InterPred: A pipeline to identify and model protein-protein interactions

Claudio Mirabello[1] and Björn Wallner[1*]

[1]Division of Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, SE-581 83 Linköping, Sweden

Short title: InterPred models protein-protein interactions

Keywords: protein structure prediction, protein modelling, machine learning, random forest, docking

*to whom correspondence should be addressed

Abstract

Protein-protein interactions (PPI) are crucial for protein function. There exist many techniques to identify PPIs experimentally, but to determine the interactions in molecular detail is still difficult and very time-consuming. The fact that the number of PPIs is vastly larger than the number of individual proteins makes it practically impossible to characterize all interactions experimentally. Computational approaches that can bridge this gap and predict PPIs and model the interactions in molecular detail are greatly needed. Here we present InterPred, a fully automated pipeline that predicts and model PPIs from sequence using structural modeling combined with massive structural comparisons and molecular docking. A key component of the method is the use of a novel random forest classifier that integrate several structural features to distinguish correct from incorrect protein-protein interaction models. We show that InterPred represents a major improvement in protein-protein interaction detection with a performance comparable or better than experimental high-throughput techniques. We also show that our full-atom protein-protein complex modelling pipeline performs better than any state of the art protein docking method on a standard benchmark set. In addition, it was also top-ranked in the latest CAPRI37 benchmark (http://duffman.it.liu.se/capri/).

InterPred source code can be downloaded from http://wallnerlab.org/InterPred

1 Introduction

Protein-protein interactions (PPI) are crucial for many cellular functions, such as signal transduction, transport, metabolism, and transcription. Since knowledge

of PPIs is important to understand both basic biology and human disease at the molecular level [1], major efforts have been devoted to experimentally characterize PPIs from the level of detecting interactions to exact molecular details of the interaction [2].

Detection and identification of protein interactions can be experimentally performed using high-throughput (HT) techniques such as yeast two-hybrid [3] and affinity purification [4]. But these methods have inherent limitations resulting in many false positives and negatives [5]. Novel methods using proximity-ligation techniques like BioID [6] are promising, but are limited to finding proteins that are part of the same complex and not necessarily in direct physical interaction. Characterization of the quaternary structure of proteins in molecular detail can still only be performed for individual proteins, using regular structural biology methods like x-ray crystallography, NMR or cryo-EM.

Complementing the experimental methods, several computational approaches have also been derived to predicted if proteins interact from the amino acid sequence [7, 8, 9], using co-evolution [10], gene co-expression [11], and phylogenetic profiles [12] or by combining different sources of information [13].

Similarly, there are many methods for predicting the quaternary structures of proteins, however still mostly restricted to dimers. These methods can be categorized as template-based or template-free methods. Template-based modeling methods identify complex structure templates by aligning the amino acid sequences [14, 15, 16] or structural models [17, 18] of the target chains against solved complex structures in the PDB (Protein Data Bank) or libraries of the complex interface [19, 20, 21, 22]. In template-free methods complex structures are constructed by assembling known or modeled monomers using protein-protein

docking [23, 24, 25, 26].

Both categories of methods have their advantages and disadvantages. Template-free modeling methods have the advantage that they can be applied to any protein pair given that the monomer structures are known or can be modelled. But in general the quality of the prediction cannot be guaranteed in particular if the monomer structures change upon binding [27]. A further limitation is that many thousands of docking models have to be sampled to find at least one correct docking model [28], making them unsuitable for genome-wide applications.

Template-based methods have in general higher accuracy if there are homologous complex templates available. The accuracy depends heavily on the evolutionary distance between the target and template, and drops rapidly when the sequence identity approaches the twilight-zone (i.e. when the sequence identity is around 30%). However, it has been shown that protein-protein interfaces are degenerate and that in distantly and even unrelated proteins often are similar [29, 30, 31]. This means that template-based modeling could potentially be expanded using analogous protein complex template structures.

Since structure is more conserved than sequence, methods that compare structural models, as opposed to sequences, against a template library should have a higher change of finding analogous templates. Indeed, by conducting the structural search more broadly to find similar interfaces from proteins not necessarily homologous, combined with features extracted from structural alignments with non-structural clues in a Bayesian classifier has enabled the identification of PPIs on a genome-wide scale (human, yeast) with accuracy that is comparable to HT experimental techniques [32]. Other methods have also reported improved performance in docking model quality using structural alignment searches [18, 22].

However, despite the seemingly large number of available methods for template-based docking using structural alignments [18, 22, 33, 34, 35], none of them, to the best of our knowledge, is available for practical large-scale use outside the lab they were developed.

In this study, we present InterPred, a computational pipeline to predict and model protein-protein interactions using structural modeling combined with massive structural comparisons and molecular docking. A key component of the method is the use of a random forest classifier to integrate several structural features to distinguish correct from incorrect protein interaction models. The method is open source and available as a stand-alone download.

2 Methods

The aim of InterPred is two-fold: 1) to decide if two protein interaction and 2) If so, predict how the two proteins interact.

In short, the InterPred pipeline consists of three steps (Figure 1): (1) target modelling, (2) template search including interaction modeling/scoring, and (3) refinement.

In the first step, the sequences of the protein targets, whose interaction is investigated, are used to build structural models using homology modeling. Next, structural alignments are used to find close and distant structural similarities to the two models in the PDB [36]. Whenever the structures similar to the two models also form a complex in the PDB, it defines an interaction template for modeling the interaction. An interaction model is then built by superimposing the representative structures to their corresponding structural homologs in the interaction

template. Based on the interaction model a set of features are calculated and used as input to a random forest classifier trained to sift out the more promising models that will go through to the final refinement step. Each step is described in detail below.

2.1 Structural Modelling of Target Sequences

The construction of structural models for the target sequences is a crucial step towards the search for structural templates; without any models it is not possible to run InterPred at all. The target sequences are used in a homology modelling system combining HHblits [37] from the HHpred suite for the template search and MODELLERv9.13 [38] to build full-atom 3D models. The Hidden Markov Model (HMM) profiles are built by searching the *uniprot20_2013_03* database bundled in the HHpred suite, using two iterations of HHblits with E-value cut-off 10^{-2} and maximum pairwise identity 90%. The HMM profile was then used to search against the HHpred PDB database clustered at 70% sequence identity (September 6, 2014) with the same settings. Models were constructed using the template with the lowest E-value and for each non-overlapping template with $E<10^{-3}$ using MODELLER. If different templates cover different regions of the target sequence, models are built separately for each region. This is particular useful when building models of multi-domain proteins, as it provides a form of automatic domain splitting. It is in fact preferable to run the structural template search with single domain structures rather than a large multi-domain structure. Because the size difference between two domains during the structural search could cause a bias towards templates that are similar to the larger domain and

potentially useful templates for the smaller domain could be missed.

