ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Protein–Protein Interaction Methods and Protein Phase Separation

Castrense Savojardo,[1,*] Pier Luigi Martelli,[1,*] and Rita Casadio[1,2]

[1]Biocomputing Group, Department of Pharmacy and Biotechnology and Interdepartmental Center "Luigi Galvani" for Integrated Studies of Bioinformatics, Biophysics, and Biocomplexity, University of Bologna, 40126 Bologna, Italy; email: rita.casadio@unibo.it

[2]Institute of Biomembranes, Bioenergetics, and Molecular Biotechnologies (IBIOM), Italian National Research Council (CNR), 70126 Bari, Italy

Annu. Rev. Biomed. Data Sci. 2020. 3:89–112

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-011720-104428

*These authors equally contributed to this article

Keywords

protein–protein interaction, protein interacting surfaces, protein phase separation, machine learning, deep learning, computational methods, Cajal body, PPI prediction, PPI networks, PPI database, intrinsically disordered proteins

Abstract

In the last decade, newly developed experimental methods have made it possible to highlight that macromolecules in the cell milieu physically interact to support physiology. This has shifted the problem of protein–protein interaction from a microscopic, electron-density scale to a mesoscopic one. Further, nowadays there is increasing evidence that proteins in the nucleus and in the cytoplasm can aggregate in membraneless organelles for different physiological reasons. In this scenario, it is urgent to face the problem of biomolecule functional annotation with efficient computational methods, suited to extract knowledge from reliable data and transfer information across different domains of investigation. Here, we revise the present state of the art of our knowledge of protein–protein interaction and the computational methods that differently implement it. Furthermore, we explore experimental and computational features of a set of proteins involved in phase separation.

1. INTRODUCTION

In cells, biochemical processes take place in heterogeneous and crowded environments that influence the efficiency of the reactivity and density distribution of participating macromolecules in biological processes and metabolic pathways—the ensemble of billions of reactions in a living organism. In eukaryotes, dozens of canonical membrane-enclosed compartments increase the complexity of the cell-internal structure by sequestering molecules and confining specific processes within the subvolumes of specialized organelles, whose roles are fundamental for the overall cell physiology (1). Some other organelles are present in the cell milieu and interestingly they are not membrane delimited. They can be regarded as open macromolecular assemblies, held together by weak macromolecular interactions, and whose structures and functions are presently under close investigation. Examples in the nuclear environment include the nucleolus, the subnuclear Cajal inclusion bodies, and the nuclear speckles, while the cytoplasmic environment includes the stress granules, the P-bodies, and the germ granules (2).

Recently, with the development of advanced tools of molecular biology, microscopy, and liquid phase separation (3, 4), it has become possible to highlight some dynamical aspects of the open macromolecular assemblies and discover that membraneless organelles are common to several types of cells working under physiological conditions (5, 6). Open organelles concentrate bio-macromolecules such as RNAs, DNAs, proteins, and other small molecules. In mammalian neurons, the subnuclear Cajal inclusion bodies contain protein and RNA components that associate in a cell cycle–dependent fashion or under specific metabolic and stress conditions (see Reference 7 and references therein).

Although the characterization of the membraneless compartments and the mechanisms of their formation has just begun, exciting results support the notion that condensation mechanisms are driven by collective protein–protein interactions (PPIs) and protein–nucleic acid interactions, in dynamic equilibria with the surroundings (2, 4) (**Figure 1**).

According to these findings, the microsized intracellular structures seem to play fundamental roles in controlling and regulating different biological and metabolic processes in different cell types. Increasing evidence supports the notion that PPI is one of the key mechanisms giving rise to membraneless organelles. Growing interest is mounting in a new field in cell biology named protein phase separation (8–10), which may indeed link microscopic characteristics to mesoscopic structural and functional characteristics of the cell milieu by tuning the spatiotemporal fine regulation of the functional encounters that lead to biological processes and metabolic pathways (9–12).

Unsolved problems still deserve attention: What drives phase transition and ultimately specific PPIs? If phase transition can be described as a nucleation phenomenon, what triggers nucleation?

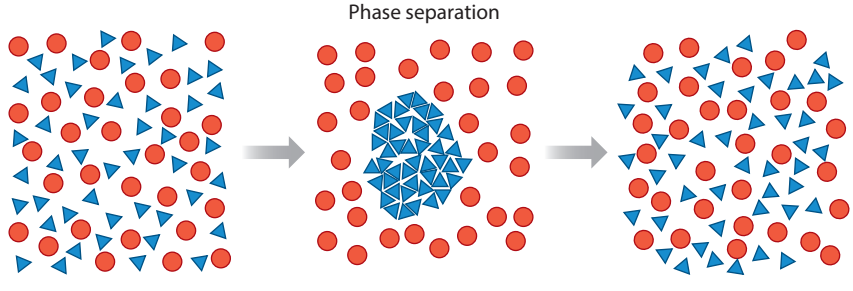


Figure 1

Simplified graphical representation of the dynamics of protein phase separation.

Do membraneless organelles dynamically associate for efficient metabolic activities? Or do they aggregate only under pathological conditions?

For decades, we have tried to uncover patterns or properties at the basis of PPI, with the ultimate goal of understanding functional mechanisms of protein homo- and heterocomplexes, and later of complex interactomes. Nowadays, in the likelihood that liquid phase condensation is an important mechanism of cell physiology and disease, it is even more urgent to understand which proteins can undergo the single to droplet phase transition for describing and modeling the emergent properties of the complex cell interior (**Figure 1**).

In the following we review our present sources of information for PPIs and briefly describe the never-ending process of generating algorithms capable of extracting information from valuable data with the aim of transferring knowledge by computing properties of never-seen-before examples (13). We also try to disclose, with computational approaches, properties common to proteins known to take part in granule assembly.

2. PROTEIN PHASE SEPARATION

Undoubtedly, what we know about PPIs and nucleic acid–protein interactions derives so far from in vitro experiments and less frequently from in vivo experiments. This is due to the technical difficulty in monitoring molecular dynamics in situ, which is to say in the specific cell environments, whose composition and complexity may vary in different cell types. Recent papers indeed deal with the problem of extrapolating in vitro data on macromolecular interactions to in vivo environments, where each protein can be regarded as one of the several partners taking part in functional interactions (4, 14). Thermodynamics and kinetics describe some typical conformational equilibria in which the environment effect can be taken into account. The real problem here is how to model environmental effects (the solvent effect, but also more specific effects due to other macromolecules in the specific environment). We are very familiar with pairwise interactions: Two molecules can attract each other either in a nonspecific or in a site-specific manner, originating transient and stable complexes, respectively. In both cases, the interaction derives from one or more elementary physical forces, including electrostatic attraction and hydrogen-bonding and solvent-mediated interactions. This classification relies on the detection in vitro of the Gibbs free energy change of complex stability, which is collected for a fraction of putative complexes in specific databases, like SKEMPI (Structural Kinetic and Energetic Database of Mutant Protein Interactions) (15) (**https://life.bsc.es/pid/skempi2/database/summary**) and PDBbind (16) (**http://www.pdbbind.org.cn/**).

However, in phase behavior, like in protein phase separation, besides pairwise interactions, three and higher multibody interactions can occur to mediate membraneless organelle formation. For example, in dilute to moderate particle volume fractions, when the pair interaction potentials are embedded within a thermodynamic perturbation theory approach, it is possible to simulate the phase behavior of nanoparticles suspended in a solution of oppositely charged polyelectrolytes. Results suggest that the fluid-like phase is metastable in such systems and that the aggregation and cluster morphologies are mediated by particle charge (17, 18). The model was later expanded to describe the interaction of protein-charged patches with polyelectrolyte complexes (19), supporting the notion that aggregates depend on the extent of charged patches. Moreover, it has been recently suggested that multivalence of adhesive domains or their inherent flexibility/disorder may lead to protein phase separation in cells, with many different properties ranging from solid-like to liquid-like assemblies (10). The basic idea is that proteins promoting phase transitions should in principle be endowed with different and flexible interaction patches to be able to interact in a multibody manner with the environment.

It is therefore urgent to revise our knowledge of PPI, keeping in mind that our modeling can also help the annotation of proteins involved in phase transition.

3. STATISTICAL ANALYSIS OF PROTEIN–PROTEIN INTERFACES

3.1. A Dataset of Functional Protein–Protein Complexes from Protein Data Bank

The Protein Data Bank (PDB) represents the main source of data concerning atomic-level information about protein–protein complexes. As of July 2019, the PDB contains about 67,000 protein–protein complexes (excluding monomeric entries and protein–DNA and protein–RNA complexes).

From the PDB, we selected a set of functional protein–protein complexes whose structures were determined by X-ray crystallography with a resolution <3.0 Å. Membrane proteins were excluded from this dataset. Moreover, we filtered out some PDB entries that our scripts could not automatically process because of inconsistencies or errors in the corresponding bio-assembly file. When multiple PDB entries reported the same protein–protein complex, we retained only a representative structure, selecting the one with the highest X-ray resolution (R-factor).

