Linköping University | Department of Physics, Chemistry and Biology
Master thesis, 30 hp | Educational Program: Physics, Chemistry and Biology
Autumn term 2016 | LITH-IFM-A-EX--16/3280—SE

Improving detection of promising unrefined protein docking complexes

Author: Malin Rörbrink

Examiner: Björn Wallner
Supervisor: Claudio Mirabello

LINKÖPINGS
UNIVERSITET

Avdelning, institution

Division, Department

Department of Physics, Chemistry and Biology
Linköping University

Datum

Date

16/12-2016

Språk

Language

□ Svenska/Swedish
☒ Engelska/English

□ _______________

Rapporttyp

Report category

□ Licentiatavhandling
☒ Examensarbete
□ C-uppsats
□ D-uppsats
□ Övrig rapport

□ ____________

URL för elektronisk version

ISBN

ISRN: LITH-IFM-A-EX--16/3280—SE

Serietitel och serienummer **ISSN**

Title of series, numbering _____________________

Titel

Title

Improving detection of promising unrefined protein docking complexes

Författare

Author

Malin Rörbrink

Sammanfattning

Abstract

Understanding protein-protein interaction (PPI) is important in order to understand cellular processes. X-ray crystallography and mutagenesis, expensive methods both in time and resources, are the most reliable methods for detecting PPI. Computational approaches could, therefore, reduce resources and time spent on detecting PPIs. During this master thesis a method, cProQPred, was created for scoring how realistic coarse PPI models are. cProQPred use the machine learning method Random Forest trained on previously calculated features from the programs ProQDock and InterPred. By combining some of ProQDock's features and the InterPred score from InterPred the cProQPred method generated a higher performance than both ProQDock and InterPred.

This work also tried to predict the quality of the PPI model after refinement and the chance for a coarse PPI model to succeed at refinement. The result illustrated that the predicted quality of a coarse PPI model also was a relatively good prediction of the quality the coarse PPI model would get after refinement. Prediction of the chance for a coarse PPI model to succeed at refinement was, however, without success.

Nyckelord

Keyword

Protein-protein interaction, machine learning

Abstract

Understanding protein-protein interaction (PPI) is important in order to understand cellular processes. X-ray crystallography and mutagenesis, expensive methods both in time and resources, are the most reliable methods for detecting PPI. Computational approaches could, therefore, reduce resources and time spent on detecting PPIs. During this master thesis a method, cProQPred, was created for scoring how realistic coarse PPI models are. cProQPred use the machine learning method Random Forest trained on previously calculated features from the programs ProQDock and InterPred. By combining some of ProQDock's features and the InterPred score from InterPred the cProQPred method generated a higher performance than both ProQDock and InterPred.

This work also tried to predict the quality of the PPI model after refinement and the chance for a coarse PPI model to succeed at refinement. The result illustrated that the predicted quality of a coarse PPI model also was a relatively good prediction of the quality the coarse PPI model would get after refinement. Prediction of the chance for a coarse PPI model to succeed at refinement was, however, without success.

Abbreviations

CPM	Combined probability of finding the interface within a specific distance from Sc and EC given nBSA
CP score	Contact preference score (predicted ProQDock score)
cProQPred	Predictor (combined version of ProQDock and InterPred)
EC	Electrostatic complementarity
E_{rep}	Van der Waals repulsive term
E_{tmr}	The difference between Rosetta total score and the van der Waals repulsive term
Fintres	Fraction of residues at the interface
FN	False negative
FP	False positive
FPR	False positive rate
InterPred score	The probability that the PPI model is true
IS score	Interface Similarity
Isc	Interface energy
Ld	Link density
nBSA	Normalized buried surface area
PDB	Protein Data Bank
PPI	Protein-protein interaction
ProQ	Estimation of the structural PPI models accuracy
ProQDock score	Predicted DockQ score
rGb	Accessibility score
ROC	Receiver Operating Characteristic
rTs	Rosetta total score
Sc	Shape complementarity
SVM	Support vector machine
TN	True negative
TP	True positive
TPR	True positive rate (Recall)

Contents

List of Figures

List of Tables

1. Introduction

This section presents the background, purpose and aim of this master thesis.

1.1 Background

Many proteins interact with each other in order to fulfill their function [1], therefore, protein-protein interactions (PPI) are an important component in cellular processes [2] [3]. PPIs represent a key part in several areas of research, such as proteomics research in order to understand disease mechanisms or for drug development [2], where the 3D structure of the complex can give information about the protein's function [1]. However, structural determination of a protein complex is technically more demanding then determining the structure of a single protein or a protein fragment [1] [4].

There are several high-throughput methods for detecting PPIs [2] [3], e.g protein chip [5] and Yeast two-hybrid [6]. Detection of PPI through high-throughput methods often generate a large number of false negatives and positives [3] [7]. Therefore, the most reliable methods for detecting PPI are X-ray crystallography and mutagenesis, which are expensive in both time and resources [8]. Thus, new methods for predicting PPIs are developed through computational approaches [3]. Construction of these PPI models constitutes the two basic challenges in structural biology; generating realistic PPI models and evaluating the accuracy of the model, i.e. scoring how realistic the model [9] is more often referred to as the models quality.

1.1.1 Generating realistic PPI models

In order to generate realistic PPI models, a number of different computational approaches have been applied, e.g. approaches using co-evolution [10] and phylogenetic profiles [11].

Methods using co-evolution for predicting PPI works through comparing the evolutionary distance between proteins sequences. In co-evolution, phylogenetic trees are built for the two presumed interacting proteins and a similarity between the trees indicate coordinated evolution, and that all cellular complex in those phylogenetic trees would have endured a similar evolutionary pressure [10]. An example could be when both proteins disappear simultaneously within one species, indicating that the proteins cannot function without each other [10]. Methods using phylogenetic profiles consider proteins with matching or similar phylogenetic profiles to have a tendency to be functionally linked [11]. Functionally linked proteins are thought to have to evolve correlated, which indicate that the proteins participate together in a structural complex or metabolic pathway [11].

The different approaches for predicting the 3D structure are either template-based or template-free methods. A template-based method uses templates, which can be of a different kind, in order to predict the PPI. InterPred is a pipeline for creating coarse PPI models based on structural templates [12]. The structural templates, which later constitutes the coarse PPI model, are obtained through structural alignments between structural models of the investigated proteins in the Protein Data Bank, PDB [12]. The created coarse PPI models might include overlaps between the two interacting proteins [12], e.g. atoms are too close to be realistic. These overlaps can be resolved in a refinement process generating a more realistic PPI model, i.e. a refined PPI model.

1.1.2 Evaluating the accuracy of PPI models

When modeling PPIs a common technique is to generate many alternative PPI models, which necessitates methods for evaluating the quality of the models [4]. Constructing these programs for identifying the correct PPI models among the incorrect is the second challenge in structural biology [9]. In order to identify the correct PPI models, several different approaches have been applied, e.g. interface composition and shape complementarity [13] combined with machine learning [9] [12].

Evaluating the PPI model quality through interface composition and shape complementary is based on the importance of hydrophobic interactions in PPI. Since hydrophobic interactions lower the entropic energy [13], which makes binding favorable compared to not binding. In order for a presumed interaction to result in PPI the shape of the protein interfaces must be complementary [13]. Therefore, hydrophobicity and shape complementary can be used to evaluate the quality of different PPI models.

Machine learning methods learn from known examples to make predictions on unknown or new examples [14]. There are several different machine learning methods, e.g. Random Forest neural networks and Support Vector Machine (SVM) [14]. An example of a program that uses machine learning to predict the quality of PPI models is ProQDock [9]. ProQDock calculates thirteen different features directly from the PPI model. These features are then used by the machine learning method SVM in order to predict the quality of the PPI model [9].

The real quality of a PPI model can be calculated, for example, with Interface Similarity score (IS score) [15]. In order to calculate the IS score both the true structure and the predicted PPI model are needed. The score indicates how well the PPI model match the native protein-protein complex. An IS score of 0.12, or higher, is considered an acceptable PPI model [12].

1.2 Purpose of the project

For computational approaches to be an option for investigation of PPI, good programs for evaluation of the PPI models quality are essential. Particularly, since generating many alternative PPI models are a common technique when modeling PPI [4].

1.2.1 Work process

This work predicts the IS score for 30 000 previously generated coarse PPI models with the machine learning method Random Forest. Features from the program ProQDock together with the InterPred score from the pipeline InterPred was used to train the Random Forest predictor. Training was performed using 5-fold cross-validation, and the number of features was optimized so predictor only contained the essential features required for generating the highest correlation between the true IS and predicted IS score.

Furthermore, this work also tried to predict the quality of the PPI model after refinement, i.e. the refined IS score. In addition, the chance for a coarse PPI model to succeed at refinement was also tried. The chance of succeeding at refinement was defined as the difference in IS and refined IS score.

1.2.2 Aim

1. Create a Random Forest predictor that can predict the IS score for coarse PPI models based on features pre-calculated directly from the coarse PPI models themselves.

2. Based on the Random Forest predictor created for prediction of the IS score for coarse PPI models, create a predictor that can predict the quality of the PPI model after refinement. A predictor that can determine which coarse PPI models that should generate realistic refined PPI models.

2. Process

This section presents the timetable for the work process. The timetable was followed-up through continually checkups and discussions with the examiner.

Week	Activity
36	Literature study
37	Literature study (focus ProQDock)
38	Run ProQDock on all unique models
39	Debug ProQDock
40	Run ProQDock on all unique coarse PPI models
	Run ProQDock on all coarse PPI models
41	Cluster the coarse PPI models
	Create predictor
42	Write report
43	Mid-term evaluation
	Parameter optimization
44	Evaluate results
45	Create graphs
46	Create complementary results
47	Write report
48	Write report
49	Proofread
	Submit report
50	Create presentation
	Present the master thesis
51	Correct the final report
52	Correct the final report
	Write reflection document

3. Theory of methods

In this section, important theories, concepts and functions are explained to provide a deeper understanding of programs and methods used in this master thesis. The exact procedures for the work process is presented in Methods, *Section 4*.

3.1 Machine learning

Machine learning methods, e.g. Random Forest and SVM, creates predictors that based on features and labels of training data can make predictions of the label for previously unseen data [14]. The prediction can either be of classification or regression type, where classification mean that the prediction is one of a numbers of classes and regression predict real-values [16]. Regardless of which machine learning method the principle is always the same, *Figure 1*.

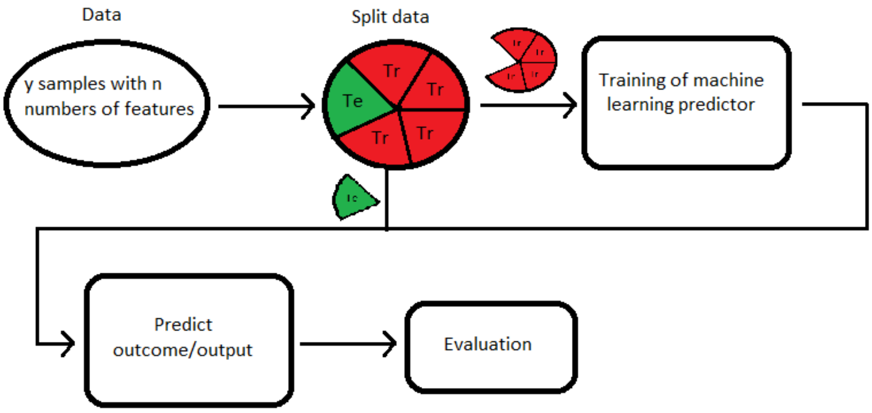

Data
y samples with n numbers of features
Split data
Te
Tr
Tr
Tr
Tr
Training of machine learning predictor
Predict outcome/output
Evaluation

Figure 1. Overview of the principle behind machine learning. During the split data step, the data is divided into test set, Te, and training set, Tr.

