Published in final edited form as:
Proteins. 2012 August ; 80(9): 2207–2217. doi:10.1002/prot.24110.

Scoring Protein Interaction Decoys using Exposed Residues (SPIDER): A Novel Multi-Body Interaction Scoring Function based on Frequent Geometric Patterns of Interfacial Residues

Raed Khashan[1], **Weifan Zheng**[1,2], and **Alexander Tropsha**[1,*]

[1]Laboratory for Molecular Modeling, the UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599.

[2]BRITE Institute & Department of Pharmaceutical Sciences, North Carolina Central University, Durham, NC 27707.

Abstract

Accurate prediction of the structure of protein-protein complexes in computational docking experiments remains a formidable challenge. It has been recognized that identifying native or native-like poses among multiple decoys is the major bottleneck of the current scoring functions used in docking. We have developed a novel multi-body pose-scoring function that has no theoretical limit on the number of residues contributing to the individual interaction terms. We use a coarse-grain representation of a protein-protein complex where each residue is represented by its side chain centroid. We apply a computational geometry approach called Almost-Delaunay tessellation that transforms protein-protein complexes into a residue contact network, or an un-directional graph where vertex-residues are nodes connected by edges. This treatment forms a family of interfacial graphs representing a dataset of protein-protein complexes. We then employ frequent subgraph mining approach to identify common interfacial residue patterns that appear in at least a subset of native protein-protein interfaces. The geometrical parameters and frequency of occurrence of each “native” pattern in the training set are used to develop the new SPIDER scoring function. SPIDER was validated using standard “ZDOCK” benchmark dataset that was not used in the development of SPIDER. We demonstrate that SPIDER scoring function ranks native and native-like poses above geometrical decoys and that it exceeds in performance a popular ZRANK scoring function. SPIDER was ranked among the top scoring functions in a recent round of CAPRI (Critical Assessment of PRedicted Interactions) blind test of protein–protein docking methods.

Keywords

Bioinformatics; Amino acids; Centroids; Statistical potential; Delaunay tessellation; Subgraph mining; Motifs; Coarse-grained; ZDOCK; CAPRI

Introduction

Protein–protein interactions are of central importance for virtually every process in a living cell. Many of the most important molecular processes in the cell such as DNA replication

[*]Corresponding author: Prof. Alexander Tropsha 327 Beard Hall, CB# 7568 The UNC Eshelman School of Pharmacy, Chapel Hill, NC 27599. Telephone: (919) 966-2955 FAX: (919) 966-0204 alex_tropsha@unc.edu.

Availability: The program to calculate the SPIDER scores is available free of charge and it can be obtained by contacting the corresponding author.

are carried out by large ensembles of structurally and functionally interacting proteins. Thus, the characterization of these interactions at the structural level improves our understanding of diseases and can provide the basis for the discovery of new drugs targeting protein-protein interactions. However, it is still challenging to solve the structures of protein complexes, which constitute only a small fraction of experimentally determined structures in the Protein Data Bank (PDB).[1] Therefore, the development of computational docking methods capable of accurate prediction of the structures of protein-protein complexes from the structures of the interacting individual proteins is of substantial interest.[2]

Computational docking usually entails two steps: first, an initial sampling of the configurational space of the interacting proteins to generate docking poses; second, pose scoring to select the putative native (or native like) protein-protein complexes. This two-step approach has been used by many research groups, and has proven efficient in docking small-molecules to their target proteins.[3] Sampling (the first step), is typically performed by using rigid-body strategies, which have significantly improved with the use of Fast Fourier Transform (FFT) algorithms.[4-8] In fact, as a result of using efficient sampling strategies, the library of poses generated by multiple docking experiments may include native-like or near-native poses. However, many scoring functions (second step) may often fail to recognize those poses because geometrical decoys often score better or even much better than native or near-native poses.[3] Comparative assessment of the accuracy of current protein-protein docking approaches suggests that there is a substantial room for improvement in scoring functions.[9, 10]

Ideally, scoring functions should provide an accurate description of the binding free energy; in particular, those functions that are based on physicochemical principles and atom level structure representation are expected to have high accuracy in discriminating the native pose from decoys. However, calculations relying on such scoring functions are usually computationally expensive and prone to error when dealing with the expected inaccuracies in the inter-protein contacts caused by the rigid-body docking approach. As an alternative, many researchers have relied on knowledge-based statistical scoring functions derived at either the residue level or the atomic level from the pair-wise propensities of intermolecular interactions at the protein-protein interfaces. [11-16] Residue-level statistical potentials are less sensitive to details of atomic arrangements, thus providing an efficient approach for scoring rigid-body docking poses at very low computational cost. However, most of the knowledge-based statistical potentials employ two-body (pair-wise), or in some cases, higher-order terms (three and four-body) to represent patterns of interacting residues, whereas realistically, it is multi-residue interactions that are responsible for the formation of stable protein complexes.[17-23] Therefore, it may be advantageous to develop a scoring function that captures multi-body interaction terms by design to improve the outcome of scoring experiments.

Herein, we introduce a novel knowledge-based statistical scoring function that captures the propensities of multiple residues to form the interface of protein-protein complexes. This scoring function exploits the constructs from computational geometry to define a network of interacting residues at the protein-protein interfaces and uses graph-mining techniques to identify frequent patterns of interacting residues at protein-protein interfaces. Thus, by design this new approach places no theoretical limit on the number of residues (equivalent to the number of terms in multi-body statistical scoring functions) forming frequent interfacial interaction patterns. We demonstrate that this novel scoring function termed SPIDER (Scoring Protein Interaction Decoys using Exposed Residues) succeeds in identifying native and native like poses within large libraries of protein complexes generated by computational docking. We also show that SPIDER achieves significantly higher enrichment of geometrically native-like poses among top scoring hits when compared to a widely used

ZRANK scoring function[24]. SPIDER was ranked among the top 6 (out of 28) scoring functions in a recent round 21 of CAPRI (Critical Assessment of PRedicted Interactions) blind test of protein–protein docking methods.[25]

Materials and Methods

We have developed a simple multi-body knowledge-based scoring function that captures geometric and compositional residues interaction patterns formed at the interfaces of X-ray characterized protein complexes. This function is termed Scoring Protein Interaction Decoys using Exposed Residues (SPIDER). The simplified workflow diagram for the development, implementation, and validation of the SPIDER method is shown in Figure 1, and the individual components of the workflow are discussed below.

