Published in final edited form as:
J Mol Biol. 2011 February 4; 405(5): 1295–1310. doi:10.1016/j.jmb.2010.11.025.

iWRAP: An interface threading approach with application to prediction of cancer related protein-protein interactions

R. Hosur[1,5], **J. Xu**[1,2], **J. Bienkowska**[1,3,*], and **B. Berger**[1,4,*]

[1] Computer Science and Artificial Intelligence Laboratory, MIT.

[2] Toyota Technological Institute at Chicago.

[3] Computational Biology Group, BiogenIdec.

[4] Department of Mathematics, MIT.

[5] Department of Materials Science and Engineering, MIT.

Abstract

Current homology modeling methods for predicting protein-protein interactions (PPIs) have difficulty in the “twilight zone” (<40%) of sequence identities. Threading methods extend coverage further into the twilight zone by aligning primary sequences for a pair of proteins to a best-fit template complex to predict an entire three-dimensional structure. We introduce a threading approach, iWRAP, which focuses on only the protein interface. Our approach combines a novel linear programming formulation for interface alignment with a boosting classifier for interaction prediction. We demonstrate its efficacy on SCOPPI, a classification of PPIs in the Protein Databank, and on the entire yeast genome. iWRAP provides significantly improved prediction of PPIs and their interfaces in stringent cross-validation on SCOPPI. Furthermore, by combining our predictions with a full-complex threader, we achieve coverage of 13% for the yeast PPIs, which is close to a 50% increase over previous methods at a higher sensitivity. As an application, we effectively combine iWRAP with genomic data to identify novel cancer related genes involved in chromatin remodeling, nucleosome organization and ribonuclear complex assembly. iWRAP is available at http://iwrap.csail.mit.edu.

Keywords

structural bioinformatics; protein-protein interactions; threading; cancer; genome annotation

1. Introduction

Protein-protein interactions (PPIs) play a central role in all biological processes. Akin to the complete sequencing of genomes, complete description of interactomes is a fundamental step towards a deeper understanding of biological processes, and has a vast potential to

*Corresponding author: bab@mit.edu, jbienkowska@gmail.com 32-G574, 32 Vassar Street, Cambridge, MA - 02139, PH: 617-253-1827, FAX:617-258-5429.

Conflict of Interest: None declared.

impact systems biology, genomics, molecular biology and therapeutics. Although high-throughput biochemical approaches for discovering PPIs have proven very successful[1,2,3,4], the coverage of experimentally determined PPI data remains poor (Table S1) and is prone to errors[5,6]. Such low coverage is partly because the set of possible PPIs to be verified is so large (50 million for a species with 10,000 genes) that any exhaustive experimental verification will take a long time, even with high-throughput techniques. While the rate of PPI discovery has leveled off in recent years (see Fig S1), the number of solved protein structural complexes has rapidly grown: there has been a 40% increase in the number of complex templates in the 14 months between the two versions of Structural Classification of Proteins database (SCOP, 1.65 and 1.69)[7]. This growing resource of structural data presents an opportunity to utilize this information for accurate PPI predictions.

There have recently been proposals to harness the information provided by structure-based computational approaches as a potentially high-quality, high-coverage data source for large-scale integrative approaches to interactome construction[8,9,10,11,12]. Prieto, Las and Rivas[13] have reviewed publicly available interaction databases of known structural data that facilitate analysis of PPIs[14,15,16]. In the absence of a solved structure for a pair of protein “query” sequences, structure-based approaches typically rely on aligning the query sequences to either sequence or structure-based “templates” for solved structures in the Protein Data Bank (PDB)[17].

In one such approach, homology modeling, two protein sequences are assumed to interact based simply on their primary sequence homology to known interacting proteins. Homology modeling has had considerable success at predicting PPIs on a genome scale[11,18,19,20] and reconstructing and predicting 3D multi-protein complexes[9]. More recently, Fukuhara and Kawabata have described HOMCOS[21,22], a web-server that performs a similar task to Aloy and Russell's InterPrets[9], again by homology modeling. MODBase is a database of homology models for protein complexes that have sequence similarity to known structures higher than 50%[23]. ADAN is a specialized database for prediction of protein-protein interactions mediated by linear motifs and utilizes position-specific matrices to assess putative interactions[24]. Other sequence-based methods utilize genetic information and multiple sequence alignments to predict specific protein-protein interactions[25,26,27,28]. However, effective use of homology modeling requires relatively high sequence similarity between the query and template protein-pairs[8].

In another popular approach, threading, the three-dimensional (3D) structure for a pair of protein query sequences is predicted by aligning their sequences to templates, based on both sequence and structure profiles, for complexes in the PDB to see if a similar structure can be found. The goodness of a query pair-template alignment is evaluated using a scoring function. The essential computational components of a PPI threading approach are: template construction, alignment of query sequences to templates, and interaction scoring. Lu et al. developed Multiprospector[29], a threading algorithm that constructs statistical potential functions to evaluate potential PPIs[30]. Singh, Xu & Berger further proposed a machine-learning based threading algorithm DBLRAP, which also performs full complex threading, and demonstrated its superiority in predicting PPIs over homology modeling and Multiprospector[8,31]. Threading identifies compatible structures for proteins that share less sequence similarity with the template; thus typically widening the range of proteins for which predictions can be made over homology modeling.

While homology modeling/threading approaches work well and have good overall accuracy when sequences are somewhat similar to their putative templates, they perform poorly in the “twilight zone” of sequence identities. In particular, they often give inaccurate alignments in

the putative interaction regions for sequences with low similarity and therefore are unable to predict interactions accurately in such cases, which we demonstrated previously for the special case of cytokines[32]. It has been observed that functional residues such as those at the interface are more conserved than non-functional ones, both in sequence[33,34,35] and structure[36,37]. Furthermore, it has been shown just recently that partial homology models, based only on interface alignments, are good candidates for templates used in docking studies[38]. Here we capitalize on these observations by performing threading on only the protein-protein interface after a suitable complex template is identified.

We introduce the program iWRAP (Interface Weighted RAPtor), which predicts whether two proteins interact by combining a novel linear programming approach for interface alignment with a boosting classifier[39] for interaction prediction. iWRAP simultaneously optimizes contacts in query sequences to templates of protein-protein interfaces, after constraining alignments to only those residues likely to be involved in the interaction. This approach is in contrast to existing threading approaches that align each sequence individually to an entire protein structure in the complex. We recently demonstrated the utility of interface threading on two cytokine receptor families by implementing LTHREADER[32], where we manually generated templates specific to this family and aligned each query sequence separately to each template. The driving hypothesis of iWRAP's approach is that more accurate prediction of protein-protein interfaces improves predictions of protein-protein interactions. We show in this paper for general PPIs that (i) more accurate interface alignments lead to improved interface contact prediction, which in turn (ii) significantly improves PPI prediction. Thus, by optimizing the interface alignments after identifying a suitable template, iWRAP exploits functional conservation at the interface to predict PPIs.

We demonstrate the efficacy of these techniques on two datasets, SCOPPI, a database that classifies protein complexes in the PDB[40], and the yeast genome. First, we use SCOPPI as our gold standard database to confirm hypothesis (i): we show that interface threading, i.e. localized threading, leads to better interface contact prediction over full-complex threaders. For difficult alignment problems and a range of sequence identity values less than 40%, iWRAP outperforms standard threading and sequence-based methods, while for easier problems the methods are comparable. Our results on the full yeast genome scan address hypothesis (ii): we demonstate that our method, which novelly uses boosting[39] to classify iWRAP's interface threading scores for PPI prediction, outperforms methods based on whole-sequence alignments. In particular, we perform a full genome scan of yeast to predict interactions, and compare iWRAP's performance on experimental data to DBLRAP, which has been shown to have the best performance amongst available structure-based PPI prediction methods[8,31].

As an application, through mapping of yeast cancer related genes and their putative interactions to the human genome, we identify interactions enriched relative to a recent yeast genetic interaction set[41]. We find that these interacting genes are involved in chromatin remodeling, ribonuclear complex assembly and nucleosome organization[42]; processes known to be critically involved in cancer. We focus on yeast cancer related genes and putative interactions since the function and interactions of yeast genes are much better understood than human genes[43]. Moreover, the malignant behavior of human cells is often caused by dis-regulation of cell cycle, growth and apoptosis processes that are conserved across eukaryotic organisms at the level of genes and their interactions[44].

iWRAP's predictions are made publicly available at its website so that they can be used for further exploration or systems-level integrative approaches.

2. Results

2.1. Overview of the threading algorithm

We develop iWRAP, an algorithm for threading query sequence pairs to only the interface of a suitable complex template. Figure 1 is a schematic of iWRAP, displaying a flowchart of the various stages of the algorithm. In the first stage, template construction, from alignments of multiple protein-protein interfaces[36], we construct specific interface profiles based on amino acid propensities, secondary structure and solvent accessibilities for discrete environmental classes of the interface.

In the second stage, alignment of a query sequence pair to a template, we utilize a profile-scoring scheme that captures amino acid sequence propensities and predicted secondary structure for the query sequences. We first identify a suitable template using a single domain threader- RAPTOR[45] (also see *PPI Prediction: yeast genome*). RAPTOR is used for whole genome scans of pairs of proteins to identify structures most compatible with each protein sequence. For each protein, we select ten top-scoring single domain structures with a threading z-score of at least 3. We then rank the complex templates composed of these single domains based on the sum of their single-domain threading z-scores. When only one sequence of the query pair matches a domain in the complex, we do not discard it. This procedure selects for each query pair at most 10 possible complex templates for threading of the interface by iWRAP. For each of these selected complex templates, iWRAP uses a local alignment of the query sequence profile to the interface template profile; this directly reflects the quality of the interface alignment, without being influenced by alignments elsewhere in the structure. We select the best interface template using a z-score that evaluates iWRAP's interface score with respect to a distribution obtained by randomizing the interface contacts.

