Published in final edited form as:
Proteins. 2010 February 1; 78(2): 400–419. doi:10.1002/prot.22550.

PIE - Efficient filters and coarse grained potentials for unbound protein-protein docking

Ravikant Dintyala[1] and **Ron Elber**[2]

[1] Department of Computer Science, Cornell University, 4130 Upson Hall, Ithaca, NY 14853

[2] Department of Chemistry and Biochemistry, Institute of Computational Engineering and Sciences, University of Texas at Austin, 1 University Station, ICES, C0200, The University of Texas at Austin, Austin TX 78712

Abstract

Identifying correct binding modes in a large set of models is an important step in protein-protein docking. We identified protein docking filter based on overlap area that significantly reduces the number of candidate structures that require detailed examination. We also developed potentials based on residue contacts and overlap areas using a comprehensive learning set of 640 two chain protein complexes with mathematical programming. Our potential showed substantially better recognition capacity compared to other publicly accessible protein docking potentials in discriminating between native and nonnative binding modes on a large test set of 84 complexes independent of our training set. We were able to rank a near native model on the top in 43 cases and within top 10 in 51 cases. We also report an atomic potential that ranks a near native model on the top in 46 cases and within top 10 in 58 cases. Our filter + potential is well suited for selecting a small set of models to be refined to atomic resolution.

Keywords

Protein Docking; Scoring function; Linear Programming; Ranking decoys

Introduction

Predicting shapes of protein complexes from structures of individual proteins is an important challenge in computational structural biology. In cells, most proteins are involved in multiple interactions with other proteins. These interactions may create functional assembly[1–3], help in signaling[4,5] and in subtle (allosteric) manipulation of protein activities. These interactions form a complex network that enables cell control of biochemical activities[6–8]. Typically the problem of predicting the structure of a complex from its constituents is separated into two groups: (i) predicting the complex from the structures of its monomers in the bound form (also called bound docking). (ii) modeling the complex when the shape of the individual proteins to-be-docked is known only approximately (unbound prediction – realistically we do not know the "right" conformation of the monomer without knowing the structure of the complex). The first type of prediction is considered relatively easy while the second case is more difficult. The level of added difficulty depends on the similarity of the starting individual structures to the final structures in the complex.

Most docking protocols [9–19] sample the conformations and relative orientations of ligand (the smaller protein) with respect to the receptor (the larger protein), select a small (usually 100 to

1000) set of promising models, refine them to atomic resolution, rescore them and present the sorted list of models of the complex. The goal of a scoring function is to quickly and effectively filter the initial conformations sampled and retain as many near native models as possible. This problem is of high significance since the number of models considered in order to sample near native structures is inversely related to the quality of starting structures.

Residue pair potentials are a natural choice of coarse graining intra and inter interactions in proteins and have been useful in predicting folds of single chain proteins. In a possible approach one derives [20–24] a statistical potential based on the frequency of contact between a given pair of species in a set of non redundant complexes and a prior for the probability of observing such a contact.

Discriminatory learning [25–27] explicitly incorporates information from incorrect binding modes and bypasses issues with choice of reference state. Tobi and Bahar [28] demonstrate that it is possible to separate docked models with the correct binding mode from mis-docked models for a set of 63 complexes using a contact potential. Bordner and Gorin [29] train a random forest based classifier on residue contacts, residue interface propensities, conservation and a Vander Waals filter. They were able to sample and rank a near native model as the top model in 4 out of 17 cases (14 of which are enzyme inhibitor complexes). Our potential (called PIE - short for protein interaction energy) learnt on a large dataset shows significantly improved performance when tested on a similar collection of complexes. Atomic potentials for protein docking were derived under the statistical framework [30,31]. Oliver et al. [32] train a SVM based scoring function on a combination of atomic, residue based, shape (area, volume), shape complementarily, conservation based features on the complexes in benchmark 2 [33]. We trained atomic potentials and they had a small improvement in performance of ranking decoys of benchmark 2 [33] compared to residue potentials (that can be evaluated faster).

The paper is organized as follows – in the section Methods we describe our filter and detail the procedure followed to design the potential; in section Results and Discussion we report the efficacy of PIE for scoring and different variations of the potential. Our improvements come from an extensive training set, filtering based of overlap areas and inclusion of overlap areas in addition to contacts as features for discriminating between native and nonnative binding modes. We emphasize and illustrate in the results section that the overlap area excludes unlikely core penetration and is not correlated with surface matching.

Methods

Computing Overlap and Solvent Accessible Areas

Our procedure for efficiently computing overlap areas is based on an extension of the algorithm described by Scott et al[34]. The main idea of the procedure is as follows (see also Figure 1):

- The area of a sphere S1 lost in overlap with sphere S2 is a function of the point (P) where the line joining the centers intersects surface of S1 and the angle (θ) the circle of intersection subtends at the center of S1. The points lost to each (P, θ) overlap can be pre-computed for a sampling of points on unit sphere.
- Scott et al. [34] use a sampling of 256 points on unit sphere (obtained by maximizing the spread of the points), the state of each point (buried vs exposed) is stored in a bitmap. The residual surface area of a sphere is computed by finding the points that are exposed after accounting for all the overlaps of the sphere with its neighbors and scaling the number based on the square of the radius.

We pre-compute the burial bitmaps for each atom on the receptor and ligand separately. For each transformation of the ligand, we compute interface atoms (intersecting atoms of other

molecule, done efficiently using neighbor lists). For each interface atom, we compute the number of points lost to overlaps with atoms of other molecule and from them we obtain the overlap area.

An approximate solvent accessible surface area can be computed by increasing the radii of each atom by the probe radius and computing the portions of surfaces of each atom that are not in overlap with any other atom. A simple extension of the method can be used to compute volumes and solvent excluded volumes (using concentric shells). We use the atomic radii (Lennard Jones contact distance σ) from the OPLS force field [35].

Filter

We observe that the overlap area for native like structures is lower than the overlap area in mis-docked structures (as illustrated in Figure 2). So we use the overlap area as a filter for discarding mis-docked structures. We tried different thresholds for filtering and we observe the best filtering when we select the top 2500 structures in terms of changes in overlap area to be re-ranked by PIE. The choice of the threshold used in the filter depends on the quality of input proteins. We obtain a similar improvement in performance when we use the radii from the AMBER force field (param98) instead of the OPLS parameters or when we use the overlap volume (the total volume that is in the intersection of the atoms in receptor with atoms in ligand) instead of the overlap area.

Model

We use residue contacts (two residues are defined to be in contact if the distance between their side chain centers of mass is less than 6.8 Å) and overlap area computed as outlined earlier. For a transformation τ ($\tau = (t, u)$ where t and u are the translation and rotation of the ligand),

$$E(\tau)=w\Delta_{OA}(\tau)+\sum_{i,j=1,i\leq j}^{20} w_{ij}n_{ij}(\tau) \quad (1)$$

where $\Delta_{OA}(\tau)$ is the total overlap area upon complex formation and $n_{ij}(\tau)$ is the number of contacts between residues of type i and j.

Model learning - Training set

We prepared an extensive set of bound and unbound 2-chain complexes. An initial set of 640 complexes was prepared according to the following procedure:

- Select 2-chain proteins in the PDB with each chain having at-least 40 residues.
- Find similar complexes - two complexes are similar if (both) the corresponding chains are homologous – greater than 35% identity in the BLAST[36] alignment or E-value is below 10^{-6}.
- Cluster complexes – starting from a list of complexes, select a complex if it is not similar to complexes already selected.
- Check for biological significance - a complex is biological if the chains were dissimilar or there is confirmation in the reference articles at PDB that the observed dimer is not an artifact of crystallization.

To prepare a set of complexes to be used for unbound-unbound docking and bound-unbound docking, we looked for unbound conformations or close homologues (sequence identity around 85%) for each protein in the set of 640 complexes. For 55 complexes we had unbound

conformations for both chains (unbound-unbound cases) and for 123 complexes we had unbound conformation for only one chain (bound-unbound cases). Whenever unbound conformations were available we used them for docking. The size distribution of proteins in our training set is illustrated in Figure 3. The list of unbound and bound complexes is provided in Table 1 and Table 2.

Model learning – Learning procedure

The goal is to obtain parameters w that minimize $\sum_{ij,ik} \eta_{ij,ik}$ such that

$$w^T(P_{ij}^+ - P_{ik}^-) > 1 - \eta_{ij,ik} \text{ and } \eta_{ij,ik} \geq 0 \text{ for all } ij \text{ and } ik \quad (2)$$

Where P_{ij}^+ is the vector of interface properties (residue-residue contacts and change in overlap area) of the j^{th} correctly docked structure for complex i and P_{ik}^- is the feature vector of the k mis-docked structure for complex i. For each complex we sampled orientations of the ligand using Patchdock [23]. A typical number of sampled orientations for one complex was 16,000 which was used to generate 160,000 inequalities (we considered upto 10 near-native structures to be discriminated from the other incorrect structures).

Docking of two homologous proteins for the unbound-unbound cases and bound-unbound cases is based on structural modeling of the chains of the actual complex. We modeled the structures of the individual chains based on the sequences and folds of the unbound protein chains (we used the Needleman-Wunsch[37] alignment between the sequence of the chains of the bound proteins and the sequence of the chains of the unbound protein and computed the structures using Modeller [38]). We generated translations and rotations using both bound proteins and modeled proteins but computed contact maps and changes in overlap areas by applying the translations and rotations on the modeled proteins. The addition of bound transformations in the unbound cases was useful in generating more positive examples. There were three procedures to generate positive examples: In the first procedure we used Patchdock to dock the modeled chains and successful docking experiments were classified as positive (see definition in the next paragraph for a positive classification). In the second generation of positive examples, we computed transformations for the docking of the native bound chains but applied them on modeled chains. In the third case we overlapped the structures of the modeled chains with the bound chains in the native complex. In each of the three procedures we collected all the positive examples that scored better than the threshold discussed below. For bound cases, the second and the third procedures were used to generate positive examples.

We separated the computed complexes into hits/positives (backbone rmsd is within 4 Å for all heavy backbone atoms of the two chains when comparing the bound complex and the modeled complex) and misses (backbone rmsd above 7 Å and less than 5% of native contacts at the interface). We dropped all other structures that correspond to a “gray” classification. The “gray” matches are likely to introduce noise to the learning process. Features from the native complex were added as a hit. We restricted the number of hits to 10 (based on rmsd and making sure that a new hit is different by at least 4 contacts from the complexes selected already). For each pair of hit and miss, we created an inequality if the set of contacts in a negative differed from the set of contacts in a positive by at-least 10 (this condition helps in simplifying cases with large rms differences and still small changes in contact maps that otherwise complicate the learning process).

We used Pf3 [39] to solve the resultant linear program with 54,126,279 constraints[3]. An exact solution (satisfying all the inequalities, (Eq. (2))) could not be found within the functional form (Eq. (1)) that we use for the potential. Hence, some of the inequalities remained unsolved. However the model potential with optimized parameters was able to solve 97% of the inequalities. The set of parameters determined (PIE640) were stable – the parameters obtained with 90% of constraints (PIE540, the potential learnt from the 540 complexes not overlapping benchmark2) are similar to the parameters obtained with all the constraints (correlation coefficient 0.96, a figure comparing the potentials is in supplementary material).

Results and Discussion

We use protein-protein docking benchmark 2.0 [33] for our evaluation as it is widely used for comparing scoring functions. We removed cases similar to the 84 cases in the benchmark from our training set. The definition used for similarity is outlined below:

- Two proteins P and P′ are similar to each other if
  - The alignment of P to P′ returned by BLAST has at-least 30% positives (positions not aligned to gaps).
  - The TM-score [40] for P to P′ is above 0.4 or there is 30% identity among interface residues (a residue is in the interface if it is in contact with a residue of the other protein) in the TM-alignment between P and P′.
- A complex C formed by proteins P, Q is similar to another complex C′ formed by proteins P′ and Q′ if either P is similar to P′ and Q is similar to Q′ or P is similar to Q ′ and Q is similar to P′.

We use a stronger criterion to remove cases overlapping benchmark 2.0 from our training set than the criterion we use to remove redundancies in our training set. It is important to use the extra measures to avoid learning examples that are going to be used for testing later. 100 complexes were removed from our initial set of 640 complexes because of overlaps with complexes in benchmark2. The set of complexes used for training the potential and the resultant potential (PIE540) are presented in the supplementary material.

We test our filtering and scoring protocol on three procedures for generating docked structures – ZDOCK [41], Patchdock [42] and our implementation of geometric hashing (along the lines of [11]).

ZDOCK Sampling

ZDOCK [41] is a procedure for modeling protein complexes that docks rigid molecules based on a combination of shape complementarity (PSC), desolvation and electrostatics that is efficiently computed using the Fast Fourier Transform (FFT) algorithm. ZRANK [43] is a scoring function that uses a more elaborate linear combination of atomic Vander Waals, electrostatics and desolvation for scoring models.

We use the models generated by ZDOCK 3.0 with ZRANK with a grid size of 1.2 Å and 6° Euler angle rotations (as provided at http://zlab.bu.edu/zdock/decoys.shtml). We also use the irmsd (irmsd is the minimum rmsd between the positions of the interface residues in the native structure and the model, described in [44]) values provided with the set for verification. A hit (defined as a model with irmsd below 4 Å) was present in the set for 79 out of the 84 cases. A hit was present in the lowest 2500 cases in terms of overlap area in 76 cases (in 3 cases all the hits were lost because of the filter).

[3]The solution took about 3 hours on 150 cores (Intel Xeon 2.33 GHz. 1GB memory per core, 8 cores per node).

The results of scoring the decoys by our filter and scoring function (FP) and our scoring function alone (P) in comparison to scoring by ZRANK (denoted by ZR) are presented in Table 3. We present the best rank of a hit (Bestrank), the number of hits in the top 10,20,100 and 2048 (TopN) along with the number of hits sampled and the number of hits filtered by the overlap area filter. The column labeled Random gives the rank for which the probability that a random scoring function will do better is at-least 0.5. This score signifies the ease of ranking a set of models. For example, consider the scenario where for each pair of proteins, half the structures for the protein-protein complex are hits, then a procedure that ranks at random will return a hit at rank 1 with probability 0.5 and a ranking procedure that is not able to rank a hit on the top on more than half of the cases is worse than a random ranking procedure.

Our top ranked model is a hit in 43 cases compared to 10 cases by ZD3.0ZR and we have a hit in top 10 models in 51 cases compared to 21 by ZD3.0ZR.). Discarding cases with clashes using our filter leads to a similar improvement in performance of ZRANK (top ranked model is a hit in 42 cases and a hit is ranked within top 10 in 53 cases). In a follow up paper Pierce and Weng [45] further refined the top ranked structures and re-ranked them, they evaluated their procedure using 2.5Å irmsd as a cutoff criterion for a positive. On a set of 27 complexes, they were able to rank a positive as top model in 10 cases compared to 7 cases by ZD3.0ZR. When restricted to the 27 complexes in the refinement and rescoring study [45], we score a near native model (irmsd < 2.5Å) as rank 1 in 23 cases.

Patchdock Sampling

Patchdock [42] is an algorithm for molecular docking that uses image segmentation (for defining patches) and object recognition (based on geometric hashing). We used Patchdock with default options (no specification of interface or complex type) for docking the input proteins in the benchmark and generating candidate structures for tests of our classification capacity.

Comparison of our potential to the default scoring function in Patchdock (PD) is presented in Table 4. On the 71 cases where Patchdock sampled a hit, our top ranked model is a hit in 27 cases and we have a hit in top 10 models in 51 cases. Discarding cases with large overlap area using our filter improves the performance of Patchdock (the best rank of a hit with the filter is better than the best rank of a hit without the filter in 48 cases while it is worse in 21 cases; in 2 cases all hits are lost because of the filter). While the improvement is significant it is not as strong as in the case of the structures generated by ZD3.0ZR.

The results for two different sampling procedures (ZDOCK and Patchdock) indicate that our filtering and scoring procedure is not sensitive to the procedure used for generating docked models.

Variation of Sampling

We implemented a sampling protocol based on geometric hashing (along the lines of Fischer et al [11] details in the supplementary material) and with our potential for scoring models we are able to sample as well as rank a near native structure (within 4 Å irmsd) as the top model in 19 cases on the benchmark. The breakdown of results by hardness is presented in Table 5. SP is the result for our sampling coupled with our coarse grained potential (PIECED540); FP+ZR is the result of our filtering and scoring of decoys generated by ZDOCK3.0 with ZRANK and ZR is the result of ZDOCK3.0 with ZRANK. While our filtering procedure enriches scoring structures sampled by ZDOCK+ZRANK, the enrichment of the filtering to our sampling (which is similar to Fisher et al[11]) is not that striking.

Comparison with other residue based scoring functions

The result of comparing the overlap filter and our scoring function to the potential of Lu et al. [23] (LLS), Tobi et al. [28] (TB) and Glaser et al[22] (GSVB) in ranking the decoys in the ZDOCK set (note that benchmark2 overlaps with the training sets of LLS and TB) is outlined in Table 6 and Table 7. In short we score a model within 4 Å irmsd as the top model in 19 cases compared to 6 by TB and 4 by LLS. However when the filter is used to discard structures with large overlap area, the result improves to 43 cases compared to 36 by TB, 29 by LLS and 7 by GSVB.

