

Supplementary Data

For

iAlign: a method for the structural comparison of
protein-protein interfaces

MU GAO AND JEFFREY SKOLNICK

Supplementary Method

The two data sets used for generating docking models were derived from the M-TASSER template library (Chen and Skolnick, 2008).

(i) *Docking Set 568 (DS568)*. From the template library, we selected all complexes whose individual proteins are less than 500 amino acids in length and conducted rigid-body docking using the bound structures from the complexes. Docking was done with the program FT-Dock (Gabb, et al., 1997) and default parameters. The top 10,000 docking models, ranked by shape complementarity, were retained for analysis. In each docking model, we calculated iRMSD, the root mean square deviation of interfacial residue C_{α} atoms observed in the native structure with respect to their positions in the model. A docking model is considered near native if its iRMSD is less than 5 Å. We kept complexes with at least 20 near native docking models, totaling 568 complexes (DS568).

(ii) *Heterodimer 218 (Hete218)*. We selected a set of 218 heterodimers whose interacting partner proteins have low/no global structural similarity. In each case, the mTM-score reported by TM-align is less than 0.5 between the two pairs of monomeric protein structures.

Supplementary Results

Discriminating docking models

In an evaluation of protein-protein docking models, a common metric is the interfacial RMSD (iRMSD), which measures the deviation of native protein-protein interface residues observed in the experimental structure relative to their positions in the models (Lensink, et al., 2007). Obviously, one needs to know the prior sequence correspondence for an iRMSD calculation. By contrast, iAlign does not use any pre-specified information concerning the sequence correspondence of the target and template interfaces. Instead, the native interface is aligned to (putative) interfaces present in docking models.

In the first test, near-native docking models ($iRMSD < 5 \text{ \AA}$) generated for 568 protein-protein complexes (set DS568) were compared to corresponding native complex structures with iAlign. For each pair of complexes, iAlign was run twice with the iTM-score and IS-score as the similarity measure, respectively. The resulting p-values of the scores were used to rank docking models. The two rankings by the iTM/IS-score were then compared separately to the ranking by iRMSD. If iAlign performs properly, one expects to see a correlation between these rankings. As shown in Fig. S3, the rankings of near-native docking models by both the iTM-score and IS-score are indeed strongly correlated with the ranking by iRMSD. The correlation was evaluated with the Spearman Correlation Coefficient (SCC) (Maritz, 1981) calculated for each complex. The value of SCC is higher than 0.7 in about 70/87% of cases using the iTM/IS-score, respectively. The mean (standard deviation) of SCCs is 0.79 (0.16) for the iTM-score and 0.84 (0.13) for the IS-score (Fig. S3B). The latter is statistically significantly higher (paired and one-tailed t-test p-value $< 2.2 \times 10^{-16}$). The result suggests that the IS-score has better discrimination between similar interfaces than the iTM-score, mainly due to the fact that the IS-score incorporates both geometric distance and contact patterns.

To examine whether iAlign returns a reasonable significance assessment not only for near-native docking models, but also for models dissimilar from the native complexes, we further compare both near-native and non-near-native models to the native complex. Since it is quite likely to obtain a native-like interface far away from the native structure of a homodimer, e.g., taken from an oligomer with high order point symmetries (C3, D4, etc), we avoided unnecessarily confusion of the results by focusing on 218 heterodimers (set Hete218). For each complex, we consider top 200 models ranked by shape complementarity. A total of 43,600 docking models were assessed with iAlign. As expected, the vast majority (94%) of these models have an insignificant IS-score with p-value > 0.01 , while only a small fraction (1.6%) of docking models resemble the native structure at a high level of similarity with p-value $< 1 \times 10^{-6}$ (Fig. S4A).

Encouragingly, docking models within a 2.5 Å iRMSD from their native structure or with more than 30% of native contacts all have a significant p-value better than 0.01, mostly less than 1×10^{-6} (Fig. S4B and C). Conversely, almost all interfaces with a highly significant p-value ($< 1 \times 10^{-6}$) have iRMSD of less than 2.5 Å and more than 30% of native contacts.