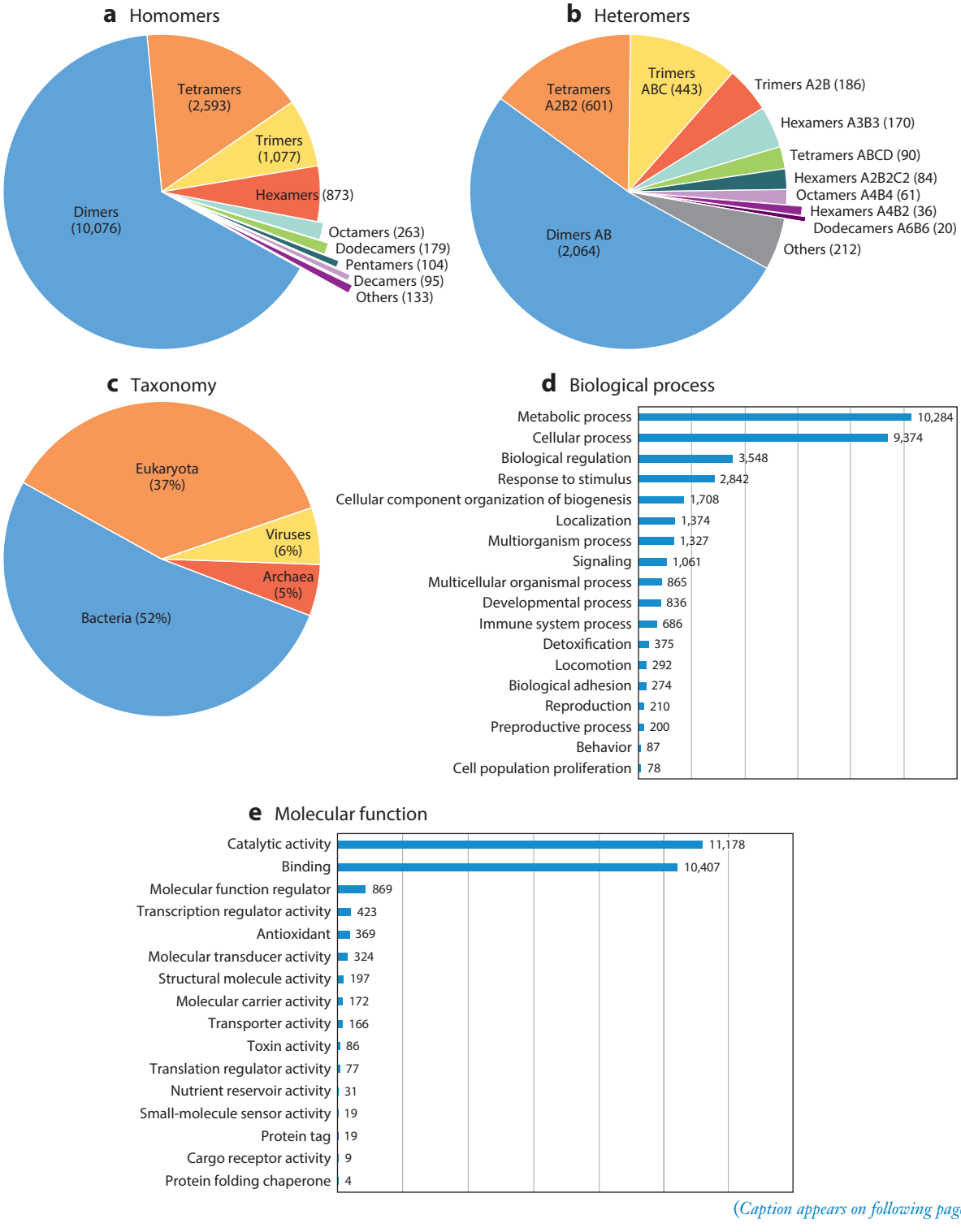
We ended up with a dataset comprising 19,360 functional complexes, 15,393 and 3,967 out of which are homomeric and heteromeric complexes, respectively. Overall, the dataset includes 60,347 PDB chains, which reduce to 24,294 when collapsing identical chains into a single entity. We used this dataset for characterizing the main features of protein–protein interfaces.

Figure 2*a*,*b* shows the distributions of quaternary structures for homomeric (**Figure 2*a***) and heteromeric (**Figure 2*b***) complexes in our dataset. As the pie chart in **Figure 2*a*** shows, 65% of all homocomplexes are dimers, followed by tetramers (17%), trimers (7%), and hexamers (6%). For heterocomplexes (**Figure 2*b***), heterodimers with global stoichiometry AB account for 52% of the entire dataset, followed by heterotetramers of the type A2B2 (15%) and heterotrimers of the ABC (11%) and A2B (5%) types. Taxonomic classification of all PDB entries in the dataset is reported in **Figure 2*c***. Half of the entries are from bacterial organisms (52%), while eukaryotic proteins account for the 37% of the dataset. Only small fractions are from viruses and Archaea (6% and 5%, respectively). **Figure 2*d*,*e*** shows the distribution of the Gene Ontology first-level annotations of the biological processes (**Figure 2*d***) and molecular functions (**Figure 2*e***) of the proteins in the dataset. Most protein complexes in the database are involved in metabolic processes or cellular processes (**Figure 2*d***) and about 58% are enzymes (**Figure 2*e***).

3.2. Definition of Protein–Protein Interfaces

Residues involved in protein–protein interfaces are all located on the surface of monomers participating in the formation of a protein complex. As a preliminary step for extracting protein–protein interface residues, monomer surfaces are identified by retaining residues with a relative solvent accessibility higher than 20% (20).

Once the monomer surface has been computed, two alternative definitions of interface residues can be applied. The first is based on the different solvent accessibility between the complex and the monomer: Interface residues are identified as the surface residues undergoing a reduction of accessible area upon complex formation. A second definition is based on the computation of inter-residue distances: Interface residues for a given monomer are those having at least one residue in another subunit at a distance below a predefined threshold. The most common distance thresholds are between 5 and 8 Å (21). Many studies have shown that the choice of the definition only slightly affects the interface composition as well as the performance of interface predictors (22, 23).



(*Caption appears on following page*)

Figure 2 (*Figure appears on preceding page*)

Statistics of a selected dataset of 15,393 homomeric and 3,967 heteromeric high-resolution representative complexes extracted from the Protein Data Bank. Panels *a* and *b* show the distribution of quaternary structures for homomeric (*a*) and heteromeric (*b*) complexes. Panel *c* shows the taxonomic distribution of the 19,360 functional protein–protein complexes in the dataset. Panels *d* and *e* show the distribution of first-level Gene Ontology biological process (*d*) and molecular function (*e*) annotations of the 19,360 protein–protein complexes included in the dataset.

For the dataset described here we adopted the first definition based on accessibility difference. Identified interfaces were classified into two separate classes: homointerfaces participating in homomeric interactions and heterointerfaces participating in heteromeric interactions. Overall, the dataset comprises 5,796,600 residues, 3,242,166 of which are located in the monomer surfaces. There are 1,041,790 interface residues in total: 928,650 are part of homointerfaces while 113,140 are in heterointerfaces.

Figure 3 shows statistics on the surface area overall involved in homo- and heterointerfaces per monomer. The areas were computed at the protein level, summing the values of absolute accessibility derived from the software DSSP (Define Secondary Structure of Proteins) (24) for each of the 24,294 nonredundant monomeric chains in the dataset. **Figure 3** shows that areas involved in homomeric interactions are in general slightly larger than those of heteromeric ones: Average areas of homo- and heterointerfaces are about 3,946 Å^2 and 3,551 Å^2, respectively. Previous studies described similar findings (see Reference 25, and references therein).

3.3. Main Features of Interaction Sites

Computational methods developed in past years for the prediction of protein–protein interfaces have adopted a large spectrum of features in an attempt to capture key characteristics of interface residues with respect to other residues in the protein surface.

A widespread feature adopted by many approaches is the residue interface propensity (26–30), defined as the ratio between the frequencies of a given residue type *R* in the interface and the

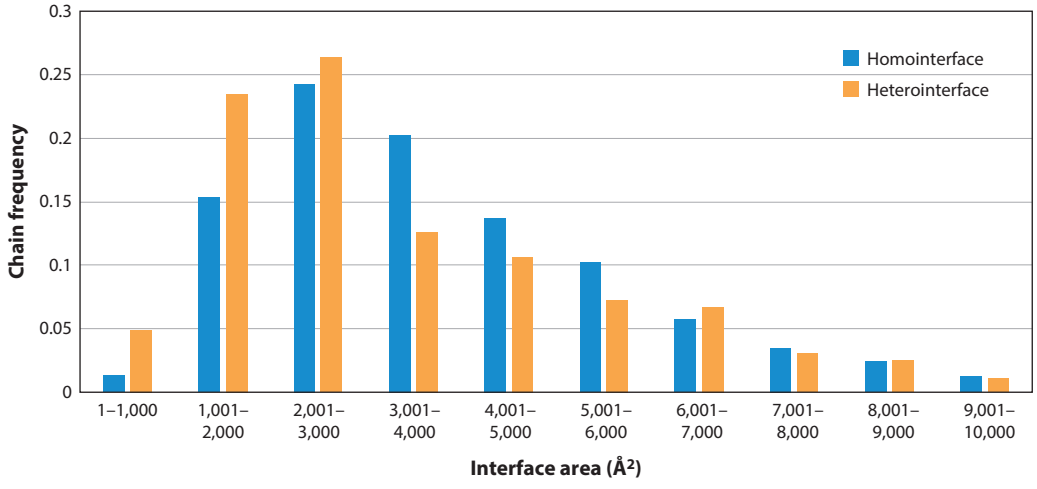


Figure 3

Statistics of surface areas (Å^2) involved in homomeric and heteromeric interactions for the 24,294 monomers included in our dataset. Areas were obtained by summing, for each monomer, accessibility values of interacting residues as computed with the software DSSP.