Examination of results for different classes of complexes

The performance of our scoring function for each class of complexes is illustrated in the Figure 4–6 (a few representatives are presented here, the remaining are in the supplementary material); a short summary is presented in Table 8. Each row in the figure (for Figures 4–6) presents the result of scoring structures that are filtered based on overlap area using different residue based potentials. The leftmost plot in a row is the plot of our energy (z score of the negative of our energy, since in our framework higher energies are better) against irmsd. The next two plots represent the result of scoring by the potential of Lu et al. [23] and Tobi et al. [28] respectively. In most cases we obtained marked funnels compared to others. By funnel we mean that the average energy (negative of the score) is monotonically increasing (on the average) as a function of the irmsd. PIE540 was learnt on the pooled set of complexes, we did not over-represent antibody-antigen interactions in our training set (there are 6 antibody heavy chain – light chain complexes and 9 antibody – antigen complexes in the set of 540 complexes used for training). Additional class specific bias in learning might improve our results.

Examination of the potential

Figure 7 illustrates the residue contact potential and Table 9 summarizes the potential learnt on our entire training set (the subset used to train potential for testing and the potential learnt are detailed in the supplementary material). Figure 8 summarizes the contribution of residue contacts to the overall energy grouped by residue type and overlap area. The value plotted is the energy of the residue type in the native structure averaged over all the 640 complexes in the training set. Contacts between hydrophobic residues contribute significantly towards the recognition of native binding modes. Though we derive our potential from a very large set of complexes, we observe non convergence of some entries to physically meaningful values (for example the score for W-I contact has a different sign compared to the score for W-L contact). We have significant contact statistics for every contact type in the native structures of complexes in our set. We also derive a statistical potential for residue contacts from the set of 640 complexes ($e_{ij} = -\log((N_{ij}/N)/((N_i/N) *(N_j/N)))$, where N_{ij} is the number of contacts between residues of type i and j, $N = \Sigma_j N_{ij}$ and $N = \Sigma_i = N_i$) and the statistical potential has the same inconsistencies. When we add additional constraints requiring more physically consistent behavior to our linear programming formulation (for instance, $w_{WI} > w_{WL}/2$ and $w_{WL} > w_{WI}/2$) the inconsistency is corrected and the resultant potential has a correlation coefficient of 0.99 with the original potential (PIE640).

Variations of features

We report the results of our attempts to add additional features as well as regroup features in Table 11. We investigate the additional features by redefining the potential and re-optimizing the linear parameters exactly like what we have done to the residue contacts and overlap area based potential.

3 residues are said to form a 3body contact if each pair of residues picked from the three are in contact with each other. To enhance the statistics we use 5 residue types – polar – G,P,N,Q,S,T,C; basic – K,R; acidic – D,E; aliphatic – A,V,L,I,M and aromatic – F,Y,W,H for

computing 3body contacts). The results are presented in row 4 and 5 of Table 11. There is a very small improvement compared to the use of pair potential.

We considered evolutionary profiles by dividing residues into 3 types based on their conservation relative to solvent-exposed residues (we use the column entropy in the profile computed from Blast with E-value cutoff of 0.001 on sequences in Swissprot[46], we sort solvent-exposed residues based on entropy and divide them into 3 types – among the top 33% of conserved residues on the surface, between 33% and 66% and below 66%) and counting contacts between residues of different types. The results are presented in row 6 of Table 11. We only observe a minor improvement with the inclusion of evolutionary profile.

For atomic potential we use the grouping into 18 atom types that was used in DARS [47] and an extension to 20 atom types (CYS C^{β} and MET S^{δ} in two new groups). The results are presented in rows 7,8,9 and 10 of Table 11. When we include overlap areas (grouped by atom type), the performances of atomic potentials are a little better than the performance of the residue potential. The potential with 20 atom types is presented in Table 12 and the potential with 18 atom types is provided in the supplementary material.

We do not observe improvement in discriminatory ability of the potential when we use the changes in solvent accessible surface area computed by DSSP [48] as a feature or if we use changes in solvent excluded surface area. Figure 9 shows the relationship between the overlap area computed by us against the changes in surface areas computed by DSSP. The overlap areas are close to zero for bound complexes while the solvent accessible areas are very different from zero.

Heterodimers vs Combined

There were 186 homodimers and 354 heterodimers in the training set (no overlaps with the test set). We trained docking potential on the 354 heterodimers separately and ranked the decoys as earlier. Our filter with the potential trained on heterodimers ranked a hit as top model in 41 cases and identified a hit in top 10 in 54. The potential learnt on the set of all 417 heterodimers in our set of non redundant complexes is presented in Table 10. The heterodimer potential is very similar to the potential learnt on the entire set (as illustrated in Figure 10, correlation coefficient for the residue contacts is 0.96).

There were studies in the past on the extent of similarity (/diversity) of interactions in homodimers and heterodimers [23,49–51]. The number of homodimers is relatively small in our set (only 223 complexes) making it more difficult to obtain the necessary statistics and rigorously investigate the similarities between homodimers and heterodimers. Nevertheless, we went ahead and trained the overlap area plus residue contact potential by solving the corresponding inequalities for homodimers. The correlation between the homodimer potential and the combined potential for the contribution of residue contacts is 0.69. The potential is provided in the supplementary material. We trained docking potential on the 186 homodimers separately and ranked the decoys as earlier. Our filter with the potential trained on homodimers ranked a hit as top model in 29 cases and identified a hit in top 10 in 45 cases.

Transient vs Combined

85 complexes in our set of 640 non-redundant complexes overlap with the set of transient complexes identified in the study of Mintseris et. al [52] and Ansari et. al [53]. There were 102 additional complexes between 2-chains in the study of Mintseris et. al and Ansari et. al where each chain was at-least 35 residues long. We identified 105 additional transient complexes in our set by examining the primary citations for each structure. We trained a potential on the 329 transient complexes (190 in our set, 102 from earlier studies and 37 complexes from

benchmark2 that do not overlap the set selected). The transient potential is highly correlated to the combined potential (correlation coefficient 0.98, comparison in Figure 11).

Comparison to other scoring functions

It is of interest to compare our potential to existing scoring functions that integrate a wide range of information. Oliver et al [32] use a wide range of biochemical and geometric features (residue contacts, atomic contacts, residue interface propensities by complex type, gap volume, buried surface area, conservation) and learn a scoring function within the SVM framework based on the cases in benchmark2. On cases in benchmark2 they were able to sample and rank a near native solution (within 5 Å C^{α} rmsd) on the top in 16 cases in contrast to 19 by our sampling and scoring procedure and 43 by our reranking of structures sampled by ZDOCK.

Summary

Our goal is to design a single potential with a physically meaningful functional form for separating native binding modes from the non native ones. Such a scoring function can be used in statistical mechanics study of dynamics and function in addition to the prediction of the structures of protein complexes. We seek a residue based potential towards this end since it is less sensitive to local conformational changes and as a result has greater radius of convergence; even though we observed higher recognition with atom based potential on our test set. We are after energy features that can be efficiently computed – at this point they are residue contacts and overlap areas (we did not obtain additional discriminatory ability from the inclusion of 3body contacts, changes in surface areas and evolutionary profiles). We use a discriminatory framework for designing potential since it allows for learning from incorrect binding modes in contrast to approaches that are based on positive learning only (such as statistical potentials). The resultant potential does a good job of recognizing native binding modes on a large and diverse set of test complexes. It is likely that our gain over existing potentials is due to the filtering based on shape complementarity. Though symmetric homodimers enrich contacts between like residues, we observe a small improvement in performance with potentials learnt on the combined set of homodimers and heterodimers compared to potential learnt exclusively on heterodimers. Our filter and potential are likely to be useful for selecting a small set of models to be refined to atomic resolution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the anonymous reviewers for their suggestions, and specifically for the comments on overlap area. This research was supported by NIH grant GM067823.

References

1. Söllner T, Bennett MK, Whiteheart SW, Scheller RH, Rothman JE. A protein assembly−disassembly pathway in vitro that may correspond to sequential steps of synaptic vesicle docking, activation, and fusion. Cell 1993;75(3):409–418. [PubMed: 8221884]
2. Arents G, Burlingame RW, Wang BC, Love WE, Moudrianakis EN. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left−handed superhelix. Proceedings of the National Academy of Sciences of the United States of America 1991;88(22):10148–10152. [PubMed: 1946434]
3. Lesné S, Koh MT, Kotilinek L, Kayed R, Glabe CG, Yang A, Gallagher M, Ashe KH. A specific amyloid−[beta] protein assembly in the brain impairs memory. Nature 2006;440(7082):352–357. [PubMed: 16541076]

4. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal Protein−Signaling Networks Derived from Multiparameter Single−Cell Data. Science 2005;308(5721):523–529. [PubMed: 15845847]

5. Hamm HE. The Many Faces of G Protein Signaling. Journal of Biological Chemistry 1998;273(2): 669–672. [PubMed: 9422713]

6. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi−Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein−protein interactions in Saccharomyces cerevisiae. Nature 2000;403(6770):623–627. [PubMed: 10688190]

7. Gandhi TKB, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 2006;38(3):285–293. [PubMed: 16501559]

8. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. A Protein Interaction Map of Drosophila melanogaster. Science 2003;302(5651):1727–1736. [PubMed: 14605208]

9. Wodak SJ, Janin J. Computer analysis of protein−protein interaction. Journal of Molecular Biology 1978;124(2):323–342. [PubMed: 712840]

10. Katchalski−Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proceedings of the National Academy of Sciences of the United States of America 1992;89(6):2195–2199. [PubMed: 1549581]

11. Fischer D, Lin SL, Wolfson HJ, Nussinov R. A Geometry−based Suite of Molecular Docking Processes. Journal of Molecular Biology 1995;248(2):459–477. [PubMed: 7739053]

12. Ruben Abagyan MT, Kuznetsov Dmitry. ICM − A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. Journal of Computational Chemistry 1994;15(5):488–506.

13. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. Journal of Molecular Biology 1997;272(1):106–120. [PubMed: 9299341]

14. Ritchie DW, Kemp GJL. Protein docking using spherical polar Fourier correlations. Proteins: Structure, Function, and Genetics 2000;39(2):178–194.

15. Chen R, Weng Z. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins: Structure, Function, and Genetics 2002;47(3):281–294.

16. Gray JJ, Moughon S, Wang C, Schueler−Furman O, Kuhlman B, Rohl CA, Baker D. Protein−Protein Docking with Simultaneous Optimization of Rigid−body Displacement and Side−chain Conformations. Journal of Molecular Biology 2003;331(1):281–299. [PubMed: 12875852]

17. Nuno P, Palma LK, Wampler John E, Moura José JG. BiGGER: A new (soft) docking algorithm for predicting protein interactions. Proteins: Structure, Function, and Genetics 2000;39(4):372–384.

18. Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF. Protein docking using continuum electrostatics and geometric fit. Protein Eng 2001;14(2):105–113. [PubMed: 11297668]

19. Camacho CJ, Vajda S. Protein docking along smooth association pathways. Proceedings of the National Academy of Sciences of the United States of America 2001;98(19):10636–10641. [PubMed: 11517309]

20. Moont G, Gabb HA, Sternberg MJE. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins: Structure, Function, and Genetics 1999;35(3):364–373.

21. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. Empirical solvent−mediated potentials hold for both intra−molecular and inter−molecular inter−residue interactions. Protein Sci 1998;7(12): 2578–2586. [PubMed: 9865952]

22. Glaser F, Steinberg DM, Vakser IA, Ben−Tal N. Residue frequencies and pairing preferences at protein−protein interfaces. Proteins: Structure, Function, and Genetics 2001;43(2):89–102.
23. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein−protein interactions. Biophysical Journal 2003;84(3):1895–1901. [PubMed: 12609891]
24. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi−chemical approximation. Macromolecules 1985;18(3):534–552.
25. Maiorov VN, Grippen GM. Contact potential that recognizes the correct folding of globular proteins. Journal of Molecular Biology 1992;227(3):876–888. [PubMed: 1404392]
26. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins: Structure, Function, and Genetics 2000;38(2): 134–148.
27. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins: Structure, Function, and Genetics 2000;40(1):71–85.
28. Tobi D, Bahar I. Optimal design of protein docking potentials: Efficiency and limitations. Proteins: Structure, Function, and Bioinformatics 2006;62(4):970–981.
29. Bordner AJ, Gorin AA. Comprehensive inventory of protein complexes in the Protein Data Bank from consistent classification of interfaces. BMC Bioinformatics 2008;9:234. [PubMed: 18474114]
30. Robert CH, Janin J. A soft, mean−field potential derived from crystal contacts for predicting protein −protein interactions. Journal of Molecular Biology 1998;283(5):1037–1047. [PubMed: 9799642]
31. Song Liu CZ, Zhou Hongyi, Zhou Yaoqi. A physical reference state unifies the structure−derived potential of mean force for protein folding and binding. Proteins: Structure, Function, and Bioinformatics 2004;56(1):93–101.
32. Martin O, Schomburg D. Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. Proteins: Structure, Function, and Bioinformatics 2008;70(4): 1367–1378.
33. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. Protein−protein docking benchmark 2.0: An update. Proteins: Structure, Function, and Bioinformatics 2005;60(2):214–216.
34. LeGrand SM, Merz KM. Rapid approximation to molecular surface area via the use of Boolean logic and look−up tables. Journal of Computational Chemistry 1993;14(3):349–352.
35. Jorgensen WL, Tirado−Rives J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. Journal of the American Chemical Society 1988;110(6):1657–1666.
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI−BLAST: a new generation of protein database search programs. Nucl Acids Res 1997;25 (17):3389–3402. [PubMed: 9254694]
37. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 1970;48(3):443–453. [PubMed: 5420325]
38. Eswar, N.; Webb, B.; Marti−Renom, MA.; Madhusudhan, MS.; Eramian, D.; Shen, MY.; Pieper, U.; Sali, A. Comparative protein structure modeling using MODELLER. In: Coligan, John E., et al., editors. Current protocols in protein science/editorial board. Vol. Chapter 2. 2007.
39. Wagner M, Meller J, Elber R. Large−scale linear programming techniques for the design of protein folding potentials. Mathematical Programming 2004;101(2):301–318.
40. Zhang Y, Skolnick J. TM−align: a protein structure alignment algorithm based on the TM−score. Nucl Acids Res 2005;33(7):2302–2309. [PubMed: 15849316]
41. Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z. Integrating statistical pair potentials into protein complex prediction. Proteins: Structure, Function, and Bioinformatics 2007;69(3):511–520.
42. Duhovny, D.; Nussinov, R.; Wolfson, HJ. Efficient Unbound Docking of Rigid Molecules. Proceedings of the Second International Workshop on Algorithms in Bioinformatics; Springer −Verlag. 2002.
43. Pierce B, Weng Z. ZRANK: Reranking protein docking predictions with an optimized energy function. Proteins: Structure, Function, and Bioinformatics 2007;67(4):1078–1086.

44. Méndez R, Leplae R, Maria LD, Wodak SJ. Assessment of blind predictions of protein−protein interactions: Current status of docking methods. Proteins: Structure, Function, and Genetics 2003;52 (1):51–67.
45. Pierce B, Weng Z. A combination of rescoring and refinement significantly improves protein docking performance. Proteins: Structure, Function, and Bioinformatics 2008;72(1):270–279.
46. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS−PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucl Acids Res 2003;31(1):365–370. [PubMed: 12520024]
47. Chuang G-Y, Kozakov D, Brenke R, Comeau SR, Vajda S. DARS (Decoys As the Reference State) Potentials for Protein−Protein Docking. Biophys J 2008;95(9):4217–4227. [PubMed: 18676649]
48. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen −bonded and geometrical features. Biopolymers 1983;22(12):2577–2637. [PubMed: 6667333]
49. Pinak Chakrabarti JJ. Dissecting protein−protein recognition sites. Proteins: Structure, Function, and Genetics 2002;47(3):334–343.
50. Ranjit Prasad Bahadur PCFRJJ. Dissecting subunit interfaces in homodimeric proteins. Proteins: Structure, Function, and Genetics 2003;53(3):708–719.
51. Anastasya Anashkina EK, Esipova Natalia, Tumanyan Vladimir. Comprehensive statistical analysis of residues interaction specificity at protein−protein interfaces. Proteins: Structure, Function, and Bioinformatics 2007;67(4):1060–1077.
52. Mintseris J, Weng Z. Atomic contact vectors in protein−protein recognition. Proteins: Structure, Function, and Genetics 2003;53(3):629–639.
53. Ansari S, Helms V. Statistical analysis of predominantly transient protein−protein interfaces. Proteins: Structure, Function, and Bioinformatics 2005;61(2):344–355.