2.2 Template Search and Interaction Modelling

The *structural template search* is performed using structural alignments with TM-align [39], for each of the representative structures from the target modelling step, against every chain in the PDB using both biological and asymmetric unit. All similar chains in PDB with TM-score > 0.5 and RMSD < 4 Å define the space of similar structures of a given target in the PDB. Potential interaction templates are defined whenever elements in the spaces of the similar structures of the two targets coincide to some experimental structure, i.e. the same PDB code is found in both spaces, and the targets are similar to chains from the same PDB entry. For each potential interaction template, a coarse interaction model is obtained in the *interaction modelling* step, by superimposing the structural representatives of the targets onto the interaction template, thereby transferring the positional relationship of the template to the targets. However, it is important to note that not all of the interaction models constructed in this way will be correct. And given multiple interaction models for the same targets, a ranking method is needed to select those that are potentially better given the features at hand.

2.3 Ranking Interaction Models

To construct a ranking of plausible interaction models, a random forest classifier was trained to predict the likelihood of interaction based on the series of features (Table 1). All features are only based on properties that can be calculated from the model itself. There are other useful information that could potentially improve

performance, like essentiality of the proteins in the interacting pair, co-expression levels, gene ontology similarity, functional similarity and phylogenetic profiles. However, none of these features will be specific to the 3D coordinates of a particular interaction model making them unsuitable for ranking interaction models. Another problem with using these type features is how to deal with cases when information is missing, which is the most likely scenario. Thus, to avoid these problems we chose to train the random forest classifier only on structural information at this stage. The structural features are described in detail below.

Interface features describe the similarity between the interface of the template and the interface of the model. We define the interface between two molecules as all residues pairs from each molecule in the complex, for which any heavy atom are within 6.5 Å. The interface features are described by the size of the interface in the template complex, model complex, and the number of residues that overlaps when the model is superimposed on the template (Figure 2).

Structural alignment features describe the quality of the structural alignment between the two target structures and their structural templates using the RMSD for the superposition, TM-score normalized by the length of target, and TM-score normalized by the length the template, for the two targets separately. In addition, a hard-wired filter was also employed, by requiring that at least one of the two TM-scores from each alignment needs to exceed 0.5 and both RMSDs need to be better than 4 Å for the interaction model to be considered at all, otherwise it is automatically discarded.

Model quality features describe the quality of the sequence alignment used to build the target structures from the sequence input and was represented by the sequence identity between the sequence templates and the target sequences for the

two models.

We also tried using the sequence similarity of the target-template structural alignment as a feature, but it did not improve performance (data not shown)

2.4 Interaction Model Refinement

As the interaction models are based purely on superpositions, they might contain severe clashes and sub optimal interactions. To produce all-atom models with no clashes and optimized interactions, interaction models were refined using RosettaDock [40] with the perturbation flag ($-dock_pert$) set to 5 12, where 5 and 12 are the standard deviations in Angstrom and degrees for the initial translation and rotation perturbation, respectively. 10,000 decoys were generated and the decoy that changed the interface the least, i.e the lowest IRMSD to starting interaction model, was selected.

2.5 Data sets

For training and testing various aspects of InterPred three different protein-protein interaction sets were used. A binary interaction set was used to train and test the method on detection, an interaction set from a recent benchmark of protein-protein interaction experiments, and finally a 3D interaction set was used to assess the refinement and the ability to predict the 3D structure of the interaction models.

2.5.1 Positive and negative binary interaction sets

The positive and negative binary interaction sets were constructed in a similar way to [13]: the set of interacting pairs, or *positive set*, is composed of protein pairs

from the yeast and human genome that have been shown to be interacting by at least two separate publications. The set of non-interacting pairs, or *negative set*, consists of human and yeast protein pairs from different cellular compartments according to Gene Ontology [41]. To make absolutely sure that none of the pairs are interacting, we were strict in picking proteins that are located in one and only one of the following compartments: membrane, mitochondria, endoplasmic reticulum, and nucleoplasm. The resulting positive set contained $30,247$ interacting pairs and the negative set contained $13,121$ non-interacting pairs.

To train the random forest classifier, relevant input interaction models were constructed by running the first part of the pipeline on the positive and negative binary interaction sets, generating on average one to two thousand interaction models per protein pair, yielding a total of 64-million models for the positive and 14-million for the negative sets, respectively. The interaction models based on the negative sets are all incorrect and can all be used as negative interaction models. For the interaction models based on the positive set on the other hand it cannot be assumed that all 64-million possible interaction models are correct. In fact, most of them will be incorrect, since the total number of possible interaction is much higher that the subset of correct interactions. To make sure that positive interaction models are correct only the 4,162 protein pairs for which obvious homologous interaction templates could be found using sequences alone were included in the *correct interaction model set*. This was done by searching for sequence homologs in the PDB for each protein in the positive set using HHblits and construct interaction models for those pairs which shared sequence templates from the same PDB. This resulted in 80,921 correct interaction models.

To prepare the sets for 10-fold cross-validation during random forest training,

the full set of sequences from the positive and negative sets were clustered using BLASTCLUST [42] and divided into 10 parts in such a way that no pair of targets from two different folds shared more than 50% sequence similarity at 90% coverage. The exact BLASTCLUST parameters are not crucial, since the training is not performed directly on sequences, but on a rather limited set of features calculated from the interaction models. The aim of the clustering was to ensure that no two sets of features from two different folds were identical.

2.5.2 Y2H benchmark set

In order to assess the performance of InterPred to predict if two proteins interact (detection), a benchmark set consisting of sequences of 184 protein pairs, where 92 pairs are directly interacting and 92 are not interacting was used [43].

2.5.3 Docking model quality benchmark

To test the capability of InterPred to produce and dock interaction models, a set of protein-protein complexes from the Protein Docking Benchmark 4.0 [44] was used. This benchmark consists of 176 bound targets and the corresponding unbound interactors.

For our comparison with the ZDOCK method, we downloaded from the developers' webpage the set of interaction decoys generated by ZDOCK for the targets in the Benchmark 4.0 set. This decoy set contains 3,600 predictions ranked by ZDOCK 3.0.2, covering all possible relative docking orientation at 15 degree resolution for each test case in the benchmark set. This is the set of decoys over which ZDOCK shows the best performance against Benchmark 4.0 [45].

2.6 Random forest classifier

The random forest classifier in the TreeBagger class from Matlab's Statistics and Machine Learning Toolbox (version R2014b) were used. An ensemble of 100 decision trees per forest was trained to recognize true vs. false interactions. The fraction of decision trees predicting a "positive" interaction determines the *InterPred score*. To find the best combination of features, several versions of the classifier were trained using different combinations of input features. The trainings were performed on the *correct interaction models set* and the full negative interaction model set using 10-fold cross-validation sets as described above. The final testing was performed on protein pairs from the full positive and negative sets using the same cross-validation set using the highest InterPred score for any interaction model for any given protein pair was taken as the score for that pair.

