Supplementary Information

Supplementary Figures

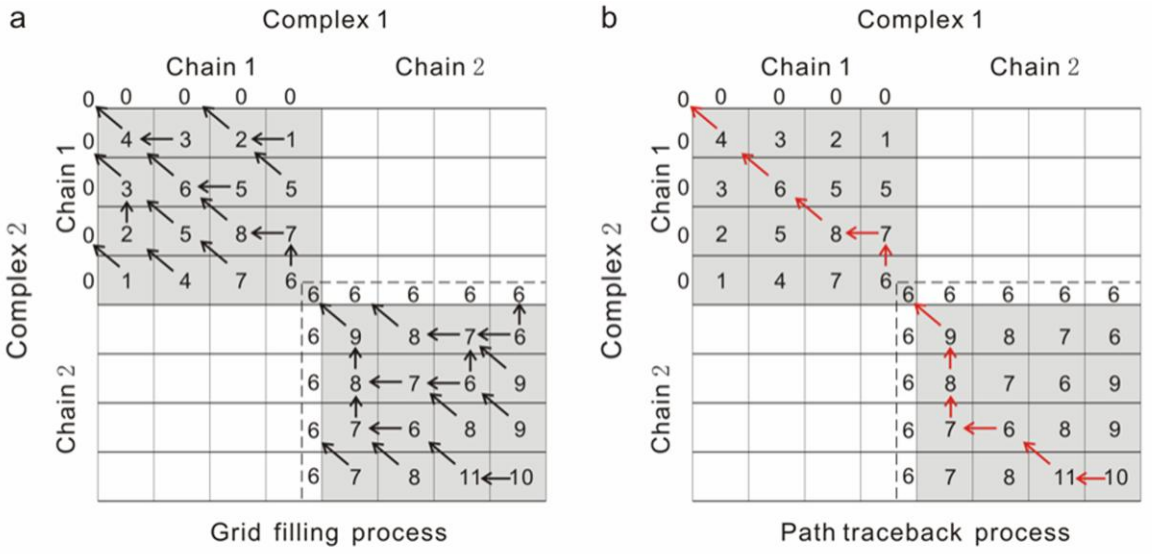

Figure S1: An illustration of the modified dynamic programming algorithm to prevent cross-chain alignment. The left panel shows the process of grid filling-up with the cross-alignment zone (empty grids) ignored. The dashed lines indicate a pseudo-layer which assumes the value of the last grid of the quadrant corresponding to Chain 1 of both complexes. The value of the pseudo-layer (6 in this example) is used as a starting score of the next quadrant corresponding to Chain 2 of both complexes. The picture on the right panel shows the trace-back path (indicated by red arrows).

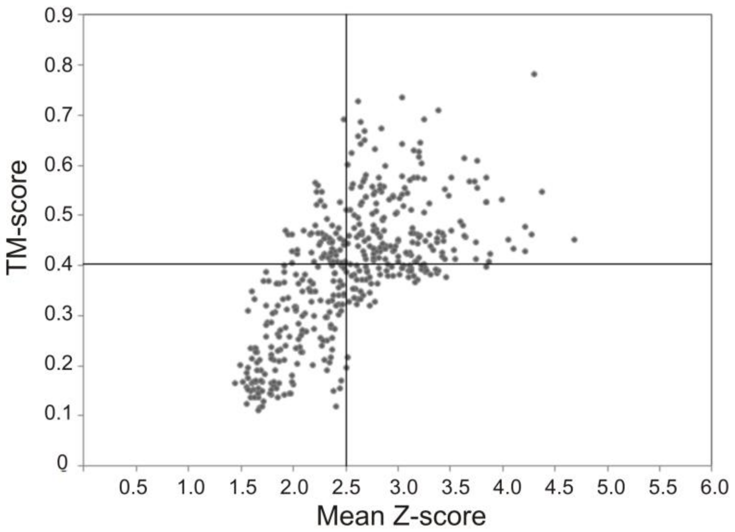

Figure S2: TM-score versus Z-score of the first COTH templates for 180 training proteins.

Supplementary Tables

Table S1. Mean distance (Å) between C_α atoms of amino acids that are in inter-chain contact.

