Supplement:
iWRAP: An interface threading approach for protein-protein interaction prediction

Raghavendra Hosur, Jinbo Xu, Jadwiga Bienkowska*, and Bonnie Berger*

*Corresponding author email: bab@mit.edu, jbienkowska@gmail.com

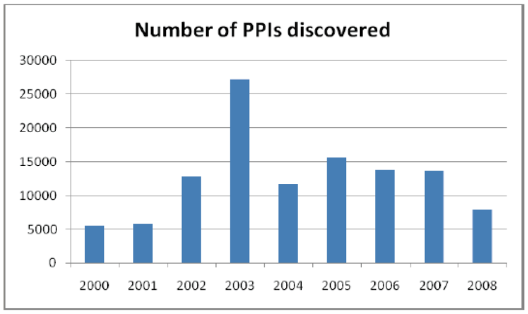


Figure S1. Rate of discovery of new eukaryotic PPI data has slowed

Organism	**Number of interactions**	**Percentage of proteins with at least 1 interaction**
Mouse	1486	6.0
Human	26640	41.8
Worm	4559	14.5
Fly	22740	52.7
Yeast	48901	93.5

Table S1. Availability of experimental PPI data for major eukaryotic organisms. Data based on phenotypic suppression/enhancement and synthetic interaction was excluded, as these experiments do not provide evidence of a direct physical interaction between proteins.

Evaluation of alignments

Calculation of information content. Besides sequence identity, information content is another popular metric used to quantify the difficulty of an alignment problem. The information content for an alignment is calculated by summing the information content of each column of the alignment. The information content of each column is calculated as given by the equation:

$$ic_j = \sum_i P_{ij} log(P_{ij}/Q_i)$$

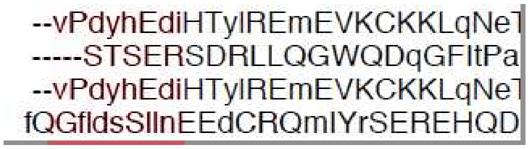
In the above equation, ic_j is the information content of column j, P_{ij} the frequency of amino acid i in column j and Q_i the background frequency of amino acid i. To get the frequency of each amino acid in a column, we count the number of occurrences of that amino acid and divide it by the length of the column. A pseudo-count of 0.01 is added to all counts to avoid zero count. The background distribution Q is taken as the interface propensities of the amino acids [4]. This distribution is quite different from the frequencies of occurrence of individual amino acids in the entire SWISSPROT [1] database. However, for the purposes of this study, information content calculated based on this distribution captures the relative hardness of each alignment.

Calculation of alignment accuracy. For an alignment of a sequence S to a template T obtained using a threading approach, the number of correct alignments is calculated by counting the number of common pairs (t,s) between the threading alignment and the alignment generated by CMAPi for T and S. The accuracy is then obtained by normalizing this count by the length of the CMAPi alignment.

Calculation of contact accuracy. Three contact accuracies are calculated for each predicted contact map. The exact accuracy, i.e., the number of correctly predicted contacts divided by the total number of true contacts. The two other accuracies allow for a shift ($|\delta|$) in the predicted contacts. For example, if (s_1, s_2) are positions of a true contact, we consider a predicted contact to be correct if it is within ($s_1 \pm \delta$, $s_2 \pm \delta$). We only report the contact accuracies with a shift of 2.

Calculation of interface RMSD. For an interface alignment of a sequence S to a template T obtained using a threading approach, the RMSD is calculated by considering only the C_α coordinates of the aligned residues. The Biopython module SuperImposer is used to calculate the minimum RMSD. Average RMSD per family pair is calculated by averaging the RMSDs for all possible template-sequence alignments within a family pair.

Results



```
--vPdyhEdiHTyIREmEVKCKKLqNeT
-----STSERSDRLLQGWQDqGFItPa
--vPdyhEdiHTyIREmEVKCKKLqNeT
fQGfIdsSIlnEEdCRQmIYrSEREHQD
```

A

Features/Position in alignment	3	4	5	6	7	8	9
Residues	VXG	PXF	DXL	YSD	HTS	SE	DLE
Sec. Struct.	C	C	C	C	H	H	H
Avg. Solv. Acc.	58	78	59	22	25	99	69

B

Figure S2. Example of an interface template. **A)** An example of a multiple interface alignment from CMAPi (only one core is shown). The upper case letters represent the contacting residues in the interface, profiles constructed from residues highligted in red are shown in B. **B)** Interface template encoding the consensus residues, consensus secondary structure class and average solvent accessibility at the highlighted (in red) alignment positions in **A**. "X" represents the gap state in the alignment.

Cross-validation within SCOPPI families

iWRAP improves contact predictions for within-family threading alignments (Fig S2a). The trend in relative improvement of iWRAP over DBLRAP is not clear when viewed as a function of information content (Fig S3b) and iracc (Fig S3c).