However, there are a few exceptions, in which a high level of interface similarity is detected but the model seems far away from the native structure. The most notable case is a docking model of a Tomato Inhibitor TI-II/Subtilisin complex (Barrette-Ng, et al., 2003). The model has an iRMSD of 17 Å, but an IS-score of 0.39 with a significant p-value of 4×10^{-9} . The apparent discrepancy is resolved when one inspects the structures of the complex and the docking model. As shown in Fig. 1, the inhibitor TI-II can interact with two subtilisin molecules simultaneously at two distinct sites in a similar fashion. While the native complex is one of these two TI-II/Subtilisin complexes, the unusual docking model mimics the other one, resulting in a high iRMSD but nonetheless a valid solution. Therefore, iAlign correctly identifies non-trivial relationships between the two interfaces.

About 4.5% of docking models exhibit an interface significantly similar to the native interfaces at a p-value between 0.01 and 1×10^{-6} . In the corresponding interface alignments, around 60% of interfacial residues of these models are identical (in amino acid type) to their correspondent residues at the native interfaces (Fig. S4D), though the model interfaces typically have larger than 5 Å iRMSD and preserve less than 30% of native contacts. These model interfaces usually overlap a part of the native interface.

Supplementary Figures

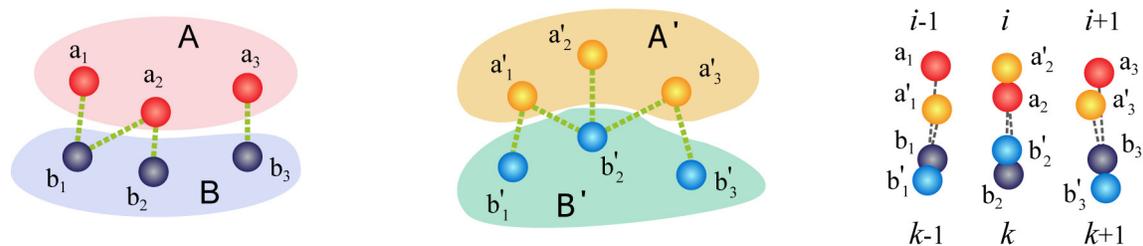


Fig. S1. Illustration of contact overlap. Interfacial residues of complex A/B (left; red/blue) and of complex A'/B' (middle; orange/cyan) are represented by spheres and labeled as x_j ($x = a, a', b, b'$, and $n = 1, 2, 3$). Interfacial contacts are shown in green dashed lines. In the structural alignment (right), a residue-residue correspondence is established between nearest neighbors (e.g., a_1 and a'_1), leading to three pairs of overlapped interfacial contacts (black dashed lines). The contact overlap factor at the i th alignment position is calculated as $f_i = (1/2 + 1/1)/2 = 0.75$; and the factor at the k th position is $f_k = (1/1 + 1/3)/2 = 0.67$.