For the third stage, scoring the putative interaction, we begin by integrating stage 2's interface-specific alignment score into a general threading scoring scheme implemented similar to RAPTOR[45]. This produces an initial contact map, which we further refine through contact map optimization in the neighborhood of interacting residues. For the fourth stage, interaction prediction, we extract features of the predicted interface (e.g. interface energy, z-score, size) to input into a boosting classifier, which then computes a probability of interaction for the two query proteins. Note that this stage is employed only for our yeast genome scans, and not for our benchmarking tests on SCOPPI. See Materials and Methods for a more detailed description of each of these stages and training and test sets.

Our algorithm builds upon our previous work LTHREADER[32], where we have shown that supervised construction of the interface templates, along with a localized scoring scheme based on sequence-specific profiles significantly improves alignment and prediction accuracies for the cytokine family. LTHREADER independently aligned each sequence to a profile representing one sequence of the interface template using a sliding-window approach. In contrast, iWRAP uses a linear programming approach (LP) to align pairs of sequences to a two-dimensional (2D) profile of a protein-protein interface and utilizes pairwise quasi-chemical scores for evaluation and optimization. Additionally, LTHREADER focused the alignments on putative interaction cores determined by predicted secondary structure, while iWRAP does not make such an assumption; it uses the LP to decide the optimal interface region. iWRAP further optimizes an objective function based on the Hadamard product of 2D contact maps, thereby simultaneously adjusting interface residues of both interacting proteins. iWRAP rigorously deals with gaps in the alignment, whereas LTHREADER aligns the entire putative interaction core to the interface profile ignoring gaps altogether. Moreover, interface templates used by iWRAP are constructed by a fully-automated procedure that uses our recent multiple interface alignment algorithm

CMAPi[36], while LTHREADER had to rely on time-consuming manually-constructed multiple interface alignments. In particular, LTHREADER chose parameters in its alignment algorithm to reflect the structural and physical constraints of the two cytokine families it was tested on. Extension of LTHREADER to other families would require the estimation of those parameters in a principled way. A detailed description of the nontrivial task of interface template construction from the CMAPi alignments is provided in Materials and Methods: *Template construction*. Finally, the combination of iWRAP's interface threading with a general single-domain threader (RAPTOR), the latter of which is used to identify most likely complexes for pairwise threading, allows PPI prediction on a genomic scale – a feature missing in LTHREADER.

2.2. Interface validation

We evaluate iWRAP on two challenges that one encounters using structural information to predict likely protein-protein interactions: sequence-interface alignment and interface contact prediction. For sequence-interface alignments, we first compare the performance of iWRAP with that of a full complex threader, DBLRAP[8], a profile-based alignment program MUSCLE[46] and our previous algorithm LTHREADER, in stringent cross-validation on SCOPPI. We then continue to compare the two superior alignment algorithms, iWRAP and DBLRAP, using several additional metrics that evaluate the absolute quality of the putative interface: Root Mean Square Deviation (RMSD) of the interface alignments, contact accuracy and interfacial energy (Definitions in SI). See Materials and Methods for a detailed description of the training and test set construction. We emphasize that in cross-validation tests, we restrict ourselves to only difficult alignments (i.e. sequence identity < 40%) because easier alignments are straightforward to address using conventional threading techniques or sequence alignment.

Cross-validation within SCOPPI families—iWRAP performs better than or competitive to other sequence and structure-based techniques in terms of average alignment accuracies (see Table 1). Average alignment accuracies are calculated by averaging the alignment accuracies computed by threading the test sequence pair to each template in the training set. iWRAP improves average alignment accuracies for roughly 80% of the families (in cross-validation tests) for which we can construct multiple interface alignments and sufficiently large training and test sets. For the remaining 20% of families, iWRAP gives equivalent or slightly lower accuracies than DBLRAP. iWRAP performs much better than techniques based on sequence alone. We compared iWRAP with profile-based alignments computed using a state-of-the-art alignment program MUSCLE[46]. Profiles for the sequences were computed by running PSI-BLAST for 5 iterations with an E-value cutoff of 0.001 against the ‘nr’ protein database[47]. Profile-based alignments, rather than pairwise alignments, were used as they have been shown to be more accurate for remote homology detection[48]. iWRAP also performs much better than our earlier algorithm LTHREADER. To evaluate the additional value of iWRAP scoring function, we used our new interface profiles along with the threading approach employed by LTHREADER. Briefly, we first align the secondary structure tags of the query and template to roughly identify the interaction cores. Then we use predicted secondary structure and predicted solvent accessibilities in a scoring function similar to LTHREADER, confining the search space to within 5 residues of the secondary structure identified as the putative interaction core. In the three cases where iWRAP performs worse than any of the three previous methods, the overall sequence similarity is rather high giving these methods a slight advantage. Following on this observation, for whole genome scans, we combine DBLRAP with iWRAP.

Interfaces predicted by iWRAP are closer to true interfaces than those predicted by DBLRAP. Below we focus on comparing iWRAP and DBLRAP, since their contact

accuracies are much better than that of MUSCLE and LTHREADER (see SI Fig S4B,C). As an example, Figure 2 illustrates the case of the interface formed in the PDB structure 1upc (Fig 2A) between chains A(12-195) and B(375-573). The template used for threading these two sequences is shown in Fig 2B, with the interface residues highlighted in green. DBLRAP completely misses the correct interface region as a result of poor alignment of chain B (Fig 2C), giving a contact accuracy of 0%. In contrast, iWRAP produces an initial interface closer to the true one, with a contact accuracy of 27% (Fig 2D). On further refinement of the contact map (see Materials and Methods: *Contact map optimization*), iWRAP's predicted interface (Fig 2E) is much closer to the true interface (Fig 2A), with 46% contact accuracy. The predicted structure of the true interface is shown in Fig 2F. It was constructed by mapping true interface residues (magenta, Fig 2A) to the template (Fig 2B) using alignments computed by iWRAP. iWRAP aligns the true interface residues to the interface of the template and is thus able to correctly identify the interacting residues. To emphasize the fact that iWRAP is an interface threading approach, rather than a full-complex approach, the rest of the structure is colored in light-gray. Additionally, the higher statistical significance of iWRAP's predicted interface energy (z-score=2.7), calculated by randomizing the interfacial contacts, as compared to DBLRAP's (z-score=-0.1), is further indicative of the improved interface prediction. The higher contact accuracies and associated z-scores enable iWRAP to improve PPI prediction over DBLRAP. A second example demonstrating iWRAP's improved interface prediction over DBLRAP is given in SI Fig S3.

More generally, iWRAP outperforms other sequence-based and threading methods at correctly predicting interfacial contacts across all template-query pairs in the test set, except for a few very small interfaces (see Fig 3A). We find that iWRAP improves over DBLRAP in predicting interfacial contacts when the number of true contacts is greater than 25-30 (see Fig 3A, right of the solid vertical line). Even when DBLRAP fails to account for 10% of the contacts, iWRAP can predict 20-30% of the contacts (see SI Fig S4A).

We investigated the variation of contact accuracy with sequence similarity at the interface for the alignments in the cross-validation set. For sequence identities between 0.2 and 0.4, iWRAP significantly improves contact prediction (Fig 3B, right of the solid vertical line). However, when the sequence identity between the template and query becomes less than 0.15, there is no consistent improvement over DBLRAP (Fig 3B, left of the solid vertical line). We have also observed that other features of the interface, namely information content and iracc (see Materials and Methods: *Training and test sets*), do not significantly influence the contact predictions (see SI Fig S4D and S4E).

We sought to further investigate iWRAP's superior performance on medium to large contact maps (>25 contacts). We hypothesize this improvement is due to the localized character of our interface profiles. We evaluated the contact density for both methods on contact maps with greater than 25 contacts, where we presume iWRAP's profiles are aiding in its superior performance (Fig 3A). Following the contact-map mining techniques of Hu et al.[49], we characterized each contact by the pattern of contacts in a 5×5 residue neighborhood around it, where the average density is the number of contacts divided by 25. We observe that iWRAP contact predictions have a higher density (0.26) on average than DBLRAP predictions (0.22), on both the training and test sets (see SI for details). Furthermore, when the interface is small, there are many feasible alignments for the interface region; this makes it difficult for iWRAP to get accurate alignments without using restraints from the whole complex. Based on this analysis, we conclude that size and density are factors in the improved performance of iWRAP, and thus may be responsible for the decreased performance in the case of fewer than 20-25 contacts.

iWRAP consistently gives lower interface energies (normalized by the number of predicted contacts) as compared to DBLRAP (Fig 3C). To predict protein interactions iWRAP and DBLRAP use the residue-level statistical potential developed by Lu et al.[30] to score putative interactions. The interaction score (energy) is obtained by summing over all the contacts in the putative interface.

We also evaluated alignments using the conventional metric of interface RMSD and confirmed that iWRAP alignments have similar or lower RMSD than DBLRAP's (see Fig 3D). Thus iWRAP improvements in alignment and contact accuracy do not affect the RMSD of the predicted interface. Note that while optimizing the parameters, RMSD was not optimized for the threading alignments.

