HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2013 April 25.

Published in final edited form as:

Nature. 2012 October 25; 490(7421): 556–560. doi:10.1038/nature11503.

Structure-based prediction of protein-protein interactions on a genome-wide scale

Qiangfeng Cliff Zhang[1,2,3,*], **Donald Petrey**[1,2,3,*], **Lei Deng**[2,3,8], **Li Qiang**[6], **Yu Shi**[7], **Chan Aye Thu**[2], **Brygida Bisikirska**[3], **Celine Lefebvre**[3,5], **Domenico Accili**[6], **Tony Hunter**[7], **Tom Maniatis**[2], **Andrea Califano**[2,3,4,5,#], and **Barry Honig**[1,2,3,#]

[1]Howard Hughes Medical Institute

[2]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032

[3]Columbia Initiative in Systems Biology, Columbia University, New York, NY10032

[4]Department of Biomedical Informatics, Columbia University, New York, NY 10032

[5]Institute of Cancer Genetics, Columbia University, New York, NY 10032

[6]Naomi Berrie Diabetes Center, Department of Medicine, College of Physicians & Surgeons of Columbia University, New York, NY 10032

[7]Molecular and Cell Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037

[8]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

Abstract

The genome-wide identification of pairs of interacting proteins is an important step in the elucidation of cell regulatory mechanisms[1,2]. Much of our current knowledge derives from high-throughput techniques such as yeast two hybrid and affinity purification[3], as well as from manual curation of experiments on individual systems[4]. A variety of computational approaches based, for example, on sequence homology, gene co-expression, and phylogenetic profiles have also been developed for the genome-wide inference of protein-protein interactions (PPIs)[5,6]. Yet, comparative studies suggest that the development of accurate and complete repertoires of PPIs is still in its early stages[7–9]. Here we show that three-dimensional structural information can be used to predict PPIs with an accuracy and coverage that are superior to predictions based on non-structural evidence. Moreover, an algorithm, PrePPI, that combines structural information with

Correspondence and requests for materials should be addressed to A.C. (califano@c2b2.columbia.edu) or B.H. (bh6@columbia.edu).

*These authors contributed equally to this work.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author contributions Q.C.Z., D.P., A.C. and B.H. designed the research; Q.C.Z. performed the computational work; Q.C.Z., D.P., A.C. and B.H. analyzed the data; L.D. set up the PrePPI web server, L.Q., Y.S., C.A.T., and B.B. performed co-IP studies, Q.C.Z., D.P., A.C. and B.H. wrote the paper including text from C.L., D.A., T.H. and T.M..

other functional clues is comparable in accuracy to high-throughput experiments, yielding over 30,000 high confidence interactions for yeast and over 300,000 for human. Experimental tests of a number of predictions demonstrate the ability of the PrePPI algorithm to identify unexpected PPIs of significant biological interest. The surprising effectiveness of three-dimensional structural information can be attributed to the use of homology models combined with the exploitation of both close and remote geometric relationships between proteins.

To date, structural information has had relatively little impact in constructing protein-protein interactomes, primarily because there is a dramatic difference between the number of proteins with known sequence and those with an experimentally known structure. For example, as of early 2010, the PDB (Protein Data Bank) provides structures for ~600 of the total complement of ~6,500 yeast proteins (~10%), while structural coverage of protein-protein complexes is even more sparse with only about 300 structures available out of the approximately 75,000 PPIs (<0.5%) recorded in publically available databases. However, ~3,600 additional yeast proteins have homology models in either the ModBase[10] or SkyBase[11] databases. Moreover, there were about 37,000 protein-protein complexes derived from multiple organisms in the PDB and PQS[12] (Protein Quaternary Structure) databases, that might be used as "templates" to model PPIs. Clearly, if structure is to be useful on a large scale, it is essential that modeling of individual proteins and of complexes be exploited.

A number of studies have used structurally characterized complexes as "templates" to construct models of complexes that might be formed between proteins that have been classified as having sequence and/or structural relationships to the proteins in the template[13–15]. Here we search more broadly for templates using geometric relationships between groups of secondary structure elements as revealed by structural alignment, independently of how they are classified. It has been demonstrated that even distantly related proteins often use regions of their surface with similar arrangements of secondary structure elements to bind to other proteins[16–18], suggesting the possibility of significantly expanding the number of putative PPIs that can be identified. It is likely that further expansion can be achieved if interactions involving unstructured regions of proteins are taken into account, but these are not considered in the current work.

Our approach to the prediction of PPIs is embodied in an algorithm we have named PrePPI (Predicting Protein-Protein Interactions) that combines structural and non-structural interaction clues using Bayesian statistics (see Figure 1 and online Methods for details). The structural component of PrePPI involves a number of steps. Briefly, given a pair of query proteins (QA and QB), we first use sequence alignment to identify structural representatives (MA and MB) that correspond to either their experimentally determined structures or homology models. We then use structural alignment to find both close and remote structural neighbors (NA_i and NB_j) of MA and MB (an average of ~1500 neighbors are found for each structure). Whenever two (e.g. NA_1 and NB_3) of the over 2 million pairs of neighbors of MA and MB form a complex reported in the PDB, this defines a template for modeling the interaction of QA and QB. Models of the complex are created by superimposing the representative structures on their corresponding structural neighbors in the template (i.e.,

MA on NA_1 and MB on NB_3). This procedure produces about 550 million "interaction models" for about 2.4 million PPIs involving about 3,900 yeast proteins and about 12 billion models for about 36 million PPIs involving about 13,000 human proteins. Note that an interaction model is based on structure-based sequence alignments of query proteins to their individual templates (Figure S1) and that we do not construct a three-dimensional model of each complex since the scoring of so many individual complexes would be prohibitively time consuming using standard energy functions (for example as used in docking[19]).

