LETTER

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[2,3,4], Li Qiang[5], Yu Shi[6], Chan Aye Thu[2], Brygida Bisikirska[3], Celine Lefebvre[3,7], Domenico Accili[5], Tony Hunter[6], Tom Maniatis[2], Andrea Califano[2,3,7,8] & Barry Honig[1,2,3]

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 present knowledge derives from high-throughput techniques such as the yeast two-hybrid assay 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, termed PrePPI, which combines structural information with 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 considerable 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.

Until now, structural information has had relatively little impact in constructing protein–protein interactomes, primarily because there is a marked difference between the number of proteins with known sequences and those with an experimentally known structure. For example, as of early 2010, the Protein Data Bank (PDB) provided structures for ~600 of the total complement of ~6,500 yeast proteins (~10%), whereas the 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 Protein Quaternary Structure[12] (PQS) databases, which might be used as 'templates' to model PPIs. Clearly, if structure is to be useful on a large scale, it is essential that modelling of individual proteins and of complexes be exploited.

A number of studies have used structurally characterized complexes as templates to construct models of the 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]. We searched 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 considerably 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 this work.

Our approach to the prediction of PPIs is embodied in an algorithm we have named PrePPI (predicting protein–protein interactions), which combines structural and non-structural interaction clues using Bayesian statistics (see Fig. 1 and 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 to homology models. We then use structural alignment to find both close and remote structural neighbours (NA_i and NB_j) of MA and MB (an average of ~1,500 neighbours are found for each structure). Whenever two (for example, NA_1 and NB_3) of the over 2 million pairs of neighbours of MA and MB form a complex reported in the PDB, this defines a template for modelling the interaction of QA and QB. Models of the complex are created by superimposing the representative structures on their corresponding structural neighbours in the template (that is, 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 (Supplementary Fig. 1) and that we do not construct a three-dimensional model of each complex because 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 (Supplementary Fig. 1). The first score, SIM, depends on the structural similarity between models of the two query proteins (that is, MA and MB) and those in the template complex (that is, 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 (for example, 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 (for example, 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

[1]Howard Hughes Medical Institute, Columbia University, New York, New York 10032, USA. [2]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, USA. [3]Center for Computational Biology and Bioinformatics, Columbia Initiative in Systems Biology, Columbia University, New York, New York 10032, USA. [4]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China. [5]Naomi Berrie Diabetes Center, Department of Medicine, College of Physicians & Surgeons of Columbia University, New York, New York 10032, USA. [6]Molecular and Cell Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA. [7]Institute of Cancer Genetics, Columbia University, New York, New York 10032, USA. [8]Department of Biomedical Informatics, Columbia University, New York, New York 10032, USA.
*These authors contributed equally to this work.