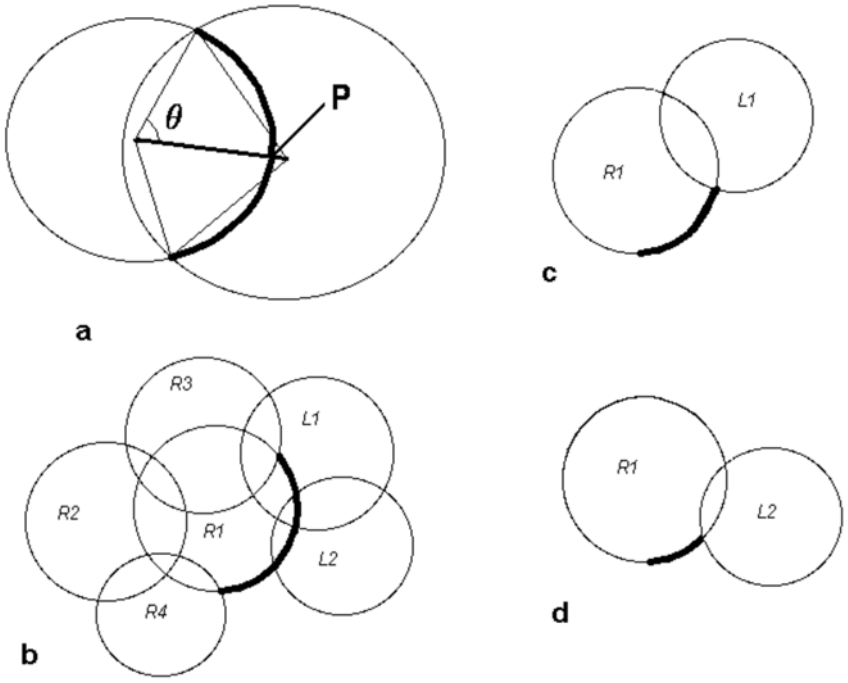

Figure 1.
Sketch of the algorithm for efficient computation of surface area - (P, θ) overlap (P is the point where the line joining the centers of the spheres intersects surface of the sphere, and θ is the angle subtended at the center) (a), exposed surface in receptor (b), exposure lost to the ligand (c,d)

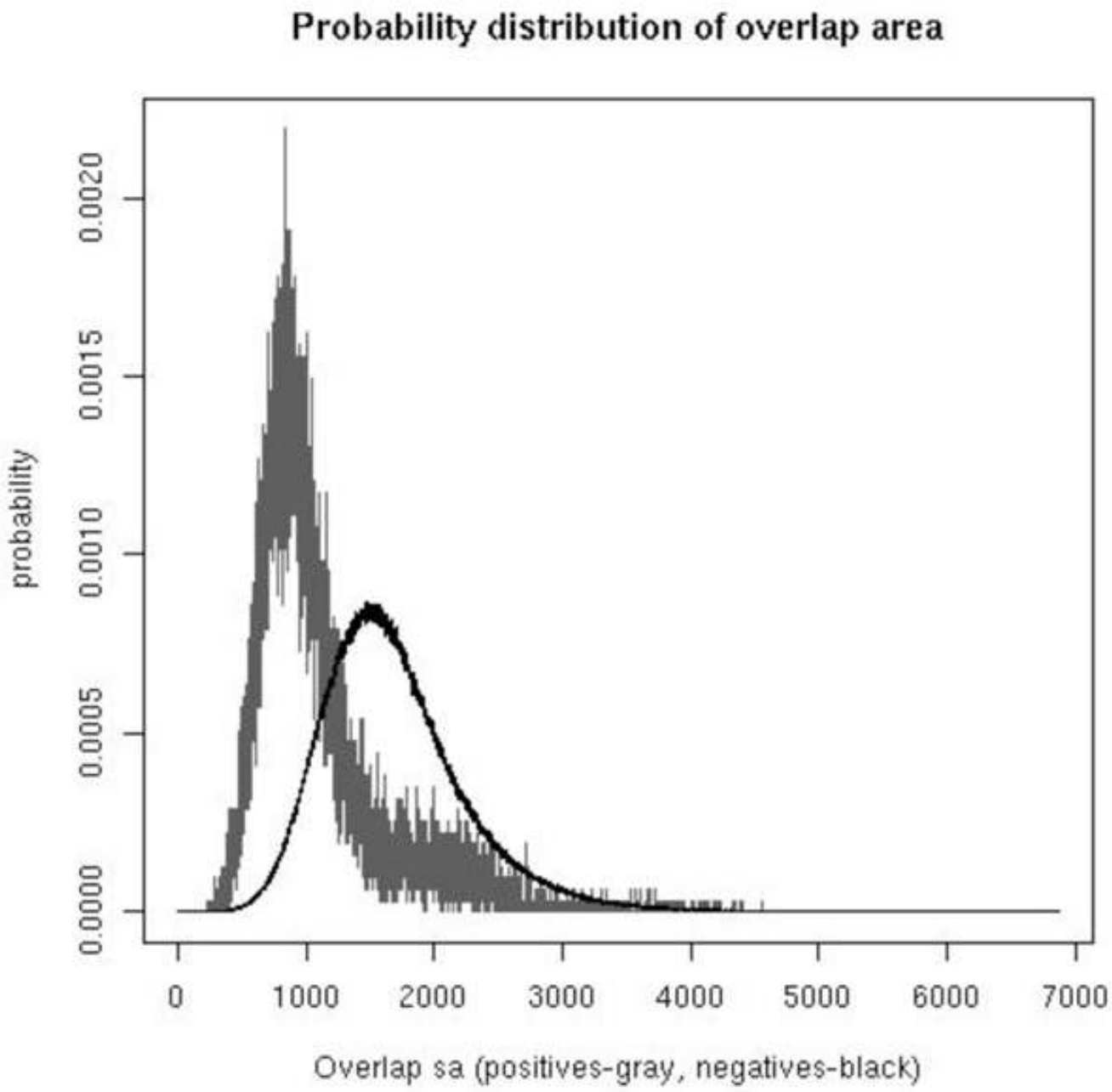

Figure 2.
Probability distribution of overlap area for native like structures (gray, curve closer to the y-axis) and misdocked structures (black) based on ZDOCK sampling on benchmark 2.0

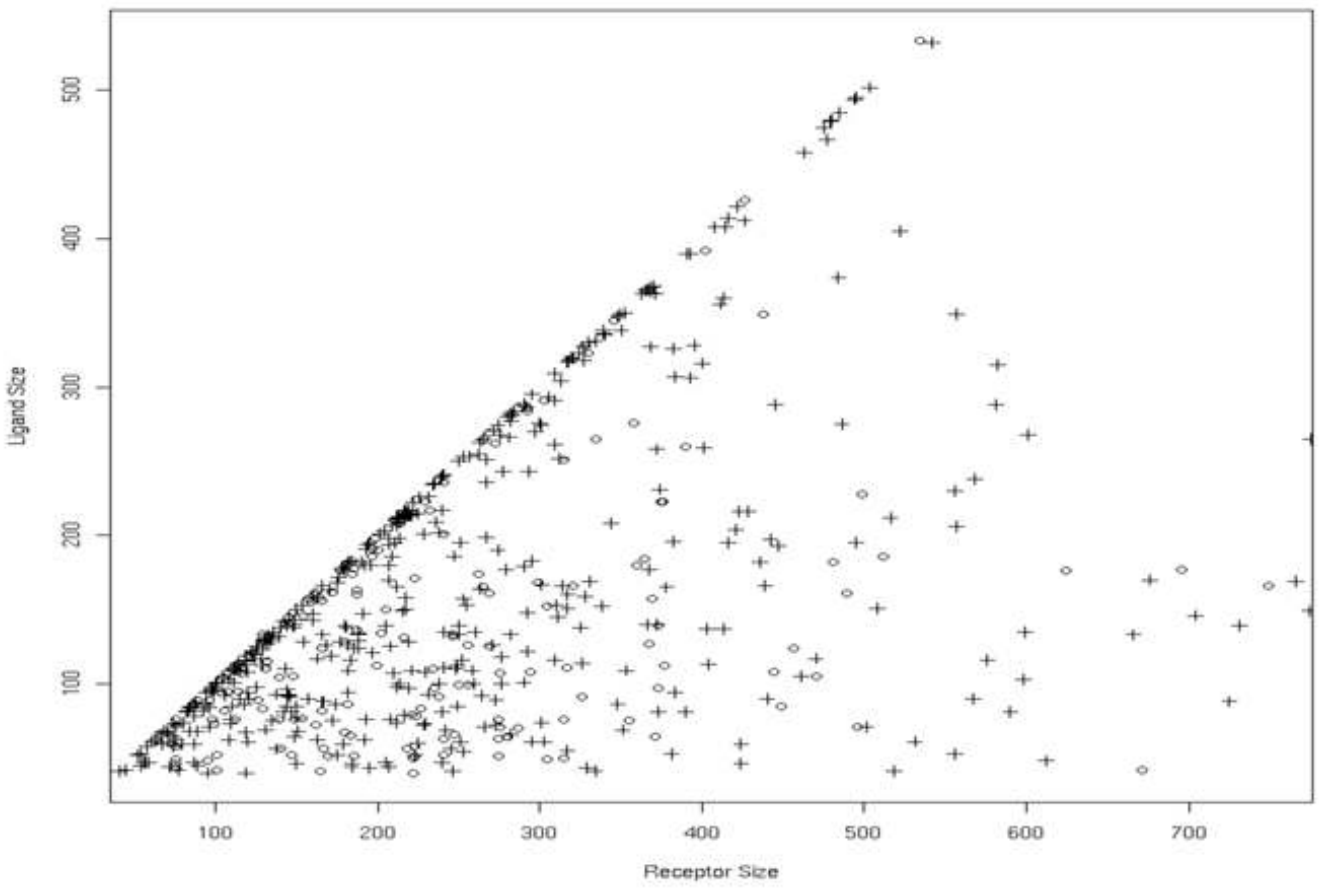

Figure 3.
Size distribution of complexes in the training set (points represented by pluses are bound cases and points represented by circles are unbound cases)

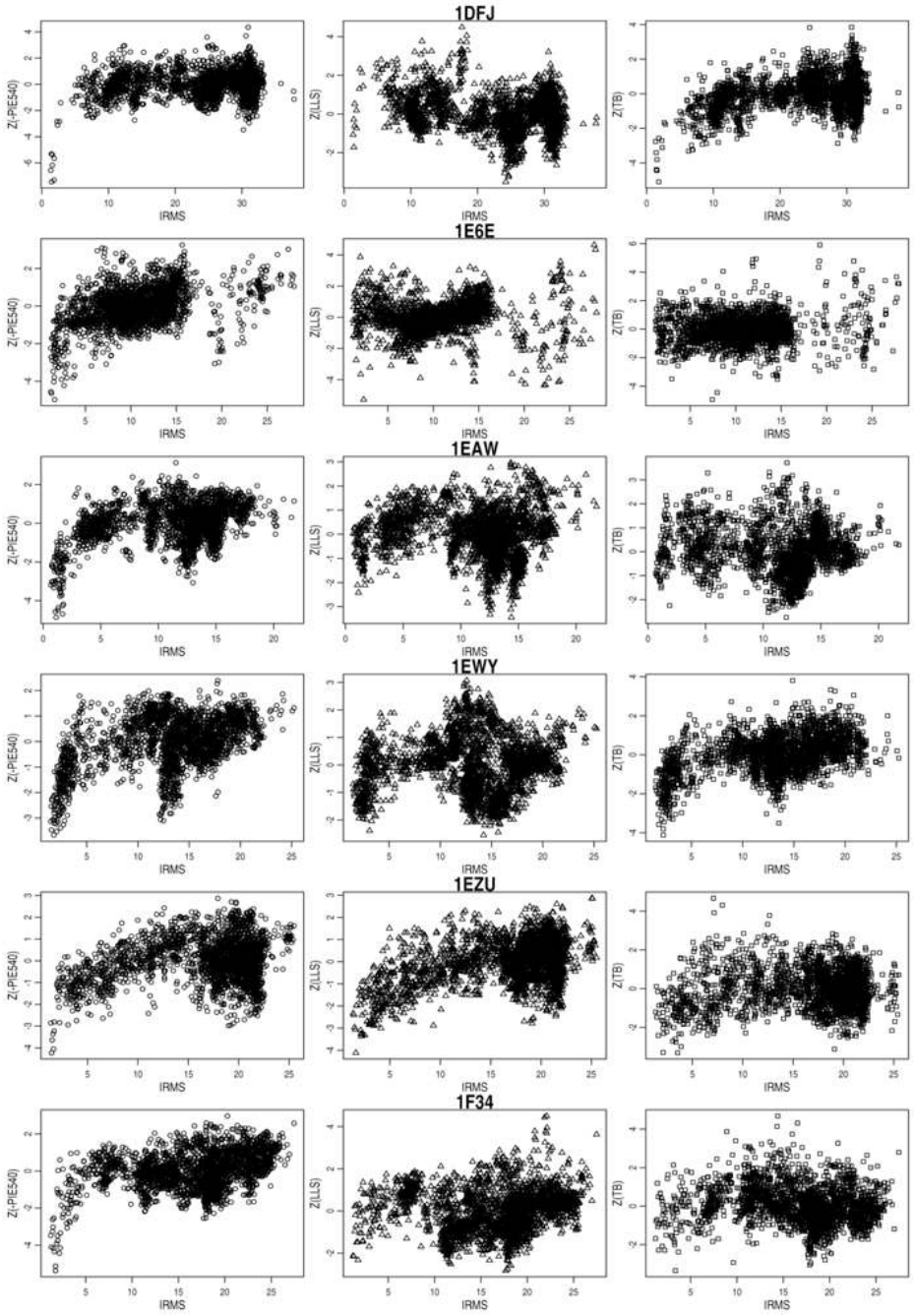

Figure 4.

Enzyme Inhibitor – In each row, we present the result of scoring models filtered based on overlap area, the left column is the result of present potential (PIE540), the result of Lu et al[23] (LLS) in the center and that of Tobi et al[28] (TB) is on the right