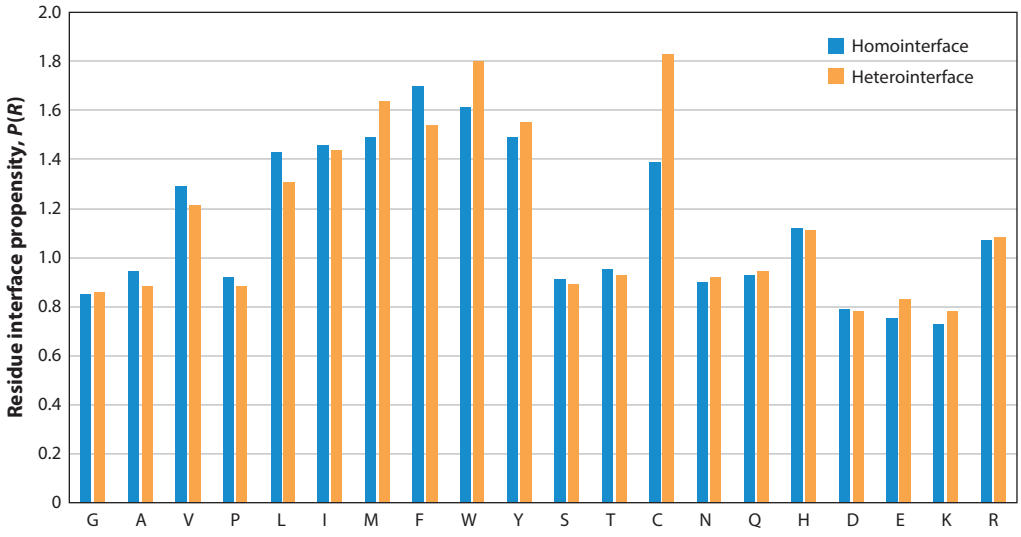


Figure 4

Residue interface propensity for homo- and heterointerfaces. Propensities were computed as the ratios of frequencies of residues in the interface and the surface (Equation 1).

surface. The most general equation is

$$P(R) = \frac{f_I(R)}{f_S(R)}, \qquad 1.$$

where $f_I(R)$ and $f_S(R)$ are the frequencies of residue R in the interface and the surface, respectively. A value of $P(R) > 1$ means that there is a propensity of residue R to be part of the protein interface. In **Figure 4** we show the distributions of residue interface propensity scores for each residue type computed separately for homo- and heterointerfaces in our PDB-derived dataset.

As a general trend, no major differences are observed between homo- and heterointerfaces. In both cases, residues D, E, and K are rarely found as part of protein interfaces, while residues L, I, M, F, W, Y, C, and R show a marked propensity to mediate PPIs. Cysteine and, to a lesser extent, tryptophan are more prone to form interfaces in heterocomplexes. Phenylalanine and leucine are conversely more abundant in homointerfaces.

Evolutionary conservation is another important feature that can be exploited in general to identify functionally important residues (31) and that was adopted to recognize PPI sites (26, 29, 30, 32–36). Conservation can be estimated by means of many different algorithms, including multiple sequence alignment (MSA) computed for a target protein using programs like PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) (37) or HHBlits (38). Shannon's entropy and its variants (31, 39) are routinely adopted to score positional conservation for each column of the MSA.

Figure 5 shows the distribution of the conservation scores for interface and surface residues in our PDB-derived dataset. In this analysis, we only used a selection of chains from our dataset obtained by reducing internal redundancy to 30% pairwise sequence identity. MSAs were generated by running PSI-BLAST for each sequence (three iterations with the *e*-value threshold set to 0.001) against the UniRef90 sequence database (40). Normalized Shannon's entropy was used to

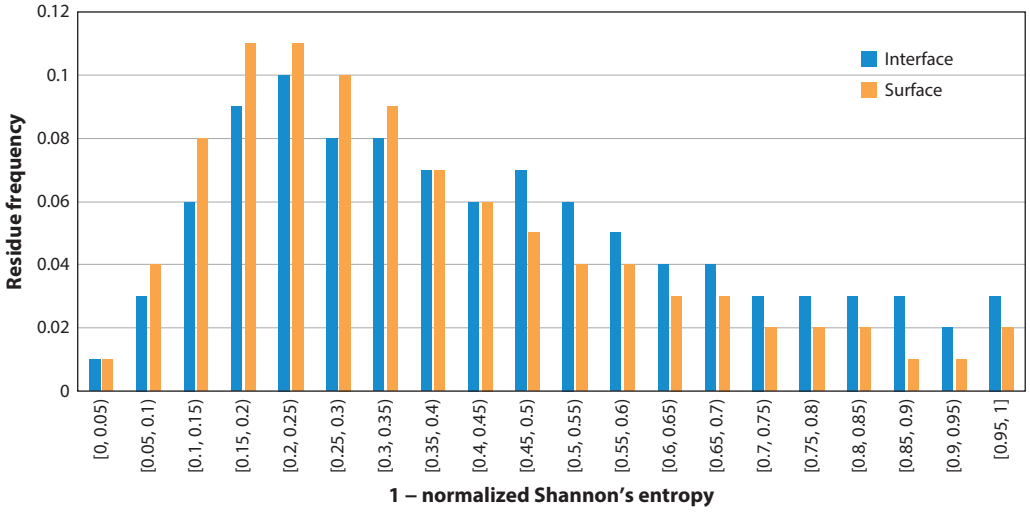


Figure 5

Distribution of evolutionary conservation scores, 1 − normalized Shannon's entropy, of interface and surface residues. Conservation scores were computed on a redundancy-reduced dataset at 30% pairwise sequence identity comprising 9,301 chains from our dataset.

score conservation for each position i in the MSA as follows:

$$NS(i) = \frac{-1}{\log 20} \sum_{j=1}^{20} P_{i,j} \times \log P_{i,j}, \qquad 2.$$

where $P_{i,j}$ is the frequency of residue type j in the i-th column of the MSA. The final conservation score is computed as

$$S(i) = 1 - NS(i). \qquad 3.$$

Results shown in **Figure 5** indicate that the interface residues tend to be slightly more conserved than the other surface residues, as already reported by other studies (41). In general, conservation alone is not sufficient to accurately discriminate interface residues from the remaining surface residues (32, 41).

4. PREDICTORS OF PPI SITES

Given the importance of understanding principles of PPIs for unraveling biological complexity, the past two decades have seen a rapid increase in methods and tools for the analysis and prediction of PPI sites.

The computational problem of PPI site prediction can be defined as follows: Given an input protein, one wishes to identify residues that are involved in interactions with other proteins. Different settings are possible. Firstly, prediction can be performed starting either from three-dimensional (3D) protein structure or directly from the protein covalent structure (primary sequence). Secondly, the identification of PPI sites can be done either in a partner-unspecific or in a partner-specific way. Partner-unspecific methods take as input the protein in its monomeric form and identify PPI sites without prior knowledge of the specific interaction partner. As an

alternative, partner-specific methods predict paired PPI sites on two input proteins known (or assumed) to physically interact (42, 43).

As in many other areas of computational biology, machine learning plays a major role in the prediction of PPI sites. About two-thirds of the tools available in the literature are completely or partially based on some machine learning algorithm leveraging the vast amount of information available in public databases, such as the PDB. However, other approaches have been described, including methods based on simple statistical inference from features (28, 29, 32, 44), scoring functions (27), and homology transfer from templates (45–48).

In the next sections we review the available tools in literature. For the sake of clarity, methods are separated into two main categories according to the required input: methods performing predictions on unbound monomeric structures and methods attempting to predict interaction sites directly on protein sequences.

4.1. Predictors of PPI Sites on Protein Structures

The development of tools to perform prediction of PPI sites starting from protein structure dates back to the pioneering work in the field by Jones & Thornton (49). In the last two decades, many tools have been released based on many different techniques. A nonexhaustive list of structure-based tools is reported in **Table 1**. The general workflow of machine learning–based approaches for predicting PPI sites on protein structures is outlined in **Figure 6**.

The success of structure-based prediction builds on top of the fact that structural features have a significant discriminative power for PPI site prediction. Features adopted by structure-based methods include different metrics for evaluating residue solvent exposure, e.g., relative solvent accessibility or geometrical features like protrusion and depth indexes, computation of surface curvature, residue electrostatic potentials, atomic B factors, and secondary structure (26, 30, 33, 36, 49–55).

Many structure-based methods are not limited to structural features but also incorporate sequence-based descriptors like residue composition, physical-chemical attributes, interface propensities, and evolutionary information in the form of sequence profiles or position-specific scoring matrices (26, 30, 36, 52–54, 56). However, in contrast to pure sequence-based approaches, protein structural information allows researchers to fully exploit the structural neighborhood of a given residue. In this way, all features (either structure-based or sequence-based) can be aggregated by averaging over spatial nearest neighbors, and this improves prediction performance in many cases (26, 27, 30, 32, 33, 48, 50, 51, 55).

Individual features are not sufficient to unambiguously identify PPI sites on protein structures. For this reason, tools adopt different techniques, routinely based on machine learning, to perform feature combination and improve discrimination power.