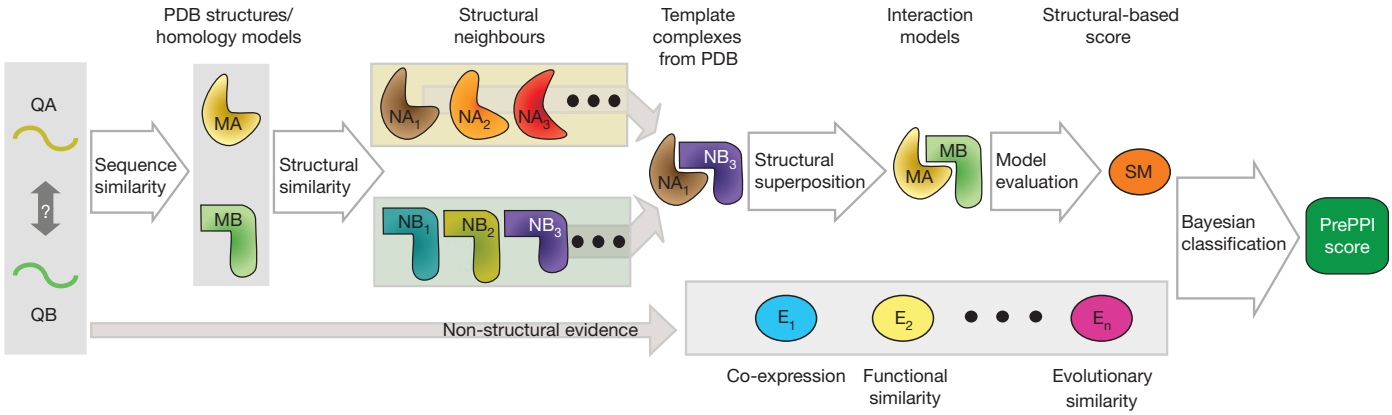

Figure 1 | Predicting protein–protein interactions using PrePPI. Given a pair of query proteins that potentially interact (QA, QB), representative structures for the individual subunits (MA, MB) are taken from the PDB, where available, or from homology model databases. For each subunit we find both close and remote structural neighbours. A 'template' for the interaction exists whenever a PDB or PQS structure contains a pair of interacting chains (for example, NA_1–NB_3) that are structural neighbours of MA and MB, respectively. A model is constructed by superposing the individual subunits, MA and MB, on their corresponding structural neighbours, NA_1 and NB_3. We assign five empirical-structure-based scores to each interaction model (Supplementary Fig. 1) and then calculate a likelihood for each model to represent a true interaction by combining these scores using a Bayesian network (Supplementary Fig. 2) 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 (for example, co-expression, functional similarity) using a naive Bayesian classifier.

individual subunits of the model[20–22]. These scores are calculated as OS, which is 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 (Supplementary Fig. 2) to yield a likelihood ratio (LR) that a candidate protein–protein complex represents a true interaction (see Methods). The network is trained on positive and negative 'gold standard' reference data sets. 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 (Supplementary Table 1); the HC sets contain 11,851 yeast interactions and 7,409 human interactions that 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 plus 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 assess quantitatively the performance of structural modelling (SM), we compared it with a number of non-structural clues previously used to infer PPIs[24–26]: (1) essentiality of the proteins in the interacting pair; (2) co-expression level; (3) gene ontology (GO) functional similarity; (4) Munich Information Centre for Protein Sequences (MIPS) functional similarity; and (5) phylogenetic profile similarity. We used the same algorithms or data for clues 1–4 as previously described[25] but developed our own phylogenetic profile algorithm (for details, see Methods and Supplementary Table 2). Briefly, a phylogenetic profile was constructed for every protein using a set of completely resolved proteomes as references. Because interacting proteins tend to co-evolve, proteins with similar profiles are predicted to interact.

As shown in Supplementary Figs 3 and 4, SM yields comparable performance to other clues over the entire range of false positive rate (FPR) values but is considerably more effective at low FPRs (for example, FPR $\leq$0.1%). This is critical as, owing 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 FPRs, SM by itself outperforms even the naive 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 a FPR of ~0.1%; see Methods), about 70% and 50%, respectively, share GO biological terms at, or more specific than, the sixth level of the GO hierarchy, suggesting that many of these interactions are real (Supplementary Fig. 5).

As mentioned earlier, PrePPI combines structural and non-structural clues using a naive Bayesian network[24–26]. Supplementary Fig. 4 shows that the performance of PrePPI 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 Supplementary Fig. 6. It is evident from the figure that combining structural and non-structural clues yields many more high-confidence predictions and identifies more interactions in the HC set than either source of information alone. As an independent test of PrePPI, we assessed its performance against one of the challenges in the 2009 Dialogue for Reverse Engineering Assessments and Methods (DREAM) workshop specifically aimed at PPI predictions[27]. As discussed in Supplementary Table 3, PrePPI outperformed all other methods for cases where structural information is available.

We compared the performance of PrePPI to that of high-throughput experiments (Supplementary Table 4) using data provided in a detailed comparison of different high-throughput techniques reported previously[23]. We used the data sets in ref. 23 to define true positives, and compiled a new negative reference set that consists of protein pairs in which each protein is annotated as localized to a different cellular compartment (see Supplementary Fig. 7 and Methods). This was essential for comparison to experimental assays because, as constructed, our N set excludes data compiled from high-throughput experiments, and hence the FPR for experimental assays is artificially zero (see also related discussion in supplementary information in ref. 23).