When constructing machine learning programs, *Figure 1*, the data will have to be divided into test- and training sets [14]. The test- and training sets contain different examples, where each example contain features and the label, e.g. a feature for predicting PPI models quality can be shape complementarity [9].

The reason for dividing the data into test- and training sets is because training and testing cannot be performed on the same data since then the program would know the label instead of predicting it [17]. Machine learning programs are constructed through training the predictor on the training sets and thereafter testing the predictor on the test set [16]. The output of the predictor is the predicted label, and the quality of the predictor is the difference between the true label and the predicted label [16].

3.1.1 Random Forest

The machine learning method Random Forest creates, just as the name suggests, a forest of random decision trees [18], where each tree is grown from the training set [19]. A decision tree for predicting if a person would play golf or not, depending on the weather could be built like *Figure 2*. The training set, *Table 1*, contain the three features; overlook, wind and humidity. Examples, in this case, indicates days where the weather was investigated and the label, play or not play, was recorded. This is an illustration of classification where you have two classes, play or not play.

Table 1. Illustration of a training set, with eight examples containing the features (outlook, wind and humidity) and the label (play/not play).

Examples	**Outlook**	**Wind**	**Humidity**	**Label**
1	Sunny	Low	High	Play
2	Rainy	Low	High	Play
3	Cloudy	High	Normal	Play
4	Sunny	Low	Normal	Play
5	Rainy	High	High	Not play
6	Sunny	High	Low	Not play
7	Cloudy	Low	Normal	Play
8	Cloudy	High	Low	Play

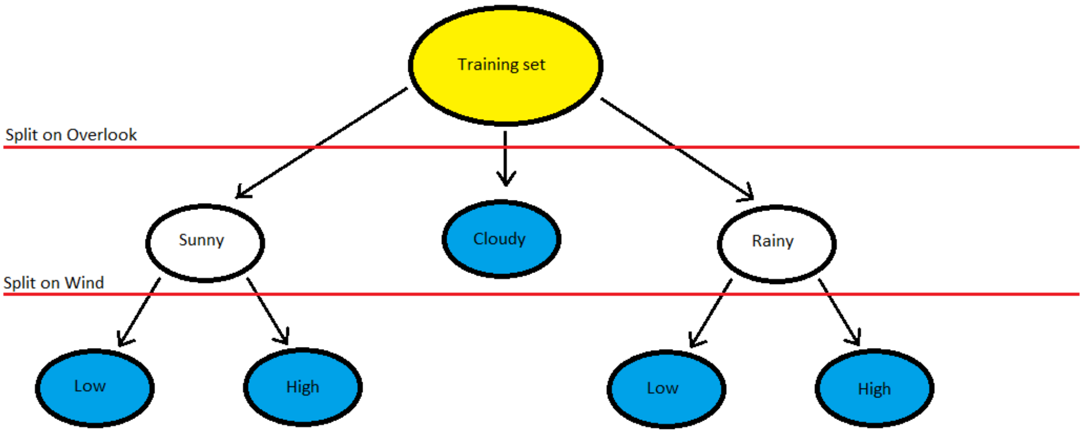


Figure 2. Overview of how a decision tree could look like for the training set presented in Table 1.

The decision tree grows from a root node, yellow node in *Figure 2*, and the data is split based on different features in order to create nodes that can predict the label [19]. The blue nodes are called leaves and in this illustration, they are all pure, i.e. all samples in the same leaf give the same label. However, when trees are allowed to become fully grown, i.e. grow until all leaves are pure, it is usually not optimal for generating the best predictions on the unseen data, and the same applies to the number of trees in the forest [20].

When the forest reaches a critical size the predictions will not become significantly better with increasing number of trees [20]. Therefore, the number of trees and max depth, the maximal distance between the root node and the leaf that is farthest away, are two parameters to optimize in order to obtain optimal predictions. The optimal parameters should be crossvalidated [20]. When combining several decision trees the label is determined by averaging all trees probabilistic prediction [19].

3.1.2 Test- and training sets

Since testing and training the machine learning predictor on the same data would result in overfitting, i.e. the predictor repeating the label it has previously seen instead of predicting it. Cross-validation is a common practice to avoid overfitting [17]. With cross-validation, the examples is divided into k number of test sets. When training the algorithm, k-1 test sets are used for training and the remaining test set is used for evaluation [9] [17]. This is repeated k times, and test sets are rotated until all test set have been used for evaluation.

In order to ensure that training and testing are not performed on the same data or data with high resemblance, all similar examples are grouped into the same test set [9]. The clustering of data in the case of interfaces can be performed through e.g. structural interface similarity by the program iAlign [21]. Where the coverage percentage for iAlign are 90 % with an error of 0.005 per query for detection of structural similar protein-protein interfaces [21]. With the threshold, P-value, 1×10^{-4} in iAlign, 0.01 % of the interfaces are classified as false positive (FP) [21], i.e. iAlign states that the interfaces of the two coarse PPI models are similar even if they are not. However, when clustering PPI models, FP are not a problem since test sets containing PPI models with no resemblance does not affect the prediction of the label, but not detecting true positives (TP) and clustering the two PPI models into different test set will result in overfitting. Thus, a higher P-value would result in more FP and therefore less false negatives (FN) which minimize the risk of overfitting.

An PPI is categorized as FN when the method falsely predicts no interactions between the two proteins and FP when the method predicts an interaction between the two proteins even though the interaction does not exist.

3.1.3 Parameter optimization

In order to obtain optimal results for the machine learning predictor, some parameters have to be optimized. For Random Forest these parameters are the number of trees and max depth [20]. These parameters can be optimized with the program GridSearchCV [22], through a grid-search over a parameter grid containing the interesting values for each parameter [23]. Every combination possible of the values for the two parameters are evaluated and ranked based on cross-validation score [24]. The cross-validation score is the accuracy of the predictor [17], and the higher the cross-validation score the better.

3.1.4 Feature importance

When constructing decision trees, Random Forest starts splitting the training set on features that will contribute more to the prediction of the label [25]. Therefore, the placement of a node for splitting on a specific feature can be used to estimate the importance of that feature [25]. Features used for splitting the data at the top of the tree, i.e. near the root node, will be of higher importance than the features used for splitting the nodes to leaves [25].

3.1.5 Evaluation

In order to evaluate the accuracy of a machine learning methods predictions, the Pearson's correlation as well as Receiver Operating Characteristic (ROC) curves, precision and recall can be calculated using the prediction and the true label [9].

Pearson's correlation is the linear relationship between the true and the predicted label ranging from -1 to 1 [26]. Where both -1 and 1 implies total linear relationship, either negative- or positive linear relationship while 0 implies no correlation between the prediction and the true label [26].

To evaluate the predictor's ability to correctly rank models, ROC curves can be used. In ROC curves the true positive rate (TPR) are plotted against the false positive rate (FPR) [9] for different cutoff points in order to visualize the tradeoff between sensitivity, TPR, and specificity [27]. TPR and FPR are calculated through Equation (1) and (2) respectively.

$$TPR = Recall = \frac{TP}{TP+FN} \tag{1}$$

$$FPR = \frac{FP}{FP+TN} \tag{2}$$

Precision visualizes how many of the predicted interactions that are actual interactions, and recall is the percentage of correctly predicted positive interaction over the total number of true positive interactions. Precision and recall are calculated through Equation (3) and (1) respectively [9]. A precision of 0.8 and a recall of 0.6 would indicate: 60 % of all positives are found and 80 % of all predictions are correct.

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

3.2 InterPred

InterPred is a pipeline whose aim is to determine if two protein interact and how the possible interaction would occur [12], i.e. create PPI models. To achieve this purpose, InterPred uses structural templates and the pipeline basically consists of three phases [12]:

Phase 1 - Target modeling
Phase 2 - Structural template identification, modeling interaction and scoring
Phase 3 - Refinement

The three different phases will be discussed in detail in *Section 3.2.1-3.2.3*.

3.2.1 Phase 1 - Target modeling

Since InterPred is a structural template based method, structural models of the target proteins structure, homology models, must be obtained in order to search for structural templates. The search and modeling of these homology models are included in InterPred's first phase, *Figure 3*.

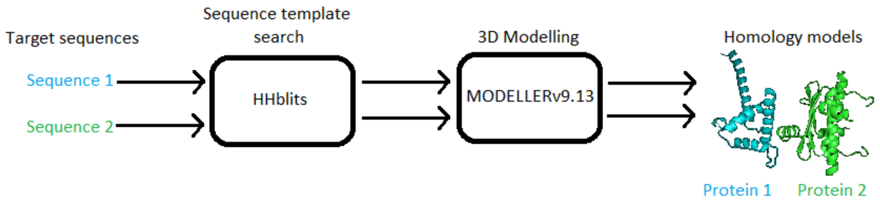


Figure 3. Overview of phase 1 in the InterPred pipeline.

The sequences for the target proteins, the proteins from the investigated interaction, are used in order to build 3D models of the proteins [12]. These 3D models are built through a homology modeling system which use HHblits [28] to search for sequence templates [12] and MODELLERv9.13 [29] to create the homology models [12].

3.2.2 Phase 2 - Structural template identification, modeling interaction and scoring

Phase 2, *Figure 4*, consist of identification of structural template, modeling the presumed interaction and scoring the coarse PPI model in order to determine its quality.

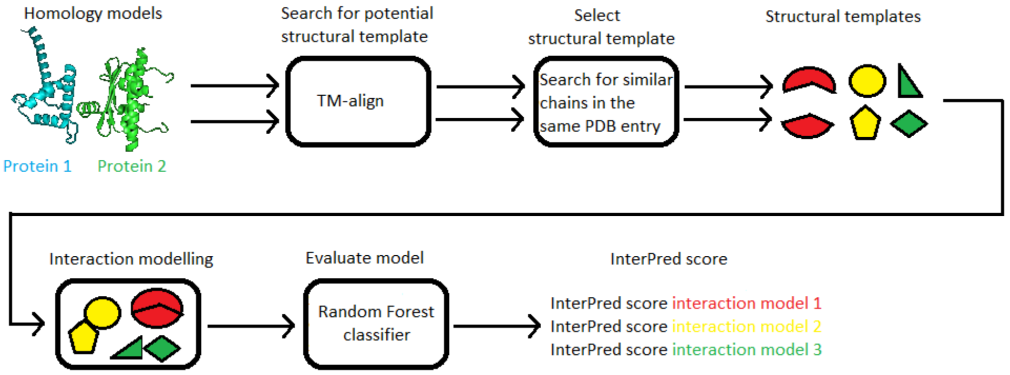

Homology models
Protein 1
Protein 2
Search for potential structural template
TM-align
Select structural template
Search for similar chains in the same PDB entry
Structural templates
Interaction modelling
Evaluate model
Random Forest classifier
InterPred score
InterPred score interaction model 1
InterPred score interaction model 2
InterPred score interaction model 3

Figure 4. Overview of phase 2 in the InterPred pipeline.