Machine learning frameworks adopted for PPI site prediction from structure include binary classifiers such as support vector machines (SVMs) (26, 30, 46, 50, 53, 57), neural networks (NNs) (33, 56), random forests (RFs) (36, 42, 54), and probabilistic graphical models like conditional random fields (CRFs) (30, 51, 52, 55). CRFs are the discriminative counterpart of hidden Markov models and they are well suited for modeling sequential data like protein sequences, as they can capture the potential relationships among adjacent residues. In the context of PPI site prediction, CRFs have been used to explicitly model the relationships along the sequence of residues in the protein surface (30, 51, 52, 55). In some cases, different machine learning approaches are combined together to exploit the complementary advantages of each technique (30, 53).

Proper selection of training data is crucial for the definition of accurate and unbiased machine learning–based approaches. Datasets for training and testing PPI site predictors are mainly

Table 1 List of structure-based methods for PPI site prediction

Name	Year	Partner specificity	Method details	URL	Reference(s)
BIPSPI (structure)	2018	Yes	XGBoost, tree boosting	**http://bipspi.cnb.csic.es/xgbPredApp/**	42
SVM plus 3D Zernike descriptors	2018	No	SVM plus 3D Zernike descriptors	Web server not available	57
ISPRED4	2017	No	SVM plus CRF	**https://ispred4.biocomp.unibo.it**	30, 56
INSPiRE	2017	No	Information transfer from knowledge base of amino acid structural neighborhoods	Web server not available	48
Dong et al.	2014	No	Pairwise CRF	Web server not available	55
PAIRpred	2014	Yes	SVM	Source code available at **http://combi.cs.colostate.edu/supplements/pairpred/**	43
Li et al.	2012	No	RF plus feature selection (mRMR plus IFS)	Web server not available	36
PrISE	2012	No	Information transfer using local surface structural similarity	**http://ailab-projects2.ist.psu.edu/prise/index.py**	47
PresCont	2012	No	Residue properties	**https://bioinf.ur.de/prescont.php**	29
PredUS	2011	No	Information transfer by structure similarity plus SVM refinement	**https://honiglab.c2b2.columbia.edu/PredUs/index_omega.html**	46
Liu et al.	2009	No	HMSVM	Web server not available	52
Sikić et al.	2009	No	RF	Web server not available	54
CRFSite	2007	No	CRF	Web server not available	51
SPPIDER	2007	No	SVM, NN	**http://sppider.cchmc.org/**	53
PINUP	2006	No	Empirical scoring function	**http://sysbio.unl.edu/services/PINUP/**	27
WHISCY	2006	No	Conservation scores and structural features	**http://milou.science.uu.nl/services/WHISCY/**	28
cons-PPISP	2005	No	Consensus NN	**http://pipe.scs.fsu.edu/ppisp.html**	33
PredPPI	2005	No	SVM	Web server not available	26
Koike & Takagi	2004	No	SVM	Web server not available	50
ProMate	2004	No	Interface score computed from combination of structural and conservation features	Web server not available	32
Jones & Thornton	1997	No	Patch analysis	Web server not available	49

Abbreviations: CRF, conditional random field; HMSVM, hidden Markov support vector machine; IFS, incremental feature selection; mRMR, maximum relevance minimum redundancy; NN, neural network; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

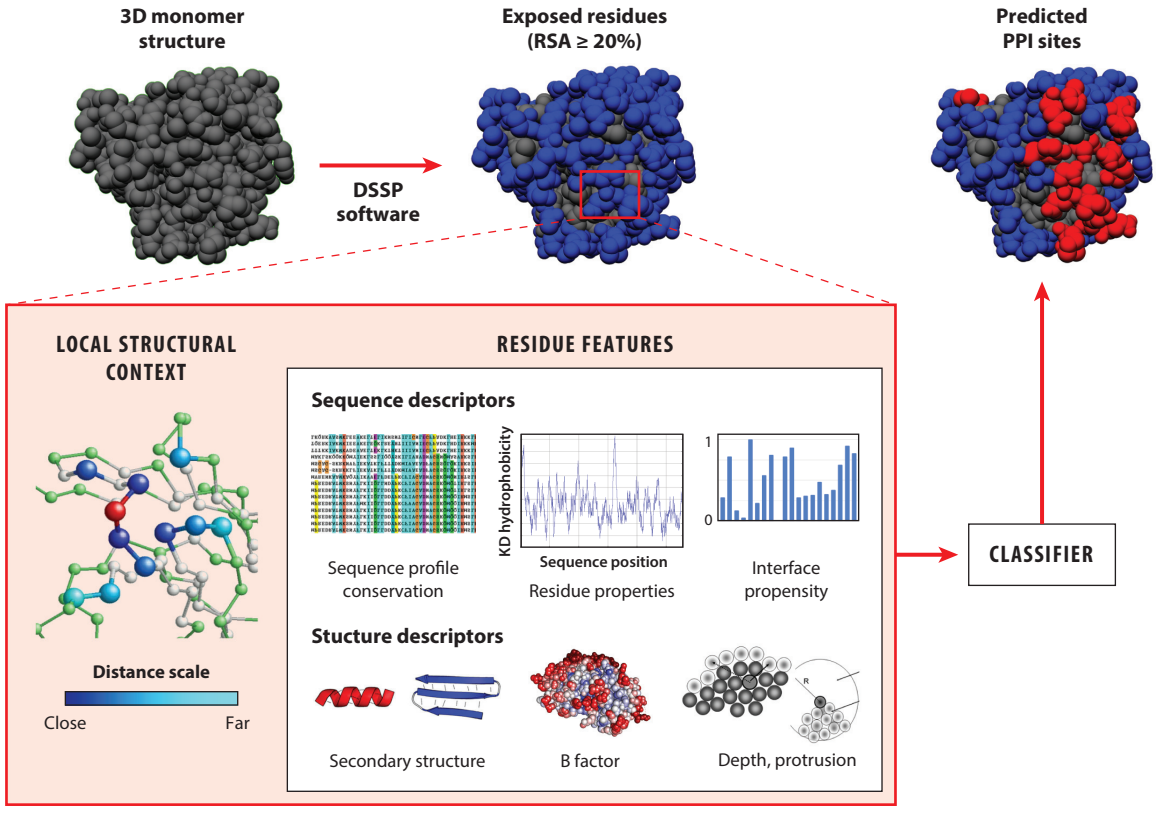


Figure 6

The general workflow of machine learning–based classifiers for predicting PPI sites from protein structure. Abbreviation: KD, Kyte–Doolittle; RSA, relative solvent accessibility.

derived from protein–protein complexes that are deposited in the PDB. In order to improve data quality, researchers often rely on derived datasets specifically designed to score methods in the context of PPIs such as the Docking Benchmark dataset (58) and data derived from the Critical Assessment of Predicted Interactions (CAPRI) (59). Raw data extracted from the PDB need to be properly filtered to exclude nonfunctional complexes and enforce constraints on experimental method, resolution, sequence redundancy, and overall quality of the structure (e.g., filtering out short chains or structures with missing residues).

One key issue when dealing with structural features is the availability of the structures of both the complex and the monomeric conformations of each subunit. Indeed, complex formation always determines a conformational change at the level of protein–protein interfaces. If features are computed on atomic coordinates of subunits extracted from the complex structure, this may easily lead to biases in the prediction, since conformational changes may introduce fingerprints for PPI sites. In order to train unbiased predictors, researchers should use monomeric structures for structural feature extraction, and the complex should be used only to identify experimental PPI sites (30). Many benchmark datasets such as the Docking Benchmark (58) explicitly also report, for each complex, the PDB entries of unbound subunits involved in the interactions.

Comparative evaluation of different structural tools for PPI site prediction is challenging since different methods use different datasets for training and testing. To date, state-of-the-art

structural PPI site predictors report Matthews correlation coefficients (MCCs) in the range of 0.4–0.5 (60, 61).

4.2. Predictors of PPI Sites from Protein Sequence

Despite their efficacy, structure-based methods have the limitation of being applicable only when the protein structure is available. To date, the vast majority of proteins, even those involved in several disease-related pathways, are known only at the level of primary sequences. In this respect, the ability to accurately identify PPI sites directly on protein sequences is of prominent importance since it opens the possibility of screening any translation product. Nevertheless, the prediction of PPI sites from sequence alone is much more challenging and difficult. In machine learning, the efficacy of the method strongly depends on the amount of input information and by definition a protein sequence contains less information than a protein structure. For this reason, the field of sequence-based PPI prediction has observed a relatively low rate of expansion compared to structure-based methods. **Table 2** reports an updated list of sequence-based tools available.

Besides the unavailability of highly informative structural features, another issue that hampers the performance of sequence-based predictors is related to the unbalanced nature of the underlying classification problem: Proteins contain few PPI sites compared to the total number of residues in the sequence. When protein structure is available, this problem can be mitigated (but not completely eliminated) by restricting the search of PPI sites to residues on the protein surface. When only the sequence is known, this filtering procedure cannot be applied, resulting in very unbalanced datasets that are difficult to process using standard classification techniques (62).

Sequence features adopted so far to develop prediction methods include evolutionary information, conservation scores, and physical-chemical properties of residues (e.g., hydrophobicity or conformational propensities). Some methods include structural features computed from the protein sequence with specific classifiers, such as predicted solvent accessibility and secondary structure, in an attempt to fill the gap with structure-based approaches (63–65). Several sequence-based PPI predictors have been developed in the past years that leverage these features and use different machine learning algorithms, including NNs (34, 64), SVMs (43, 63, 66–68), RFs (42, 64), naïve Bayes (35), and logistic regression (65, 69).