As can be seen in the receiver operating characteristic (ROC) curves reported in Fig. 2a and Supplementary Fig. 8, PrePPI performance is generally comparable, although somewhat better overall, than high-throughput methods for most data sets that were tested. Figure 2b shows a Venn diagram in which the PrePPI data set is based on an LR cutoff of 600 (FPR $\approx$ 0.1%). Results for other LRs and additional reference sets are shown in Supplementary Fig. 9. As can be seen, many of the interactions inferred by PrePPI are different from those identified

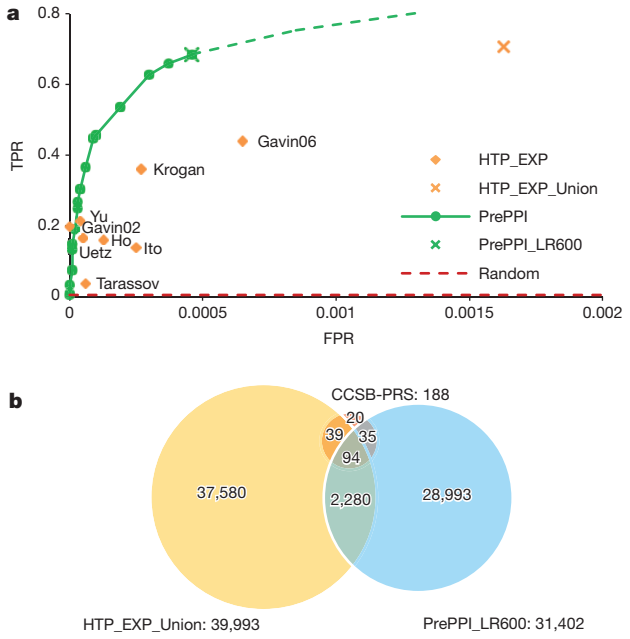

Figure 2 | ROC curve and Venn diagram for PrePPI predictions and high-throughput experiments in yeast. **a**, ROC curve. TPR, true positive rate. **b**, Venn diagram. The CCSB-PRS positive reference interaction set is defined in ref. 23 and described in Methods. High-throughput experiments are labelled with the first author of the relevant publication (Supplementary Table 4). The number of interactions in each set is given after the set label in the Venn diagram.

by high-throughput 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 (Supplementary Table 5). 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 assays, was carried out in four separate laboratories, leading to confirmation of 15 of these interactions (Supplementary Figs 10–14 and Supplementary Table 6). Specifically, the investigators in each laboratory queried the PrePPI database for previously uncharacterized interactions involving proteins of interest and that, as much as possible, had relatively high SM and PrePPI scores (see Supplementary Table 6 for more information).

One set of predictions involves potential PPIs formed between the nuclear receptor peroxisome proliferator-activated receptor γ (PPAR-γ) and other transcription factors. PPAR-γ has a pivotal role in regulating glucose and lipid metabolism, the 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 PPAR-γ and the transcription factors LXR-β (also known as NR1H2), PAX7, PDX1, NKX2-2 and HHEX (Supplementary Table 6). Except for HHEX, all of the interactions were validated (Supplementary Fig. 10). The predicted interaction with nuclear receptor LXR-β 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 so far unrecognized convergence of signalling and metabolic pathways regulated by these two nuclear receptors. The interaction between the ligand-binding domain of PPAR-γ and the homeodomains of PAX7, PDX1 and NKX2-2 are new observations that require further studies, as they suggest that PPAR-γ may have a role in endocrine progenitor and pancreatic β-cell development.

A second set of examples involves suppressor of cytokine signalling (SOCS3), an SH2-domain-containing protein that negatively regulates cytokine-induced signal transduction. Until now, 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 (Supplementary Fig. 11A, B). PrePPI also predicts the formation of a complex between SOCS3 and BTK, a cytoplasmic tyrosine kinase important in B-lymphocyte development, differentiation and signalling, and this interaction was also validated (Supplementary Fig. 11C). 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 indicate that the PrePPI scoring function accounts, at least in part, for the binding preference of closely related protein domains (Supplementary Fig. 15, see also later).