Cross-validation across SCOPPI families—In addition to cross-validation tests within the same SCOPPI family we have tested the ability of iWRAP to accurately predict interfaces when threaded complexes are from SCOPPI family pairs sharing only one SCOP family (e.g. **b.47.1.2**_g.3.15.1 and **b.47.1.2**_g.68.1.1). For these across-family threading tests, we restricted ourselves to alignments having a high iracc score (> 0.75, see Materials and Methods: *Training and test sets*), thereby ensuring similar binding patterns. Successful threading of across-family pairs allows us to address PPI predictions when a template complex for the same SCOPPI family does not exist. However, in such cases, it is possible that the interaction can be predicted using a similar interface for another PPI. It is known that despite lack of overall structural similarity some proteins interact with different protein partner using a very similar interface; for example, interaction mimicry has been observed in host-pathogen interactions[50].

Most threading methods rely on a template database, which might not be completely representative and might not have an appropriate template for every query sequence. While traditional cross-validation strategies do not perform across-family tests, we do so in order to try to address the problem of the limited number of templates available for genome-wide PPI predictions.

For across-family predictions iWRAP predicts the interacting residues more accurately than DBLRAP for 75% of SCOPPI family pairs (see SI Table S2) in the cross-validation test. Despite the high iracc score (>0.75) for such alignments, the binding patterns might be relatively different, leading to a poorer overall prediction by DBLRAP. However, for cases when DBLRAP fails to predict even 10% of contacts, iWRAP can account for nearly 20-30% of the true contacts (see Fig 3E). This suggests that using iWRAP for PPI prediction with templates of complexes sharing one SCOP family can increase the coverage of predictions.

2.3. PPI Prediction: yeast genome

We have applied iWRAP for genome-scale analysis to predict the yeast interaction network. In cross-validation tests above, we used templates in the training set to thread query sequences in the test set. For the yeast genome scan we use a single sequence threader, RAPTOR, to identify suitable templates for each sequence in the query pair using z-score > 3.0. If we do not have an interface template for a SCOPPI family composed of the SCOP families corresponding to any combination of these templates, we use DBLRAP to thread the two sequences onto a conventional full-complex template (see SI for details). Once the putative interface is determined, we use interface-specific scores to predict the interaction between the proteins (stage 4). See Materials and Methods for a detailed description of the classifier employed to predict an interaction.

In order to evaluate our predictions, we compute a receiver operating characteric (ROC) curve by varying the probability cutoff for predicting an interaction. When comparing ROC curves against other homology/structure-based PPI predictors, we find that iWRAP consistently outperforms HOMCOS, Multiprospector and DBLRAP. Multiprospector reports a sensitivity of 20% at a specificity of 80%, whereas iWRAP achieves a sensitivity of 56% at 80% specificity (see Fig 4). HOMCOS reports a recall of 80% with a precision of 10%. In contrast, iWRAP achieves a precision of 27% at the 80% recall level (see SI Fig S6). Struct2Net[8,31] uses the DBLRAP threading program for prediction of interactions from structural data. When comparing against Struct2Net (only yeast predictions), we find that iWRAP dominates Struct2Net at all accuracy levels (see Fig 4).

Interface threading requires multiple structural data for an interaction, which is not always available. By using interface threading in conjunction with DBLRAP, our method, i.e. iWRAP+DBLRAP(boost), achieves a coverage of 13% for the yeast interactome. This is close to a 50% increase in coverage over previous methods[31], without any compromise in sensitivity (Fig 4). Here, coverage is defined as the percentage of high-confidence interactions in Biogrid[51] for which a method can make a prediction. iWRAP makes predictions for 9752 high-confidence interactions in Biogrid (involving around 3400 proteins), whereas DBLRAP makes predictions for 5832 interactions (involving around 2700 proteins). 3920 are unique to iWRAP's interface threading predictions; this results in close to a 50% increase in coverage compared to DBLRAP. In addition, iWRAP predicts about 100,000 novel interactions in the yeast genome; the cutoff (= 0.9) for identifying a positive interaction is chosen based on the distribution of interaction probabilities (see SI Fig S7). We note that around 60% of our predictions come from across family threading– that's not surprising given the limited template database; it is more likely to have a good match to one sequence of the query, than to both of them.

To further analyze iWRAP's performance, we looked at the 640 proteins involved in the high-confidence interactions from Biogrid uniquely predicted by iWRAP. One finding from a GO term enrichment analysis using Amigo[52] revealed that this set was enriched for proteins functioning as structural constituents of the ribosome (GO: 0003735, P-value < 10*e* – 6). Additionally, iWRAP makes predictions for proteins within functional complexes involving nuclear proteins such as the ‘U5 snRNP complex’ and ‘SAS complex’. Amongst the type of functional complexes that both iWRAP and DBLRAP predict, we find that iWRAP's predictions are significantly enriched for the following complexes (> 6 fold over DBLRAP): ‘Rtt109p/Vps75p complex’ (12 fold over DBLRAP), ‘signal peptidase complex’ (11 fold) and ‘GPI-anchor transamidase complex’ (9 fold). The full list of such complexes and complexes unique to iWRAP predictions is given in SI *Genomic Predictions*. The annotation of these complexes, including their memberships, were taken from a manually-curated dataset compiled by Pu et al.[53]. Finally, we investigated the templates selected for the unique predictions made by iWRAP. Table 2 gives a summary of the most frequent templates used for predicting these interactions. While DBLRAP selects one representative complex for each SCOPPI family, multiple templates can be selected by iWRAP from within a family. This contributes to iWRAP's improved prediction accuracy as features for only the most significant interface are considered for PPI prediction. Furthermore, as noted earlier in cross-validation tests (Fig 3A), size of the interface template is correlated with iWRAP's accuracy: larger interfaces lead to more confident predictions. From Table 2, the average probability computed by iWRAP for interface templates of size less than 20 contacts (mean=0.20, std.dev=0.13) is half of the average probability computed for templates greater than 20 (mean=0.40, std.dev=0.20).

2.4. iWRAP predicts novel cancer-related interactions

We demonstrate that iWRAP can be used to identify important targets for experimental investigation through an application to yeast homologs of human cancer-related genes. We integrate enrichment and functional analysis to enumerate bona fide candidates for further investigation (Fig 5). Recently, a large scale double-mutant study has revealed a genetic interaction map for yeast[41]. However, the set of interesting genes for any detailed study of a disease (e.g. cancer) is still large. In contrast to this approach, we use iWRAP predictions to identify the most important targets for further study. It has been shown that structure-based scores are one of the most significant predictors, as compared to co-localization, co-expression and GO term enrichment, for general PPI prediction[54,8]. We employ these criteria to prioritize and validate our targets (Fig 5A). For the set of yeast genes related to cancer identified in CYGD[55], we first filter the predicted interactions based on co-localization. iWRAP identifies 727 interactions for the disease genes (out of ~54000 possible interactions). After discarding predictions between proteins that are not co-localized; 301 putative interactions remain for further analysis. We then identify genes enriched for GO processes, with the genetic interaction set as the background. Note that this is a much more stringent criterion than using the whole genome as the background; the latter yields many more putative interacting genes. We used AmiGO[52] to filter genes based on a p-value cutoff of 0.01 (corrected for multiple hypothesis testing). The enrichment analysis narrows down the list of candidate genes to 28. Note that we are using both co-localization and enrichment as filters to select the most important candidate genes; we treat both of them as equally important. For genes that were significantly enriched (~4 fold, see SI Table S3), we used IsoBase[56] to identify their human funtional orthologs. To exploit the more comprehensive yeast genome annotation, we carried out the enrichment on the yeast predictions before mapping them onto the human genome. We found that these enriched genes are differentially expressed in cancer-vs-normal tissues[57]. Furthermore, using BLAST we were able to identify similar proteins (E-value < 10) in a database of cancer-related proteins[58]. We hypothesize that these novel interactions are directly involved in cancer-related pathways, and should be investigated further (Fig 5B).

Amongst the genes predicted by iWRAP as interacting with known cancer promoting genes, particularly interesting are genes coding for ribosomal proteins associated with either the small (RPS) or large (RPL) subunit (Fig 5B). Mutations in several of these proteins, including RPS17 and RPL5 identified by iWRAP, have been very recently implicated in congenital abnormalities and predisposition to cancer, known as Diamond Blackfan Anemia (DBA)[59]. The expression disregulation of RPS and RPL genes have also been observed in pancreatic cancer and stromal displasia[60] and in colorectal cancer[61]. In addition, there are two (human DEAD box) helicases DDX23 and DDX55 (Fig 5B) in the set of putative interactions. Even though there is limited research on various human helicases they are believed to be involved in embryogenesis and cell growth and have recently been shown to be involved in tumorigenesis[62]. Furthermore, iWRAP predicts an interaction between XPA (RAD14) and SMARCA5; the latter has been shown to be critical for regulating the genetic program required for normal differentiation[63].

3. Discussion

We introduce the program iWRAP and show that integrating interface profiles into a localized scoring scheme aids in interfacial contact prediction. We introduce the use of across-family templates to mitigate the limited number of templates, and also capture convergently evolved interface motifs. We apply our approach to predict interacting proteins encoded by the entire yeast genome. Furthermore, by integrating our predictions in a combined functional and enrichment study of cancer related genes in yeast, we show that iWRAP can uncover novel, biologically relevant interactions.