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 2A). Following the contact-map mining techniques of BYSTROFF, we characterized each contact by the pattern of contacts in a 5x5 residue neighborhood around it. The average density in this neighborhood is calculated by dividing the number of contacts by 25. We observe that iWRAP predictions have a higher density on average than DBLRAP predictions, on both the training and testing sets. We then separately clustered patterns of contacts predicted by iWRAP and DBLRAP and represented each cluster with a representative contact pattern (as done in BYSTROFF). In the cross-validation set, patterns unique to iWRAP predictions have an average density of 0.24 whereas patterns unique to DBLRAP have an average density of 0.21. We find a similar trend in the training set (for template-query less than 40% sequence identity): iWRAP predictions (density 0.278) are denser than DBLRAP predictions (0.226), which is a difference of 1-2 contacts. Average density of patterns common for both is 0.12 in the cross-validation set and 0.22 in the training set. We conclude that density is a factor in the improved performance, and thus may be a factor in the decreased performance in the case of fewer than 20-25 contacts.

Cross-validation across SCOPPI families

iWRAP improves contact accuracy prediction for around 75% of the families in the across-family cross-validation test (Table S2). Threading using templates having one common family with the query pair help us increase coverage in the yeast interactome prediction. Of the 110 SCOPPI families having atleast three complexes in a binding mode, we are able to thread around 30 families after filtering based on sequence identity ($< 40\%$) and iracc score (> 0.75).

Examination of the across-family threading results provides some interesting observations (see SI Table S2). For SCOPPI family pair **a.56.1.1**_d.133.1.1 we obtain the best query as **a.56.1.1**_d.41.1.1. This indirectly indicates a similarity in binding patterns between the two families; an observation also suggested by the ABAC database of convergent evolved interaction motifs ABAC. Further evidence for such an inference is demonstrated by the family **d.185.1.1**_f.23.14.1. In our results, the best query for this template is from the family **d.185.1.1**_f.23.12.1, which is also noted in the ABAC database. This suggests that threading across-families might capture similar binding patterns and can be used in genome-scale PPI predictions.

Methods

Templates

For each family pair in SCOPPI, the coordinates are obtained from the listed PDB IDs. In order to exclude interfaces formed due to crystallization, we select interfaces with more than five contacts. Furthermore, PDB models with resolutions lower than 2.5 Å are selected whenever possible. From an interface made up of two domains, three templates are constructed. One is the complex template (dimer), which consists of residue pairs (one on each domain) which have at least one of their heavy atoms at a distance less than 4.5 Å. Three templates are constructed from an interface in a PDB [2] file. A "dimer" template is the template describing the interface residues (see main text). Two additional templates are constructed by extracting the C_α and C_β coordinates for individual domains from the PDB entry. In addition to spatial coordinates, these two templates have information about solvent accessibilities and secondary structure, computed using the program DSSP [5]. These are in the form similar to the templates used by RAPTOR [9].

Multiple interface alignment

Unlike profiles used in prediction of single chain protein structure, construction of profiles for PPI prediction is challenging because interactions between the two protein sequences complicates their treatment as independent alignments. In addition, profiles based on sequence alignments alone do not effectively capture the multiple binding modes exhibited within the same family. As demonstrated in Pulim et al. for the special case of cytokines [7], profiles based on a contact-map representation and alignment of interfaces are better suited for PPI prediction. Templates and profiles are constructed using these multiple interface alignments (see Template Construction in Main Text) for every family pair having atleast 3 "inter-domain" interfaces. These consist of domains on two different chains in the PDB file. Since we are interested in templates for PPI prediction, we consider only inter-domain interfaces. This has the added advantage of filtering out (dimer) interfaces formed due to crystallization. On the other hand, true homodimers will be excluded from our analysis.

Genomic predictions: *S.cerevisiae*

For genomic predictions, we used a two phase approach to identify templates for threading. In the first phase, each of the two query proteins is threaded (using RAPTOR) against the non-redundant database (<40% sequence identity) of proteins in SCOP1.75 [6]. This database contains around 10000 templates. We then select the top templates for each query protein by ranking them by z-scores and using a z-score cutoff of 3.0. At the end of this phase, we end up with 10-15 templates for each protein. In the second phase, we check to see if we have a dimer with the SCOP domains represented by any one of these templates. In case we don't find such a template, we look for a dimer template which has one SCOP family common with one of the templates for the two query proteins. In case of multiple such dimer templates, we use the template with the highest sequence identity to the query proteins. This ensures that even for across-family threading, we utilize structurally similar templates. Our database has around 2000 total dimer templates (compared to around 2200 non-redundant dimers for Struct2Net).

Once we have the threading alignments for two yeast query proteins using iWRAP, we extract the following features from the results: template lengths (ltmpa,ltmb), sequence lengths (lseqa,lseqb), predicted number of contacts (cmap), total interface energy (total.energy), normalized interface energy (energy), z-scores for the threading alignments (alnza, alnzb) and z-score for the interface energy (z). In addition, we use the features sum of threading z-scores (total.z), square root of the product of sequence lengths (piab), total interface energy normalized by piab (energy_pi) and number of contacts normalized by piab (cmap_piab). The negative examples are generated as in Struct2Net [8].

The variable importance plot is shown in Fig S5 and recall-precision curves are shown in Fig S6. As was observed by Singh et al. [8], the size of the sequences (piab), total interfacial energy (total.energy), normalized interfacial energy (energy_pi) are the most significant predictors. In addition, we find that sum of alignment z-scores (total.z) and the number of predicted contacts are important features which were not used in [8].