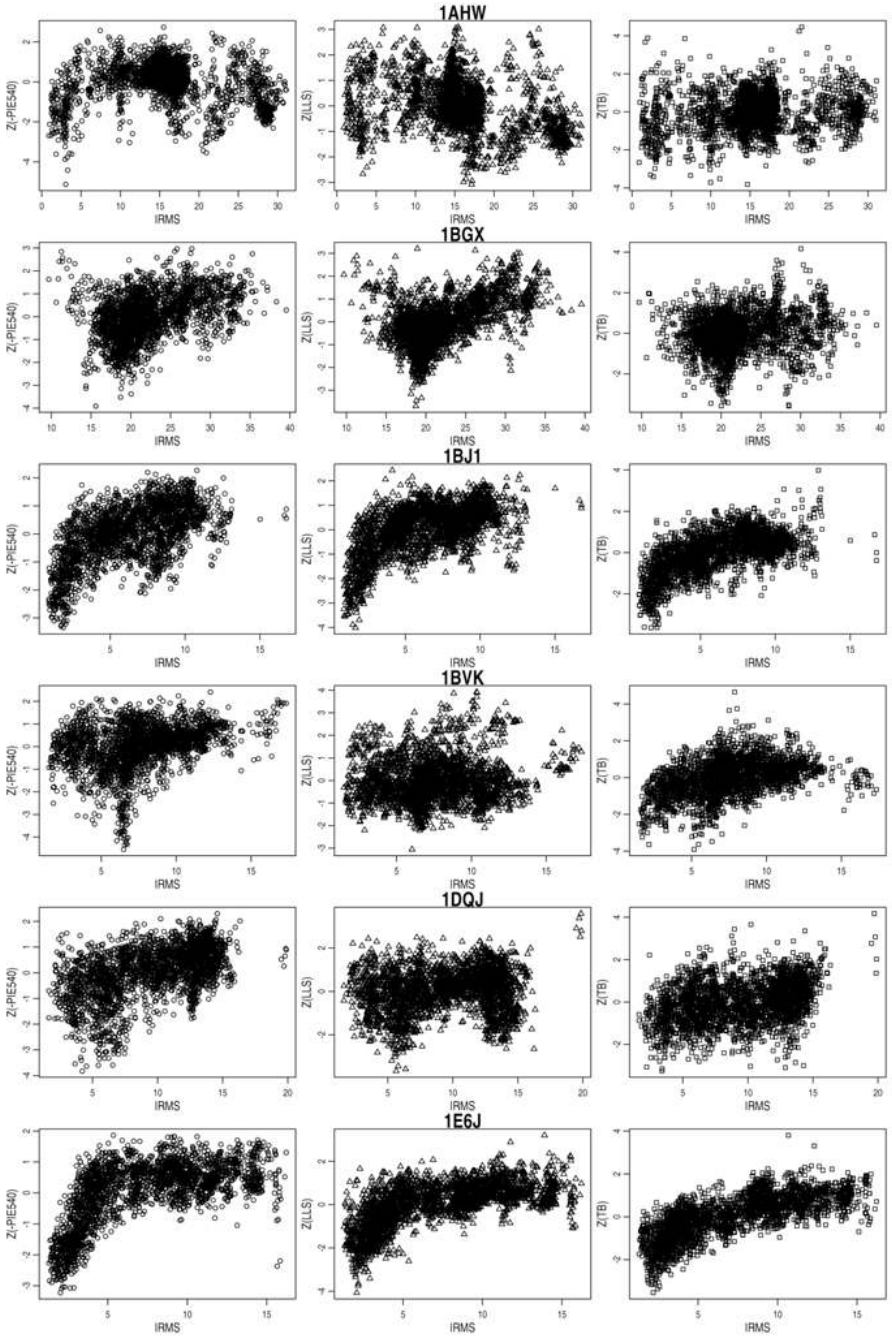

Figure 5.
The same as Figure 4 but this time for Antibody Antigen

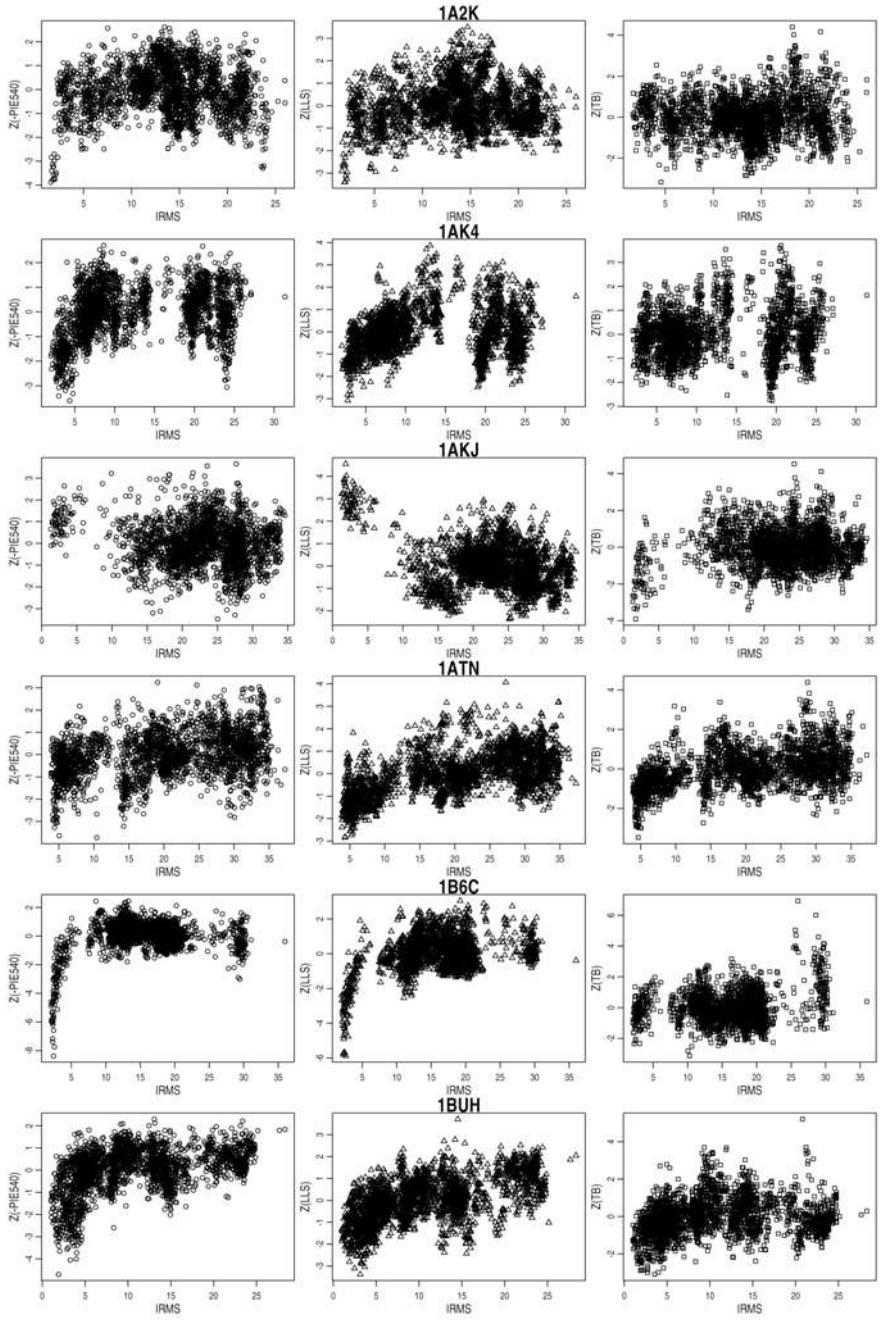

Figure 6.
The same as Figure 4 but this time for complexes in the Others category

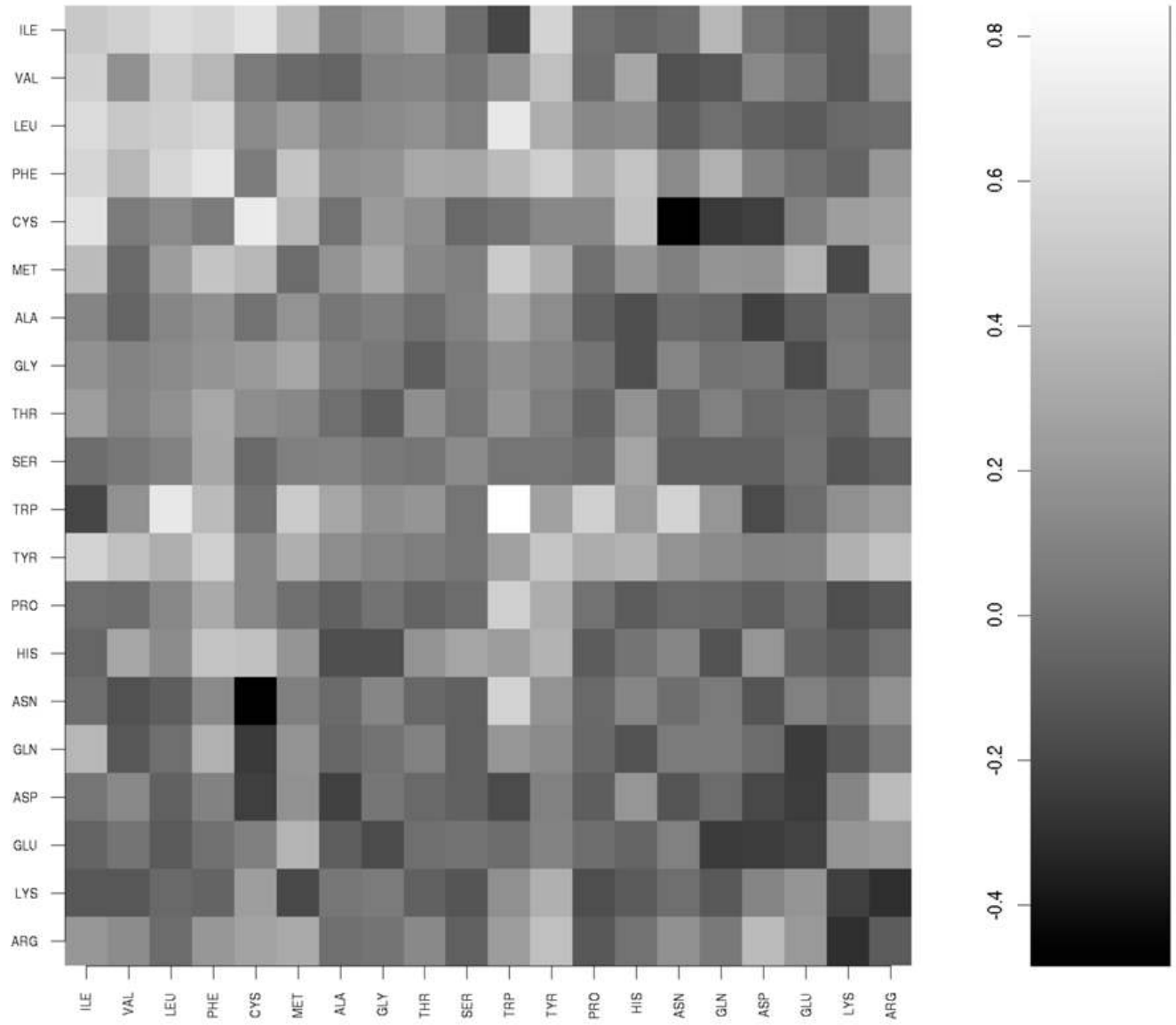

Figure 7.
Contribution of Residue contacts (darker shades indicate unfavorable contribution to binding)

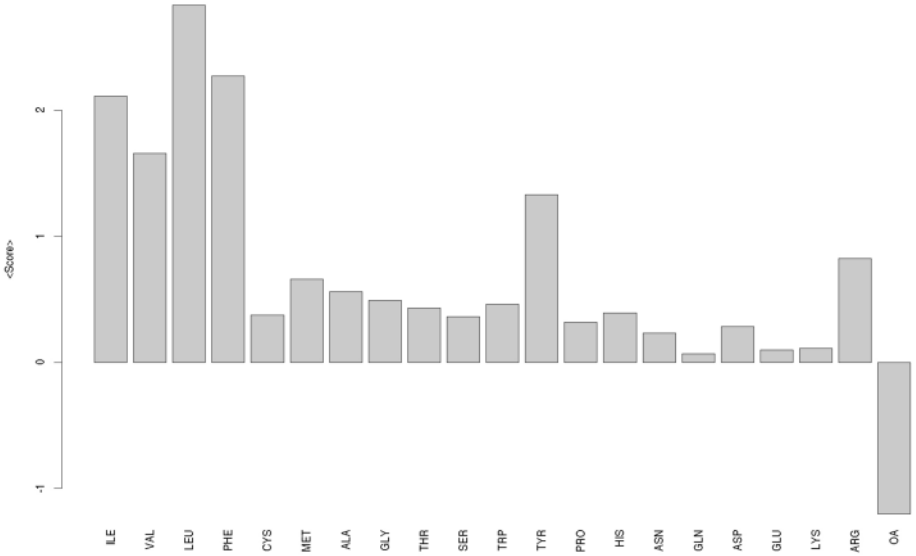

Figure 8.
Breakdown of contact score by residue type and the contribution of overlap area (OA). The value plotted is the score of the residue type in the native structure averaged over all the 640 complexes in the training set.

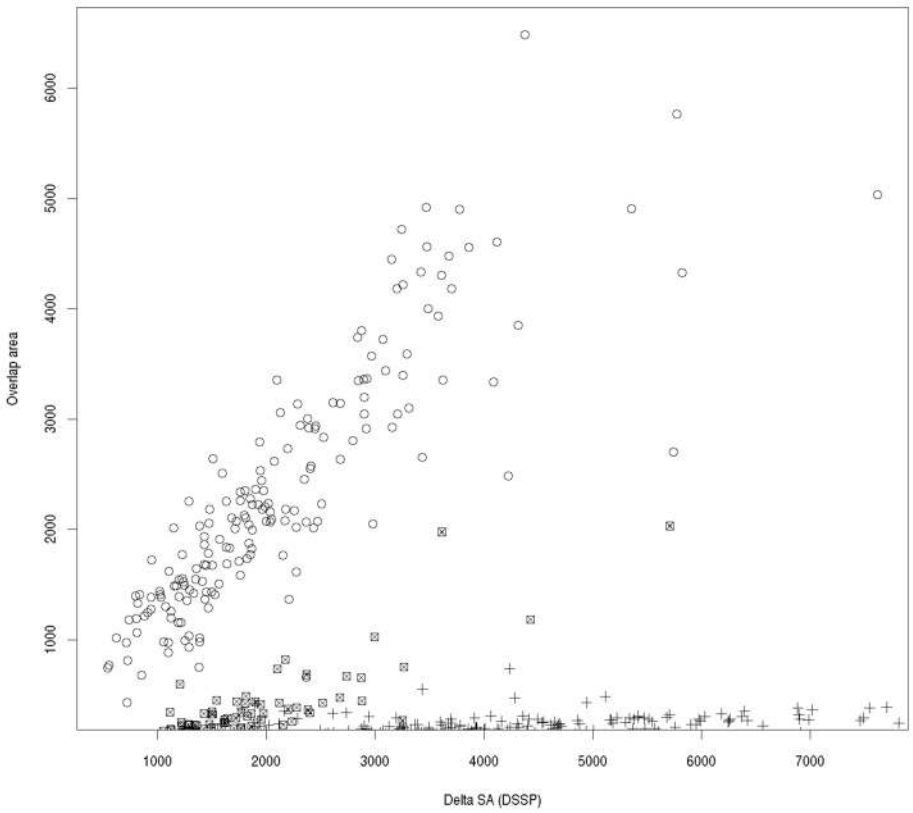

Figure 9.
Comparison of overlap area and change in surface area computed by DSSP (points represented by pluses are bound cases, points represented by circles are unbound cases and points represented by squares are cases in the benchmark2). Bound cases have almost 0 overlap areas, cases in benchmark2 have low overlap area while unbound cases have higher overlap areas.

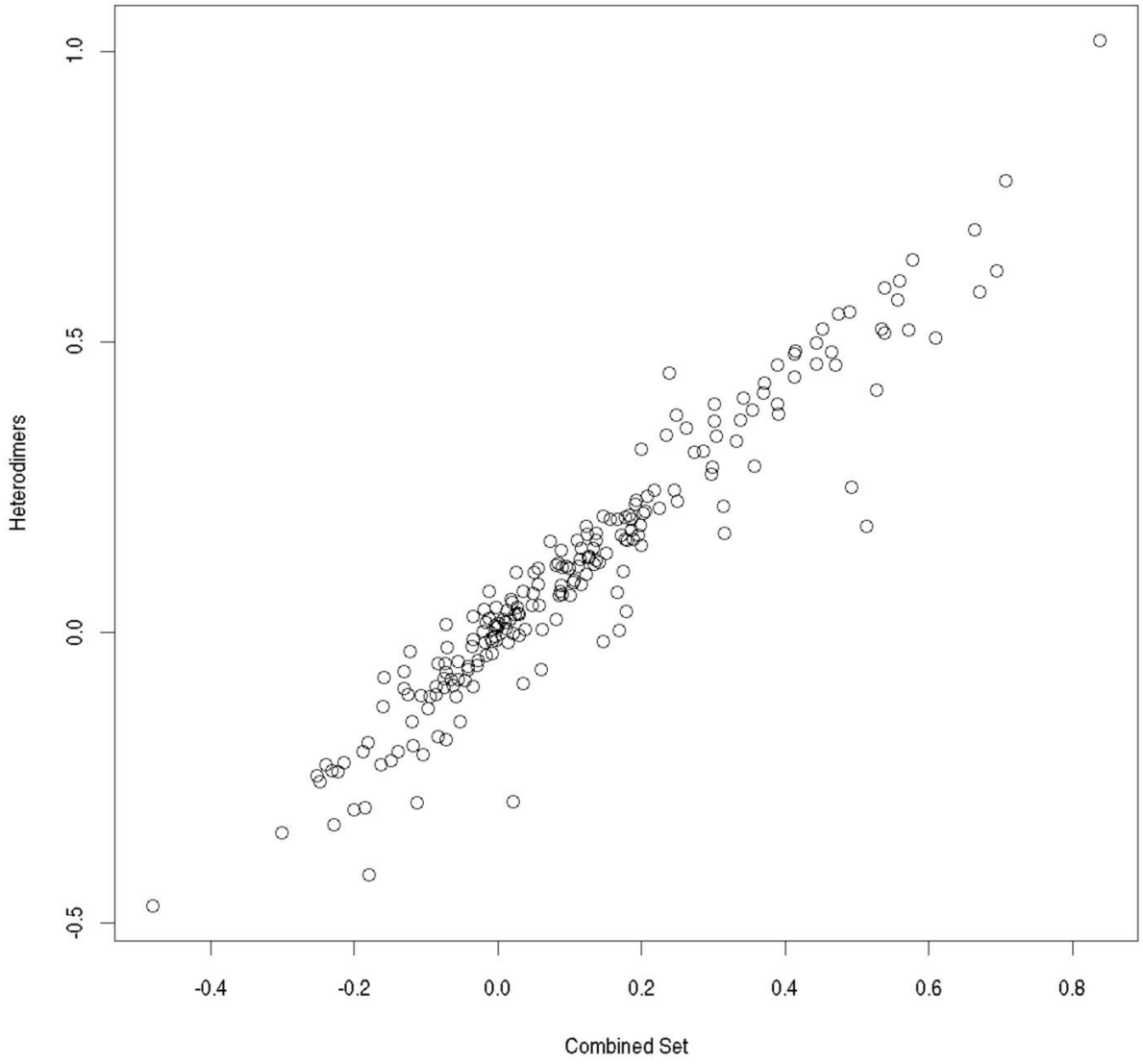

Figure 10.
Comparison of the residue contact weights for the potentials learnt on the combined set and the set of heterodimers. The value plotted is the score of a contact type in the combined potential vs the score in the potential learnt on the set of heterodimers. The correlation coefficient between the potentials is 0.96 (linear fit for heterodimer potential based on the combined potential has a non zero y-intercept).