Once an interaction model has been created, it is evaluated using a combination of five empirical scores that measure properties derived from alignments of the individual monomers to their templates (Figure S1). The first score, SIM, depends on the structural similarity between models of the two query proteins (i.e. MA and MB) and those in the template complex (i.e. NA_1 and NB_3). The next two scores determine whether the interface in the template complex actually exists in the model. They are calculated as SIZ, the number and COV, the fraction of interacting residue pairs in the template (e.g. NA_1-NB_3) that align to some pair of residues in the model (MA-MB). The final two scores reflect whether the residues that appear in the model interface have properties consistent with those that mediate known PPIs (e.g., residue type, evolutionary conservation, or statistical propensity to be in protein-protein interfaces). This information is obtained from three publically available servers that predict interfacial residues based on the sequence and structure of the individual subunits of the model[20–22]. These scores are calculated as OS, identical to SIZ but with the additional requirement that both residues in an interacting pair of the template align to predicted interfacial residues in MA and MB and OL, the number of template interfacial residues that align to predicted interfacial residues in MA and MB. We note that although the interaction models produced by our procedure can reveal the approximate locations of potential interfaces, they will not, in general, be accurate at atomic resolution.

The five empirical scores are combined using a Bayesian network (Figure S2) to yield a likelihood ratio (LR) that a candidate protein-protein complex represents a true interaction (see Methods online). The network is trained on positive and negative "gold standard" reference datasets. Similar to two recent studies[23,24], we combine interaction data from multiple databases to ensure a broad coverage of true interactions. We divide these sets into high-confidence (HC) and low-confidence (LC) subsets (Table S1); the HC sets contain 11,851 yeast interactions and 7,409 human interactions which have more than one publication supporting their existence; interactions with only one supporting publication compose the LC set. All potential PPIs in a given genome *not* in the HC+LC set form the negative (N) reference set. Using the Bayesian network classifier trained on the yeast HC set, we select the best interaction model with the highest LR for each PPI.

To quantitatively assess the performance of structural modeling (SM), we compared it with a number of non-structural clues previously used to infer PPIs[24–26]: a) essentiality of the proteins in the interacting pair; b) co-expression level; c) Gene Ontology (GO) functional similarity; d) MIPS functional similarity; and e) phylogenetic profile similarity. We used the same algorithms or data for other clues as Gerstein and coworkers[25] but developed our own phylogenetic profile algorithm (see details in Methods online and Table S2). Briefly, a phylogenetic profile was constructed for every protein using a set of completely resolved

proteomes as references. Since interacting proteins tend to co-evolve, proteins with similar profiles are predicted to interact.

As shown in Figure S3 and S4, SM yields comparable performance to other clues over the entire range of false positive rate (FPR) but is considerably more effective at low FPR (e.g. FPR ≤ 0.1%). This is critical since, due to the huge number of negative interactions, only very low FPR rates can produce a small enough number of false positives to be used effectively in practice. At low FPR, SM by itself outperforms even the naïve Bayesian classifiers that combine all non-structure-based clues (NS). Looking specifically at the thousands of high confidence SM predictions in the LC and the N sets with an LR score > 600 (a value used in Ref. 25 and corresponding in our study to FPR of ~0.1%, see Methods online), about 70% and 50%, respectively, share GO biological term at, or more specific than, the 6th level of the GO hierarchy, suggesting that many of these interactions are real (Figure S5).

As mentioned above, PrePPI combines structural and non-structural clues using a naïve Bayesian network[24–26]. Figures S4 shows that PREPPI's performance is superior to that obtained from structural and non-structural evidence alone implying that the two sources of information are largely complementary. This point can be clearly seen in the Venn diagrams of high confidence (LR > 600) predictions shown in Figure S6. It is evident from the figure that combining structural and non-structural clues yields many more high confidence predictions and identifies more HC interactions than either source of information alone. As an independent test of PrePPI, we assessed its performance against one of the challenges in the 2009 DREAM (Dialogue for Reverse Engineering Assessments and Methods) workshop specifically aimed at PPI predictions[27]. As discussed in Table S3, PrePPI outperformed all other methods for cases where structural information is available.

We have compared the performance of PrePPI to that of high-throughput (HT) experiments (Table S4) using data provided in a detailed comparison of different HT techniques reported by Vidal and coworkers[23]. We used their datasets to define true positives and compiled a new negative reference set which consists of protein pairs where each protein in a pair is annotated as localized to a different cellular compartment (see Figure S7 and Methods online). This was essential for comparison to experimental assays, since, as constructed, our N set excludes data compiled from HT experiments, and hence the FPR for experimental assays is artificially zero (see also related discussion in SOM of Ref. 23).