For the combined predictor, we used DBLRAP's threading alignments to extract features used in Struct2Net, and trained a classifier as above. The two predictions were combined by using a common cutoff to compute the combined ROC curve.

References

[1] A Bairoch and R Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28:45–48, 2000.

[2] H Berman, J Westbrook, Z Feng, G Gilliland, T Bhat, H Weissig, I Shindyalov, and P Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[3] M Costanzo, A Baryshnikova, J Bellay, Y Kim, E Spear, and et al. The genetic landscape of a cell. *Science*, 327:425–431, 2010.

[4] A Fernández, L Ridgway Scott, and H Scheraga. Amino-acid residues at protein-protein interfaces: Why is propensity so different from relative abundance? *J. Phys. Chem. B.*, 107:9929–9932, 2003.

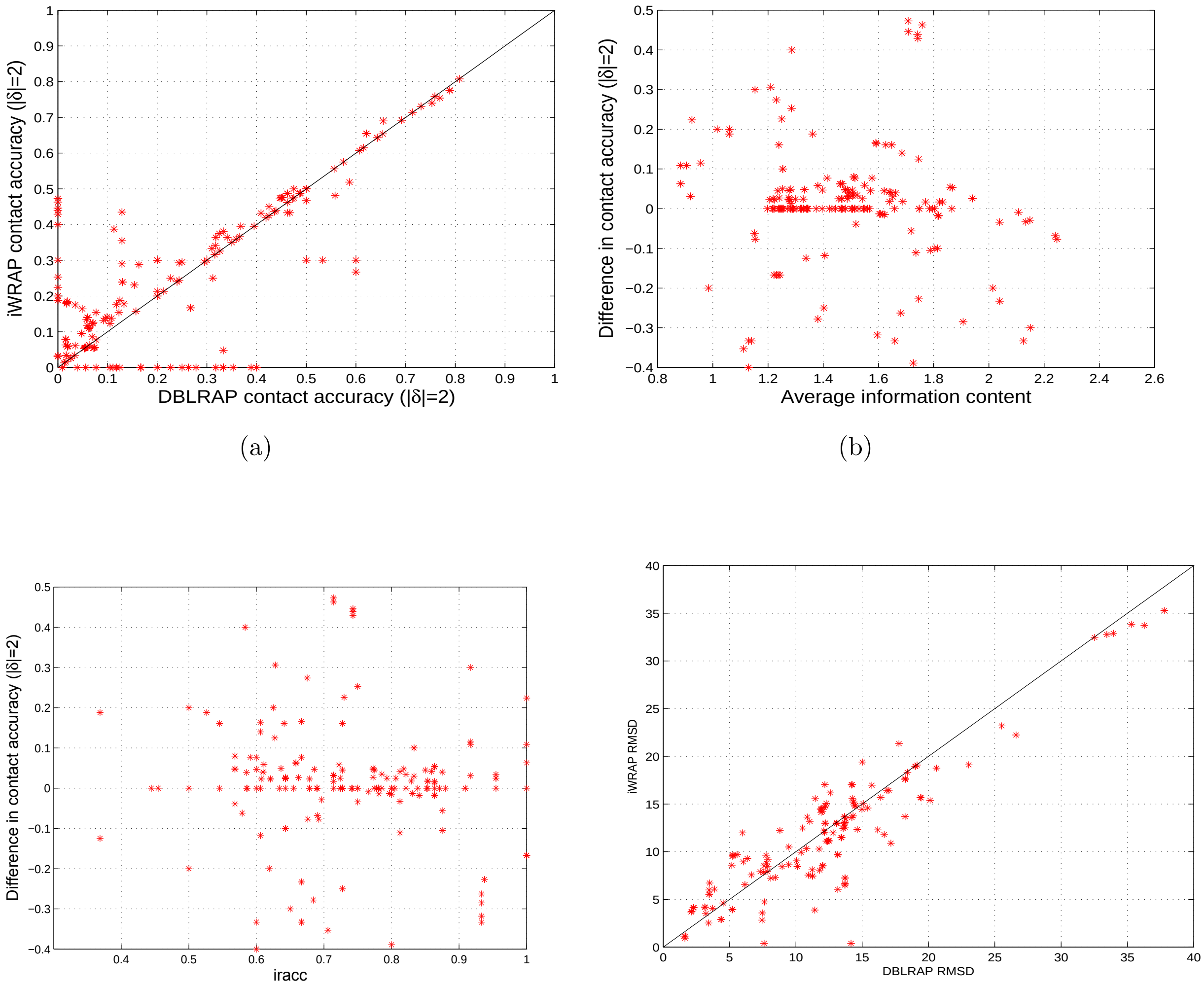
[5] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[6] A.G Murzin, S.E Brenner, T Hubbard, and C Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[7] L Pulim, J Bienkowska, and B Berger. LTHREADER: Prediction of ligand-receptor interactions using localized threading. *Proceedings of the Pacific Symposium on Biocomputing*, 12:64–75, 2007.

[8] R Singh, J Xu, and B Berger. Struct2net: Integrating structure into protein-protein interaction prediction. *Proceedings of the Pacific Symposium on Biocomputing*, 11:403–414, 2006. `http://struct2net.csail.mit.edu/`.