While we have optimized the two new parameters (α and ω_{gap}) in our threading scoring function specific to interface predictions (see Materials and Methods), it would be interesting to see if simultaneously optimizing the other parameters, already optimized separately in the fold recognition score of RAPTOR, improves accuracies even further. In particular, we expect the sequence profile and secondary structure scores to be the most important for very low sequence identities; as we have shown in Fig 3B, the interface profiles may not be sufficient to pinpoint the exact interaction core in such cases. As noted in *Cross-validation within SCOPPI families*, for sparse contacts and small interfaces in long sequences, the localized nature of iWRAP can miss the interaction core, thus identifying an incorrect interacting surface. In such cases, a pre-processing step with DBLRAP to roughly identify the interface region could be beneficial before using the localized threading algorithm.

In this paper, we have focused on SCOPPI families having more than three complexes in a binding mode. In addition, we have not considered complexes formed by domains in the same SCOP family, which rules out homodimers (as handled by HOMCOS). Combining interface threading with DBLRAP effectively addresses limitations of small number of SCOPPI-derived interface templates. Furthermore, for families having only one solved complex, we plan to utilize interface profiles computed from PSI-BLAST as input to our localized algorithm. We believe that an expanded template database and a full optimization of the scoring function parameters will improve iWRAP's predictive abilities even further.

Our program iWRAP makes accurate PPI predictions that are independent of all the non-structure-based approaches and may thus be combined with any of them. iWRAP is designed to handle template-query pairs having low sequence similarity, making it complementary to other PPI databases like MODBase[23]. A key advantage of iWRAP is that, apart from the PDB data used for constructing templates, the prediction algorithm only requires protein sequence data as input. It can thus be applied to proteins for which no functional data is available.

4. Materials and Methods

4.1. Stage 1: Template construction

We utilize the SCOPPI classification of protein-protein interfaces to construct interface profiles. SCOPPI classifies interfaces based on sequence and structural similarity of the interface[40]. In addition, for each interacting SCOP family pair, SCOPPI provides a sequence alignment of other interfaces in the same SCOP family pair. Here we use this classification of interfaces to construct our own multiple interface alignments for each SCOP family pair using CMAPi[36]. CMAPi employs a contact-map representation to efficiently align multiple interfaces and thereby improves alignments, as compared with SCOPPI and other sequence/structure-based alignment programs, especially in cases where the sequence identity between aligned structures is low[36]. A contact map is a binary matrix representation of the residue-residue interactions between two proteins. If the distance between any two heavy atoms of the two residues is less than 4.5Å, the corresponding entry in the contact map is one, and zero otherwise.

We construct interface profiles from these interface alignments by computing a unique set of consensus environment classes, one for each interface alignment position (see SI Fig S2). An environment class is a combination of a secondary structure (SS) class, an amino acid class and average solvent accessibility (across the alignment at that position). We use the classification as defined by Rice et al. (1997), which, briefly, consists of three SS classes, two solvent accessibility classes and seven amino acid classes. Rice et al. also provide a table, H3P2, which provides amino acid/SS preferences for these environmental classes. The

profiles computed from a multiple interface alignment represent the environment information at the interface across the multiple structures in the alignment. Since the consensus contact map constructed by CMAPi includes all contacts across the aligned complexes, our interface profiles are robust to small variations in inter-residue distances.

4.2. Stage 2: Aligning query sequences to templates

The goal in this stage is to align query sequence profiles to interface template profiles, constructed in stage 1. We obtain query sequence profiles from PSIBLAST[47] and query secondary structure (SS) predictions from PSIPred[64]. Once we identify a suitable template, we score individual query-template alignments using Rice et al.'s H3P2 table (see above), which, in the context of single structure alignment, quantifies the preference of aligning a query sequence/SS profile to a template profile. However, since our query SS's are predicted, we instead use H3P2 scores weighted by the PSIPred SS probability distribution at a query sequence position.

$$H3P2score\,(t, s\,(t)) = \sum_{ss=C,H,E} P\,(ss)\,H3P2\,(s\,(t)\,, ss, t) \tag{1}$$

Here t is the template position, $s(t)$ is the query sequence position aligned to template position t, ss is C(coil), H(helix) or E(beta strand), $P(ss)$ is the probability of a secondary structure class at position $s(t)$ given by PSIPred and $H3P2(s(t), ss, t)$ is the H3P2 table score of aligning query $s(t)$ having ss to the template position t. While Eq. 1 represents the score for one aligned position, the total alignment score is calculated by summing over all aligned positions. Note that we utilize only one state 'C' to model loops. We currently do not distinguish between coil and other structural loops such as beta turns or tight turns.

4.3. Stage 3: Interface scoring

The goal in this stage is to integrate the interface profile scoring scheme from stage 2 into a general threading approach to obtain a score for a putative interaction. Our solution employs a LP strategy motivated by that used by RAPTOR for single-domain threading. We begin by constructing our objective function. For each sequence in the query pair, in addition to the RAPTOR single-domain threading score, we include the interface profile score (see stages 1,2 above) of aligning the query sequence, s, with the interface template profile:

$$E_{iWRAP} = E_{RAP} - \alpha E_{CMAPi} - \omega_{gap} GAP_{RAP} \tag{2}$$

$$E_{CMAPi} = \sum_{t} H3P2score\,(t, s\,(t)) \tag{3}$$

$$E_{RAP} = E_m + E_s + E_g + E_p + E_{ss} \tag{4}$$

E_{iWRAP} is the interface threading energy function (scoring function); E_{CMAPi} is the interface profile score; $H3P2score$ is the alignment score from the H3P2 table (see stage 2, Eq 1); and GAP_{RAP} is the total gap (opening+extension) score used by RAPTOR. E_{RAP} is the threading score employed by RAPTOR. This includes environment fitness score E_s based on solvent accessibility, secondary structure compatibility score E_{ss}, sequence profile scores calculated from PSI-BLAST E_m, an affine gap penalty E_g and a pairwise within-domain interaction

score E_p[45] . To score an alignment to an interface template position represented by a gap state, we use the mean negative score in the H3P2 table (i.e. mean of the unfavorable alignment scores). To take into account possible gaps at the interface, we add a weighted negative penalty ($\omega_{gap}GAP_{RAP}$) to the score. Note that parameters α and ω_{gap} are optimized independently based on our training set, as described in *Training and test sets*. To obtain the alignment, the E_{iWRAP} score is minimized independently for each of the two query sequences using the implementation of RAPTOR, which utilizes an open-source optimization library (COIN)[65] (Fig 6, left).

Contact map optimization—From the independent interface threading above, we produce an initial query contact map (Fig 6, right). We further refine this contact map by incorporating residue-residue interaction specificity and optimizing similarity of the binding patterns in query and template. We carry out optimization in the neighborhood of interacting residues using a residue-residue interaction score[30]. A 10×10 sub-matrix in the contact map around an interacting pair defines this local neighborhood. For each contact (S_1, S_2) in the initial contact map, we maximize the Hadamard product between two matrices: one, a sub-matrix around the predicted contact in the query contact map (Q_{cmap}) and two, a sub-matrix around the corresponding template contact in the template contact map (T_{cmap}). If (T_1, T_2) is the corresponding template contact, then this optimization can be written as:

$$A=\arg \max_{s_1=S_1+d_1, s_2=S_2+d_2} \sum_{d_1,d_2 \in [-5,5]} \delta\left(Q_{cmap}(s_1,s_2), T_{cmap}(T_1+d_1, T_2+d_2)\right) \quad (5)$$

where 'A' represents the set of possible contacts that maximize the Hadamard product, δ is the kronecker-delta function and d_1, d_2 are the sub-matrix indices. This optimization maximizes (around each contact) the similarity of binding patterns in the template and query contact maps. For residues aligned to gaps, we allow the alignment to shift so that the nearest non-gapped position is used in the Hadamard product optimization. Since each Hadamard optimization is performed independently, one template contact could be mapped to multiple contacts in the query contact map. To avoid one to many mappings, for each template contact, we rank the possible predicted contacts using the quasi-chemical residue-residue interaction scoring potential of Lu et al.[30] ('E_{pwqc}') and choose the top ranking unique one:

$$optimizedContact=\operatorname{argmin}_{c \in A} \quad E_{pwqc}(c) \quad (6)$$

The final contact map is the set of these *optimizedContacts* (Fig 6, right). Additionally, the significance of the predicted interaction score is measured by calculating a z-score with respect to a distribution generated by randomizing the interfacial contacts. The total score (energy) of the interface and the associated z-score are used in predicting interactions in stage 4.

4.4. Stage 4: PPI prediction

The goal in this stage is to predict whether the two query proteins interact based on the interface score computed in stage 3. Since only a few protein pairs interact *in vivo*, the main challenge here is to discriminate true interactions from false ones. To achieve this goal, we extract a vector of scores '$X_{Interface}$' that quantifies the quality of the predicted interface[8] and feed this vector to a boosting classifier, which computes a probability 'p' of the interaction:

$$p = f\left(X_{Interface}\right) \tag{7}$$

$$X_{Interface} = \{tA, tB, sA, sB, cmap, E, e, zA, zB, z\_e, tZ, E\_pi, cmap\_pi, piAB\} \tag{8}$$

We extract the following features, i.e. '$X_{Interface}$', from the putative interface: template sequence lengths (*tA, tB*), query sequence lengths (*sA, sB*), predicted number of contacts (*cmap*), total interface energy computed from the pairwise potential ($E = \Sigma_{c\ \varepsilon\ optimizedContacts} E_{pwqc}(c)$), normalized interface energy (*e*), z-scores for the threading alignments (z*A*, z*B*) and z-score for the interface energy (*z_e*). In addition, we use the features sum of threading z-scores (*tZ*), square root of the product of sequence lengths (*piAB*), total interface energy normalized by *piAB* (*E_pi*) and number of contacts normalized by *piAB* (*cmap_pi*).