2.7 Performance measures

To assess performance the following quality measures were used.

FPR, the *False Positive Rate* is defined as:

$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP}. \tag{1}$$

TPR, the *True Positive Rate* (or *Sensitivity*) is defined as:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}. \tag{2}$$

Where P is the number of positive examples in the set (e.g. number of interacting pairs), N the number of negative examples, TP the number of positives

correctly identified as such by a classifier and FP the number of negatives incorrectly identified as such by a classifier (i.e. negatives incorrectly classified as positives).

3 Results and Discussion

In this work we have developed InterPred, a method that predicts if two proteins interact and if so also produce a three-dimensional interaction model. The exact details of the different parts of InterPred are described in Methods. Here, we give the main results and provide a benchmark to existing methods, both in terms of the ability to detect protein-protein interactions and to model the interaction in molecular detail.

3.1 Detection: Training of the Random Forest

To train the random forest classifier we constructed a negative and positive training set (see Methods) and run the first part of the pipeline to construct relevant interaction models using these pairs. In total 78 million interaction models for $43,368$ protein pairs were generated. Of these, 20 millions were generated from $1,397$ yeast pairs and 44 millions were generated from $2,792$ human pairs for the positive set. The remaining 14 million models were generated from a mix of $13,121$ human and yeast pairs taken from different cellular compartments to ensure that no true interactions were included in the negative set.

Features were calculated from the interaction models and several random forests were trained using different combinations of features. The impact of the different features used in the training on the quality of the predictions were compared in a

Receiver Operating Characteristic (ROC) curve (Figure 3), where the FPR (Eq. 1) is plotted against the TPR (Eq. 2) by varying the cutoff on the InterPred score. As more features are taken into account, there is an ever-increasing area under the ROC curve.

Since the ratio between positive and negative examples in protein-protein interaction is very low, i.e. when considering all the possible combinations of pairs taken from a random set of proteins, the number of non-interacting pairs will be several orders of magnitude larger than the number of interacting pairs). It is then vital that a system designed to detect or predict interactions will be capable of doing so at very low FPRs. The subplot in Figure 3 represents a zoomed in version of the area of the ROC curve (highlighted in green) at low FPRs. Here, the contribution of adding more features to represent the interactions is more evident.

3.2 Detection: Homology modelling

The first step in the InterPred pipeline is to use state-of-the-art homology modelling techniques to build 3D structures from the input sequences (see Methods). It is important to model as much as possible of the input sequence, since the pipeline is based on structural comparisons. This step also includes a domain parser that splits the target sequence into domains based on the available template hits. No attempt is made to model the complete multi-domain protein, instead each domain is used as a separate input to the pipeline.

The full set of 43,368 target pairs contained 24,237 different proteins. For 23,397 of these at least one domain could be modeled, using the domain parser an additional 7,960 (34%) domains could be modeled. These additional domains

effectively almost double the number of interactions models for both the positive and negative sets. Indicating the importance of model as much as possible of the target sequence.

It is well known that the quality of homology models depends to high degree on the sequence similarity to known structural templates. We include the sequence identity between the two target sequences and their respective templates at the homology modelling stage among the features for the prediction. The ROC curve in Figure 3 shows that the predictions are more accurate if such features are included, especially at higher FPRs.

3.3 Detection: Y2H benchmark results

Here, the ability of InterPred to correctly detect interacting pairs of proteins is benchmarked. The result is compared to a number of experimental high-throughput methods using a previous benchmark set containing 184 protein pairs, where 92 pairs are directly interacting and 92 are not [43]. To ensure a fair comparison, we removed from InterPred's training set any pair where both partners have sequence identity $\geq 50\%$ to any target pair in the benchmark. To avoid self-hits, e.g. target pairs for which the complex structure has already been experimentally resolved, we removed any structure in the PDB library that has sequence identity $\geq 90\%$ to any target in the benchmark.

As we can see from the ROC curves in Figure 4, InterPred performs considerably better than any other high-throughput method at all FPR ranges. In particular, it correctly predicts the interaction for 25% more targets than the best high-throughput method with no false positives (FPR=0).

3.4 Coarse modelling: Impact of using structure instead of sequence

A common technique to construct models of protein-protein interactions is to use template-based docking. It is conceptually similar to the InterPred pipeline, but with the important difference that only the sequence is used to find remote homologous structures. Here, we use the sequence to build a 3D model and search with the structure to find not only remote homologous but also similar structural interfaces.

To measure the added value of using structural information, we performed a "template reduction", by removing template pairs at different thresholds from 90% to 20% sequence similarity, and finally templates that cannot be detected at all by PSI-BLAST (NA, E-value ¿ 0.1). Figure 5 shows the percentage of targets from the Docking Benchmark 4.0 that could be correctly modelled, i.e. coarse interaction model with Interface Similarity score (IS-score) [46] ≥ 0.12, at different sequence similarity thresholds; by requiring that only one template (yellow) or both templates (blue) to be under the sequence similarity threshold. In reality the first case is probably most realistic, since in order to properly model an interaction both templates needs to be found. The results show that approximately 26% of the targets could be modelled even with all templates found by PSI-BLAST removed ("NA" bars). In the twilight zone around 30% sequence similarity, when sequence alignments are not reliable, InterPred is capable of finding structural templates for up to 40% of the targets in the benchmark (23% when both templates need to satisfy the threshold condition).

3.5 Coarse modelling: Evaluating the InterPred score

Next, we wanted to evaluate how well the InterPred score correlates to the effective quality of the coarse models that are produced when no obvious templates are available. To this end, we filtered out all templates with >30% sequence identity to any of the target pairs in Docking Benchmark 4.0.

InterPred was used to construct coarse interaction models that were scored using the random forest. For each target, the top 10 coarse models by InterPred score were compared to the native complex using IS-score. In Figure 6 we show the percentage of targets that can be modelled with at least one acceptable model (IS-score ≥ 0.12) among the top 10 for different InterPred score thresholds. Each bin contains a subset of targets for which InterPred was able to find a model at a given InterPred score. To compare, we also used models generated with ZDOCK 3.0.2 using 15 degree sampling, downloaded from the ZDOCK webpage (see Methods).

Even though ZDOCK is an *ab initio* method for docking, it is still one of the most widely used rigid-body docking methods that are currently available for academic use. Moreover, *ab initio* docking is usually the obvious choice when no obvious templates are available for a given complex such as in this case, where all templates with >30% sequence idenity have been removed.