Preparing the internal training and external testing datasets

We have selected two unrelated datasets from two different research groups to train and test the SPIDER score, respectively. We used the Dockground[26] dataset of 508 protein complexes for internal training, and the ZDOCK benchmark[27] composed of 124 protein complexes for external testing to validate the scoring function. To avoid any biases in external testing, the Dockground training set was curated to eliminate both complexes that were found in the testing set as well as complexes of proteins that had greater than 70% sequence similarity to those in the testing set. Sequence similarity was calculated via “blastclust”, the program used in the PDB. According to “blastclust”, all binding partners (protein chains) are involved in calculating the sequence similarity. Therefore, the resulting training set included only 241 out of initial 508 complexes. The external ZDOCK testing benchmark dataset included both native structures as well as 54,000 computationally generated decoys for each protein complex. In addition, the information provided with ZDOCK includes predictions calculated using their internal scoring function, which allows us to test and compare ZDOCK results vs. those generated with our new SPIDER scoring function.

Graph Representation of the Protein-Protein Interfaces and Application of Frequent Subgraph Mining Technique

The initial step in SPIDER is elucidating the frequent patterns of multiple interacting residues at the protein-protein interface. Thus, using the internal training dataset defined as described above, we use the following steps to find the frequent interfacial residue patterns (see Figure 2 illustrating major steps):

1. For each X-ray characterized native protein complex in the training set, the interfacial residues (represented by side chain centroids) are defined as those that have at least one residue from the opposite protein chain within a user defined threshold of 10.0 Å. We then employ Almost-Delaunay Tessellation developed by Snoeyink and co-workers[28] and used in our previous studies of single-chain proteins.[19] Almost-Delaunay is a modification of Delaunay tessellation, which is a standard computational geometry technique to analyze the geometry of point objects described in two-, three-, or higher dimensional spaces. When applied to a set of randomly distributed points in 2D space, Delaunay tessellation generates an aggregate of space-filling, non-overlapping, irregular triangles (Delaunay simplices) with the original points appearing as vertices; the same approach applied to points in 3D space, e.g., side chain centroids of amino acid residues, yields Delaunay tetrahedral. Almost Delaunay was developed to incorporate the imprecision of the point coordinates in defining the tessellation pattern; thus, Almost Delaunay features a special parameter, ε, that reflects the coordinate error. Almost Delaunay always generates a larger number of Delaunay simplices than

Delaunay tessellation; previous studies suggested that the use of ε = 0.5 affords the best balance between the number of Delaunay simplices resulting from tessellation and the computational efficiency and accuracy of the SNAPP[29] scoring function developed by us previously to evaluate single chain protein conformations in fold recognition experiments.

2. For each protein complex, the interface is represented by labeled graph where nodes are residue side chain centroids and edges connect these centroids, thus, forming an interaction residue network at the interface of each protein complex that can also be treated as a labeled un-directional graph (see Figure 2(B)).

3. After defining interfacial graphs for all protein complexes in the training set, we use efficient fast frequent subgraph mining (FFSM[30, 31]) technique to find frequent common subgraphs that occur in at least a certain fraction (called support value) of the interfaces formed by native complexes (see Figure 2(C) and 2(D)). Furthermore, we eliminate those subgraphs that are found only as part of bigger (parent) subgraphs; the remaining subgraphs are known as "non-coherent", or "closed" subgraphs. This step is crucial since it allows to avoid overlapping subgraphs, and thus, prevent overweighting and having false positives when scoring decoys.

We used a support value of ~5% with FFSM to identify subgraphs that occurred in no less than 5% of the training set proteins; i.e., in no less than 10 protein complexes of the 241 complexes in the training set. This threshold was chosen after some experimentation, which included changing the support value within the range of 1% and 10% and looking at the number of common subgraphs corresponding to the support value (Figure 3). We found that an arbitrary value of 5% resulted in identifying ca. 25,025 "native" subgraphs.

These subgraphs represent the multi-residue interfacial patterns that appear frequently in our internal training dataset. An example of a frequent interfacial residue pattern corresponding to a subgraph mined by FFSM that appears in 10 protein complexes is shown in Figures 2(C) and 2(D). The pattern contains 7 interfacial residues represented as nodes, and 6 edges representing a connection (i.e., interaction) between these nodes. Notice that although this pattern appears in 10 protein complexes, it does not necessarily have the same geometry in all of these 10 complexes. Figure 2(C) and 2(D) present examples of two different geometries for the same pattern.

Deriving the Scoring Function Using Frequent Residue Interaction Patterns

The frequent interaction patterns act as structural motifs characteristic of protein-protein interfaces. As mentioned above we called such patterns "native" since they occur at the interfaces of X-ray characterized protein complexes. For each of these patterns we have stored both the Cartesian coordinates of the composing nodes and its geometric frequency of occurrence in the internal training set of protein complexes. These patterns were used in the scoring function as described below. Figure 2(C) gives example of one of the frequent patterns of interacting residues at the protein complexes interface.

Given a test set of protein complexes, the first step in calculating the SPIDER score for a given pose of a protein complex is identifying the interface in that complex. As discussed above, we define the interface as a network of interactions (derived with the help of Almost Delaunay Tessellation) formed by the interacting interfacial residues that are separated by no more than 10.0 Å. Then, we look in that interface for the presence of "native" subgraph patterns, which were found in the complexes in the training dataset. All such patterns are then used in scoring the protein complex as we shall discuss later. But at this stage, we only look for matching subgraphs, not taking into account the patterns' geometry.

To convert the frequency of observed interactions into the scoring function terms, we assume that the higher is the number of the frequent “native” patterns found at the interface of a protein complex and the more frequent these native patterns are, the higher should be the score for this pose. We have also realized that the better is the geometrical fit between the pattern of interaction in a pose and the matching “native” pattern (i.e. the smaller the Root Mean Square Deviation (RMSD), between matching patterns corresponding to matching subgraphs), the higher the score should be as well. [Note: to avoid confusing the RMSD of matching patterns with the RMSD of generated poses with respect to native pose, we will use the terms RMSD^{Pattern} and RMSD^{Pose}, respectively]. Also, we have assumed that the score should be influenced by the size of the frequent pattern identified for the pose, i.e., the score should be higher for bigger patterns. In addition, the larger the number and the fraction of interfacial residues covered with “native” patterns, the higher the score is. Finally, the higher the number of “native” patterns used averaged over the number of covered interfacial residues, the higher the score is. Taking all these considerations into account, we have derived the following formula to score a pose:

$$\text{Score}=\left\|\sum_{\mathrm{i}}^{\mathrm{N}}\sum_{\mathrm{j}}^{\mathrm{M}}|P_i|/RMS\,D_{ij}^{Pattern}\right\|+\|X_1\|+\|X_2\|+\|X_3\|+\|X_4| \qquad (1)$$

where N is the total number of frequent (”native”) patterns found at the interface, M is the frequency of the pattern *i* in the training set, and therefore is the number of modes of interaction (number of different internal geometric coordinate sets) for that pattern, $|P_i|$ is the size of the pattern P_i (i.e., total number of residues in the pattern), and RMSD $^{Pattern}{}_{ij}$ is calculated for the best fit between pattern P_i in the test complex and the matching “native” pattern. The first summation is over all patterns that are found at the interface. The second summation reflects the frequency of each pattern and the different modes of interaction for each pattern. Also, to avoid dividing by zero, an epsilon value of $1*10^{-60}$ is added to the RMSD^{Pattern}. (This value is chosen based on the smallest empirical RMSD^{Pattern} value that was found in our studies.) An RMSD^{Pattern}cutoff value of 0.1, 0.5, and 1.0 Å are used to decide if the pattern should be included in the scoring function or not. This cutoff value defines the applicability domain of our knowledge-based scoring function as we will explain later.

