Supplementary Information

High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling

Jing Yang, Richard Jang, Yang Zhang, and Hong-Bin Shen

Dependency of performance on the number of TM helices and selected contacts

We separated the training dataset of 60 TM proteins into 5 subsets according to the number of TM helices (i.e. 3-4, 5-6, 7, 8-10, and >10). The performance comparison with TMHcon (Fuchs *et al*., 2009) of residue contact prediction on the 5 subsets is listed in Table S8. As can be seen, MemBrain performs better than TMHcon on all the 5 subsets. For TMHcon, the best performance is on proteins with seven TM helices, while MemBrain achieves the best performance on proteins with five to six TM helices. Interestingly, MemBrain performs better on large proteins (>=8 TM helices) than small proteins (<=4 TM helices) while TMHcon performs poorly on large proteins. This may be due to the combination of PSICOV (Jones *et al*., 2012) into the machine learning predictors. Before combining PSICOV, the ensemble classifier OSC achieves 42.9% and 49.0% prediction accuracy on proteins with eight to ten TM helices and more than ten TM helices respectively. When combined with PSICOV, MemBrain achieves 61.0% and 62.5% prediction accuracy respectively, which is a significant improvement.

In the above comparisons, we have focused on the prediction of the top $L/5$ contacts. In Figure S9, we plotted the data of prediction performance versus coverage. As expected, the prediction accuracy increases at the expense of decreasing the prediction coverage, indicating the higher the predicted probabilities, the more confident the outputs will be. We also extracted the performance on the top $L/2$ and top L predictions to compare MemBrain with TMhhcp (Wang *et al*., 2011). As shown in Table S9, for the both top $L/2$ and top L cutoffs, MemBrain outperforms TMhhcp visibly.

Table S1. Performance on different groups of proteins according to MSAs size.

MSAs size	Number of proteins	Accuracy (%)		Coverage (%)		Accuracy (δ=4) (%)	
		P [a]	M [b]	P [a]	M [b]	P [a]	M [b]
Group 1: (0,250]	5	12.8	23.6	3.4	5.7	67.5	85.6
Group 2: (250,500]	5	28.0	73.5	5.4	14.9	60.3	95.9
Group 3: (500,1000]	16	39.1	72.0	7.0	11.8	76.3	92.2
Group 4: (1000,5000]	22	48.4	61.2	6.5	8.8	74.6	88.8
Group 5: >5000	12	52.8	61.5	8.8	10.5	82.0	90.7

[a] P denotes PSICOV.

[b] M denotes MemBrain.

Table S2. Performance of individual classifiers and combined classifier.

Method	Accuracy (%)	Coverage (%)	Accuracy (δ=4) (%)
OET1	47.8	7.8	78.5
OET2	45.5	7.4	78.0
OET3	46.8	7.4	78.4
OET4	46.0	7.4	78.2
OET5	45.8	7.3	78.1
OETs [a]	48.2	7.8	79.4
SVM1	47.6	7.8	84.2
SVM2	48.7	8.2	85.2
SVM3	49.0	8.1	84.8
SVM4	46.9	7.7	84.0
SVM5	48.8	8.1	83.1
SVMs [b]	50.7	8.5	84.7
OSC [c]	52.8	8.7	85.3
PSICOV	42.1	6.7	74.7
MemBrain [d]	62.0	10.2	90.4

[a] Combining five independent OET-KNN classifiers.

[b] Combining five independent SVM classifiers.

[c] Fusing OETs and SVMs according to Eq.(11) in main text.

[d] Fusing OSC and PSICOV according to Eq.(12) in main text.

Table S3. Performance comparisons of feature level fusion versus decision level fusion. In the feature level fusion, we treat correlated mutation scores (CMs) as the input features for OET1 and SVM1. A sliding window covering neighboring residue pairs was used to encode the residue pair {*i*, *j*}, i.e., {*i*-*n*, *j*-*n*}, ..., {*i*+*n*, *j*+*n*} for parallel TM helices and {*i*-*n*, *j*+*n*}, ..., {*i*+*n*, *j*-*n*} for anti-parallel TM helices. In the decision level fusion, we linearly combined the prediction probabilities from OET1 and SVM1 with those from PSICOV to make final predictions.

Method	Accuracy (%)	Coverage (%)	Accuracy (δ=4) (%)
OET1 [a]	47.8	7.8	78.5
OET1←CMs (*n*=0) [b]	47.5	7.7	78.7
OET1←CMs (*n*=2) [b]	47.4	7.7	79.2
OET1←CMs (*n*=4) [b]	47.4	7.7	79.3
OET1+PSICOV [c]	57.0	9.3	85.1
SVM1 [a]	47.6	7.8	84.2
SVM1←CMs(*n*=0) [b]	54.8	9.0	87.0
SVM1←CMs(*n*=2) [b]	54.6	9.0	87.5
SVM1←CMs(*n*=4) [b]	58.1	9.4	88.5
SVM1+PSICOV [c]	59.1	9.6	89.5

[a] Refer to Table S2.

[b] Feature level fusion; CMs are encoded as feature vectors fed into OET1 and SVM1.

[c] Decision level fusion; PSICOV is treated as an independent predictor, and its outputs are combined linearly with the outputs from OET1 and SVM1.

Table S4. Performance comparison of TMH-TMH interaction prediction.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Comparison on the Training Dataset				
TMHcon [a]	78.0	45.1	88.2	0.372
SVMcon	63.7	31.9	88.4	0.249
SVMSEQ	65.9	36.5	87.9	0.290
PSICOV	74.1	65.2	79.5	0.453
MemBrain [b]	88.2	57.1	93.1	0.544
MemBrain [c]	90.1	56.2	94.5	0.555
Comparison on the Independent Dataset				
TMHcon	76.7	39.5	88.5	0.322
MEMPACK	80.4	27.0	93.7	0.278
TMhhcp1	79.1	54.5	86.2	0.430
TMhhcp2	80.4	53.7	88.9	0.435
SVMcon	78.2	24.5	95.3	0.291
SVMSEQ	68.0	29.9	90.3	0.259
PSICOV	80.4	62.6	85.3	0.493
MemBrain	87.9	56.3	92.5	0.526

[a] TMHcon used *p*-value rather than MCC, the MCC is calculated according to their original reported data.

[b] Results obtained from 4-fold cross-validation.

[c] Results obtained from jackknife cross-validation.