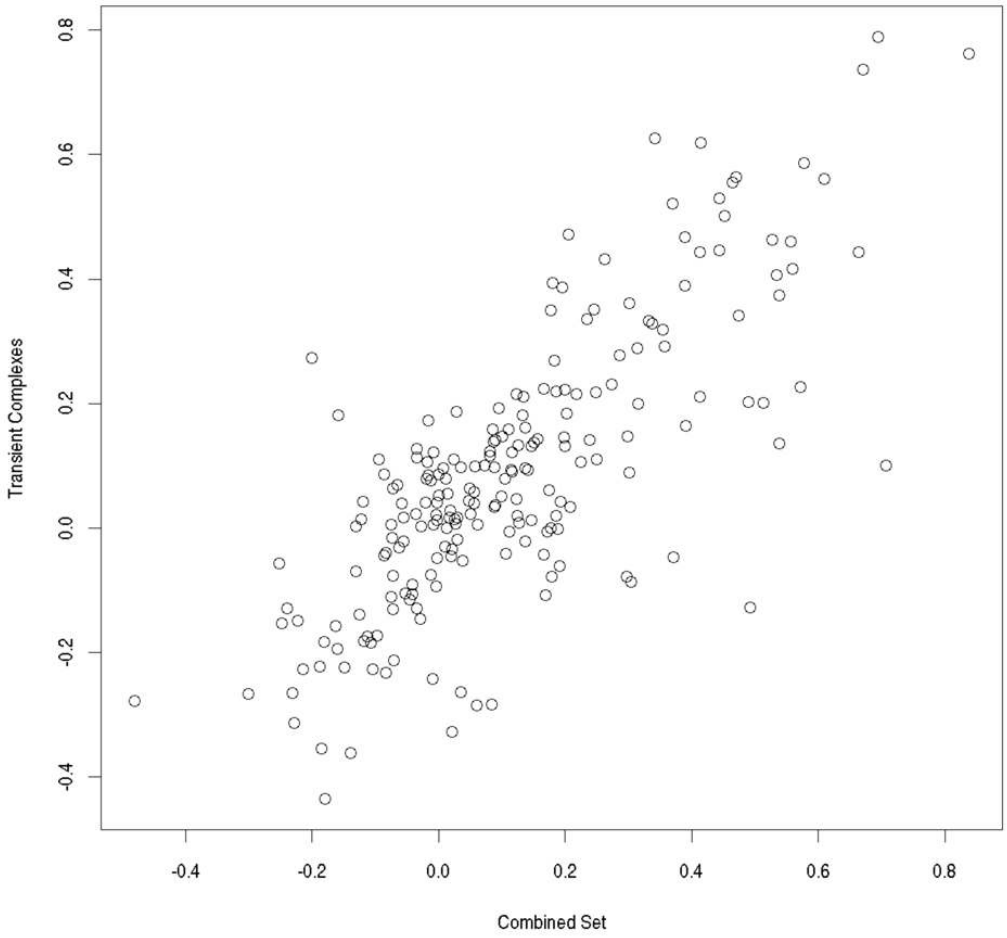

Figure 11.

Comparison of the residue contact weights for the potentials learnt on the combined set and the set of transient complexes. The value plotted is the score of a contact type in the combined potential vs the score in the potential learnt on the set of transient complexes. The correlation coefficient between the potentials is 0.98

Table 1

List of Unbound Complexes (there are 55 unbound-unbound cases and 123 bound-unbound cases).

Receptor	Ligand	Complex	Receptor	Ligand	Complex	Receptor	Ligand	Complex
1fns_H	1ikf_L	15c8(H:L)	1jhl_H	43c9_A	1a7n(H:L)	2dd8_L	1mcw_W	1a8j(L:H)
1ud7_A	1ud7_A	1aar(A:B)	1za5_A	1za5_A	1ap5(A:B)	1f90_H	1a4j_L	1ay1(H:L)
1gke_A	1gke_A	1bm7(A:B)	1oss_A	1ca0_D	1brb(E:I)	1xcm_A	1rrb_-	1c1y(A:B)
1mrc_H	1iai_M	1cf8(H:L)	1xd8_A	1xd8_A	1cp2(A:B)	1oss_A	1pi2_-	1d6r(A:I)
1xap_A	1g2n_A	1dkf(B:A)	1ds6_B	1mh1_-	1doa(B:A)	1hi2_A	1hi2_A	1dyt(B:A)
1dby_A	1dby_A	1fb0(A:B)	1bye_A	1bye_A	1gnw(A:B)	1bnf_A	1fw7_A	1gou(A:B)
1udv_A	1udv_A	1h0x(A:B)	1sct_B	1sct_B	1hbi(A:B)	1e8y_A	2fn4_A	1he8(A:B)
1qo3_C	1qo3_C	1ja3(A:B)	1xsm_-	1xsm_-	1jk0(A:B)	1tfg_-	1tfg_-	1kla(A:B)
3bmp_A	1s4y_A	1lx5(A:B)	2b3a_A	2b3a_A	1lxd(A:B)	2fjh_H	1a4j_L	1mim(H:L)
1rqf_A	1rqf_A	1qf8(A:B)	1oss_A	1ifg_A	1slu(B:A)	1ii3_A	1ii3_A	1snd(A:B)
2fep_A	2fep_A	1sxh(A:D)	1euv_A	1euv_B	1tgz(A:B)	1tid_A	1h4y_A	1th8(A:B)
1q90_A	1pla_-	1tkw(B:A)	1c7g_A	1c7g_A	1tpl(A:B)	2a46_A	2iov_A	1uis(A:B)
2prk_-	1spb_P	1v5i(A:B)	1xxq_A	1xxq_A	1vbp(A:B)	1ud7_A	2dah_A	1wr1(A:B)
1dqs_A	1dqs_A	1xah(B:A)	1bit_-	1shp_-	1yc0(A:I)	2b1k_A	1jzd_C	1z5y(E:D)
2c2v_C	1ud7_A	1zgu(A:B)	1ekx_A	2at1_B	2atc(A:B)	1ud7_A	1z96_A	2g3q(B:A)
1a4j_L	2dtg_A	2g60(L:H)	1ayz_A	1kps_B	2grn(A:B)	1b33_B	1b33_B	2j96(A:B)
2sfa_-	1ytp_A	4sgb(E:I)	1a4f_B	1out_A	1a4f(B:A)	1a50_B	1xcf_A	1a50(B:A)
1ab8_A	1azs_B	1ab8(A:B)	1ac6_A	1qle_L	1ac6(A:B)	1bit_-	1acb_I	1acb(E:I)
1all_B	1b33_B	1all(B:A)	1oss_A	1an1_I	1an1(E:I)	1ant_I	1oyh_I	1ant(I:L)
1oss_A	1avw_B	1avw(A:B)	1py3_A	1ay7_B	1ay7(A:B)	1bi7_A	1d9s_A	1bi7(A:B)
1bis_B	1exq_A	1bis(B:A)	1bqq_T	1rm8_A	1bqq(T:M)	1aqm_-	1bvn_T	1bvn(P:T)
1bxi_B	1gxg_A	1bxi(B:A)	1j6z_A	1c0f_S	1c0f(A:S)	1c5m_D	1kig_L	1c5m(D:F)
1oss_A	1c9p_B	1c9p(A:B)	2csn_-	1cki_B	1cki(A:B)	1ah2_-	1cse_I	1cse(E:I)
2sfa_-	1cso_I	1cso(E:I)	2fv8_A	1cxz_B	1cxz(A:B)	1d9c_A	1fg9_A	1d9c(A:B)
1kfx_L	1df0_B	1df0(A:B)	1dfj_I	1e21_A	1dfj(I:E)	1ii4_E	1djs_B	1djs(A:B)
1vfa_B	1dl7_L	1dl7(H:L)	1ek6_A	1lrk_A	1ek6(A:B)	1euv_A	2io3_B	1euv(A:B)
1exb_A	1eod_A	1exb(A:E)	1exu_A	1ddh_B	1exu(A:B)	1kzz_A	1f3v_A	1f3v(B:A)

Receptor	Ligand	Complex	Receptor	Ligand	Complex	Receptor	Ligand	Complex
1yat_-	1fap_B	1fap(A:B)	1fxt_A	1ud7_A	1fxt(A:B)	1g4u_S	2j0v_A	1g4u(S:R)
1z93_A	1g6v_K	1g6v(A:K)	1gla_G	2gpr_-	1gla(G:F)	1grn_B	1mh1_-	1grn(B:A)
1hdm_A	1k8i_B	1hdm(A:B)	1yvn_A	1hlu_P	1hlu(A:P)	1hjo_A	1hx1_B	1hx1(A:B)
1i7q_A	1i1q_B	1i1q(A:B)	1i7y_B	1b33_B	1i7y(B:A)	1igf_H	1ind_L	1ind(H:L)
1in8_A	1ixs_A	1ixs(B:A)	1j7v_R	1vlk_-	1j7v(R:L)	1g9k_A	1jiw_I	1jiw(P:I)
1jtd_B	1g6a_A	1jtd(B:A)	134l_-	1jtt_A	1jtt(L:A)	2fn4_A	1k8r_B	1k8r(A:B)
1k9o_I	1ym0_A	1k9o(I:E)	2axh_A	1kgc_D	1kgc(E:D)	1ksg_A	1ksj_B	1ksg(A:B)
1ks6_A	1ktz_A	1ktz(B:A)	1kxp_D	1hlu_A	1kxp(D:A)	1r8u_B	1l3e_A	1l3e(B:A)
1miq_A	1ls5_B	1ls5(A:B)	1m4u_A	1waq_A	1m4u(A:L)	1miu_A	1iyj_A	1miu(A:B)
1ml0_A	1bo0_-	1ml0(A:D)	1nw9_B	1xb0_A	1nw9(B:A)	1o97_D	1o94_C	1o97(D:C)
1aoh_A	1ohz_B	1ohz(A:B)	1oph_A	1oss_A	1oph(A:B)	1ud7_A	1otr_A	1otr(B:A)
1oxb_A	1ab6_A	1oxb(A:B)	1r8u_B	1p4q_A	1p4q(B:A)	1p5u_A	1p5v_B	1p5v(A:B)
1gvk_B	1ppf_I	1ppf(E:I)	1ah2_-	1r0r_I	1r0r(E:I)	1mv9_A	1r1k_D	1r1k(A:D)
1r8s_E	1e0s_A	1r8s(E:A)	1rke_A	1qkr_A	1rke(A:B)	2fjh_H	1rzf_L	1rzf(H:L)
1s9i_A	1s9i_A	1s9i(A:B)	1scj_A	1spb_P	1scj(A:B)	1shy_B	2asu_B	1shy(B:A)
1rbl_A	1svd_M	1svd(A:M)	1urd_A	1svx_A	1svx(B:A)	1c9b_B	1tba_A	1tba(B:A)
1tej_B	1rmr_A	1tej(B:A)	1u7d_A	1zh0_A	1u7d(A:B)	2boo_A	1udi_I	1udi(E:I)
1p98_A	1uel_B	1uel(A:B)	2gmi_A	1ur6_B	1ur6(A:B)	1usu_A	1usv_B	1usu(A:B)
1ltx_R	1vg0_B	1vg0(A:B)	1nf1_A	1wq1_R	1wq1(G:R)	1wrd_A	1ud7_A	1wrd(A:B)
1p3c_A	1x7a_L	1x7a(C:L)	1xzv_A	1y01_A	1y01(B:A)	1ah2_-	1y1k_I	1y1k(E:I)
2j0v_A	1yhm_B	1yhm(A:B)	1yke_B	1ykh_A	1ykh(B:A)	1yx5_A	1ud7_A	1yx5(A:B)
1oiv_A	1z0j_B	1z0j(A:B)	2erj_A	1z92_A	1z92(B:A)	2a1a_B	1kl9_A	2a1a(B:A)
2apo_A	2aus_D	2apo(A:B)	2aq2_B	1d9k_B	2aq2(B:A)	2arp_F	1nys_B	2arp(F:A)
2ayo_A	1ud7_A	2ayo(A:B)	2b0z_A	1j3s_A	2b0z(A:B)	2b42_A	1h1a_A	2b42(A:B)
1tyj_A	2b59_B	2b59(A:B)	2d9q_B	1bgd_-	2d9q(B:A)	1ud7_A	2den_A	2den(B:A)
2dx5_A	1ud7_A	2dx5(A:B)	2e31_A	2ass_A	2e31(A:B)	1yt4_A	2g2u_B	2g2u(A:B)
1z22_A	2gzh_B	2gzh(A:B)	2hd5_A	1ud7_A	2hd5(A:B)	1got_A	2ihb_B	2ihb(A:B)
2ipa_B	1nw2_A	2ipa(B:A)	2prk_-	2sic_I	2sic(E:I)	1adu_A	1adt_-	1adu(A:B)
1f6f_B	1axi_A	1bp3(B:A)	2iw8_A	1puc_-	1buh(A:B)	1gz7_A	1gz7_A	1cle(A:B)
2btz_A	2dne_A	1y8n(A:B)	1ywt_A	1ywt_A	1yz5(B:A)	1htl_A	1j2j_A	2a5d(B:A)

Receptor	Ligand	Complex	Receptor	Ligand	Complex	Receptor	Ligand	Complex
1yh2_A	2bf8_B	2bf8(A:B)	1a7h_A	1ha4_A	1a7h(A:B)	1e96_B	1ryf_A	1e96(B:A)
2h57_A	1j2j_B	1j2j(A:B)	1eth_A	1lpa_A	1lpa(B:A)	1efu_A	1xb2_B	1xb2(A:B)
2fju_B	1z2c_A	2fju(B:A)						

Table 2

List of 462 Bound Complexes

11ba(B:A)	1a64(A:B)	1a6d(A:B)	1afw(B:A)	1ah8(B:A)	1aiq(A:B)	1aof(B:A)	1aok(A:B)	1avb(A:B)	1azt(B:A)
1azv(A:B)	1b2p(A:B)	1b3q(B:A)	1b6d(A:B)	1b70(B:A)	1b8m(B:A)	1bdm(B:A)	1bft(A:B)	1bj3(A:B)	1bjy(A:B)
1bk5(A:B)	1bkn(A:B)	1bsl(B:A)	1cd0(A:B)	1cqk(A:B)	1cqp(A:B)	1d0c(A:B)	1djn(A:B)	1djo(A:B)	1dpg(A:B)
1dug(A:B)	1e8a(B:A)	1eaj(A:B)	1ecz(A:B)	1eej(A:B)	1eo2(B:A)	1eq9(A:B)	1equ(A:B)	1ern(B:A)	1es0(B:A)
1ext(A:B)	1eyj(A:B)	1f6b(B:A)	1fn9(A:B)	1fo4(A:B)	1fst(B:A)	1fx0(A:B)	1fxw(F:A)	1fzv(A:B)	1g2q(A:B)
1gpz(A:B)	1gqf(A:B)	1gzp(A:B)	1h0t(A:B)	1h8f(A:B)	1hci(A:B)	1hf9(A:B)	1hw1(A:B)	1i49(A:B)	1ibg(L:H)
1iho(A:B)	1ihr(B:A)	1ihv(A:B)	1ijj(B:A)	1ixm(B:A)	1iz5(A:B)	1j2e(A:B)	1j2f(B:A)	1j3n(A:B)	1j7d(B:A)
1jff(B:A)	1jkg(B:A)	1jl9(B:A)	1jlw(A:B)	1jm6(A:B)	1jya(B:A)	1kba(A:B)	1kzq(A:B)	1l5h(B:A)	1m08(A:B)
1m2v(B:A)	1m48(B:A)	1mmi(A:B)	1msv(A:B)	1n1e(A:B)	1n1j(A:B)	1n3o(A:B)	1n7h(B:A)	1n9j(A:B)	1nch(B:A)
1nou(A:B)	1o0v(B:A)	1oe7(A:B)	1on1(A:B)	1p4k(A:C)	1p53(A:B)	1p9s(A:B)	1pbi(A:B)	1pg3(B:A)	1qav(B:A)
1qpp(B:A)	1qup(A:B)	1r2f(A:B)	1r46(B:A)	1rj9(B:A)	1rlu(A:B)	1sdd(B:A)	1sg3(A:B)	1sox(A:B)	1spp(B:A)
1taf(B:A)	1tij(A:B)	1tw2(B:A)	1u4q(A:B)	1u7i(A:B)	1u9o(A:B)	1uii(A:B)	1uix(A:B)	1uol(A:B)	1utu(A:B)
1v4e(A:B)	1v5w(A:B)	1vbm(A:B)	1vyu(B:A)	1vzg(A:B)	1wio(A:B)	1wpq(A:B)	1wu9(B:A)	1wwa(X:Y)	1x13(A:B)
1x1v(B:A)	1xng(B:A)	1xv8(A:B)	1y4s(A:B)	1ypq(B:A)	1zvf(A:B)	2a4w(B:A)	2a72(A:B)	2a87(A:B)	2aao(A:B)
2adg(B:A)	2ajs(H:L)	2aq0(A:B)	2ask(A:B)	2axh(A:B)	2bdw(A:B)	2bhl(A:B)	2bjn(B:A)	2bm4(A:B)	2bnx(B:A)
2c0j(A:B)	2c36(A:B)	2c7b(A:B)	2ckl(A:B)	2co3(B:A)	2cwx(A:E)	2f1f(A:B)	2f2l(X:A)	2f3g(A:B)	2gb5(A:B)
2ged(A:B)	2ghv(E:C)	2giy(A:B)	2h34(A:B)	2h9a(A:B)	2hkn(A:B)	2hp4(A:B)	2ic2(B:A)	3req(A:B)	3ygs(P:C)
1a22(B:A)	1a6u(H:L)	1abr(B:A)	1ai4(B:A)	1aj7(H:L)	1aks(A:B)	1aqk(H:L)	1aro(P:L)	1atn(A:D)	1aui(A:B)
1b34(A:B)	1b41(A:B)	1baf(H:L)	1bdj(A:B)	1bfv(H:L)	1bh8(B:A)	1bkd(S:R)	1bmq(A:B)	1bpl(B:A)	1bun(A:B)
1cc1(L:S)	1cee(A:B)	1cf4(A:B)	1cgi(E:I)	1cgs(L:H)	1ci6(A:B)	1d4v(B:A)	1d5s(A:B)	1dhk(A:B)	1dj7(A:B)
1dqd(H:L)	1dtd(A:B)	1dtw(A:B)	1dwl(B:A)	1e0a(A:B)	1e44(B:A)	1e9y(B:A)	1efv(A:B)	1eg9(A:B)	1ep1(A:B)
1euc(B:A)	1f02(I:T)	1f2t(A:B)	1f34(A:B)	1f45(A:B)	1f60(A:B)	1fai(H:L)	1fc2(D:C)	1fle(E:I)	1fm0(E:D)
1fm2(B:A)	1fpp(B:A)	1fq1(B:A)	1fs0(G:E)	1g4y(R:B)	1gfw(A:B)	1ggr(A:B)	1gh6(B:A)	1gl4(A:B)	1h1v(A:G)
1h2s(A:B)	1h2t(C:Z)	1h32(A:B)	1h59(A:B)	1hcn(B:A)	1htr(B:P)	1i1r(A:B)	1i4e(A:B)	1i72(A:B)	1iar(B:A)
1ira(Y:X)	1ire(B:A)	1itb(B:A)	1j6t(A:B)	1jma(A:B)	1jow(B:A)	1jql(A:B)	1jsd(A:B)	1jv2(A:B)	1jw9(B:D)
1k28(A:D)	1ka9(F:H)	1kac(A:B)	1kbh(B:A)	1kmi(Z:Y)	1l4d(A:B)	1ldj(A:B)	1luj(A:B)	1lzw(B:A)	1m10(B:A)
1mco(H:L)	1mhm(A:B)	1npe(A:B)	1nql(A:B)	1nrj(B:A)	1nt2(B:A)	1nun(B:A)	1o2f(A:B)	1o5e(H:L)	1o6s(A:B)
1oe9(A:B)	1of5(A:B)	1oo0(A:B)	1oo9(A:B)	1ory(A:B)	1pdk(A:B)	1pg5(A:B)	1psk(L:H)	1qbk(B:C)	1qdl(A:B)