[9] J Xu, M Li, D Kim, and Y Xu. RAPTOR: Optimal protein threading by linear programming. *J Bioinform Comput Biol*, 1:95–117, 2003.



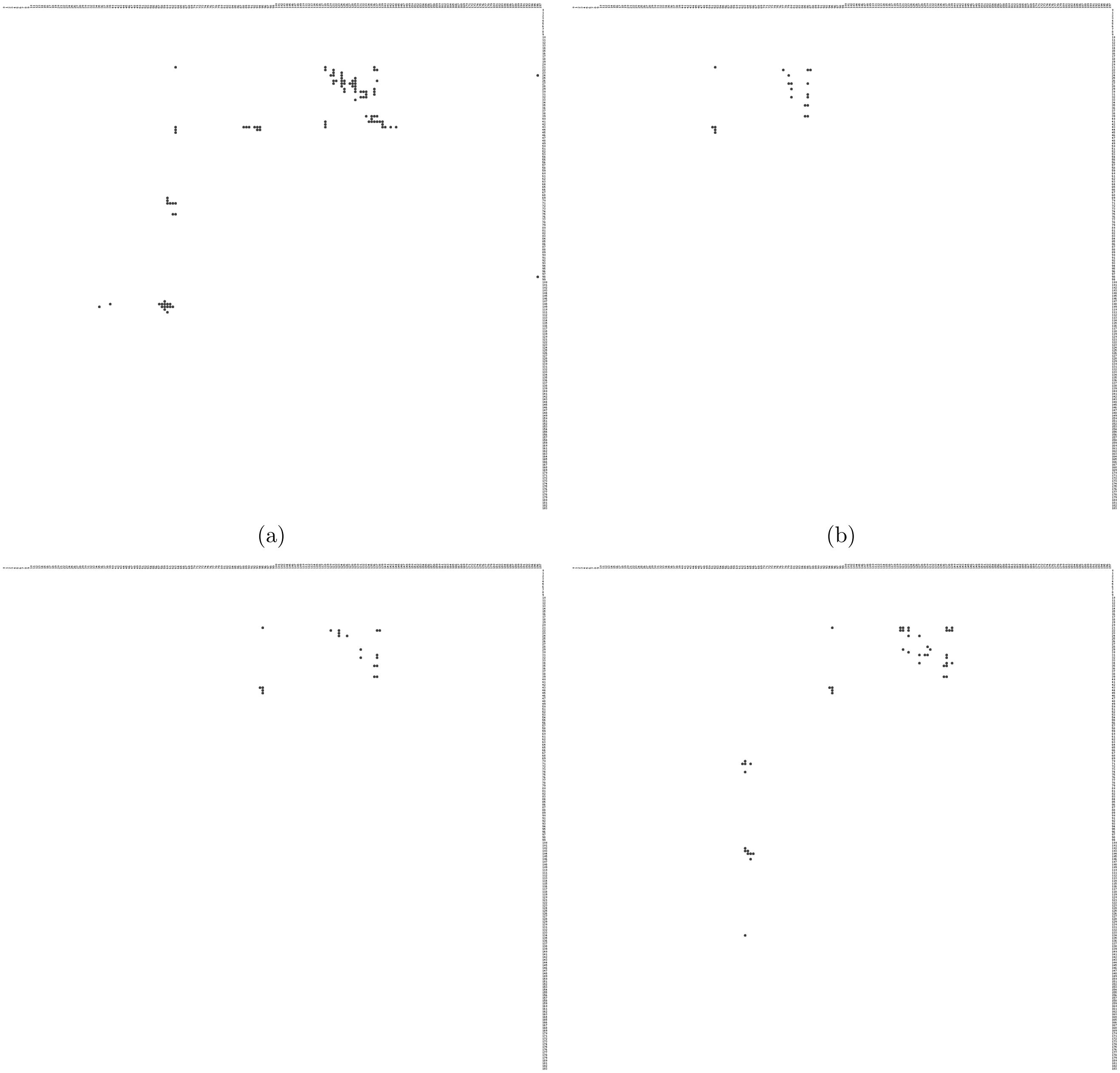
(a)

(b)

(c)

(d)

Figure S3. Improvements in contact accuracy ($|\delta| = 2$) by iWRAP. **a)**Even when DBLRAP predicts less than 20% of contacts, iWRAP can predict significantly more number of contacts close to the true contact map. Relative improvement in contact accuracy ($|\delta| = 2$) as a function of information content and iracc: **b)** Although the trend is not very clear, iWRAP only seems to perform poorly as compared to DBLRAP when the information content per position in the true alignment is high. **c)** The performance improvement of iWRAP is not strongly dependent on iracc. But the lower iracc alignments (i.e. harder alignments) are the alignments that give the highest improvements in accuracies. **d)** RMSD comparison between iWRAP and DBLRAP- better contact prediction by iWRAP does not affect RMSD of the predicted interface. All the results are for template-query pairs having a sequence identity less than 40% at interface.



(a)

(b)

(c)

(d)

Figure S4. Example of improvement in contact accuracy by iWRAP. **a)** Contact map representation for the true interface of 1upcA12-1upcB375 **b)** Contact map predicted by DBLRAP. **c)** Initial contact map predicted by iWRAP from threading. **d)** Final contact map predicted by iWRAP after contact map optimization. In Figure 2 of the main document, these contacts are mapped onto the actual structure.

