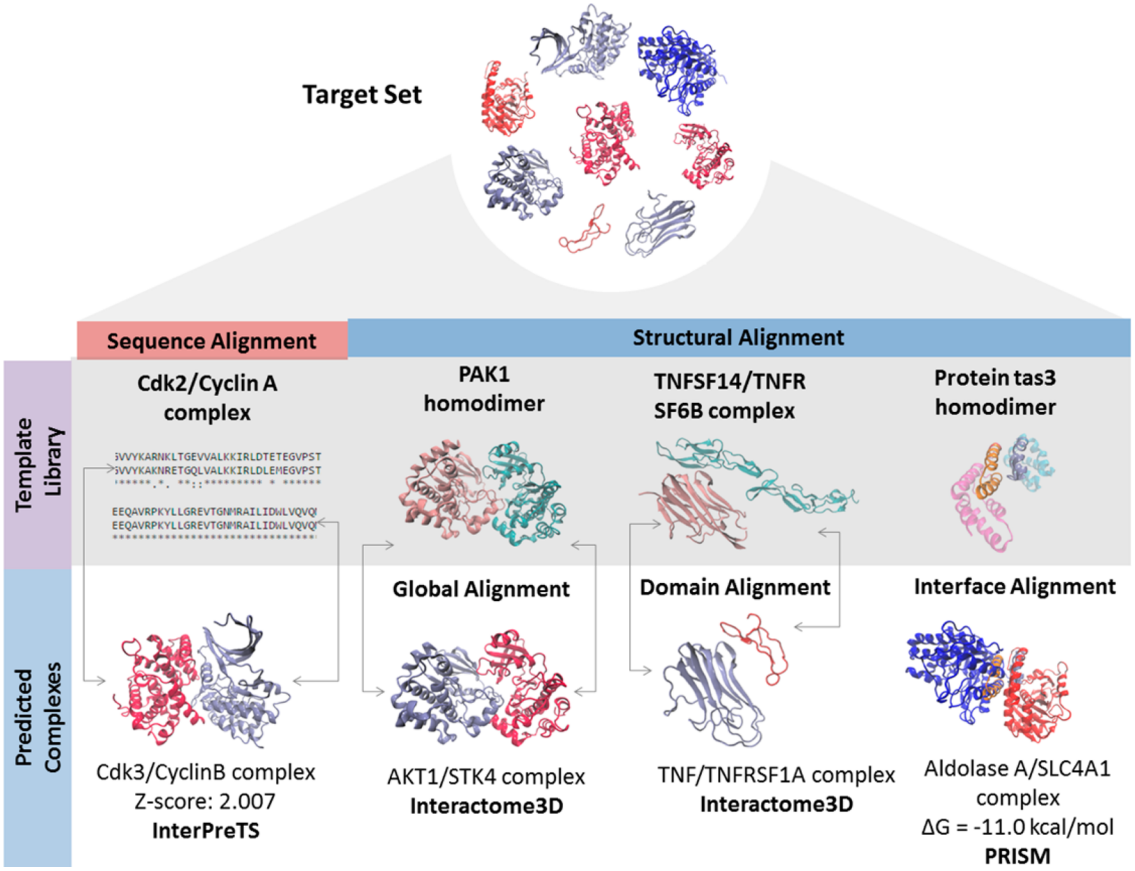


Figure 10. Conceptual illustration of template-based approaches. Three different methods and their predicted complexes are presented. For a constructed template library and target set, the user can utilize sequence similarity-based approaches such as InterPreTS to find the predicted complex as in Cdk3/CyclinB complex. Structural alignment is another option where global protein structure or functional domain similarity is searched in Interactome3D[73] and template interface partners are spatially searched on the target protein surfaces in PRISM.[119]

Table 7. Template-Based PPI Prediction Approaches

name	web link	modeling approach	refinement and scoring
Struct2Net[198]	http://groups.csail.mit.edu/cb/struct2net/webserver/	structure-based threading	logistic regression to evaluate the probability of interaction
iWrap[123]	http://iwrap.csail.mit.edu	threading on only the protein−protein interface	boosting classifier are used to compute a probability of the interaction
INstruct[199]	http://instruct.yulab.org/about.html	domain homology	no scoring
InterPreTS	http://www.russelllab.org/cgi-bin/tools/interprets.pl	homology modeling	empirical potentials and statistical significance
COTH[200]	http://zhanglab.ccmb.med.umich.edu/COTH/	multimeric threading	combination of alignment score, solvent accessibility, hydrophobicity, match between query, and native protein interfaces
HOMCOS[201]	http://homcos.pdbj.org/	homology modeling	no scoring
Interactome3D[73]	http://interactome3d.irbbarcelona.org/	both global structural homology and domain−domain templates	no scoring
PrePPI[124]	https://bhapp.c2b2.columbia.edu/PrePPI/	template-based	Bayesian classification
PRISM[119]	https://prism.ccbb.ku.edu.tr/prism_protocol/	protein−protein interface templates.	flexible refinement and energy calculation
MULTIPROSPECTOR[202]	N/A	sequence threading	total energy, interfacial energy and z-scores
M-Tasser[203]	N/A	sequence threading	iterative refinement and clustering.
ISearch[122]	N/A	domain-domain interface templates.	no scoring

accordingly. The improvements made by the refinement approaches have been demonstrated in CAPRI challenges as well. Some of the available docking and refinement approaches are listed in Table 6.

Because blind docking is computationally heavy and produces many possible orientations of protein complexes, application of classical docking techniques is challenging at the proteome level. Therefore, knowledge-based approaches have emerged for structural assembly finding which consider

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

available signatures of protein binding.[110,118] These approaches will be reviewed in the next section as template-based docking approaches.

9.2. Template-Based Prediction of Protein Assemblies

The structural aspects and physicochemical properties of protein interactions inspired the idea that evolutionary information in terms of sequence or structural similarity of proteins and binding interfaces can be used to model unknown assemblies of proteins.[119] Usually homologous protein pairs have a tendency to use the same binding interfaces. Another aspect of binding implies that the interface region of some protein pairs can be structurally conserved, although their global structures are completely different.[46] Template-based approaches, in principle, depend on inferring information from experimentally solved protein complex structures to predict new structural assemblies of target proteins.[110,118,120] Stages of template-based prediction are as follows: template library preparation, target set selection, template-to-target similarity search, refinement, and scoring. The most important stage in these types of approaches is the template library preparation, because the prediction performance is limited by the diversity of the templates. As the coverage of the PDB increases the performance of template-based approaches also improves. The annual statistics of the PDB show that the number of experimentally solved structures increases exponentially. The target set can be composed of two proteins, or all components of a pathway, or the whole proteome. Since template-based approaches are computationally efficient, prediction both at the proteome level and at a low level is possible. Also, the solution space is limited in template-based approaches; therefore, it does not produce many possible conformations as in docking. The template-to-target similarity search stage can be divided into two at the very top level: (i) global similarity and (ii) local similarity-based approaches. The similarity can be searched by threading, sequence alignment, or structural alignment. For structural alignment, the overall protein structure of the template and target can be used, or domain matches can be searched for, or only the interface region of the template complex can be searched for on the target surfaces. A summary of template-based methods is illustrated in Figure 10. Also, a list of some available template-based approaches is tabulated in Table 7. In sequence-based approaches, if target structures are similar to each complementary chain of a template protein complex in sequence then their binding orientation is assumed to be the same. Here, the matching is solely dependent on sequence similarity scores. Although sequence similarity implies functional similarity between proteins, completely different proteins in sequence can have similar global folds. Therefore, the prediction space of sequence similarity-based approaches is limited.

Proteins having low sequence similarity can have similar global folds. Searching for global structural similarity of target proteins on the template protein complexes is an alternative method to model protein assemblies. Besides the global structural similarity, the presence of domains in template proteins is another measure for prediction and can be considered in this classification. Interactome3D is a database constructed by a hybrid method searching for sequence and structural similarity of the complete proteome to the structurally solved experimental complexes as well as the domain occurrence in the target-to-template comparisons.[73] When all interactions deposited in a set of PPI databases are considered, Interactome3D can model 54.9% of these interactions with complete or partial structures.[73]

Threading is an approach to model protein structures based on fold similarity by using known protein structures as template.[121] In this way, target protein sequence is threaded onto a template structure in a homology-independent manner. In threading-based methods to predict protein assemblies, two stages are available. First, each target sequence is separately threaded to the structurally solved template monomers and candidate structures for each target sequence are found. Then two target sequences join each other for dimeric threading to structurally solved template protein complexes. Finally, best ranking templates from monomeric threading are superimposed onto the best ranking templates from dimeric threading, and a final list of predicted complexes is obtained. Scoring functions which consider energy calculations and statistical measures are used to assess the predicted complex.

Completely different protein pairs can interact via similar 3D interface motifs which opened a new way for modeling protein complexes. Several methods using interface templates have been developed for protein interaction prediction. If partner chains of a protein interface are structurally similar to the surface region on target protein structures, the matching target proteins can be superimposed onto the template interface. In this way, a putative protein complex can be obtained. Then this putative complex can be refined and ranked based on a scoring function. PRISM is the first method using interface templates for modeling protein complexes.[119] The first step is template library preparation where all protein interfaces in the PDB are extracted and redundant interfaces are removed. Then surface regions of target proteins are extracted to be used for structural similarity search. Each interface is separated into the partner chains, and these chains are structurally aligned to the target protein surfaces. The PRISM approach additionally considers the structural conservation of binding hot spots in template interfaces while structural matching. If two targets are similar to the complementary chains of an interface and if there is at least one matching hot spot between the target surface and the template chain then these two targets are assumed to be interacting. Matching targets are superimposed onto the template interfaces; in this way, the first set of putative protein assemblies is obtained. Then PRISM applies a refinement procedure where putative models having steric clashes are eliminated from the final list. Also, with a flexible refinement protocol, side chains and the backbone of the final complex are optimized. For scoring and ranking, binding energies for the modeled complexes are calculated. The flexible refinement stage makes the predictions physically more accurate. Later, new methods like ISEARCH,[122] IWrap,[123] PrePPI,[124] iLoops,[125] KBDock,[126] PAIRPred,[127] EVcomplex[108] used similar approaches based on interface knowledge.