A subcategory of sequence-based methods comprises approaches based on analysis of protein coevolution. These approaches are all partner specific and identify pairs of interacting residues between two input partners. The assumption at the basis of these tools is that residue positions that are detected as covarying across protein–protein interfaces are in physical contact in the protein complex (70–73). Recent comparative benchmarks performed on state-of-the-art sequence-based predictors report MCC scores in the range of 0.13–0.23, much lower than achieved with structure-based methods (65, 74, 75).

5. DATABASES OF BINDING AFFINITIES AND THEIR PREDICTION

PPI thermodynamics is determined by measuring binding affinities and other thermodynamic parameters (76). Two updated databases are freely available: PDBbind (16; **http://www.pdbbind-cn.org/**) and SKEMPI (15; **https://life.bsc.es/pid/skempi2**). PDBbind is a collection of experimentally measured binding affinity data for different types of biomolecular complexes deposited in PDB; SKEMPI version 2.0 lists, for each PDB protein–protein complex, binding affinities in case of variations, the associated kinetic rate constants, and the entropy and enthalpy change

Table 2 List of sequence-based methods for PPI site prediction

Name	Year	Partner specificity	Method details	URL	Reference(s)
SCRIBER	2019	No	Multilevel logistic regression	**http://biomine.cs.vcu.edu/servers/SCRIBER/**	65
SeRenDIP	2019	No	RF	**http://www.ibi.vu.nl/programs/serendipwww/**	64, 75
BIPSPI (sequence)	2018	Yes	XGBoost, tree boosting	**http://bipspi.cnb.csic.es/xgbPredApp/**	42
SSWRF	2016	No	RF plus SVM	**http://202.119.84.36:3079/SSWRF-PPI/SSWRF-PPI.html**	63
EVComplex	2014	Yes	Direct coupling analysis based on mean field approximation	**https://evcouplings.org/complex**	73
GREMLIN	2014	Yes	Direct coupling analysis based on maximization of pseudo-likelihoods	**http://gremlin.bakerlab.org/cplx_submit.php**	71
LORIS	2014	No	*L*1 logistic regression	Web server not available	69
PAIRpred	2014	Yes	SVM	Source code available at **http://combi.cs.colostate.edu/supplements/pairpred/**	43
NPS-HomPPI	2011	No	Transfer by sequence similarity (partner-unspecific)	**http://ailab-projects2.ist.psu.edu/NPSHOMPPI/**	45
PS-HomPPI	2011	Yes	Transfer by sequence similarity (partner-specific)	**http://ailab-projects2.ist.psu.edu/PSHOMPPIv1.3/**	45
Chen & Li	2010	No	SOM plus SVM	Web server not available	66
PSIVER	2010	No	Naïve Bayes classifier with kernel density estimation	**https://mizuguchilab.org/PSIVER/**	35
Weigt et al.	2009	Yes	Residue coevolution in MSAs by direct-coupling analysis based on message-passing algorithms	Web server not available	72
ISIS	2007	No	NN	Web server not available	34
Res et al.	2005	No	SVM	Web server not available	68
Yan et al.	2004	No	SVM plus Bayesian classifier	Web server not available	67
Gallet et al.	2000	No	Hydrophobicity distribution along the sequence	Web server not available	44
Pazos et al.	1997	Yes	Correlated mutations	Web server not available	70

Abbreviations: MSA, multiple sequence alignment; NN, neural network; RF, random forest; SOM, self-organizing map; SVM, support vector machine; XGBoost, extreme gradient boosting.

values, when available. Thermodynamic data are adopted as training sets for several computational methods, which predict binding affinities with different strategies based on 3D structures of complexes or amino acid sequences or based on changes in binding affinities of the complexes upon variations. These methods, which include docking methods, have been recently reviewed (77) and are benchmarked during CAPRI (59) experiments.

6. DETECTION OF PPIs AT THE PROTEOME SCALE

Data on PPIs can be collected at a mesoscopic level, which allows all the putative direct and undirect interactions occurring in the cell milieu to be represented at large.

In the last three decades, different techniques have been developed for high-throughput experimental assays of PPIs taking place in cells or in cell compartments at a given moment of the cell's life (78, 79). These techniques have allowed the collection of different sets of species-specific interactions, routinely represented as networks, where nodes and links represent proteins and detected interactions, respectively. No single technique is nowadays sufficient to capture all the possible interactions between the proteins expressed in a cell. This is mainly due to the wide range of affinities characterizing the interactions of different protein pairs and to the dynamic nature of protein–protein networks that exhibit a large adaptability to environmental conditions and external signals (80, 81). Available experimental techniques for high-throughput detection of PPIs can be divided into two major classes: methods for detecting binary interactions between proteins and methods for detecting proteins that are part of the same complex (82).

Methods in the first class implement different strategies for assaying the interaction between two specific proteins, routinely indicated as bait and prey. The prototypical technique is the yeast two-hybrid screening (Y2H), originally developed by Fields & Song in 1989 (83). Y2H is performed in vivo, mainly in yeast cells, but similar procedures have been also set in mammalian cells (84). Bait and prey proteins (or protein fragments) are fused to two different domains of the same transcription factor. The interaction between bait and prey in the nucleus of the yeast cell reconstitutes the transcription factor and therefore promotes the expression of a reporter gene. The procedure can be automatized by building large libraries of baits and preys and it allows for proteome-scale assays for binary interactions (85). However, several problems limit the sensitivity of Y2H. Among them are (*a*) the necessity that both bait and prey localize in the nucleus, (*b*) the possibility that fused domains hide interaction sites, and (*c*) the differences in protein conformation or states induced by the expression system, including posttranslational modifications and binding to cofactors (78). Moreover, false-positive detection can arise from nonspecific interactions of overexpressed proteins.

As alternatives to binary interaction assays, methods in the second class are suited to detect the proteins that are part of a complex captured and purified from a cell lysate. The most effective capture techniques include affinity purification (AP) and cofractionation (Co-frac). Independently of the capture technique, proteins of the complex are routinely recognized with mass spectrometry (MS) assays.

AP techniques capture complexes by means of an immobilized antibody directed toward a specific bait, recognized from a specific native epitope or a tag fused to its N terminus (86). Alternatively, cofractionation techniques capture complexes by means of different chemico-physical separation procedures that are fully independent of the definition of baits (87). Sensitivity of these methods is mainly limited by the amount of the complex in the sample. Moreover, experimental settings (e.g., elution buffer) can have a strong effect in selecting interactions on the basis of their strength.

6.1. Experimental Human Interactome

Triggered by advances in high-throughput techniques, large efforts have been made to define species-specific interactomes, comprising all the possible interactions between proteins expressed by an organism, to be used as a reference for understanding genotype–phenotype relationships. Recently, the updated versions of two human experimental interactomes have been released: (*a*) the binary interactome map HuRI (human reference interactome), obtained with three Y2H assays

Table 3 Size of two experimental human interactomes recently released

	HuRI[a]	BioPlex2.0[b]
Number of proteins[c]	8,470	10,844
Number of interactions	51,907	55,498

[a]Database downloaded from **http://interactome.baderlab.org** (88) in August 2019.
[b]Database downloaded from **https://bioplex.hms.harvard.edu** (89) in August 2019.
[c]Splicing isoforms have been collapsed.

on a library spanning 17,408 human protein–coding genes (88), and (*b*) BioPlex 2.0 [biophysical interactions of ORF(open reading frame)eome-derived complexes], deriving from the analysis of AP-MS experiments after the capture of complexes nucleated on 6,251 different protein baits (89).

Table 3 reports the dimensions of the two networks, after collapsing splicing variants and deleting interactions with proteins marked as UNKNOWN. The size of the HuRI and BioPlex2.0 networks are comparable, covering respectively 36% and 46% of the 23,413 protein-coding genes of the GRCh38.p12 human genome assembly (primary and alternative assembly). Also, the numbers of detected interactions are similar.

However, when compared, the two networks show a very small overlap (**Table 4**). The two databases share 4,827 proteins, accounting for 56.9% and 44.5% of HuRI and BioPlex2.0, respectively. Restricting the dataset to the links between shared proteins, the interaction numbers decrease to 16,133 for HuRI and 12,610 for BioPlex2.0. Noticeably, the number of shared interactions is very low: The overlap between the two large networks reduces to 829 links involving some 1,123 proteins.

These results confirm similar analyses performed on previous versions of the human interactomes (82) and agree with trends already observed when analyzing interactomes of different species (80, 90). Several causes contribute to the poor overlap among different interactomes. These may include (*a*) the different approaches taken by Y2H and AP-MS experiments, which aim to detect binary interactions and associations in the same complex, respectively; (*b*) error rates of the experimental procedures, which are only partially mitigated by the control experiments; (*c*) the techniques' biases in preferentially detecting interactions of a particular type (e.g., interactions of a particular strength or taking place in a particular compartment); (*d*) levels of protein expression in the different expression systems; and (*e*) different posttranslational processing in different expression systems.

Owing to the complexity of PPIs networks and the complementarity of experimental methods, it is not possible to rely on a single technique for compiling a comprehensive interactome, and integrative approaches must be adopted.

6.2. Comprehensive Databases of PPIs

Different databases have been implemented over the years for collecting PPIs from different sources, including high- and low-throughput experiments and literature-curated mining (91, 92).

