Supporting Information

Gao and Skolnick 10.1073/pnas.1012820107

SI Methods

Alignment Algorithm. Previously, we described the heuristic algorithm implemented in iAlign for finding the optimal sequential alignment between two interfaces (1). Briefly, the algorithm has two major phases: In the first phase, several guessed solutions are generated through gapless alignments, secondary structure comparison, and fragment alignments. Starting from these guessed alignments, dynamic programming is iteratively applied during the second phase. The scoring matrix of dynamic programming is defined as

$$s_{ij} = \begin{cases} 1/(1+d_{ij}^2/d_0^2) & \text{for the iTM-score} \\ (f_{ij}+\delta)/(1+d_{ij}^2/d_0^2) & \text{for the IS-score.} \end{cases} \qquad \textbf{[S1]}$$

Here, d_{ij} is the distance between the ith residue of one structure and the jth residue of the other structure, and $f_{ij} \equiv (c_{ij}/a_i + c_{ij}/b_j)/2$, where a_i and b_j are the numbers of interfacial contacts of the ith and jth residues, respectively, and c_{ij} is the number of pairs of overlappable contacts. A contact between residues i and m is defined as overlappable to a contact between residues j and n if the distances between both i and j and between m and n are less than an empirical distance $d^* \equiv \min(1.5[\min(L_T,L_Q)]^{0.3}+3.5,8)$ Å. The small constant δ, set at 0.01, is introduced to prevent a score of zero. The best sequential alignment according to a specified scoring function (iTM/IS-score) is reported.

To allow a nonsequential alignment between two interfaces, iAlign continues to the third phase of iteratively searching for an optimal nonsequential alignment. The problem of finding an optimal nonsequential alignment (or match) is converted to the linear sum assignment problem (LSAP), which is also equivalent to the problem of finding a maximum weight matching in a weighted bipartite graph. To solve LSAP efficiently, we implemented the shortest augmenting path algorithm (2), which has a polynomial time complexity of $O(N^3)$, where $N = \max(L_T,L_Q)$. In our scenario, the goal of the LSAP procedure is to minimize the object function,

$$\min \sum_{k=1}^{N_a} t_{\phi(k)\varphi(k)}, \qquad \textbf{[S2]}$$

where $\phi(k)$ and $\varphi(k)$ define the functions mapping the kth aligned position to the original residue index in the query and the template structure, respectively, and the cost function $t_{ij} \equiv 2 - s_{ij}$, where s_{ij} is given by Eq. **S1**. The negative sign is introduced because we search for the minimum in LSAP instead of the maximum in dynamic programming, and the addition of 2 meets the requirement of nonnegative cost.

Application of the LSAP procedure with the scoring matrix defined in **S2** yields an alignment. The match between two residues is pruned if their distance is larger than d^*, which essentially introduces insertion/deletions. The iTM/IS-score is subsequently calculated for the alignment, and the superposition corresponding to the iTM/IS-score is used to obtain a new cost function, which in turn generates a new alignment. The procedure is repeated until the alignment converges or reaches an upper limit of 30 iterations.

Statistical Significance. The statistical significance of IS-scores from nonsequential alignments is estimated by comparing about 1.8 million random protein–protein complex pairs, selected such that the two complexes lack structurally related monomers; i.e., their monomeric TM-score <0.35 (1). The means (SD) of the maximum iTM/IS-scores for random interfaces are 0.278(0.071)/0.207(0.031), respectively.

We further estimate the statistical significance by modeling the distributions of IS-scores using Gumbel distributions, which are over maximum values and suitable to our cases because the IS-scores are the maxima of many alignments. Fig. S4 shows the observed and modeled distributions of scores at various lengths. Each distribution is modeled by the Gumbel distribution

$$P(z) = \exp[z - \exp(z)], \qquad \textbf{[S3]}$$

where $z = (s-\mu)/\sigma$. The variable s denotes the IS-score; μ and σ are the location and the scale parameters, respectively. These parameters are estimated through linear regression fits

$$\mu = a + b\ln(L_Q) + c\ln(L_T) \qquad \sigma = c + d\ln(L_Q) + f\ln(L_T). \qquad \textbf{[S4]}$$