Other parameters used: X_1 is the number of interfacial residues that match the native patterns. X_2 is the fraction of interfacial residues that match the native patterns. X_3 is the number of native patterns found at the interface. Finally, X_4 is the number of native patterns found at the interface divided by the number of interfacial residues that match residues in the native patterns; i.e., the average number of patterns matched per one interfacial residue.

Validation of the Scoring Function

In order to test the ability of the SPIDER pose scoring function to accurately identify the native pose (as determined by *X*-ray) among those deviating from the native structure (i.e., generated computationally), we have developed a study design (Figure 1) based on the following considerations. Ideally, a good scoring function should be able to rank the poses closest-to-native on top of all non-native poses. Thus, we use the following sensible methods to validate such ability: First, we look at the rank of the nearest-native pose among all poses for each protein complex. We obtain the average rank of all complexes using our scoring function and, as a matter of benchmarking, compare it to that of the standard ZRANK scoring function. In addition, we look at the rank of the native pose among all poses for each protein complex and obtain the average rank of the native poses across all protein complexes for comparison purposes as well. Second, following the ZDOCK benchmarking approach, we count the number of poses ranked by SPIDER within certain fraction of all 54,000 decoys that also have less than 2.5 Å RMSD of the native pose. This cutoff value of 2.5 Å has been used in most studies to define a “native-like” pose, or a “hit”. Therefore, given the values $\mathrm{RMSD}^{\mathrm{Pose}}$ for the decoys of each protein complex in our external training set, we can find the number of hits among the 54,000 decoys for each protein complex. Thus, for validation of our scoring function and comparison with the standard ZRANK scoring function, the number of hits found in top 500, 2000, 5000, … etc., selected by each scoring function among the 54,000 poses is reported in a receiver operating characteristic (ROC) curve for each scoring function. The curve will give us an idea about each function’s ability to distinguish hits.

A third comparative validation approach is to obtain the ranks of all the hits in each protein complex dataset. We report the rank of the first hit identified (along with its $\mathrm{RMSD}^{\mathrm{Pose}}$), the last hit, and the average of the ranks of all hits for each scoring function (using the aforementioned definition of a hit). Finally, we look at the correlation coefficient between the $\mathrm{RMSD}^{\mathrm{Pose}}$ of the poses and their score. A good scoring function should be able to correlate the pose $\mathrm{RMSD}^{\mathrm{Pose}}$ with its rank. Herein, in addition to calculating this correlation for all 54,000 decoy poses, we also calculate the correlation coefficient for those poses that have $\mathrm{RMSD}^{\mathrm{Pose}}$ below 5 Å. In summary, we suggest that the four independent validation methods should provide sufficient means for the unbiased evaluation of the relative performance of SPIDER vs. the popular ZRANK scoring function.

Results and Discussion

The uniqueness of our approach is that we use multi-body statistical scoring function expecting that the use of higher-order terms should improve the accuracy of scoring. This is possible because capturing multiple interacting residues in an interfacial subgraph pattern affords the discovery of multi-residue motifs (or patterns of interactions) that are hard if not impossible to identify with alternative approaches that place an upper limit on the number of nodes involved in interfacial interaction (e.g., pair-wise scoring functions). It is also feasible that placing a limit on the number of residues involved in the development of knowledge-based statistical scoring functions can lead to patterns (motifs) that only appear as part of larger patterns, leading to noise (false positives) when scoring candidate poses. Furthermore, the assumption that pair-wise interactions are additive and their linear summation yields the accurate total multi-body score has been clearly demonstrated (experimentally and in simulation) to be incorrect; in fact, these interactions can be cooperative (i.e., more favorable than the independent pair-wise interactions), or anti-cooperative. [32-35] Thus, there is a demand for a multi-body interaction scoring functions that model these non-additive dependent interactions implicitly as a whole rather than treating them as a summation of independent lower order interactions. Consequently, a more accurate scoring can be achieved using higher-order multi-body scoring function, which is the objective of this study.

Defining Applicability Domain

The applicability domain of a model is predicated on the inherently limited diversity of the physico-chemical, structural or any feasible feature space (e.g., types of patterns occurring at the interface of protein-protein complexes) characterizing the training set used for model development. In theory, to avoid unjustified extrapolation, the prediction for new objects (e.g., new protein-protein complexes) can be obtained reliably if they are located in the same feature space as the training set objects. This issue is well known in the field of Quantitative Structure Activity Relationship modeling.[36, 37]

We address the issue of the applicability domain for the new scoring function by analyzing the terms involved in the SPIDER scoring function as defined in Eq. 1. We need to make sure that the protein complex we are predicting does have certain similarity to the dataset that was used in training the scoring function; i.e., there exists geometric similarity between the native patterns and the interaction patterns in our target protein complex. The similarity threshold would represent the domain of applicability of the scoring function, and it can be defined using certain cutoff values for geometric similarity as explained below.

Defining the applicability domain provides us with the confidence concerning the rank-scoring of poses generated for a protein complex. For example, if we are facing a protein complex that contains patterns of interactions that are not geometrically similar to the "native" patterns, then, we should be alarmed that our confidence in this case is low. Geometric similarity is measured by $\mathrm{RMSD}^{\mathrm{Pattern}}$, and thus, we have defined cutoff values for the $\mathrm{RMSD}^{\mathrm{Pattern}}$ below. We use an enrichment curve to address the effect of this parameter on the ability of the scoring function to rank hits. As we will explain later, we use different cutoff values for $\mathrm{RMSD}^{\mathrm{Pattern}}$; notice that the smaller the cutoff values are, the more accurate are our predictions but at the expense of the number of complexes we can predict that are within the applicability domain. On the other hand, using higher cutoff values, we are able to predict more complexes (but with lower accuracy).