	G	**A**	**V**	**L**	**I**	**S**	**T**	**C**	**M**	**P**	**D**	**N**	**E**	**Q**	**K**	**R**	**H**	**F**	**Y**	**W**
G	3.3	5.5	5.6	6.1	6.0	5.6	5.6	5.3	5.7	5.6	5.8	5.6	5.6	5.9	6.5	6.5	5.8	6.2	6.4	7.0
A	5.5	3.8	6.1	6.5	6.6	5.3	5.7	5.4	5.9	6.1	5.8	6.1	6.5	5.7	6.3	6.8	5.8	6.6	6.3	6.8
V	5.5	5.4	4.1	7.0	6.3	5.9	6.0	5.5	6.3	5.8	5.9	6.2	6.3	6.0	6.5	6.5	5.9	7.1	6.7	7.1
L	5.6	5.9	6.8	4.3	6.9	5.5	6.3	5.8	6.9	6.5	5.9	6.6	6.1	6.2	6.5	7.3	6.1	7.4	7.1	7.7
I	5.5	5.9	6.4	6.9	4.3	5.6	6.0	5.9	6.6	6.1	5.7	6.0	6.2	6.0	6.2	7.0	6.4	7.9	7.0	7.3
S	5.3	5.5	6.3	5.8	6.0	4.4	7.3	5.5	6.6	6.5	6.0	7.2	6.0	6.0	6.5	7.5	6.0	6.6	6.8	6.7
T	5.5	5.5	5.7	6.5	5.9	6.8	4.4	5.4	6.1	6.0	5.6	7.7	5.8	6.6	6.5	7.6	6.5	6.5	6.9	7.2
C	5.4	5.7	6.3	6.7	6.1	5.4	5.8	4.6	6.2	5.8	5.3	5.7	5.7	5.8	5.8	6.4	5.8	6.8	7.1	6.7
M	6.3	6.3	6.7	7.5	7.0	7.1	6.7	6.4	4.5	6.4	6.5	6.9	7.2	7.3	8.8	7.9	6.7	7.5	7.2	9.0
P	5.2	5.4	5.9	6.7	5.9	6.3	5.9	5.4	6.2	4.9	5.5	6.1	6.0	6.2	6.1	8.1	5.9	6.3	7.2	7.1
D	5.6	5.9	6.5	6.5	6.8	6.0	6.3	5.5	7.1	5.8	4.9	6.9	7.1	6.6	7.7	8.0	6.7	6.4	7.5	7.1
N	5.5	5.8	6.3	7.1	6.1	6.8	7.8	5.5	6.2	5.9	6.3	4.3	6.5	6.4	7.6	8.3	6.4	6.6	7.3	6.7
E	6.1	6.0	6.6	6.7	6.9	6.6	6.2	5.1	6.5	6.2	6.6	6.7	4.5	6.5	7.7	8.0	6.9	7.1	7.4	8.6
Q	6.3	5.9	6.8	7.4	6.8	6.5	6.6	6.2	7.2	6.6	6.9	6.8	6.9	4.6	6.9	7.5	7.2	7.4	8.0	7.5
K	5.8	6.0	6.1	6.6	6.4	6.0	6.2	5.4	7.1	6.0	7.1	6.4	6.8	6.4	4.2	7.2	6.9	7.0	7.2	6.9
R	6.5	6.8	7.2	7.1	7.2	7.5	7.7	6.0	7.9	7.6	7.8	7.8	8.0	7.2	7.3	4.8	7.3	7.9	8.0	8.5
H	5.9	6.3	6.3	7.1	7.0	6.2	6.9	6.5	6.8	6.5	7.4	6.7	7.3	7.0	8.3	8.2	4.1	7.1	7.6	7.6
F	6.0	6.2	7.6	7.6	7.5	6.3	7.2	6.4	7.3	6.3	6.3	6.8	6.7	6.9	7.9	7.9	7.0	5.8	7.6	8.8
Y	6.3	6.7	7.0	7.2	7.4	6.8	7.0	6.3	7.4	7.1	7.2	7.2	7.7	7.3	7.6	7.8	7.3	7.7	5.0	8.3
W	7.1	7.9	8.6	9.0	7.9	7.8	9.2	6.4	8.5	7.5	8.6	7.6	8.7	6.8	9.4	9.7	7.9	8.9	7.9	5.8

Table S2. Standard deviation (Å) of C_α distance for amino acids that are in inter-chain contact.