1qge(D:E)	1qgk(A:B)	1qs0(A:B)	1r8o(A:B)	1rf8(A:B)	1s70(A:B)	1sc5(A:B)	1shw(B:A)	1sko(A:B)	1sq2(L:N)
1stf(E:I)	1t0h(B:A)	1t0p(A:B)	1t6b(X:Y)	1ta3(B:A)	1tdq(A:B)	1te1(A:B)	1tf0(A:B)	1tfk(A:B)	1ti8(A:B)
1tlh(A:B)	1tmq(A:B)	1tnr(A:R)	1txq(A:B)	1u0s(Y:A)	1u5s(A:B)	1u7e(A:B)	1ujw(A:B)	1ukv(G:Y)	1us7(A:B)
1uzx(A:B)	1vra(B:A)	1w98(A:B)	1wa8(A:B)	1whs(A:B)	1wmh(A:B)	1wpx(A:B)	1wql(A:B)	1wx2(A:B)	1wyw(A:B)
1x3w(A:B)	1xdt(T:R)	1xew(Y:X)	1xg2(A:B)	1xou(B:A)	1xrs(A:B)	1xt9(A:B)	1xtg(A:B)	1y56(A:B)	1y64(B:A)
1y8x(A:B)	1ycs(B:A)	1yrt(A:B)	1yuk(B:A)	1yvb(A:I)	1z00(A:B)	1z1d(B:A)	1z3e(A:B)	1zbd(A:B)	1zbx(A:B)
1zhh(A:B)	1zl8(B:A)	1zlh(A:B)	1zls(H:L)	1zun(B:A)	2a78(B:A)	2acm(A:B)	2aho(A:B)	2b3t(B:A)	2bez(C:F)
2blf(A:B)	2c0l(A:B)	2c1m(A:B)	2c7m(B:A)	2cg5(A:B)	2co6(B:A)	2cyz(B:A)	2d5r(A:B)	2d74(A:B)	2dfx(I:E)
2dvw(A:B)	2e30(A:B)	2f4m(A:B)	2fbj(H:L)	2few(A:B)	2ffk(A:B)	2fh5(B:A)	2fho(B:A)	2fom(B:A)	2ftx(A:B)
2fwl(B:A)	2fyl(B:A)	2g16(B:A)	2g77(A:B)	2ga9(D:A)	2gsk(A:B)	2gy7(B:A)	2hi7(A:B)	2hle(A:B)	2hym(A:B)
2ie4(A:C)	2ipp(B:A)	2iw5(A:B)	2iwt(B:A)	2j2z(A:B)	2kin(A:B)	2nxn(A:B)	2o2v(A:B)	2o3b(A:B)	2oob(B:A)
3eza(A:B)	2ns1(A:B)	2ehb(A:D)	2hdi(A:B)	2icn(A:B)	2j5y(A:B)	2jqr(A:B)	2jss(A:B)	2jt4(B:A)	2ju0(A:B)
2jxc(A:B)	2k3s(A:B)	2k6d(B:A)	2nqd(B:A)	2nts(P:A)	2nz8(B:A)	2o8v(A:B)	2ocf(A:D)	2ot3(A:B)	2ozn(A:B)
2p1m(B:A)	2p43(A:B)	2p7v(A:B)	2pa8(D:L)	2pjh(B:A)	2pjw(V:H)	2pnz(B:A)	2px9(A:B)	2q2e(B:A)	2q97(A:T)
2qc1(B:A)	2qkl(A:B)	2qkw(B:A)	2qwn(A:B)	2raw(A:B)	2rd0(A:B)	2rd7(A:C)	2rlz(A:B)	2rms(A:B)	2rnr(B:A)
2uuy(A:B)	2v1y(B:A)	2v3b(A:B)	2v8s(E:V)	2v9t(B:A)	2ver(A:N)	2vol(A:B)	2vrw(B:A)	2vsm(A:B)	2z0d(A:B)
2z2r(A:B)	2z59(A:B)	2z5b(B:A)	2z64(A:C)	2zf5(O:Y)	2zfd(A:B)	3beg(A:B)	3bei(B:A)	3bes(R:L)	3bn3(B:A)
3buz(A:B)	3by4(A:B)	3byh(A:B)	3c5x(A:C)	3c98(A:B)	3cbj(A:B)	3cf4(A:G)	3cqc(A:B)	3daw(A:B)	3ddc(A:B)
3dgp(B:A)	3eb6(B:A)								

Table 3

Scoring Decoys generated by ZDOCK+ZRANK (ZR) compared to our scoring function (P) (PIE540 was used), our filter with our scoring function (FP) and our filter with ZRANK (FZR). TopN is the number of hits within top N. The column labeled Random gives the rank for which the probability that a random scoring function will do better is at-least 0.5 (as explained in the text).

Case	Bestrank				Top10	Top20	Top100	hits filtered	Total hits	Random
	FP	P	F+ZR	ZR	FP/P/FZR/ZR	FP/P/FZR/ZR	FP/P/FZR/ZR			
1A2K	1	1	1	1038	9/4/10/0	15/9/15/0	21/20/21/0	145	558	68
1ACB	1	21	1	780	9/0/9/0	17/0/18/0	74/6/69/0	218	492	77
1AHW	1	6	1	27	8/2/9/0	11/3/16/0	37/10/54/6	168	346	109
1AK4	2	23	1	1315	7/0/5/0	11/0/13/0	63/14/72/0	196	253	148
1AKJ	728	3963	23	175	0/0/0/0	0/0/0/0	0/0/1/0	65	233	161
1ATN	373	2693	105	6872	0/0/0/0	0/0/0/0	0/0/0/0	5	5	6991
1AVX	1	1	1	11	10/6/10/0	20/8/20/2	93/42/99/7	426	742	51
1AY7	81	277	36	74	0/0/0/0	0/0/0/0	1/0/2/2	257	397	95
1B6C	1	1	1	1	10/7/10/9	20/13/20/16	86/71/84/41	113	473	80
1BJ1	1	4	1	19	10/3/10/0	20/7/20/1	98/60/100/5	646	1604	24
1BUH	1	3	1	353	8/4/9/0	14/5/17/0	82/32/84/0	360	421	89
1BVK	37	102	6	116	0/0/1/0	0/0/2/0	5/0/8/0	326	409	92
1BVN	1	1	1	10	10/10/10/1	20/18/20/3	58/66/59/14	74	528	71
1CGI	1	4	1	22	10/2/10/0	20/6/20/0	98/41/97/15	360	846	45
1D6R	33	173	17	2347	0/0/0/0	0/0/1/0	1/0/1/0	3	32	1158
1DE4	1	4	4	426	2/3/2/0	4/3/3/0	16/17/18/0	32	92	406
1DFJ	1	1	1	2	8/10/7/3	10/20/7/3	10/95/8/10	11	302	124
1DQJ	5	47	11	1620	1/0/0/0	1/0/2/0	9/1/20/0	214	263	143
1E6E	1	1	1	3	10/9/10/4	19/18/16/9	60/58/60/30	152	413	91
1E6J	1	13	1	1	10/0/10/7	20/3/20/10	98/35/100/34	714	1152	33
1E96	54	51	67	24	0/0/0/0	0/0/0/0	1/2/5/5	29	196	191
1EAW	1	18	1	1	10/0/10/4	20/2/20/4	64/15/74/16	216	531	71
1EER	38	309	29	330	0/0/0/0	0/0/0/0	1/0/1/0	5	8	4482
1EWY	1	3	1	21	10/4/7/0	17/9/14/0	55/33/69/9	260	510	74

Case	Bestrank				Top10	Top20	Top100	hits filtered	Total hits	Random
	FP	P	F+ZR	ZR	FP/P/FZR/ZR	FP/P/FZR/ZR	FP/P/FZR/ZR			
1EZU	1	1	1	2247	7/2/8/0	10/3/9/0	21/9/20/0	93	330	114
1F34	1	2	1	62	10/6/10/0	20/7/20/0	33/33/34/1	60	150	249
1F51	86	936	65	3	0/0/0/2	0/0/0/2	3/0/6/7	65	299	126
1FC2	1	1	1	154	4/9/3/0	4/14/3/0	4/33/4/0	4	62	601
1FQ1	20	95	1	15260	0/0/1/0	1/0/1/0	1/1/1/0	8	9	4003
1FQJ	724	3479	933	491	0/0/0/0	0/0/0/0	0/0/0/0	19	24	1538
1FSK	1	2	1	1	9/5/9/9	19/10/19/17	96/54/95/28	478	818	46
1GCQ	214	3149	246	922	0/0/0/0	0/0/0/0	0/0/0/0	100	140	267
1GHQ	0	48852	0	2982	0/0/0/0	0/0/0/0	0/0/0/0	0	2	15817
1GP2	6	46	9	133	1/0/1/0	3/0/5/0	23/2/19/0	141	276	136
1GRN	12	1486	11	558	0/0/0/0	2/0/3/0	9/0/12/0	122	163	230
1HE1	4	91	2	36	2/0/4/0	6/0/10/0	18/1/36/3	193	253	148
1HE8	188	1120	208	75	0/0/0/0	0/0/0/0	0/0/0/1	6	6	5892
1HIA	0	3	0	1998	0/2/0/0	0/2/0/0	0/4/0/0	0	59	631
1I2M	1	1	1	473	5/3/8/0	10/4/12/0	29/16/25/0	49	80	466
1I4D	4	19	3	1349	1/0/4/0	3/1/5/0	9/2/21/0	102	351	107
1I9R	9	124	9	38	1/0/1/0	4/0/2/0	24/0/15/5	74	331	113
1IB1	0	52519	0	50414	0/0/0/0	0/0/0/0	0/0/0/0	0	1	27001
1IJK	2	11	1	444	7/0/6/0	13/4/12/0	43/19/47/0	98	113	331
1IQD	1	2	1	1	10/9/10/9	20/17/20/14	100/72/100/44	433	795	48
1JPS	1	6	1	1	10/1/9/1	19/2/18/2	54/10/58/10	202	385	98
1K4C	18	22	62	162	0/0/0/0	2/0/0/0	10/5/3/0	512	1302	29
1K5D	20	288	6	84	0/0/2/0	1/0/3/0	21/0/31/1	97	141	265
1KAC	1	7	1	11	7/1/6/0	11/2/9/1	24/6/19/3	92	157	238
1KKL	5	1	2	70	4/6/8/0	7/8/11/0	15/26/18/1	27	159	235
1KLU	90	1074	137	13333	0/0/0/0	0/0/0/0	1/0/0/0	18	18	2040
1KTZ	67	588	169	397	0/0/0/0	0/0/0/0	3/0/0/0	80	90	415
1KXP	1	3	1	12	4/7/5/0	5/14/5/2	6/74/7/6	17	282	133
1KXQ	1	3	1	14	8/3/10/0	17/7/20/2	37/31/40/4	69	156	240

Case	Bestrank				Top10	Top20	Top100	hits filtered	Total hits	Random
	FP	P	F+ZR	ZR	FP/P/FZR/ZR	FP/P/FZR/ZR	FP/P/FZR/ZR			
1MAH	1	3	4	3	6/3/2/3	10/4/3/6	62/28/31/15	460	668	57
1ML0	1	1	1	1	10/7/10/6	18/15/20/10	49/42/63/33	183	529	71
1MLC	1	8	1	5	8/1/8/3	12/1/9/3	48/7/40/8	514	576	65
1N2C	16	103	15	4505	0/0/0/0	1/0/3/0	7/0/7/0	8	51	729
1NCA	54	923	30	14	0/0/0/0	0/0/0/1	2/0/7/5	87	123	304
1NSN	1	8	1	468	8/2/5/0	11/5/8/0	28/13/20/0	64	139	269
1PPE	1	1	1	1	10/10/10/10	20/20/20/19	62/83/71/72	576	2410	16
1QA9	118	807	96	1850	0/0/0/0	0/0/0/0	0/0/1/0	26	29	1276
1QFW	29	430	26	192	0/0/0/0	0/0/0/0	18/0/17/0	85	107	349
1RLB	1	4	1	1	10/3/9/6	16/7/16/8	61/29/70/31	212	1292	29
1SBB	848	9740	760	3639	0/0/0/0	0/0/0/0	0/0/0/0	19	26	1421
1TMQ	1	1	1	71	10/8/7/0	18/14/13/0	54/39/51/1	176	353	106
1UDI	1	1	1	2	7/6/7/1	10/9/11/2	19/20/21/4	83	306	123
1VFB	63	44	77	437	0/0/0/0	0/0/0/0	1/2/1/0	215	326	115
1WEJ	1	1	1	2	10/9/10/1	20/15/19/4	98/63/90/9	428	715	53
1WQ1	29	105	32	296	0/0/0/0	0/0/0/0	3/0/3/0	5	139	269
2BTF	1	2	1	151	10/7/10/0	20/13/20/0	68/49/78/0	192	275	136
2HMI	1	181	1	272	4/0/9/0	8/0/14/0	39/0/50/0	201	331	113
2JEL	1	1	1	42	10/10/10/0	20/20/20/0	74/87/70/1	296	1020	37
2MTA	1	25	1	57	5/0/10/0	8/0/18/0	34/6/62/2	157	553	68
2PCC	1	4	3	238	4/3/3/0	7/5/6/0	41/29/48/0	201	276	136
2QFW	1	9	1	6	10/1/8/2	19/1/17/3	84/18/75/23	183	509	74
2SIC	1	1	1	1	10/9/10/8	20/18/20/9	100/95/98/20	288	764	49
2SNI	1	1	1	114	10/8/10/0	20/12/20/0	87/48/96/0	314	523	72
2VIS	146	889	215	8	0/0/0/1	0/0/0/1	0/0/0/4	136	703	54
7CEI	1	1	1	3	10/9/10/7	19/16/20/11	79/59/94/53	399	965	39

Table 4

Scoring decoys generated by Patchdock (PD).