Identifying the native pose and the nearest-native pose

Our goal is to identify the native pose as well as the nearest-native pose among all computationally generated poses for each protein complex. The rank score of these two poses in each protein complex as evaluated by the scoring function is recorded, and then the average of the ranks is calculated and used to compare between the SPIDER score and the standard ZRANK score. Figure 4 shows the results for each scoring function using an enrichment curve. The curve shows the effect of $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoffs on the ability of SPIDER score to rank native and nearest-native poses; it also shows the effect of these cutoffs on the percentage of cases that can be predicted using the SPIDER score.

The results show that in the worst case scenario, using the SPIDER multi-body interaction scoring function, the nearest-native pose can be identified for 100% of the cases in top 12,500 ranked poses compared to 20,000 ranked poses using the standard ZRANK score. On the other hand, when using lower cutoff values for $\mathrm{RMSD}^{\mathrm{Pattern}}$, we limit the number of complexes that are scored but also observe a much higher rank order of the nearest native pose. For instance, with the applicability domain threshold set very conservatively at 0.5 Å, we score only 18% of all poses (i.e., a fraction within the applicability domain) but the nearest native pose has an average rank of 1000.

In Figure 4, we also show the rank of the native pose for each protein complex in the testing set among all computationally generated decoys. Although in practice we are scoring non-native poses generated by docking rather than native, we still use this metric for comparative evaluation of the SPIDER scoring function. Notice that when comparing the curve for the native pose to that of the nearest-native pose, the former is shifted to the left, thus giving higher rank to the native pose compared to the nearest-native pose. In Figure 5, we show the 3D structure for one of the complexes highlighting the native, nearest-native, and highest ranked poses.

The ability of SPIDER to distinguish hits in the top "N" poses

In this validation method, we look for the number of hits (defined as poses with $\mathrm{RMSD}^{\mathrm{Pose}}$ value below 2.5 Å) a scoring function can find in the top, e.g., 500, 2000, 5000, … etc., number of poses among the 54,000 decoys for each protein complex. The more hits we find, the higher is our chance to find the correct pose of interaction for a protein complex. Figure 6 shows the results obtained for the two scoring functions we are comparing in this study.

Notice that the ZRANK score (in dashed lines) is not affected by the $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoff values since it is not involved in the internal development of the score. On the other hand, looking at SPIDER score, we notice that the probability to distinguish hits (represented by larger area under the ROC curve) increase as we move from lower $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoffs (0.3 Å) to higher $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoffs (≥0.6 Å). We believe that this is due to the fact that more hits are being selectively pulled in and identified as we use higher cutoff values. It is also clear that at higher $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoffs (≥0.6 Å) SPIDER outperforms ZRANK.

Recall that in the previous section the lower $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoffs were better in identifying native and nearest-native poses. The observations from both sections inform us that our ability to distinguish hits (with $\mathrm{RMSD}^{\mathrm{Pose}}$ ≤2.5Å) is better with higher $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoffs, yet at the same time, if we are looking only for the native (or nearest-native) pose, we need to use lower $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoffs, and in this case, our ability to do so will be lower.

The rank of the first hit, last hit, and the average rank of all hits

To augment our understanding of how these close-to-native hits are ranked and where they are located among all 54,000 computationally generated poses, we analyze the ranks of these hits. We look at the distribution of these hits among all poses by reporting the rank of the first hit found (along with its $\mathrm{RMSD}^{\mathrm{Pose}}$), the rank of the last hit found, and the average of the ranks of all hits. This should provide a better illustration as to how these hits are being ranked by each scoring function, and provide a way to compare between the two scoring functions we have used in this study.

In Table I, we notice that when using the SPIDER score at lower $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoffs the ranks of the hits (first hit, last hit, and the average rank of hits) are significantly better compared to the ZRANK score. Let's analyze the results using the SPIDER score at a low $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoff value such as 0.5 Å. The first hit was found in the top 500 ranked poses in 60% of the complexes, which we were able to predict. This is 5 times better than 2,500 which is the rank of the first hit using ZRANK score. Also, the average rank of all hits was 1,285 in those 60% of all complexes compared to 20,000 using ZRANK. Finally, for those 60% cases, all hits were found within the top 3,600 ranked poses among the 54,000 poses. For comparison, it took ZRANK to screen almost the entire 54,000 poses to find all the hits.

Now let's consider the worst case scenario, which is using high (or no) cutoff values. Using the SPIDER score, we could not find the first hit until we screened 4,500 poses (or 3,500 using 0.7 Å $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoff value), unlike ZRANK where the first hit was found in the top 2,500 poses. This indicates a loss in sensitivity to close-to-native poses using high $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoff values. Yet at the same time, the average of the hits using SPIDER score was 14,000, which is better than 20,000 for the ZRANK. Also, it took the SPIDER score to screen 32,000 poses to find all hits while it took ZRANK to screen almost the entire collection of 54,000 poses to find all hits.

The correlation between $\mathrm{RMSD}^{\mathrm{Pose}}$ and the score

Another method for validating the scoring function is to look at the correlation coefficient between the score of a pose and it's $\mathrm{RMSD}^{\mathrm{Pose}}$. As the $\mathrm{RMSD}^{\mathrm{Pose}}$ of the pose decreases, we expect the score for that pose to be better. In other words, as the pose gets closer to the native pose, it should rank higher. We do not expect to see a good correlation as the pose gets further away from the native pose since the scoring function is designed to identify only close-to-native poses. Figure 7 show the correlation coefficient for both scoring functions for all poses. Although the correlation coefficients at higher $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoff values were not significant (0.5 and above), we are still showing them as a way to compare the performance of SPIDER vs. ZDOCK scoring functions. As expected using lower

RMSD$^{\text{Pattern}}$ cutoff values does afford higher correlation coefficients, but with a great loss in the percentage of cases we are able to predict.

Performance of the SPIDER scoring function in the round 21 of CAPRI

This CAPRI round 21 was designed with the goal of scoring protein-protein complexes only.[25] Therefore, 87 designed complexes (using Spanish influenza hemagglutinin as target),[38] and 120 naturally occurring complexes were supplied by CAPRI organizers to 28 participating research groups; the challenge was to classify the complexes as binders or non-binders. For each participating research group, the receiver operator characteristic (ROC) curve was generated, plotting the true-positive rate versus the false-positive rate. Using these rates, the accuracy of the classification was calculated as the area under the curve (AUC, in percentage) for each group. Our group (using SPIDER score) had an accuracy of 83%, ranking among the top 6 (out of 28) scoring groups, with the highest accuracy being 86%.