Table S5. Protein structure modeling of 13 GPCRs by I-TASSER with or without using MemBrain contact predictions with RMSD and TM-score calculated in whole chain [a].

PDBID	L [b]	RMSD (Å)/TM [c]	RMSD(Å)/TM [d]
1u19A	348	19.9/0.450	16.9/0.542
2rh1A	282	22.8/0.208	15.5/0.465
2y00A	286	9.1/0.466	8.0/0.575
2z73A	350	22.9/0.196	16.7/0.590
3emlA	286	20.9/0.194	23.0/0.290
3oduA	282	12.5/0.556	10.6/0.694
3pblA	272	18.1/0.266	18.3/0.361
3rzeA	267	16.0/0.257	6.7/0.597
3vw7A	284	15.7/0.488	13.8/0.558
4dajA	264	5.6/0.702	5.0/0.769
4djhA	286	8.7/0.638	9.4/0.714
4ea3A	278	9.1/0.614	6.5/0.744
4grvA	298	12.3/0.561	9.0/0.650
Average	291	14.9/0.430	12.3/0.581

[a] All GPCR templates and homologous templates with sequence identity >30% were excluded.

[b] Number of residues in the entire GPCR chain.

[c] RMSD and TM-score of the first model by I-TASSER without using MemBrain predictions.

[d] RMSD and TM-score of the first model by I-TASSER using MemBrain predictions.

Table S6. Protein structure modeling of 13 GPCRs by I-TASSER using GPCR templates and MemBrain contact predictions with RMSD and TM-score calculated in the transmembrane regions [a].

PDBID	L [a]	L_{TM} [b]	Acc (*L*/5) [c]	Acc (*L*) [d]	RMSD [e]	TM-score [f]
1u19A	348	169	0.52	0.36	1.3	0.934
2rh1A	282	180	0.58	0.35	1.5	0.937
2y00A	286	180	0.61	0.35	1.6	0.925
2z73A	350	181	0.69	0.46	1.3	0.947
3emlA	286	186	0.35	0.25	2.3	0.870
3oduA	282	182	0.72	0.36	2.2	0.875
3pblA	272	174	0.59	0.36	1.5	0.931
3rzeA	267	176	0.57	0.31	1.7	0.915
3vw7A	284	182	0.56	0.36	2.5	0.834
4dajA	264	177	0.54	0.30	2.0	0.890
4djhA	286	177	0.57	0.37	2.1	0.882
4ea3A	278	177	0.51	0.29	1.8	0.907
4grvA	298	182	0.61	0.36	2.1	0.885
Average	291	178	0.57	0.35	1.8	0.902

[a] Homologous templates with sequence identity >30% were excluded.

[b] Number of residues of the whole-chain.

[c] Number of residues in the transmembrane regions.

[d] Accuracy of the top *L*/5 contact predictions by MemBrain.

[e] Accuracy of the top *L* contact predictions used by I-TASSER.

[f] RMSD (Å) of the first model by I-TASSER using GPCR templates and MemBrain predictions.

[g] TM-score of the first model by I-TASSER using GPCR templates and MemBrain predictions.

Table S7. Performance of the top *L*/5 contact predictions for each range on the 22 CASP9 targets.

Targets	SVMSEQ [a]			PSICOV			SVMSEQ+PSICOV		
	Accuracy (%)			Accuracy (%)			Accuracy (%)		
	S [b]	M [c]	L [d]	S [b]	M [c]	L [d]	S [b]	M [c]	L [d]
T0529	28.1	37.7	3.5	7.0	4.4	4.4	28.1	37.7	5.3
T0531	23.1	0.0	38.5	7.7	7.7	0.0	23.1	0.0	30.8
T0534	28.6	22.1	7.8	10.4	6.5	10.4	28.6	16.9	15.6
T0537	17.1	17.1	22.4	22.4	39.5	67.1	17.1	18.4	50.0
T0544	37.0	22.2	18.5	18.5	11.1	33.3	44.4	25.9	40.7
T0547	40.2	18.0	59.8	19.7	19.7	35.3	41.8	18.0	61.5
T0550	45.6	33.8	22.1	7.4	14.7	23.5	45.6	33.8	39.7
T0553	46.4	17.9	7.1	17.9	14.3	25.0	50.0	17.9	32.1
T0555	36.7	26.7	13.3	16.7	13.3	30.0	43.3	30.0	43.3
T0561	25.0	6.3	9.4	3.1	3.1	6.3	25.0	6.3	12.5
T0571	37.7	24.6	13.0	13.0	17.4	24.6	37.7	26.1	18.8
T0578	24.2	36.4	39.4	3.0	12.1	9.1	24.2	36.4	36.4
T0581	25.9	11.1	3.7	3.7	3.7	0.0	22.2	11.1	3.7
T0604	51.8	45.5	13.6	19.1	24.6	50.9	50.9	45.5	32.7
T0608	51.8	41.1	10.7	30.4	50.0	48.2	53.6	41.1	35.7
T0616	61.9	28.6	4.8	33.3	0.0	28.6	61.9	23.8	28.6
T0618	13.9	19.4	11.1	5.6	8.3	0.0	19.4	19.4	11.1
T0621	8.8	8.8	11.8	5.9	2.9	5.9	8.8	8.8	11.8
T0624	87.5	37.5	18.8	12.5	18.8	6.3	87.5	37.5	6.3
T0629	4.7	0.0	32.6	14.0	11.6	7.0	4.7	0.0	32.6
T0637	20.7	20.7	10.3	6.9	0.0	27.6	27.6	20.7	31.0
T0639	23.1	11.5	3.9	3.9	3.9	3.9	23.1	15.4	3.9
Average	33.6	22.1	17.1	12.8	13.1	20.3	34.9	22.3	26.6

[a] Predictions were extracted from http://www.predictioncenter.org/download_area/CASP9/predictions/.

[b] S denotes short-range contacts defined as sequence separation of two residues between 6 and 11 residues.

[c] M denotes medium-range contacts defined as sequence separation of two residues between 12 and 23 residues.

[d] L denotes long-range contacts defined as sequence separation of two residues more than 23 residues.

Table S8. Performance comparison on different number of TM helices.

TMH	Accuracy (%)		Coverage (%)		Accuracy (δ=4) (%)	
	T [a]	M [b]	T [a]	M [b]	T [a]	M [b]
3-4	33.1	47.7	7.8	11.2	77.7	84.2
5-6	25.1	74.0	4.2	12.4	72.4	94.4
7	40.3	69.7	5.0	8.5	93.5	97.2
8-10	19.0	61.0	2.6	9.5	71.9	92.8
>10	20.9	62.5	2.2	6.7	80.1	88.9

[a] T denotes TMHcon.