We train a boosting classifier on known high-confidence interactions from Biogrid to learn an accurate function '*f*'. Our method is based on AdaBoost, which involves improving the overall classification by appropriately weighting outputs of a series of rules of thumb, or base classifiers; we use classification trees as the base classifiers[39] (see SI for details). Using this trained model a probability of interaction is computed, which indicates iWRAP's confidence in predicting an interaction between the query proteins: 1 indicates maximum confidence and 0 indicates no confidence. Note that this stage is used only for our genome scans, where we have no *a priori* knowledge of interaction between the query proteins.

4.5. Training and test sets

For each SCOPPI family (i.e. SCOP family pair), the set of complexes is divided into a training set and a test set; a leave-one-out cross-validation (LOOC) procedure is employed to optimize the parameters. A complex in the test set has an interface sequence identity less than 40% with each of the complexes in the training set. The complexes from the training set are used in constructing the multiple interface alignments with CMAPi, and subsequently the interface profiles. We use the training set to optimize the two parameters in the scoring function, α and ω_{gap} from Eqn 2. The parameters are varied alternatively to maximize the alignment accuracy of the threading alignments, where CMAPi alignments are used as the gold-standard. At each iteration, α is varied in intervals of 5, and ω_{gap} is varied in intervals of 0.1. The parameter value which gives the maximum alignment accuracy is chosen at each iteration. After an initial broad sweep for α, the parameters typically converge within 20 iterations.

In addition to LOOC testing within a SCOPPI family, we consider the performance of iWRAP on complexes having similar binding patterns (as given by an iracc of greater than 0.75) across families. Interacting residue accuracy (iracc) gives a measure of similarity in binding patterns between two interfaces: an iracc of one indicates very similar interfaces, and zero highly dissimilar interfaces[32]. For across-family cross-validation, we restrict ourselves to SCOPPI family pairs sharing one SCOP family. Notice that the parameter optimization has been carried out independently for each SCOPPI family, and hence alignments across SCOPPI family pairs are independent of the training process.

In order to train the classifier in stage 4 for our genomic scans, we constructed the set of training examples as in Struct2Net[8,31]. Briefly, the set of positive examples was taken as the high-confidence interactions in Biogrid[51]. Any two proteins separated by at least three edges in the interaction network constructed from Biogrid were considered as non-interacting, and included in the negative set. For our predictions on yeast, the training set

consisted of 3500 positive and 16000 negative examples. Our test set had 720 positive and 3000 negative examples.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Thanks to Rohit Singh, Vinay Pulim and Daniel Park for help with data and code. Thanks to Jerome Waldispuhl for critically reading the manuscript.

Funding

National Institute of Health (NIH) grant 1R01GM081871.

References

1. Li S, Armstrong C, Bertin N, Ge H, Milstein S, et al. A map of the interactome network of the metazoan *c. elegans*. Science. 2004; 303:540–543. [PubMed: 14704431]
2. Rual J, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. Towards a proteome-scale map of human protein-protein interaction network. Nature. 2005; 437:1173–1178. [PubMed: 16189514]
3. Uetz P, Giot L, Cagney G, Mansfield T, Judson R, et al. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. Nature. 2000; 403:623–627. [PubMed: 10688190]
4. Giot L, Bader J, Brouwer C, Chaudhari A, Kuang B, et al. A protein interaction map of *drosophila melanogaster*. Science. 2003; 302:1727–1736. [PubMed: 14605208]
5. Björkland A, Light S, Hedin L, Elofsson A. Quantitative assessment of the structural bias in protein-protein interaction assays. Proteomics. 2008; 8:4657–4667. [PubMed: 18924110]
6. Sontag D, Singh R, Berger B. Probabilistic modeling of systematic errors in two-hybrid experiments. Proceedings of the Pacific Symposium on Biocomputing. 2007; 12:445–457.
7. Murzin A, Brenner S, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology. 1995; 247:536–540. [PubMed: 7723011]
8. Singh, R.; Xu, J.; Berger, B. Struct2net: Integrating structure into protein-protein interaction prediction; Proceedings of the Pacific Symposium on Biocomputing. 2006. p. 403-414.http://struct2net.csail.mit.edu/
9. Aloy P, Russell R. Interrogating protein interactions networks through structural biology. Proceedings of the National Academy of Sciences. 2002; 99:5896–5901.
10. Kim P, Lu L, Xia Y, Gerstein M. Relating three-dimensional structures to protein networks provides evolutionary insights. Science. 2006; 314:1938–1941. [PubMed: 17185604]
11. Aloy P, Russell R. Structural systems biology: modelling protein interactions. Nature Reviews Molecular Cell Biology. 2006; 7:188–197.
12. Aytuna A, Gursoy A, Keskin O. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics. 2005; 21:2850–2855. [PubMed: 15855251]
13. Prieto C, De L, Rivas J. Structural domain-domain interactions: assessment and comparison with protein-protein interaction data to improve the interactome. Nucleic Acids Research. 2006; 34:W298–W302. [PubMed: 16845013]
14. Stein A, Russell R, Aloy P. 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Research. 2005; 33:D413–D417. [PubMed: 15608228]
15. Je erson E, Walsh T, Roberts T, Barton G. Snappi-db: a database and api of structures, interfaces and alignments for protein-protein interactions. Nucleic Acids Research. 2007; 35:D580–D589. [PubMed: 17202171]

16. Finn R, Marshall M, Bateman A. ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. Bioinformatics. 2005; 21:410–412. [PubMed: 15353450]
17. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P. The protein data bank. Nucleic Acids Research. 2000; 28:235–242. [PubMed: 10592235]
18. Ben-Hur A, Noble W. Kernel methods for predicting protein-protein interactions. Bioinformatics. 2005; 21(Suppl 1):i38–46. [PubMed: 15961482]
19. Deng M, Mehta S, Sun F, Chen T. Inferring domain-domain interactions from protein-protein interactions. Genome Research. 2002; 12:1540–1548. [PubMed: 12368246]
20. Betel D, Breitkreuz K, Isserlin R, Dewar-Barch D, Tyers M, Hogue C. Structure-templated predictions of novel protein interactions from sequence information. PLOS Computational Biology. 3:e182.
21. Fukuhara N, Go N, Kawabata T. Prediction of interacting proteins from homology-modeled complex structure using sequence and structure scores. Biophysical Journal. 2007; 3:13–26.
22. Fukuhara N, Go N, Kawabata T. HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. Nucleic Acids Research (Web Server Issue). 2008; 36:W185–189.
23. Pieper U, Eswar N, Webb B, Eramian D, Kelly L, Barkan D, Carter H, Mankoo P, Karchin R, Marti-Renom M, Davis F, Sali A. MODBASE: A database of annotated comparative protein structure models and associated resources. Nucleic Acids Research. 2009; 37:D347–D354. [PubMed: 18948282]
24. Encinar J, Fernandez-Ballester G, Sanchez I, Hurtado-Gomez E, Stricher F, Beltrao P, Serrano L. ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. Bioinformatics. 2009; 25:2418–2424. [PubMed: 19602529]
25. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Current Opinion in Structural Biology. 2002; 12:368–373. [PubMed: 12127457]
26. Valencia A, Pazos F. *In silico* two-hybrid system for the selection of physically interacting protein pairs. Proteins. 2002; 47:219–227. [PubMed: 11933068]
27. Burger L, Nimwegen E. Accurate prediction of protein-protein interactions from sequence alignments using a bayesian method. Molecular Systems Biology. 4:165. [PubMed: 18277381]
28. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein-protein interactions based only on sequences information. Proceedings Of The National Academy Of Sciences. 2007; 104:4337–4341.
29. 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]
30. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophysical Journal. 2003; 84:1895–1901. [PubMed: 12609891]
31. Singh R, Park D, Xu J, Hosur R, Berger B. Struct2net: a web service to predict protein-protein interactions using a structure-based approach. Nucleic Acids Research (Web Server Issue). 2010:1–8.
32. Pulim L, Bienkowska J, Berger B. LTHREADER: Prediction of extra-cellular ligand-receptor interactions in cytokines using localized threading. Protein Science. 2008; 17:279–292. [PubMed: 18096641]
33. Caffrey D, Somaroo S, Hughes J, Mintseris J, Huang E. Are protein-protein interfaces more conserved in sequence than rest of the protein surface? Protein Science. 2004; 13:190–202. [PubMed: 14691234]
34. Capra J, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007; 23:1875–1882. [PubMed: 17519246]
35. Fraser H, Hirsh A, Steinmetz L, Scharfe C, Feldman M. Evolutionary rate in the protein interaction network. Science. 2002; 296:750–752. [PubMed: 11976460]
36. Pulim V, Bienkowska J, Berger B. Optimal contact map alignment of protein-protein interfaces. Bioinformatics. 2008; 24:2324–2328. [PubMed: 18710876]
37. Zhang Q, Petrey D, Norel R, Honig B. Protein interface conservation across structure space. Proceedings of the National Academy of Sciences. 2010; 107:10896–10901.