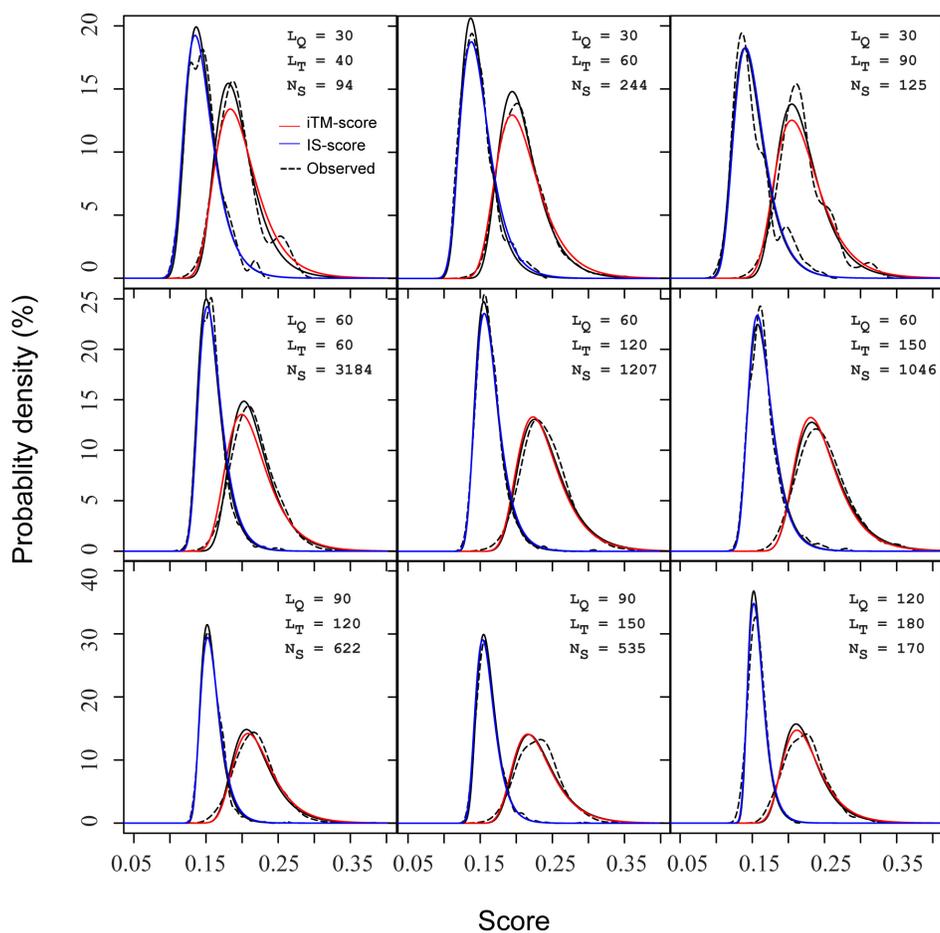


Fig. S2. Distributions of the iTM/IS-scores among random interfaces of various lengths. Dashed black lines are the observed probability density, and the solid black lines are direct fits using the Gumbel distributions. Blue and red lines are the probability densities calculated for the IS- and iTM-scores with statistical models described by Eq. 5 and 6. L_Q and L_T represents the length of query, template, and N_S is the number of samples from unrelated interface pairs.

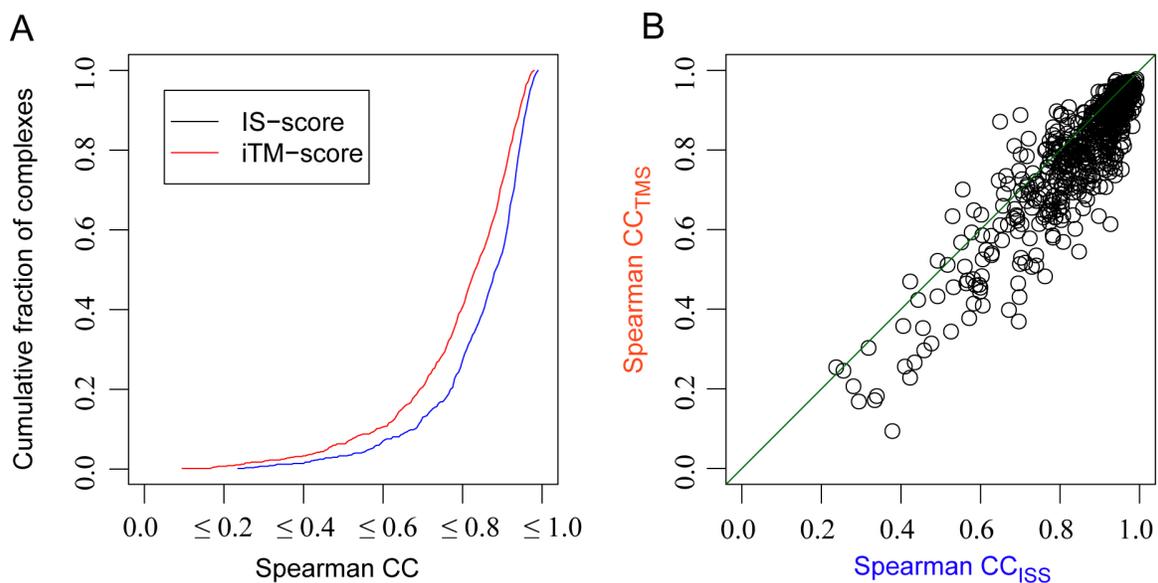


Fig. S3. Protein-protein docking models compared to native complexes by iAlign. **(A)** Cumulative distributions of 568 complexes versus Spearman Correlation Coefficients. For each complex, the SCC is between the ranking of its docking models by the p-value of iTM/IS-score and the ranking of same models by iRMSD. **(B)** SCC according to iTM-score p-value versus SCC according to IS-score p-value.