As can be seen in the ROC (receiver operating characteristic) curves reported in Figure 2A and Figure S8, PrePPI performance is generally comparable, although somewhat better overall, than HT methods for most data sets that were tested. Figure 2B shows a Venn diagram in which the PrePPI dataset is based on an LR cutoff of 600 (FPR ≈ 0.1%). Results for other LRs and additional reference sets are shown in Figure S9. As can be seen, many of the interactions inferred by PrePPI are different from those identified by HT assays. Methods that combine both approaches may thus prove to be highly effective in expanding the coverage of PPIs.

At an LR cutoff of 600, PrePPI predicts 31,402 high confidence interactions for yeast and 317,813 interactions for human. These, as well as predictions with lower LR scores, are available in a database from the PrePPI website (http://bhapp.c2b2.columbia.edu/PrePPI/). As a further validation of PrePPI we tested its performance on the approximately 24,000 new interactions involving human proteins that were added to public databases after August 2010 (Table S5). Among these interactions, 1,644 are predicted by PrePPI to have an LR>600 (based on a Bayesian classifier derived from pre-2009 data on yeast) so that they essentially correspond to experimental validation of true predictions.

Specific experimental validation of 19 individual PrePPI predictions, using co-immunoprecipitation (co-IP) assays, was carried out in four separate labs, leading to confirmation of 15 of these interactions (Figure S10~14, Table S6). Specifically, the investigators in each lab queried the PrePPI database for previously uncharacterized interactions involving proteins of interest and which, as much as possible, had relatively high SM and PrePPI scores (see Table S6 for more information). Here we briefly discuss some of our findings with emphasis on the structural domains predicted by PrePPI to form the protein-protein interface.

One set of predictions involves potential PPIs formed between the nuclear receptor peroxisome proliferator-activated receptor gamma (PPARγ) and other transcription factors. PPARγ plays a pivotal role in regulating glucose and lipid metabolism, inflammatory response and tumorigenesis and is known to heterodimerize with Retinoid X Receptors (RXRs) and to recruit cofactors to regulate target gene transcription. PrePPI predicts high confidence interactions between PPARγ and the transcription factors LXRβ, PAX7, PDX1, NKX2.2 and HHEX (Table S6). Except for HHEX, all of the interactions were validated (Figure S10). The predicted interaction with nuclear receptor LXRβ might have been expected based on the ability of these proteins to heterodimerize through their ligand binding domains. Nevertheless, this specific interaction had not previously been characterized and suggests a heretofore unrecognized convergence of signaling and metabolic pathways regulated by these two nuclear receptors. The interaction between the ligand binding domain of PPARγ and the homeodomains of PAX7, PDX1 and NKX2.2 are fundamentally new observations that require further studies, as they suggest that PPARγ may have a role in endocrine progenitor and pancreatic beta-cell development.

A second set of examples involves the suppressor of cytokine signaling protein, SOCS3, an SH2 domain-containing protein that negatively regulates cytokine-induced signal transduction. To date, the mechanism of the inhibitory function of SOCS3 has been primarily established for its involvement in the JAK/STAT pathway. PrePPI predicts that SOCS3 forms complexes with GRB2 and RAF1, two key components in the Ras/MAPK pathway, and these interactions were confirmed experimentally (Figure S11A and B). PrePPI also predicts the formation of a complex between of SOCS3 and BTK, a cytoplasmic tyrosine kinase important in B-lymphocyte development, differentiation, and signaling, and this interaction was also validated (Figure S11C). The SOCS3 GRB2 interaction is predicted to be mediated by their SH2 domains, whereas the SOCS3 interaction with BTK is predicted to be mediated by an SH2-SH3 domain interaction. Analysis of the predicted binding preferences of SH2 domains as well as results on other protein families indicates that the

PrePPI scoring function accounts, at least in part, for the binding preference of closely related protein domains (Figure S15, see also below).

A third group of novel observations involves the identification of kinases that interact with the clustered protocadherin proteins (protocadherin α, β and γ – PCDHα, β and γ). The PCDHs have six cadherin-like extracellular domains and unique cytoplasmic domains. They assemble into large complexes at the cell surface, and associate with a variety of proteins, including signaling adaptors, kinases and phosphatases. Analysis of potential PCDH-kinase PPIs confirmed published interactions between PCDHα and γ with the tyrosine kinase RET, and predicted interactions with ROR2, VEGFR2 and ABL1 (Tables S6, Figure S12 – experiments done in mice). PrePPI predicts that these PPIs are mediated by the extracellular cadherin domains and Ig domains, a result that was confirmed experimentally (Figure S12A~D). A hydrophobic residue, Phe 64, of the ROR2-Ig domain is predicted to be in the center of the interface it forms with PCDHα4. Mutating this Phe to an Ala, a smaller hydrophobic residue, has no detectable effect on binding while mutating it to charged residues significantly weakens the interaction (Figure S12B and C). These results suggest that, in addition to predicting binary interactions, PrePPI has the potential to reveal novel and unsuspected interfaces.

The fourth group of experiments was carried out with the goal of identifying new components of large protein-protein complexes. We validated two previously uncharacterized interactions between the special AT-rich sequence-binding protein SATB2 and the Emerin “proteome” complex 32, and one involving the pre-mRNA-processing factor PRPF19 and the centromere chromatin complex (Figure S13). It is important to emphasize that each of the PPIs detected must be confirmed through appropriate *in vivo* experiments. Taken together, however, these findings suggest that PrePPI has sufficient accuracy and sensitivity to provide a wealth of novel hypotheses that can drive biological discovery.