Conclusions

We have developed a knowledge-based scoring function that incorporates multi-body interactions between protein complexes. This scoring function uses frequent patterns of interaction of interfacial residues that appear in native protein complexes as a way to predict whether certain pose is close to native. The scoring function evaluates how similar “geometrically” are the interacting residues in a pose to those “native” patterns derived from native poses. It also takes into account the frequency of these “native” patterns as they appear in native complexes, and the number of matched residues at the interface of the pose we are scoring. We have implemented a way of providing how confident we are in the score by using cutoff values for the RMSD$^{\text{Pattern}}$ of matched patterns. In other words, by limiting ourselves to score only those poses that have interactions patterns very similar to the “native” patterns (i.e., have low RMSD$^{\text{Pattern}}$, below 0.5 Å), we become more confident that the poses we are able to score are more likely to be, indeed, close to the native pose. The uniqueness of this approach comes from the following features:

1. The use of Almost Delaunay Tessellation as a way of identifying interfacial residues has a great advantage over Delaunay Tessellation based methods. This is because Almost Delaunay tessellation not only finds residues that are in direct contact with the interface, but it also allows some flexibility in choosing those residues due to errors in the position of these residues that may appear in low resolution X-ray complexes.
2. Using subgraph mining to capture natural multi-residue interaction patterns for scoring. Thus, the score is implicitly based on objective multi-body terms.
3. Using “closed” subgraph in scoring; thus, small-size pattern of interacting residues (such as pair-wise interaction patterns) that do not occur except as part of a multiple number of interacting residues are eliminated since they are not considered “closed” subgraphs. This helps us to avoid scoring false positive poses.
4. In calculating the score, we take into account the geometric frequency of the common motifs that represent patterns of interacting residues. So, not only the patterns defined by their composition have to be frequent, but also their geometry has to be frequent as well.
5. Providing a way of measuring how confident we are in the score by applying cutoff values to define the applicability domain.

The SPIDER scoring function overcomes the limitations of alternative scoring functions when computing interaction scores for highly complex structures such as protein complex decoys. As stated previously, interactions are generally not independent; scoring interactions, e.g., using pairwise statistical potentials can be non-additive. Thus, scoring candidate poses using higher-order interaction terms should be more accurate than using a linear summation of lower-order interactions. This is true because higher-order interactions can be cooperative or anti-cooperative summation of the lower-order interactions. [32-35] Thus, the use of higher-order multi-body interactions in the development of statistical scoring functions is highly preferred and of greater advantage over current lower-order interactions based scoring functions.

We will continue to experiment with various approaches to improve the SPIDER scoring function. One possible future direction is to analyze the frequency and contributions of different types of multi-body interactions, followed by optimizing weights of such contributions to the total scoring function, similar to the approach employed recently by Gniewek et al.[39] Furthermore, the native pose selection may be significantly improved by taking into account the computation of fluctuational entropy in the computation of the free energy, as proposed by Zimmermann et al.[40]

The SPIDER scoring function can be used alone (with lower $\mathrm{RMSD}^{\mathrm{Pattern}}$, below 0.5 Å), or as a filtering tool (with higher $\mathrm{RMSD}^{\mathrm{Pattern}}$, above 0.5 Å) to drastically reduce the number of docking candidates needed, allowing a more extensive scoring using energy-based scoring function methods similar to what we observed in our most recent studies of protein-ligand complexes.[41, 42]

In summary, we believe that we have used sensible, solid, and unbiased methods to verify the new SPIDER scoring function. We suggest that these validation methods demonstrate clearly that the new scoring function can out-perform the popular ZRANK scoring function using different independent metrics. We show that SPIDER is able to identify close-to-native poses when used within its applicability domain. The use of multi-body interaction terms within the scoring function based on frequent geometric patterns of interaction of interfacial residues is a novel, simple approach that affords more accurate scoring of protein complex decoys as compared to a commonly used scoring function. We expect that SPIDER can be employed both independently as well as, e.g., a pre-filter before using the fine resolution energy-based methods.

Acknowledgments

The authors acknowledge the financial support of the studies reported in this paper by the NIH (GM068665 and ARRA Supplement GM066940-06S). We would like to thank the ZDOCK team for making the decoy poses as well as their score/rank available to the public; Professor Vakser and his group for providing the Dockground dataset to be used for training scoring functions; Dr. Luke Huan at the University of Kansas for providing the subgraph mining program (FFSM); Dr. Deepak Bandyopadhyay (GSK) for providing the program for computing Almost-Delaunay Tessellation.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
2. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking. Protein Sci. 2005; 14:1328–1339. [PubMed: 15802647]
3. Pons C, Talavera D, de lC X, Orozco M, Fernandez-Recio J. Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking. J Chem Inf Model. 2011; 51:370–377. [PubMed: 21214199]
4. Garzon JI, Lopez-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J, Chacon P. FRODOCK: a new approach for fast rotational protein-protein docking. Bioinformatics. 2009; 25:2544–2551. [PubMed: 19620099]

5. Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z. Integrating statistical pair potentials into protein complex prediction. Proteins. 2007; 69:511–520. [PubMed: 17623839]

6. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol. 1997; 272:106–120. [PubMed: 9299341]

7. 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. Proc Natl Acad Sci U S A. 1992; 89:2195–2199. [PubMed: 1549581]

8. Ritchie DW, Kemp GJ. Protein docking using spherical polar Fourier correlations. Proteins. 2000; 39:178–194. [PubMed: 10737939]

9. Lensink MF, Mendez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins. 2007; 69:704–718. [PubMed: 17918726]

10. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a Critical Assessment of PRedicted Interactions. Proteins. 2003; 52:2–9. [PubMed: 12784359]

11. Zhang C, Liu S, Zhou H, Zhou Y. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci. 2004; 13:400–411. [PubMed: 14739325]

12. Huang SY, Zou X. An iterative knowledge-based scoring function for protein-protein recognition. Proteins. 2008; 72:557–579. [PubMed: 18247354]

13. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophys J. 2003; 84:1895–1901. [PubMed: 12609891]

14. Glaser F, Steinberg DM, Vakser IA, Ben Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins. 2001; 43:89–102. [PubMed: 11276079]

15. Lo CL, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol. 1999; 285:2177–2198. [PubMed: 9925793]

16. Moont G, Gabb HA, Sternberg MJ. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins. 1999; 35:364–373. [PubMed: 10328272]

17. Bernauer J, Aze J, Janin J, Poupon A. A new protein-protein docking scoring function based on interface residue properties. Bioinformatics. 2007; 23:555–562. [PubMed: 17237048]

18. Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol. 1996; 3:213–221. [PubMed: 8811483]

19. Krishnamoorthy B, Tropsha A. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. Bioinformatics. 2003; 19:1540–1548. [PubMed: 12912835]

20. Li X, Liang J. Geometric cooperativity and anticooperativity of three-body interactions in native proteins. Proteins. 2005; 60:46–65. [PubMed: 15849756]

21. Godzik A, Skolnick J. Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc Natl Acad Sci U S A. 1992; 89:12098–12102. [PubMed: 1465445]