	G	**A**	**V**	**L**	**I**	**S**	**T**	**C**	**M**	**P**	**D**	**N**	**E**	**Q**	**K**	**R**	**H**	**F**	**Y**	**W**
G	0.8	2.1	2.3	2.6	2.5	2.0	2.1	1.1	1.2	2.3	2.5	1.5	1.7	2.1	2.7	2.0	1.6	2.4	2.1	2.4
A	2.3	1.3	2.9	2.8	2.8	1.7	1.9	1.1	1.7	2.5	2.4	2.2	3.2	2.0	3.2	3.2	1.8	2.8	2.0	2.4
V	2.3	2.0	1.3	2.9	2.3	2.4	2.7	1.2	1.7	2.1	2.5	2.6	2.9	2.2	2.9	2.5	1.7	2.8	2.5	2.1
L	2.2	2.3	2.8	1.5	2.3	1.8	2.7	1.3	2.3	2.8	2.4	2.7	2.4	2.3	3.2	3.1	1.8	2.6	2.2	2.3
I	2.5	2.6	2.7	2.6	1.4	2.1	2.2	2.1	1.9	2.6	2.5	2.5	3.0	2.1	2.8	3.0	2.3	3.2	2.6	2.2
S	2.1	2.3	2.9	1.9	0.6	1.4	3.5	1.2	3.0	2.9	3.0	3.3	2.4	1.8	3.1	3.1	1.8	2.7	2.2	2.2
T	2.3	2.2	2.2	2.7	1.8	3.3	1.5	1.2	1.8	2.7	1.6	3.5	1.6	2.7	2.9	2.9	2.3	2.0	2.4	2.7
C	2.2	2.5	2.8	2.8	2.3	1.7	2.5	1.8	1.3	2.4	1.9	1.9	2.1	1.7	2.0	2.5	1.3	2.6	2.7	2.4
M	2.5	2.6	2.6	2.8	2.8	2.7	2.7	2.2	1.6	2.5	2.5	2.8	3.3	3.2	3.4	3.3	2.3	2.4	2.0	2.9
P	1.8	2.1	2.2	2.9	2.1	2.7	2.3	1.3	2.0	1.9	1.6	2.4	2.7	1.6	2.6	3.6	1.4	2.2	2.2	2.4
D	2.6	2.9	3.2	2.9	3.5	2.6	2.9	2.1	3.7	2.3	1.3	3.1	3.6	2.8	3.4	2.8	1.9	2.9	2.7	2.7
N	2.1	2.3	2.8	3.2	2.5	3.1	3.6	1.3	2.1	2.0	2.8	1.9	3.0	2.2	4.2	3.2	2.1	2.6	2.7	1.5
E	2.6	2.6	3.1	2.9	3.1	2.7	2.4	1.6	2.5	2.4	2.7	2.6	1.4	2.4	2.8	2.9	2.0	3.1	2.3	3.1
Q	2.9	2.6	3.1	3.2	2.9	2.7	2.8	2.3	3.1	2.9	3.0	2.7	3.1	1.2	3.1	2.5	2.7	2.9	2.9	2.9
K	2.2	2.7	2.7	2.8	3.0	2.1	2.3	1.4	3.3	2.4	3.1	2.5	2.6	2.0	1.6	3.1	2.8	3.0	2.5	2.1
R	2.4	2.9	3.1	2.7	2.8	2.8	3.1	1.7	3.4	2.9	2.9	2.9	2.9	2.6	3.2	1.9	2.6	3.2	2.4	3.0
H	2.3	2.7	2.1	2.8	2.6	2.3	2.8	2.2	2.3	2.6	2.9	2.4	2.6	2.7	4.1	3.1	1.1	2.5	2.2	1.8
F	2.1	2.2	3.1	2.6	2.6	2.1	3.0	1.4	2.0	2.0	2.4	2.3	2.7	2.1	3.7	3.0	2.2	1.6	2.3	2.4
Y	2.0	2.7	2.7	2.0	2.6	2.2	2.6	1.8	2.2	2.3	2.0	2.5	2.9	2.2	2.9	3.1	2.3	2.6	1.3	2.7
W	3.1	3.3	3.6	3.4	2.7	3.0	3.7	2.2	3.1	2.9	3.6	3.0	3.8	2.4	4.1	4.1	2.9	3.2	2.7	1.9