Template Pair	**DBLRAP**			**iWRAP**		
	Avg_Acc	Best_Query	Max_Acc	Avg_Acc	Best_Query	Max_Acc
a.56.1.1_d.133.1.1	0.04	a.56.1.1_d.41.1.1	0.74	0.09	a.56.1.1_d.41.1.1	0.09
d.185.1.1_f.23.13.1	0.24	f.21.1.2_f.23.13.1	0.67	0.14	f.21.1.2_f.23.13.1	0.67
d.185.1.1_f.23.14.1	0.06	d.185.1.1_f.23.12.1	0.14	0.20	d.185.1.1_f.23.12.1	0.43
d.145.1.3_d.15.4.2	0.11	d.133.1.1_d.15.4.2	0.31	0.32	d.145.1.3_d.41.1.1	0.50
d.185.1.1_f.21.1.2	0.0	-	0.0	0.08	d.185.1.1_f.23.12.1	0.14
c.81.1.1_d.58.1.5	0.0	-	0.0	0.14	b.52.2.2_d.58.1.5	0.14
b.1.18.8_c.37.1.8	0.03	a.87.1.1_c.37.1.8	0.09	0.06	a.87.1.1_c.37.1.8	0.13
b.85.3.1_b.92.1.1	0.03	b.85.3.1_c.1.9.2	0.06	0.18	b.85.3.1_c.1.9.2	0.20
b.47.1.2_g.3.11.1	0.04	b.23.1.1_g.3.11.1	0.07	0.0	-	0.0
a.56.1.1_d.41.1.1	0.12	d.15.4.2_d.41.1.1	0.23	0.12	d.15.4.2_d.41.1.1	0.15
b.47.1.2_g.3.15.1	0.05	b.47.1.2_g.68.1.1	0.11	0.10	b.47.1.2_g.68.1.1	0.24
b.47.1.2_g.8.1.2	0.11	b.47.1.2_g.68.1.1	0.36	0.03	b.47.1.2_g.3.15.1	0.08
d.133.1.1_d.15.4.2	0.05	d.145.1.3_d.15.4.2	0.06	0.0	-	0.0
b.23.1.1_g.3.11.1	0.01	b.47.1.2_g.3.11.1	0.022	0.03	b.47.1.2_g.3.11.1	0.04
f.23.12.1_f.23.13.1	0.0	f.23.12.1_f.23.14.1	0.0	0.07	f.23.12.1_f.23.14.1	0.11
f.21.1.2_f.27.1.1	0.04	f.21.1.2_f.23.13.1	0.08	0.02	f.23.13.1_f.27.1.1	0.05
d.145.1.3_d.41.1.1	0.01	d.145.1.3_d.15.4.2	0.02	0.04	d.145.1.3_d.15.4.2	0.06
f.23.13.1_f.28.1.1	0.01	f.23.13.1_f.32.1.1	0.02	0.09	f.23.13.1_f.32.1.1	0.22
c.36.1.10_c.36.1.6	0.0	-	0.0	0.10	c.36.1.10_c.48.1.1	0.13
c.36.1.10_c.48.1.1	0.0	-	0.0	0.10	c.36.1.6_c.48.1.1	0.20
f.23.13.1_f.32.1.1	0.0	-	0.0	0.35	f.23.13.1_f.28.1.1	0.38
f.21.1.2_f.23.13.1	0.07	d.185.1.1_f.23.13.1	0.12	0.06	d.185.1.1_f.23.13.1	0.11
a.7.3.1_d.15.4.2	0.02	d.145.1.3_d.15.4.2	0.02	0.0	-	0.0
b.33.1.1_f.32.1.1	0.03	f.23.13.1_f.32.1.1	0.03	0.0	-	0.0
f.27.1.1_f.32.1.1	0.0	-	0.0	0.06	f.23.13.1_f.32.1.1	0.07
b.49.2.3_c.1.6.1	0.10	b.49.2.2_c.1.6.1	0.22	0.14	b.49.2.2_c.1.6.1	0.26
f.23.12.1_f.23.14.1	0.03	f.23.12.1_f.23.13.1	0.03	0.03	f.23.12.1_f.23.13.1	0.03
b.49.2.2_c.1.6.1	0.24	b.49.2.3_c.1.6.1	0.38	0.15	b.49.2.3_c.1.6.1	0.29

Table S2. Cross-validation results (contact accuracies $|\delta| = 2$) for interfaces having one common family.