22. Godzik A, Kolinski A, Skolnick J. Topology fingerprint approach to the inverse protein folding problem. J Mol Biol. 1992; 227:227–238. [PubMed: 1522587]

23. Zhu H, Sommer I, Lengauer T, Domingues FS. Alignment of non-covalent interactions at protein-protein interfaces. PLoS One. 2008; 3:e1926. [PubMed: 18382693]

24. Pierce B, Weng Z. ZRANK: reranking protein docking predictions with an optimized energy function. Proteins. 2007; 67:1078–1086. [PubMed: 17373710]

25. Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, Zhou HX, Mitchell JC, Demerdash ON, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko JS, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Aze J, Soner S, Ovali SK, Ozbek P, Tal NB, Haliloglu T, Hwang H, Vreven T, Pierce BG, Weng Z, Perez-Cano L, Pons C, Fernandez-Recio J, Jiang F, Yang F, Gong X, Cao L, Xu X, Liu B, Wang P, Li C, Wang C, Robert CH, Guharoy M, Liu S, Huang Y, Li L, Guo D, Chen Y, Xiao Y, London N, Itzhaki Z, Schueler-Furman O, Inbar Y, Potapov V, Cohen M, Schreiber G, Tsuchiya Y, Kanamori E, Standley DM, Nakamura H, Kinoshita K, Driggers CM, Hall RG, Morgan JL, Hsu VL, Zhan J, Yang Y, Zhou Y, Kastritis PL, Bonvin AM, Zhang W, Camacho CJ, Kilambi KP, Sircar A, Gray JJ, Ohue M, Uchikoga N, Matsuzaki Y, Ishida T, Akiyama Y, Khashan R, Bush S, Fouches D, Tropsha A, Esquivel-

Rodriguez J, Kihara D, Stranges PB, Jacak R, Kuhlman B, Huang SY, Zou X, Wodak SJ, Janin J, Baker D. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol. 2011; 414:289–302. [PubMed: 22001016]

26. Douguet D, Chen HC, Tovchigrechko A, Vakser IA. DOCKGROUND resource for studying protein-protein interfaces. Bioinformatics. 2006; 22:2612–2618. [PubMed: 16928732]

27. Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. Proteins. 2010; 78:3111–3114. [PubMed: 20806234]

28. Bandyopadhyay D, Snoeyink J. Almost-Delaunay simplices: Robust neighbor relations for imprecise 3D points using CGAL. Computational Geometry. 2007; 38:4–15.

29. Tropsha A, Carter CW Jr. Cammer S, Vaisman II. Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins. Methods Enzymol. 2003; 374:509–544. [PubMed: 14696387]

30. Huan, J.; Wang, W.; Prins, J. Efficient Mining of Frequent Subgraph in the Presence of Isomorphism. Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM); 2003. p. 549-552.

31. Huan, J.; Wang, W.; Bandyopadhyay, D.; Snoeyink, J.; Prins, J.; Tropsha, A. Mining Family Specific Residue Packing Patterns from Protein Structure Graphs. Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB); 2004. p. 308-315.

32. Rank JA, Baker D. A desolvation barrier to hydrophobic cluster formation may contribute to the rate-limiting step in protein folding. Protein Sci. 1997; 6:347–354. [PubMed: 9041636]

33. Shimizu S, Chan HS. Anti-cooperativity and cooperativity in hydrophobic interactions: Three-body free energy landscapes and comparison with implicit-solvent potential functions for proteins. Proteins. 2002; 48:15–30. [PubMed: 12012334]

34. Czaplewski C, Rodziewicz-Motowidlo S, Liwo A, Ripoll DR, Wawak RJ, Scheraga HA. Molecular simulation study of cooperativity in hydrophobic association. Protein Sci. 2000; 9:1235–1245. [PubMed: 10892816]

35. Ben-Naim A. Statistical potentials extracted from protein structures: Are these meaningful potentials? J Chem Phys. 1997; 107:3698–3706.

36. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inf. 2010; 29:476–488.

37. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. Altern Lab Anim. 2005; 33:445–459. [PubMed: 16268757]

38. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science. 2011; 332:816–821. [PubMed: 21566186]

39. Gniewek P, Leelananda SP, Kolinski A, Jernigan RL, Kloczkowski A. Multibody coarse-grained potentials for native structure recognition and quality assessment of protein models. Proteins. 2011; 79:1923–1929. [PubMed: 21560165]

40. Zimmermann MT, Leelananda SP, Gniewek P, Feng Y, Jernigan RL, Kloczkowski A. Free energies for coarse-grained proteins by integrating multibody statistical contact potentials with entropies from elastic network models. J Struct Funct Genomics. 2011; 12:137–147. [PubMed: 21674234]

41. Hsieh JH, Yin S, Wang XS, Liu S, Dokholyan NV, Tropsha A. Cheminformatics Meets Molecular Mechanics: A Combined Application of Knowledge-Based Pose Scoring and Physical Force Field-Based Hit Scoring Functions Improves the Accuracy of Structure-Based Virtual Screening. J Chem Inf Model. 2011

42. Hsieh JH, Yin S, Liu S, Sedykh A, Dokholyan NV, Tropsha A. Combined application of cheminformatics- and physical force field-based scoring functions improves binding affinity prediction for CSAR data sets. J Chem Inf Model. 2011; 51:2027–2035. [PubMed: 21780807]

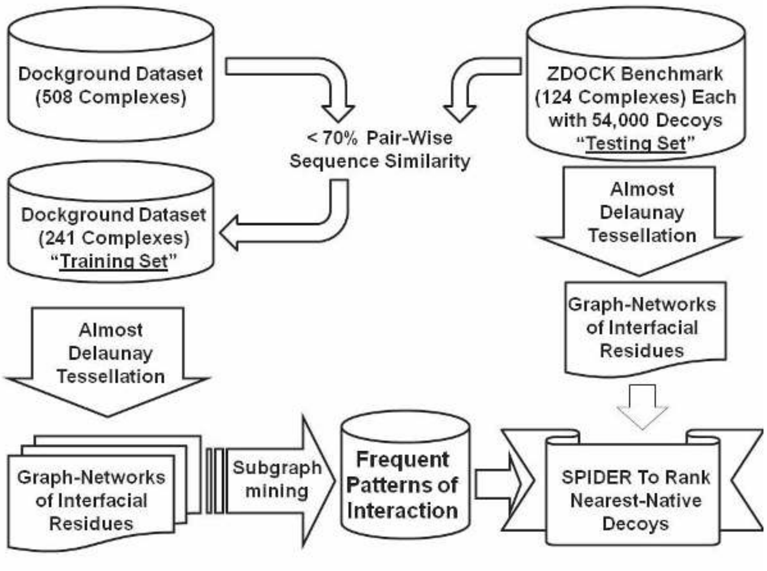
Dockground Dataset (508 Complexes)
ZDOCK Benchmark (124 Complexes) Each with 54,000 Decoys "Testing Set"
< 70% Pair-Wise Sequence Similarity
Dockground Dataset (241 Complexes) "Training Set"
Almost Delaunay Tessellation
Almost Delaunay Tessellation
Graph-Networks of Interfacial Residues
Graph-Networks of Interfacial Residues
Subgraph mining
Frequent Patterns of Interaction
SPIDER To Rank Nearest-Native Decoys

Figure 1.