[b] M denotes MemBrain.

Table S9. Performance comparison on the top *L*/2 and top *L* predictions.

Cutoff	Method	Accuracy (%)	Coverage (%)	Accuracy (δ=4) (%)
Top *L*/2	TMhhcp1	42.8	17.4	81.8
	TMhhcp2	37.5	15.0	79.3
	MemBrain	54.2	22.0	85.7
Top *L*	TMhhcp1	34.6	27.6	77.9
	TMhhcp2	30.2	24.0	76.4
	MemBrain	46.5	36.8	83.2

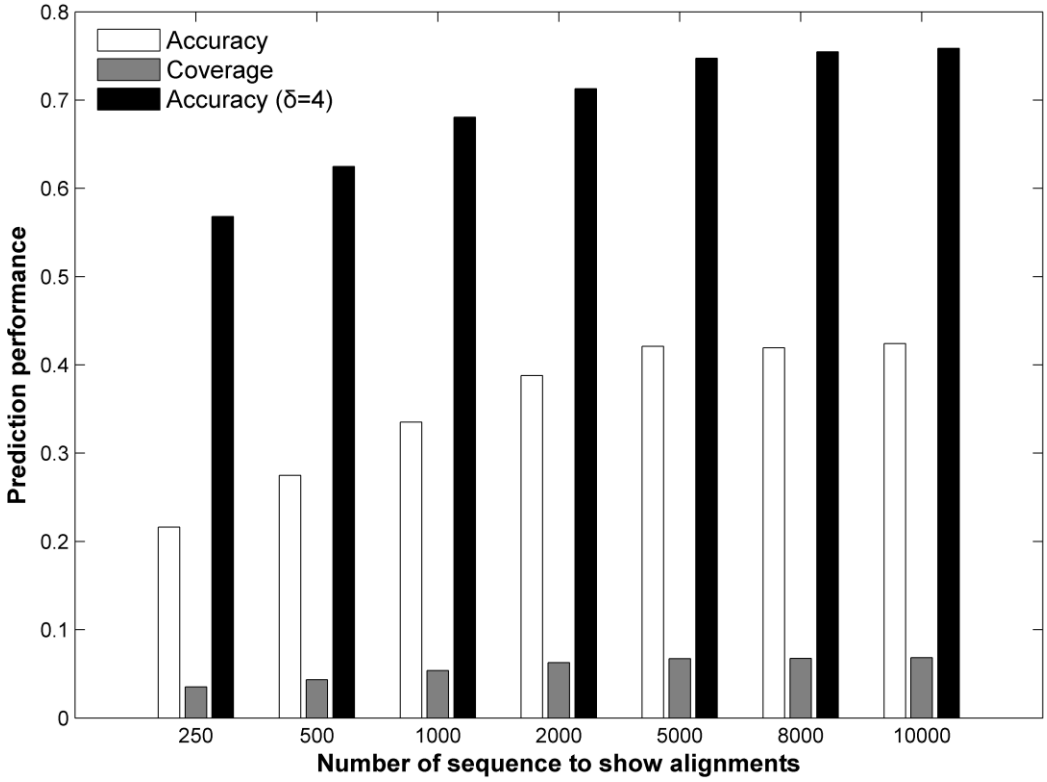


Figure S1. Performance of PSICOV depends on the number of homologous sequences searched by PSI-BLAST. When we set the number of aligned sequences to 250 with the -b parameter in PSI-BLAST program, the accuracy is only 21.6% with a coverage rate of 3.5%. When we increase this parameter, the prediction performance improves as well. When it reaches 5,000, the accuracy and the coverage are 42.1% and 6.7% respectively, which are 20.5% and 3.2% higher than those obtained at 250. We then tried to further increase this parameter to greater than 5,000, but found that the prediction performance did not change much. In particular, the prediction accuracy even reduced a little in the case of 8,000 compared to 5,000.

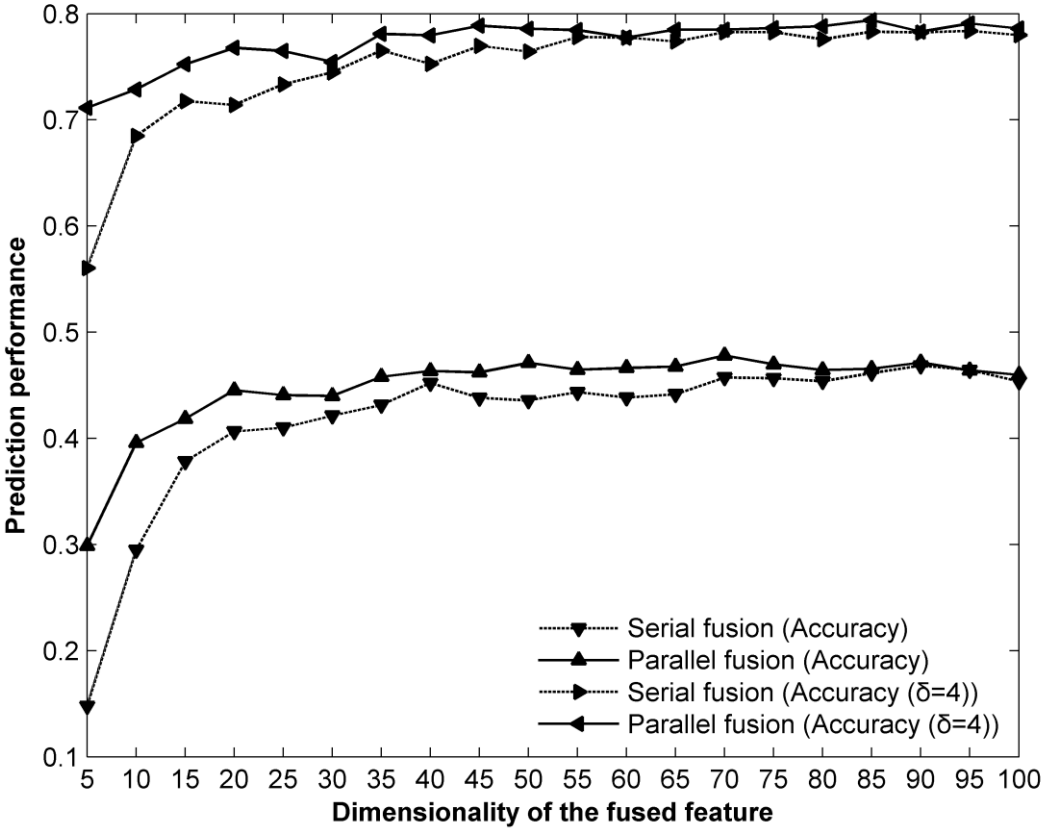


Figure S2. Performance comparisons of OET-KNN classifier for serial and parallel fusions on different reduced dimensionalities. As can be seen, the prediction performances of parallel fusion are consistently better than those of serial fusion using PCA algorithm, and thus the parallel fusion with reduced dimensionality of 70 is used for OET-KNN classifier.