The parameters a to f, given in Table S1, were obtained by linear fitting to the location and scale parameters, through maximum likelihood estimates with the EVD package in the R project (http://www.r-project.org/). Finally, the p-value is calculated using the formula

$$p\text{-value} = 1 - \exp[-\exp(-z)]. \qquad \textbf{[S5]}$$

Graph Analysis. The shortest path between a node and the rest of the nodes in a graph is computed with Dijkstra's algorithm (3). The search of the largest strongly connected component (LSCC) at the kth neighbor cutoff is converted to the problem of searching the largest clique in an undirected graph and subsequently solved with a branch-and-bound algorithm (4).

1. Gao M, Skolnick J (2010) iAlign: A method for the structural comparison of protein–protein interfaces. *Bioinformatics* 26(18):2259–2265.
2. Derigs U (1985) The shortest augumenting path method for solving assignment problems—Motivation and computational experience. *Algorithms and Software for Optimization*, ed Monma CL (Baltzer, Basel), Vol 4, pp 57–102.
3. Dijkstra EW (1959) A note on two problems in connexion with graphs. *Numer Math* 1:269–271.
4. Ostergard PRJ (2002) A fast algorithm for the maximum clique problem. *Discrete Appl Math* 120(1-3):197–207.

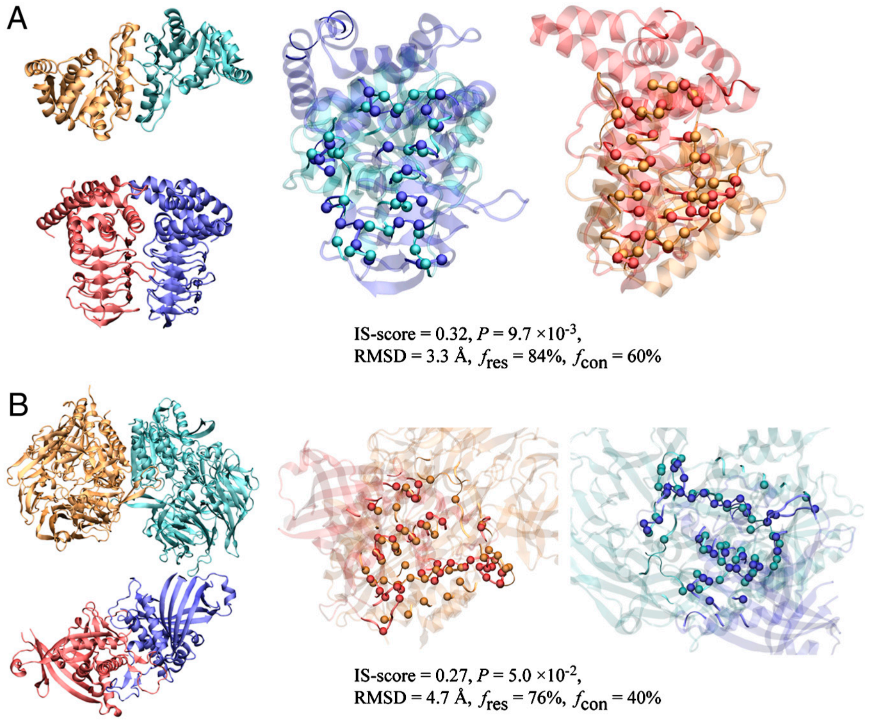
A
IS-score = 0.32, $P = 9.7 \times 10^{-3}$,
RMSD = 3.3 Å, f_{res} = 84%, f_{con} = 60%
B
IS-score = 0.27, $P = 5.0 \times 10^{-2}$,
RMSD = 4.7 Å, f_{res} = 76%, f_{con} = 40%

Fig. S1. Additional examples of similar protein–protein interface pairs with different secondary structure types identified by iAlign. Coordinates of structures were taken from the Protein Data Bank. The template (cyan/orange) and target (blue/red) proteins are (*A*) oxidoreductase (1jay_AB) and serine acetyltransferase (1t3d_AC), and (*B*) DPP-IV (2buc_AB) and RNA triphosphatase Cet1p (1d8i_BC). The *Right* snapshot shows the optimal interface alignment reported by iAlign. The interfacial alignments are illustrated separately for each side of the interfaces; and the C_α atoms of aligned residues are represented in spheres. For clarity, the interface/noninterface regions are shown in solid/transparent colors, respectively.