10. COMPARISON OF THE AVAILABLE APPROACHES

Because the performance of each prediction approach is evaluated based on different gold standards and the input required to run each method is different, it is difficult to make a head-to-head comparison. However, each approach reviewed here has its own advantages and disadvantages over others. In general, interolog and gene/domain fusion-based methods are better in predicting permanent interactions compared to other types of interactions and complete genome sequences or domains of several organisms are necessary for the prediction. The network topology, gene cluster, gene neighborhood, and

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

coexpression profile-based approaches are more useful to predict functional interactions instead of direct physical interactions. Residue coupling and coevolution-based approaches perform well in predicting direct and physical interactions. However, the main limitation in this type of approach is the availability of a large number of evolutionarily related sequences to the target proteins. All these approaches attempt addressing the question of which proteins interact with which others. If a user aims to identify the possible binding region on the structure or sequence of a single protein, the methods listed in Table 5 should be browsed. This type of approach does not address the question of which protein pairs interact; rather it predicts possible binding patches. If atomic details of the interaction and structural assemblies are necessary in addition to predicting pairwise interactions, the user needs to browse Table 6 or 7. Template-based approaches are computationally more effective than ab initio docking at the proteome level and their false positive rates are relatively lower.[128,129] However, the performance of template-based approaches is directly dependent on the quality and the coverage of the templates.[49,130] On the basis of the availability of the method either as downloadable data set or a web server for batch run or source codes, the user can choose the method. Depending on the requirements of each prediction method and users' expectation from the output, methods can be run individually or in combination. We should also note that some docking methods are very fast and easily available as a server. A handy guide is provided in Figure 11 for selection of an appropriate database or tool for specific purposes.

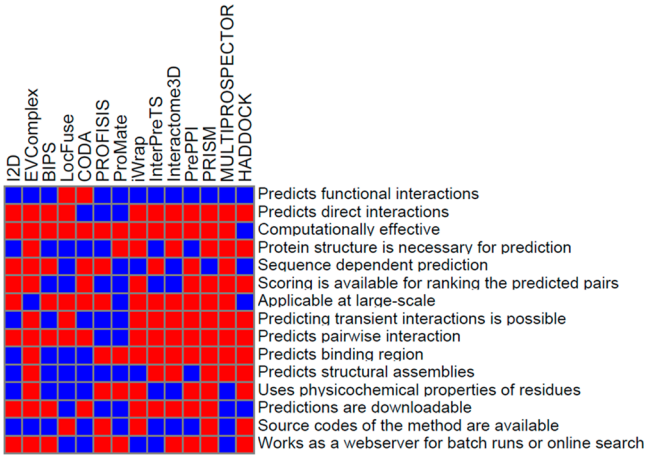


Figure 11. Comparison of some available tools/servers/databases based on a set of annotations. Entries colored red show that the corresponding tool is capable of doing the corresponding annotation (red = "Yes", blue = "No").

11. PPI DATABASES

The primary source of PPI databases is the literature that contains accumulated results of high-throughput and low-throughput experiments of protein interactions. There are various publicly available databases that catalogue PPI information (listed in Table 8). These databases are useful to organize the available data obtained from multiple sources. As interaction−detection methods are improved and the number of sequenced genomes and structurally solved proteins increases, the size of each database enlarges in parallel. As of 2015, over 100 PPI databases and resources, most of them operated independently, are available (as listed in http://pathguide.org). Major problems with the databases are the redundancy, differences in curation and annotation, and interoperability. Given the advantages/drawbacks and differences in each different experimental PPI detection techniques and the diversity of the data available in databases, integration of all information from different sources in an appropriate way emerges toward having a more confident protein interactome. Methodological differences between interaction data sources and low overlap ratios between different experimental PPI databases are other reasons for the necessity of integrating multiple databases and scoring each interaction based on the evidence of being a real interaction.

The major standardization approaches for organizing separate resources include HUPO-PSI (Human Proteome Organization Proteomics Standard Initiative)[131] and IMEx Consortium (International Molecular Exchange Consortium).[132] HUPO-PSI defines a common data format to exchange PPI information, i.e., HUPO-PS-MI XML is now used by many PPI databases to share their interaction data. The IMEx consortium addresses the redundancy of PPI information recorded in different databases and the differences due to curation of the same primary literature. The major databases DIP (Database of Interacting Proteins),[133] IntAct,[134] and MINT (Molecular Interaction Database)[135] are among the core founding partners of IMEx. IMEx provides a common curation guideline to eliminate duplication of curation efforts and to eliminate discrepancies due to methodological differences of independent curation efforts. It provides a nonredundant set of protein interactions through a common interface. Recently, MINT has merged with IntAct to further improve the efficiency of curation and organization efforts. The other major PPI databases include BioGrid (Biological General Repository for Interaction Data set)[136] and HPRD (Human Protein Reference Database).[137] BioGrid is not a part of the IMEx consortium, but an observer member contains an archive of genetic interactions in addition to protein interactions. HPRD is a specialized database providing only curated human protein−protein interaction and other proteomic information including post-translational modifications and tissue expression. An enlarged list of the major PPI databases and their features are listed in Table 8.

Two other major resources that integrate data from PPI databases with advanced search operations are IRefWeb[138] and String (Search Tool for Interacting Genes and Proteins).[90,139] IRefWeb is an interactive web interface to access all the PPI data from major primary databases consolidated by IRefIndex (Interaction Reference Index). IRefWeb allows one to access details of interaction data and supporting evidence such as the number of publications supporting the interaction, the type of experimental detection, and the agreement across databases. String is also a resource combining interactions from primary databases; however, it provides not only direct protein−protein interactions but also protein−protein associations (indirect) based on experimental evidence or predictions using text mining. In addition, its web interface provides visualization of protein−protein associations as a network with visual annotations describing the relations between the proteins and the corresponding references.

Logically, the next step is to represent protein−protein interactions as networks of proteins and provide visualization and analysis tools to relate such networks to biological

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

Table 8. List of Databases Organizing Experimental and Literature-Curated PPIs

name	web link	quality assessment method	number of interactions	number of proteins
DIP[133]	http://dip.doe-mbi.ucla.edu/	curated	78 191	27 098
MINT[135]	http://mint.bio.uniroma2.it/mint/	curated	241 458	35 553
IntAct[134]	http://www.ebi.ac.uk/intact/	curated	456 489	83 574
HPRD[137]	http://www.hprd.org/	curated	41 327	30 047
BIND	http://bind.ca	curated		
MIPS[204]	http://mips.helmholtz-muenchen.de/proj/ppi/	curated		
CORUM[205]	http://mips.helmholtz-muenchen.de/genre/proj/corum	a resource of manually annotated protein complexes from mammalian organisms		
BioGRID[136]	http://thebiogrid.org/	curated protein and genetic interactions from publications.	345 577	53 561
CCSB Interactome Database[12]	http://interactome.dfci.harvard.edu/	high-throughput Y2H, not curated	4303	13 944
InWeb[206]	http://www.broadinstitute.org/mpg/dapple/dapple.php	not curated, confidence scoring	428 430	12 793
STRING[90,139]	http://string-db.org/	not curated, confidence scoring		>5 million
MiMI[207]	http://mimi.ncibi.org/MimiWeb/AboutPage.html	quality assessment and scoring	3.5 million	3.7 million
HIPPIE[208]	http://cbdm.mdc-berlin.de/tools/hippie/information.php	quality assessment and scoring	72 916	11 836
iRefWeb[138]	http://wodaklab.org/iRefWeb	quality assessment and scoring	~18 000 (for human)	~222 098 (for human)
HitPredict[209]	http://hintdb.hgc.jp/htp/	quality assessment and scoring	176 983	36 930
IMID	www.integrativebiology.org	quality assessment and scoring		
HAPPI	http://discern.uits.iu.edu:8340/HAPPI/	quality assessment and scoring	2 922 202	32 125
HUPO	http://www.psidev.info/groups/molecular-interactions	quality assessment and scoring		
Pathway Databases				
KEGG	http://www.kegg.jp/	curated		
BioCarta	http://www.biocarta.com/genes/index.asp			
Reactome[210]	http://www.reactome.org/	curated	7041 (in human)	7460 (in human)
ConsensusPathDB[211]	http://consensuspathdb.org/		416 872	154 537
SPIKE	http://www.cs.tau.ac.il/~spike/	curated	20 412	34 338
NCI-PID	http://pid.nci.nih.gov/index.shtml	curated	9248	

processes and pathways. Table 8 provides some of the very popular protein−protein interaction and pathway databases.