A third group of novel observations involves the identification of kinases that interact with the clustered protocadherin proteins (protocadherin α, β and γ (PCDH-α, -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 signalling adaptors, kinases and phosphatases. Analysis of potential PCDH-kinase PPIs confirmed published interactions between PCDH-α and -γ with the tyrosine kinase RET, and predicted interactions with ROR2, VEGFR2 and ABL1 (Supplementary Table 6 and Supplementary Fig. 12; experiments done in mice). PrePPI predicts that these PPIs are mediated by the extracellular cadherin domains and immunoglobulin (Ig) domains, a result that was confirmed experimentally (Supplementary Fig. 12A–D). A hydrophobic residue, Phe 64, of the ROR2 Ig domain is predicted to be in the centre of the interface it forms with PCDH-α4. Mutating this Phe to an Ala, a smaller hydrophobic residue, has no detectable effect on binding, whereas mutating it to charged residues considerably weakens the interaction (Supplementary Fig. 12B, 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 (Supplementary Fig. 13). 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 modelling, were unanticipated, but should not come as a complete surprise. Most protein complexes in the PDB have structural neighbours 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 modelling and PrePPI. The first is the marked expansion in the number of interactions that can be modelled, owing 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 (versus 36 million when homology models are used) could have been modelled. Similarly, had we limited ourselves to structural neighbours taken from the same Structural Classification of Proteins (SCOP) fold, only ~225 thousand interactions could have been modelled, as opposed to 36 million.

As might be expected, predictions based on structural modelling that use only PDB structures or close structural neighbours are more likely to recover known interactions (defined by their presence in databases) than those that only use homology models or remote structural relationships (Supplementary Fig. 16). However, the latter, on their own, yield a marked 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 led to an even greater expansion in high-confidence predictions when combined with functional, evolutionary and other sources of evidence (Supplementary Fig. 16).

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—that is, specificity—is obtained from the properties of the predicted interface, for example, the statistical propensity of certain amino acids to appear in interfaces[20,21] (and, additionally, from non-structural clues; for example, are the two proteins co-expressed). As examples, our analysis of the SH2 and GTPase families shows that the structural modelling (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 (Supplementary Fig. 15).

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 to make reliable predictions and to improve prediction specificity (Supplementary Figs 15 and 16).

Figure 3 provides two examples of the use of remote structural relationships and homology models. In Fig. 3a, an HC set interaction of serine/threonine-protein kinase D1 (PRKD1) and protein kinase C-ε (PRKCE) is recovered by structural modelling using a complex of two proteins in the ubiquitin pathway (not kinases) as template. Note that PRKD1 and PRKCE 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-δ (EEF1D) and the von Hippel–Lindau tumour suppressor (VHL) using the same template as that used in Fig. 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 EEF1D and VHL were found to interact using mass spectroscopy[29] and by co-immunoprecipitation experiments reported here (Supplementary Fig. 14).

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 high-throughput 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 have an important role in molecular systems biology.