The accuracy and range of applicability of PrePPI, and the crucial role of structural modeling, were unanticipated, but should not come as a complete surprise. Most protein complexes in the PDB have structural neighbors that share binding properties[17], and protein interface space may well be close to “complete” in terms of the packing orientations of secondary structure elements[18]. Moreover, these elements can be identified with geometric alignment methods[17,28], a fact that has been exploited in the approach introduced here. Although the information required to predict whether two proteins interact appears to be present in the PDB, the question has been how to mine the data.

Three key elements are responsible for the success of structural modeling and PrePPI. The first is the significant expansion of the number of interactions that can be modeled, due to the use of both homology models and remote structural relationships. About 8,600 PDB structures but more than 31,000 models are found as representatives of at least one domain of ~14,100 human proteins. If we had only used experimentally determined structures in our analysis, a total of only ~2.5 million human PPIs (vs. 36 million when homology models are used) could have been modeled. Similarly, had we limited ourselves to structural neighbors taken from the same SCOP fold, only ~225 thousand interactions could have been modeled, as opposed to 36 million.

As might be expected, predictions based on the structural modeling that use only PDB structures or close structural neighbors are more likely to recover known interactions (defined by their presence in databases) than those that only use homology models or remote structural relationships (Figure S16). However the latter, on their own, yield a dramatic expansion in the total number of interaction models and, consequently many more high confidence predictions and known interactions. Most importantly, in the calculation of the PrePPI score, the huge number of low confidence structural interaction models lead to an even greater expansion in high-confidence predictions when combined with functional, evolutionary and other sources of evidence (Figure S16).

The second key element in our strategy is the efficiency of our scoring scheme for interaction models which allows us to evaluate an extremely large number of models while still discriminating among closely related family members. Discrimination among complexes involving members of the same protein family, i.e. specificity, is obtained from the properties of the predicted interface, e.g. the statistical propensity of certain amino acids to appear in interfaces[20,21] (and, additionally, from non-structural clues, e.g. are the two proteins co-expressed). As examples, our analysis of the SH2 and GTPase families shows that the structural modeling (and PrePPI scores) for these closely related proteins produce a wide range of LRs with the higher LRs associated with a higher probability of being a known interaction (Figure S15).

The third element responsible for the success of PrePPI is the Bayesian evidence integration method that allows independent and possibly weak interaction clues to be combined so as to make reliable predictions and to improve prediction specificity (Figure S15~16).

Figure 3 provides two examples of the use of remote structural relationships and homology models. In Figure 3A, an HC set interaction of serine/threonine-protein kinase D1 (PKD1) and protein kinase C epsilon (PKCε) is recovered by structural modeling using a complex of two proteins in the ubiquitin pathway (not kinases) as template. Note that PKD1 and PKCε are not sequence homologues of the two corresponding ubiquitin pathway proteins and are classified as belonging to different SCOP folds. However, the interaction model has a significant SM score (LR=130) arising from both local structural similarity and a conserved interface. Figure 3B describes a prediction of an LC set interaction between the elongation factor 1-delta (EF1δ) and the von Hippel-Lindau tumor suppressor (VHL) using the same template as that used in Figure 3A. Again, there is no sequence relationship between the target and the template proteins, and they are classified into different folds. Nevertheless, the interaction model has an LR of 70. We note that the EF1δ and VHL were found to interact using mass spectroscopy[29] and by co-IP experiments reported here (Figure S14).

The exploitation of homology models and of remote structural relationships implies that each new structure that is determined experimentally can be used to detect large numbers of new functional relationships even if the protein in question is of only limited biological interest on its own. In this regard, our approach has benefitted from structural genomics initiatives which produced a large increase in the coverage of sequence families that did not have structural representatives[30]. We note that PrePPI appears in many cases to offer a viable alternative to HT experiments yielding, in addition to a likelihood of a given

interaction, a model (albeit a crude one) of the domains and residues that form the relevant protein-protein interface. This should in turn facilitate the generation of experimentally testable hypotheses as to the presence of a true physical interaction. In conclusion, our study suggests the ability to add a structural "face" for a large number of PPIs and that Structural Biology can play an important role in molecular Systems Biology.

Methods

Proteins and domains

We obtained the yeast proteome from UniProt[31], and parsed its 6,521 proteins into 7,792 domains using the SMART online server[32]. Similarly, for human, we identified 20,318 unique proteome members, producing 49,851 individual domains.