12. PROTEIN−PROTEIN INTERACTION NETWORKS AND VISUALIZATION

As reviewed in the previous sections, huge amounts of interaction data have been accumulated in multiple databases which are both experimental and computational. PPI data can be considered as a graph where each node represents a protein and each edge represents the interaction between two proteins. Visualization is crucial for a better understanding and graph theoretic analysis of these data. The aim is to create intuitive and interactive visualization of the PPI network data that will not be lost within the complexity of these networks.[140] While some protein interaction databases provide their built-in web-based visualization resources, such as String database (Figure 12), there are many independent types of software to visualize protein interaction networks as well. Cytoscape is the most popular software for visualization, analysis, and modeling of protein interaction networks.[141] Most of the listed PPI databases are accessible through web services of Cytoscape, and they can be downloaded directly and visualized by Cytoscape. Besides the complete networks of multiple organisms, available functional pathways can also be downloaded through the web services. The network analysis plug-in helps in calculating many topological parameters and centrality measures such as node degree, betweenness centrality, network diameter, clustering coefficient, and shortest path lengths, and these calculations do not require expertise in graph theory from the user. Nodes and edges can be colored according to their biological attributes or topological parameters. Merging, intersecting, or comparing multiple networks are also possible with Cytoscape. It is also possible to install external plugins. For example, a newly released plug-in, CytoStruct, enables combining Cytoscape's network visualization of PPIs with molecular viewers and add a layer of structural analysis of proteins and their interactions.[142] Usually visualization of the whole interaction network of a specific organism gives a hairball-like structure as illustrated in the central network in Figure 1 and visually not so informative. In general, this overall network is divided into functionally-related subnetworks, components of the same protein complexes, proteins in the same cellular compartment, or functional pathways. Network clustering helps in finding high-order protein complexes. Also, reverse engineering and network optimization methods help in finding biologically meaningful subnetworks.[143−146] Some functionally related proteins come together and construct pathways. In Figure 13 a Cytoscape visualization of the insulin signaling pathway is illustrated where nodes are colored based on their cellular localization, and node shapes are selected

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

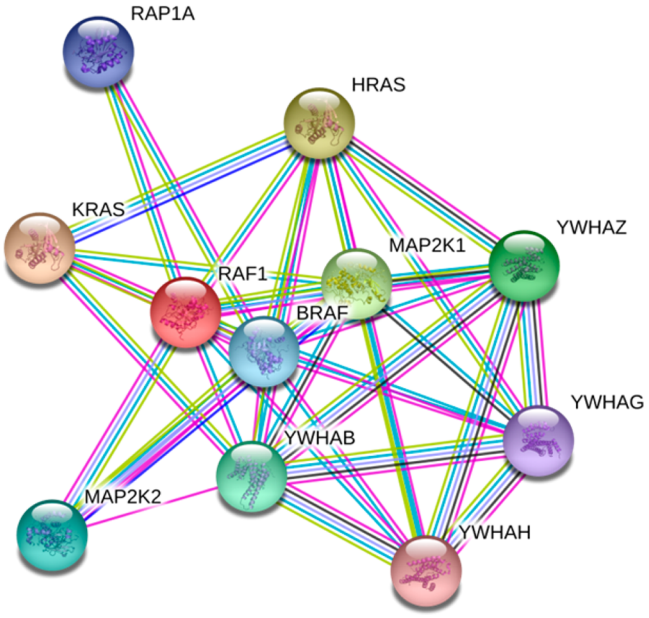


Figure 12. Snapshots from the built-in visualization part of the String database.[90,139] Each node represents a protein, and each edge represents the interaction between two proteins. Edges are composed of different colored lines representing the evidence of the interaction whether it is retrieved from databases or it is experimental data or obtained by text mining.

according to the molecule type. Different from the undirected PPI networks, interactions have a direction in pathway maps, i.e. inhibition and activation. Edge shapes are drawn according to their cellular activity where bar-ending edges represent inhibition and arrow-ending edges represent stimulation.

13. CONCLUSION

With the advancements in technology and techniques used in experiments, more and more PPI data have become available. In parallel, computational methods emerge to validate and complete the missing interactions. Consequently, large amounts of experimental and computational PPI data have been accumulated in diverse sets of databases.

In this review, we aim to introduce protein−protein interactions and provide a broad and informative survey of methods for predicting such interactions, databases available, and tools to analyze the data using various approaches. The review aims to give an unbiased view of the field. We emphasize that each method, either experimental or computational, has its own advantages/disavantages. Apparently, a single method is limited with its reliability and coverage. It is also important to represent protein−protein interactions as networks of proteins and provide visualization and analysis tools to relate the networks to biological processes and pathways and to see how proteins coordinate in pathways. For better results hybrid approaches/metaservers are emerging for such purposes. It is clear that the community is pursuing to integrate and

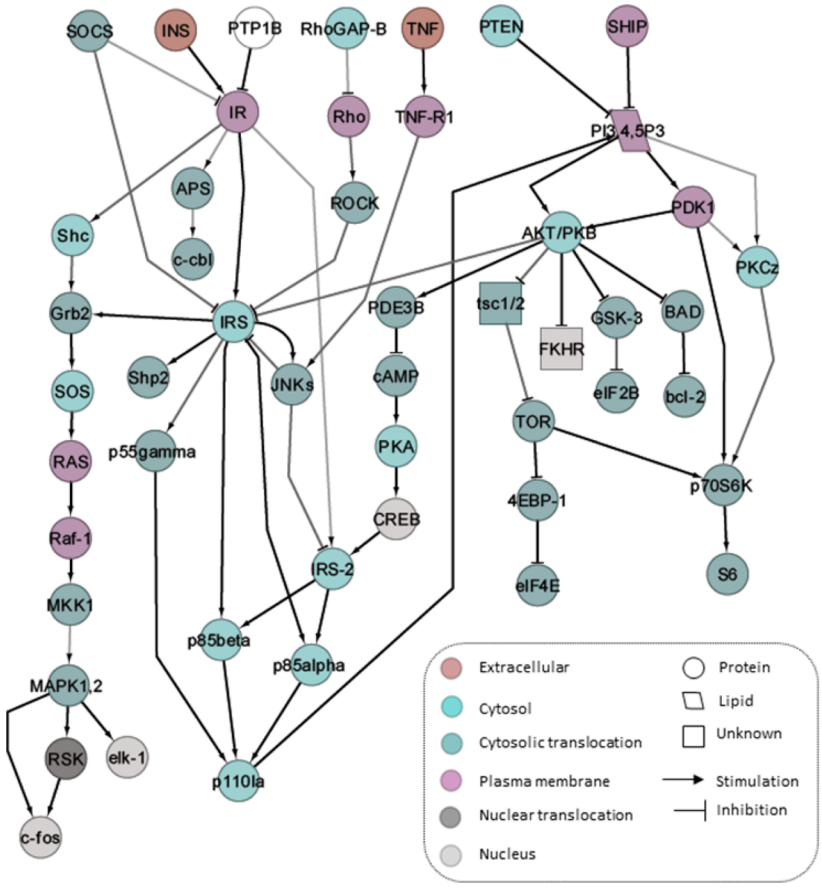


Figure 13. Insulin signaling pathway drawn with the Cytoscape network visualization tool.[141] Pathway information has been obtained from the Database of Cell Signaling at Science Signaling. Nodes represent each entity in the pathway, and edges represent their interactions. These interactions can be inhibition or stimulation as the legend illustrates. Nodes have been colored according to their cellular localization.

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

standardize differently annotated data in an organized way defining a common data format to exchange PPI information. It will be especially important to integrate experimental and computational techniques that complement each other in an organized way.

There are many efforts toward building proteome-scale interactome. However, there are several difficulties: First, there is no clear definition of an interactome. Usually it is defined as the set of molecular interactions in an organism. However, these will change from one cell type to another even within an organism. Further, interactions are dynamic depending on the signal coming from outside sources; therefore, interactions are time and condition dependent. Additionally, proteins are in a crowded cellular environment, and they collide with each other randomly; therefore, it is not certain which interactions are "biological" and which are simply driven by diffusion in vivo. The ultimate goal is to put all these findings and data in the cellular environment, consider their interactions with different compartments in the cell, i.e., the membrane, mitochondria, and the cytoskeleton, and merge these data with their time dependence, stability, affinity, and dynamics to gain further insight into cellular mechanisms. We expect that in the future a well-defined proteome-scale map of protein interactions will be put together in high resolution obtained by the integrative approaches toward "the human interactome".

AUTHOR INFORMATION

Corresponding Authors

*E-mail: okeskin@ku.edu.tr.
*E-mail: ntuncbag@metu.edu.tr.
*E-mail: agursoy@ku.edu.tr.

Notes

The authors declare no competing financial interest.

Biographies

Currently, Ozlem Keskin is a professor in the Chemical and Biological Engineering Department at Koc University, Istanbul. She was a postdoctoral fellow at the National Cancer Institute−National Institutes of Health, U.S.A., from 1999 to 2001. She received her Ph.D. degree in Chemical Engineering in 1999, at Bogazici University, Istanbul. She is a member of the Science Academy, Turkey, and recipient of several awards including the Science Award, Turkey, 2012, and UNESCO-L'OREAL Co-Sponsored Fellowship Award for Young Women in Life Sciences, 2005. She is an Associate Editor in Plos Comp Biol, Plos One, and BMC Structural Biology. Her work focuses on understanding the principles of protein−protein interactions (PPIs), the molecular mechanisms, physical principles, and dynamics of macromolecular systems. She coheads the Computational Systems Biology (COSBI) group with Prof. Gursoy, aiming to construct protein interactomes by integrating atomistic details of protein−protein interfaces (http://home.ku.edu.tr/~okeskin).

Nurcan Tuncbag works in the Department of Health Informatics at Middle East Technical University in Ankara, Turkey since 2014. Her research focus is on structural modeling of protein interactions and revealing how the genome and the networks of interactions among proteins and genes are altered in cells during cancer. She received her B.S. degree in Chemical Engineering from Istanbul Technical University (ITU) in 2005 and the M.S. and Ph.D. degrees in Computational Science and Engineering from Koc University in 2007 and 2010, respectively. In 2010, she joined the Department of Biological Engineering at Massachusetts Institute of Technology (MIT) as a postdoctoral associate.

Attila Gursoy is a Professor of Computer Science in the Department of Computer Engineering, Koç University. He received his Ph.D. degree from the University of Illinois at Urbana−Champaign in Computer Science in 1994. He was a postdoctoral research associate at the Theoretical and Computational Biophysics Group at Beckman Institute, UIUC, from 1994 to 1996. His research is in the area of computational biology and high-performance computing, particularly on protein−protein interactions and signaling networks. He acts as an academic editor of PLOS ONE. He serves as an executive committee member of the Informatics group of TUBITAK (the National Scientific and Technological Research Council of Turkey). He is a founding member of the Turkish Bioinformatics Association. He received the Werner-von-Siemens Excellence Award for Science and Innovation for his work on high-performance computational biology. He is a member of Science Academy (Turkey).