Table 4 Overlap between two experimental human interactomes recently released

	HuRI[a]	BioPlex2.0[b]
Number of shared proteins	4,827 (56.9%)	4,827 (44.5%)
Number of interactions among shared proteins	16,133 (31.1%)	12,610 (22.7%)
Number of shared interactions	829 (5.1%)	829 (6.6%)

[a]Database downloaded from **http://interactome.baderlab.org** (88) on July 27, 2019.
[b]Database downloaded from **https://bioplex.hms.harvard.edu** (89) on July 27, 2019.

Table 5 Comprehensive databases of PPIs

	Species	Proteins	Interactions	Human proteins[a]	Interactions in human
IntAct[b]	1,529	113,357	593,007	21,795	288,419
BioGRID[c]	69	71,266	677,442	21,624	384,954

[a]Splicing isoforms have been collapsed.
[b]Database downloaded from **https://www.ebi.ac.uk/intact/** on July 27, 2019.
[c]Database downloaded from **https://downloads.thebiogrid.org** on July 27, 2019. Only physical interactions have been considered (version 3.5.175).

A complete review of available databases is beyond the scope of this review, and here we consider only two widely adopted resources (**Table 5**). IntAct (93) is the centralized database, located at EMBL-EBI (European Bioinformatics Institute of the European Molecular Biology Laboratory), collecting data from all the resources of the IMEx (International Molecular Exchange) consortium, an international collaboration of curation efforts for PPIs. It collects data for more than 1,500 different organisms. BioGRID (Biological General Repository for Interaction Datasets; 94) is a public database that archives genetic and protein interaction data from 69 model organisms, including humans. **Table 5** compares the content of the two databases, considering in the case of BioGRID only physical interactions. When restricting to humans, it is evident that the number of proteins reported in databases largely surpasses the coverage achieved by single experimental procedures. However, merging information from different sources is not a trivial process, and curators have to devise criteria optimizing the trade-off between comprehensiveness and accuracy of detected interactions.

While the ratio between the number of represented proteins and the number of coding genes can be used to estimate the completeness in the proteome space, it is very hard to assess the completeness in the interaction space, since the total number of interactions taking place in a cell is still unknown. For humans this number is estimated to range between 650,000 (95) and 1,000,000 when interactions mediating posttranslational modifications are taken into account (96). It is therefore evident that a part of the interactome still sits in the shade. Different computational procedures are available for inferring new putative PPIs.

7. COMPUTATIONAL METHODS FOR INFERRING NETWORKS OF PPIs

Several strategies have been adopted to complement the available experimental interactomes by leveraging the available databases of PPIs to infer novel interacting protein pairs. The problem can be generally posed as the prediction of the edges connecting a set of proteins known at the sequence or structure levels. When structures are known or can be derived with modeling procedures, or at least when their structural domains can be mapped onto the sequence, a prediction can be performed by searching for similar structures or domains in the set of complexes reported in the PDB. On this basis, structurally annotated interactomes have been implemented (97, 98). Although leading to highly reliable predictions, 3D knowledge can be exploited only for a limited subset of proteins. In the absence of structural information, different data can be integrated to perform the prediction. Among them are (*a*) phylogenetic profiles, which infer the interaction from the concomitant presence (or absence) of orthologous protein pairs in different genomes (99); (*b*) the detection in different genomes of rearrangement events that lead to two or more proteins fusing into a unique multidomain protein (100); (*c*) the detection of correlated expression levels of different proteins in different conditions (101); and (*d*) the transfer of information from orthologous sequences (102). Versions of these methods are incorporated in the pipeline at the basis of the comprehensive STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)

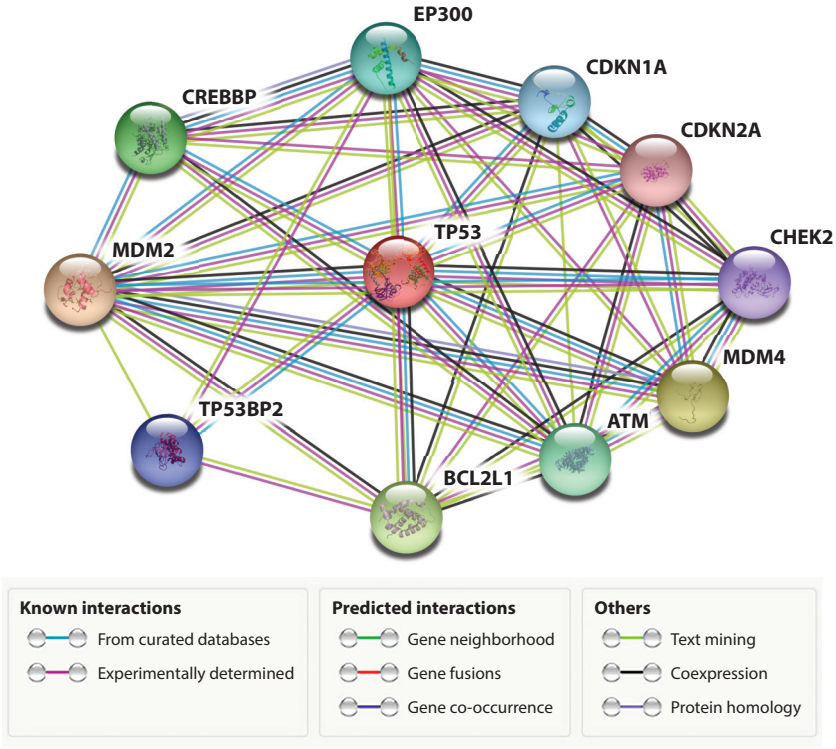


Figure 7

The human TP53 interaction network extracted from the human interactome included in the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database (**https://string-db.org/**). Only the first shell of interactors of TP53 is considered. Edges represent protein–protein associations and the meanings of their colors are indicated in the figure key. The availability of the protein 3D structure is marked by filling nodes with 3D representations.

database of physical and functional interactions to complement data deriving from low- and high-throughput experiments and from automatic literature mining (103). **Figure 7** shows an example of a PPI network available in STRING: the TP53 interaction network.

Prediction of interacting protein pairs recently took advantage of the development of computational frameworks for efficiently extracting partial correlation coefficients from covariance matrices deriving from large multiple sequence alignments. This technique has been applied to the refinement of existing interactomes (104) and to the de novo prediction of bacterial interactomes (105). These procedures exploit the idea of correlated mutations [i.e., possible interaction sites can be revealed by the high correlation among mutations in a species-matched multiple sequence alignment of protein pairs (70)] and can be reliably applied only when large and diverse multiple sequence alignments can be derived from available sequence databases.

Recently, advances in machine learning procedures facilitated the building of computational frameworks that can extract complex information from protein sequences and profiles of interacting and noninteracting proteins. In this context, it is fundamental to collect pairs of proteins that are unlikely to interact, and a dedicated dataset (Negatome 2.0) has been released to this aim (106). On this basis, inference systems are available for the prediction of interacting protein pairs (107–109).

Table 6 Predicting PPI sites and flexible regions of proteins of the Cajal body

UniProt code	Gene	Length	PPIs	Flexible sites[a]	Flexible PPIs	IntAct interactors	BioGRID interactors
P38432	*COIL*	576	149	244	14	123	110
Q9BUR4	*WRAP53*	548	166	165	24	41	54
Q16637	*SMN1*	294	90	145	44	268	213
P55199	*ELL*	621	64	194	3	36	54
Q06787	*FMR1*	632	106	243	49	294	84
Q14331	*FRG1*	258	44	90	13	14	19
Q15020	*SART3*	963	115	199	4	125	211
Q5JVS0	*HABP4*	413	112	313	87	25	105
Q5W0Q7	*USPL1*	1,092	136	169	4	24	25
Q6NT76	*HMBOX1*	420	42	151	12	130	70
Q7L014	*DDX46*	1,031	170	228	55	27	65
Q7Z6G8	*ANKS1B*	1,248	184	304	22	21	10
Q8WWY3	*PRPF31*	499	75	256	46	220	193
Q96JC9	*EAF1*	268	33	164	15	105	76
Q96RS0	*TGS1*	853	100	263	3	43	45
Q9BPY3	*FAM118B*	351	33	66	8	48	51
Q9H089	*LSG1*	658	106	173	5	22	31
Q9H814	*PHAX*	394	61	91	20	44	64
Q9NPE3	*NOP10*	64	21	34	11	31	37
Q9UBY9	*HSPB7*	170	41	39	14	24	14
Q9Y2F5	*ICE1*	2,266	131	547	7	23	25
Q9Y4X5	*ARIH1*	557	147	126	53	81	198
P14678	*SNRPB*	240	74	146	27	144	171
Q14978	*NOLC1*	699	77	573	15	90	152
P54253	*ATXN1*	815	394	242	130	351	266

[a]Number of predicted flexible sites with MobiDB (112).

Finally, from the current knowledge of partially solved interactomes, it is possible to extract rules of interaction patterns. New possible interaction pairs can be then inferred from the application of these rules, for example, (*a*) "proteins similar to interacting proteins are likely to interact as well" (110) and (*b*) "proteins interact if one of them is similar to the other's partners" (111).