Structures

Structural representatives of the entire protein or different individual domains were either taken directly from the PDB[33], where available, or from the ModBase[10] and SkyBase[11] homology model databases. PDB structures were identified by sequence homology, using a single iteration of PSI-BLAST[34] and an E-value cutoff 0.0001; matching structures in the PDB were required to have >90% sequence identity and cover >80% of the query target (the entire protein or any domain). Homology models were selected based on two criteria: a) an E-value less than 1e-6, or b) an E-value less than 1 and either a structure-based pG score ≥ 0.3, for SkyBase models[35], or a ModPipe protein quality score MPQS ≥ 0.5, for ModBase models. When multiple structures were available for a target/domain we choose only one representative using: a) the PDB structure with the best resolution, if available; b) otherwise, the ModBase model with the highest MPQS score; or c) the SkyBase model with the highest pG score. Based on these criteria, we identified 1,361 PDB structures and 7,222 homology models for 4,193 different yeast proteins. Among these, 627 proteins could be matched to a PDB structure and 3,662 to a homology model, with some proteins having both. For human, 14,132 proteins were matched to 8,582 PDB structures and 30,912 models. Specifically, 4,286 proteins were matched to a PDB structure and 11,266 were matched to a homology model, with some proteins matched to both.

Structural neighbors

We used a structural alignment tool Ska[36] to identify structural neighbors. Ska allows alignments to be considered significant even if only three secondary structural elements are well aligned. At a PSD[37] (protein structure distance) cutoff of 0.6, we identified 1,448 neighbors (both close and remote) per structure for 7,875 structures of 3,911 yeast proteins and 1,553 neighbors per structure for 36,743 structures of 13,545 human proteins.

Template complexes

As of February, 2010, there were about 37,000 protein-protein complexes involving multiple organisms in the PDB and PQS[12] databases. We used 28,408 and 29,012 complexes as templates during our modeling of yeast and human interactions, respectively. PQS terminated updates after Aug. 2009, and has been replaced by the PISA (Protein interfaces, surfaces and assemblies) server [38] which will be used in future work.

Interaction modeling

Given a pair of proteins or domains, we built their interaction model by superimposing their structures with the corresponding structural neighbors in the templates (Figure 1). For yeast, we built 550 million models for 2.4 million potential PPIs, and for human, we built 12 billion models for 36 million potential PPIs. We calculated five structure-based scores for each model (Figure S1) and used a Bayesian network to combine these scores into a likelihood ratio (LR) to evaluate an interaction model (Figure S2) based on the HC and the N reference sets (Table S1).

Non-structural clues

For the yeast proteome, we downloaded the raw data for four different clues; protein essentiality (ES), co-expression (CE), GO[39] similarity and MIPS[40] similarity, from the Gerstein lab (http://networks.gersteinlab.org/intint/supplementary.htm). We also implemented a measure of phylogenetic profile (PP) similarity based on that introduced in reference [41] (see below). We calculate a likelihood ratio (LR) for each non-structure clue based on our HC and N reference sets. For the human proteome, we calculated three different clues following the protocol of Gerstein and colleagues for GO and CE and as described below for PP. For CE, we used the expression dataset (GDS1962), which is one of the most comprehensive microarray studies of 19,803 human genes under 180 different conditions[42], from the Gene Expression Omnibus[43].

Phylogenetic profile (PP) similarity

Similar to Enault et. al.[44], we calculated a continuous score between 0 and 1 to measure the occurrence of a protein and/or domain in 1,156 reference organisms of complete proteome information from UniProt. These scores form a phylogenetic profile vector (PPV), and the Pearson correlation coefficient (PCC) was used to define the similarity between two vectors. For proteins with multiple domains, each domain's PPV is calculated independently, and the highest PCC score of different domain pairs is selected as the similarity score between two proteins. Similarity scores for pairs of proteins/domains with >40% sequence identity and, of course, for homomeric protein/domain pairs were not calculated.

The Naïve Bayes Classifier

We combine the different types of clues with each other and structural modeling into a single Naïve Bayes PPI classifier[24–26]:

$$\mathrm{LR}(c_1, c_2, \cdots, c_n) = \prod_{i=1}^{n} \mathrm{LR}(c_i)$$

10-fold cross validation

We randomly divided the positive and negative reference sets into 10 subsets of equal size. Each time, we used 9 subsets to train the classifier, and obtained the LR for each protein pair, i.e., interaction, in the excluded subset from the trained classifier. We repeated the procedure 10 times using different subsets as training and testing datasets and finally

obtained an LR for each interaction. We counted the number of true positives (predictions in the HC set) and false positives (predictions in the N set) and calculated the prediction TPR (true positive rate) =TP/(TP+FN) and the FPR (false positive rate) =FP/(FP+TN) to plot the receiver operating characteristic (ROC) curves. In all cases, we have removed structural interaction models based on a template that corresponds to an actual crystal structure of the two target proteins.

Comparison with high-throughput (HT) experiments

We retrieved eight HT experiment datasets for yeast and three for human (Table S4). In our comparison, in addition to the HC sets, we also use the same reference interaction sets used in the comparative study of different HT techniques. These include ~1,300 PPIs (CCSB-BGS) and a subset of 188 highly reliable PPIs that are referenced in at least four manuscripts (CCSB-PRS). We compiled a new negative reference set, which consists of 440,000 yeast and 1,750,000 human protein pairs where each protein in a pair is annotated as localized to a different cellular compartment (Figure S7).

New protein interaction dataset

We used 23,779 human protein interactions newly deposited into databases after Aug. 2010 as independent validations of PrePPI predictions, which were based on pre-2010 data (Table S5).