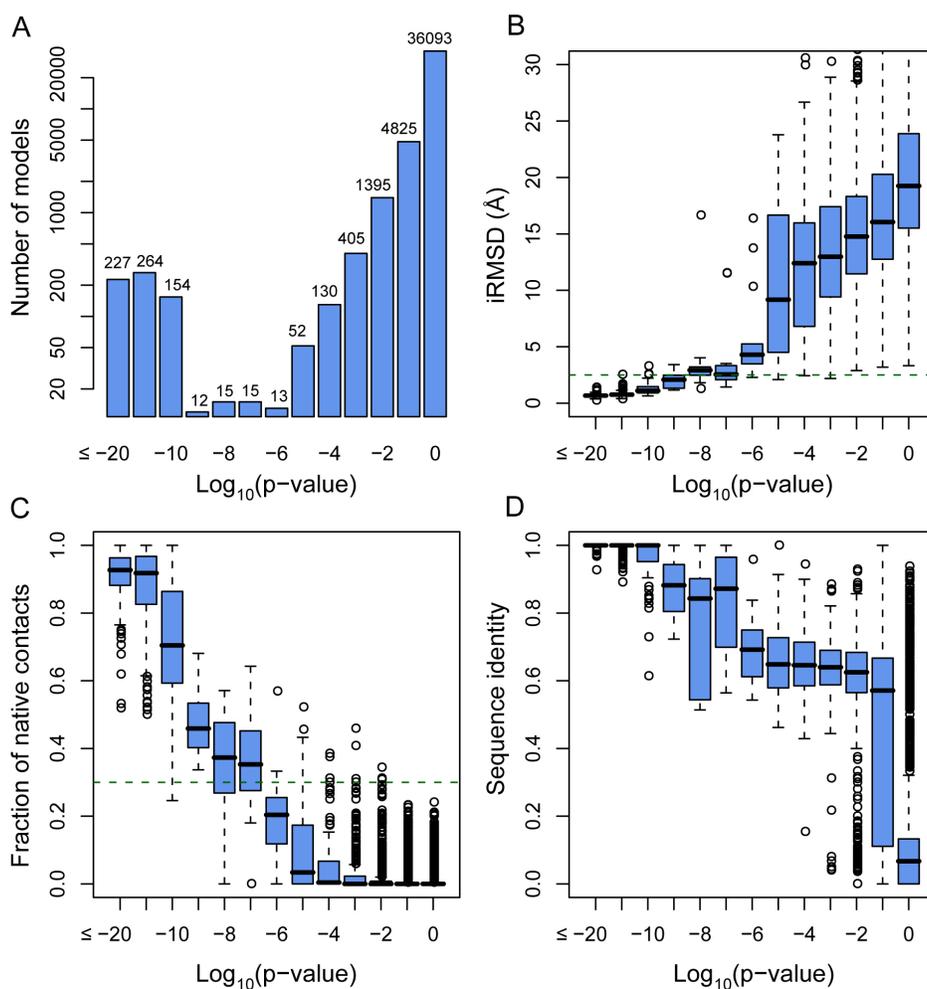


Fig. S4. Interface similarity between docking models and the native structure of 218 heterodimers. (A) Overall distribution of docking models according to the p-values of the IS-score reported by iAlign. Box plots of docking models according to (B) interfacial RMSD, (C) fraction of native contacts preserved in models, and (D) sequence identity over the aligned regions.