Table S3. Summary of I-RMSD (Å) of predicted models by different methods.*

Name	ZDOCK-exp	ZDOCK-model	COTH	COTH-exp	COTH-model
1avxA-1avxB	2.16 (5)	4.43 (2)	4.44 (1)	5.00	4.81
1ay7A-1ay7B	11.77 (6)	13.87 (4)	8.96 (3)	11.64	12.53
1bvnP-1bvnT	1.97 (10)	3.64 (7)	4.53 (1)	6.11	6.64
1cgiE-1cgiI	9.63 (1)	12.57 (2)	3.86 (8)	4.30	4.78
1d6rA-1d6rI	8.04 (4)	11.09 (1)	13.54 (10)	14.86	15.30
1dfjE-1dfjI	2.09 (10)	2.22 (1)	7.48 (1)	7.87	7.74
1e6eA-1e6eB	2.33 (5)	3.83 (5)	1.52 (3)	2.42	3.14
1eawA-1eawB	3.21 (1)	3.87 (4)	7.86 (6)	8.89	7.18
1ewyA-1ewyC	2.26 (7)	4.51 (8)	6.37 (4)	8.72	9.50
1f34A-1f34B	13.62 (10)	16.74 (10)	11.25 (1)	13.75	14.58
1mahA-1mahF	1.35 (8)	2.34 (1)	2.43 (2)	3.55	3.54
1ophA-1ophB	6.20 (2)	6.62 (3)	7.57 (9)	9.40	10.01
1ppeE-1ppeI	1.29 (1)	3.40 (4)	5.08 (1)	6.66	6.85
1tmqA-1tmqB	12.33 (7)	13.67 (7)	14.29 (1)	14.13	14.18
1udiE-1udiI	3.86 (2)	7.86 (6)	2.36 (3)	3.68	3.92
2b42B-2b42A	1.85 (1)	4.60 (5)	1.04 (4)	3.07	3.68
2o8vA-2o8vB	11.05 (8)	14.89 (5)	4.09 (5)	4.74	4.86
2pccA-2pccB	9.44 (3)	12.85 (2)	7.31 (5)	8.89	9.42
2sicE-2sicI	1.69 (4)	2.94 (1)	3.42 (6)	2.48	3.50
2sniE-2sniI	6.42 (8)	7.08 (1)	3.08 (4)	4.35	4.91
2uuyA-2uuyB	2.49 (4)	4.19 (9)	3.48 (1)	3.40	3.70
7ceiA-7ceiB	2.22 (6)	5.74 (10)	5.93 (1)	7.76	8.38
1ak4A-1ak4D	6.77 (9)	9.52 (2)	10.60 (7)	11.55	9.97
1b6cA-1b6cB	2.77 (1)	4.35 (3)	2.72 (10)	3.79	4.15
1buhA-1buhB	13.97 (5)	14.44 (1)	7.06 (8)	8.30	8.72
1e96A-1e96B	2.72 (7)	5.71 (8)	9.45 (9)	11.90	12.72
1efnB-1efnA	8.29 (8)	9.61 (7)	2.17 (1)	1.94	2.32
1fc2C-1fc2D	5.36 (7)	8.94 (3)	3.50 (7)	5.87	6.67
1fqjA-1fqjB	14.71 (3)	17.26 (2)	19.40 (4)	18.52	18.82
1gcqB-1gcqC	9.68 (9)	12.44 (5)	5.02 (5)	6.10	6.31
1ghqA-1ghqB	11.5 (7)	13.25 (4)	6.90 (6)	6.35	6.53
1glaG-1glaF	2.50 (1)	6.38 (2)	8.68 (2)	8.64	8.65
1gpwA-1gpwB	2.03 (8)	2.56 (3)	4.46 (9)	6.42	7.08
1he1C-1he1A	8.03 (8)	10.89 (1)	4.35 (2)	7.21	8.17
1j2jA-1j2jB	3.93 (3)	5.41 (8)	8.83 (10)	11.40	12.26
1kacA-1kacB	2.18 (10)	2.51 (10)	3.43 (1)	4.00	4.81
1ktzA-1ktzB	9.46 (1)	13.19 (6)	6.58 (3)	8.92	9.69
1kxpA-1kxpD	3.27 (1)	4.69 (2)	3.08 (4)	4.98	5.29
1qa9A-1qa9B	12.65 (2)	14.72 (10)	8.18 (6)	10.72	11.57
1s1qA-1s1qB	13.32 (1)	16.62 (1)	19.03 (5)	20.58	21.09
1sbbA-1sbbB	9.73 (4)	9.78 (5)	3.53 (4)	2.70	2.98
1t6bX-1t6bY	6.48 (6)	8.84 (6)	4.68 (1)	5.86	6.89
1xd3A-1xd3B	4.06 (6)	4.78 (8)	4.81 (8)	6.38	6.90
1z0kA-1z0kB	2.06 (3)	2.37 (8)	8.39 (3)	9.13	9.38
1z5yD-1z5yE	8.31 (1)	10.98 (7)	1.36 (3)	2.41	3.67
1zhiA-1zhiB	9.56 (9)	12.31 (1)	13.22 (2)	15.75	16.59
2ajfA-2ajfE	11.59 (10)	12.36 (9)	14.89 (5)	17.82	18.79
2btfA-2btfP	13.03 (8)	13.65 (7)	5.76 (7)	7.92	8.65