Co-immunoprecipitation in mammalian cells

After 48 hrs post-transfection with indicated expression plasmids, HEK-293T cells were lysed in lysis buffer (20 mM HEPES pH 7.9, 100 mM NaCl, 0.2 mM EDTA, 1.5 mM MgCl2, 10 mM KCl, 20% glycerol and 0.1% Triton-X100 for Fig. S10~S11; 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, and 1% NP-40 for Fig. S12; and 1x Cell Lysis Buffer (Cell Signaling) for Fig S13, respectively) supplemented with Protease Inhibitor Cocktail (Roche). Cell lysates were sonicated and pre-cleared with 30 μL of Protein G Sepharose (GE, Sweden) before incubating with 15 μL anti-Flag M2 or 40 μL anti-HA Affinity Gel (Sigma-Aldrich) overnight at 4 °C with shaking. Agarose beads were washed 4 times with lysis buffer. Lysates (input) and immunoprecipitates were denatured in reducing protein sample buffer and analyzed by SDS-PAGE and immunoblotted (IB) with anti-Flag (Sigma-Aldrich), anti-HA (Roche), anti-PPARγ (Santa Cruz), anti-ABL1 (Santa Cruz), anti-ROR2 (Cell Signaling), or anti-VEGFR2 (abcam) antibodies as indicated.

Protein analysis from brain

Crude membrane fractions were prepared from brains of P0 to P5 wild type mice or *pcdhg*$^{del/del}$ mice provided by Xiaozhong Wang. The brain tissues were homogenized in a buffer A (5 mM Tris-HCl (pH 7.4), 0.32 M sucrose, 1 mM EDTA, 50 mM DTT) supplemented with the complete Protease Inhibitor Cocktail. The nuclei and insoluble debris were collected by a low speed centrifugation at 1000 × g for 10 minutes and subsequently the supernatant was collected by centrifugation at 22,000 × g for 30 minutes. The pellet was washed in the buffer A and solubilized in lysis buffer (Pierce). Crude membrane fraction (supernatant) was collected by centrifugation at 22,000 × g for 20 minutes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by NIH grants GM030518 and GM094597 (BH), CA121852 (AC and BH), DK057539 (DA), CA082683 (TH), R01NS043915 (TM). L.D. thanks CSC (China Scholarship Council) scholarship 2010626059. We thank Ursula Pieper from Andrej Sali's laboratory for help with ModBase, Hunjoong Lee for help with SkyBase.

References

1. Bonetta L. Protein-protein interactions: Interactome under construction. Nature. 2010; 468:851–854. [PubMed: 21150998]
2. Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. Cell. 2011; 144:986–998. [PubMed: 21414488]
3. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Comput Biol. 2007; 3:e42. [PubMed: 17397251]
4. Reguly T, et al. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J Biol. 2006; 5:11. [PubMed: 16762047]
5. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol. 2007; 3:e43. [PubMed: 17465672]
6. Salwinski L, Eisenberg D. Computational methods of analysis of protein-protein interactions. Current Opinion in Structural Biology. 2003; 13:377–382. [PubMed: 12831890]
7. von Mering C, et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 2002; 417:399–403. [PubMed: 12000970]
8. Braun P, et al. An experimentally derived confidence score for binary protein-protein interactions. Nature Methods. 2009; 6:91–97. [PubMed: 19060903]
9. Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics. 2002; 1:349–356. [PubMed: 12118076]
10. Pieper U, et al. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Research. 2006; 34:D291–295. [PubMed: 16381869]
11. Mirkovic N, Li Z, Parnassa A, Murray D. Strategies for high-throughput comparative modeling: applications to leverage analysis in structural genomics and protein family organization. Proteins. 2007; 66:766–777. [PubMed: 17154423]
12. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends in Biochemical Sciences. 1998; 23:358–361. [PubMed: 9787643]
13. Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99:5896–5901. [PubMed: 11972061]
14. Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins. 2002; 49:350–364. [PubMed: 12360525]
15. Davis FP, et al. Protein complex compositions predicted by structural similarity. Nucleic Acids Research. 2006; 34:2943–2952. [PubMed: 16738133]
16. Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O. Architectures and functional coverage of protein-protein interfaces. Journal of Molecular Biology. 2008; 381:785–802. [PubMed: 18620705]
17. Zhang QC, Petrey D, Norel R, Honig BH. Protein interface conservation across structure space. Proc Natl Acad Sci U S A. 2010; 107:10896–10901. [PubMed: 20534496]
18. 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. [PubMed: 21149688]