38. Kundrotas P, Vakser I. Accuracy of protein-protein binding sites in high-throughput template-based modeling. PLOS Computational Biology. 2010; 6:e1000727. [PubMed: 20369011]
39. Culp M, Johnson K, Michailidis G. ada: A R package for stochastic boosting. Journal of Statistical Software. 17
40. Winter C, Henschel A, Kim W, Schroeder M. SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Research (Database issue). 2006; 34:310–314.
41. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear E, et al. The genetic landscape of a cell. Science. 2010; 327:425–431. [PubMed: 20093466]
42. T. GO Consortium. Gene ontology: tool for the unification of biology. Nature Genetics. 2000; 25:25–29. [PubMed: 10802651]
43. Lehner B, Fraser A. A first-draft human protein-interaction map. Genome Biology. 2004; 5:63.
44. Madeo F, Herker E, Wissing S, Jungwirth H, Eisenberg T, Fröhlich K. Apoptosis in yeast. Current Opinion in Microbiology. 2004; 7:655–660. [PubMed: 15556039]
45. Xu J, Li M, Kim D, Xu Y. RAPTOR: Optimal protein threading by linear programming. J Bioinform Comput Biol. 2003; 1:95–117. [PubMed: 15290783]
46. Edgar R. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004; 32(5):1792–1797. [PubMed: 15034147]
47. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25:3389–3402. [PubMed: 9254694]
48. Eddy S. Hidden markov models. Current Opinion in Structural Biology. 1996; 6:361–365. [PubMed: 8804822]
49. Hu, J.; Shen, X.; Shao, Y.; Bystroff, C.; Zaki, M. Mining protein contact maps. BIOKDD02: Workshop on Data Mining in Bioinformatics.
50. Stebins C, Galán J. Structural mimicry in bacterial virulence. Nature. 2001; 412:701–705. [PubMed: 11507631]
51. Stark C, Breitkreutz B, Reguly T, Boucher L, Brietkreutz A, Tyers M. BIOGRID: A general repository for interaction datasets. Nucleic Acids Research. 2006; 34:D535–539. [PubMed: 16381927]
52. Carbon S, Ireland A, Mungall C, Shu S, Marshall B, Lewis S, AmiGO HUB W. P. W. g. Amigo: online access to ontology and annotation data. Bioinformatics. 2009; 25:288–289. [PubMed: 19033274]
53. Pu S, Wong J, Turner B, Cho E, Wodak S. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Research. 2009; 37:825–831. [PubMed: 19095691]
54. Bandyopadhyay S, Kelley R, Krogan N, Ideker T. Functional maps of protein complexes from quantitative genetic interaction data. PLOS Computational Biology. 2008; 4:e1000065. [PubMed: 18421374]
55. Güldener U, Münsterkõtter M, Kastenmüller G, Strack N, van Helden J, et al. CYGD: the comprehensive yeast genome database. Nucleic Acids Research. 2005; 33:D364–D368. [PubMed: 15608217]
56. Singh, R.; Xu, J.; Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection; Proceedings of the National Academy of Sciences. 2008. p. 12763-12768.http://isobase.csail.mit.edu/
57. Rhodes D, Yu J, Deshpande N, Varambally R, Ghosh D, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. 2004; 6:1–6. [PubMed: 15068665]
58. Huang Y, Hang D, Lu L, Tong L, Gerstein M, Montelione G. Targeting the human cancer pathway protein interaction network by strutural genomics. Molecular and Cellular Proteomics. 2008; 7:2048–2060. [PubMed: 18487680]
59. Lipton J, Ellis S. Diamond blackfan anemia 2008-2009: broadening the scope of ribosome biogenesis disorders. Current Opinion in Pediatrics. 2010; 22:12–19. [PubMed: 19915471]
60. Crnogorac-Jurvevic T, Efthimiou E, Capelli P, Blaveri E, Baron A, et al. Gene expression profiles of pancreatic cancer and stromal desmoplasia. Oncogene. 2001; 20:7437–7446. [PubMed: 11704875]

61. Lai M, Xu J. Ribosomal proteins and colorectal cancer. Current Genomics. 2007; 8:43–49. [PubMed: 18645623]

62. Parsyan A, Shahbazian D, Martineau Y, Petroulakis E, Alain T, et al. The helicase protein dhx29 promotes translation initiation, cell proliferation, and tumorigenesis. Proceedings Of The National Academy Of Sciences. 2009; 106:22217–22222.

63. Stopka T, Zakova D, Fuchs O, Kubrova O, Blafkova J, et al. Chromatin remodeling gene smarca5 is dysregulated in primitive hematopoietic cells of acute leukemia. Leukemia. 2000; 14:1247–1252. [PubMed: 10914549]

64. Jones D. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology. 1999; 292:195–202. [PubMed: 10493868]

65. Lougee-Heimer R. The common optimization interface for operations research. IBM Journal of Research and Development. 2003; 47:57–66.

66. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research. 2003; 13:2498–2504. [PubMed: 14597658]

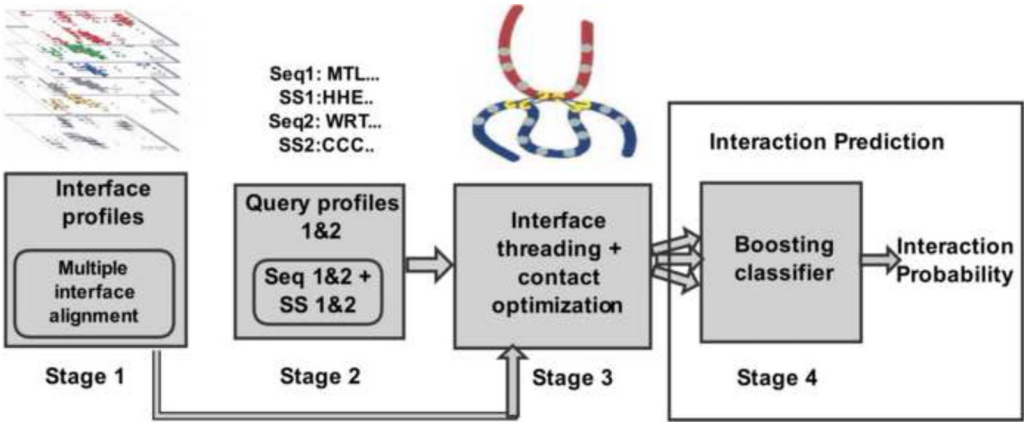

Figure 1. Schematic of the iWRAP algorithm

Interface profiles are constructed from a multiple contact map alignment in Stage 1. Individual interfaces are represented as colored contact maps, and the aligned interface is shown at the bottom in black and white (Stage 1, top). Query profiles consisting of the sequence and secondary structure propensities (Stage 2) is threaded onto the interface template (Stage 3; query1 is shown in red, query2 in blue, interface residues in yellow). From the putative interface in Stage 3, a number of features of the interface are used in predicting an interaction in Stage 4 (for only the yeast genome scan).

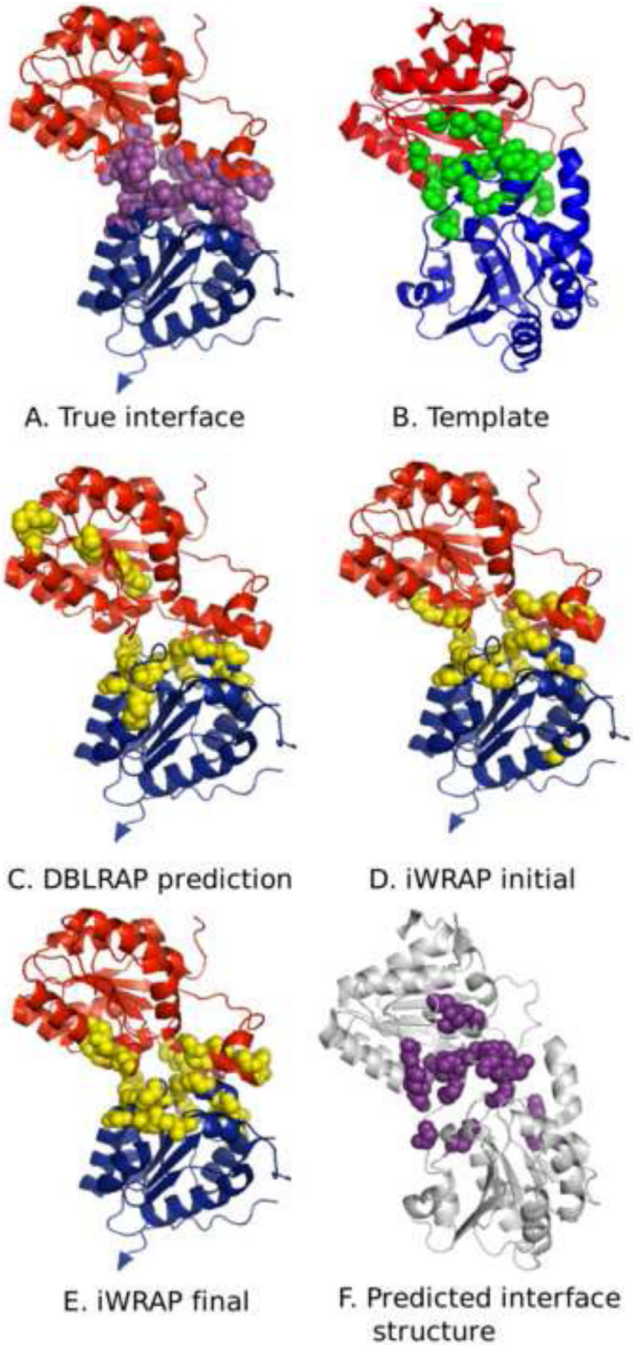

Figure 2. Example of improved contact predictions by iWRAP in within-family cross-validation

PDB 1upc chains A(12-195) and B(375-573) are threaded to the template 1qpbAB. **A**) The true interface computed from the PDB structure of 1upc has roughly 50 contacts. The interface residues are shown as purple spheres, chain B is shown in red and chain A in blue. **B**) The template (1qpbAB) used for threading the query sequences; the interface residues are shown in green. **C**) The interface residues (yellow spheres) predicted by DBLRAP. DBLRAP fails to align the interface region of one interacting partner due to low sequence homology between the query and template (contact accuracy = 0%). **D**) Initial interface (yellow spheres) predicted by iWRAP after threading (contact accuracy = 27%). iWRAP uses interface profiles constructed from a multiple alignment of the interfaces 1mczHG,

1jscAB, 1ozhDC and 1qpbAB; the profiles are then mapped onto the template 1qpbAB. **E**) Final interface (yellow spheres) predicted by iWRAP after contact map optimization. This step refines the contact map, resulting in contacts closer to the true interface. The final contact map is closer to the true contact map (contact accuracy = 46%). **F**) Predicted interface structure obtained by mapping true interface residues from A onto the template structure in B using iWRAP alignments.