A workflow for the preparation of the training and testing datasets as well as the derivation of interaction patterns which are employed in SPIDER for scoring/ranking.

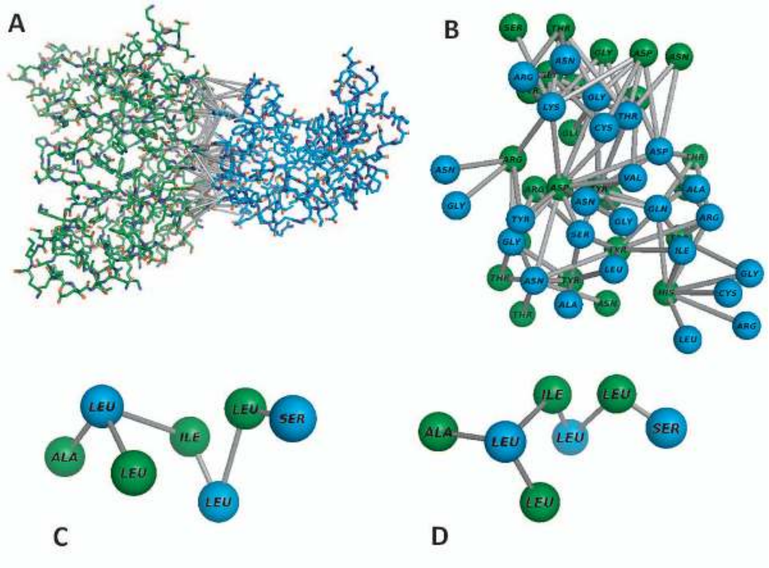

Figure 2.

(A) Protein-protein interface identified for protein dimer (PDB id: 1a2y) using Almost-Delaunay Tessellation. The gray lines connect the interacting side chain centroids at the interface of the two chains in the protein complex, thus forming a contact residue network. (B) A graph formed by the interfacial contact network extracted from 1a2y complex. (C)and (D) are examples of a frequent pattern for multiple interacting residues' centroids derived using subgraph mining. This pattern contains seven residues (SER, ALA, ILE, and four LEU), and was found in 10 protein complexes (1b0n, 1ci6, 1fs1, 1nkp, 1nql, 1wmi, 1x3w, 1xkp, 1xou, and 2a6q). The same pattern can appear in different geometries in different protein complexes. In (C), the pattern appears with certain geometry in one of the protein complexes ("1b0n"). In (D), the same pattern appears with different geometry in another protein complex ("2a6q"). All geometries of all frequent patterns are stored to be used in deriving the SPIDER score.

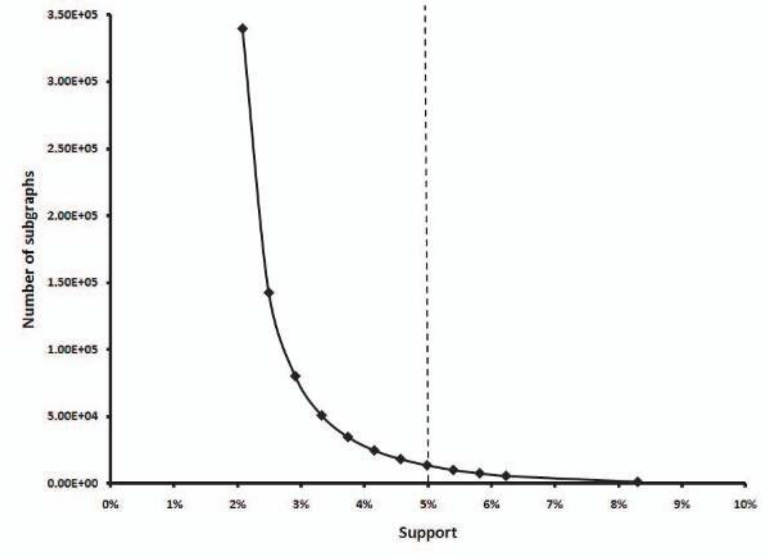

Figure 3.

The number of frequent subgraphs mined from the training set as a function of the support value.

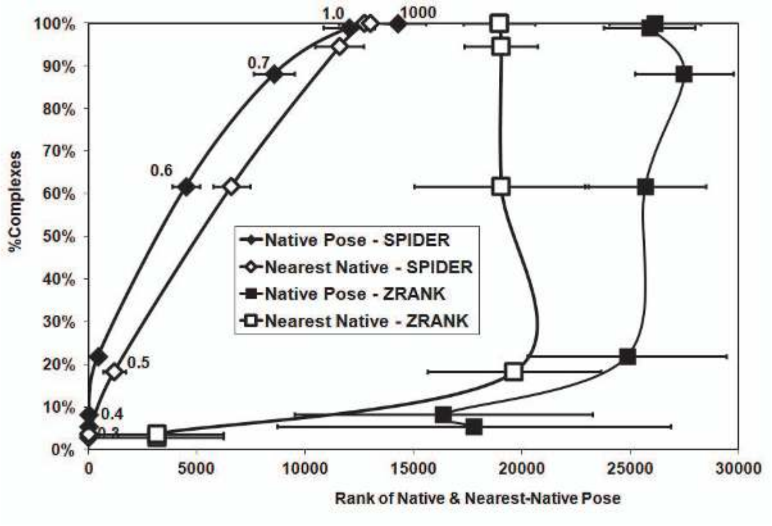

Figure 4.

The influence of the applicability domain on the accuracy of prediction and the number of complexes that can be predicted. Average rank order of the native as well as the nearest-neighbor decoy poses across all complexes in the test set using both SPIDER and ZRANK. Several RMSD^{Pattern} cutoff values are employed as shown by values next to each square (0.3 Å, 0.4 Å, 0.5 Å, 0.6 Å, 0.7 Å, 1.0 Å, and above 1 Å; error bars are shown as well). Notice that ZRANK ranks native and nearest-neighbor poses much worse that SPIDER using even the least restrictive applicability domaincutoff of 1,000 Å.