19. Wass MN, Fuentes G, Pons C, Pazos F, Valencia A. Towards the prediction of protein interaction partners using physical docking. Mol Syst Biol. 2011; 7:469. [PubMed: 21326236]
20. Chen HL, Zhou HX. Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. Proteins-Structure Function and Bioinformatics. 2005; 61:21–35.
21. Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. Nucleic Acids Res. 2006; 34:3698–3707. [PubMed: 16893954]
22. Zhang QC, et al. PredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res. 2011; 39:W283–287. [PubMed: 21609948]
23. Yu H, et al. High-quality binary protein interaction map of the yeast interactome network. Science. 2008; 322:104–110. [PubMed: 18719252]
24. Lefebvre C, et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Molecular Systems Biology. 2010; 6:377. [PubMed: 20531406]
25. Jansen R, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003; 302:449–453. [PubMed: 14564010]
26. von Mering C, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005; 33:D433–437. [PubMed: 15608232]
27. Stolovitzky G, Prill RJ, Califano A. Lessons from the DREAM2 Challenges. Annals of the New York Academy of Sciences. 2009; 1158:159–195. [PubMed: 19348640]
28. Keskin O, Nussinov R, Gursoy A. PRISM: protein-protein interaction prediction by structural matching. Methods in Molecular Biology. 2008; 484:505–521. [PubMed: 18592198]
29. Ewing RM, et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. Molecular Systems Biology. 2007; 3:89. [PubMed: 17353931]
30. Levitt M. Nature of the protein universe. Proc Natl Acad Sci U S A. 2009; 106:11079–11084. [PubMed: 19541617]
31. Apweiler R, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research. 2004; 32:D115–119. [PubMed: 14681372]
32. Letunic I, Doerks T, Bork P. SMART 6: recent updates and new developments. Nucleic Acids Research. 2009; 37:D229–D232. [PubMed: 18978020]
33. Berman HM, et al. The Protein Data Bank. Nucleic Acids Research. 2000; 28:235–242. [PubMed: 10592235]
34. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]
35. Sanchez R, Sali A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci U S A. 1998; 95:13597–13602. [PubMed: 9811845]
36. Petrey D, Honig B. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. Methods Enzymol. 2003; 374:492–509. [PubMed: 14696386]
37. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. Journal of Molecular Biology. 2000; 301:665–678. [PubMed: 10966776]
38. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007; 372:774–797. [PubMed: 17681537]
39. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 2000; 25:25–29. [PubMed: 10802651]
40. Mewes HW, Albermann K, Heumann K, Liebl S, Pfeiffer F. MIPS: a database for protein sequences, homology data and yeast genome information. Nucleic Acids Research. 1997; 25:28–30. [PubMed: 9016498]
41. Huynen M, Snel B, Lathe W 3rd, Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Research. 2000; 10:1204–1210. [PubMed: 10958638]
42. Sun L, et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell. 2006; 9:287–300. [PubMed: 16616334]

43. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. 2011; 39:D1005–1010. [PubMed: 21097893]

44. Enault F, Suhre K, Claverie JM. Phydbac “Gene Function Predictor”: a gene annotation tool based on genomic context analysis. BMC Bioinformatics. 2005; 6:247. [PubMed: 16221304]

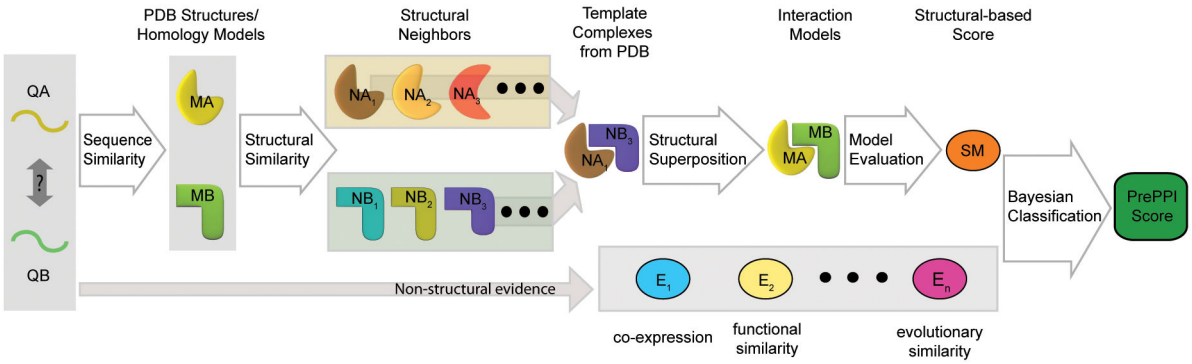

Figure 1. Predicting protein-protein interactions using PrePPI

Given a pair of query proteins that potentially interact (QA, QB), representative structures for the individual subunits (MA, MB) are taken from the PDB, where available, or from homology model databases. For each subunit we find both close and remote structural neighbors. A “template” for the interaction exists whenever a PDB or PQS structure contains a pair of interacting chains (e.g. NA1-NB3) that are structural neighbors of MA and MB, respectively. A model is constructed by superposing the individual subunits, MA and MB, on their corresponding structural neighbors, NA1 and NB3. We assign five empirical structure-based scores to each interaction model (Figure S1) and then calculate a likelihood for each model to represent a true interaction by combining these scores using a Bayesian Network (Figure S2) trained on the HC and the N interaction reference sets. We finally combine the structure-derived score (SM) with non-structural evidence associated with the query proteins (e.g., co-expression, functional similarity) using a naïve Bayesian classifier.