ACKNOWLEDGMENTS

N.T. thanks the TUBITAK-Marie Curie Co-funded Brain Circulation Scheme (114C026) and Young Scientist Award Program of the Science Academy (Turkey) for support. O.K. and A.G. are members of the Science Academy (Turkey). We acknowledge partial funding from Tubitak projects (114M196 and 113E164).

REFERENCES

(1) Wang, E. T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; Mayr, C.; Kingsmore, S. F.; Schroth, G. P.; Burge, C. B. Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* **2008**, *456*, 470−476.
(2) Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S.; et al. Ensembl 2014. *Nucleic Acids Res.* **2014**, *42*, D749−755.
(3) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. A Draft Map of the Human Proteome. *Nature* **2014**, *509*, 575−581.
(4) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-Spectrometry-Based Draft of the Human Proteome. *Nature* **2014**, *509*, 582−587.
(5) Berggard, T.; Linse, S.; James, P. Methods for the Detection and Analysis of Protein-Protein Interactions. *Proteomics* **2007**, *7*, 2833−2842.
(6) Nooren, I. M.; Thornton, J. M. Diversity of Protein-Protein Interactions. *EMBO J.* **2003**, *22*, 3486−3492.
(7) Hart, G. T.; Ramani, A. K.; Marcotte, E. M. How Complete Are Current Yeast and Human Protein-Interaction Networks? *Genome Biol.* **2006**, *7*, 120.
(8) Hakes, L.; Pinney, J. W.; Robertson, D. L.; Lovell, S. C. Protein-Protein Interaction Networks and Biology–What's the Connection? *Nat. Biotechnol.* **2008**, *26*, 69−72.
(9) Menche, J.; Sharma, A.; Kitsak, M.; Ghiassian, S. D.; Vidal, M.; Loscalzo, J.; Barabasi, A. L. Disease Networks. Uncovering Disease-Disease Relationships through the Incomplete Interactome. *Science* **2015**, *347*, 1257601.
(10) Stumpf, M. P.; Thorne, T.; de Silva, E.; Stewart, R.; An, H. J.; Lappe, M.; Wiuf, C. Estimating the Size of the Human Interactome. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 6959−6964.
(11) Venkatesan, K.; Rual, J. F.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, K. I.; et al. An Empirical Framework for Binary Interactome Mapping. *Nat. Methods* **2009**, *6*, 83−90.
(12) Rolland, T.; Tasan, M.; Charloteaux, B.; Pevzner, S. J.; Zhong, Q.; Sahni, N.; Yi, S.; Lemmens, I.; Fontanillo, C.; Mosca, R.; et al. A Proteome-Scale Map of the Human Interactome Network. *Cell* **2014**, *159*, 1212−1226.

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

(13) Petschnigg, J.; Snider, J.; Stagljar, I. Interactive Proteomics Research Technologies: Recent Applications and Advances. *Curr. Opin. Biotechnol.* **2011**, *22*, 50−58.
(14) Stynen, B.; Tournu, H.; Tavernier, J.; Van Dijck, P. Diversity in Genetic in Vivo Methods for Protein-Protein Interaction Studies: From the Yeast Two-Hybrid System to the Mammalian Split-Luciferase System. *Microbiol. Mol. Biol. Rev.* **2012**, *76*, 331−382.
(15) Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 4569−4574.
(16) Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. A Comprehensive Analysis of Protein-Protein Interactions in Saccharomyces Cerevisiae. *Nature* **2000**, *403*, 623−627.
(17) Huang, H.; Jedynak, B. M.; Bader, J. S. Where Have All the Interactions Gone? Estimating the Coverage of Two-Hybrid Protein Interaction Maps. *PLoS Comput. Biol.* **2007**, *3*, e214.
(18) Suter, B.; Kittanakom, S.; Stagljar, I. Two-Hybrid Technologies in Proteomics Research. *Curr. Opin. Biotechnol.* **2008**, *19*, 316−323.
(19) Yu, H.; Braun, P.; Yildirim, M. A.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.; Simonis, N.; et al. High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science* **2008**, *322*, 104−110.
(20) Bensimon, A.; Heck, A. J.; Aebersold, R. Mass Spectrometry-Based Proteomics and Network Biology. *Annu. Rev. Biochem.* **2012**, *81*, 379−405.
(21) Heck, A. J. Native Mass Spectrometry: A Bridge between Interactomics and Structural Biology. *Nat. Methods* **2008**, *5*, 927−933.
(22) Leitner, A.; Walzthoeni, T.; Kahraman, A.; Herzog, F.; Rinner, O.; Beck, M.; Aebersold, R. Probing Native Protein Structures by Chemical Cross-Linking, Mass Spectrometry, and Bioinformatics. *Mol. Cell. Proteomics* **2010**, *9*, 1634−1649.
(23) Jager, S.; Cimermancic, P.; Gulbahce, N.; Johnson, J. R.; McGovern, K. E.; Clarke, S. C.; Shales, M.; Mercenne, G.; Pache, L.; Li, K.; et al. Global Landscape of Hiv-Human Protein Complexes. *Nature* **2012**, *481*, 365−370.
(24) Gavin, A. C.; Maeda, K.; Kuhner, S. Recent Advances in Charting Protein-Protein Interaction: Mass Spectrometry-Based Approaches. *Curr. Opin. Biotechnol.* **2011**, *22*, 42−49.
(25) Pu, S.; Vlasblom, J.; Turinsky, A.; Marcon, E.; Phanse, S.; Trimble, S. S.; Olsen, J.; Greenblatt, J.; Emili, A.; Wodak, S. J. Extracting High Confidence Protein Interactions from Affinity Purification Data: At the Crossroads. *J. Proteomics* **2015**, *118*, 63−80.
(26) Braun, P.; Tasan, M.; Dreze, M.; Barrios-Rodiles, M.; Lemmens, I.; Yu, H.; Sahalie, J. M.; Murray, R. R.; Roncari, L.; de Smet, A. S.; et al. An Experimentally Derived Confidence Score for Binary Protein-Protein Interactions. *Nat. Methods* **2009**, *6*, 91−97.
(27) Miller, B. W.; Lau, G.; Grouios, C.; Mollica, E.; Barrios-Rodiles, M.; Liu, Y.; Datti, A.; Morris, Q.; Wrana, J. L.; Attisano, L. Application of an Integrated Physical and Functional Screening Approach to Identify Inhibitors of the Wnt Pathway. *Mol. Syst. Biol.* **2009**, *5*, 315.
(28) Ozbabacan, S. E.; Engin, H. B.; Gursoy, A.; Keskin, O. Transient Protein-Protein Interactions. *Protein Eng., Des. Sel.* **2011**, *24*, 635−648.
(29) La, D.; Kong, M. S.; Hoffman, W.; Choi, Y. I.; Kihara, D. Predicting Permanent and Transient Protein-Protein Interfaces. *Proteins: Struct., Funct., Genet.* **2013**, *81*, 805−818.
(30) Mintseris, J.; Weng, Z. P. Structure, Function, and Evolution of Transient and Obligate Protein-Protein Interactions. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 10930−10935.
(31) Xu, Z. H.; Horwich, A. L.; Sigler, P. B. The Crystal Structure of the Asymmetric Groel-Groes-(Adp)(7) Chaperonin Complex. *Nature* **1997**, *388*, 741−750.
(32) Yusupov, M. M.; Yusupova, G. Z.; Baucom, A.; Lieberman, K.; Earnest, T. N.; Cate, J. H. D.; Noller, H. F. Crystal Structure of the Ribosome at 5.5 Angstrom Resolution. *Science* **2001**, *292*, 883−896.
(33) Plowman, S. J.; Hancock, J. F. Ras Signaling from Plasma Membrane and Endomembrane Microdomains. *Biochim. Biophys. Acta, Mol. Cell Res.* **2005**, *1746*, 274−283.
(34) Pogoryelov, D.; Krah, A.; Langer, J. D.; Yildiz, O.; Faraldo-Gomez, J. D.; Meier, T. Microscopic Rotary Mechanism of Ion Translocation in the F(O) Complex of Atp Synthases. *Nat. Chem. Biol.* **2010**, *6*, 891−899.
(35) Jones, S.; Thornton, J. M. Principles of Protein-Protein Interactions. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 13−20.
(36) Levy, E. D.; Teichmann, S. Structural, Evolutionary, and Assembly Principles of Protein Oligomerization. *Prog. Mol. Biol. Transl* **2013**, *117*, 25−51.
(37) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, W.; Garner, E. C.; Obradovic, Z. Intrinsically Disordered Protein. *J. Mol. Graphics Modell.* **2001**, *19*, 26−59.
(38) Varadi, M.; Vranken, W.; Guharoy, M.; Tompa, P. Computational Approaches for Inferring the Functions of Intrinsically Disordered Proteins. *Frontiers in molecular biosciences* **2015**, *2*, 45.
(39) Berlow, R. B.; Dyson, H. J.; Wright, P. E. Functional Advantages of Dynamic Protein Disorder. *FEBS Lett.* **2015**, *589*, 2433−2440.
(40) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114*, 6589−6631.
(41) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(42) Janin, J. Specific Versus Non-Specific Contacts in Protein Crystals. *Nat. Struct. Biol.* **1997**, *4*, 973−974.
(43) Krissinel, E.; Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **2007**, *372*, 774−797.
(44) Xu, Q. F.; Dunbrack, R. L. The Protein Common Interface Database (Protcid)-a Comprehensive Database of Interactions of Homologous Proteins in Multiple Crystal Forms. *Nucleic Acids Res.* **2011**, *39*, D761−D770.
(45) Duarte, J. M.; Srebniak, A.; Scharer, M. A.; Capitani, G. Protein Interface Classification by Evolutionary Analysis. *BMC Bioinf.* **2012**, *13*, 334.
(46) Keskin, O.; Tsai, C. J.; Wolfson, H.; Nussinov, R. A New, Structurally Nonredundant, Diverse Data Set of Protein-Protein Interfaces and Its Implications. *Protein Sci.* **2004**, *13*, 1043−1055.
(47) Valdar, W. S.; Thornton, J. M. Conservation Helps to Identify Biologically Relevant Crystal Contacts. *J. Mol. Biol.* **2001**, *313*, 399−416.
(48) Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. A Dissection of Specific and Non-Specific Protein-Protein Interfaces. *J. Mol. Biol.* **2004**, *336*, 943−955.
(49) Gao, M.; Skolnick, J. Structural Space of Protein-Protein Interfaces Is Degenerate, Close to Complete, and Highly Connected. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 22517−22522.
(50) Keskin, O.; Gursoy, A.; Ma, B.; Nussinov, R. Principles of Protein-Protein Interactions: What Are the Preferred Ways for Proteins to Interact? *Chem. Rev.* **2008**, *108*, 1225−1244.
(51) Bogan, A. A.; Thorn, K. S. Anatomy of Hot Spots in Protein Interfaces. *J. Mol. Biol.* **1998**, *280*, 1−9.
(52) Clackson, T.; Wells, J. A. A Hot Spot of Binding Energy in a Hormone-Receptor Interface. *Science* **1995**, *267*, 383−386.
(53) Tuncbag, N.; Keskin, O.; Gursoy, A. Hotpoint: Hot Spot Prediction Server for Protein Interfaces. *Nucleic Acids Res.* **2010**, *38*, W402−406.
(54) Kortemme, T.; Baker, D. A Simple Physical Model for Binding Energy Hot Spots in Protein-Protein Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 14116−14121.