The fraction of acceptable models for InterPred increases from 45% for the lowest thresholds to around 80% for the highest. Thus, it is clear that, the higher the InterPred score, the higher is the chance that the interaction model is correct.

If we compare InterPred and ZDOCK, the two methods perform similar up to InterPred score 0.4. But for InterPred score > 0.4 it progressively becomes more and more advantageous to use InterPred, as the fraction of correctly pre-

dicted targets peaks at 78% when at least one model with InterPred score ≥ 0.9 is available. The performance of ZDOCK remains roughly constant around 45% across all thresholds, meaning that the performance is fairly stable across the targets. The reason to not replace ZDOCK with a line is that we wanted to compare exactly the same targets for the two methods.

3.6 Refinement

In the final step of the InterPred pipeline the coarse interaction models were used as *starting points* for docking using the RosettaDock low and high-resolution protocol. To assess the refinement, the 176 test cases in the Docking Benchmark 4.0 were used as input to InterPred to generate 389,552 possible coarse interaction models (self-hits removed). The coarse models were ranked by InterPred score and the top 10 coarse models for each test case were selected for refinement.

Each selected coarse interaction model was used as a starting structure in a docking simulation using RosettaDock (see Methods). For each starting structure 10,000 decoys were generated (e.g. up to 100,000 decoys per test case if 10 coarse models are available) and their quality was evaluated using IS-score [46]. Two thresholds (0.12 and 0.17) have been applied to IS-score to distinguish between "Bad", "Acceptable" and "Near-native" models as defined in [34].

For each test case, the docked models (decoys) from RosettaDock need to be ranked in order to select only the more promising ones. We tested several measures to rank the models, including the Rosetta energy score, interface score (Isc), IRMSD and RMSD between the docked model and the corresponding starting model. In Figure 7 we show a scatter plot with IS-score before and after docking,

for models selected by Rosetta Score, IRMSD, and as comparison also the best docked model. In most cases the best docked model is better than the starting model. However, it is clear that neither the Rosetta Score nor interface score (data not shown) is able to select these good models. In fact, the best measure we came up with is to select the models whose interface changes the least, i.e the models with the lowest IRMSD to the starting structure.

Figure 8 shows the percentage of test cases for which at least one Acceptable (yellow) or Near-native (blue) model is produced among the top 1 to 1,000 models. To avoid selecting models from only one starting structure, the ranking was constructed by sorting each of the docking rounds for the 10 starting models separately and then picking top models from each starting structure, with rank 1 being the the best overall, rank 10 consists of top 1 from each starting structure, rank 50 is top 5 from each starting structure and so on. Strikingly, for around 52% of the test cases the rank 1 model represents an at least Acceptable model, and a Near-native model for around 45% of the test cases. This increases to around 65% and 60%, respectively, when considering the top 10 ranked models (top 1 from each of the 10 starting structures).

We used the same 15 degree sampling decoy set from ZDOCK as above to perform a comparison of the docking quality. ZDOCK has been shown to be the best *ab initio* method for protein docking in a recent review [28].

The results for ZDOCK are shown in Figure 8 (light blue and yellow bars) when the rank 1 to 1,000 decoys by ZDOCK score are evaluated.

There is a stark difference between InterPred and ZDOCK for rank 1 or rank 10, where InterPred fares considerably better than ZDOCK in number of test cases docked at Acceptable and Near-native quality. For rank 1 and rank 10 Interpred

are able to dock 150% and 40%, respectively, as many targets at Acceptable quality compared to ZDOCK. For rank 50 the number of Acceptable targets is similar but the number of targets with Near-native quality is higher for InterPred. For rank 100 and above ZDOCK performs better than InterPred. The reason for this is that while the InterPred models are docked locally around a given starting point, ZDOCK samples the whole conformational space among its 3,600 predictions. This means that by considering lower ranked docked models you will almost always find a correct model. However, we would argue that it is not really realistic to consider more than the top 10 ranked models.

4 Conclusions

We have presented InterPred, a tool that predicts if two proteins are interacting from their sequence and builds a full-atom model of the interaction. Starting from two protein sequences, 3D models are constructed and the structural neighborhoods of each 3D model are explored by comparing them against all possible structural domains in the PDB using structural alignment. Candidate coarse interaction models are constructed by superposition whenever structural neighborhoods coincide. These models are ranked using a novel random forest classifier that distinguish correct from incorrect interactions based on features calculated from the coarse interaction models. Finally, if the goal is to generate an all-atom description of the interaction, the top-ranked models are used as starting points in an all-atom docking procedure to generate the final docked conformation selected as the lowest IRMS to the starting structure.

Our results show that the use of close and remote structural interaction tem-

plates represents a major improvement when comparing to methods where only the sequence (and/or sequence profiles) are used to predict interactions.

We have tested the detection capabilities of InterPred on an independent benchmark that have previously been used to compare Y2H methods. The results show how InterPred is a valid alternative to Y2H, with a 25% improvement over the TPR of the best Y2H method in the low FPR region (FPR < 0.20).

Furthermore, we have tested the capability of InterPred in generating full-atom docked models on the Docking Benchmark 4.0, and showed that it performs better than the best available docking pipeline (ZDOCK 3.0.2) for the top ranked docking decoys with an improvement of around 150% for the first ranked decoy with 45-50% target interactions correctly modelled.

The InterPred score is also a useful predictor of the success rate for modeling the protein-protein interaction in molecular detail starting from the coarse-grained interaction model. Molecular docking starting from interaction models with InterPred scores >0.5 produces acceptable quality models for 50% of the targets, while a score >0.9 produces acceptable quality models for almost 80% of the targets.

InterPred is available as a standalone download from http://wallnerlab.org/InterPred/, and should be useful for anyone working with protein-protein interactions.

5 Acknowledgement

This was was supported by Swedish Research Council grant number 2012-5270, and the Swedish e-Science Research Center.

References

1. Wang, P. I. and Marcotte, E. M. It's the machine that matters: Predicting gene function and phenotype from protein networks. Journal of proteomics 73(11):2277–2289, October, 2010.

2. Shoemaker, B. A. and Panchenko, A. R. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS computational biology 3(3):e42, March, 2007.

3. Parrish, J. R., Gulyas, K. D., and Finley, R. L. Yeast two-hybrid contributions to interactome mapping. Current opinion in biotechnology 17(4):387–393, August, 2006.

4. Vasilescu, J. and Figeys, D. Mapping protein-protein interactions by mass spectrometry. Current opinion in biotechnology 17(4):394–399, August, 2006.