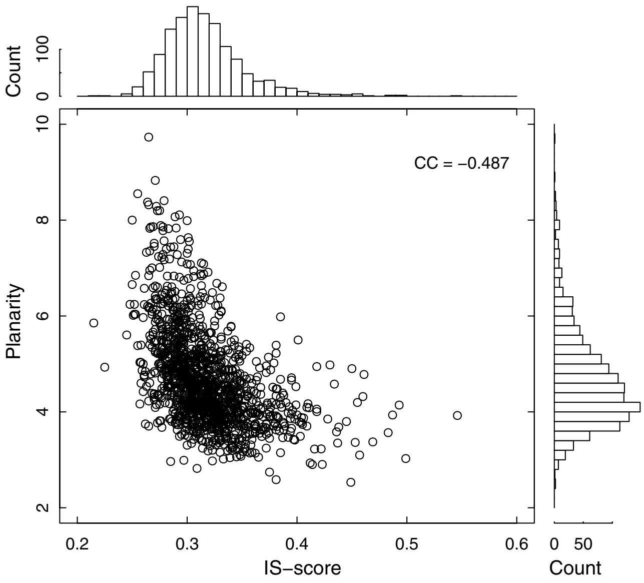

Fig. S2. Planarity of protein–protein interface versus interface similarity measured by IS-score. The IS-score is of the best match found for each representative interface from PDB150. Planarity is defined as the root-mean-square deviation of C_α atoms from the best-fit plane through the interface. The best-fit plane is calculated through principal component analysis with the program SURFNET (1).

[1] Laskowski RA (1995) SURFNET—A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graphics* 13(5):323–330.

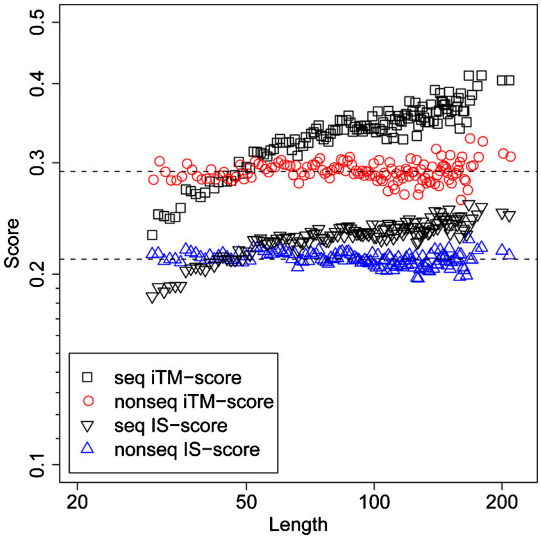

Fig. S3. Means of both iTM-scores and IS-scores among random interfaces of similar lengths. For a given length, we consider all random pairs whose lengths are between 95 and 105% of the length. All scores were calculated for the optimal nonsequential alignments with different scaling factor d_0. Seq iTM/IS-scores denote iTM/IS-scores calculated with the scaling factor d_0 originally used for sequential alignment, whereas nonseq iTM/IS-scores denote iTM/IS-scores calculated with adjusted d_0. The adjustment yields approximately length-independent iTM/IS-scores.