Case	Bestrank				Top10	Top20	Top100	hits filtered	#sampled	Total hits	Random
	FP	P	F+PD	PD	FP/P/FPD/PD	FP/P/FPD/PD	FP/P/FPD/PD				
1A2K	1	4	10	291	1/1/1/0	2/1/1/0	4/3/2/0	11	24150	16	1024
1ACB	1	1	450	406	2/2/0/0	2/3/0/0	6/6/0/0	8	13458	14	651
1AHW	2	3	270	168	4/2/0/0	4/3/0/0	10/5/0/0	19	20711	19	742
1AK4	269	448	1937	2160	0/0/0/0	0/0/0/0	0/0/0/0	1	12480	2	3656
1AKJ	114	233	240	545	0/0/0/0	0/0/0/0	0/0/0/0	3	22671	10	1519
1AVX	1	1	119	382	8/5/0/0	10/8/0/0	18/12/0/0	32	25203	36	481
1AY7	3	3	24	321	3/2/0/0	6/3/0/0	10/8/1/0	20	8745	21	284
1B6C	1	1	62	99	7/5/0/0	9/7/0/0	11/12/1/1	11	17691	13	919
1BJ1	2	2	155	4632	3/3/0/0	4/3/0/0	6/5/0/0	7	17352	10	1163
1BUH	1	1	15	431	8/7/0/0	14/9/1/0	21/17/3/0	24	12292	25	337
1BVK	5	13	72	347	1/0/0/0	2/1/0/0	8/4/1/0	25	15526	25	425
1BVN	1	1	16	32	6/6/0/0	8/9/1/0	17/17/1/1	19	24239	32	520
1CGI	1	4	162	131	5/3/0/0	9/5/0/0	11/10/0/0	13	10663	15	482
1D6R	12	42	149	67	0/0/0/0	1/0/0/0	2/1/0/1	27	8312	28	204
1DQJ	2	3	431	1313	2/1/0/0	6/3/0/0	10/6/0/0	13	16619	13	863
1E6E	1	3	512	39	1/1/0/0	3/2/0/0	4/4/0/1	6	24736	11	1511
1E6J	1	2	211	1513	3/2/0/0	5/5/0/0	12/9/0/0	20	14105	20	481
1E96	6	12	15	511	1/0/0/0	2/1/1/0	10/6/2/0	15	15205	17	608
1EAW	1	1	11	59	5/4/0/0	11/4/1/0	16/12/1/1	27	13387	30	306
1EER	0	6	0	1273	0/1/0/0	0/1/0/0	0/1/0/0	0	10513	2	3080
1EWY	1	2	136	32	6/4/0/0	10/6/0/0	15/16/0/2	17	14902	19	534
1EZU	2	4	1077	1206	1/1/0/0	1/1/0/0	2/1/0/0	2	22766	6	2484
1F34	1	1	222	7	9/8/0/1	14/13/0/2	16/20/0/4	17	17986	25	492
1F51	6	18	78	365	2/0/0/0	2/1/0/0	5/3/1/0	9	25292	11	1545
1FC2	22	23	296	2319	0/0/0/0	0/0/0/0	1/1/0/0	1	5163	1	2582
1FQJ	74	250	27	318	0/0/0/0	0/0/0/0	1/0/1/0	7	18380	12	1032

Case	Bestrank				Top10	Top20	Top100	hits filtered	#sampled	Total hits	Random
	FP	P	F+PD	PD	FP/P/FPD/PD	FP/P/FPD/PD	FP/P/FPD/PD				
1FSK	1	2	98	420	5/3/0/0	8/4/0/0	16/13/1/0	22	19513	22	606
1GCQ	32	76	108	2172	0/0/0/0	0/0/0/0	1/1/0/0	8	7575	8	629
1GHQ	200	1014	617	5368	0/0/0/0	0/0/0/0	0/0/0/0	8	21859	12	1227
1GRN	1	3	47	40	1/1/0/0	1/1/0/0	13/2/3/1	30	19834	30	454
1HE1	3	4	18	10	2/2/0/1	4/3/1/1	7/7/1/2	13	16253	18	614
1HE8	59	204	86	4443	0/0/0/0	0/0/0/0	1/0/1/0	4	40790	5	5281
1HIA	218	395	910	525	0/0/0/0	0/0/0/0	0/0/0/0	3	9948	4	1583
1I2M	0	1398	0	2542	0/0/0/0	0/0/0/0	0/0/0/0	0	24304	1	12153
1I4D	20	13	91	268	0/0/0/0	1/1/0/0	4/4/1/0	12	51824	22	1608
1I9R	8	8	664	42	1/1/0/0	1/1/0/0	3/3/0/1	5	1626	5	211
1IJK	1	1	371	3874	3/1/0/0	5/3/0/0	7/8/0/0	7	29135	10	1952
1IQD	1	1	406	3228	5/4/0/0	5/5/0/0	6/5/0/0	6	22307	6	2434
1JPS	3	4	174	1185	4/4/0/0	6/5/0/0	8/7/0/0	13	12902	13	670
1K4C	42	42	135	152	0/0/0/0	0/0/0/0	1/1/0/0	4	1585	4	253
1KAC	24	72	636	179	0/0/0/0	0/0/0/0	3/1/0/0	21	16883	25	462
1KKL	6	8	1663	1835	1/1/0/0	1/1/0/0	1/1/0/0	2	22886	2	6704
1KLU	8	24	928	6434	2/0/0/0	3/0/0/0	6/3/0/0	7	20167	8	1674
1KTZ	77	258	1605	10811	0/0/0/0	0/0/0/0	1/0/0/0	1	13210	1	6606
1KXP	1	1	541	29	2/2/0/0	3/3/0/0	3/3/0/2	3	1385	3	286
1KXQ	2	1	316	11	1/3/0/0	1/3/0/1	1/7/0/1	4	26025	13	1352
1MAH	14	55	347	438	0/0/0/0	1/0/0/0	5/1/0/0	12	18776	12	1054
1ML0	1	1	9	24	3/5/1/0	6/5/1/0	9/8/1/1	12	24656	17	986
1MLC	9	23	563	847	1/0/0/0	3/0/0/0	6/4/0/0	10	21397	10	1433
1NCA	14	15	575	469	0/0/0/0	2/2/0/0	3/3/0/0	6	4098	6	448
1NSN	4	13	329	1254	3/0/0/0	4/2/0/0	4/4/0/0	5	18534	5	2400
1PPE	1	1	165	12	4/5/0/0	5/6/0/2	9/9/0/2	17	3422	18	130
1QA9	235	570	382	2379	0/0/0/0	0/0/0/0	0/0/0/0	10	18547	10	1243
1QFW	4	10	77	1457	2/1/0/0	3/1/0/0	4/3/1/0	7	15288	7	1442
1RLB	4	7	737	4845	1/2/0/0	1/2/0/0	3/2/0/0	4	26251	13	1364

Case	Bestrank				Top10	Top20	Top100	hits filtered	#sampled	Total hits	Random
	FP	P	F+PD	PD	FP/P/FPD/PD	FP/P/FPD/PD	FP/P/FPD/PD				
1SBB	27	134	56	3079	0/0/0/0	0/0/0/0	1/0/1/0	13	20662	13	1073
1TMQ	1	1	308	3	5/3/0/1	7/6/0/1	13/13/0/1	18	34451	26	907
1UDI	1	1	145	28	4/3/0/0	5/5/0/0	11/8/0/2	14	15276	16	648
1VFB	7	8	28	1541	1/1/0/0	2/1/0/0	10/3/1/0	36	16145	37	300
1WEJ	1	2	27	2152	5/4/0/0	9/8/0/0	18/17/2/0	26	16115	26	424
1WQ1	20	53	84	208	0/0/0/0	1/0/0/0	2/2/1/0	8	27324	9	2026
2BTF	1	1	54	4994	2/2/0/0	3/3/0/0	4/5/1/0	6	27702	7	2612
2HMI	10	10	1164	1469	1/1/0/0	1/1/0/0	1/1/0/0	1	1735	1	868
2JEL	1	1	391	288	7/4/0/0	10/8/0/0	20/18/0/0	27	14280	30	327
2MTA	1	1	58	10	4/2/0/1	8/4/0/1	21/16/3/1	33	27774	41	466
2PCC	10	15	327	1132	1/0/0/0	3/2/0/0	8/6/0/0	17	23572	17	942
2QFW	2	10	97	1018	2/1/0/0	3/2/0/0	12/4/1/0	19	13733	19	492
2SIC	1	1	1	109	9/9/1/0	10/13/2/0	21/23/3/0	27	22514	39	397
2SNI	1	4	145	580	6/4/0/0	10/6/0/0	18/10/0/0	22	12387	25	339
2VIS	6	6	107	253	1/1/0/0	1/1/0/0	3/3/0/0	3	619	3	128
7CEI	3	4	281	117	2/2/0/0	5/4/0/0	11/10/0/0	17	6731	17	269

Table 5

Result of our protocol grouped by ease (according to benchmark2) – SP is the result for our sampling coupled with our coarse grained potential, FP+ZR is the result of our filtering and scoring of decoys generated by ZDOCK3.0 with ZRANK and ZR is the result of ZDOCK3.0 with ZRANK. The columns indicate the number of cases where a hit was ranked at the top/top10/top100, the last column indicates the number of cases where a hit was sampled.

Category	Rank1	Top10	Top100	With Hits
	SP|FP+ZR|ZR	SP|FP+ZR|ZR	SP|FP+ZR|ZR	S|ZR
Easy	19|39|10	33|44|21	43/55|38	60/63
Medium	1|2|0	7|5|0	8|9|3	11|11
Difficult	0|2|0	0|2|0	3|4|0	4|5

Table 6

Summarized results for comparison of our filter (F) and scoring function (P) to Tobi et al[28] (TB), Lu et al[23] (LLS) and Glaser et al[22] (GSVB) in ranking structures generated by ZDOCK3.0 with ZRANK.

Potential	Rank1	Top10	Top100
P	19	40	54
TB	6	14	39
LLS	4	10	27
FP	43	51	68
F+TB	36	55	65
F+LLS	29	44	57
F+GSVB	7	24	49

Table 7

Comparison of our scoring function (P) to Tobi et al[28] (TB) and Lu et al[23] (LLS).

Case	Bestrank			Bestrank			Random	Random	Top10	Topf10	Top100	Topf100
	P	TB	LLS	FP	F+TB	F+LLS	no filter	filtered	P/TB/LLS	P/TB/LLS	P/TB/LLS	P/TB/LLS
1A2K	1	2852	45	1	73	1	68	12	4/0/0	9/0/6	20/0/1	21/1/25
1ACB	21	25	9	1	2	1	77	8	0/0/1	9/6/9	6/1/7	74/22/74
1AHW	6	149	3530	1	4	5	109	11	2/0/0	8/3/1	10/0/0	37/28/8
1AK4	23	1030	189	2	20	1	148	9	0/0/0	7/0/5	14/0/0	63/7/33
1AKJ	3963	16	46479	728	1	2352	161	27	0/0/0	0/7/0	0/7/0	0/30/0
1ATN	2693	6214	350	373	240	10	6991	324	0/0/0	0/0/1	0/0/0	0/0/3
1AVX	1	29	50	1	1	1	51	5	6/0/0	10/6/10	42/2/9	93/56/93
1AY7	277	103	1404	81	25	555	95	7	0/0/0	0/0/0	0/0/0	1/9/0
1B6C	1	24	83	1	5	1	80	16	7/0/0	10/2/10	71/7/1	86/8/87
1BJ1	4	3	56	1	1	1	24	3	3/1/0	10/10/10	60/18/8	98/98/100
1BUH	3	302	74	1	1	1	89	5	4/0/0	8/9/10	32/0/1	82/54/73
1BVK	102	215	1228	37	2	3	92	6	0/0/0	0/2/1	0/0/0	5/43/9
1BVN	1	1	1	1	1	1	71	24	10/3/4	10/10/10	66/37/36	58/44/39
1CGI	4	14	3	1	1	1	45	5	2/0/1	10/9/10	41/7/9	98/84/94
1D6R	173	16942	7028	33	1880	2386	1158	516	0/0/0	0/0/0	0/0/0	1/0/0
1DE4	4	325	174	1	81	20	406	54	3/0/0	2/0/0	17/0/0	16/1/4
1DFJ	1	10	2361	1	1	99	124	153	10/1/0	8/6/0	95/17/0	10/10/1
1DQJ	47	252	2580	5	1	37	143	9	0/0/0	1/4/0	1/0/0	9/31/8
1E6E	1	505	187	1	4	1	91	12	9/0/0	10/1/1	58/0/0	60/11/8
1E6J	13	5	5	1	1	1	33	3	0/3/1	10/10/10	35/21/8	98/97/92
1E96	51	105	67	54	60	4	191	60	0/0/0	0/0/1	2/0/3	1/2/3
1EAW	18	528	268	1	9	26	71	9	0/0/0	10/1/0	15/0/0	64/1/5
1EER	309	9156	43898	38	152	2203	4482	324	0/0/0	0/0/0	0/0/0	1/0/0
1EWY	3	65	366	1	1	3	74	7	4/0/0	10/9/1	33/1/0	55/69/19
1EZU	1	126	27	1	1	1	114	19	2/0/0	7/6/9	9/0/8	21/20/43
1F34	2	28	3371	1	1	11	249	29	6/0/0	10/1/0	33/1/0	33/10/4

Case	Bestrank			Bestrank			Random	Random	Top10	Topf10	Top100	Topf100
	P	TB	LLS	FP	F+TB	F+LLS	no filter	filtered	P/TB/LLS	P/TB/LLS	P/TB/LLS	P/TB/LLS
1F51	936	403	2542	86	9	463	126	27	0/0/0	0/1/0	0/0/0	3/10/0
1FC2	1	2	12	1	1	1	601	398	9/2/0	4/2/4	33/6/12	4/4/4
1FQ1	95	4122	19813	20	193	648	4003	208	0/0/0	0/0/0	1/0/0	1/0/0
1FQJ	3479	2593	38696	724	51	2083	1538	90	0/0/0	0/0/0	0/0/0	0/1/0
1FSK	2	2	200	1	1	1	46	4	5/5/0	9/10/10	54/16/0	96/91/84
1GCQ	3149	1797	747	214	12	2	267	18	0/0/0	0/0/6	0/0/0	0/7/30
1GHQ	48852	46172	41821	0	0	0	15817	0	0/0/0	0/0/0	0/0/0	0/0/0
1GP2	46	17	66	6	1	1	136	13	0/0/0	1/10/9	2/5/2	23/65/49
1GRN	1486	453	7807	12	1	109	230	15	0/0/0	0/3/0	0/0/0	9/14/0
1HE1	91	34	2287	4	4	79	148	9	0/0/0	2/4/0	1/1/0	18/21/1
1HE8	1120	8534	42390	188	396	2210	5892	273	0/0/0	0/0/0	0/0/0	0/0/0
1HIA	3	48	289	0	0	0	631	0	2/0/0	0/0/0	4/1/0	0/0/0
1I2M	1	3	1191	1	6	72	466	36	3/1/0	5/2/0	16/4/0	29/22/1
1I4D	19	112	89	4	10	1	107	17	0/0/0	1/1/4	2/0/2	9/7/26
1I9R	124	2342	761	9	116	18	113	24	0/0/0	1/0/0	0/0/0	24/0/5
1IB1	52519	34467	53385	0	0	0	27001	0	0/0/0	0/0/0	0/0/0	0/0/0
1IJK	11	153	221	2	1	10	331	18	0/0/0	7/5/1	19/0/0	43/36/19
1IQD	2	1	18	1	1	1	48	4	9/4/0	10/10/10	72/34/22	100/86/70
1JPS	6	1491	12510	1	44	316	98	9	1/0/0	10/0/0	10/0/0	54/1/0
1K4C	22	1475	1084	18	3	14	29	4	0/0/0	0/4/0	5/0/0	10/24/15
1K5D	288	458	5454	20	2	299	265	18	0/0/0	0/3/0	0/0/0	21/19/0
1KAC	7	545	3906	1	7	19	238	19	1/0/0	7/1/0	6/0/0	24/10/2
1KKL	1	1096	1	5	101	9	235	64	6/0/5	4/0/1	26/0/15	15/0/10
1KLU	1074	6960	9872	90	229	664	2040	95	0/0/0	0/0/0	0/0/0	1/0/0
1KTZ	588	16699	3873	67	494	49	415	22	0/0/0	0/0/0	0/0/0	3/0/1
1KXP	3	33	1	1	1	1	133	100	7/0/6	4/2/2	74/8/24	6/4/7
1KXQ	3	13	191	1	1	113	240	25	3/0/0	8/9/0	31/6/0	37/51/0
1MAH	3	26	11	1	1	3	57	4	3/0/0	6/8/8	28/3/6	62/71/49
1ML0	1	40	9	1	3	1	71	10	7/0/1	10/1/10	42/2/5	49/15/79

Case	Bestrank			Bestrank			Random	Random	Top10	Topf10	Top100	Topf100
	P	TB	LLS	FP	F+TB	F+LLS	no filter	filtered	P/TB/LLS	P/TB/LLS	P/TB/LLS	P/TB/LLS
1MLC	8	31	3458	1	1	54	65	4	1/0/0	8/9/0	7/1/0	48/72/3
1N2C	103	1092	348	16	59	48	729	208	0/0/0	0/0/0	0/0/0	7/3/3
1NCA	923	40	22862	54	2	591	304	20	0/0/0	0/7/0	0/5/0	2/34/0
1NSN	8	441	21857	1	1	837	269	27	2/0/0	8/8/0	13/0/0	28/26/0
1PPE	1	2	1	1	1	1	16	4	10/1/9	10/9/10	83/11/61	62/68/97
1QA9	807	4636	51767	118	64	2418	1276	66	0/0/0	0/0/0	0/0/0	0/1/0
1QFW	430	1298	21667	29	8	993	349	21	0/0/0	0/1/0	0/0/0	18/8/0
1RLB	4	1	2	1	1	1	29	9	3/6/1	10/8/10	29/35/19	61/57/80
1SBB	9740	9832	28003	848	259	1152	1421	90	0/0/0	0/0/0	0/0/0	0/0/0
1TMQ	1	59	145	1	1	1	106	10	8/0/0	10/4/9	39/4/0	54/24/29
1UDI	1	1	19	1	1	2	123	21	6/5/0	7/6/3	20/10/6	19/20/20
1VFB	44	137	627	63	1	2	115	9	0/0/0	0/2/1	2/0/0	1/12/4
1WEJ	1	69	39	1	1	1	53	5	9/0/0	10/10/5	63/3/1	98/91/33
1WQ1	105	670	7363	29	279	1363	269	324	0/0/0	0/0/0	0/0/0	3/0/0
2BTF	2	68	880	1	1	1	136	10	7/0/0	10/7/3	49/4/0	68/61/21
2HMI	181	20	663	1	1	1	113	9	0/0/0	4/10/7	0/3/0	39/59/29
2JEL	1	1	219	1	1	1	37	6	10/6/0	10/10/10	87/31/0	74/95/33
2MTA	25	42	20	1	3	1	68	12	0/0/0	5/1/10	6/3/5	34/12/66
2PCC	4	8	1351	1	1	3	136	9	3/1/0	4/4/3	29/2/0	41/56/19
2QFW	9	11	448	1	1	4	74	10	1/0/0	10/10/6	18/10/0	84/72/39
2SIC	1	24	7	1	1	1	49	7	9/0/2	10/8/9	95/5/8	100/49/51
2SNI	1	24	15	1	2	1	72	6	8/0/0	10/6/4	48/12/7	87/50/14
2VIS	889	66	70	146	2	8	54	13	0/0/0	0/1/1	0/1/1	0/25/6
7CEI	1	1	143	1	1	6	39	5	9/4/0	10/9/1	59/21/0	79/82/3

Table 8

Result of our filtering and scoring for different categories of complexes.