The homology models from phase 1 are used to identify structural templates through structural alignments with TM-align [30] against every chain in the PDB [12]. From the potential structural templates, obtained by TM-align, only the templates where both target templates have similarities with chains from the same PDB entry are selected as structural templates [12]. Each pair of structural templates is then used in the interaction modeling step, where the coarse PPI models are generated [12]. Since the quality of the coarse PPI models will vary, a scoring function is needed in order to rank the coarse PPI models and select the best ones.

The coarse PPI models are scored by a Random Forest classifier trained on features, giving one of two classes, Yes/No interaction, to predict the probability that the coarse PPI model interact [12] (InterPred score). The features used for training the Random Forest classifier could be calculated directly from the PPI models and are derived from the model quality, structural alignment and interface [12]. Model quality measures the quality of the homology models [12]. The structural alignment feature defines the structural alignment quality between the structures of both targets and their respective structure templates, and the feature for interface describe the similarity between the structural template interfaces and the coarse PPI model's interface [12]. InterPred filter all models with an InterPred score under 0.5 because they have a low probability of generating high quality PPI models [12].

3.2.3 Phase 3 - Refinement

The coarse PPI models might include severe clashes [12], i.e. overlap between the two interacting proteins. Since clashes suggest that the coarse PPI model is not entirely realistic a refinement step is useful for trying to reduce the clashes in order to find more realistic models, which is the third and last phase in the InterPred pipeline, *Figure 5*.

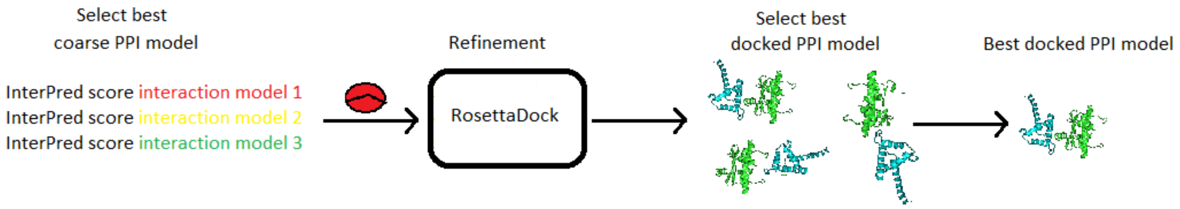

Select best
coarse PPI model
InterPred score interaction model 1
InterPred score interaction model 2
InterPred score interaction model 3
Refinement
RosettaDock
Select best
docked PPI model
Best docked PPI model

Figure 5. Overview of phase 3 in the InterPred pipeline.

InterPred uses RosettaDock [31] to refine the selected coarse PPI model [12].

3.3 ProQDock

ProQDock is a program using SVM to predict PPI models quality, ProQDock score [9]. The ProQDock score is the estimated DockQ score, which indicates the quality of the PPI model [15]. However, in order to calculate the DockQ score the true structure, as well as the predicted PPI model are needed which is the reason for ProQDock to predict an estimated value for DockQ score instead of calculating it. The Pearson's correlation between DockQ and IS score is 0.98 [15].

The predicted DockQ score, ProQDock score, is predicted from thirteen features calculated directly from the PPI models [9]. ProQDock's features describe different aspects of a PPI in order to try to give accurate prediction of the PPI model's quality. Description of the features will be discussed in detailed in *Section 3.3.1-3.3.9*.

3.3.1 Shape complementarity

A necessary condition for protein-protein binding is shape complementarity (Sc) at the protein interface [13]. Sc is defined as how complementary both proteins interfaces are physically. For example, a good Sc could be the match of a concave protein interface with a convex protein interface [32], *Figure 6*.

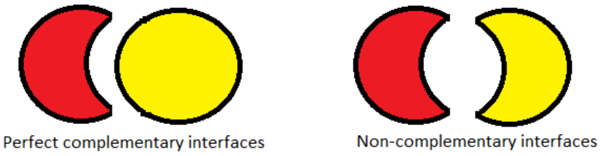

Perfect complementary interfaces
Non-complementary interfaces

Figure 6. Illustration of perfect complementary interfaces and non-complementary interfaces.

ProQDock uses the Sc program from the CCP4 package [33] to calculate shape complementarity for the protein interface. Sc is represented by a score between -1 and 1, where -1 represent non-complementarity interfaces and 1 correspond to perfect complementary between the two proteins interfaces [9].

3.3.2 Electrostatic complementarity

Electrostatic complementarity (EC) measure if the electrostatic potentials of the protein interfaces are complementary, since, the EC has effect on the stability of the native protein-protein complex [32] [34] [35] [36]. A complementary EC can be visualized as the attraction between a positive and a negative charge.

Calculating EC is time consuming and obtained with DelPhi [37] which calculate the electrostatic potential for each point at the surface of the interface by solving the linearized Poisson-Boltzmann equation iteratively [9]. The score for EC is ranked between -1 and 1, where a higher score indicates higher complementarity in the electrostatic potential for the two interfaces [9].

3.3.3 Size of the interface

The size of interfaces varies between different PPIs [38] [39]. Since a smaller interface easier generate a higher Sc and EC than a larger interface, that has to take more residues into account, size of the interface should be taken into consideration when evaluating Sc and EC. ProQDock measures the size of the interface in two different ways, through normalized buried surface area (nBSA) and fraction of residues buried at the interface (Fintres) [9].

nBSA measure the fraction of exposed surface area that becomes buried due to binding, while Fintres measure the fraction of buried interface residues [9].

3.3.4 Probability of finding interface within a specific distance

As mentioned in the *Section 3.3.3* above, Sc and EC should be considered in proportion to the size of the interface. ProQDock measures the combined probability of finding the interface within a specific distance from Sc and EC given nBSA [9]; (CPM). In order to calculate CPM a reference in terms of information about nBSA, Sc and EC from native structures in a database is necessary. ProQDock constructs the reference through three steps [9]:

1. Calculation of nBSA, Sc and EC for 1 879 native structure from the database.
2. Categorizing the interfaces into small, medium and large based on the distribution of nBSA.
3. Categorizing the Sc and EC based on distribution into intervals of 0.005.

ProQDock uses the reference in order to calculate the probability of finding the interface within a specific distance of Sc and EC, Equation (6). Where Equation (4) and (5) are the probability of finding the interface within a specific distance of Sc and EC respectively [9].

$$P(Sc|nBSA) \tag{4}$$

$$P(EC|nBSA) \tag{5}$$

$$CPM = \log(P(Sc|nBSA)) + \log(P(EC|nBSA) \tag{6}$$

In order to avoid Equation (6) to be undefined, ProQDock defines CPM as 0 if either Equation (4) or (5) is zero [9].

3.3.5 Link density

An interesting feature to note for protein-protein interaction is the number of contacts between residues at the interface of the two proteins. ProQDock measures the number of contacts between the two proteins through link density (Ld) [39]. Ld is defined as the ratio between actual contact points, links, and the theoretical maximum number of links [9] [39]. In ProQDock a link is defined as any couple of heavy atom from different residues within 6 Å [9], where the residues have to belong to different proteins. The theoretical maximum number of links are calculated through Equation (7), where A and B represent the number of residues found at the interface of the interacting proteins respectively [9]. Note, Ld measure ratio of contact points between the two proteins, unfortunately, these contact points can be located outside the interfaces if the model is constructed such as the interacting parts are located outside the interfaces, i.e. poorly constructed model.

$$Theoretical\ maximum\ number\ of\ links = A * B \qquad (7)$$

3.3.6 Interface contact preference score

Some inter-residue contact is preferred at the interface [38] and various residues have different effect on single protein or PPI [40]. Therefore, the number of specific contact between residues at the interfaces are interesting to study. Specific contact refers to interactions between residues of a certain sort, e.g. Ala-Ala or Glu-Asn and so on. ProQDock measures the number of times each specific contact exist between the two interfaces and the result is weighted by literature values [41] for the contact preference [9]. The contact score was used to train an SVM to predict the ProQDock score by the interface preference score alone [9]; (CP score).

3.3.7 Burial of residues

If a model has a lot of e.g. hydrophobic residues at the surface the model is probably incorrect, since hydrophobic residues are more usual in the core than the surface [36] [38]. By measuring the burial of residues at the protein complex surface an indication of the quality of the model can be obtained. Given a particular solvent accessibility, ProQDock's feature burial residues (rGb) measure the propensity of a specific residue type with that specific solvent accessibility [9]. Note, rGb are measured for the entire surface i.e. not only for the interfaces.

3.3.8 Structural model accuracy

A feature for estimating the accuracy of the protein-protein structure is necessary to examine how realistic the complex structure is. In ProQDock the accuracy for the structural complex is estimated by the program ProQ2 [42], which uses structural and sequence based features to predict the accuracy of the structural model through SVM [43]; (ProQ).

3.3.9 Rosetta energy terms

Rosetta is a molecular modeling software [44] [45] that scores molecular conformations in order to identify optimal energy functions [45] [46], which in turn calculate the energy terms. ProQDock uses four Rosetta energy terms: Rosetta total score (rTs), interface energy (Isc), van der Waals repulsive term (E_{rep}) and the Rosetta total score minus the van der Waals repulsive term (E_{tmr}), as features in order to predict the quality of PPI models. The lower the value of the energy term, the better the PPI model.

rTs is Rosetta's most common energy term and represents all Rosetta's energy terms merged into one score. Isc is the distribution of intermolecular bonds at the interface, in other words, the excess of binding energy [9] at the interface, indicating how willingly the proteins are to interact. E_{rep} represent the steric repulsion of van der Waals forces. It is only calculated if the interatomic distance between two atoms is less than the sum of van der Waals radii for the atoms' and if the torsion angles of a single residue not affect the interatomic distance [47]. E_{rep} can give an indication if the proteins in the model clash instead of being placed side by side, where overlap suggests that the model is not entirely realistic. Small overlaps might be resolved through refinement of the PPI model. Since even minor clashes can have a high impact on the rTs [9] it is, therefore, essential to study the E_{tmr}, which is the total energy score without the contribution of E_{rep}, in order to allow some clashes in the PPI models.

4. Methods

This section describes the procedures for the work process.

4.1 PPI models

The 30 000 coarse PPI models used were previously generated, obtained from the InterPred study [12], through the InterPred pipeline for a set of protein-protein complexes from the Protein Docking Benchmark 4 [48].

In the IntePred study [12] only some of the best coarse PPI models for each target underwent refinement, generating 10 000 refined PPI models for each coarse PPI model [12]. Out of these refined PPI models, only the best for each coarse PPI model was selected to represent the coarse PPI model after refinement. The best refined PPI model was defined as the one with the smallest difference in interface between the coarse- and refined PPI model [12].

4.2 Test- and training sets

The 30 000 coarse PPI models were divided into 5 clusters for 5-fold cross-validation, where the coarse PPI models were clustered based on interface similarity by the program iAlign. In order to identify all similar coarse PPI models the threshold in iAlign was set to the P-value 1×10^{-3}, and all combinations of coarse PPI models with P-value under or equal to the threshold were considered similar and grouped in the same cluster.