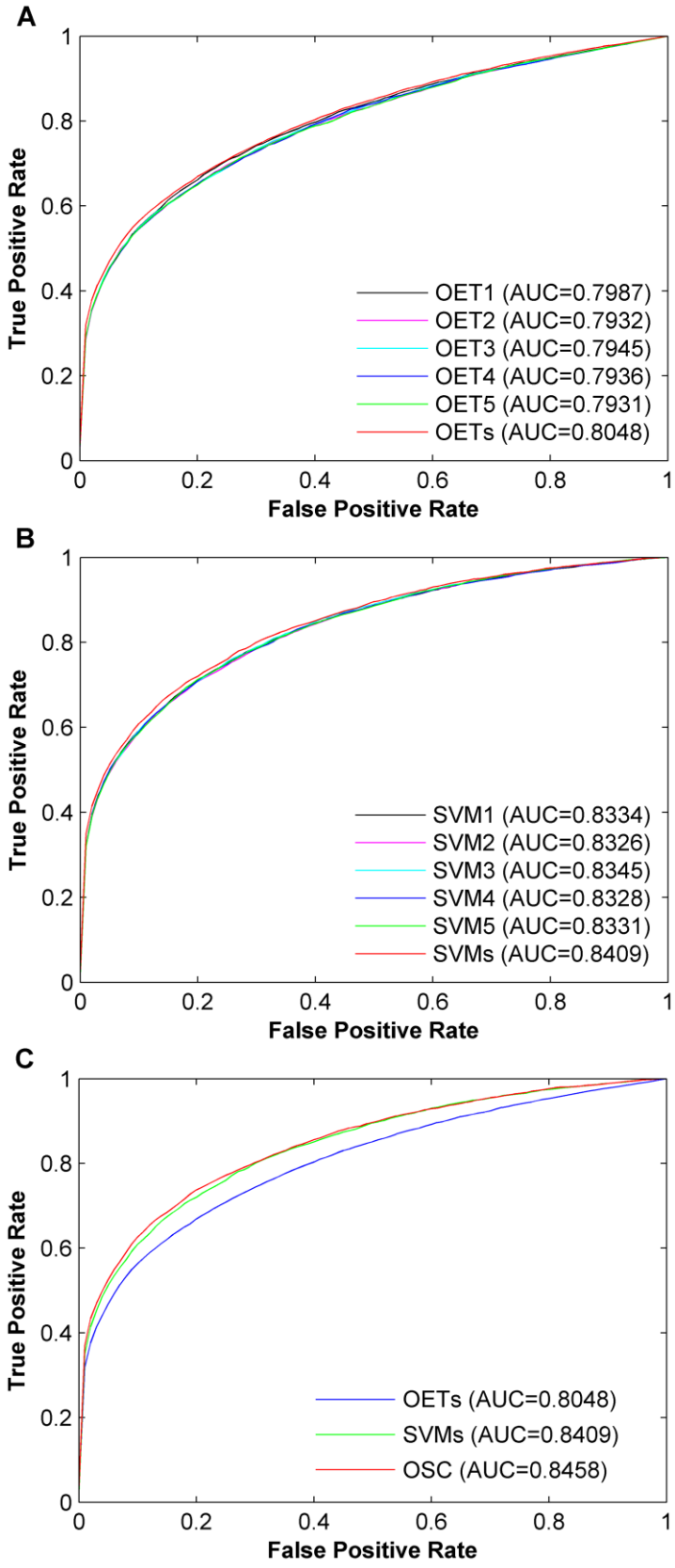


Figure S3. ROC curves of individual classifiers and combined classifier. (A) ROC curves of OET-KNN classifiers and OETs. (B) ROC curves of SVM classifiers and SVMs. (C) ROC curves of OETs, SVMs, and OSC.