5. Braun, P. Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. Proteomics 12(10):1499–1518, May, 2012.

6. Roux, K. J., Kim, D. I., and Burke, B. BioID: a screen for protein-protein interactions. Current protocols in protein science / editorial board, John E. Coligan ... [et al.] 74:Unit 19.23., 2013.

7. Lu, L., Lu, H., and Skolnick, J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins: Structure, Function, and Bioinformatics 49(3):350–364, November, 2002.

8. You, Z.-H., Lei, Y.-K., Zhu, L., Xia, J., and Wang, B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC bioinformatics 14 Suppl 8:S10, 2013.

9. Hamp, T. and Rost, B. Evolutionary profiles improve protein–protein interaction prediction from sequence. Bioinformatics 31(12):1945–1950, 2015.

10. Pazos, F. and Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng 14(9):609–614, September, 2001.

11. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences of the United States of America 96(6):2896–2901, March, 1999.

12. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America 96(8):4285–4288, April, 1999.

13. Zhang, Q. C., Petrey, D., Garzón, J. I., Deng, L., and Honig, B. PrePPI: a structure-informed database of protein-protein interactions. Nucleic acids research 41(Database issue):D828–33, January, 2013.

14. Chen, H. and Skolnick, J. M-TASSER: an algorithm for protein quaternary structure prediction. Biophys J 94(3):918–928, Feb, 2008.

15. Mukherjee, S. and Zhang, Y. Protein-protein complex structure predictions by multimeric threading and template recombination. Structure 19(7):955–966, Jul, 2011.

16. Guerler, A., Govindarajoo, B., and Zhang, Y. Mapping monomeric threading to protein-protein structure prediction. J Chem Inf Model 53(3):717–725, Mar, 2013.

17. Davis, F. P., Braberg, H., Shen, M.-Y., Pieper, U., Sali, A., and Madhusudhan, M. S. Protein complex compositions predicted by structural similarity. Nucleic acids research 34(10):2943–2952, 2006.

18. Kundrotas, P. J. and Vakser, I. A. Global and local structural similarity in protein–protein complexes: Implications for template-based docking. Proteins: Structure, Function, and Bioinformatics 81(12):2137–2142, 2013.

19. Gao, M. and Skolnick, J. ialign: a method for the structural comparison of protein–protein interfaces. Bioinformatics 26(18):2259–2265, 2010.

20. Sinha, R., Kundrotas, P. J., and Vakser, I. A. Docking by structural similarity at proteinprotein interfaces. Proteins: Structure, 2010.

21. Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat Protoc 6(9):1341–1354, August, 2011.

22. Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O., and Gursoy, A. Prism: a web server and repository for prediction of protein–protein interactions and

modeling their 3d complexes. Nucleic acids research 42(W1):W285–W289, 2014.

23. Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. Journal of the American Chemical Society 125(7):1731–1737, February, 2003.

24. Kozakov, D., Brenke, R., Comeau, S. R., and Vajda, S. Piper: an fft-based protein docking program with pairwise potentials. Proteins: Structure, Function, and Bioinformatics 65(2):392–406, 2006.

25. Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R., and Weng, Z. Integrating statistical pair potentials into protein complex prediction. Proteins: Structure, Function, and Bioinformatics 69(3):511–520, 2007.

26. Pierce, B. G., Wiehe, K., Hwang, H., Kim, B.-H., Vreven, T., and Weng, Z. Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers. Bioinformatics 30(12):1771–1773, 2014.

27. Moreira, I. S., Fernandes, P. A., and Ramos, M. J. Protein-protein docking dealing with the unknown. J Comput Chem 31(2):317–342, Jan, 2010.

28. Huang, S.-Y. Exploring the potential of global protein–protein docking: an overview and critical assessment of current programs for automatic ab initio docking. Drug Discovery Today 20(8):969–977, 2015.

29. Tuncbag, N., Gursoy, A., Guney, E., Nussinov, R., and Keskin, O. Architectures and functional coverage of protein-protein interfaces. Journal of Molecular Biology 381(3):785–802, September, 2008.

30. Zhang, Q. C., Petrey, D., Norel, R., and Honig, B. H. Protein interface conservation across structure space. Proceedings of the National Academy of Sciences of the United States of America 107(24):10896–10901, June, 2010.

31. Gao, M. and Skolnick, J. Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. Proceedings of the National Academy of Sciences of the United States of America 107(52):22517–22522, December, 2010.

32. Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., and Honig, B. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490(7421):556–560, October, 2012.

33. Sinha, R., Kundrotas, P. J., and Vakser, I. A. Protein docking by the interface structure similarity: how much structure is needed? PloS one 7(2):e31349, 2012.

34. Tuncbag, N., Keskin, O., Nussinov, R., and Gursoy, A. Fast and accurate modeling of protein–protein interactions by combining template-interface-based docking with flexible refinement. Proteins: Structure, Function, and Bioinformatics 80(4):1239–1249, 2012.

35. Szilagyi, A. and Zhang, Y. Template-based structure modeling of protein-protein interactions. Curr Opin Struct Biol 24:10–23, Feb, 2014.

36. Ber man, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. Nucleic acids research 28(1):235–242, January, 2000.

37. Remmert, M., Biegert, A., Hauser, A., and Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods 9(2):173–175, February, 2012.

38. Webb, B. and Sali, A. Comparative Protein Structure Modeling Using MODELLER. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] 47:5.6.1–5.6.32, 2014.

39. Zhang, Y. and Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research 33(7):2302–2309, 2005.

40. Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of Molecular Biology 331(1):281–299, August, 2003.

41. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic acids research 43(Database issue):D1049–56, January, 2015.

42. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25(17):3389–3402, 1997.

43. Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Roncari, L., De Smet, A.-S., et al. An experimentally derived confidence score for binary protein-protein interactions. Nature methods 6(1):91–97, 2009.

44. Hwang, H., Vreven, T., Janin, J., and Weng, Z. Protein-protein docking benchmark version 4.0. Proteins 78(15):3111–3114, Nov, 2010.

45. ZDOCK Decoy Sets Webpage. `https://zlab.umassmed.edu/zdock/perf_decoys.shtml`.

46. Gao, M. and Skolnick, J. New benchmark metrics for protein-protein docking methods. Proteins: Structure, Function, and Bioinformatics 79(5):1623–1634, 2011.