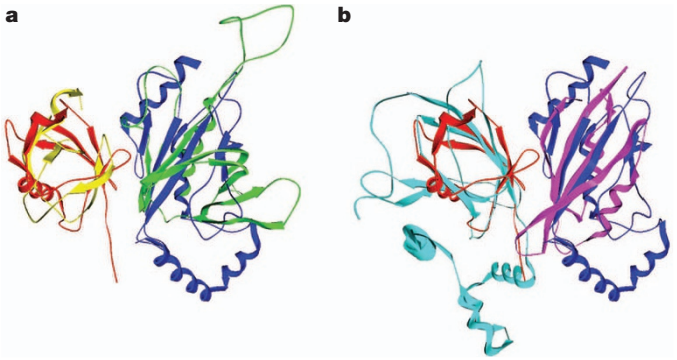

Figure 3 | Models for the PPI formed between PRKD1 and PRKCE, and EEF1D and VHL using homology models and remote structural relationships. **a**, Model for PRKD1 and PRKCE. **b**, Model for EEF1D and VHL. The same template complex of ubiquitin-conjugating enzyme E2D 3 (UBE2D3) and ubiquitin (PDB accession: 2FUH; A and B chain, shown in blue and red, respectively) was used in both cases. The structures of the PH domain of PRKD1 and the GNE domain of EEF1D (shown in green and purple) are homology models from ModBase; the structure of a C1 domain of PRKCE (yellow) is a homology model from SkyBase; the structure of VHL (cyan) is from PDB (accession 1LM8; V chain). In each case, the relevant homology models are structurally superimposed on one of the two templates in the UBE2D3–ubiquitin complex.

METHODS SUMMARY

Details of the PrePPI algorithm, protein datasets, and experimental validations are available in Methods.

Full Methods and any associated references are available in the online version of the paper.

Received 29 June 2011; accepted 10 August 2012.
Published online 30 September 2012.

1. Bonetta, L. Protein–protein interactions: interactome under construction. *Nature* **468,** 851–854 (2010).
2. Vidal, M., Cusick, M. E. & Barabasi, A. L. Interactome networks and human disease. *Cell* **144,** 986–998 (2011).
3. Shoemaker, B. A. & Panchenko, A. R. Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PLOS Comput. Biol.* **3,** e42 (2007).
4. Reguly, T. *et al.* Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* **5,** 11 (2006).
5. Shoemaker, B. A. & Panchenko, A. R. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLOS Comput. Biol.* **3,** e43 (2007).
6. Salwinski, L. & Eisenberg, D. Computational methods of analysis of protein–protein interactions. *Curr. Opin. Struct. Biol.* **13,** 377–382 (2003).
7. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417,** 399–403 (2002).
8. Braun, P. *et al.* An experimentally derived confidence score for binary protein–protein interactions. *Nature Methods* **6,** 91–97 (2009).
9. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* **1,** 349–356 (2002).
10. Pieper, U. *et al.* MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **34,** D291–D295 (2006).
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* **66,** 766–777 (2007).
12. Henrick, K. & Thornton, J. M. PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23,** 358–361 (1998).

13. Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA* **99,** 5896–5901 (2002).
14. Lu, L., Lu, H. & Skolnick, J. MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. *Proteins* **49,** 350–364 (2002).
15. Davis, F. P. *et al.* Protein complex compositions predicted by structural similarity. *Nucleic Acids Res.* **34,** 2943–2952 (2006).
16. Tuncbag, N., Gursoy, A., Guney, E., Nussinov, R. & Keskin, O. Architectures and functional coverage of protein–protein interfaces. *J. Mol. Biol.* **381,** 785–802 (2008).
17. Zhang, Q. C., Petrey, D., Norel, R. & Honig, B. H. Protein interface conservation across structure space. *Proc. Natl Acad. Sci. USA* **107,** 10896–10901 (2010).
18. Gao, M. & Skolnick, J. Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proc. Natl Acad. Sci. USA* **107,** 22517–22522 (2010).
19. Wass, M. N., Fuentes, G., Pons, C., Pazos, F. & Valencia, A. Towards the prediction of protein interaction partners using physical docking. *Mol. Syst. Biol.* **7,** 469 (2011).
20. Chen, H. L. & Zhou, H. X. Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins* **61,** 21–35 (2005).
21. Liang, S., Zhang, C., Liu, S. & Zhou, Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.* **34,** 3698–3707 (2006).
22. Zhang, Q. C. *et al.* PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res.* **39,** 283–287 (2011).
23. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322,** 104–110 (2008).
24. Lefebvre, C. *et al.* A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.* **6,** 377 (2010).
25. Jansen, R. *et al.* A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302,** 449–453 (2003).
26. von Mering, C. *et al.* STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33,** D433–D437 (2005).
27. Stolovitzky, G., Prill, R. J. & Califano, A. Lessons from the DREAM2 challenges. *Ann. NY Acad. Sci.* **1158,** 159–195 (2009).
28. Keskin, O., Nussinov, R. & Gursoy, A. PRISM: protein–protein interaction prediction by structural matching. *Methods Mol. Biol.* **484,** 505–521 (2008).
29. Ewing, R. M. *et al.* Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3,** 89 (2007).
30. Levitt, M. Nature of the protein universe. *Proc. Natl Acad. Sci. USA* **106,** 11079–11084 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work is supported by National Institutes of Health grants GM030518 and GM094597 (B.H.), CA121852 (A.C. and B.H.), DK057539 (D.A.), CA082683 (T.H.), R01NS043915 (T.M.). L.D. thanks the China Scholarship Council scholarship 2010626059. We thank U. Pieper from A. Sali's laboratory for help with ModBase, and H. Lee for help with SkyBase.