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

(55) Darnell, S. J.; LeGault, L.; Mitchell, J. C. Kfc Server: Interactive Forecasting of Protein Interaction Hot Spots. *Nucleic Acids Res.* **2008**, *36*, W265−269.
(56) Ofran, Y.; Rost, B. Protein-Protein Interaction Hotspots Carved into Sequences. *PLoS Comput. Biol.* **2007**, *3*, e119.
(57) Tuncbag, N.; Gursoy, A.; Keskin, O. Identification of Computational Hot Spots in Protein Interfaces: Combining Solvent Accessibility and Inter-Residue Potentials Improves the Accuracy. *Bioinformatics* **2009**, *25*, 1513−1520.
(58) Zhu, X.; Mitchell, J. C. Kfc2: A Knowledge-Based Hot Spot Prediction Method Based on Interface Solvation, Atomic Density, and Plasticity Features. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 2671−2683.
(59) Gonzalez-Ruiz, D.; Gohlke, H. Targeting Protein-Protein Interactions with Small Molecules: Challenges and Perspectives for Computational Binding Epitope Detection and Ligand Finding. *Curr. Med. Chem.* **2006**, *13*, 2607−2625.
(60) Huo, S.; Massova, I.; Kollman, P. A. Computational Alanine Scanning of the 1:1 Human Growth Hormone-Receptor Complex. *J. Comput. Chem.* **2002**, *23*, 15−27.
(61) Keskin, O.; Ma, B.; Nussinov, R. Hot Regions in Protein−Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues. *J. Mol. Biol.* **2005**, *345*, 1281−1294.
(62) Li, X.; Keskin, O.; Ma, B. Y.; Nussinov, R.; Liang, J. Protein-Protein Interactions: Hot Spots and Structurally Conserved Residues Often Locate in Complemented Pockets That Pre-Organized in the Unbound States: Implications for Docking. *J. Mol. Biol.* **2004**, *344*, 781−795.
(63) Wells, J. A.; McClendon, C. L. Reaching for High-Hanging Fruit in Drug Discovery at Protein-Protein Interfaces. *Nature* **2007**, *450*, 1001−1009.
(64) Morelli, X.; Bourgeas, R.; Roche, P. Chemical and Structural Lessons from Recent Successes in Protein-Protein Interaction Inhibition (2p2i). *Curr. Opin. Chem. Biol.* **2011**, *15*, 475−481.
(65) Ma, B.; Elkayam, T.; Wolfson, H.; Nussinov, R. Protein-Protein Interactions: Structurally Conserved Residues Distinguish between Binding Sites and Exposed Protein Surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 5772−5777.
(66) Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin, J. Dissecting Subunit Interfaces in Homodimeric Proteins. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 708−719.
(67) Kar, G.; Gursoy, A.; Keskin, O. Human Cancer Protein-Protein Interaction Network: A Structural Perspective. *PLoS Comput. Biol.* **2009**, *5*, e1000601.
(68) Nishi, H.; Hashimoto, K.; Panchenko, A. R. Phosphorylation in Protein-Protein Binding: Effect on Stability and Function. *Structure* **2011**, *19*, 1807−1815.
(69) Keskin, O.; Nussinov, R. Similar Binding Sites and Different Partners: Implications to Shared Proteins in Cellular Pathways. *Structure* **2007**, *15*, 341−354.
(70) Tuncbag, N.; Gursoy, A.; Guney, E.; Nussinov, R.; Keskin, O. Architectures and Functional Coverage of Protein-Protein Interfaces. *J. Mol. Biol.* **2008**, *381*, 785−802.
(71) Kim, P. M.; Lu, L. J.; Xia, Y.; Gerstein, M. B. Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights. *Science* **2006**, *314*, 1938−1941.
(72) Tuncbag, N.; Kar, G.; Gursoy, A.; Keskin, O.; Nussinov, R. Towards Inferring Time Dimensionality in Protein-Protein Interaction Networks by Integrating Structures: The P53 Example. *Mol. BioSyst.* **2009**, *5*, 1770−1778.
(73) Mosca, R.; Ceol, A.; Aloy, P. Interactome3d: Adding Structural Details to Protein Networks. *Nat. Methods* **2013**, *10*, 47−53.
(74) Acuner Ozbabacan, S. E.; Gursoy, A.; Nussinov, R.; Keskin, O. The Structural Pathway of Interleukin 1 (Il-1) Initiated Signaling Reveals Mechanisms of Oncogenic Mutations and Snps in Inflammation and Cancer. *PLoS Comput. Biol.* **2014**, *10*, e1003470.
(75) Engin, H. B.; Guney, E.; Keskin, O.; Oliva, B.; Gursoy, A. Integrating Structure to Protein-Protein Interaction Networks That Drive Metastasis to Brain and Lung in Breast Cancer. *PLoS One* **2013**, *8*, e81035.
(76) Kar, G.; Keskin, O.; Nussinov, R.; Gursoy, A. Human Proteome-Scale Structural Modeling of E2-E3 Interactions Exploiting Interface Motifs. *J. Proteome Res.* **2012**, *11*, 1196−1207.
(77) Kuzu, G.; Keskin, O.; Gursoy, A.; Nussinov, R. Constructing Structural Networks of Signaling Pathways on the Proteome Scale. *Curr. Opin. Struct. Biol.* **2012**, *22*, 367−377.
(78) Schreiber, G.; Keating, A. E. Protein Binding Specificity Versus Promiscuity. *Curr. Opin. Struct. Biol.* **2011**, *21*, 50−61.
(79) Kastritis, P. L.; Bonvin, A. M. On the Binding Affinity of Macromolecular Interactions: Daring to Ask Why Proteins Interact. *J. R. Soc., Interface* **2013**, *10*, 20120835.
(80) Qin, S.; Pang, X.; Zhou, H. X. Automated Prediction of Protein Association Rate Constants. *Structure* **2011**, *19*, 1744−1751.
(81) Schreiber, G.; Fleishman, S. J. Computational Design of Protein-Protein Interactions. *Curr. Opin. Struct. Biol.* **2013**, *23*, 903−910.
(82) Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting Protein-Protein Interactions Based Only on Sequences Information. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 4337−4341.
(83) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer, 2007.
(84) Ofran, Y.; Rost, B. Isis: Interaction Sites Identified from Sequence. *Bioinformatics* **2007**, *23*, e13−16.
(85) Blohm, P.; Frishman, G.; Smialowski, P.; Goebels, F.; Wachinger, B.; Ruepp, A.; Frishman, D. Negatome 2.0: A Database of Non-Interacting Proteins Derived by Literature Mining, Manual Annotation and Protein Structure Analysis. *Nucleic Acids Res.* **2014**, *42*, D396−400.
(86) Enright, A. J.; Iliopoulos, I.; Kyrpides, N. C.; Ouzounis, C. A. Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events. *Nature* **1999**, *402*, 86−90.
(87) Marsh, J. A.; Hernandez, H.; Hall, Z.; Ahnert, S. E.; Perica, T.; Robinson, C. V.; Teichmann, S. A. Protein Complexes Are under Evolutionary Selection to Assemble Via Ordered Pathways. *Cell* **2013**, *153*, 461−470.
(88) Morilla, I.; Lees, J. G.; Reid, A. J.; Orengo, C.; Ranea, J. A. Assessment of Protein Domain Fusions in Human Protein Interaction Networks Prediction: Application to the Human Kinetochore Model. *New Biotechnol.* **2010**, *27*, 755−765.
(89) Dandekar, T.; Snel, B.; Huynen, M.; Bork, P. Conservation of Gene Order: A Fingerprint of Proteins That Physically Interact. *Trends Biochem. Sci.* **1998**, *23*, 324−328.
(90) Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguez, P.; Bork, P.; von Mering, C.; Jensen, L. J. String V9.1: Protein-Protein Interaction Networks, with Increased Coverage and Integration. *Nucleic Acids Res.* **2013**, *41*, D808−815.
(91) Muley, V. Y.; Ranjan, A. Evaluation of Physical and Functional Protein-Protein Interaction Prediction Methods for Detecting Biological Pathways. *PLoS One* **2013**, *8*, e54325.
(92) Folador, E. L.; Hassan, S. S.; Lemke, N.; Barh, D.; Silva, A.; Ferreira, R. S.; Azevedo, V. An Improved Interolog Mapping-Based Computational Prediction of Protein-Protein Interactions with Increased Network Coverage. *Integr. Biol. (Camb.)* **2014**, *6*, 1080−1087.
(93) Sahu, S. S.; Weirick, T.; Kaundal, R. Predicting Genome-Scale Arabidopsis-Pseudomonas Syringae Interactome Using Domain and Interolog-Based Approaches. *BMC Bioinf.* **2014**, *15* (Suppl 11), S13.
(94) Krishnadev, O.; Srinivasan, N. Prediction of Protein-Protein Interactions between Human Host and a Pathogen and Its Application to Three Pathogenic Bacteria. *Int. J. Biol. Macromol.* **2011**, *48*, 613−619.
(95) Garcia-Garcia, J.; Schleker, S.; Klein-Seetharaman, J.; Oliva, B. Bips: Biana Interolog Prediction Server. A Tool for Protein-Protein Interaction Inference. *Nucleic Acids Res.* **2012**, *40*, W147−151.
(96) Pellegrini, M.; Marcotte, E. M.; Thompson, M. J.; Eisenberg, D.; Yeates, T. O. Assigning Protein Functions by Comparative Genome