2hleA-2hleB	2.34 (1)	3.25 (4)	4.91 (1)	6.99	7.35
2hqsA-2hqsH	12.22 (10)	13.65 (8)	3.93 (4)	4.65	4.92
2oobA-2oobB	5.25 (5)	5.35 (1)	7.94 (7)	10.19	10.95
2i25N-2i25L	8.57 (5)	11.62 (5)	3.94 (1)	5.44	5.95
1kxqH-1kxqA	2.02 (1)	4.96 (6)	2.97 (6)	2.60	3.20
1acbE-1acbI	4.65 (3)	4.85 (10)	4.26 (2)	5.00	5.75
1m10A-1m10B	17.47 (4)	18.09 (3)	10.47 (1)	12.07	12.61
1nw9B-1nw9A	8.43 (6)	9.53 (1)	3.82 (1)	3.56	4.47
1grnA-1grnB	16.78 (9)	17.05 (8)	17.59 (7)	18.95	19.41
1he8B-1he8A	32.26 (6)	35.75 (7)	26.22 (4)	27.65	28.12
1i2mA-1i2mB	13.50 (2)	15.32 (2)	9.75 (7)	11.99	12.74
1wq1R-1wq1G	8.12 (9)	10.55 (3)	13.32 (1)	15.04	15.61
1xqsA-1xqsC	9.16 (5)	9.42 (3)	11.27 (6)	12.47	10.47
2cfhA-2cfhC	8.79 (1)	12.08 (7)	2.22 (1)	1.56	2.78
2h7vA-2h7vC	12.04 (1)	14.83 (1)	5.18 (6)	6.66	7.16
2hrkA-2hrkB	10.27 (5)	12.46 (10)	13.86 (8)	16.38	17.21
2nz8A-2nz8B	5.57 (6)	7.03 (7)	14.57 (6)	15.04	16.16
1fq1A-1fq1B	13.56 (8)	13.89 (5)	14.22 (3)	15.75	16.26
1pxvA-1pxvC	13.87 (3)	17.42 (6)	5.59 (5)	5.75	6.81
1atnA-1atnD	17.54 (9)	17.60 (1)	17.81 (1)	20.51	21.40
1bkdR-1bkdS	15.58 (2)	16.06 (3)	10.73 (4)	13.00	13.75
1h1vA-1h1vG	19.06 (7)	20.12 (10)	23.10 (10)	23.08	23.08
1ibrA-1ibrB	9.00 (5)	9.29 (5)	4.72 (2)	4.81	5.00
1iraY-1iraX	21.93 (1)	25.07 (2)	7.91 (6)	7.60	7.70
1r8sA-1r8sE	7.57 (7)	10.99 (10)	14.43 (7)	15.95	16.11
1y64A-1y64B	19.62 (4)	21.33 (1)	14.09 (5)	16.84	17.76
2c0lA-2c0lB	9.81 (3)	9.83 (2)	5.70 (3)	7.37	7.92
2ot3B-2ot3A	4.67 (7)	5.55 (8)	4.03 (9)	4.50	4.97
1r0rE-1r0rI	7.68 (7)	9.59 (1)	12.63 (4)	13.83	14.23
Average I-RMSD	8.47 (5.0)	10.30 (4.8)	7.92 (4.3)	9.17	9.64
Median I-RMSD	8.29	9.78	6.37	7.76	7.92

*****The table shows comparison of the docking methods (ZDOCK-exp and ZDOCK-model) and threading based methods (COTH, COTH-exp and COTH-model) in terms of I-RMSD. The values in the table indicate the I-RMSD of the best in top 10 (as ranked by the independent programs) models while the values in parentheses indicate the rank of the models. The ranks for COTH-exp and COTH-model are the same as that for COTH.

Experimental Procedures

1. Protein-Protein Interface Prediction by BSpred

To better identify the orientation of one protein chain relative to another in protein-protein complexes, prior knowledge about the interface residues of both chains is helpful. Accordingly, we develop a new machine-learning method called BSpred, which is capable of predicting the binding status of each residue from the amino acid sequence alone.