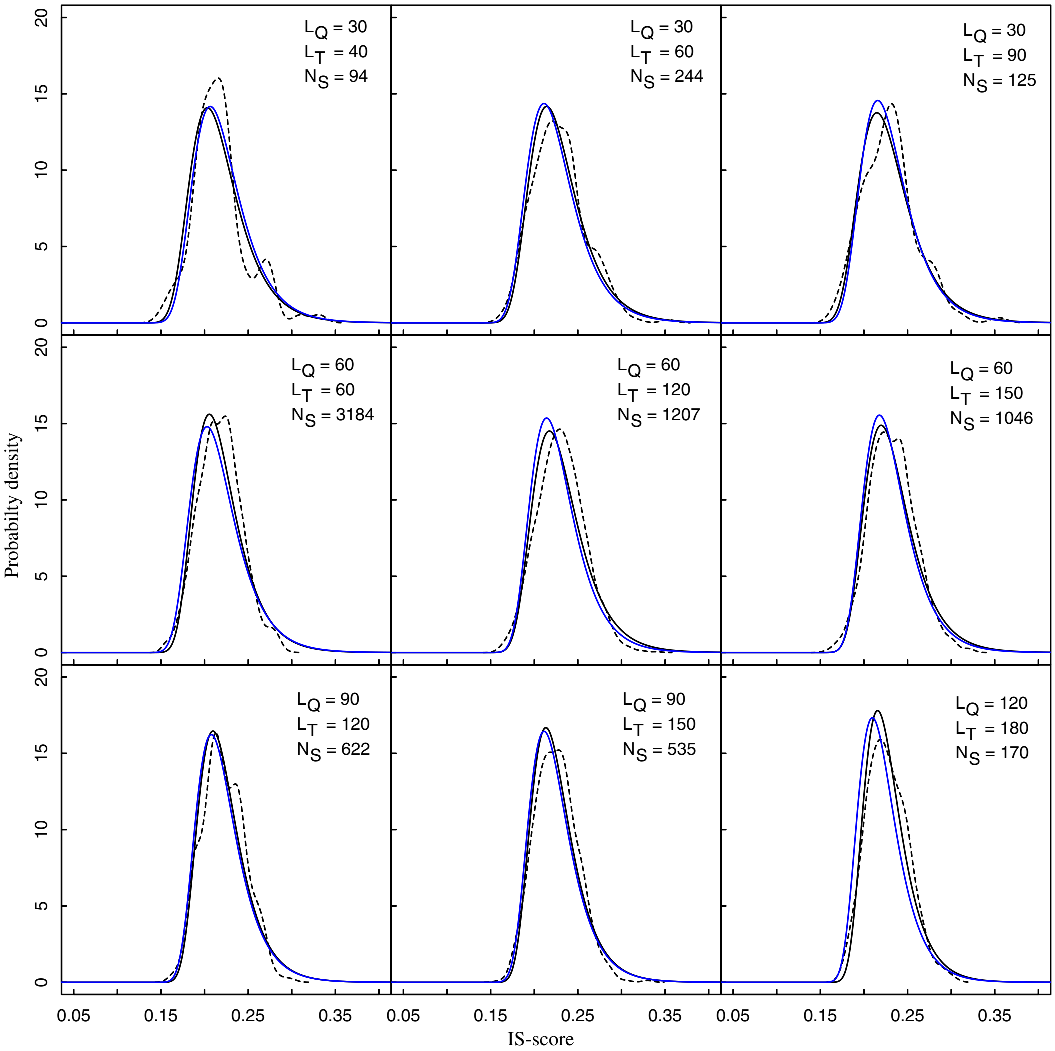

Fig. S4. Distributions of the IS-scores from nonsequential alignments among random interfaces of various lengths. Dashed black lines are the observed probability density, and the solid black lines are direct fits using the Gumbel distributions. Blue lines are the probability densities calculated for the IS-scores with statistical models described by Eqs. **S3** and **S4**. L_Q and L_T represent the length of query and length of template, respectively, and N_S is the number of samples from unrelated interface pairs.

Table S1. Parameters for calculating the location and scale parameters in Eq. S4

	IS-score	
Parameters	$L_Q < 55$	$L_Q \geq 55$
a	0.1776	0.2017
b	−0.0038	−0.0160
c	0.0113	0.0163
d	0.0397	0.0432
e	−0.0031	−0.0032
f	−0.0009	−0.0013