8. PREDICTING INTERACTING SITES IN PROTEINS INVOLVED IN PHASE TRANSITIONS

In the following, we test some of the PPI predictors to investigate their efficacy in exploring features of proteins known to undergo a phase transition. We consider a set of proteins known to be involved in Cajal bodies in human cell. In **Table 6**, we list 25 proteins whose sequence was deposited in the September 2019 Uniprot Knowledgebase release and that contain a `Cajal body cellular component` annotation. Most of the proteins have only a known sequence. For this reason, we predicted PPI sites with a predictor that was developed in house to take sequence as input and that was based on a deep learning procedure [an adaptation of our ISPRED4 (30)]. We are perfectly aware of all the limitations of sequence-based PPI predictors, as explained above. According to several papers (see Reference 10 and references therein), proteins involved in phase

Table 7 Correlation between the number of interactors, PPIs, and flexible sites of the Cajal granule proteins

	PPIs[a]	Flexible sites	Flexible PPIs[b]
IntAct	0.4	0.05	0.59
BioGRID	0.41	0.12	0.59

[a]Significant at 5%.
[b]Significant at 1%.

transitions are endowed with intrinsically disordered regions (IDRs) and we adopted MobiDB (112) as a source of annotation. We also searched the human interactomes for possible interactors and listed in **Table 6** the number of interacting proteins (interactors) reported for each protein in IntAct and BioGRID. Interestingly, all of the Cajal body human proteins have a much larger number of interactors than average (13 and 18 per human protein in IntAct and in BioGRID, respectively). The number of interactors per protein moderately correlates with the number of residues in IDRs (flexible sites); correlation increases if the number of PPIs is considered and reaches a satisfactory value when considering the number of residues that can be annotated both as PPIs and as IDRs (**Table 7**).

This suggests that eventually the inherent flexibility of the residues makes it possible to adjust the interacting surface protein to multiple partners. Previously we clarified that, on average, for proteins involved in the human cell cycle, the number of interacting partners correlates with the number of predicted interaction patches (113), a result that is different from the one here reported. It appears that in granules the inherent flexibility of residues in interaction patches matters. This result is partial and preliminary in order to show what we can do for the annotation of proteins involved in phase separation. In our opinion, this is sufficient for the time being and, with the present amount of knowledge, to stir further experiments with the aim of solving the electronic structure of proteins involved in phase transition phenomena.

9. FINAL REMARKS

From our brief overview, we can presently conclude that our knowledge of PPIs is limited, given the relatively few complexes solved with atomic resolution and the results of large-scale proteomic experiments, which are affected by the adopted experimental approach.

Apparently, computational biology can offer a wide spectrum of computational methods to address the problem of interface annotation with good performance when trained on 3D complexes. Methods predicting interactions from protein sequences perform less well. We have listed presently available state-of-the-art methods that can help in functionally annotating new protein sequences and in steering new experiments for increasing our knowledge of the cell's interior complexity. However, it is important to highlight that international benchmarking experiments, like CAPRI, indicate that predictive methods are far from perfect and that there is always room for improvement. All of the methods are based on the assumption that input takes into consideration previous and possibly highly curated knowledge, and in doing so, they are heuristic and not theoretically based. Their efficacy is therefore dependent on their ability to extract general principles of association between input knowledge and expected output. Along this line, it is evident that more structural data of protein complexes will largely improve the capability of computational methods by increasing their performance.

Presently, we also face the problem of protein phase separation, and this may also affect proteomic and interactomic results. For these proteins, poorly exploited from the experimental point

of view, here we have described how in the case of proteins of the Cajal body the number of possible interactors, as derived from two independent interactomic databases, correlates well with the number of protein residues predicted to be at the interface, which are also predicted in flexible regions. This requirement is not exclusive. Even in the case of proteins involved in phase separation, the acquisition of more experimental data will help in developing computational tools suited to their functional annotation.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Rivas G, Minton AP. 2016. Macromolecular crowding in vitro, in vivo, and in between. *Trends Biochem. Sci.* 41:970–81
2. Shin Y, Brangwynne CP. 2017. Liquid phase condensation in cell physiology and disease. *Science* 357:1253–65
3. Rivas G, Minton AP. 2018. Toward an understanding of biochemical equilibria within living cells. *Biophys. Rev.* 10:241–53
4. Jain S, Wheeler JR, Walters RW, Agrawal A, Barsic A, Parker R. 2016. ATPase-modulated stress granules contain a diverse proteome and substructure. *Cell* 164:487–98
5. Putnam A, Cassani M, Smith J, Seydoux G. 2019. A gel phase promotes condensation of liquid P granules in *Caenorhabditis elegans* embryos. *Nat. Struct. Mol. Biol.* 26:220–26
6. Schuster BS, Reed EH, Parthasarathy R, Jahnke CN, Caldwell RM, et al. 2018. Controllable protein phase separation and modular recruitment to form responsive membraneless organelles. *Nat. Commun.* 9:2985–97
7. Neugebauer KM. 2017. Special focus on the Cajal body. *RNA Biol.* 14(6):669–70
8. Alberti S, Gladfelter A, Mittag T. 2019. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* 176:419–34
9. Alberti S. 2017. Phase separation in biology. *Curr. Biol.* 27:R1089–107
10. Boeynaems S, Alberti S, Fawzi NL, Mittag T, Polymenidou M, et al. 2018. Protein phase separation: a new phase in cell biology. *Trends Cell Biol.* 28:420–35
11. Alberti S, Dormann D. 2019. Liquid–liquid phase separation in disease. *Annu. Rev. Genet.* 53:171–94
12. Banani SF, Lee HO, Hyman AA, Rosen MK. 2017. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* 18:285–98
13. Baldi P. 2018. Deep learning in biomedical data science. *Annu. Rev. Biomed. Data Sci.* 1:181–205
14. Shahid S, Hassan MI, Islam A, Ahmad F. 2017. Size-dependent studies of macromolecular crowding on the thermodynamic stability, structure and functional activity of proteins: in vitro and in silico approaches. *Biochim. Biophys. Acta Gen. Subj.* 1861(2):178–97
15. Jankauskaite J, Jiménez-García B, Dapkunas J, Fernández-Recio J, Moal IH. 2019. SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* 35(3):462–69
16. Wang R, Fang X, Lu Y, Yang CY, Wang S. 2005. The PDBbind database: methodologies and updates. *J. Med. Chem.* 48(12):4111–19
17. Levy ED. 2010. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.* 403:660–70
18. Pandav G, Pryamitsyn VA, Errington JE, Ganesan V. 2015. Multibody interactions, phase behavior and clustering in nanoparticle-polyelectrolyte mixtures. *J. Phys. Chem. B* 119:14536–50
19. Samanta R, Ganesan V. 2018. Influence of protein charge patches on the structure of protein–polyelectrolyte complexes. *Soft Matter* 14:3748–59

20. Rost B, Sander C. 1994. Conservation and prediction of solvent accessibility in protein families. *Proteins* 20:216–26
21. Cazals F. 2010. Revisiting the Voronoi description of protein–protein interfaces. *Pattern Recognit. Bioinform.* 6282:419–30
22. Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML. 2009. Progress and challenges in predicting protein–protein interaction sites. *Brief. Bioinform.* 10(3):233–46
23. de Vries SJ, Bonvin AM. 2008. How proteins get in touch: interface prediction in the study of biomolecular complexes. *Curr. Protein Pept. Sci.* 9(4):394–406
24. Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–637
25. Janin J, Bahadur RP, Chakrabarti P. 2008. Protein–protein interaction and quaternary structure. *Q. Rev. Biophys.* 41(2):133–80
26. Bradford JR, Westhead DR. 2005. Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics* 21(8):1487–94
27. Liang S, Zhang C, Liu S, Zhou Y. 2006. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.* 34(13):3698–707
28. de Vries SJ, van Dijk AD, Bonvin AM. 2006. WHISCY: What information does surface conservation yield? Application to data-driven docking. *Proteins* 63(3):479–89
29. Zellner H, Staudigel M, Trenner T, Bittkowski M, Wolowski V, et al. 2012. PresCont: predicting protein-protein interfaces utilizing four residue properties. *Proteins* 80(1):154–68
30. Savojardo C, Fariselli P, Martelli PL, Casadio R. 2017. ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model. *Bioinformatics* 33(11):1656–63
31. Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics* 23(15):1875–82
32. Neuvirth H, Raz R, Schreiber G. 2004. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. *J. Mol. Biol.* 338(1):181–99
33. Chen H, Zhou HX. 2005. Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins* 61(1):21–35
34. Ofran Y, Rost B. 2007. ISIS: interaction sites identified from sequence. *Bioinformatics* 23(2):e13–16
35. Murakami Y, Mizuguchi K. 2010. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics* 26(15):1841–48
36. Li BQ, Feng KY, Chen L, Huang T, Cai YD. 2012. Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS. *PLOS ONE* 7(8):e43927
37. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–402
38. Remmert M, Biegert A, Hauser A, Söding J. 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9(2):173–75
39. Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9(1):56–68
40. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23:1282–88
41. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. 2004. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* 13(1):190–202
42. Sanchez-Garcia R, Sorzano COS, Carazo JM, Segura J. 2019. BIPSPI: a method for the prediction of partner-specific protein–protein interfaces. *Bioinformatics* 35(3):470–77
43. Minhas Fu, Geiss BJ, Ben-Hur A. 2014. PAIRpred: partner-specific prediction of interacting residues from sequence and structure. *Proteins* 82(7):1142–55
44. Gallet X, Charloteaux B, Thomas A, Brasseur R. 2000. A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.* 302(4):917–26
45. Xue LC, Dobbs D, Honavar V. 2011. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinform.* 12:244