The input features of BSpred are the following 1) The Position Specific Scoring Matrix (PSSM) generated by PSI-BLAST search using an E-value cutoff =0.001. 2) The secondary structure (SS) of the query sequence, predicted by PSI-PRED (Jones, 1999), which is to detect the SS preference at the interface residues. The SS is represented by a 3-element vector ([0 0 1] for random coil, [0 1 0] for alpha helix, [1 0 0] for beta strands). 3) The solvent accessibility (SA) predicted from an independent neural network predictor (Chen and Zhou, 2005; Wu et al., 2007). The predicted solvent accessibility (whether buried or exposed) is a 2-element vector ([0 1] for buried, [1 0] for exposed). 4) The distinctive hydrophobicity of amino acids in protein-protein interfaces. Each amino acid is assigned by a hydrophobicity score, taken from the Eisenberg hydrophobicity scale (Eisenberg et al., 1984) which lies between 0 and 1 for all the amino acids. The NN software used in BSpred is from Fast Neural Network (FNN) (Nissen and Nemerson). By trial and error, we choose 3 layers with 50 hidden neurons for NN, which gives the best performance on training data. The training algorithm of NN is the standard Back-Propagation (BP) algorithm.

For prediction of interface residues, the neighboring residues around a central residue also contribute to the formation of interface (Ofran and Rost, 2003, 2007). We therefore use a window size of 21 to specify the *i*th residue, which includes residue indices from *i*-10 to *i*+10. Since there are 26 (=20+3+2+1) feature values for a residue, the number of features for a window around the trained residue is 546 (=21 × 26). At the N and C terminals, the input values for the neighboring residues which are not present are represented by 0. The NN output value is between -1 and 1 for each residue where larger values indicate higher confidence for that particular residue to be at the interface. Accordingly, a carefully optimized cutoff value (to obtain a balance between accuracy and coverage of prediction) is selected based on the performance on a set of training proteins which is non-redundant to the testing proteins of this work. Any residue with an output value higher than the cutoff is considered as an interface residue. We found that the NN output cutoff =-0.1 have the best balance of accuracy and coverage.

Based on the observation that interface residues are often sequentially clustered together (Ofran and Rost, 2003), we introduce a second-step post-processing for smooth filtering of raw neural network predictions, i.e. a residue with NN output score >-0.1 is finally considered as an interface residue only if at least 6 other residues in its direct sequence neighborhood (from i-3 to i+3) are also predicted to be interface residues (NN output score>-0.1). For the N-terminal and C-terminal residues, at least 3 neighboring residues should be at the interface. Also, since an interface residue must be solvent exposed at the monomer structure, any predicted interface residues which were not predicted to be solvent exposed are eliminated from our final interface predictions.

The method has been tested on a set of 150 single-chain proteins which are non-redundant to the training proteins and are known to participate in dimer formation. For assessment of the interface and chain orientation predictions, we define the *Accuracy* and *Coverage* of interface residues as

$$Accuracy = \frac{\text{No. of residues correctly predicted to be interface residues}}{\text{No. of residues predicted to be interface residues}} \quad (1)$$

$$Coverage = \frac{\text{No. of residues correctly predicted to be interface residues}}{\text{No. of actual interface residues in native}} \quad (2)$$

where an "actual interface residue" is defined as the residue whose Cα atom lies within 10Å of any Cα atoms of any residues in the opposite chain.

The final accuracy of the interface prediction by BSpred is 65.6% with coverage of 13.7%. The BSpred program and the on-line server are freely available at http://zhanglab.ccmb.med.umich.edu/BSpred.

2. Dynamic Programming Alignment for Dimeric Proteins

COTH conducts the query-template complex alignments by first joining the component sequences into artificial chains and then aligns both query chains simultaneously onto the templates. The simultaneous alignments can efficiently increase the cooperativity of multiple-chain alignments as well as facilitate the interface matches. However, the unphysical cross-alignments, i.e. residues on one chain of a complex are aligned to the residues on both chains of another complex, can be created from the conventional dynamic programming (DP) (Needleman and Wunsch, 1970; Smith and Waterman, 1981). To prevent the cross-alignments, we implemented a modified Needleman-Wunsch DP algorithm, as demonstrated in Figure S1. The regions in the DP matrix corresponding to cross-chain alignments are ignored. For example, if Chains 1 and 2 of Complex 1 are to be aligned to Chains 1 and 2 of Complex 2 respectively, the DP matrix regions corresponding to aligning chain 2 of Complex 1 with chain 1 of Complex 2 are omitted when filling up the alignment paths during DP. The filling up of the DP matrix is implemented in a three-step process (Figure S1a): 1) The region corresponding to the first chains of both complexes is filled up. 2) A pseudo-layer uniformly assumes the value of the last cell of the preceding block; by doing this the gap extension penalty is ignored at the respective first residues of the second chains. 3) The region corresponding to the second chains of both complexes is filled up starting from the pseudo-layer values. Finally, an alignment of the two complexes is constructed by tracing back the maximum alignment scores from the right-bottom lattice (Figure S1b). Since the algorithm neglects half of the lattice grids, it saves ~50% of the CPU time while preventing the cross alignments.