All coarse PPI models went through the program ProQDock.
Each cluster was transformed into a matrix, test set, containing values for all feature subtracted from the ProQDock's result files, complemented with the InterPred score from the InterPred pipeline and the true IS score.

Some coarse PPI models were not compatible with the program ProQDock and were therefore removed from the test set which resulted in a total number of 28 837 coarse PPI models, where each test set contained between 5 620 and 5 830 coarse PPI models. These test sets were referred to as Coarse test- and Coarse training set.

Since only the best coarse PPI models for each target had undergone refinement some coarse PPI models in the Coarse test sets had no corresponding refined PPI model. In order to predict the refined IS score the test sets were revised so they also contained the refined IS score, Refined test- and Refined training sets. The coarse PPI models that did not have a corresponding refined PPI model was removed. Resulting in a total of 1 595 coarse PPI models, where each Refined test set contained between 190 to 420 coarse PPI models.

Training was performed through 5-fold cross-validation where each test set had its complementary training set containing of the remaining test sets, e.g. test set 1 and its complementary training set 1 containing test set 2, 3, 4, 5. The predictor for predicting the IS score was trained on the Coarse training set and tested on the Coarse test set, while the predictor for predicting the refined IS score was trained on the Refined training set and tested on Refined test set.

4.3 Random Forest regressor

Random Forest regressor version 0.18.1 from scikit learn's Python module was used to create the predictor. The parameters stated were random state set to 42, verbose 1, the number of jobs was set to 5, and the number of trees and max depth was optimized through the program GridSearchCV. Since, the predictor used features from both ProQDock and InterPred, e.g. a combined version of ProQDock and InterPred, this method was called cProQPred.

4.3.1 Grid SearchCV

Max depth and the number of trees used in the cProQPred predictor was optimized through scikit learn's program GridSearchCV with 5-fold cross-validation. The combinations tested were all number of trees between 100-700 with steps of 50 and max depth 5-70 in steps of 5.

The parameter combinations were ranked based on its cross-validation score, and the Pearson's correlation between the true IS and predicted IS score for each parameter combination was noted. A statistic t-test with a 95 % confidence was used to investigate if the correlation from cProQPred predictors with different parameter settings were significantly different. When two correlations were considered to be not significantly different the parameter setting with the lowest number of trees and the minimum max depth was selected. For further experiments, parameter setting 100 trees and a maximum depth of 10 was selected. The Coarse test sets and Coarse training sets were used for cross-validation.

4.4 Feature importance

The importance of different features for predicting the IS score was obtained through the feature importance function in Random Forest. Given each feature its importance in percentage based on Coarse training sets.

4.5 Predictor evaluation

Several versions of the cProQPred predictor were trained with different cutoff point for feature importance, i.e. all features with an importance over the cutoff was included. The different cProQPred predictors were evaluated with Pearson's correlation between the true IS and predicted IS score, obtained through the function pearsonr in SciPY's version 0.14.0. The correlation for each Coarse test set and an overall correlation for all Coarse test sets was calculated.

The TP, FP, TN and FN were calculated through sweeping a cutoff between 0 and 1, classifying all predicted IS score over the cutoff as predicted interactions. A true IS score equal or over 0.12 was classified as a true interaction, and the TP, FP, TN and FN were calculated through comparing the predicted interaction/no interaction with the truth for each cutoff point.

4.5.1 ROC curve

The TPR and FPR for different relevant cProQPred predictors, ProQDock and InterPred were calculated and presented in a ROC curve for performance evaluation.

4.5.2 Precision and recall

The precision and recall was calculated for different relevant cProQPred predictors, ProQDock and InterPred, and visualized in a graph.

4.6 Predicting refinement quality

Predicted the quality of a refinement, i.e. how realistic a generated refined PPI model would be. The quality of a refinement was obtained through predicting the refined IS score, *Section 4.6.1*, and chance of succeeding at refinement, *Section 4.6.2*. The cProQPred predictor created in *Section 4.3-4.3.1* was used.

4.6.1 Refined IS score

Predicted the presumed refined IS score a coarse PPI model would get if it had undergone refinement. Refined test sets and Refined training sets were used for cross-validation.

4.6.2 Chance of succeeding at refinement

Predicted the chance for a coarse PPI model to succeed at refinement by predict the difference between the IS score of the coarse model and its refined IS score. Refined test sets and Refined training sets were used for cross-validation.

4.7 Correlation - IS score and refined IS score

The relationship between the IS score, refined IS score and the predicted scores from different relevant cProQPred predictors was investigated through Pearson's correlation and scatter plots. Investigated relationships presented in *Table 2*.

Table 2. The relationships investigated through scatter plots and Pearson's correlation.

True IS score	True refined IS score
True IS score	Predicted IS score
True refined IS score	Predicted refined IS score
True refined IS score	Predicted IS score
True IS score	Predicted refined IS score
Predicted IS score	Predicted refined IS score

5. Results

This section presents the results for all parts of the methods.

5.1 GridSearchCV

The cross-validation score from GridSearchCV and the Pearson's correlation between the true IS and predicted IS score for each parameter combination is presented in *Figure 7*. A t-test with 95 % confidence based on the correlation from the predictor with the best ranked parameter combination generated the interval 0.7804 – 0.7695, for within the correlation is not significantly different to the correlation obtained from the predictor with the best ranked parameter combination. The upper and lower limit of the confidence interval is presented in *Figure 7* by the yellow lines.

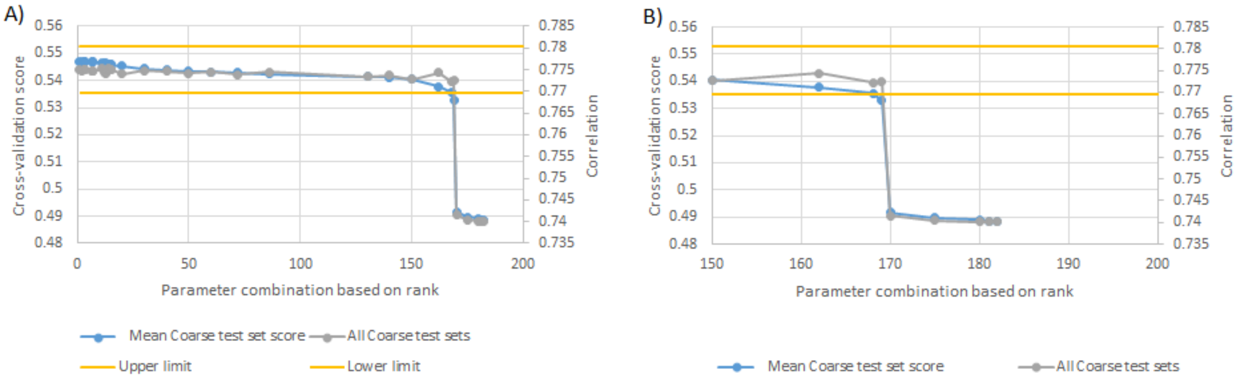


Figure 7. Graphs over the cross-validation score, blue graph, and correlation, gray graph, for each parameter combination investigated. The yellow lines display the upper- and lower limit for the 95 % confidence interval for correlation to not be significantly different to the correlation for the best ranked parameter combination. The cross-validation score is read on the primary y-axis and correlation and the upper- and lower limit for the confidence interval is read on the secondary y-axis. A) The entire graph. B) Zoom in on the critical area.

Figure 7.B) illustrate that the correlation for all parameter combinations with a lower rank than 170 are within the confidence interval, yellow lines. Therefore, all parameters combinations with a lower rank than 170 were sorted based on the number of features and max depth. The combination with the lowest number of trees and minimal max depth was selected as the optimal parameter setting, giving the parameter setting 100 trees and max depth 10.

5.2 Feature importance

Each features' importance is presented in *Table 3* and *Figure 8*.

Table 3. Table presenting all features and their respectively importance.

Feature	**Importance (%)**
Sc	21.04
EC	6.09
rGb	3.52
Ld	1.6
nBSA	19.85
Fintres	3.01
CP score	1.65
CPM	0.99
rTs	0.0
Isc	0.13
E_{rep}	0.25
E_{tmr}	0.34
ProQ	5.46
ProQDock score	1.08
InterPred score	34.99

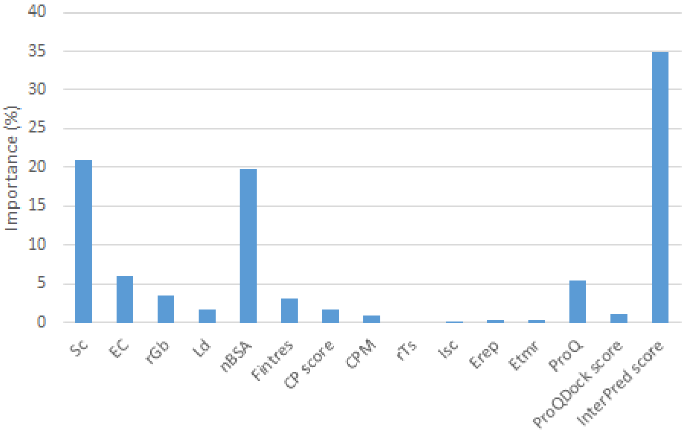


Figure 8. Each features' importance presented in percent.

Figure 8 illustrate that only a few features have a high importance, where Sc, nBSA and the InterPred score have by far the highest importance while five features: the energy terms rTs, Isc, E_{rep}, E_{tmr} and CMP, have an importance under 1 %.

5.3 Predictor evaluation

The correlation between the true IS and predicted IS score from different cProQPred predictors, containing features with an importance over a specific cutoff point, is presented in *Figure 9*.

The 95 % confidence interval for evaluating if the different cProQPred predictors correlation is not significantly different to the correlation obtained by the cProQPred predictor using all features, generated the interval 0.7669-0.7780, yellow lines in *Figure 9*. All predictors with a correlation between these two lines are not significantly different to the predictor using all features.

The different cProQPred predictors will hereby be referred to as cProQPred F > the cutoff point for feature importance, e.g. cProQPred F >10 for the cProQPred only trained on features with an importance over 10 %. The cProQPred trained on all features will be referred to as cProQPred F >0.

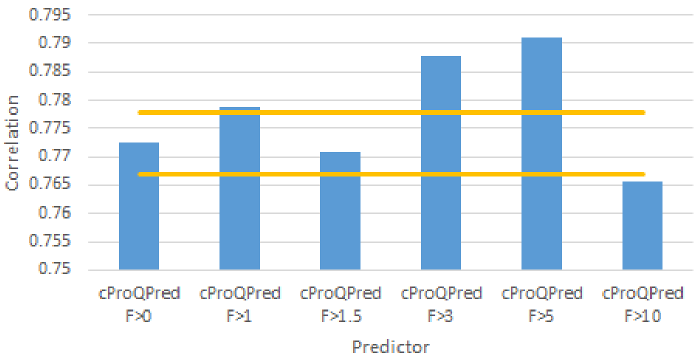


Figure 9. Correlation between the true IS and predicted IS score from predictors only trained on features with an importance over different cutoff points.

Figure 9 illustrate that cProQPred F >3, F >5 are above the upper limit of the confidence interval, and cProQPred F >1 are located just above the upper limit. cProQPred F >1.5 are within in the interval and cProQPred F >10 are just under the lower limit of the confidence interval.