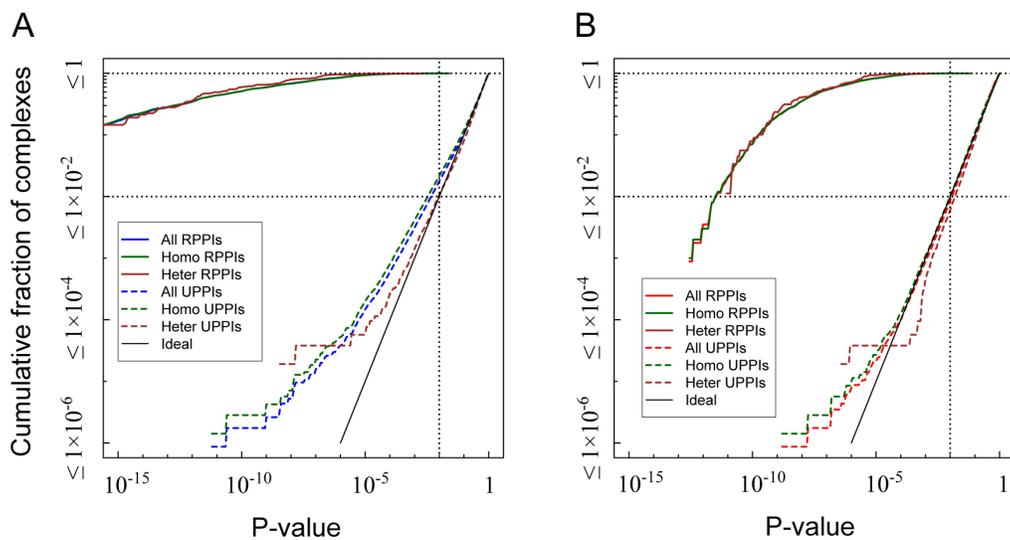


Fig. S5. Cumulative fraction of complexes versus the estimated p-values for (A) iTM-score and (B) IS-score by iAlign. Results are from pairwise comparison of 1,517 dimers, including 327 heterodimers and 1,190 homodimers. Related/un-related protein-protein interfaces are denoted as RPPIs and UPPIs, respectively.

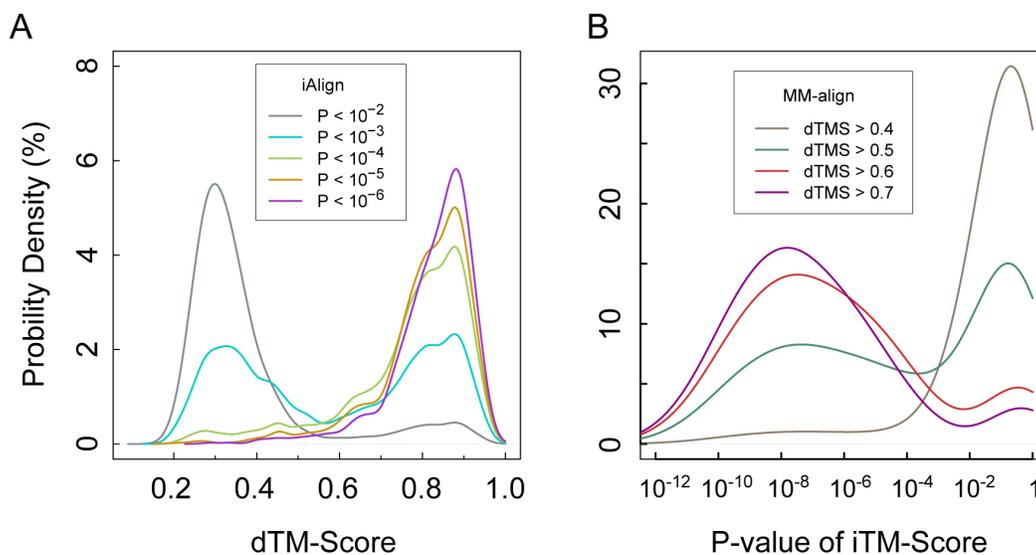


Fig. S6. Interface alignment by iAlign versus global alignment by MM-align. The statistics includes both related and unrelated pairs from Dimer1517. **(A)** Distributions of the dTM-score of protein-protein complex pairs, whose interfaces are significantly similar at different levels according to the p-values of iTM-scores. dTM-score is the best global TM-score from aligning two protein-protein complexes with the program MM-align. **(B)** Distributions of the p-values from the interface alignment of protein-protein complex pairs, whose global structures are significantly similar at different levels according to dTMS. Probability densities were estimated with a Gaussian kernel and a bandwidth window of 0.02 in (A) and of 1.0 in (B), in which p-values were first converted to $\log_{10}(\text{p-value})$.