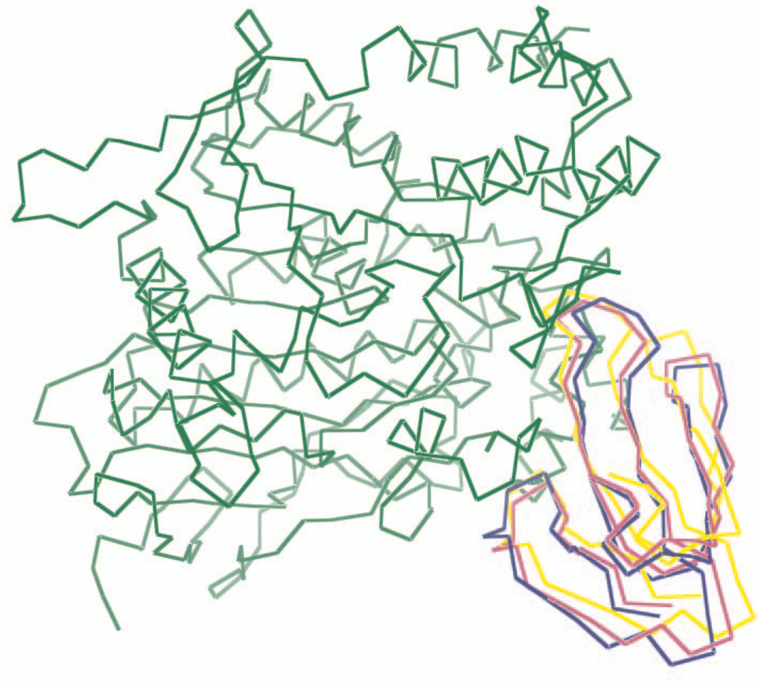

Figure 5.

The native structure for the dimer protein complex “1mah” is displayed in green and blue. The larger monomer (in green) is treated as a receptor, while the smaller monomer (in blue) is the ligand. The graph show the decoy pose (of the docked ligand) that is the closest to the native, with RMSD^{Pose} 0.86 Å, in magenta. Also, the pose that was ranked the highest using SPIDER (with RMSD^{Pose} 1.25 Å) is shown in yellow.

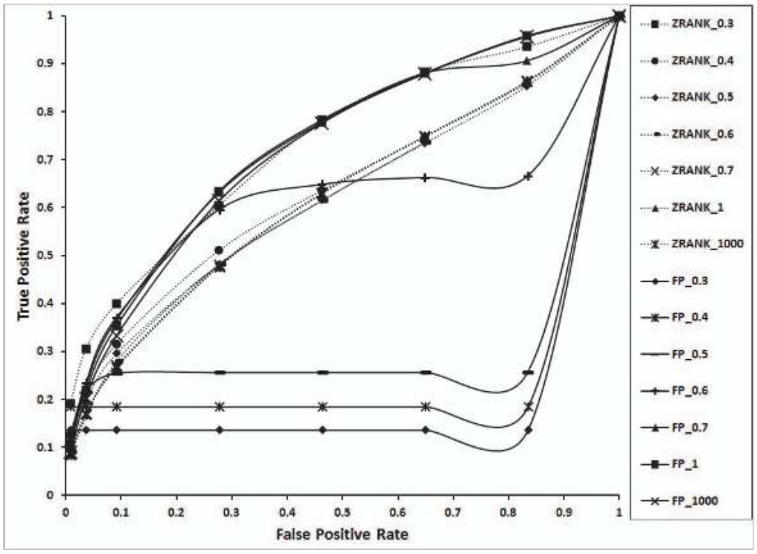

Figure 6.

The Receiver Operating Characteristic (ROC) curves for each scoring function (SPIDER and ZRANK) is used to compare each function's ability to distinguish true positive hits. The predictability of a scoring function is measured by the number of hits the function can identify in the top: 500, 2000, 5000, 15000, 25000, 35000, 45000, and 54000 poses. In addition, the effect of several RMSD$^{\text{Pattern}}$ cutoff values (0.3 Å, 0.4 Å, 0.5 Å, 0.6 Å, 0.7 Å, 1.0 Å, and above 1.0 Å) is demonstrated as well.

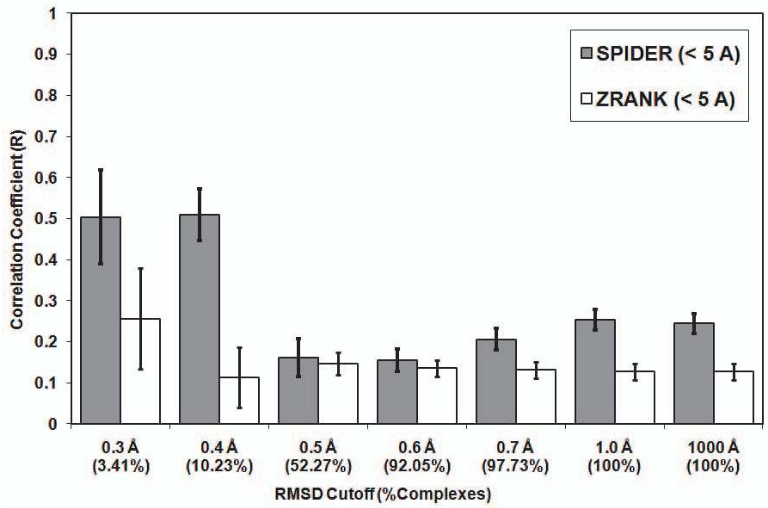

Figure 7.

The correlation coefficient between the score of a pose and the $\mathrm{RMSD}^{\mathrm{Pose}}$ of that pose as a function of different $\mathrm{RMSD}^{\mathrm{Pattern}}$ cutoff values.

Distribution of Hits using SPIDER and ZRANK scoring function

$RMSD^{Pattern}$(%Complexes)	Rank of First Hit ($RMSD^{Pose}$)		Average Rank of Hits		Rank of Last Hit	
	SPIDER	ZRANK	SPIDER	ZRANK	SPIDER	ZRANK
0.3 Å (4.35%)	1.00 (1.40 Å)	360.00 (1.80 Å)	9	13943	18.25	50961
0.4 Å (18.48%)	20.00 (1.81 Å)	277.18 (1.72 Å)	37.86	18486	76.71	52489
0.5 Å (73.91%)	547.37 (1.94 Å)	1935.00 (1.80 Å)	1286	20150	2362.3	49859
0.6 Å (94.57%)	1834.70 (1.94 Å)	2025.90 (1.82 Å)	7301.8	20291	16089	48930
0.7 Å (98.91%)	3521.40 (1.95 Å)	2388.40 (1.83 Å)	12046	19954	27733	47608
1.0 Å (100%)	4452.50 (2.00 Å)	2435.90 (1.84 Å)	13672	20001	31653	47658
1000 Å (100%)	4646.20 (1.96 Å)	2435.90 (1.84 Å)	14391	20001	32273	47658