To visualize the different Coarse test sets impact on correlation, *Table 4*, the correlation between the true IS and predicted IS scores for each Coarse test set in the cProQPred F >0 – F >10 predictors were calculated, illustrated in *Figure 10*. The correlations for the Coarse test sets within each cProQPred predictors was investigated with a t-test, 95 % confidence, to see if the correlation was significantly different between Coarse test sets.

Table 4. Table over the correlations for all Coarse test sets from different cProQPred predictors.

Test set	**cProQPred F > 0**	**cProQPred F > 1**	**cProQPred F > 1.5**	**cProQPred F > 3**	**cProQPred F > 5**	**cProQPred F > 10**
Coarse test set 1	0.6143	0.6320	0.6377	0.6508	0.5648	0.6692
Coarse test set 2	0.7851	0.7855	0.7820	0.7756	0.8133	0.7804
Coarse test set 3	0.8983	0.9062	0.8913	0.9206	0.9218	0.8956
Coarse test set 4	0.6535	0.6603	0.6402	0.7149	0.7494	0.6996
Coarse test set 5	0.7466	0.7462	0.7479	0.7513	0.7507	0.7495
All Coarse test sets	0.7724	0.7789	0.7709	0.7878	0.7910	0.7656

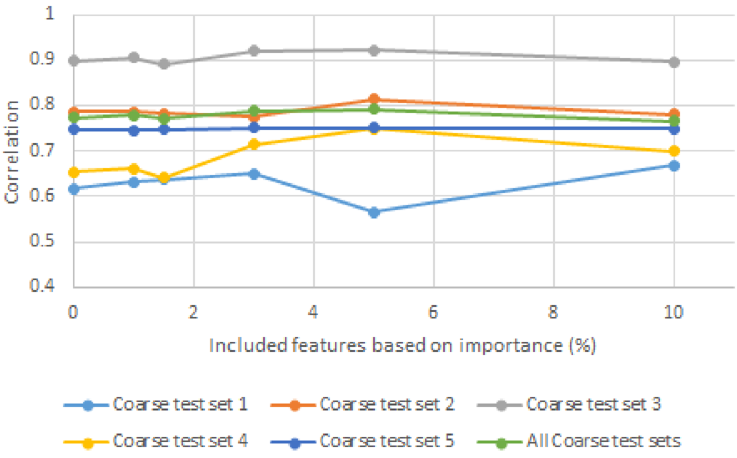


Figure 10. Graph over the correlations between true IS and predicted IS score for all Coarse test sets from different cProQPred predictors.

The result from the t-test illustrated that all cProQPred predictors include Coarse test sets which generate significantly different correlations. Regardless of which cProQPred predictor used, Coarse test set 1 always generate the lowest correlation and Coarse test set 3 always generate the highest correlation, *Table 4*.

The highest correlation for all Coarse test sets, except Coarse test set 1, are obtained for the cProQPred F >5, *Figure 10*. While the difference in correlation between the Coarse test sets is smallest for cProQPred F >10. For this reason, the cProQPred F >10 and F >5 are interesting for predictor performance evaluation together with cProQPred F >0 as a reference.

5.3.1 ROC curve

ROC curve for comparison of the cProQPred F >0, F >5 and F >10 predictors and how these perform compared with ProQDock and Interpred, *Figure 11*.

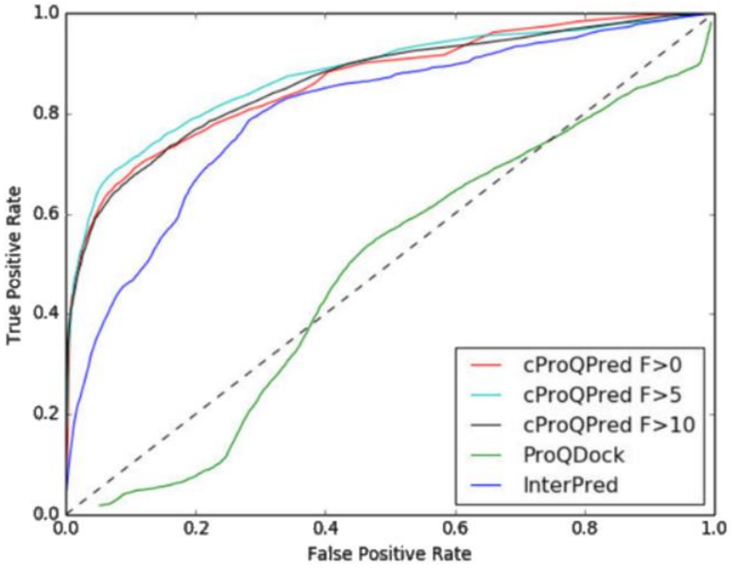


Figure 11. ROC curve for comparing the performance of cProQPred F >0, F >5, F >10, ProQDock and InterPred.

All cProQPred predictors, red-, black- and cyan graphs in *Figure 11*, finds more positives for the same number of false positives then both InterPred and ProQDock. ProQDock lies around the dashed line which represent random classification of interaction or non-interaction. A zoom in on relevant areas is presented in *Figure 12*. Scatter plots for visualization of the relationship between the true IS and predicted IS scores for the cProQPred predictors and true IS and predicted label for ProQDock and InterPred is presented in *Figure 13*.

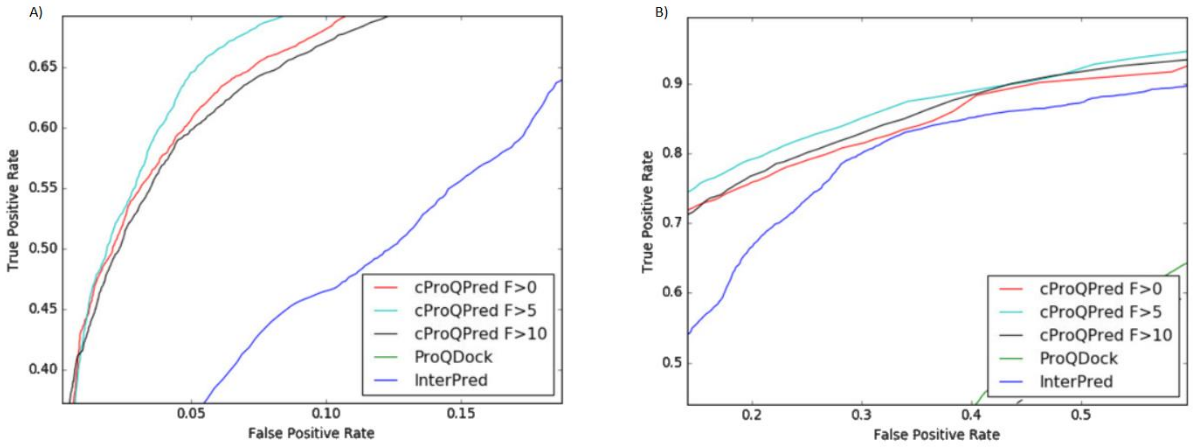


Figure 12. Zoom in of the ROC curve on the area: A) where the cProQPred predictors deviates from the y-axis, B) visualizing the performance of the cProQPred predictors and InterPred.

Figure 12 illustrate that cProQPred F >5 generate the highest true positive rate with the lowest false positive rate. However, the other cProQPred predictors perform almost the same as cProQPred F >5, and in the later part in *Figure 12. B)* InterPred also exhibit similar performance.

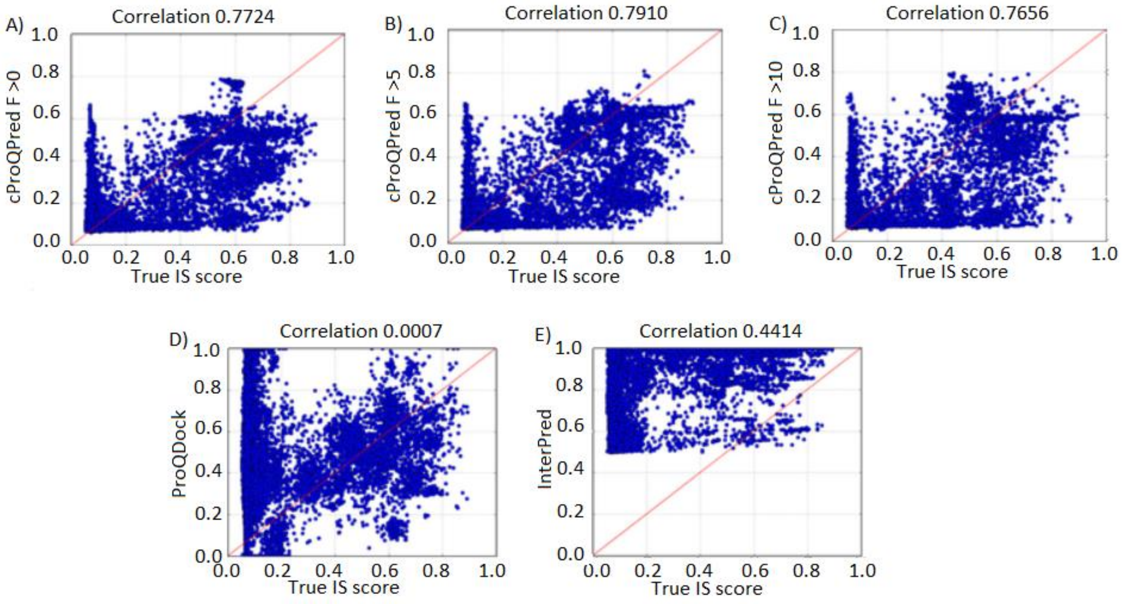


Figure 13. Scatter plots visualizing the relationship between the true IS score and predicted label generated from: A) cProQPred F >0, B) cProQPred F >5, C) cProQPred F>10, D) ProQDock.and E) InterPred

The scatter plots for cProQPred, *Figure 13.A)-C)*, are quite similar and centered around the red line which represents a perfect prediction of the IS score. Their respective correlation is also quite similar. ProQDock only has a correlation just above zero, *Figure 13.D)*, mostly because of many high scoring false positives. InterPred filters InterPred score under 0.5 resulting in a lower correlation than all cProQPred predictors, *Figure 13.E)*.

5.3.2 Precision and recall

The relationship between precision and recall for cProQPred F >0, F >5, F >10, ProQDock and InterPred are presented in *Figure 14*.