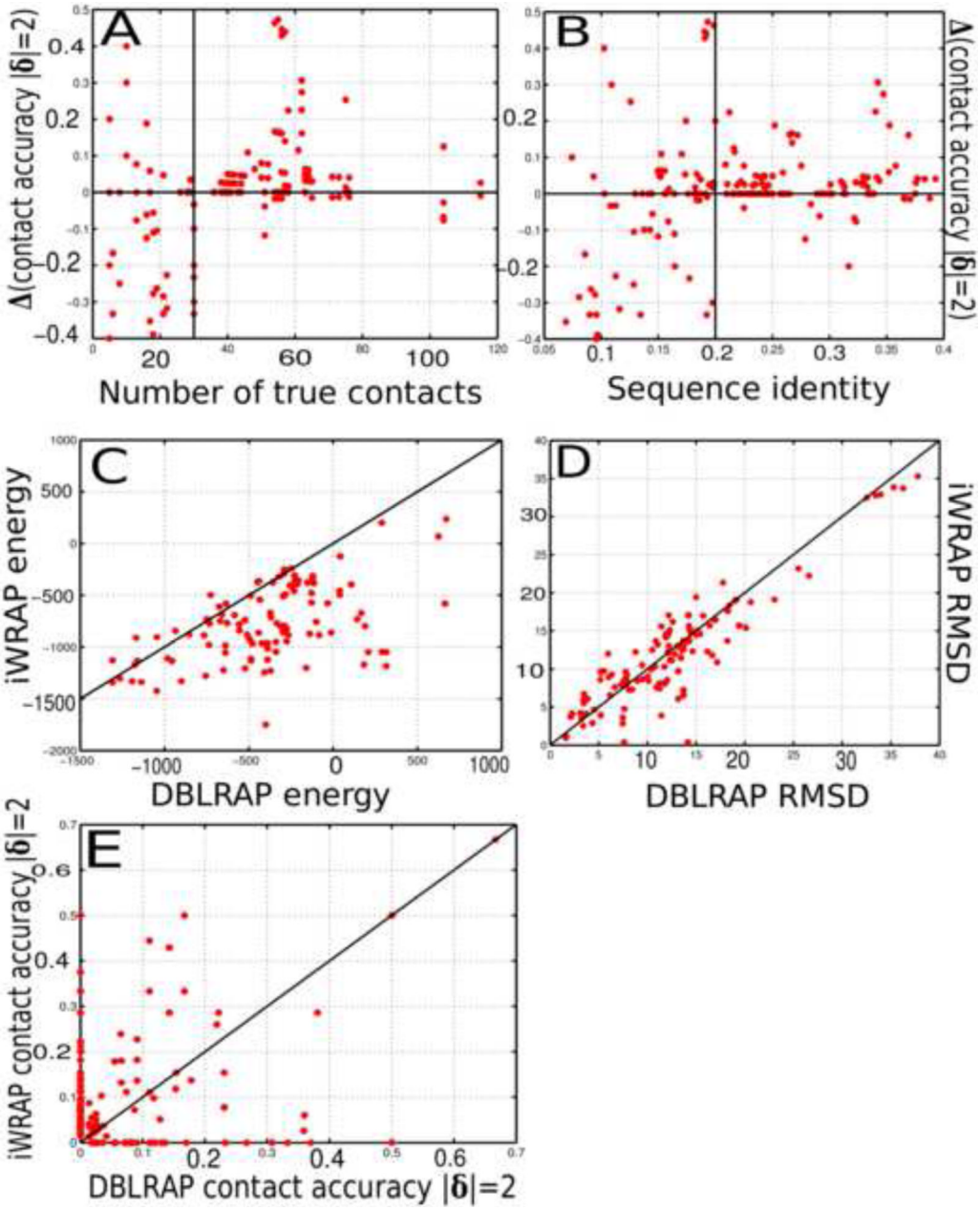

Figure 3. Interface alignment and contact validation

Panels A, B, C and D are cross-validation results on within SCOPPI family threading. Δ(contact accuracy |δ|=2) is the difference in contact accuracies (|δ|=2) between iWRAP and DBLRAP. **A**) Contact accuracy improvement of iWRAP relative to DBLRAP as a function of number of true contacts at the interface. **B**) Contact accuracy improvement of iWRAP relative to DBLRAP as a function of sequence identity at the interface. **C**) iWRAP consistently achieves lower average interface energies as compared to DBLRAP. **D**) RMSD comparison between iWRAP and DBLRAP- better contact prediction by iWRAP does not affect RMSD of the predicted interface. **E**) Cross-validation results for interfaces sharing only one SCOP family (see *Cross-validation across SCOPPI families*). See SI for calculation of contact accuracies and interface energy.

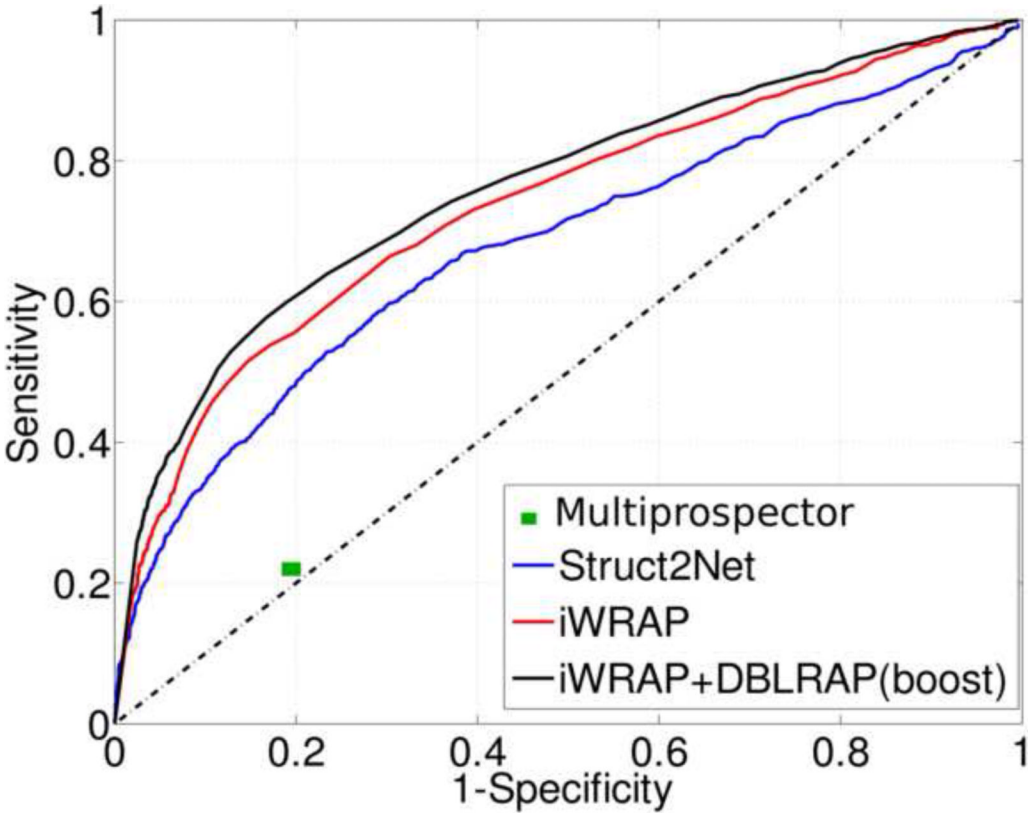

Figure 4. Results on the yeast genome

Sensitivity vs specificity for Multiprospector, iWRAP, Struct2Net and iWRAP+DBLRAP (combined method). In the combined method, DBLRAP threading results are boosted and combined with iWRAP predictions. Here sensitivity = (true positives)/(true positives + false negatives) and specificity= (true negatives)/(true negatives + false positives).