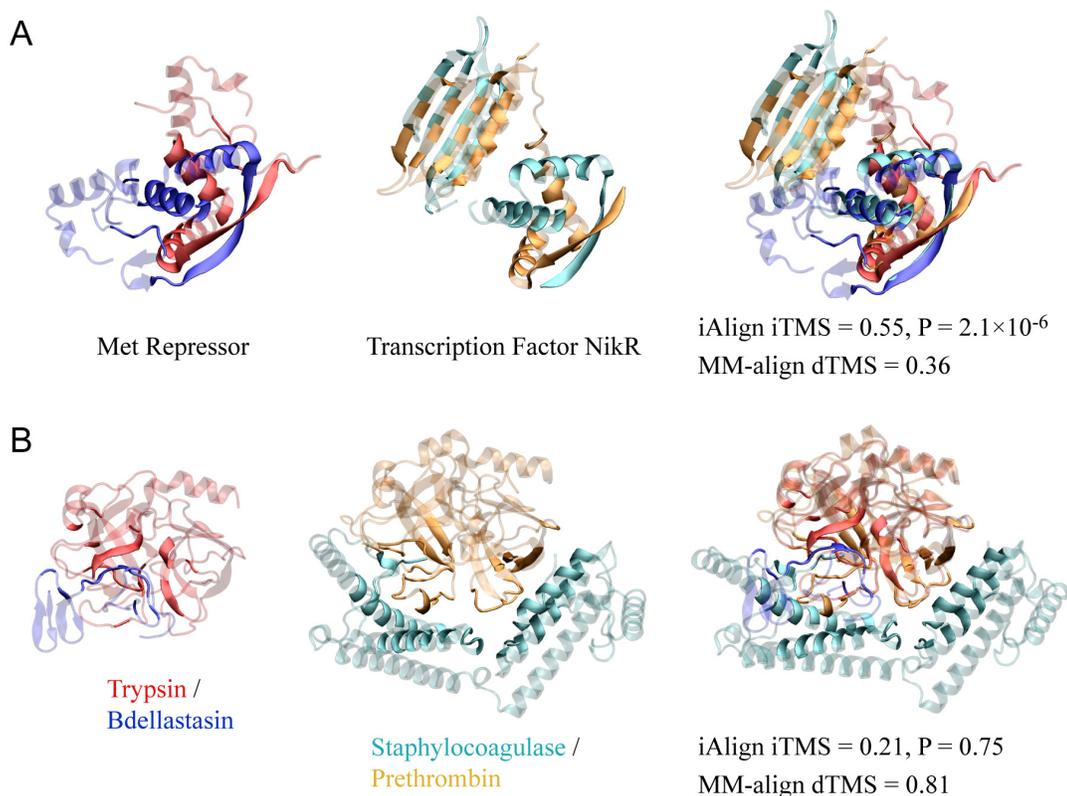


Fig. S7. Two examples illustrating advantages of interface alignment over global alignment. **(A)** Two related transcription factors, MetR (left; PDB 1cmb) and NikR (middle; PDB 1q5v) are both homodimers, and they share similar interfaces (right) as detected by iAlign at high significance. But the global complex alignment by MM-align reports an insignificant dTM-score of 0.36, due to the fact that non-interface regions adopt completely different folds. **(B)** Structural comparison between a Trypsin/Bdellastasin complex (left; PDB 1eja) and a Prethrombin/Staphylocoagulase complex (middle; PDB 1nu9). Bdellastasin is an inhibitor to Trypsin, whereas Staphylocoagulase activates Prethrombin to Thrombin, a Trypsin-like protease from the same SCOP superfamily. Although the two Trypsin-like structures aligned very well, yielding a high dTM-score of 0.81 by MM-align (right), the two protein-protein interfaces are obviously dissimilar. The interface alignment by iAlign correctly characterizes this dissimilarity with an insignificant score. In all molecular images, interface regions are shown in solid colors, and non-interface regions are dimmed.

Supplementary Table

Tab. S1. Parameters for calculating the location and scale parameters in Eq. 6.

Parameters	iTM-score		IS-score	
	$L_Q < 45$	$L_Q \geq 45$	$L_Q < 45$	$L_Q \geq 45$
<i>a</i>	-0.0690	0.2092	-0.0242	0.1635
<i>b</i>	0.0460	-0.0368	0.0391	-0.0085
<i>c</i>	0.0260	0.0344	0.0071	0.0057
<i>d</i>	0.0355	0.0416	0.0342	0.0440
<i>e</i>	-0.0050	-0.0042	-0.0058	-0.0076
<i>f</i>	0.0024	0.0007	0.0013	0.0006

Supplementary References

Barrette-Ng, I.H., et al. (2003) Structural basis of inhibition revealed by a 1 : 2 complex of the two-headed tomato inhibitor-II and subtilisin Carlsberg, *J. Biol. Chem.*, **278**, 24062-24071.

Chen, H.L. and Skolnick, J. (2008) M-TASSER: An algorithm for protein quaternary structure prediction, *Biophys. J.*, **94**, 918-928.

Gabb, H.A., et al. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information, *J. Mol. Biol.*, **272**, 106-120.

Humphrey, W., et al. (1996) VMD: visual molecular dynamics, *J. Mol. Graphics*, **14**, 33-38.

Lensink, M.F., et al. (2007) Docking and scoring protein complexes: CAPRI 3rd edition, *Proteins: Struct. Funct. Bioinform.*, **69**, 704-718.

Maritz, J.S. (1981) *Distribution-Free Statistical Methods*. Chapman & Hall.