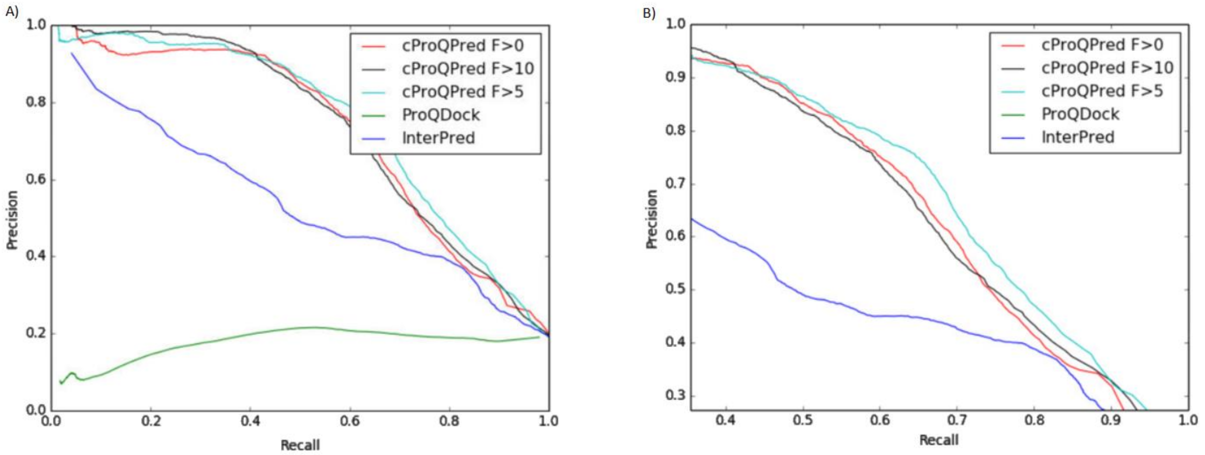


Figure 14. A) Graphs visualizing the relationship between the precision and recall for the investigated methods and B) zoom in on the relevant area.

The cProQPred predictors have both the highest precision and highest recall, while InterPred has more a linear relationship between precision and recall with lower values for both. ProQDock, however, never generate a higher precision than just above 0.2.

5.4 Predicting refinement quality

This section presents the results from predicting the refinement quality.

5.4.1 Refined IS score

The correlation between the true refined IS and predicted refined IS score is presented in *Figure 15*.

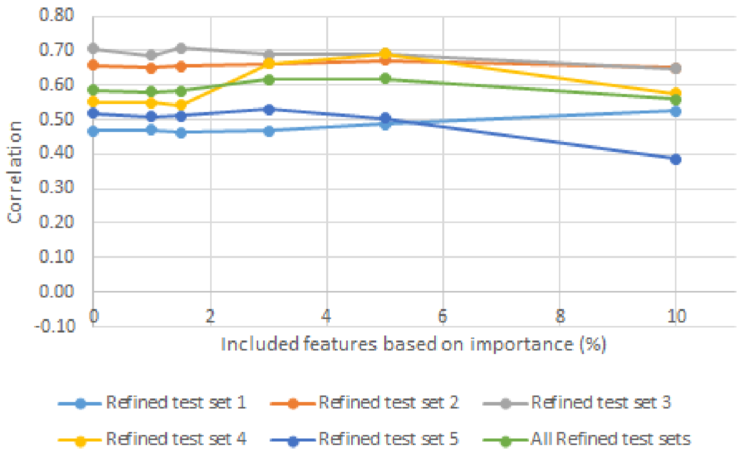


Figure 15. Graph over the correlation between the true refined IS and predicted refined IS score for all Dock test sets from different cProQPred.

The correlations for all Refined test sets from cProQPred F >0 – F >5, and all Refined test sets except Refined test set 5 for cProQPred F >10 have a correlation over 0.5, *Figure 15*. The Refined test set from cProQPred have the correlation closest to each other.

5.4.2 Chance of succeeding at refinement

The correlation between the prediction of the difference in IS score, refined IS score and true difference are presented in *Figure 16*.

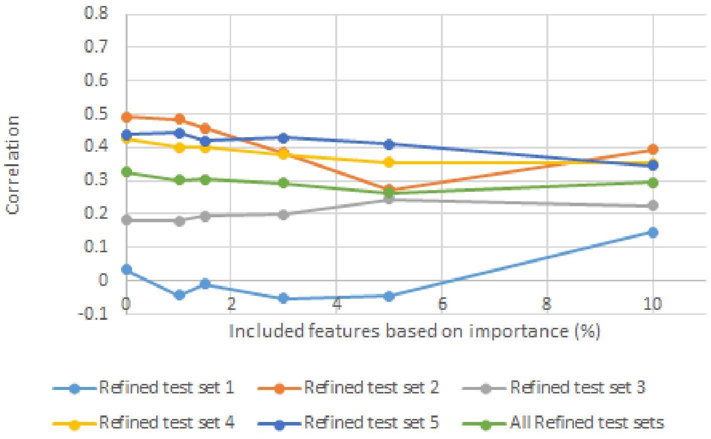


Figure 16. Graph over the correlation between the predicted difference in IS, refined IS score and real difference for all Dock test sets from different cProQPred predictors.

Only Refined test set 2, 4 and 5 have a correlation over 0.4 for cProQPred F >0 - F >1.5, and Refined test set 5 for cProQPred F >5. However, the correlation is low for all predictors when all Refined test sets are combined, green graph *Figure 16*.

5.5 Correlation - IS score and refined IS score

The correlation between all possible combinations of true IS, true refined IS, predicted IS and predicted refined IS score, all generated from cProQPred predictors, is presented in *Table 5* and *Figure 17*. For scatter plots visualizing the relationship between all combinations see *Appendix A*.

Table 5. Table over the correlation between all possible combinations of true IS, true refined IS, predicted IS and predicted refined IS score.

x-axis	y-axis	Included features	Correlation	ID number
True IS score	True refined IS score		0.8491	1
True IS score	Predicted IS score	> 10 %	0.7656	2
True IS score	Predicted IS score	> 5 %	0.791	3
True refined IS score	Predicted refined IS score	> 10 %	0.5583	4
True refined IS score	Predicted refined IS score	> 5 %	0.619	5
True refined IS score	Predicted IS score	> 10 %	0.5615	6
True refined IS score	Predicted IS score	> 5 %	0.6137	7
True IS score	Predicted refined IS score	> 10 %	0.5883	8
True IS score	Predicted refined IS score	> 5 %	0.6489	9
Predicted IS score	Predicted refined IS score	> 10 %	0.9191	10
Predicted IS score	Predicted refined IS score	> 5 %	0.9154	11

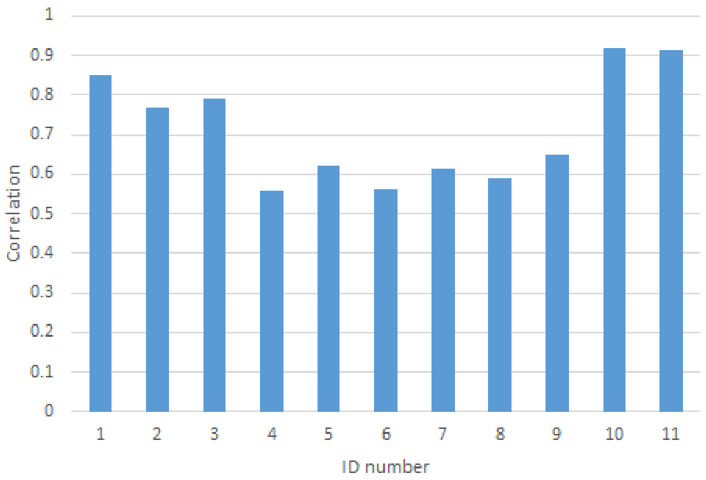


Figure 17. Correlation between all possible combinations of true IS, true refined IS, predicted IS and predicted refined IS score.

The best correlations are obtained between the predicted IS and predicted refined IS score, true IS and true refined IS score and true IS and predicted IS score, which all generate correlations between 0.766 - 0.919, *Table 5*.

A t-test with 95 % confidence illustrate that the best correlation with the true IS score is obtained with the predicted IS score from the cProQPred F >5. The best correlation with the true refined IS score is obtained with the predicted refined IS score and predicted IS score both from cProQPred F >5. This because it is no significant difference in correlation between the true refined IS and predicted refined IS score and true refined IS and predicted IS score.

6. Discussion

This section includes discussion of main results, work process, impact in a broad sense and future perspectives.

6.1 Results

The main results are discussed in *Section 6.1.1 - 6.1.5*.

6.1.1 Feature importance

There was only a few features with a high importance, *Figure 8*. This indicates that only some features have a high impact when predicting the IS score. The most important features were the InterPred score, Sc and nBSA with an importance of 34.99, 21.04 and 19.86 % respectively. EC, 6.09 %, and ProQ, 5.46, also had an importance over 5 %.

It is interesting that both the InterPred score and ProQ have a high impact on the prediction of IS score since they are relative similar features. The InterPred score indicated the chance for a coarse PPI model to interact in reality [12] and ProQ indicates the quality of the coarse PPI model [9]. However, it is not surprising that the coarse PPI model's quality would have an impact when predicting the IS score since the IS score actually indicates how well the model match the native protein-protein complex.

Five features; CPM, rTs, Isc, E_{rep}, E_{tmr}, had an importance under 1 %, indicating that these features would not affect the predicted score. This might be because of other features contribute equally to the same result. Since decision trees start splitting the training set on features into nodes based on their contribution to the predicted label [25], when two features contribute equal to the same result, the feature which is presented first in the training set will be selected for splitting the data giving it a higher importance than the other feature, thus end up further down in the decision tree. Out of the five features with an importance under 1 %, all energy terms were found and another theory for their low importance might be that calculation of energy terms are more suitable for refined PPI models than coarse PPI models.

6.1.2 Predictor evaluation

The correlation between the true IS and predicted IS score from different cProQPred predictors, *Figure 9*, illustrateed that cProQPred F >3 and F >5 was significantly better than cProQPred F >0, with correlations of 0.79, 0.78 and 0.772 respectively. The correlation is fairly high which means that these predictors can predict the IS score well.

cProQPred F >10 was significantly inferior to cProQPred F >0, with a correlation of 0.766. Even if the correlation was significant inferior to cProQPred F >0, it was still high.
An advantage with this predictor is that it does not include the feature EC, which is a time consuming feature. It, therefore, becomes a balance between obtaining best possible correlation and computational time.

The Coarse test sets for all cProQPred predictors, *Figure 10*, generated correlations which were significantly different. Regardless of cProQPred predictor, Coarse test set 3 always generated the highest correlation and Coarse test set 1 the lowest. In order to avoid overfitting, cross-validation [17] was used, where all similar coarse PPI models were grouped into the same cluster [9], Coarse test set. Since coarse PPI models with high resemblance were clustered into the same Coarse test set, this might result in some Coarse test set constitutes of more challenging coarse PPI models to predict the IS score for, resulting in lower correlation for the Coarse test set.

When clustering the coarse PPI models based on interface similarity, by iAlign, the threshold was 1×10^{-3}. Since a threshold of 1×10^{-4} results in 0.01 % FP classification of the interfaces [21] a higher threshold would minimize the risk of clustering similar PPI models into different Coarse test sets. Therefore, there is no reason to believe that the overall high correlation for all cProQPred predictors is generated due to overfitting.

ROC curve, Precision and recall
The ROC curve, *Figure 11*, for comparing the performance of cProQPred F >0, F >5, F >10 with ProQDock and InterPred, illustrated that all cProQPred predictors performed better than both ProQDock and InterPred. This result was visualized more clearly in the precision, recall graph, *Figure 14*, where cProQPred F >5 had the highest precision and recall. However, the other cProQPred predictors were located just under the cProQPred F >5, indicating that these perform quite similar. The investigation of feature importance illustrated that both some features from ProQDock and the InterPred score had a high impact when predicting the IS score. It was, therefore, expected that cProQPred generated a higher performance than the individual programs.