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

Analysis: Protein Phylogenetic Profiles. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 4285−4288.
(97) Luo, Q.; Pagel, P.; Vilne, B.; Frishman, D. Dima 3.0: Domain Interaction Map. *Nucleic Acids Res.* **2011**, *39*, D724−729.
(98) Pazos, F.; Valencia, A. Similarity of Phylogenetic Trees as Indicator of Protein-Protein Interaction. *Protein Eng., Des. Sel.* **2001**, *14*, 609−614.
(99) Lua, R. C.; Marciano, D. C.; Katsonis, P.; Adikesavan, A. K.; Wilkins, A. D.; Lichtarge, O. Prediction and Redesign of Protein-Protein Interactions. *Prog. Biophys. Mol. Biol.* **2014**, *116*, 194.
(100) Ge, H.; Liu, Z.; Church, G. M.; Vidal, M. Correlation between Transcriptome and Interactome Mapping Data from Saccharomyces Cerevisiae. *Nat. Genet.* **2001**, *29*, 482−486.
(101) Jansen, R.; Greenbaum, D.; Gerstein, M. Relating Whole-Genome Expression Data with Protein-Protein Interactions. *Genome Res.* **2002**, *12*, 37−46.
(102) Przulj, N.; Wigle, D. A.; Jurisica, I. Functional Topology in a Network of Protein Interactions. *Bioinformatics* **2004**, *20*, 340−348.
(103) Chua, H. N.; Sung, W. K.; Wong, L. Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions. *Bioinformatics* **2006**, *22*, 1623−1630.
(104) Goldberg, D. S.; Roth, F. P. Assessing Experimentally Derived Interactions in a Small World. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 4372−4376.
(105) Phan, H. T.; Sternberg, M. J. Pinalog: A Novel Approach to Align Protein Interaction Networks–Implications for Complex Detection and Function Prediction. *Bioinformatics* **2012**, *28*, 1239−1245.
(106) Pazos, F.; Valencia, A. In Silico Two-Hybrid System for the Selection of Physically Interacting Protein Pairs. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 219−227.
(107) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of Direct Residue Contacts in Protein-Protein Interaction by Message Passing. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 67−72.
(108) Hopf, T. A., Scharfe, C. P., Rodrigues, J. P., Green, A. G., Kohlbacher, O., Sander, C., Bonvin, A. M., Marks, D. S. Sequence Co-Evolution Gives 3d Contacts and Structures of Protein Complexes, *eLife* **2014** , 3.10.7554/eLife.03430
(109) Qin, S.; Zhou, H. X. Meta-Ppisp: A Meta Web Server for Protein-Protein Interaction Site Prediction. *Bioinformatics* **2007**, *23*, 3386−3387.
(110) Vakser, I. A. Protein-Protein Docking: From Interaction to Interactome. *Biophys. J.* **2014**, *107*, 1785−1793.
(111) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein-Protein Docking Benchmark Version 4.0. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 3111−3114.
(112) Vreven, T.; Moal, I. H.; Vangone, A.; Pierce, B. G.; Kastritis, P. L.; Torchala, M.; Chaleil, R.; Jimenez-Garcia, B.; Bates, P. A.; Fernandez-Recio, J.; Bonvin, A. M.; Weng, Z. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* **2015**, *427*, 3031−3041.
(113) Janin, J. Protein-Protein Docking Tested in Blind Predictions: The Capri Experiment. *Mol. BioSyst.* **2010**, *6*, 2351−2362.
(114) Lensink, M. F.; Wodak, S. J. Docking, Scoring, and Affinity Prediction in Capri. *Proteins: Struct., Funct., Genet.* **2013**, *81*, 2082−2095.
(115) Pierce, B. G.; Wiehe, K.; Hwang, H.; Kim, B. H.; Vreven, T.; Weng, Z. Zdock Server: Interactive Docking Prediction of Protein-Protein Complexes and Symmetric Multimers. *Bioinformatics* **2014**, *30*, 1771−1773.
(116) Chaudhury, S.; Berrondo, M.; Weitzner, B. D.; Muthu, P.; Bergman, H.; Gray, J. J. Benchmarking and Analysis of Protein Docking Performance in Rosetta V3.2. *PLoS One* **2011**, *6*, e22477.
(117) Mashiach, E.; Nussinov, R.; Wolfson, H. J. Fiberdock: Flexible Induced-Fit Backbone Refinement in Molecular Docking. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1503−1519.
(118) Tuncbag, N.; Gursoy, A.; Keskin, O. Prediction of Protein-Protein Interactions: Unifying Evolution and Structure at Protein Interfaces. *Phys. Biol.* **2011**, *8*, 035006.
(119) Tuncbag, N.; Gursoy, A.; Nussinov, R.; Keskin, O. Predicting Protein-Protein Interactions on a Proteome Scale by Matching Evolutionary and Structural Similarities at Interfaces Using Prism. *Nat. Protoc.* **2011**, *6*, 1341−1354.
(120) Kundrotas, P. J.; Zhu, Z.; Janin, J.; Vakser, I. A. Templates Are Available to Model Nearly All Complexes of Structurally Characterized Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 9438−9441.
(121) Szilagyi, A.; Zhang, Y. Template-Based Structure Modeling of Protein-Protein Interactions. *Curr. Opin. Struct. Biol.* **2014**, *24*, 10−23.
(122) Gunther, S.; May, P.; Hoppe, A.; Frommel, C.; Preissner, R. Docking without Docking: Isearch–Prediction of Interactions Using Known Interfaces. *Proteins: Struct., Funct., Genet.* **2007**, *69*, 839−844.
(123) Hosur, R.; Xu, J.; Bienkowska, J.; Berger, B. Iwrap: An Interface Threading Approach with Application to Prediction of Cancer-Related Protein-Protein Interactions. *J. Mol. Biol.* **2011**, *405*, 1295−1310.
(124) Zhang, Q. C.; Petrey, D.; Garzon, J. I.; Deng, L.; Honig, B. Preppi: A Structure-Informed Database of Protein-Protein Interactions. *Nucleic Acids Res.* **2013**, *41*, D828−833.
(125) Planas-Iglesias, J.; Marin-Lopez, M. A.; Bonet, J.; Garcia-Garcia, J.; Oliva, B. Iloops: A Protein-Protein Interaction Prediction Server Based on Structural Features. *Bioinformatics* **2013**, *29*, 2360−2362.
(126) Ghoorah, A. W.; Devignes, M. D.; Smail-Tabbone, M.; Ritchie, D. W. Kbdock 2013: A Spatial Classification of 3d Protein Domain Family Interactions. *Nucleic Acids Res.* **2014**, *42*, D389−395.
(127) Minhas, F.; Geiss, B. J.; Ben-Hur, A. Pairpred: Partner-Specific Prediction of Interacting Residues from Sequence and Structure. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 1142−1155.
(128) Tuncbag, N.; Keskin, O.; Nussinov, R.; Gursoy, A. Fast and Accurate Modeling of Protein-Protein Interactions by Combining Template-Interface-Based Docking with Flexible Refinement. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 1239−1249.
(129) Vreven, T.; Hwang, H.; Pierce, B. G.; Weng, Z. Evaluating Template-Based and Template-Free Protein-Protein Complex Structure Prediction. *Briefings Bioinf.* **2014**, *15*, 169−176.
(130) Gao, Y.; Douguet, D.; Tovchigrechko, A.; Vakser, I. A. Dockground System of Databases for Protein Recognition Studies: Unbound Structures for Docking. *Proteins: Struct., Funct., Genet.* **2007**, *69*, 845−851.
(131) Orchard, S.; Hermjakob, H. Data Standardization by the Hupo-Psi: How Has the Community Benefitted? *Methods Mol. Biol.* **2011**, *696*, 149−160.
(132) Orchard, S.; Kerrien, S.; Abbani, S.; Aranda, B.; Bhate, J.; Bidwell, S.; Bridge, A.; Briganti, L.; Brinkman, F. S.; Cesareni, G.; et al. Protein Interaction Data Curation: The International Molecular Exchange (Imex) Consortium. *Nat. Methods* **2012**, *9*, 345−350.
(133) Salwinski, L.; Miller, C. S.; Smith, A. J.; Pettit, F. K.; Bowie, J. U.; Eisenberg, D. The Database of Interacting Proteins: 2004 Update. *Nucleic Acids Res.* **2004**, *32*, 449D−451.
(134) Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N. H.; Chavali, G.; Chen, C.; del-Toro, N.; et al. The Mintact Project–Intact as a Common Curation Platform for 11 Molecular Interaction Databases. *Nucleic Acids Res.* **2014**, *42*, D358−363.
(135) Licata, L.; Briganti, L.; Peluso, D.; Perfetto, L.; Iannuccelli, M.; Galeota, E.; Sacco, F.; Palma, A.; Nardozza, A. P.; Santonico, E.; Castagnoli, L.; Cesareni, G. Mint, the Molecular Interaction Database: 2012 Update. *Nucleic Acids Res.* **2012**, *40*, D857−861.
(136) Chatr-Aryamontri, A.; Breitkreutz, B. J.; Oughtred, R.; Boucher, L.; Heinicke, S.; Chen, D.; Stark, C.; Breitkreutz, A.; Kolas, N.; O'Donnell, L.; Reguly, T.; Nixon, J.; Ramage, L.; Winter, A.; Sellam, A.; Chang, C.; Hirschman, J.; Theesfeld, C.; Rust, J.; Livstone, M. S.; Dolinski, K.; Tyers, M. The Biogrid Interaction Database: 2015 Update. *Nucleic Acids Res.* **2015**, *43*, D470.
(137) Keshava Prasad, T. S.; Goel, R.; Kandasamy, K.; Keerthikumar, S.; Kumar, S.; Mathivanan, S.; Telikicherla, D.; Raju, R.; Shafreen, B.;