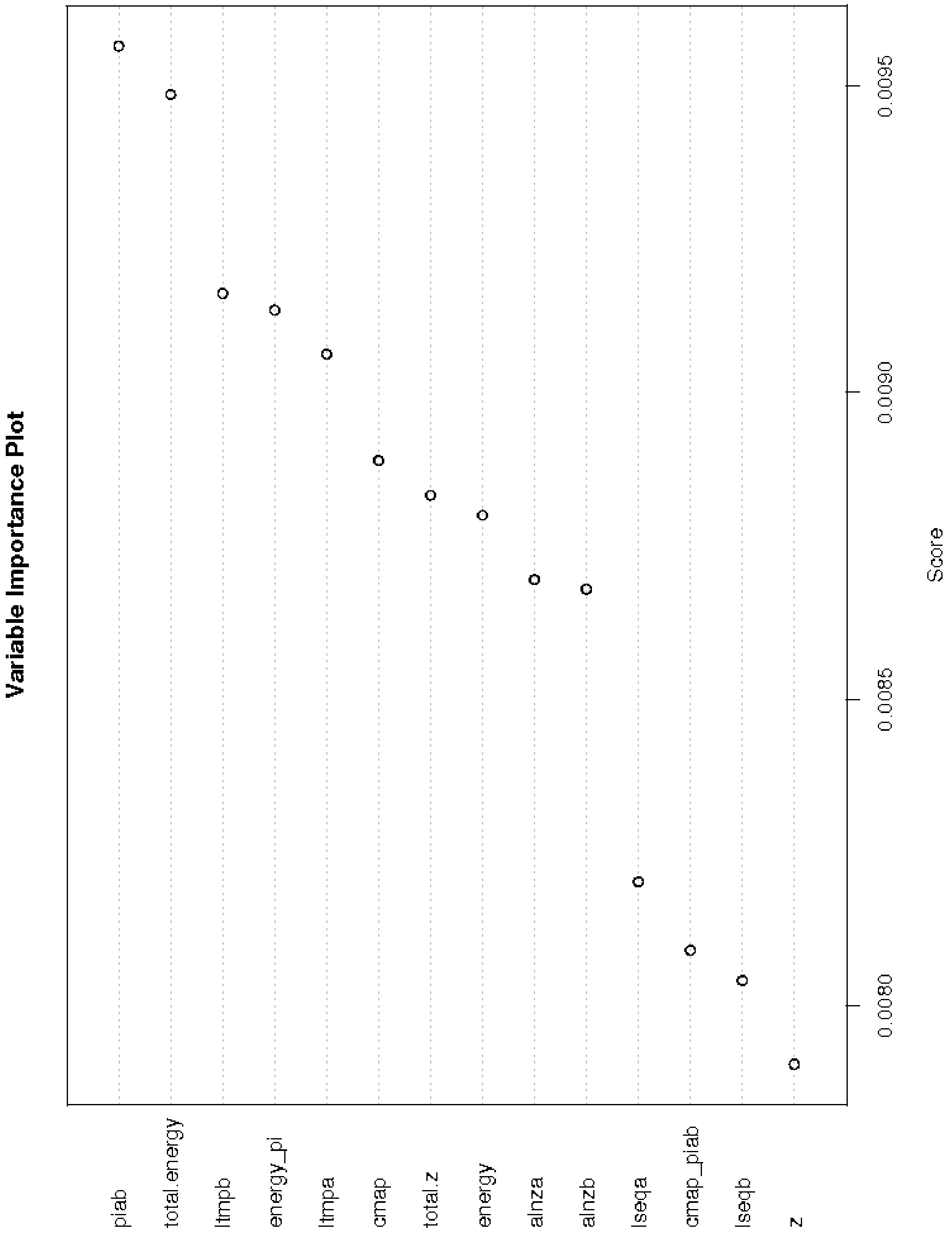


Figure S5. Variable Importance plot for the boosting classifier employed in iWRAP predictions.