The ROC curve illustrated that ProQDock performed the same as random classification. This was explained by the scatter plot, *Figure 13.D)*, which illustrated that ProQDock include a lot of outliers. The reason for this inferior results for ProQDock was probably due to the reason that ProQDock was trained on refined PPI models [9] and not on coarse PPI models as cProQPred.

InterPred had no predicted InterPred scores under 0.5, *Figure 13.E)*, because InterPred filtered all coarse PPI models with an InterPred score under 0.5, since they had a low probably of generating high quality coarse PPI models [12]. This effect the correlation between the true IS and InterPred score, and was probably the reason for InterPred's inferior correlation.

6.1.3 Predicting refinement quality

Since the Refined test sets only include between 190 - 420 coarse PPI models while the Coarse test sets include between 5 620 - 5 830 coarse PPI models, the predictors for predicting the refined IS score will be trained on fewer coarse PPI models. This might affect the predictions and moreover the correlation between the true refined IS and predicted refined IS score. The small size of the Refined test sets can, therefore, be considered as a weakness.

Refined IS score
The correlation between the true refined IS and predicted refined IS score for all test sets combined was around 0.6, *Figure 15*, for all cProQPred predictors. This classifies as a fairly good result considering it predicts the presumed refined IS score a coarse PPI model would get if it had undergone refinement.

Chance of succeeding at refinement
The correlation between the predicted difference in IS, refined IS score and true difference, *Figure 16*, had a low correlation when all Dock tests were combined, and the correlations for each Refined test set was significantly different. The spread in correlations for the different

Refined test sets was expected because of the spread in correlation for the different Coarse test sets. The distribution of correlation was, however, more spread for the Refined test sets which might be because they include fewer coarse PPI models.

Refined test set 1 always generated the lowest correlation, as well as Coarse test set 1 always generated the lowest correlation. However, the highest correlation was generated for different Refined test sets, Refined test set 2 and Refined test set 5, which differs from the prediction of IS score were Coarse test set 3 always generated the highest correlation. This difference might be because Refined test set 3 was the smallest Refined test set and perhaps included more challenging coarse PPI models for refined IS score prediction.

The correlation for all Refined test sets combined was low which probably have something to do with information loss. Since the cProQPred predictors performed good when predicting the IS score and fairly good when predicting the refined IS score, the correlation for predicting the difference in IS, refined IS score and real difference was expected to be ok. However, the correlation was low, indicating that some information was lost.

6.1.4 Correlation - IS score and refined IS score

The true IS and true refined IS score had a correlation of 0.849, *Table 5*. This high correlation was probably the consequence of the definition of best refined PPI model, which was chosen to represent the coarse PPI model after refinement. The best refined PPI model had the smallest difference in interface from the coarse PPI model [12], the IS and refined IS score should, therefore, be quite similar resulting in a high correlation.

A t-test with 95 % confidence illustrated that the best correlation with the true IS score was obtained with the predicted IS score from cProQPred F >5. The best correlation with the refined IS score was obtained with the predicted IS score or predicted refined IS score from cProQPred F >5.

The true refined IS score can be estimated through either the predicted IS score or refined IS score. Since the predicted IS and predicted refined IS score had a correlation of 0.919, from cProQPred F >5, indicating that the predicted IS score and predicted refined IS score was practically the same. Making the original predicted IS score a relative good prediction of the refined IS score.

6.2 Conclusions

The first aim of this master thesis was to create a Random Forest predictor that could predict the IS score for coarse PPI models based on features pre-calculated. This has been achieved through creating a Random Forest predictor, cProQPred, using some features generated from the program ProQDock and the InterPred score from the InterPred pipeline. The cProQPred predictor had a better performance than both ProQDock and InterPred.

Which features from ProQDock, that should be included in cProQPred was evaluated through their respectively importance for predicting the IS score. The best predictor considering correlation between true IS and predicted IS score was cProQPred F >5, only trained on features with an importance over 5 %. However, cProQPred F >5 include the time consuming feature EC. The best predictor considered the time for calculating features would, therefore, be cProQPred F >10, only trained on features with an importance over 10 %.
cProQPred F >10 generate an overall correlation of 0.766 and is, therefore, a fairly good method for predicting the IS score.

The second aim was to create a predictor that could predict the quality of a refinement. This was conducted in two different ways; by predicting the refined IS score and predicting the chance for a coarse PPI model to succeeding at refinement.

Prediction of the refined IS score with the cProQPred F >5 generated an overall correlation of 0.619 between the true refined IS and predicted refined IS score. However, the true refined IS score could be estimated equally good with the predicted IS score. The correlation between the predicted IS and predicted refined IS score was 0.919, for cProQPred F >5. The high correlation indicated that the predicted IS and predicted refined IS score was practically the same. A relative good prediction of the refined IS score can, therefore, be made by the original predicted IS score.

A coarse PPI model's chance of succeeding at refinement was predicted by the difference in IS and refined IS score, which overall generated a low correlation. Since cProQPred performed good when predicting the IS score and fairly good when predicting the refined IS score, the predicted difference was expected to have an ok correlation. The correlation was, however, low indicating a loss in information.

A limitation of cProQPred is that it uses pre-calculated features from the programs ProQDock and InterPred which makes it dependent on these programs. It would be preferable if cProQPred could calculate the necessary features directly from PPI models instead of generating them from other programs.

6.3 Analysis of the work process

The work process has functioned well thanks to continuously checkups and discussions with the examiner. Since this type of projects largely concerns solving problems which occur along the way and partial results determine the next step, a timetable was quite hard to design. During this project it has, therefore, been important to be open to chance, with a clear vision on the aim and what should be achieved.

Only a week into the project the aim changed from scoring coarse PPI models quality by creating a predictor trained on a few features which the predictor should calculate by itself to instead create a predictor trained on features from ProQDock. The new aim first focused on retraining ProQDock on the coarse PPI models but later developed to a predictor trained on features from both ProQDock and InterPred. This led to more changes in the timetable and repeating previously conducted steps. The new aim, retraining of ProQDock, appeared to be an easy way to get the project started but instead followed debugging of ProQDock which resulted in some small side projects.

The timetable has for this reason been a living document that was constantly updated and reformulated.

6.4 Impact in a broad sense

PPIs are the focus in several areas of research, for example drug development or to understand disease mechanisms [2]. Since X-ray crystallography and mutagenesis are the most reliable methods for detecting PPI, which both are expensive in time and resources [8], other methods for identification of PPI such as computational approaches are a necessity [3]. In order for computational approaches to be an option for detection of PPI, programs for evaluating the quality of the PPI models are essential. Especially, since a common technique when modeling PPI is to generate many alternative PPI models [4]. Good computational programs for generating and scoring PPI models could, therefore, reduce time and resources spent on detecting PPI.

6.5 Future perspectives

cProQPred is dependent on the programs ProQDock and InterPred to generated the necessary features. Since cProQPred performed better than both program, it would, therefore, be relevant to produce a program which by itself can calculate the relevant features directly from the coarse PPI models and predict the IS score. The new program or further developed cProQPred would then be independent and the only needing the coarse PPI models to predict their IS score.

The different features' importance was investigated, in order to determine which features should be included in the predictor based on each feature's contribution to the prediction of IS score. It would, therefore, be interesting to investigate if the features have a different importance for predicting the refined IS score. Perhaps different features contribute more to the prediction of refined IS score then IS score. If this would be the case, a predictor with another feature composition might generate a higher correlation between true refined IS and predicted refined IS score.

It would also be interesting to investigate appropriate ways of determining if a coarse PPI model has a chance of succeeding at refinement.

7. Acknowledgments

Thanks to Björn Wallner, associate professor department of physics, chemistry and biology at Linköping University, for the opportunity to deepen my knowledge in bioinformatics through this master thesis, and eagerly answering my questions.

Thanks to Björn Wallner and Claudio Mirabello, post doc. department of physics, chemistry and biology at Linköping University, for the transparency into your program InterPred and the provision of the coarse- and refined PPI models.

8. References

[1] Šikić M., et al., "Prediction of protein-protein interaction sites in sequences and 3D structures by random forest," *PLOS Computational Biology,* vol. 5, pp. 1-9, 2009. DOI: 10.1371/journal.pcbi.1000278

[2] Wong L., "Detection of interaction between proteins through rotation forest and local phase quantization descriptors," *International Journal of Molecular Sciences,* vol. 17, 2015. DOI: 10.3390/ijms17010021

[3] Xia J.-F., et al., "Predicting protein-protein interactions from protein sequences using meta predictor," *Amino Acids,* vol. 39, pp. 1595-1599, 2010.

[4] Geppert T., et al., "Protein-protein docking by shape-complementarity and property matching," *Journal of Computational Chemistry,* vol. 31, pp. 1919-1928, 2010. DOI: 10.1002/jcc.21479

[5] Zhu H., et al., "Global analysis of protein activities using proteome chips," *Science,* vol. 193, pp. 2101-2105, 2001. DOI: 10.1126/science.1062191

[6] Parrish JP., Gulyas KD., Finley RL. Jr., "Yeast two hybrid contributions to interactome mapping," *Current opinion in biotechnology,* vol. 14, no. 4, pp. 387-393, 2006. DOI: 10.1016/j.copbio.2006.06.006

[7] You Z.-H., et al., "Using manifold embedding for assessing and prediciting protein interactions from high-throughput experimental data," *Bioinformatics,* vol. 26, pp. 274-2751, 2010. DOI: 10.1093/bioinformatics/btq510

[8] Esmaielbeiki R., et al., "Progress and challenges in predicting protein interfaces," *Briefings in Bioinformatics,* vol. 17(1), pp. 117-131, 2016. DOI: 10.1093/bib/bbv027

[9] Sankar B., Wallner B., "Finding correct protein-protein docking models using ProQDock," *Bioinformatics,* vol. 32, pp. 262-270, 2016. DOI: 10.1093/bioinformatics/btw257

[10] Pazos F., Valencis A., "Similarity of phylogenetic trees as indicator of protein-protein interaction," *Protein Engineering,* vol. 14, pp. 609-614, 2001.

[11] Pellegrini M., et al., "Assigning protein functions by comparative genome analysis: Protein phylogenetics profiles," *National Academy of Sciences of the United States of America,* vol. 96, pp. 4285-4288, 1999.

[12] Wallner B., Mirabello C., "InterPred: A pipeline to identify and model protein-protein interactions," 2016. DOI: 10.1101/080754

[13] Chang S., et al., "Amino acid network and its scoring application in protein-protein docking," *Biophysical Chemistry,* vol. 134, pp. 111-118, 2008. DOI: 10.1016/j.bpc.2007.12.005

[14] Walsh I., et al., "Correct machine learning on protein sequences: a peer-reviewing perspective," *Briefings in Bioinformatics,* vol. 17, pp. 831-840, 2016. DOI: 10.1093/bib/bbv082

[15] Basu S., WallnerB., "DockQ: A quality measure for protein-protein docking models," *PLOS ONE,* vol. 11, pp. 1-9, 2016. DOI: 10.1371/journal.pone.0161879

[16] Ridder D. D., et al., "Pattern recognition in bioinformatics," *Briefings in bioinformatics,* vol. 14, pp. 633-647, 2013. DOI: 10.1093/bib/bbt020

[17] scikit learn, "Cross-validation: evaluating estimator preformance," *3.1 Cross-validation: evaluating estimator preformance,* vol. version 0.18.1, 2010-2016.