46. Zhang QC, Deng L, Fisher M, Guan J, Honig B, Petrey D. 2011. PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res*. 39:W283–87
47. Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V. 2012. Predicting protein-protein interface residues using local surface structural similarity. *BMC Bioinform*. 13:41
48. Jelínek J, Škoda P, Hoksza D. 2017. Utilizing knowledge base of amino acids structural neighborhoods to predict protein-protein interaction sites. *BMC Bioinform*. 18(Suppl. 15):492
49. Jones S, Thornton JM. 1997. Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* 272(1):133–43
50. Koike A, Takagi T. 2004. Prediction of protein–protein interaction sites using support vector machines. *Protein Eng. Des. Sel.* 17(2):165–73
51. Li MH, Lin L, Wang XL, Liu T. 2007. Protein–protein interaction site prediction based on conditional random fields. *Bioinformatics* 23(5):597–604
52. Liu B, Wang X, Lin L, Tang B, Dong Q, Wang X. 2009. Prediction of protein binding sites in protein structures using hidden Markov support vector machine. *BMC Bioinform*. 10:381
53. Porollo A, Meller J. 2007. Prediction-based fingerprints of protein–protein interactions. *Proteins* 66(3):630–45
54. Sikić M, Tomić S, Vlahovicek K. 2009. Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. *PLOS Comput. Biol.* 5(1):e1000278
55. Dong Z, Wang K, Dang TK, Gültas M, Welter M, et al. 2014. CRF-based models of protein surfaces improve protein-protein interaction site predictions. *BMC Bioinform*. 15:277
56. Fariselli P, Pazos F, Valencia A, Casadio R. 2002. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* 269(5):1356–61
57. Daberdaku S, Ferrari C. 2018. Exploring the potential of 3D Zernike descriptors and SVM for protein–protein interface prediction. *BMC Bioinform*. 19(1):35
58. Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, et al. 2015. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.* 427(19):3031–41
59. Lensink MF, Velankar S, Wodak SJ. 2017. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins* 85(3):359–77
60. Aumentado-Armstrong TT, Istrate B, Murgita RA. 2015. Algorithmic approaches to protein-protein interaction site prediction. *Algorithms Mol. Biol.* 10:7
61. Esmaielbeiki R, Krawczyk K, Knapp B, Nebel JC, Deane CM. 2016. Progress and challenges in predicting protein interfaces. *Brief. Bioinform.* 17(1):117–31
62. Xue LC, Dobbs D, Bonvin AM, Honavar V. 2015. Computational prediction of protein interfaces: a review of data driven methods. *FEBS Lett*. 589(23):3516–26
63. Wei Z, Han K, Yang J, Shen H, Yu D. 2016. Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests. *Neurocomputing* 193:201–12
64. Hou Q, De Geest PFG, Vranken WF, Heringa J, Feenstra KA. 2017. Seeing the trees through the forest: sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics* 33(10):1479–87
65. Zhang J, Kurgan L. 2019. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* 35(14):i343–53
66. Chen P, Li J. 2010. Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. *BMC Bioinform*. 11:402
67. Yan C, Dobbs D, Honavar V. 2004. A two-stage classifier for identification of protein–protein interface residues. *Bioinformatics* 20(Suppl. 1):i371–78
68. Res I, Mihalek I, Lichtarge O. 2005. An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics* 21(10):2496–501
69. Dhole K, Singh G, Pai PP, Mondal S. 2014. Sequence-based prediction of protein–protein interaction sites with L1-logreg classifier. *J. Theor. Biol.* 348:47–54
70. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. 1997. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* 271(4):511–23

71. Ovchinnikov S, Kamisetty H, Baker D. 2014. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* 3:e02030
72. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *PNAS* 106(1):67–72
73. Hopf TA, Schärfe CP, Rodrigues JP, Green AG, Kohlbacher O, et al. 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3:e03430
74. Zhang J, Kurgan L. 2018. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief. Bioinform.* 19(5):821–37
75. Hou Q, De Geest P, Griffioen CJ, Abeln S, Heringa J, Feenstra KA. 2019. SeRenDIP: sequential remastering to derive profiles for fast and accurate predictions of PPI interface positions. *Bioinformatics* 35:4794–96
76. Gromiha M, Yugandhar K, Jemimah S. 2017. Protein–protein interactions: scoring schemes and binding affinity. *Curr. Opin. Struct.* 44:31–38
77. Keskin O, Tuncbag N, Gursoy A. 2016. Predicting protein–protein interactions from the molecular to the proteome level. *Chem. Rev.* 116:4884–909
78. Snider J, Kotlyar M, Saraon P, Yao Z, Jurisica I, Stagljar I. 2015. Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.* 11:848
79. Cafarelli TM, Desbuleux A, Wang Y, Choi SG, De Ridder D, Vidal M. 2017. Mapping, modeling, and characterization of protein-protein interactions on a proteomic scale. *Curr. Opin. Struct. Biol.* 44:201–10
80. Jensen LJ, Bork P. 2008. Not comparable, but complementary. *Science* 322:56–57
81. De Las Rivas J, Fontanillo C. 2012. Protein–protein interaction networks: unraveling the wiring of molecular machines within the cell. *Brief. Funct. Genom.* 11:489–96
82. Luck K, Sheynkman GM, Zhang I, Vidal M. 2017 Proteome-scale human interactomics. *Trends Biochem. Sci.* 42:342–54
83. Fields S, Song O. 1989. A novel genetic system to detect protein–protein interactions. *Nature* 340:245–46
84. Luo Y, Batalao A, Zhou H, Zhu L. 1997. Mammalian two-hybrid system: a complementary approach to the yeast two-hybrid system. *Biotechniques* 22:350–52
85. Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U. 2009. Yeast two-hybrid, a powerful tool for systems biology. *Int. J. Mol. Sci.* 10:2763–88
86. Dunham WH, Mullin M, Gingras AC. 2012. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* 12:1576–90
87. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, et al. 2012. A census of human soluble protein complexes. *Cell* 150:1068–81
88. Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, et al. 2020. A reference map of the human binary protein interactome. *Nature* 580:402–8
89. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, et al. 2017. Architecture of the human interactome defines protein communities and disease networks. *Nature* 545:505–9
90. Wodak SJ, Vlasblom J, Turinsky AL, Pu S. 2013. Protein–protein interaction networks: the puzzling riches. *Curr. Opin. Struct. Biol.* 23:941–53
91. Szklarczyk D, Jensen LJ. 2015. Protein-protein interaction databases. *Methods Mol. Biol.* 1278:39–56
92. Miryala SK, Anbarasu A, Ramaiah S. 2018. Discerning molecular interactions: a comprehensive review on biomolecular interaction databases and network analysis tools. *Gene* 642:84–94
93. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, et al. 2014. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42:D358–63
94. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, et al. 2019. The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47:D529–41
95. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, et al. 2008. Estimating the size of the human interactome. *PNAS* 105:6959–64
96. Tompa P, Davey NE, Gibson TJ, Babu MM. 2014. A million peptide motifs for the molecular biologist. *Mol. Cell* 55:161–69
97. Mosca R, Céol A, Aloy P. 2013. Interactome3D: adding structural details to protein networks. *Nat. Methods* 10:47–53

98. Dapkunas J, Timinskas A, Olechnovic K, Margelevicius M, Diciunas R, Venclovas C. 2017. The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures. *Bioinformatics* 33:935–37
99. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *PNAS* 96:4285–88
100. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402:86–90
101. Jansen R, Greenbaum D, Gerstein M. 2002. Relating whole-genome expression data with protein-protein interactions. *Genome Res*. 12:37–46
102. Garcia-Garcia J, Schleker S, Klein-Seetharaman J, Oliva B. 2012. BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic Acids Res*. 40:W147–51
103. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 47:D607–13
104. Hosur R, Peng J, Vinayagam A, Stelzl U, Xu J, et al. 2012. A computational framework for boosting confidence in high-throughput protein-protein interaction datasets. *Genome Biol*. 13:R76
105. Cong Q, Anishchenko I, Ovchinnikov S, Baker D. 2019. Protein interaction networks revealed by proteome coevolution. *Science* 365:185–89
106. Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, et al. 2014. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res*. 42:D396–400
107. Hashemifar S, Neyshabur B, Khan AA, Xu J. 2018. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* 34:i802–10
108. Romero-Molina S, Ruiz-Blanco YB, Harms M, Münch J, Sanchez-Garcia E. 2019. PPI-Detect: a support vector machine model for sequence-based prediction of protein-protein interactions. *J. Comput. Chem.* 40:1233–42
109. Yao Y, Du X, Diao Y, Zhu H. 2019. An integration of deep learning with feature embedding for protein–protein interaction prediction. *PeerJ* 7:e7126
110. Li Y, Ilie L. 2017. SPRINT: ultrafast protein-protein interaction prediction of the entire human interactome. *BMC Bioinform*. 18:485
111. Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, et al. 2019. Network-based prediction of protein interactions. *Nat. Commun*. 10(1):1240
112. Piovesan D, Tabaro F, Paladin L, Necci M, Micetic I, et al. 2018. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res*. 46:D471–76
113. Bartoli L, Martelli PL, Rossi I, Fariselli P, Casadio R. 2010. The prediction of protein-protein interacting sites in genome-wide protein interaction networks: the test case of the human cell cycle. *Curr. Protein Pept. Sci.* 11:601–8