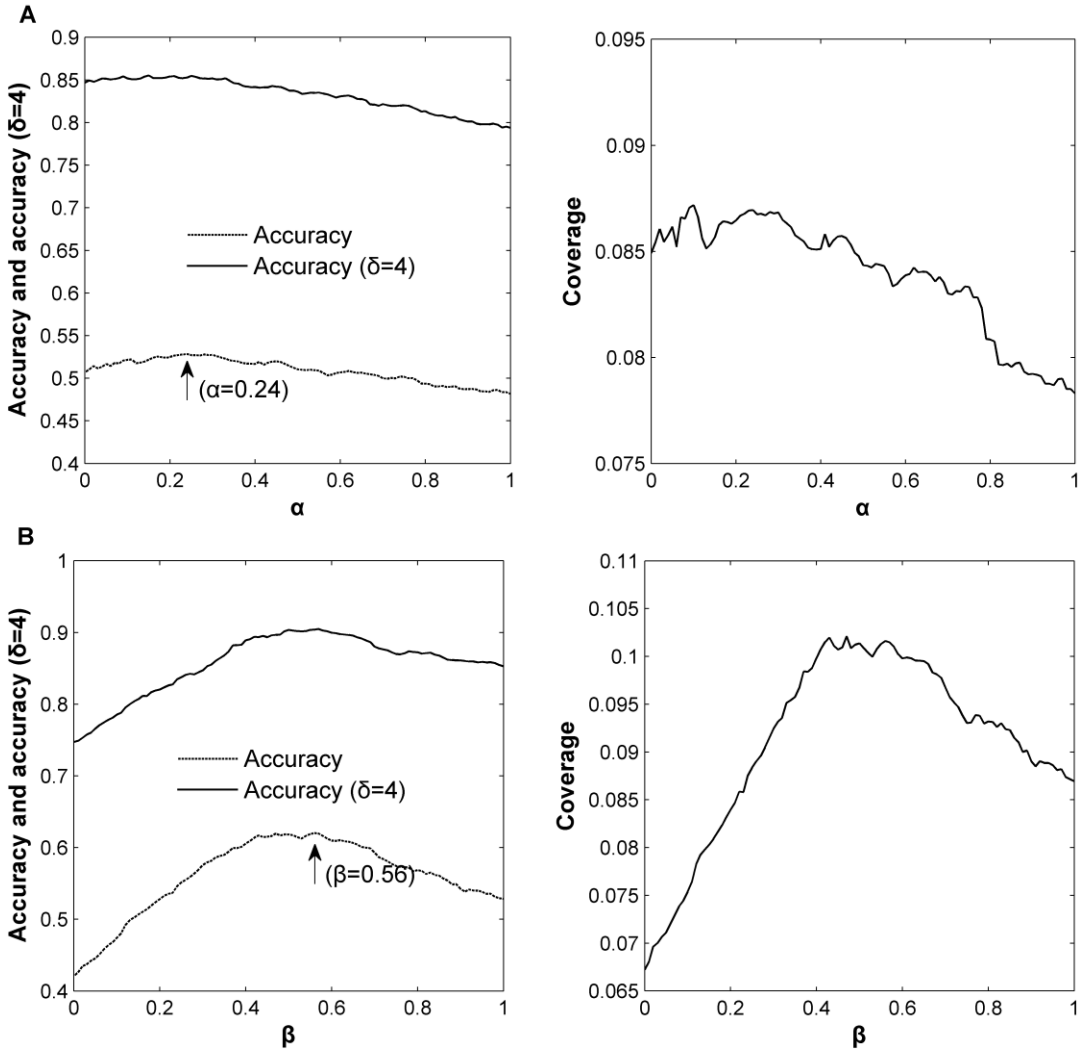


Figure S4. Performance of different weights for fusing. (A) Weights α for combining OETs and SVMs. (B) Weights β for combining OSC and PSICOV.

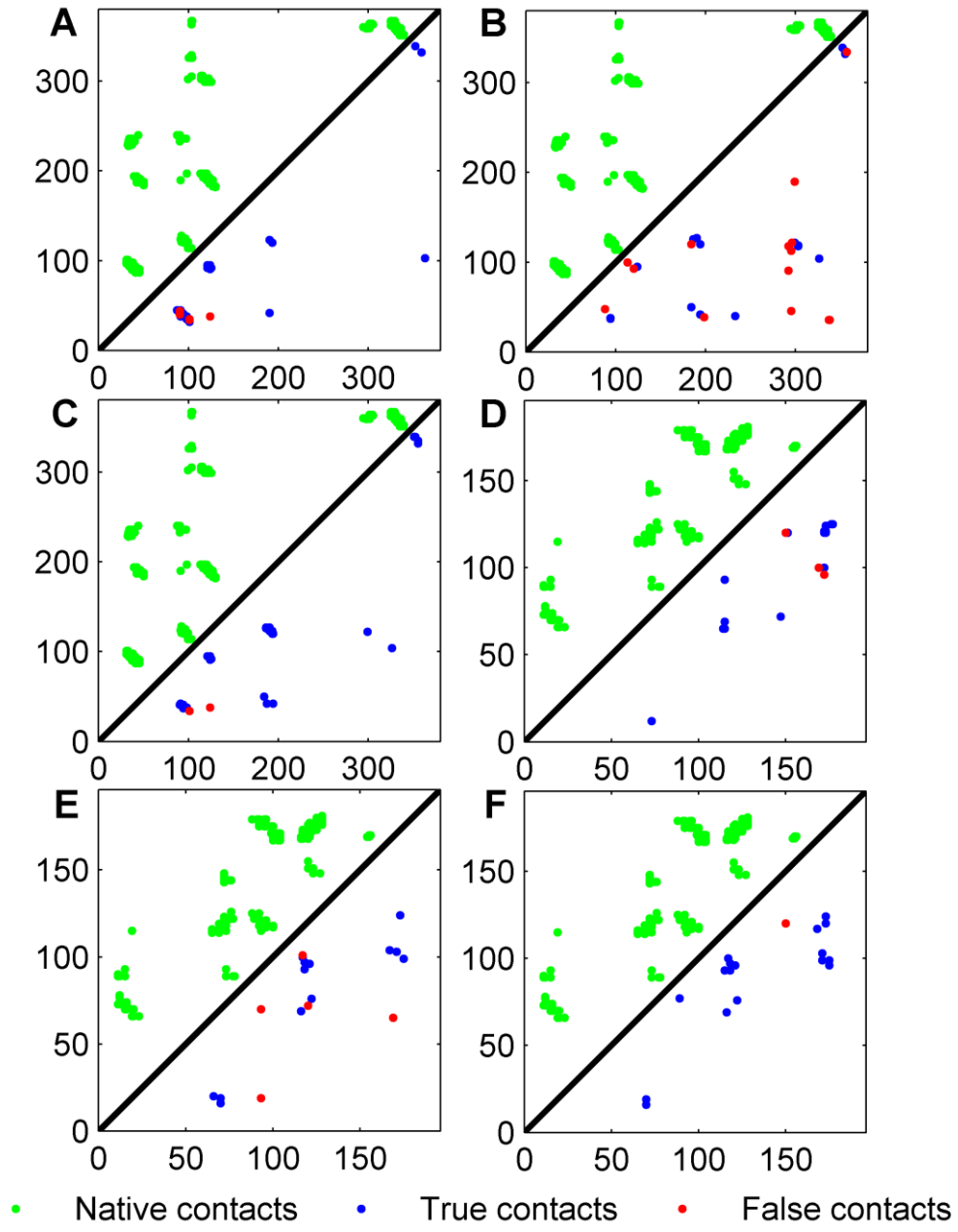


Figure S5. Top *L*/5 contacts predicted by OSC, PSICOV, and MemBrain. (A)-(C) Contact maps of protein 1bccC predicted by OSC, PSICOV and MemBrain respectively. (D)-(F) Contact maps of protein 2nr9A predicted by OSC, PSICOV and MemBrain respectively. As can be seen in Figures A-C, the predicted 14 spurious contacts by PSICOV are successfully eliminated with the assistance of OSC, but two new pseudo contacts are induced as well. Meanwhile, PSICOV reduces four false contacts predicted by OSC. Finally, only two false positives are predicted by MemBrain on 1bccC. In Figures D-F, the predicted 15 out of 18 contacts are native contacts obtained by OSC, while 13 out of 18 contacts are native contacts in PSICOV. The complementation of OSC and PSICOV improves the prediction performance by the fact that only one spurious contact out of 18 predicted contacts is predicted in the final MemBrain model on 2nr9A.