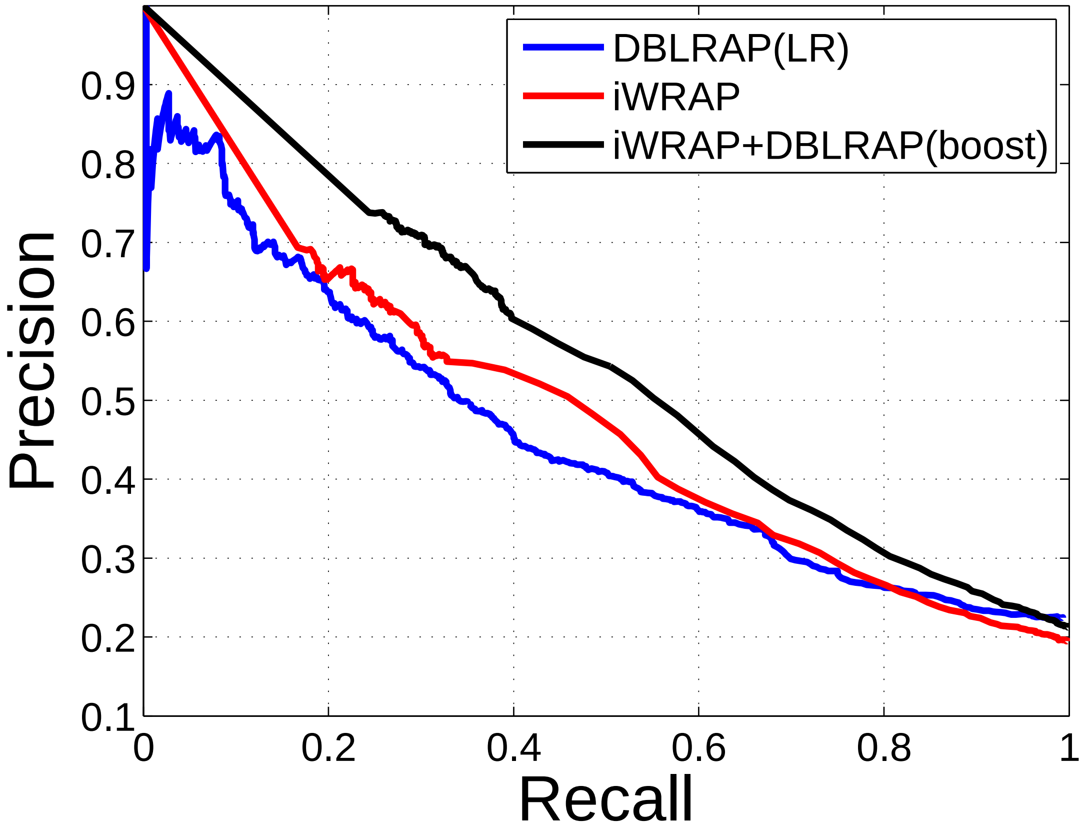


Figure S6. Precision vs recall for the *S.cerevisiae* predictions. Here, precision=true positives/(true positives + false positives) and recall = true positives/(true positives + false negatives).

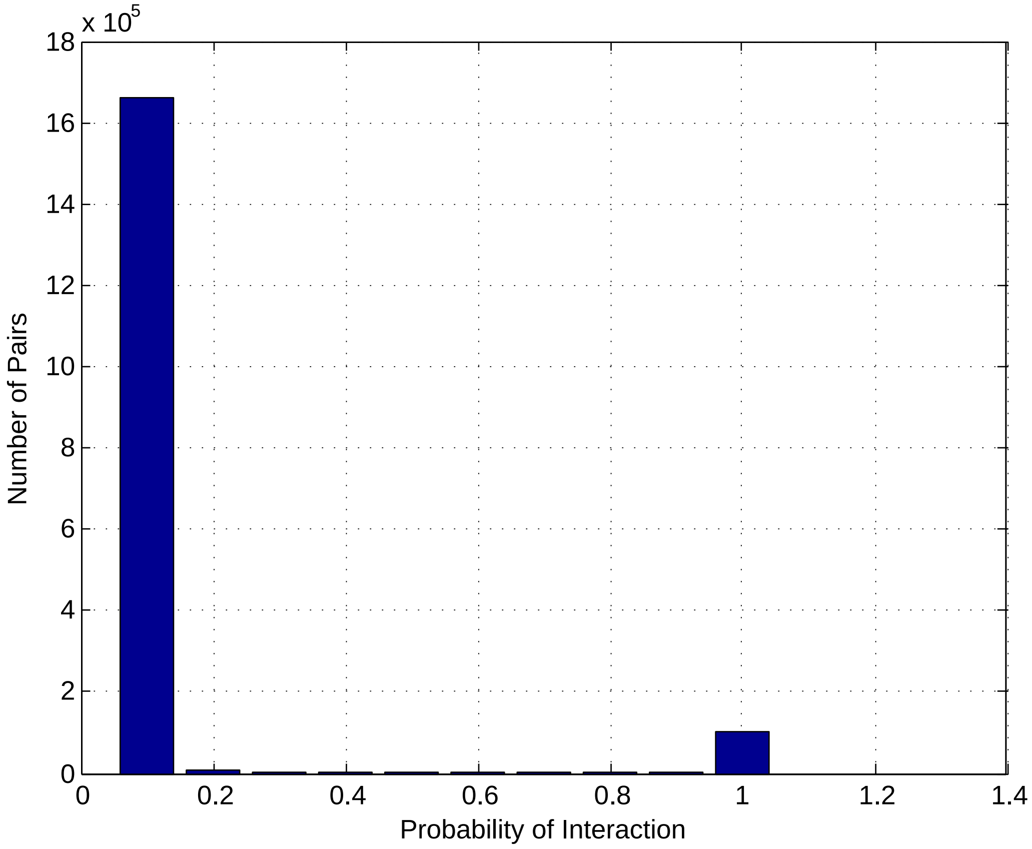


Figure S7. Distribution of interaction scores obtained for the yeast genome scan. Cutoff for an interaction was chosen as 0.9 based on this distribution.