3. Template Selection and Target Classification

The significance of a threading alignment in COTH is assessed by Z-score:

$$\mathrm{Z-score} = \frac{R_{\mathrm{score}} - \langle R_{\mathrm{score}} \rangle}{\sqrt{\langle R^2_{\mathrm{score}} \rangle - \langle R_{\mathrm{score}} \rangle^2}} \quad \text{(S3)}$$

where R_{score} is the raw alignment score R'_{score} from the dynamic programming normalized by the length of the query dimer sequence (L_{query}) i.e. R_{score}=R'_{score}/L_{query}. Because the dynamic programming of COTH uses an unique path for both chains, the overall raw alignment score and the Z-score has a bias towards the larger of the two chains, especially when the receptor is significantly larger than the ligand; this may lead to artificially high Z-scores even though the ligand is poorly aligned. To balance this bias, we rank the COTH models based on the mean Z-score of the ligand, the receptor and the complex:

$$\text{Mean Z-score} = (\text{Z-score}_{\text{complex}} + \text{Z-score}_{\text{receptor}} + \text{Z-score}_{\text{ligand}}) / 3 \qquad \text{(S4)}$$

In Figure S2, we show the mean Z-score versus the TM-score of the first COTH models of the 180 training proteins. There is a positive correlation between Z-score and TM-score with correlation coefficient=0.77. Accordingly, we categorize the query proteins into “easy” or “hard” targets based on the Z-score, i.e. when a query has at least one template alignment with an average Z-score >2.5 we define it as an “easy” target; and otherwise it is labeled as a “hard” target. When considering templates above 0.4 to be reliable, the false positive and false negative rates of Z-score=2.5 are 8.2% and 5.1%, respectively, for the training proteins. When applying this definition of Z-score to the 500 test proteins, we have 296 cases that are “easy” targets and 204 that are “hard” targets. The average TM-score for “easy” and “hard” proteins are 0.478 and 0.245, respectively. These data demonstrate that the Z-score can be used as a reliable indicator of the template quality.

When superimposing and combining monomer and dimer templates, we take the top-five templates from MUSTER for each chain; each of the monomer templates is then superimposed on the top-ten dimer templates from COTH, which results in 250 dimeric structures (=$5\times5\times10$). To rank the 250 structures, we structurally aligned each of the structure to other 249 structures by the multimeric structure alignment program MM-align (Mukherjee and Zhang, 2009) and calculate the average TM-score of the structure compared with others. The structure of the highest TM-score to other template, which means a consensus, is selected as the final COTH model.

4. Alternative Threading Approaches

To assess in detail the strength and weakness of the COTH approach, we used several alternative threading programs which are described here. PSI-BLAST(Altshucl et al., 1997) was run on our complex structure library after joining the two chains together using the BLASTP program. The templates were ranked according to the PSI-BLAST E-value. The second control used in this work was C-PPA algorithm which is an extension of our monomer threading algorithm PPA (Zhang, 2007). C-PPA use an alignment score consisting of the sequence profile-profile alignment and the predicted secondary structure from PSIPRED (Jones, 1999). In C-PPA, both chains are simultaneously aligned to the template structures present in our complex libraries by the same modified DP as described in Figure S1. The third control was C-MUSTER which is an extension of the monomer threading algorithm of MUSTER (Wu and Zhang, 2008) which consists of multiple resource of structure information predicted from sequences. One of the major differences between C-MUSTER and COTH is that C-MUSTER does not include the interface prediction from BSpred and the monomer-chain recombinations.

5. Contact distance of inter-chain residues based on C-alpha atoms

Since threading alignments provide only C_α traces, we defined two residues in opposite chains to be in contact by an amino acid specific 20×20 C_α distance matrix. To obtain the matrix, we calculated the average C_α distance of all inter-chain residues that are in contact from our library of protein complex structures, where the two residues in opposite chains are defined as in contact if the distance between any two heavy atoms is less than 5 Å. Tables S1 and S2 are the average distance and the standard deviation of the C_α distances.