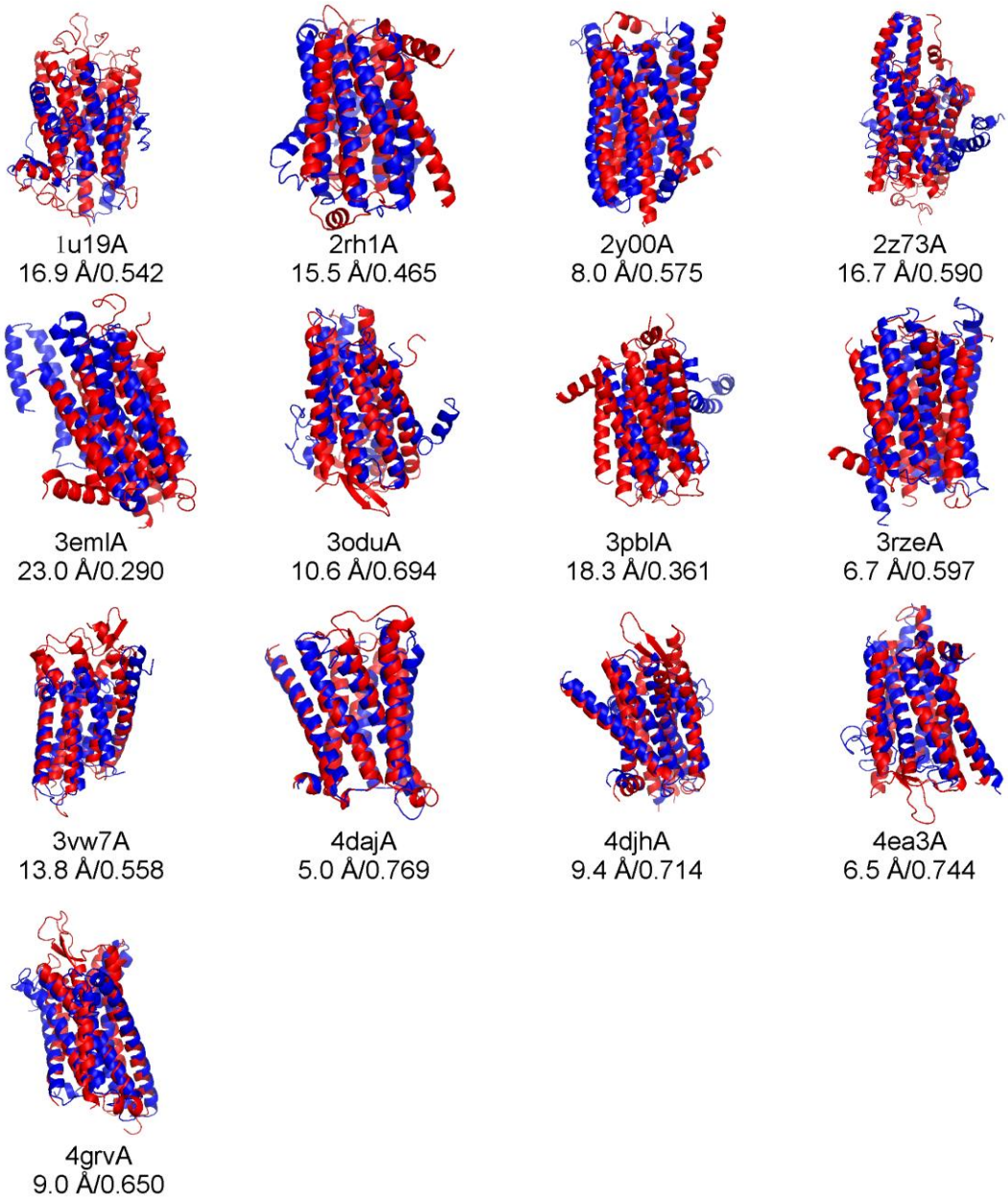
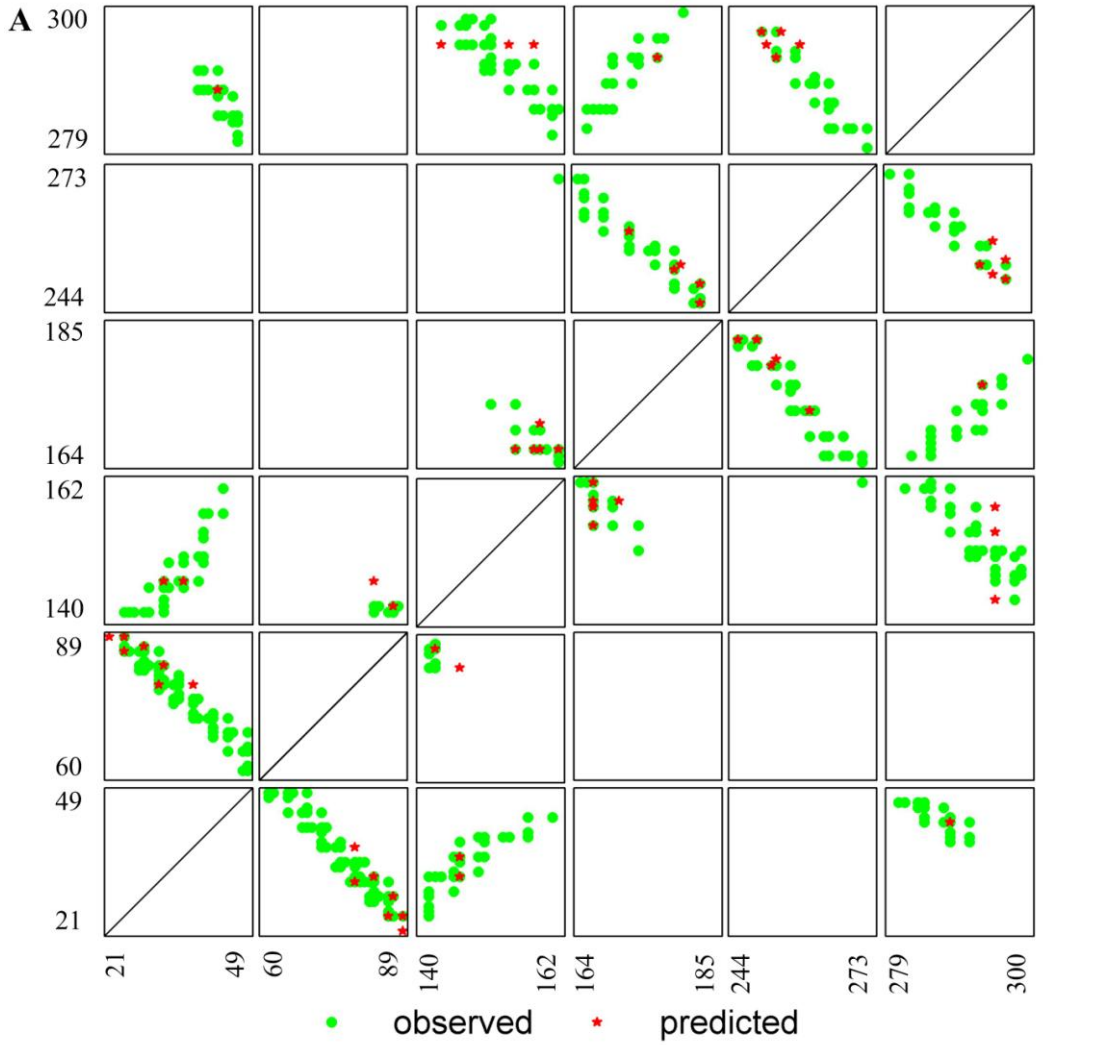