List of Figures

1 Overview of the InterPred pipeline 30

2 The interface features captures the similarity between the interface of the template and the model. It is described by the number of residues in the template complex interface (solid lines in A), the number of residues model complex interface (solid lines in B), and the number of residues that are shared when the template and model are superimposed (double lines in C). 31

3 Performance measured by ROC curves for random forest classifiers trained with different features. 32

4 ROC curves comparing InterPred to a number of mass-spec methods on a benchmark of 184 human protein pairs. 33

5 Percentage of acceptable coarse models built by templates under a given sequence identity as found by PSI-BLAST. 34

6 Fraction of correctly (Acceptable or better) modelled targets from Docking Benchmark 4.0 by InterPred (yellow) for different InterPred score. ZDOCK (blue) is included as a reference using the targets present in each bin. A target is considered correct if it has at least one acceptable model among the top 10. 35

7 IS-scores before and after docking using different measures for ranking and selection. Top 10 docked models for each target by Rosetta Score (red), IRMDS (green), and best docked model by IS-score (blue). 36

8 The percentage of successfully docked targets, when the top rank models are selected by IRMSD, for InterPred (blue/yellow) and ZDOCK (light blue/yellow). 37

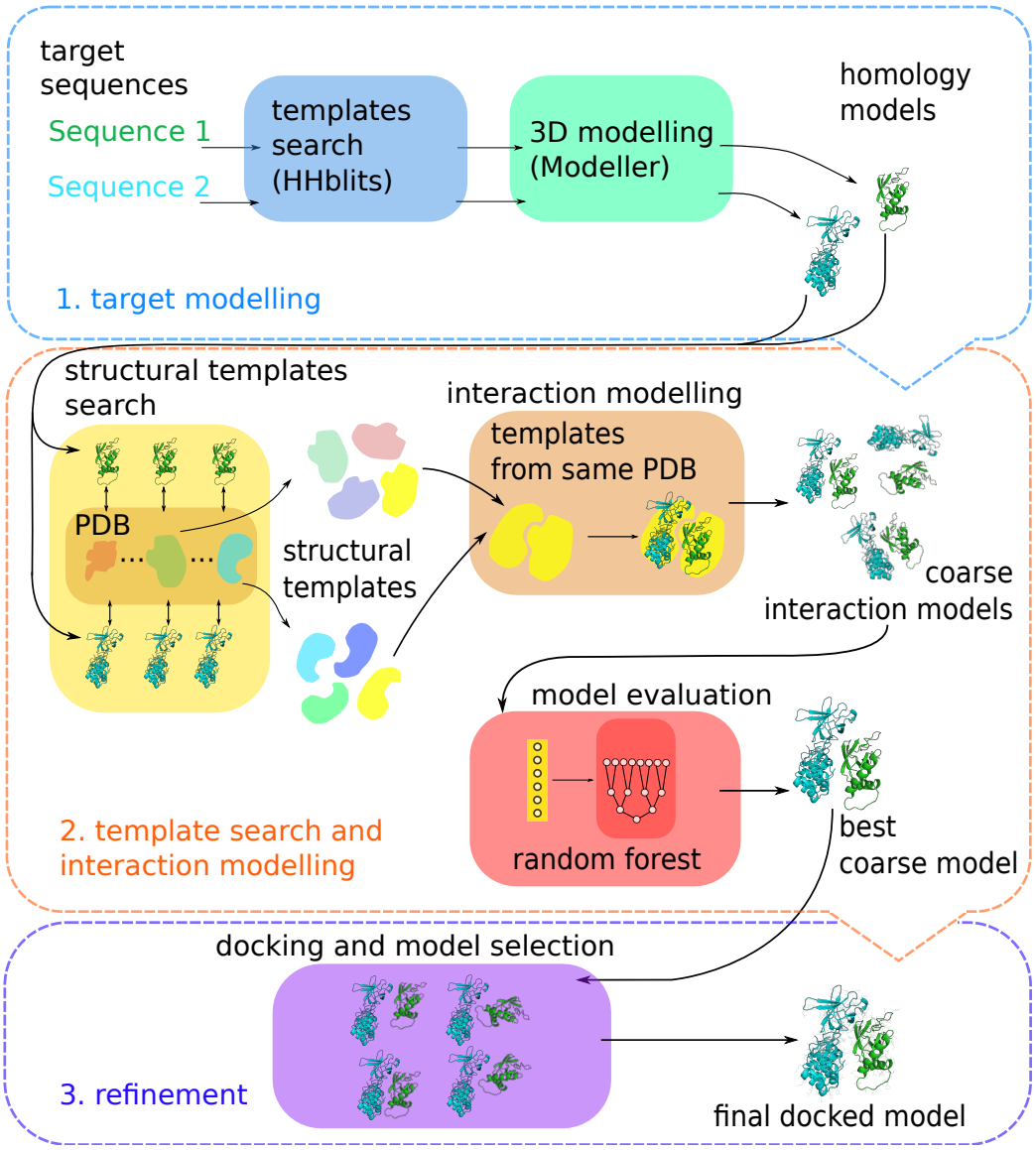

Figure 1: Overview of the InterPred pipeline

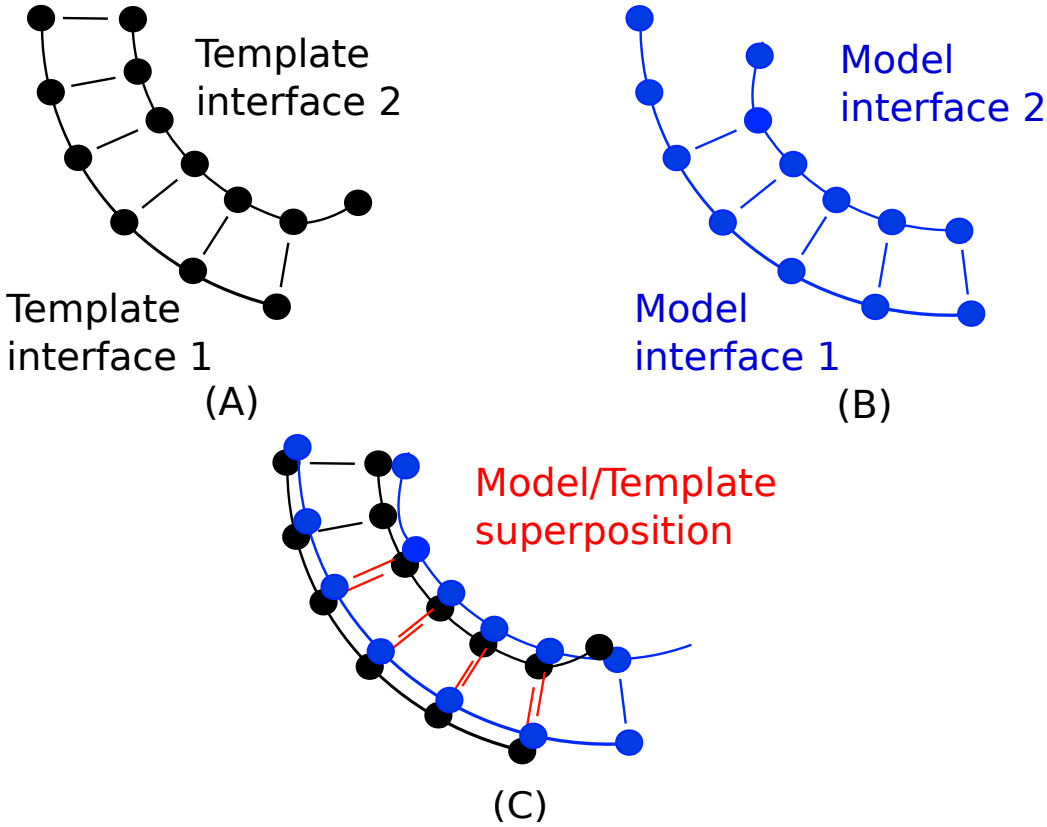

Figure 2: The interface features captures the similarity between the interface of the template and the model. It is described by the number of residues in the template complex interface (solid lines in A), the number of residues model complex interface (solid lines in B), and the number of residues that are shared when the template and model are superimposed (double lines in C).