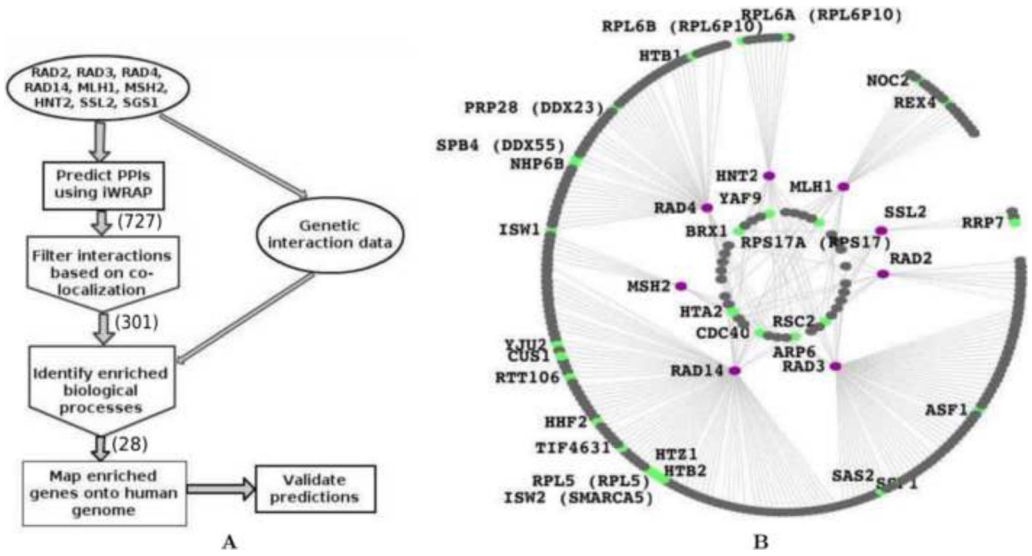

Figure 5. iWRAP predicts novel, bona fide interactions

A) Enrichment analysis was carried out to identify high-confidence interactions. Genes filtered by co-localization and significantly enriched compared to the genetic interaction set were validated using the Oncomine and HCPIN databases. Number of genes remaining after each stage are indicated in parantheses. **B**) The analysis in A reveals a set of high-confidence genes (green) predicted to be interacting with yeast homologs of cancer related genes (purple). Human orthologs of genes for which there is literature providing evidence of implication in cancer have been indicated in parentheses. Genes interacting with only one “cancer” (purple) gene are in the outermost circle, whereas those interacting with more are in the innermost circle. Genes which are not significantly enriched are colored in grey, however, these predicted interactions could also reveal novel biological insights. The figure was created using Cytoscape[66].

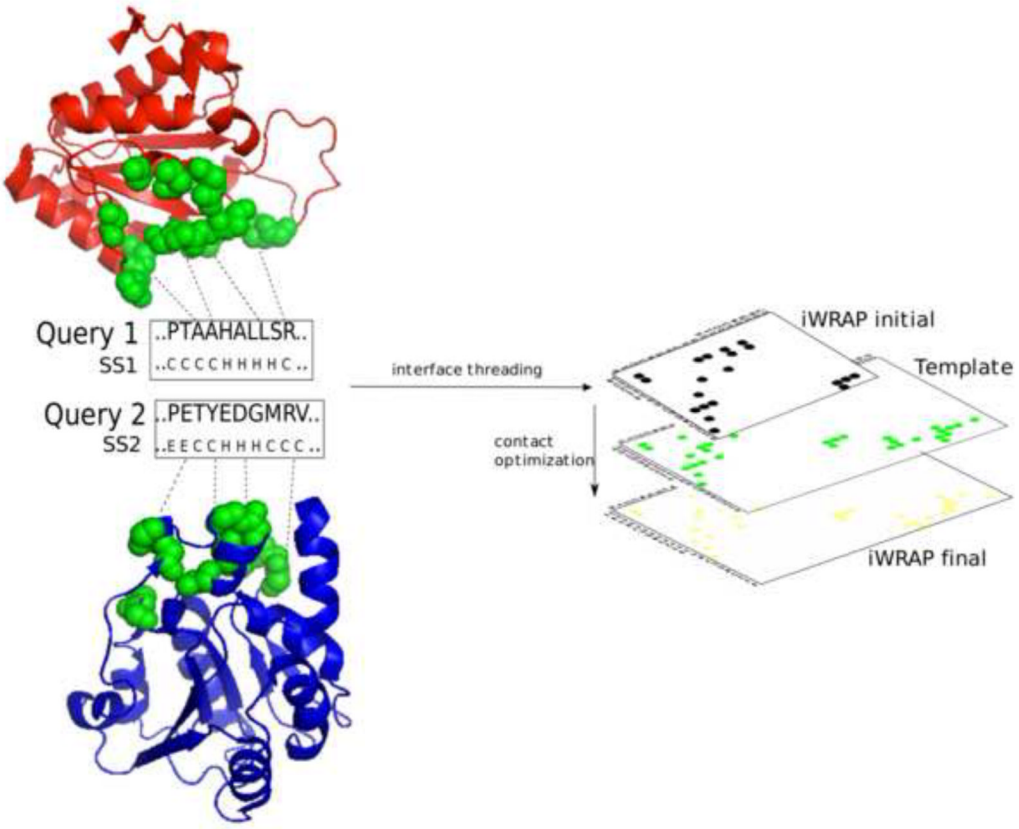

Figure 6. Schematic of interface threading and contact optimization

For the example shown in Figure 2, the query proteins are individually aligned to the template (left) using a local alignment to the interface (dashed lines). For scoring this alignment, we use the interface profiles computed from the multiple-interface alignments, predicted secondary structure for the query pair and the single-domain threading score of RAPTOR. Minimizing this alignment score produces an initial contact map, 'iWRAP initial', which is further refined using Hadamard product optimization and quasi-chemical pairwise residue potentials to produce 'iWRAP final' (right).

Table 1

Comparison of alignment accuracies

Cross-validation results (alignment accuracies) on SCOPPI families having non-redundant complexes (< 40% sequence identities at interface and at least 5 contacts) in the test-set (see Materials and Methods). See SI Table S4 for a detailed biological summary, including types of interactions, for these families.

SCOPPI Family	Seq.ID (%)	LTHREADER (%)	MUSCLE (%)	DBLRAP (%)	iWRAP (%)
f.24.1.1_f.25.1.1	10	1	4	14	**22**
b.47.1.2_g.8.1.1	18	**34**	24	0	32
b.47.1.2_g.3.15.1	7	2	3	0	**8**
a.56.1.1_d.133.1.1	5	12	3	**30**	27
c.81.1.1_d.58.1.5	5	0	7	29	**32**
a.74.1.1_d.144.1.7	11	16	10	19	**26**
c.1.12.1_c.49.1.1	12	17	**29**	24	13
c.55.1.1_d.109.1.1	21	2	**19**	13	**19**
a.80.1.1_c.37.1.20	15	3	3	9	**27**
d.133.1.1_d.87.2.1	11	0	0	15	**24**
a.137.2.1_b.70.1.1	10	1	4	28	**31**
d.171.1.1_h.1.8.1	28	**28**	13	**28**	19
e.18.1.1_e.19.1.1	6	0	7	21	**45**
c.2.1.4_c.23.12.1	15	1	20	**25**	21
b.47.1.2_g.3.2.1	35	12	18	6	**21**
d.122.1.2_d.14.1.3	12	1	5	**15**	10
b.6.1.2_f.24.1.1	20	4	27	32	**37**
Average	14	8	11	18	**24**

Table 2

The most frequent templates used by iWRAP for threading sequences involved in high-confidence interactions in Biogrid unique to iWRAP

The most frequent templates used by iWRAP for threading sequences involved in high-confidence interactions in Biogrid, which are unique to iWRAP. Column 2 gives the size of the interface template (i.e. number of contacts), column 3 the number of threaded pairs in the test set, and column 4 the average predicted probability of interaction in the test set. A template id '1v55B2-1v55A2' represents the interface formed by SCOP domains in chain B and chain A in the PDB complex '1v55'.

SCOPPI Family	Template	Size of Interface	Number of interactions in test set	Average Probability
f.17.2.1_f.24.1.1	1m56H30-1m56G14	135	40	0.297
f.17.2.1_f.24.1.1	1qleB1-1qleA17	132	23	0.398
f.17.2.1_f.24.1.1	1v55B2-1v55A2	124	33	0.481
f.17.2.1_f.24.1.1	1fftG27-1fftF52	96	18	0.183
b.40.4.1_d.104.1.1	1asyA68-1asyB205	63	5	0.400
b.40.4.1_d.104.1.1	1b8aB1001-1b8aA104	51	16	0.624
b.40.4.1_d.104.1.1	1g51A1-1g51B1105	46	9	0.667
b.40.4.1_d.104.1.1	1n9wB1-1n9wA111	43	14	0.428
a.56.1.1_d.133.1.1	1jrpE85-1jrpF124	61	8	0.000
c.55.1.1_d.109.1.1	1yagA147-1yagG1	45	8	0.732
c.55.1.1_d.109.1.1	1h1vA147-1h1vG412	32	7	0.281
b.40.2.2_d.19.1.1	1d5mC2-1d5mA4	41	12	0.180
d.185.1.1_f.23.12.1	1bgyM234-1bgyQ1	22	24	0.333
d.185.1.1_f.23.12.1	1bccA233-1bccE1	22	16	0.499
a.39.1.5_c.37.1.9	1dfkZ3-1dfkA6	19	32	0.258
a.39.1.5_c.37.1.9	1dflX4-1dflB5	14	23	0.277
a.80.1.1_c.37.1.20	1sxjA548-1sxjB7	16	18	0.397
a.80.1.1_c.37.1.20	1iqpC233-1iqpD2	15	10	0.100
a.80.1.1_c.37.1.20	1jr3B243-1jr3E1	10	10	0.300
d.185.1.1_f.23.12.1	1kb9A240-1kb9E31	7	6	0.000