[18] Ho T. K., et al., "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 832-844, 1998. DOI: 10.1109/34.709601

[19] scikit learn, "Forests of randomized trees," *1.11. Ensemble methods,* vol. version 0.18, pp. 1.11.2-1.11.2.1, 2010-2016.

[20] scikit learn, "Parameters," *1.11. Ensemble methods,* vol. version 0.18, p. 1.11.2.3, 2010-2016.

[21] Gao M., Skolnick J., "iAlign: a method for the structural comparison of protein-protein interfaces," *Structural bioinformatics,* vol. 26, no. 18, pp. 2259-2265, 2010. DOI: 10.1093/bioinformatics/btq404

[22] scikit learn, "GridSearchCV," *sklearn.grid_search,* vol. version 0.17.1, 2010-2014.

[23] scikit learn, "GridSearchCV," *model_selection.GridSearchCV,* vol. version 0.18.1, 2010-2016.

[24] scikit learn, "Exhaustive Grid Search," *3.2 Tuning the hyper-parameters of an estimator,* vol. version 0.18.1, p. 3.2.1, 2010-2016.

[25] scikit learn, "Feature importance evaluation," *1.11 Ensemble methods,* vol. version 0.18, p. 1.11.2.5, 2010-2016.

[26] SciPY, "pearsonr," *Statistical function,* 2015.

[27] scikit learn, "Receiver operating characteristic (ROC)," *3.3 Model evaluation: quantifying the quality of predictions,* vol. version 0.18.1, p. 3.3.2.12, 2010-2016.

[28] Remmert M., et al., "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods,* vol. 9, pp. 173-175, 2012. DOI: 10.1038/nmeth.1818

[29] Webb B., Sali A., "Comparative protein structure modeling using MODELLER," *Current protocols in bioinformatics,* vol. 47, pp. 5.6.1-5.6.32, 2014. DOI: 10.1002/0471250953.bi0506s47

[30] Zhang Y., Skolnick J., "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic acids research,* vol. 33, pp. 2302-2309, 2005. DOI: 10.1093/nar/gki524

[31] Gray J. J., et al., "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations," *Journal of Molecular Biology,* vol. 331, pp. 281-299, 2003. DOI: 10.1016/S0022-2836(03)00670-3

[32] Li Y., et al., "The role of shape complementarity in the protein-protein interaction," *Scientific Reports,* vol. 3, pp. 1-7, 2013. DOI: 10.1038/srep03271

[33] Winn M. D., et al., "Overview of the CCP4 suite and current developments," *Acta Crystallographica Section D, Biological Crystallography,* vol. 67, pp. 235-242, 2011. DOI: 10.1107/S0907444910045749

[34] Ganguly D., Zhang W., Chen J., "Electrostatically accelerated encounter and folding for facile recognition of intrinsically disordered proteins," *PLOS Computational Biology,* vol. 9, pp. 1-13, 2013. DOI: 10.1371/journal.pcbi.1003363

[35] Tsai M., et al., "Electrostatics, structure prediction, and the energy landscapes for protein folding and binding," *Protein Science,* vol. 25, pp. 255-269, 2015. DOI: 10.1002/pro.2751

[36] Sheinerman F. B., et al., "Electrostatic aspects of protein-protein interaction," *Current opinion in Structural Biology,* vol. 10, pp. 153-159, 2000.

[37] Li L., et al., "DelPhi: a comprehensive suite for DelPhi software and associated resources," *BMC Biophysics,* vol. 5, 2012. DOI: 10.1186/2046-1682-5-9

[38] Conte L. L., et al., "The atomic structure of protein-protein recognition sites," *Journal of Molecular Biology,* vol. 285, pp. 2177-2198, 1999. DOI: 10.1006/jmbi.1998.2439

[39] Sankar B., et al., "Mapping the distribution of packing topologies within protein interiors show predominant preference for specific packing motifs," *BMC Bioinformatics,* vol. 12, pp. 195-220, 2011.

[40] Villar H. O., Kauvar L. M., "Amino acid preferences at protein binding sites," *FEBS Letters,* pp. 125-130, 1994. DOI: 10.1016/0014-5793(94)00648-2

[41] Glaser F., "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins-structure function and bioinformatics,* vol. 43, pp. 89-102, 2001.

[42] Ray A., et al., "Improved model quality assessment using ProQ2," *BMC Bioinformatics,* vol. 13, 2012. DOI: 10.1186/1471-2105-13-224

[43] Uziela K., Wallner B., "ProQ2: estimation of model accuracy implemented in Rosetta," *Bioinformatics,* vol. 32, pp. 1411-1413, 2016. DOI: 10.1093/bioinformatics/btv767

[44] Leaver-Fay A., et al., "Rosetta3: An object-oriented software suite for the simulation and design of macromolecules," *Methods Enzymol,* vol. 487, pp. 545-574, 2011. DOI: 10.1016/B978-0-12-381270-4.00019-6

[45] Khare S. D., Witehead T. A., "Introduction to the Rosetta special collection," *PLOS ONE,* vol. 10, pp. 1-5, 2015. DOI: 10.1371/journal.pone.0144326

[46] Bazzoli A., "Enhancements to the Rosetta energy function enable improved identification of small molecules that inhibit protein-protein interactions," *POLS ONE,* 2015.

[47] Rohl C. A., et al., "Protein structure prediction using Rosetta," *In Numerical Computer Methods, Part D, Methods in Enzymology,* vol. 383, pp. 66-93, 2004. DOI: 10.1016/S0076-6879(04)83004-0

[48] Hwang H., et al., "Protein-protein docking benchmark version 4.0," *Proteins,* vol. 78, no. 15, pp. 3111-3114, 2010. DOI: 10.1002/prot.22830

[49] McCoy A. J., et al., "Electrostatic complementarity at protein/protein interfaces," *Journal of Molecular Biology,* vol. 268, pp. 570-584, 1997. DOI: 10.1006/jmbi.1997.0987

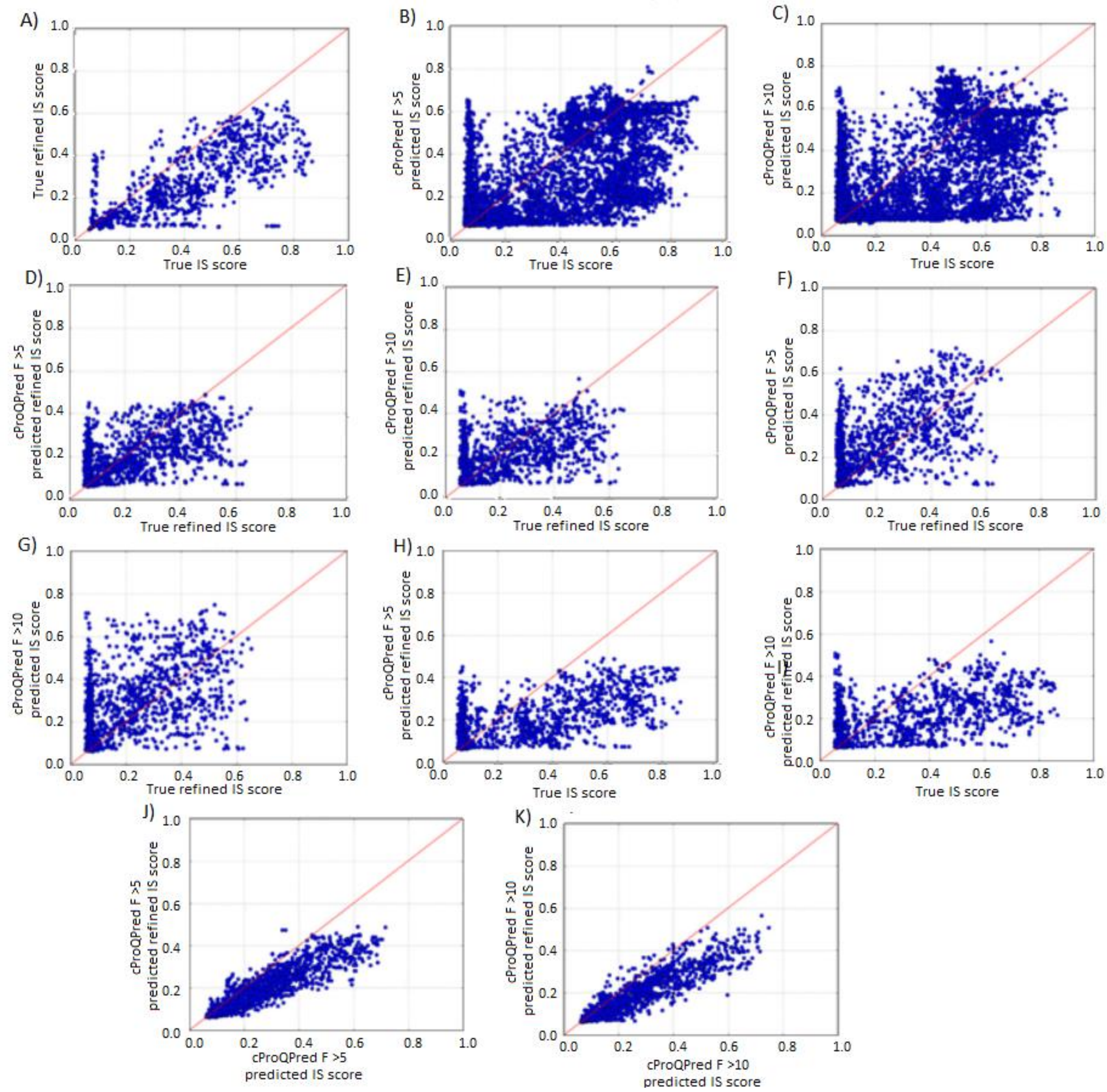
[50] You Z.-H., et al., "Predicting protein-protein interactions from primary protein sequences using multi-scale local feature representation scheme and the random forest," *PLOS ONE,* vol. 10, 2015. DOI: 10.1371/journal.pone.0125811

[51] Pons C., et al., "Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking," *Journal of Chemical Information and Modeling,* vol. 51, pp. 370-377, 2011. DOI: 10.1021/ci100353e

[52] Gavin A., et al., "Functional organization of the yeast proteome by systematic analysis of protein complex," *Nature,* vol. 415, pp. 141-147, 2002. DOI: 10.1038/415141a

[53] scikit learn, "Receiver Operating Characteristic (ROC)," *Receiver Operating Characteristic (ROC),* vol. version 0.18.1, 2010-2016.

Appendix A. Scatter plots – Relationship between IS score and refined IS score


A)
True refined IS score
True IS score
B)
cProPred F >5 predicted IS score
True IS score
C)
cProQPred F >10 predicted IS score
True IS score
D)
cProQPred F >5 predicted refined IS score
True refined IS score
E)
cProQPred F >10 predicted refined IS score
True refined IS score
F)
cProQPred F >5 predicted IS score
True refined IS score
G)
cProQPred F >10 predicted IS score
True refined IS score
H)
cProQPred F >5 predicted refined IS score
True IS score
cProQPred F >10 predicted refined IS score
True IS score
J)
cProQPred F >5 predicted refined IS score
cProQPred F >5 predicted IS score
K)
cProQPred F >10 predicted refined IS score
cProQPred F >10 predicted IS score