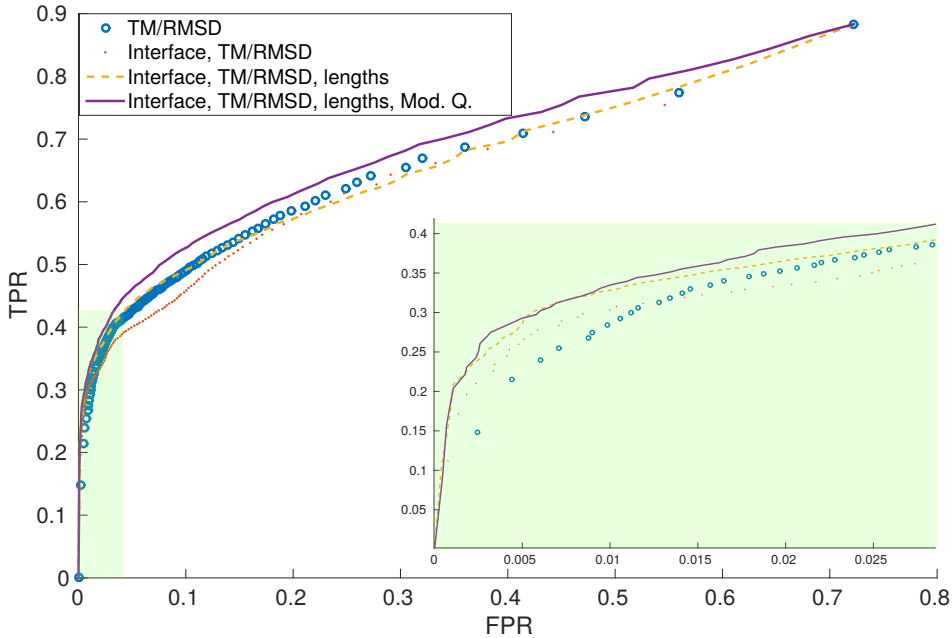

Figure 3: Performance measured by ROC curves for random forest classifiers trained with different features.

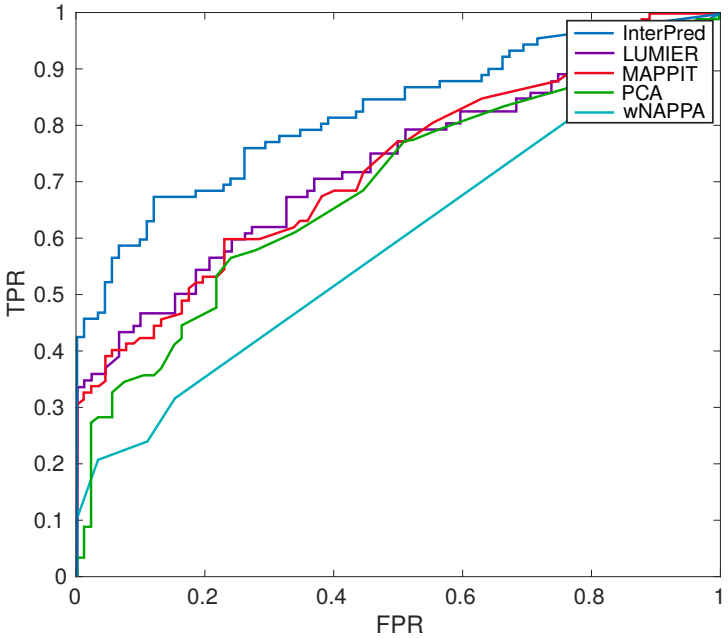

Figure 4: ROC curves comparing InterPred to a number of mass-spec methods on a benchmark of 184 human protein pairs.

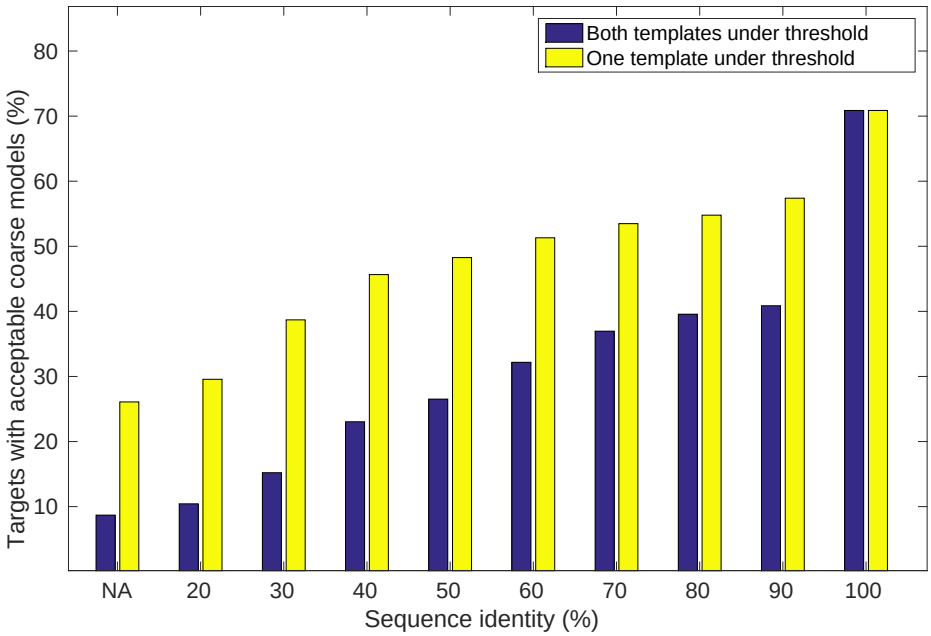

Figure 5: Percentage of acceptable coarse models built by templates under a given sequence identity as found by PSI-BLAST.