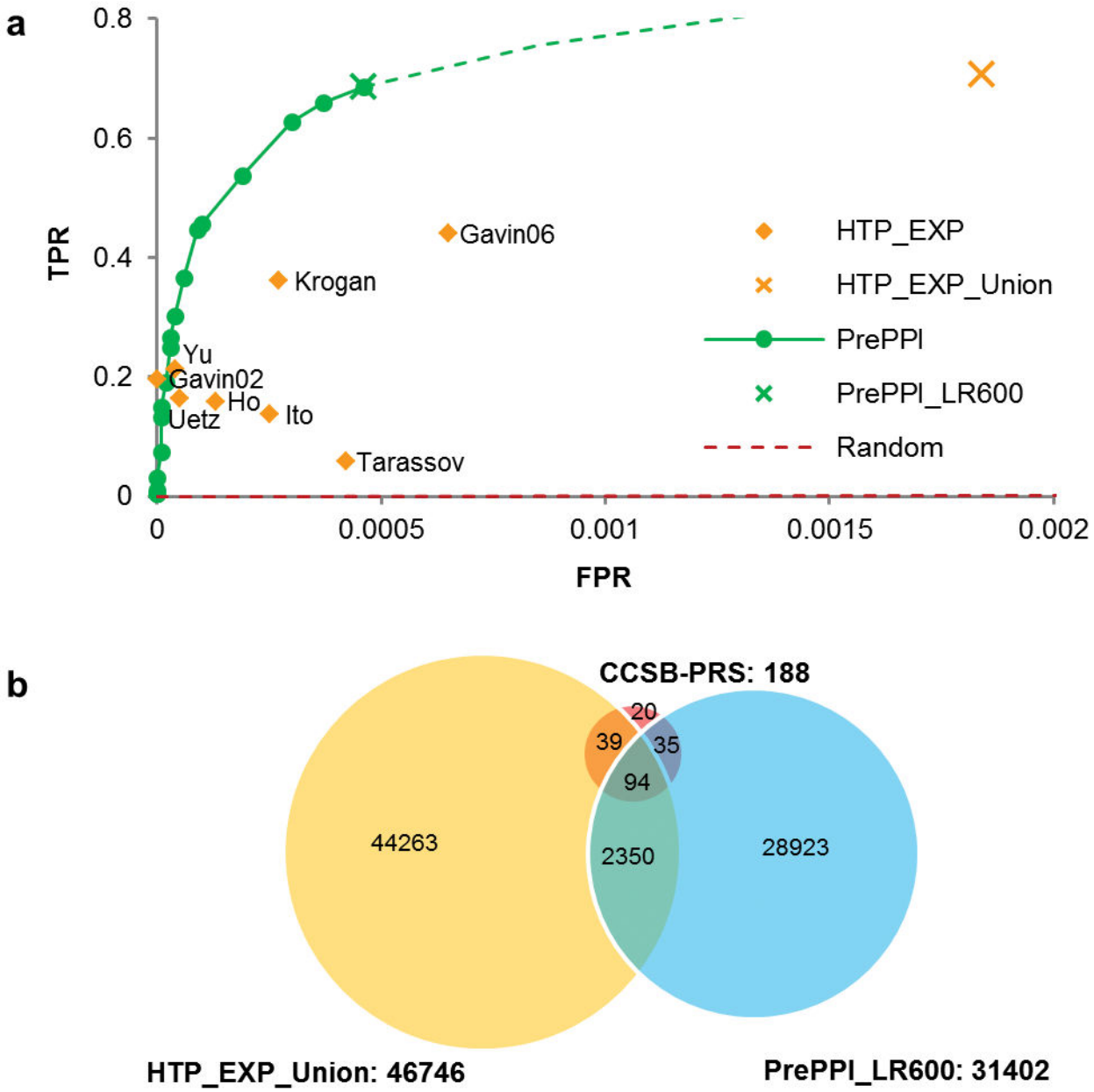

Figure 2. ROC curve (A) and Venn diagram (B) for PrePPI predictions and high-throughput (HT) experiments for yeast

HT experiments are labeled with the first author of the relevant publication (Table S4). The number of interactions in each set is given after the set label in the Venn diagram.

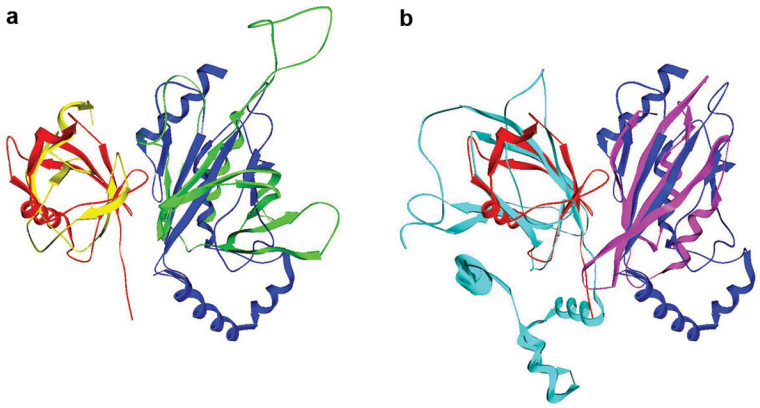

Figure 3. Models for the PPI formed between (A) PKD1 and PKCε, and (B) EF1δ and VHL using homology models and remote structural relationships

The same template complex of ubiquitin-conjugating enzyme E2D 3 and ubiquitin (PDB code: 2fuh A and B chain, shown in blue and red respectively) was used in both cases. The structures of the PH domain of PKD1 and the GNE domain of EF1δ (shown in green and purple) are homology models from ModBase; the structure of a C1 domain of PKCε (yellow) is a homology model from SkyBase; the structure of VHL (cyan) is from PDB (1lm8 V chain). In each case, the relevant homology models are structurally superimposed on one of the two templates in the E2-ubiqutin complex.