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

Venugopal, A.; et al. Human Protein Reference Database–2009 Update. *Nucleic Acids Res.* **2009**, *37*, D767−772.
(138) Turner, B.; Razick, S.; Turinsky, A. L.; Vlasblom, J.; Crowdy, E. K.; Cho, E.; Morrison, K.; Donaldson, I. M.; Wodak, S. J. Irefweb: Interactive Analysis of Consolidated Protein Interaction Data and Their Supporting Evidence. *Database* **2010**, *2010*, baq023.
(139) Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; et al. String V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.* **2015**, *43*, D447.
(140) Gehlenborg, N.; O'Donoghue, S. I.; Baliga, N. S.; Goesmann, A.; Hibbs, M. A.; Kitano, H.; Kohlbacher, O.; Neuweger, H.; Schneider, R.; Tenenbaum, D.; Gavin, A. C. Visualization of Omics Data for Systems Biology. *Nat. Methods* **2010**, *7*, S56−68.
(141) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498−2504.
(142) Nepomnyachiy, S.; Ben-Tal, N.; Kolodny, R. Cytostruct: Augmenting the Network Visualization of Cytoscape with the Power of Molecular Viewers. *Structure* **2015**, *23*, 941−948.
(143) Huang, S. S.; Fraenkel, E. Integrating Proteomic, Transcriptional, and Interactome Data Reveals Hidden Components of Signaling and Regulatory Networks. *Sci. Signaling* **2009**, *2*, ra40.
(144) Tuncbag, N.; Braunstein, A.; Pagnani, A.; Huang, S. S.; Chayes, J.; Borgs, C.; Zecchina, R.; Fraenkel, E. Simultaneous Reconstruction of Multiple Signaling Pathways Via the Prize-Collecting Steiner Forest Problem. *J. Comput. Biol.* **2013**, *20*, 124−136.
(145) Tuncbag, N.; McCallum, S.; Huang, S. S.; Fraenkel, E. Steinernet: A Web Server for Integrating 'Omic' Data to Discover Hidden Components of Response Pathways. *Nucleic Acids Res.* **2012**, *40*, W505−509.
(146) Yeger-Lotem, E.; Riva, L.; Su, L. J.; Gitler, A. D.; Cashikar, A. G.; King, O. D.; Auluck, P. K.; Geddie, M. L.; Valastyan, J. S.; Karger, D. R.; Lindquist, S.; Fraenkel, E. Bridging High-Throughput Genetic and Transcriptional Data Reveals Cellular Responses to Alpha-Synuclein Toxicity. *Nat. Genet.* **2009**, *41*, 316−323.
(147) Xu, Q.; Canutescu, A. A.; Wang, G.; Shapovalov, M.; Obradovic, Z.; Dunbrack, R. L., Jr. Statistical Analysis of Interface Similarity in Crystals of Homologous Proteins. *J. Mol. Biol.* **2008**, *381*, 487−507.
(148) Higurashi, M.; Ishida, T.; Kinoshita, K. Pisite: A Database of Protein Interaction Sites Using Multiple Binding States in the Pdb. *Nucleic Acids Res.* **2009**, *37*, D360−364.
(149) Krissinel, E.; Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **2007**, *372*, 774−797.
(150) Gong, S.; Yoon, G.; Jang, I.; Bolser, D.; Dafas, P.; Schroeder, M.; Choi, H.; Cho, Y.; Han, K.; Lee, S.; et al. Psibase: A Database of Protein Structural Interactome Map (Psimap). *Bioinformatics* **2005**, *21*, 2541−2543.
(151) Basse, M. J.; Betzi, S.; Bourgeas, R.; Bouzidi, S.; Chetrit, B.; Hamon, V.; Morelli, X.; Roche, P. 2p2idb: A Structural Database Dedicated to Orthosteric Modulation of Protein-Protein Interactions. *Nucleic Acids Res.* **2013**, *41*, D824−827.
(152) Cukuroglu, E.; Gursoy, A.; Nussinov, R.; Keskin, O. Non-Redundant Unique Interface Structures as Templates for Modeling Protein Interactions. *PLoS One* **2014**, *9*, e86738.
(153) de Beer, T. A.; Berka, K.; Thornton, J. M.; Laskowski, R. A. Pdbsum Additions. *Nucleic Acids Res.* **2014**, *42*, D292−296.
(154) Laskowski, R. A.; Hutchinson, E. G.; Michie, A. D.; Wallace, A. C.; Jones, M. L.; Thornton, J. M. Pdbsum: A Web-Based Database of Summaries and Analyses of All Pdb Structures. *Trends Biochem. Sci.* **1997**, *22*, 488−490.
(155) Mosca, R.; Ceol, A.; Stein, A.; Olivella, R.; Aloy, P. 3did: A Catalog of Domain-Based Interactions of Known Three-Dimensional Structure. *Nucleic Acids Res.* **2014**, *42*, D374−379.
(156) Finn, R. D.; Miller, B. L.; Clements, J.; Bateman, A. Ipfam: A Database of Protein Family and Domain Interactions Found in the Protein Data Bank. *Nucleic Acids Res.* **2014**, *42*, D364−373.
(157) Shoemaker, B. A.; Zhang, D.; Tyagi, M.; Thangudu, R. R.; Fong, J. H.; Marchler-Bauer, A.; Bryant, S. H.; Madej, T.; Panchenko, A. R. Ibis (Inferred Biomolecular Interaction Server) Reports, Predicts and Integrates Multiple Types of Conserved Interactions for Proteins. *Nucleic Acids Res.* **2012**, *40*, D834−840.
(158) Winter, C.; Henschel, A.; Kim, W. K.; Schroeder, M. Scoppi: A Structural Classification of Protein-Protein Interfaces. *Nucleic Acids Res.* **2006**, *34*, D310−D314.
(159) Teyra, J.; Paszkowski-Rogacz, M.; Anders, G.; Pisabarro, M. T. Scowlp Classification: Structural Comparison and Analysis of Protein Binding Regions. *BMC Bioinf.* **2008**, *9*, 9.
(160) Thorn, K. S.; Bogan, A. A. Asedb: A Database of Alanine Mutations and Their Effects on the Free Energy of Binding in Protein Interactions. *Bioinformatics* **2001**, *17*, 284−285.
(161) Fischer, T. B.; Arunachalam, K. V.; Bailey, D.; Mangual, V.; Bakhru, S.; Russo, R.; Huang, D.; Paczkowski, M.; Lalchandani, V.; Ramachandra, C.; Ellison, B.; Galer, S.; Shapley, J.; Fuentes, E.; Tsai, J. The Binding Interface Database (Bid): A Compilation of Amino Acid Hot Spots in Protein Interfaces. *Bioinformatics* **2003**, *19*, 1453−1454.
(162) Ji, Z. L.; Chen, X.; Zhen, C. J.; Yao, L. X.; Han, L. Y.; Yeo, W. K.; Chung, P. C.; Puy, H. S.; Tay, Y. T.; Muhammad, A.; Chen, Y. Z. Kdbi: Kinetic Data of Bio-Molecular Interactions Database. *Nucleic Acids Res.* **2003**, *31*, 255−257.
(163) Moal, I. H.; Fernandez-Recio, J. Skempi: A Structural Kinetic and Energetic Database of Mutant Protein Interactions and Its Use in Empirical Models. *Bioinformatics* **2012**, *28*, 2600−2607.
(164) Kastritis, P. L.; Moal, I. H.; Hwang, H.; Weng, Z.; Bates, P. A.; Bonvin, A. M.; Janin, J. A Structure-Based Benchmark for Protein-Protein Binding Affinity. *Protein Sci.* **2011**, *20*, 482−491.
(165) Kumar, M. D.; Gromiha, M. M. Pint: Protein-Protein Interactions Thermodynamic Database. *Nucleic Acids Res.* **2006**, *34*, D195−D198.
(166) Guney, E.; Tuncbag, N.; Keskin, O.; Gursoy, A. Hotsprint: Database of Computational Hot Spots in Protein Interfaces. *Nucleic Acids Res.* **2008**, *36*, D662−D666.
(167) Cukuroglu, E.; Gursoy, A.; Keskin, O. Hotregion: A Database of Predicted Hot Spot Clusters. *Nucleic Acids Res.* **2012**, *40*, D829−833.
(168) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The Foldx Web Server: An Online Force Field. *Nucleic Acids Res.* **2005**, *33*, W382−388.
(169) Kim, D. E.; Chivian, D.; Baker, D. Protein Structure Prediction and Analysis Using the Robetta Server. *Nucleic Acids Res.* **2004**, *32*, W526−531.
(170) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The Ftmap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. *Nat. Protoc.* **2015**, *10*, 733−755.
(171) Assi, S. A.; Tanaka, T.; Rabbitts, T. H.; Fernandez-Fuentes, N. Pcrpi: Presaging Critical Residues in Protein Interfaces, a New Computational Tool to Chart Hot Spots in Protein Interfaces. *Nucleic Acids Res.* **2010**, *38*, e86.
(172) Deng, L.; Zhang, Q. C.; Chen, Z.; Meng, Y.; Guan, J.; Zhou, S. Predhs: A Web Server for Predicting Protein-Protein Interaction Hot Spots by Using Structural Neighborhood Properties. *Nucleic Acids Res.* **2014**, *42*, W290−295.
(173) Meireles, L. M.; Domling, A. S.; Camacho, C. J. Anchor: A Web Server and Database for Analysis of Protein-Protein Interaction Binding Pockets for Drug Discovery. *Nucleic Acids Res.* **2010**, *38*, W407−411.
(174) Kruger, D. M.; Gohlke, H. Drugscoreppi Webserver: Fast and Accurate in Silico Alanine Scanning for Scoring Protein-Protein Interactions. *Nucleic Acids Res.* **2010**, *38*, W480−486.
(175) Ashkenazy, H.; Erez, E.; Martz, E.; Pupko, T.; Ben-Tal, N. Consurf 2010: Calculating Evolutionary Conservation in Sequence and Structure of Proteins and Nucleic Acids. *Nucleic Acids Res.* **2010**, *38*, W529−533.
(176) Valdar, W. S. Scoring Residue Conservation. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 227−241.