GO Term	P-value	Sample freq.(%)	Background freq.(%)
cellular macromolecule metabolic process	5.39e-13	76.4	52.9
macromolecule metabolic process	8.68e-12	80.0	58.0
cellular metabolic process	3.67e-11	85.2	64.9
gene expression	8.44e-11	52.8	31.4
primary metabolic process	3.86e-10	85.6	66.3
metabolic process	4.75e-9	88.8	71.5
regulation of cellular process	1.32e-8	47.2	28.3
regulation of biological process	2.57e-8	50.0	31.0
regulation of gene expression	4.85e-8	33.2	17.2
regulation of cellular biosynthetic process	5.05e-8	34.0	17.8
cellular process	7.64e-8	96.4	83.9
regulation of biosynthetic process	8.19e-08	34.0	18.0
regulation of macromolecule biosynthetic process	9.32e-08	33.2	17.4
cellular RNA metabolic process	1.49e-07	37.6	21.0
cellular biosynthetic process	1.73e-07	57.2	38.4
biosynthetic process	2.20e-07	57.6	38.9
cellular nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.68e-07	51.6	33.3
macromolecule biosynthetic process	5.92e-07	48.4	30.8
regulation of cellular metabolic process	8.41e-07	35.2	19.7
macromolecular complex subunit organization	9.46e-07	22.0	9.8
cellular macromolecule biosynthetic process	1.12e-06	48.0	30.7
regulation of primary metabolic process	1.96e-06	35.2	20.0
regulation of macromolecule metabolic process	3.12e-06	34.4	19.5
regulation of metabolic process	7.07e-06	36.0	21.1
macromolecular complex assembly	7.94e-06	18.4	7.9
cellular macromolecular complex subunit organization	9.88e-06	18.4	7.9
compound metabolic process	1.78e-05	58.4	41.7
cellular macromolecular complex assembly	3.07e-05	14.8	5.9
biological regulation	4.31e-05	51.6	35.6
regulation of transcription	2.01e-04	25.2	13.7
regulation of cellular transcription	2.01e-04	25.2	13.7
cellular transcription	3.64e-04	26.0	14.5
transcription	3.64e-04	26.0	14.5
regulation of cellular nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	4.78e-04	26.0	14.6
regulation of nitrogen compound metabolic process	5.47e-04	26.0	14.6
transcription from RNA polymerase II promoter	1.27e-03	16.8	8.0
nucleosome organization	1.72e-03	6.0	1.6
transcription, DNA-dependent	3.22e-03	22.0	12.2
cellular transcription, DNA-dependent	3.22e-03	22.0	12.2
RNA biosynthetic process	3.69e-03	22.0	12.2
regulation of transcription from RNA polymerase II promoter	3.86e-03	13.2	5.9
cellular component assembly	4.65e-03	19.6	10.5
ribonucleoprotein complex assembly	9.87e-03	6.4	1.9

Table S3. Enrichment results for the predicted PPIs. The genetic interaction set from [3] was used as the background.

SCOP 1	Name	GO F.	SCOP 2	Name	GO F.	Type
f.24.1.1	cytochrome c oxidase subunit I-like	copper ion binding	f.25.1.1	cytochrome c oxidase subunit-III like	copper ion binding	Perm,Trans
b.47.1.2	Eukaryotic proteases	trypsin activity	g.8.1.1	Small Kuntiz-type inhibitors & BPTI-like toxins	serine-type endopeptidase inhibitor activity	Perm,Trans
b.47.1.2	Eukaryotic proteases	trypsin activity	g.3.15.1	Huristasin-like	serine-type endopeptidase inhibitor activity	Trans
a.56.1.1	CO dehydrogenase ISp C-domain like	electron transporter activity	d.133.1.1	Molybdenum cofactor binding domain	electron transporter activity	Perm, Trans
c.81.1.1	Formate dehydrogenase/DMSO reductase, domains 1-3	oxidoreductase activity	d.58.1.5	Ferredoxin domains from multidomain proteins	electron transporter activity	Perm, Trans
a.74.1.1	Cyclin	ATP binding	d.144.1.7	Protein Kinases catalytic subunit	protein kinase activity	Trans
c.1.12.1	Pyruvate kinase	Kinase activity	c.49.1.1	Pyruvate kinase C-terminal domain	kinase activity	Perm
c.55.1.1	Actin/HSP70	unfolded protein binding	d.109.1.1	Gelsolin-like	actin binding	Trans
a.80.1.1	DNA polymerase III clamp loader subunits, C-terminal domain	DNA binding	c.37.1.20	Extended AAA-ATPase domain	ATP binding	Perm, Trans
d.133.1.1	Molybdenum cofactor binding domain	electron transporter activity	d.87.2.1	CO dehydrogenase flavoprotein C-terminal domain-like	electron transporter activity	Perm, Trans
a.137.2.1	Methanol dehydrogenase subunit	alcohol dehydrogenase activity	b.70.1.1	Quinoprotein alcohol dehydrogenase-like	oxidoreductase activity	Trans
d.171.1.1	Fibrinogen C-terminal domain-like	–	h.1.8.1	Fibrinogen coiled-coil central regions	–	Perm, Trans
e.18.1.1	Nickel-iron hydrogenase, large subunit	ferredoxin hydrogenase activity	e.19.1.1	Nickel-iron hydrogenase, small subunit	ferredoxin hydrogenase activity	Perm, Trans
c.2.1.4	Formate/glycerate dehydrogenases, NAD-domain	oxidoreductase activity	c.23.12.1	Formate/glycerate dehydrogenases, substrate-binding domain	oxidoreductase activity	Perm
b.47.1.2	Eukaryotic proteases	trypsin activity	g.3.2.1	Plant inhibition of proteinases and amylases	serine-type endopeptidase inhibitor activity	Perm, Trans
d.122.1.2	DNA gyrase/MutL, N-terminal domain	ATP binding	d.14.1.3	DNA gyrase/MutL, second domain	ATP binding	Perm
b.6.1.2	Periplasmic domain of cytochrome c oxidase subunit II	copper ion binding	f.24.1.1	cytochrome c oxidase subunit I-like	copper ion binding	Perm, Trans

Table S4. Biological summary of the SCOPPI families given in Table 1. Here, Go F. = GO function annotation, Type=type of interaction, Perm=Permanent, Trans= Transient.