Category	Rank1	Top10	Top100	With Hits
Enzyme	19	20	22	23
Antibody	13	15	20	21
Other	11	16	26	35

Table 9

due based scoring function (PIE640)

	ILE	VAL	LEU	PHE	CYS	MET	ALA	GLY	THR	SER	TRP	TYR	PRO	HIS	ASN	GLN	ASP	GLU	LYS	ARG
C	0.492	0.538	0.609	0.577	0.663	0.415	0.111	0.175	0.25	−0.02	−0.2	0.556	−0.004	−0.053	−0.009	0.389	0.026	−0.066	−0.119	0.208
L	0.538	0.178	0.489	0.39	0.056	−0.029	−0.06	0.101	0.106	0.035	0.18	0.444	−0.012	0.299	−0.149	−0.125	0.128	0.021	−0.123	0.146
U	0.609	0.489	0.528	0.572	0.136	0.246	0.117	0.137	0.172	0.087	0.694	0.337	0.126	0.147	−0.086	−0.002	−0.072	−0.107	−0.034	−0.021
E	0.577	0.39	0.572	0.671	0.061	0.464	0.177	0.193	0.305	0.297	0.413	0.534	0.314	0.47	0.137	0.357	0.095	0.007	−0.062	0.205
S	0.663	0.056	0.136	0.061	0.707	0.39	0.011	0.225	0.157	−0.04	0.019	0.123	0.124	0.452	−0.481	−0.253	−0.231	0.085	0.249	0.274
T	0.415	−0.029	0.246	0.464	0.39	−0.019	0.186	0.301	0.123	0.082	0.513	0.342	0	0.199	0.084	0.192	0.183	0.371	−0.185	0.315
A	0.111	−0.055	0.117	0.177	0.011	0.186	0.035	0.081	−0.01	0.088	0.302	0.151	−0.071	−0.159	−0.027	−0.046	−0.223	−0.084	0.038	0
Y	0.175	0.091	0.137	0.193	0.225	0.301	0.081	0.047	−0.09	0.05	0.166	0.105	0.018	−0.16	0.115	0.013	0.03	−0.18	0.057	0.024
R	0.25	0.196	0.172	0.305	0.157	0.123	−0.01	−0.087	0.166	0.029	0.195	0.073	−0.059	0.189	−0.042	0.088	−0.035	−0.003	−0.073	0.132
R	−0.016	0.035	0.087	0.297	−0.04	0.082	0.088	0.05	0.029	0.141	0.026	0.028	−0.013	0.287	−0.072	−0.076	−0.073	0.017	−0.131	−0.075
P	−0.2	0.18	0.694	0.413	0.019	0.513	0.302	0.166	0.195	0.026	0.839	0.262	0.539	0.238	0.559	0.202	−0.18	−0.018	0.169	0.234
R	0.556	0.414	0.337	0.534	0.123	0.342	0.151	0.105	0.073	0.028	0.262	0.475	0.333	0.369	0.186	0.134	0.089	0.099	0.354	0.444
O	−0.004	−0.052	0.126	0.314	0.124	0	−0.07	0.018	−0.06	−0.01	0.539	0.333	0.01	−0.095	−0.035	−0.042	−0.084	−0.008	−0.163	−0.12
S	−0.053	0.289	0.147	0.47	0.452	0.199	−0.16	−0.16	0.189	0.287	0.238	0.369	−0.095	0.021	0.114	−0.14	0.2	−0.055	−0.104	0.014
N	−0.009	−0.149	−0.086	0.137	−0.48	0.084	−0.03	0.115	−0.04	−0.07	0.559	0.186	−0.035	0.114	−0.008	0.056	−0.131	0.089	−0.002	0.177
N	0.389	−0.125	−0.002	0.357	−0.25	0.192	−0.05	0.013	0.088	−0.08	0.202	0.134	−0.042	−0.14	0.056	0.059	−0.017	−0.248	−0.112	0.05
P	0.026	0.128	−0.072	0.095	−0.23	0.183	−0.22	0.03	−0.04	−0.07	−0.18	0.089	−0.084	0.2	−0.131	−0.017	−0.188	−0.239	0.111	0.413
U	−0.066	0.021	−0.107	0.007	0.085	0.371	−0.08	−0.18	−0	0.017	−0.02	0.099	−0.008	−0.055	0.089	−0.248	−0.239	−0.214	0.198	0.218
S	−0.119	−0.123	−0.034	−0.062	0.249	−0.185	0.038	0.057	−0.07	−0.13	0.169	0.354	−0.163	−0.104	−0.002	−0.112	0.111	0.198	−0.228	−0.3
G	0.208	0.146	−0.021	0.205	0.274	0.315	0	0.024	0.132	−0.08	0.234	0.444	−0.12	0.014	0.177	0.05	0.413	0.218	−0.3	−0.098
OVERLAP		−0.002																		

Table 10

mer scoring function (PIEHetero)

ILE	VAL	LEU	PHE	CYS	MET	ALA	GLY	THR	SER	TRP	TYR	PRO	HIS	ASN	GLN	ASP	GLU	LYS	ARG
0.25	0.515	0.507	0.641	0.692	0.484	0.113	0.106	0.226	0.017	−0.305	0.572	−0.007	−0.154	−0.011	0.459	0.043	−0.082	−0.196	0.234
0.515	0.036	0.552	0.376	0.11	−0.057	−0.082	0.063	0.089	0.07	0.158	0.462	0.071	0.285	−0.221	−0.107	0.13	−0.002	−0.033	0.2
0.507	0.552	0.417	0.52	0.158	0.244	0.082	0.125	0.166	0.07	0.623	0.366	0.128	−0.016	−0.107	0.008	0.014	−0.109	0.027	0.001
0.641	0.376	0.52	0.586	0.005	0.482	0.16	0.227	0.337	0.272	0.479	0.522	0.217	0.46	0.17	0.286	0.114	0.022	−0.091	0.209
0.692	0.11	0.158	0.005	0.778	0.392	0.007	0.214	0.194	−0.025	0.051	0.183	0.169	0.522	−0.471	−0.247	−0.239	0.063	0.374	0.31
0.484	−0.057	0.244	0.482	0.392	0.04	0.194	0.364	0.1	0.022	0.182	0.404	0.009	0.315	0.119	0.221	0.201	0.428	−0.302	0.171
0.113	−0.082	0.082	0.16	0.007	0.194	−0.089	0.114	0.012	0.08	0.394	0.136	−0.025	−0.077	−0.049	−0.083	−0.24	−0.053	0.004	0.015
0.106	0.063	0.125	0.227	0.214	0.364	0.114	0.047	−0.093	0.066	0.069	0.086	0.056	−0.128	0.145	0.037	0.03	−0.189	0.047	0.03
0.226	0.089	0.166	0.337	0.194	0.1	0.012	−0.093	0.194	−0.006	0.167	0.157	−0.111	0.16	−0.059	0.141	−0.094	0.043	−0.069	0.145
0.017	0.07	0.07	0.272	−0.025	0.022	0.08	0.066	−0.006	0.121	0.103	0.032	0.024	0.313	−0.186	−0.08	−0.054	0.023	−0.068	−0.095
−0.31	0.158	0.623	0.479	0.051	0.182	0.394	0.069	0.167	0.103	1.019	0.351	0.593	0.446	0.606	0.205	−0.417	−0.018	0.004	0.34
0.572	0.462	0.366	0.522	0.183	0.404	0.136	0.086	0.157	0.032	0.351	0.549	0.329	0.412	0.176	0.118	0.065	0.11	0.383	0.498
−0.01	0.071	0.128	0.217	0.169	0.009	−0.025	0.056	−0.111	0.024	0.593	0.329	0.017	−0.111	−0.012	−0.065	−0.18	−0.037	−0.228	−0.154
−0.15	0.285	−0.016	0.46	0.522	0.315	−0.077	−0.128	0.16	0.313	0.446	0.412	−0.111	−0.291	0.126	−0.205	0.15	−0.051	−0.21	−0.018
−0.01	−0.221	−0.107	0.17	−0.471	0.119	−0.049	0.145	−0.059	−0.186	0.606	0.176	−0.012	0.126	−0.016	0.083	−0.097	0.113	−0.014	0.198
0.459	−0.107	0.008	0.286	−0.247	0.221	−0.083	0.037	0.141	−0.08	0.205	0.118	−0.065	−0.205	0.083	−0.064	−0.04	−0.258	−0.293	0.103
0.043	0.13	0.014	0.114	−0.239	0.201	−0.24	0.03	−0.094	−0.054	−0.417	0.065	−0.18	0.15	−0.097	−0.04	−0.205	−0.228	0.158	0.439
−0.08	−0.002	−0.109	0.022	0.063	0.428	−0.053	−0.189	0.043	0.023	−0.018	0.11	−0.037	−0.051	0.113	−0.258	−0.228	−0.224	0.184	0.244
−0.2	−0.033	0.027	−0.091	0.374	−0.302	0.004	0.047	−0.069	−0.068	0.004	0.383	−0.228	−0.21	−0.014	−0.293	0.158	0.184	−0.332	−0.345
0.234	0.2	0.001	0.209	0.31	0.171	0.015	0.03	0.145	−0.095	0.34	0.498	−0.154	−0.018	0.198	0.103	0.439	0.244	−0.345	−0.131
ERLAP	−0.002																		

Table 11

Performance of scoring function upon variation of features (we rank the structures that pass the overlap area filter, there were 76 cases that had at-least one hit in the 2500 structures passing the filter) – sa(dssp) is the change in surface area computed using DSSP, sa(excluded) is the changes in surface area accounting for regions not accessible to a probe of radius 1.4Å; contacts, atomcontacts, atomcontacts(20) indicate number of contacts between residues, atoms - grouped into 18 types and atoms - grouped into 20 types respectively; 3body is the number of contacts formed by 3 residues simultaneously (residues grouped into 5 types); evolutionary_profile is the number of contacts between residues of different levels of conservation (see text for details).

Variation	#features	Top1	Top10	Top100
overlapsa	1	17	48	66
contacts	210	36	49	62
overlaparea + contacts	211	43	51	68
contacts + 3body	245	37	51	64
overlaparea + contacts + 3body	246	43	53	68
contacts + evolutionary_profile	216	36	49	64
atomcontacts	171	40	51	64
atomcontacts + overlaparea(by atom type)	189	45	58	69
atomcontacts(20)	210	40	52	64
atomcontacts(20)+overlaparea(by atom type)	230	46	58	67
sa(excluded)+contacts	230	39	50	65
sa(dssp)+contacts	230	36	49	64

Table 12

with 20 atom types

C^{α}	C	O	GC^{α}	C^{β}	KN^{η}	KC^{δ}	DO^{δ}	RN^{η}	NN^{δ}	RN^{ε}	SO^{γ}	HN^{ε}	YC^{ζ}	FC^{ζ}	LC^{δ}	CS^{γ}	CC^{β}	MS^{δ}
0.0038	−0.01	−0.006	0.0008	−0.0008	−0.0008	−0.0076	−0.0016	0.0002	−0.002	0.0073	−0.0015	−0.004	0.0085	−0.0027	−0.0035	0.0191	−0.025	−0.012
0.011	0.015	−0.004	−0.009	0.021	0.002	−0.007	0.013	0.002	−0.014	0.006	−0.014	−0.007	0.012	0.019	0.006	−0.068	−0.117	0.049
−0.008	0.004	−0.025	0.036	−0.012	−0.01	−0.021	−0.028	0.028	0.005	−0.014	−0.005	0.041	−0.027	−0.002	−0.021	0.199	−0.064	−0.048
0.004	0.029	−0.015	−0.034	0	0.025	0.046	0.014	−0.037	−0.016	−0.015	−0.007	−0.025	−0.006	−0.01	0.006	0.058	−0.04	0.041
−0.025	−0.015	0.005	−0.005	−0.021	−0.006	0.001	−0.04	0.023	0.019	0.032	0.023	0.008	0.005	0.013	0.022	0.12	−0.034	−0.053
0.036	−0.034	−0.005	0.128	0.024	0.017	0.03	−0.009	0.052	0.017	−0.058	0.044	−0.017	0.046	0.006	0.036	0.208	−0.075	0.356
−0.012	0	−0.021	0.024	0.007	−0.011	−0.045	−0.009	0.003	−0.013	−0.018	0.003	−0.007	0.038	0.017	0.012	0.102	0.029	0
−0.01	0.025	−0.006	0.017	−0.011	−0.022	−0.01	0.021	0.011	0.012	0.004	−0.02	−0.004	0.021	0.003	0.004	0.159	−0.067	0.02
−0.021	0.046	0.001	0.03	−0.045	−0.01	−0.101	0.1	−0.043	−0.033	−0.129	0.02	−0.075	0.094	−0.025	−0.09	0.305	−0.328	−0.135
−0.028	0.014	−0.04	−0.009	−0.009	0.021	0.1	−0.012	0.053	0.004	0.024	0.006	0.018	0.001	−0.007	0.002	0.03	−0.077	0.084
0.028	−0.037	0.023	0.052	0.003	0.011	−0.043	0.053	0.025	−0.014	−0.062	0.015	−0.038	0.021	0.003	0.015	0.086	−0.069	0.02
0.005	−0.016	0.019	0.017	−0.013	0.012	−0.033	0.004	−0.014	0.021	0.042	−0.001	0.007	0.005	0.005	0	−0.145	−0.103	0.08
−0.014	−0.015	0.032	−0.058	−0.018	0.004	−0.129	0.024	−0.062	0.042	0.017	−0.047	0.033	−0.007	−0.013	−0.046	−0.022	−0.002	0.17
−0.005	−0.007	0.023	0.044	0.003	−0.02	0.02	0.006	0.015	−0.001	−0.047	0.098	0.028	−0.043	0.006	0.032	0.022	−0.03	0.157
0.041	−0.025	0.008	−0.017	−0.007	−0.004	−0.075	0.018	−0.038	0.007	0.033	0.028	0.007	−0.013	0.008	0.019	0.007	0.043	0.055
−0.027	−0.006	0.005	0.046	0.038	0.021	0.094	0.001	0.021	0.005	−0.007	−0.043	−0.013	0.049	0.015	−0.015	−0.051	−0.003	−0.101
−0.002	−0.01	0.013	0.006	0.017	0.003	−0.025	−0.007	0.003	0.005	−0.013	0.006	0.008	0.015	0.018	0.037	0.045	0.026	0.013
−0.021	0.006	0.022	0.036	0.012	0.004	−0.09	0.002	0.015	0	−0.046	0.032	0.019	−0.015	0.037	0.16	0.037	0.05	0.007
0.199	0.058	0.12	0.208	0.102	0.159	0.305	0.03	0.086	−0.145	−0.022	0.022	0.007	−0.051	0.045	0.037	0.85	0.198	0.003
−0.064	−0.04	−0.034	−0.075	0.029	−0.067	−0.328	−0.077	−0.069	−0.103	−0.002	−0.03	0.043	−0.003	0.026	0.05	0.198	0.007	0.312
−0.048	0.041	−0.053	0.356	0	0.02	−0.135	0.084	0.02	0.08	0.17	0.157	0.055	−0.101	0.013	0.007	0.003	0.312	−0.306