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. analysed the data; L.D. set up the PrePPI web server, L.Q., Y.S., C.A.T. and B.B. performed co-immunoprecipitation 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.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.C. (califano@c2b2.columbia.edu) or B.H. (bh6@columbia.edu).

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: (1) an *E* value less than 1×10^{-6}; or (2) 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 chose only one representative using: (1) the PDB structure with the best resolution, if available; (2) the ModBase model with the highest MPQS score; or (3) the SkyBase model with the highest pG score. On the basis of 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 neighbours. We used structural alignment tool Ska[36] to identify structural neighbours. Ska allows alignments to be considered significant even if only three secondary structural elements are well aligned. At a protein structure distance[37] (PSD) cutoff of 0.6, we identified 1,448 neighbours (both close and remote) per structure for 7,875 structures of 3,911 yeast proteins, and 1,553 neighbours 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 modelling of yeast and human interactions, respectively. PQS terminated updates after August 2009, and has been replaced by the protein interfaces, surfaces and assemblies (PISA) server[38], which will be used in future work.

Interaction modelling. Given a pair of proteins or domains, we built their interaction model by superimposing their structures with the corresponding structural neighbours in the templates (Fig. 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 (Supplementary Fig. 1) and used a Bayesian network to combine these scores into an LR to evaluate an interaction model (Supplementary Fig. 2) based on the HC and the N reference sets (Supplementary Table 1).

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 laboratory (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 later). We calculate an 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 previously described[25] for GO and CE, and as described later for PP. For CE, we used the expression data set (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 similarity. Using a similar method to that previously described[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 naive Bayes classifier. We combine the different types of clues with each other and structural modelling into a single naive Bayes PPI classifier[24–26]:

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

Tenfold cross validation. We randomly divided the positive and negative reference sets into ten subsets of equal size. Each time, we used nine subsets to train the classifier, and obtained the LR for each protein pair, that is, interaction, in the excluded subset from the trained classifier. We repeated the procedure ten times using different subsets as training and testing data sets 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 = TP/(TP+FN), and the FPR (false positive rate) = FP/(FP+TN), to plot the 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 experiments. We retrieved eight high-throughput experiment data sets for yeast and three for human (Supplementary Table 4). In our comparison, in addition to the HC sets, we also use the same reference interaction sets used in the comparative study of different high-throughput 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 in which each protein in a pair is annotated as localized to a different cellular compartment (Supplementary Fig. 7).

New protein interaction data set. We used 23,779 human protein interactions newly deposited into databases after August 2010 as independent validations of PrePPI predictions, which were based on pre-2010 data (Supplementary Table 5).

Co-immunoprecipitation in mammalian cells. Forty-eight hours after 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 $MgCl_2$, 10 mM KCl, 20% glycerol and 0.1% Triton-X100 for Supplementary Figs 10 and 11; 20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA and 1% NP-40 for Supplementary Fig. 12; and 1× Cell Lysis Buffer (Cell Signaling) for Supplementary Fig. 13) supplemented with Protease Inhibitor Cocktail (Roche). Cell lysates were sonicated and pre-cleared with 30 µl of Protein G Sepharose (GE) 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 four times with lysis buffer. Lysates (input) and immunoprecipitates were denatured in reducing protein sample buffer, analysed by SDS–PAGE and immunoblotted with anti-Flag (Sigma-Aldrich), anti-HA (Roche), anti-PPAR-γ (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 postnatal day (P)0 to P5 wild-type mice or *Pcdhg*$^{del/del}$ mice provided by X. 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 dithiothreitol) supplemented with the Complete Protease Inhibitor Cocktail. The nuclei and insoluble debris were collected by a low-speed centrifugation at 1,000*g* for 10 min and subsequently the supernatant was collected by centrifugation at 22,000*g* for 30 min. 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 min.

31. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32,** D115–D119 (2004).
32. Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res.* **37,** D229–D232 (2009).
33. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
34. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
35. Sanchez, R. & Sali, A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA* **95,** 13597–13602 (1998).
36. Petrey, D. & Honig, B. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.* **374,** 492–509 (2003).
37. Yang, A. S. & 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. *J. Mol. Biol.* **301,** 665–678 (2000).
38. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372,** 774–797 (2007).
39. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
40. Mewes, H. W., Albermann, K., Heumann, K., Liebl, S. & Pfeiffer, F. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.* **25,** 28–30 (1997).
41. Huynen, M., Snel, B., Lathe, W. III & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10,** 1204–1210 (2000).
42. Sun, L. *et al.* Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* **9,** 287–300 (2006).
43. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.* **39,** D1005–D1010 (2011).
44. Enault, F., Suhre, K. & Claverie, J. M. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. *BMC Bioinformatics* **6,** 247 (2005).

CORRECTIONS & AMENDMENTS

CORRIGENDUM

doi:10.1038/nature11977

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

Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng, Li Qiang, Yu Shi, Chan Aye Thu, Brygida Bisikirska, Celine Lefebvre, Domenico Accili, Tony Hunter, Tom Maniatis, Andrea Califano & Barry Honig

Nature **490,** 556–560 (2012); doi:10.1038/nature11503

In this Letter, one of the points shown in Fig. 2 and Supplementary Figs 8, 9 and Supplementary Table 4 reflects the presence of interactions that had been erroneously deposited from a previous publication[1] into the IntAct database. We have now used the MINT database to retrieve these interactions, and Fig. 2 is corrected here (shown below as Fig. 1). The error in IntAct was corrected on 9 November 2012 in consultation with the original authors of the paper. We thank S. Michnick for bringing this to our attention. We also thank M. Maletta for pointing out that Supplementary Fig. 10C was mislabelled and erroneously indicated that NKX2-2 protein was not included in the experiment. See Supplementary Information to the original paper for corrected versions of Supplementary Figs 8–10C and Supplementary Table 4. These errors do not affect the results or conclusion of the paper, and have been corrected in the HTML and PDF of the original paper.

1. Taeassov, K. *et al.* An *in vivo* map of the yeast protein interactome. *Science* **320,** 1465–1470 (2008).

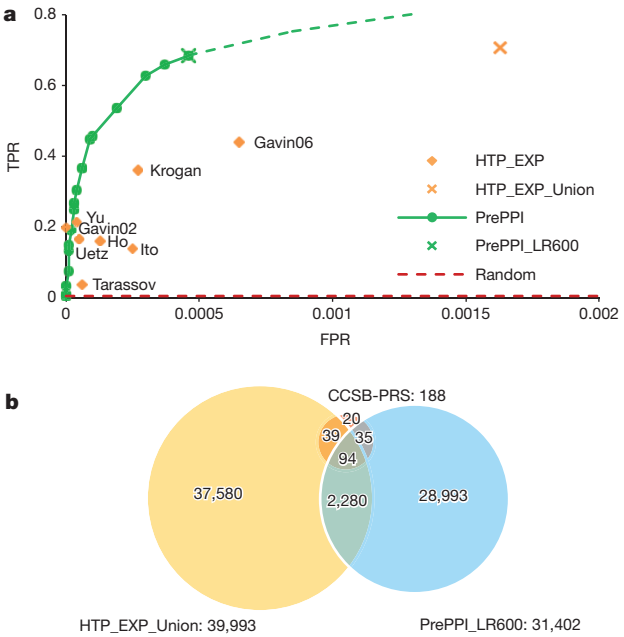

Figure 1 | This is the corrected Fig. 2.