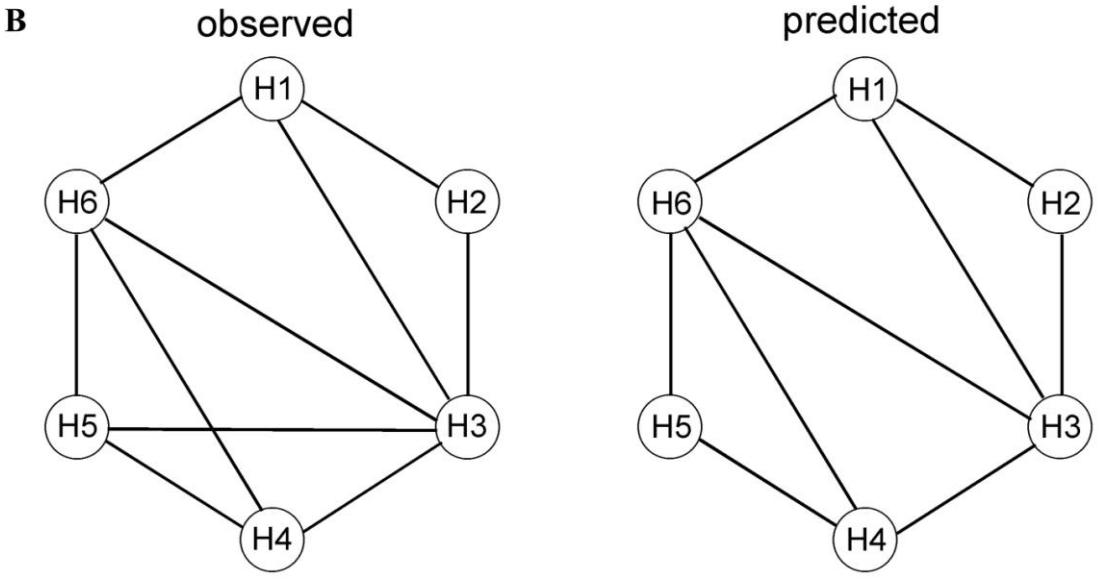