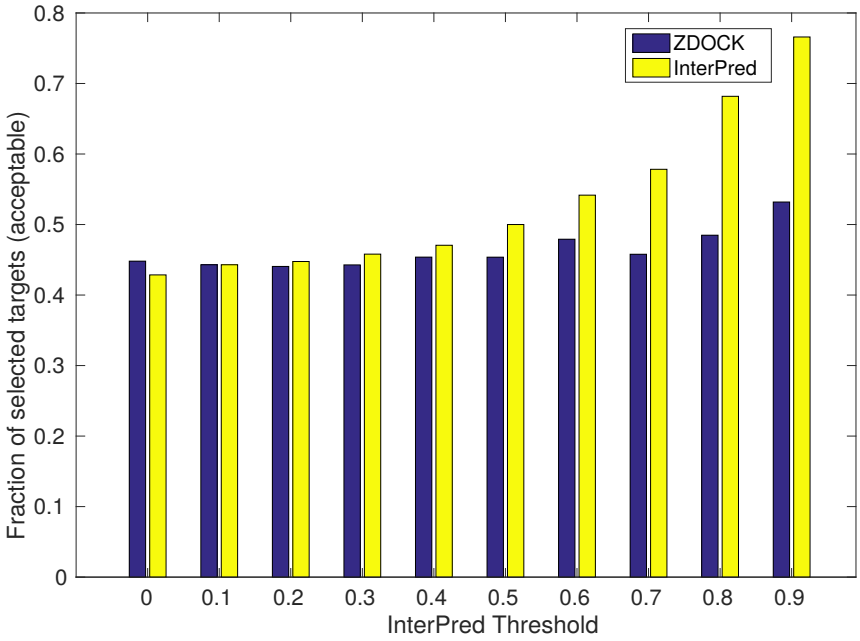

Figure 6: Fraction of correctly (Acceptable or better) modelled targets from Docking Benchmark 4.0 by InterPred (yellow) for different InterPred score. ZDOCK (blue) is included as a reference using the targets present in each bin. A target is considered correct if it has at least one acceptable model among the top 10.

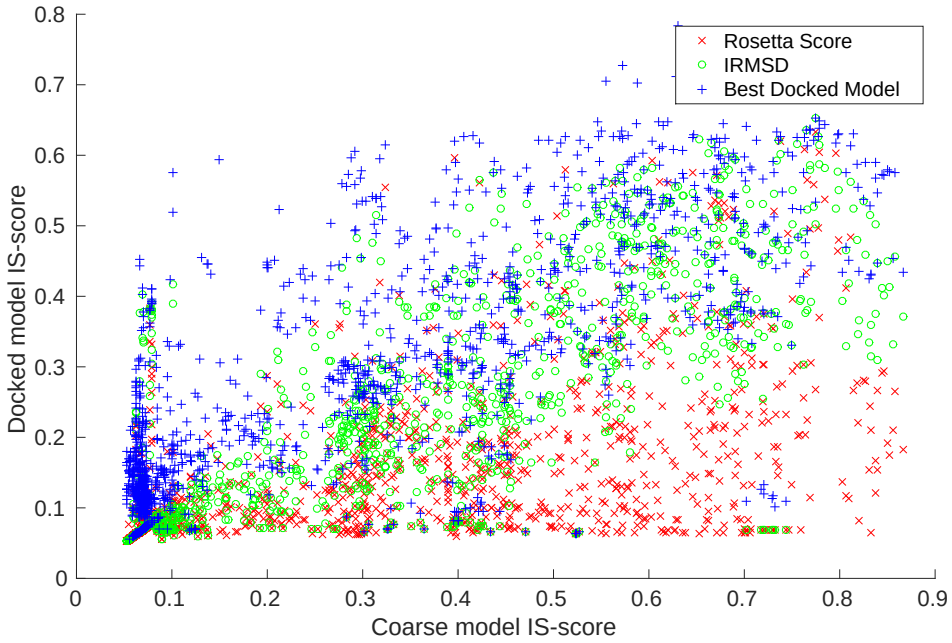

Figure 7: IS-scores before and after docking using different measures for ranking and selection. Top 10 docked models for each target by Rosetta Score (red), IRMDS (green), and best docked model by IS-score (blue).

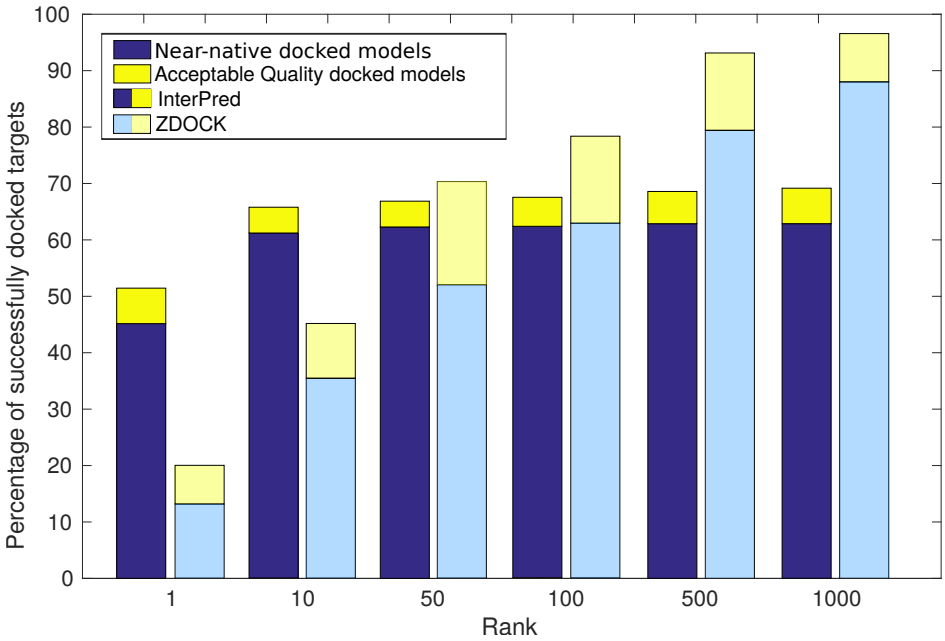

Figure 8: The percentage of successfully docked targets, when the top rank models are selected by IRMSD, for InterPred (blue/yellow) and ZDOCK (light blue/yellow).

List of Tables

1 Description of features used to train the random forest classifier. . 39

Feature group	Description	#
Interface	Interface similarity	3
Structural	Structural alignment scores (TM/RMSD)	6
Sequence lengths	Sequence lengths of targets and templates	4
Model quality	Sequence identity from the template modeling of the two targets	2

the number of feature descriptors for a particular feature group

Table 1: Description of features used to train the random forest classifier.