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909

(177) Zhu, H.; Sommer, I.; Lengauer, T.; Domingues, F. S. Alignment of Non-Covalent Interactions at Protein-Protein Interfaces. *PLoS One* **2008**, *3*, e1926.
(178) Cheng, S.; Zhang, Y.; Brooks, C. L., 3rd Pcalign: A Method to Quantify Physicochemical Similarity of Protein-Protein Interfaces. *BMC Bioinf.* **2015**, *16*, 33.
(179) Gao, M.; Skolnick, J. Ialign: A Method for the Structural Comparison of Protein-Protein Interfaces. *Bioinformatics* **2010**, *26*, 2259−2265.
(180) Shulman-Peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. J. Multibind and Mappis: Webservers for Multiple Alignment of Protein 3d-Binding Sites and Their Interactions. *Nucleic Acids Res.* **2008**, *36*, W260−264.
(181) Bernauer, J.; Bahadur, R. P.; Rodier, F.; Janin, J.; Poupon, A. Dimovo: A Voronoi Tessellation-Based Method for Discriminating Crystallographic and Biological Protein-Protein Interactions. *Bioinformatics* **2008**, *24*, 652−658.
(182) Zhu, H.; Domingues, F. S.; Sommer, I.; Lengauer, T. Noxclass: Prediction of Protein-Protein Interaction Types. *BMC Bioinf.* **2006**, *7*, 27.
(183) Soner, S.; Ozbek, P.; Garzon, J. I.; Ben-Tal, N.; Haliloglu, T. Dynaface: Discrimination between Obligatory and Non-Obligatory Protein-Protein Interactions Based on the Complex's Dynamics. *PLoS Comput. Biol.* **2015**, *11*, e1004461.
(184) Negi, S. S.; Schein, C. H.; Oezguen, N.; Power, T. D.; Braun, W. Interprosurf: A Web Server for Predicting Interacting Sites on Protein Surfaces. *Bioinformatics* **2007**, *23*, 3397−3399.
(185) Shazman, S.; Celniker, G.; Haber, O.; Glaser, F.; Mandel-Gutfreund, Y. Patch Finder Plus (Pfplus): A Web Server for Extracting and Displaying Positive Electrostatic Patches on Protein Surfaces. *Nucleic Acids Res.* **2007**, *35*, W526−530.
(186) Neuvirth, H.; Raz, R.; Schreiber, G. Promate: A Structure Based Prediction Program to Identify the Location of Protein-Protein Binding Sites. *J. Mol. Biol.* **2004**, *338*, 181−199.
(187) Zhang, Q. C.; Deng, L.; Fisher, M.; Guan, J.; Honig, B.; Petrey, D. Predus: A Web Server for Predicting Protein Interfaces Using Structural Neighbors. *Nucleic Acids Res.* **2011**, *39*, W283−287.
(188) Murakami, Y.; Jones, S. Sharp2: Protein-Protein Interaction Predictions Using Patch Analysis. *Bioinformatics* **2006**, *22*, 1794−1795.
(189) de Vries, S. J.; van Dijk, A. D.; Bonvin, A. M. Whiscy: What Information Does Surface Conservation Yield? Application to Data-Driven Docking. *Proteins: Struct., Funct., Genet.* **2006**, *63*, 479−489.
(190) Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J. Cluspro: An Automated Docking and Discrimination Method for the Prediction of Protein Complexes. *Bioinformatics* **2004**, *20*, 45−50.
(191) Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. Patchdock and Symmdock: Servers for Rigid and Symmetric Docking. *Nucleic Acids Res.* **2005**, *33*, W363−367.
(192) de Vries, S. J.; van Dijk, M.; Bonvin, A. M. The Haddock Web Server for Data-Driven Biomolecular Docking. *Nat. Protoc.* **2010**, *5*, 883−897.
(193) Torchala, M.; Moal, I. H.; Chaleil, R. A.; Fernandez-Recio, J.; Bates, P. A. Swarmdock: A Server for Flexible Protein-Protein Docking. *Bioinformatics* **2013**, *29*, 807−809.
(194) Lyskov, S.; Gray, J. J. The Rosettadock Server for Local Protein-Protein Docking. *Nucleic Acids Res.* **2008**, *36*, W233−238.
(195) Andrusier, N.; Nussinov, R.; Wolfson, H. J. Firedock: Fast Interaction Refinement in Molecular Docking. *Proteins: Struct., Funct., Genet.* **2007**, *69*, 139−159.
(196) Pierce, B.; Weng, Z. Zrank: Reranking Protein Docking Predictions with an Optimized Energy Function. *Proteins: Struct., Funct., Genet.* **2007**, *67*, 1078−1086.
(197) Cheng, T. M.; Blundell, T. L.; Fernandez-Recio, J. Pydock: Electrostatics and Desolvation for Effective Scoring of Rigid-Body Protein-Protein Docking. *Proteins: Struct., Funct., Genet.* **2007**, *68*, 503−515.
(198) Singh, R.; Park, D.; Xu, J.; Hosur, R.; Berger, B. Struct2net: A Web Service to Predict Protein-Protein Interactions Using a Structure-Based Approach. *Nucleic Acids Res.* **2010**, *38*, W508−515.
(199) Meyer, M. J.; Das, J.; Wang, X.; Yu, H. Instruct: A Database of High-Quality 3d Structurally Resolved Protein Interactome Networks. *Bioinformatics* **2013**, *29*, 1577−1579.
(200) Mukherjee, S.; Zhang, Y. Protein-Protein Complex Structure Predictions by Multimeric Threading and Template Recombination. *Structure* **2011**, *19*, 955−966.
(201) Fukuhara, N.; Kawabata, T. Homcos: A Server to Predict Interacting Protein Pairs and Interacting Sites by Homology Modeling of Complex Structures. *Nucleic Acids Res.* **2008**, *36*, W185−189.
(202) Lu, L.; Lu, H.; Skolnick, J. Multiprospector: An Algorithm for the Prediction of Protein-Protein Interactions by Multimeric Threading. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 350−364.
(203) Chen, H.; Skolnick, J. M-Tasser: An Algorithm for Protein Quaternary Structure Prediction. *Biophys. J.* **2008**, *94*, 918−928.
(204) Pagel, P.; Kovac, S.; Oesterheld, M.; Brauner, B.; Dunger-Kaltenbach, I.; Frishman, G.; Montrone, C.; Mark, P.; Stumpflen, V.; Mewes, H. W.; Ruepp, A.; Frishman, D. The Mips Mammalian Protein-Protein Interaction Database. *Bioinformatics* **2005**, *21*, 832−834.
(205) Ruepp, A.; Waegele, B.; Lechner, M.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; Mewes, H. W. Corum: The Comprehensive Resource of Mammalian Protein Complexes–2009. *Nucleic Acids Res.* **2010**, *38*, D497−501.
(206) Lage, K.; Karlberg, E. O.; Storling, Z. M.; Olason, P. I.; Pedersen, A. G.; Rigina, O.; Hinsby, A. M.; Tumer, Z.; Pociot, F.; Tommerup, N.; Moreau, Y.; Brunak, S. A Human Phenome-Interactome Network of Protein Complexes Implicated in Genetic Disorders. *Nat. Biotechnol.* **2007**, *25*, 309−316.
(207) Jayapandian, M.; Chapman, A.; Tarcea, V. G.; Yu, C.; Elkiss, A.; Ianni, A.; Liu, B.; Nandi, A.; Santos, C.; Andrews, P.; Athey, B.; States, D.; Jagadish, H. V. Michigan Molecular Interactions (Mimi): Putting the Jigsaw Puzzle Together. *Nucleic Acids Res.* **2007**, *35*, D566−571.
(208) Schaefer, M. H.; Fontaine, J. F.; Vinayagam, A.; Porras, P.; Wanker, E. E.; Andrade-Navarro, M. A. Hippie: Integrating Protein Interaction Networks with Experiment Based Quality Scores. *PLoS One* **2012**, *7*, e31826.
(209) Patil, A.; Nakai, K.; Nakamura, H. Hitpredict: A Database of Quality Assessed Protein-Protein Interactions in Nine Species. *Nucleic Acids Res.* **2011**, *39*, D744−749.
(210) Vastrik, I.; D'Eustachio, P.; Schmidt, E.; Gopinath, G.; Croft, D.; de Bono, B.; Gillespie, M.; Jassal, B.; Lewis, S.; Matthews, L.; Wu, G.; Birney, E.; Stein, L. Reactome: A Knowledge Base of Biologic Pathways and Processes. *Genome Biol.* **2007**, *8*, R39.
(211) Kamburov, A.; Pentchev, K.; Galicka, H.; Wierling, C.; Lehrach, H.; Herwig, R. Consensuspathdb: Toward a More Complete Picture of Cell Biology. *Nucleic Acids Res.* **2011**, *39*, D712−717.

DOI: 10.1021/acs.chemrev.5b00683
Chem. Rev. 2016, 116, 4884−4909