Figure S6. Superposition of the first model (blue) and the X-ray structure (red) in the whole-chain for 13 known GPCRs. Models are generated by I-TASSER with contact restraints from MemBrain and all GPCR templates have been excluded.


A
300
279
273
244
185
164
162
140
89
60
49
21
21
49
60
89
140
162
164
185
244
273
279
300
observed
predicted


B
observed
predicted
H1
H2
H3
H4
H5
H6
H1
H2
H3
H4
H5
H6

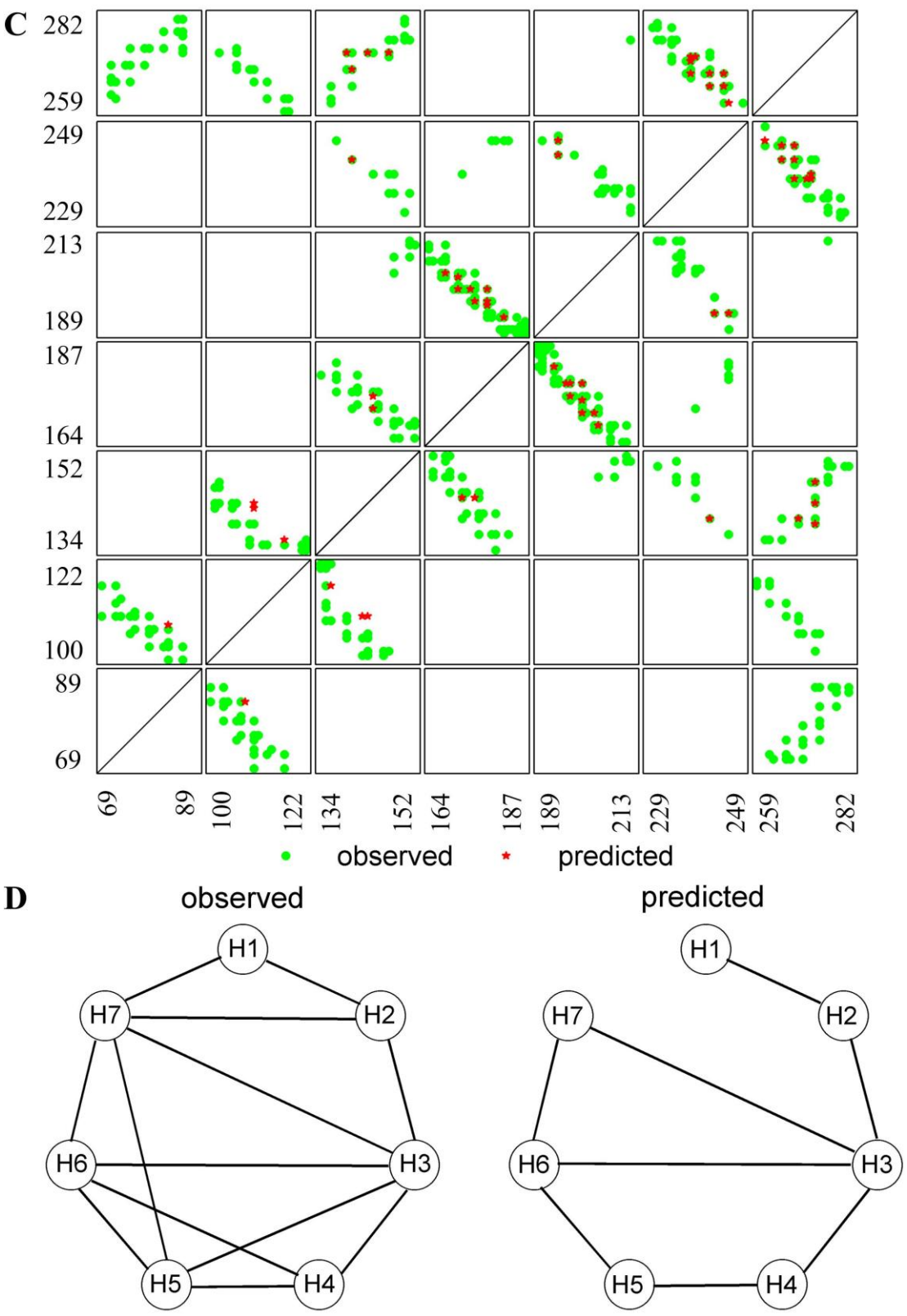


Figure S7. Observed and predicted contact maps and helix interaction patterns by MemBrain. (A) Predicted contact map of protein 3qf4A. (B) Predicted helix interaction pattern of protein 3qf4A. (C) Predicted contact map of protein 3ug9A. (D) Predicted helix interaction pattern of protein 3ug9A.

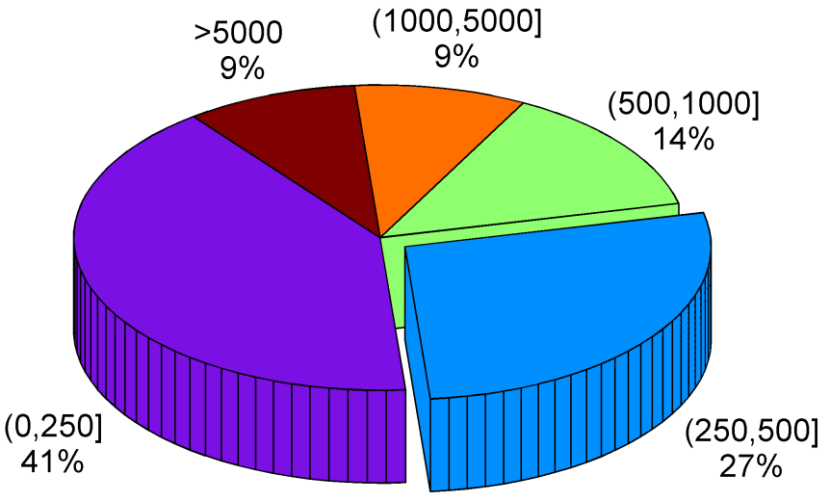


Figure S8. Distributions of homology sizes searched by PSI-BLAST against UniRef90 database for the 22 targets in CASP9.

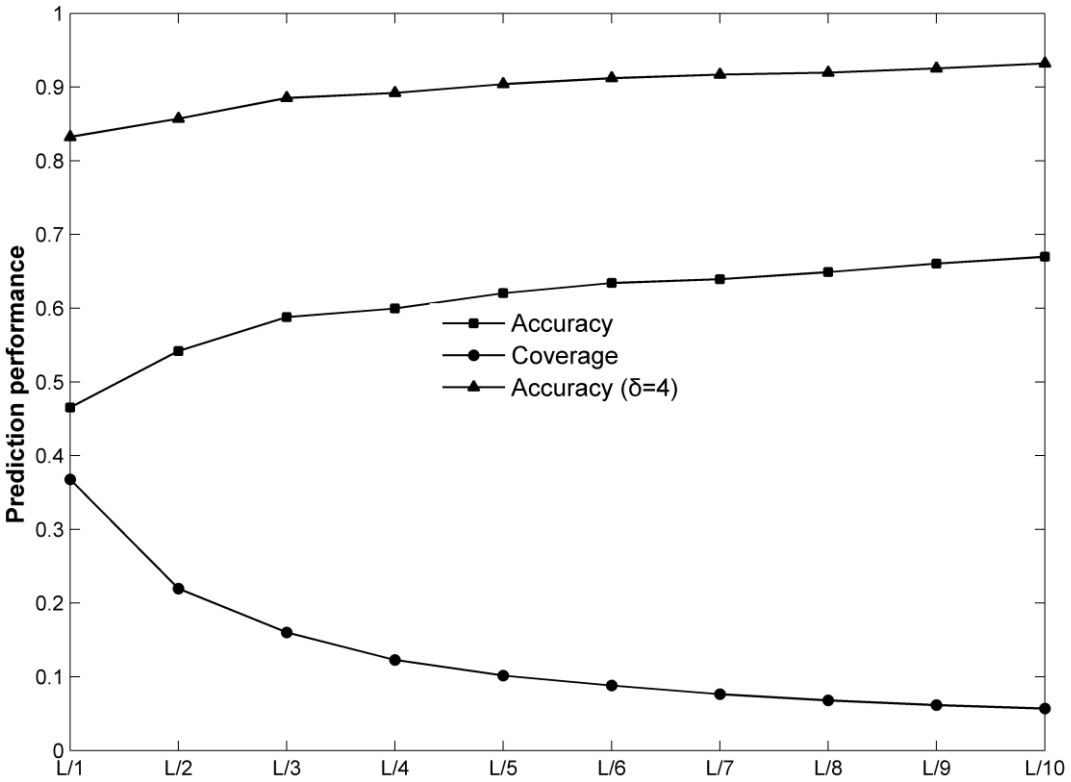


Figure S9. Performances of different cutoffs for the number of selected contacts.

REFERENCES

Fuchs,A. *et al*. (2009) Prediction of helix–helix contacts and interacting helices in polytopic membrane proteins using neural networks. *Proteins*, **74**, 857-871.

Jones,D.T. *et al*. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184-190.

Wang,X.F. *et al*. (2011) Predicting residue-residue contacts and helix-helix interactions in transmembrane proteins using an integrative feature-based random forest approach. *PLoS One*, **6**, e26767.