6. Structural superposition and combination of monomer templates.

Most proteins in the PDB library have been solved in monomer form (Berman et al., 2000). As a result, the number of available structures in monomer structure library (38,884) is much higher than that in dimer structure library (6,118). The structure space is far more complete in the tertiary structure library than in the complex structure library. We therefore expect to further improve the COTH threading alignments by combining both the tertiary and quaternary structure libraries. To achieve this, we first use the normal COTH threading procedure as described above to identify the template frames of complex structures. Meanwhile, we thread the monomer sequence (the individual chains of the dimer) to the tertiary structure library by the extended MUSTER algorithm. Finally, the MUSTER monomer templates, which usually have a better tertiary structure quality than that on the COTH-threading templates, are superimposed by TM-score rotation matrix on the COTH complex templates, based on the commonly aligned residues. It should be noted here that only the aligned regions in the MUSTER and COTH threading alignments was used for superposition (but not the original PDB structures of the templates structurally aligned together). If no commonly aligned residues are present between the MUSTER and COTH threading template (which actually never happened in any of our training or testing proteins), the program simply discards the superposition step and retains the original COTH threading template alignments.

The final complex model consists of the re-oriented structures of the MUSTER templates. If there are regions which are aligned by COTH threading but not aligned by MUSTER, the structural coordinates are not copied to the final models because these regions may have steric clashes with the MUSTER templates though the copy increases the coverage. The advantage of the superimposition step is that the resultant template retains the information regarding the relative orientation of the chains extracted by the COTH threading alignment while the tertiary structure qualities of the individual chains are significantly improved since the MUSTER templates have been generated from a much larger structure library. However, it is possible that in some rare cases, the combination step may result in unphysical inter-chain clashes (the inter-chain Cα-Cα distance <3.8Å). To rule out the clashes, the COTH program automatically discards the residues from the chain of higher alignment coverage which has a distance <3.8 Å to any residues in another chain.

7. Summary of threading and docking models using an I-RMSD cutoff 4 Å.

For the comparisons of the DOCKING based methods (ZDOCK-exp and ZDOCK-model) and COTH-based methods (COTH, COTH-exp and COTH-model), we have used an I-RMSD cutoff 5 Å to define successful cases, which is higher than the 4 Å cutoff used in CAPRI for defining an acceptable model. This is partly because threading approaches start from sequence only and the modeling accuracy is generally lower than the CAPRI docking models that start from experimental unbound structures.

Nevertheless, we also calculated the models using the 4 Å cutoff as in CAPRI. The number of the acceptable hits for the models by ZDOCK-exp, ZDOCK-model, COTH, COTH-exp and COTH-model are 23, 11, 19, 14 and 12, respectively. The hit distribution is similar as that using the cutoff 5 Å shown in Table 3. Again, ZDOCK-exp has the highest number of acceptable hits. But when using predicted monomer models, COTH-model slightly outperforms ZDOCK-model by 1 more hit. COTH-exp is worse than ZDOCK-exp in this resolution region, which highlights the necessity of fine-tuning the COTH models by maximizing the interface area. If we combine the two methods by taking 5 models from each approach, the number of acceptable hits increases to 27 when using experimental unbound monomer structures. Similarly, if we combine ZDOCK-model and COTH-model by taking the top 5 models from each approach, the number of acceptable hits increases to 16. These data demonstrate again the complement of the two modeling approaches.

References

Altshucl, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI_BLAST: a new generation of protein database search programs. Nucleic Acids Research *25*, 3389-3402.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Chen, H., and Zhou, H.X. (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res *33*, 3193-3199.

Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci U S A *81*, 140-144.

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology *292*, 195-202.

Mukherjee, S., and Zhang, Y. (2009). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res *37*, e83.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443-453.

Nissen, S., and Nemerson, E. Fast artificial neural network. Available at http://fann.sourceforge.net.

Ofran, Y., and Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. FEBS Lett *544*, 236-239.

Ofran, Y., and Rost, B. (2007). ISIS: interaction sites identified from sequence. Bioinformatics *23*, e13-16.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.

Wu, S., Skolnick, J., and Zhang, Y. (2007). Ab initio modelling of small proteins by iterative TASSER simulations. BMC Biol *5*, 17.

Wu, S., and Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins *72*, 547-556.

Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. Proteins *69*, 108-117.