Improving Scoring Functions for Protein Docking by Machine Learning and Learning Data

Dissertation zur Erlangung des akademischen Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie
der Freien Universität Berlin

vorgelegt von

Florian Krull

aus Lüneburg

Oktober 2015

Die vorliegende Arbeit wurde unter Anleitung von Prof. Dr. E.-W. Knapp im Zeitraum vom 01.03.2007 bis 28.10.2015 am Institut für Chemie der Freien Universität Berlin im Fachbereich Biologie, Chemie, Pharmazie durchgeführt.

1. Gutachter: Prof. Dr. Ernst-Walter Knapp, Freie Universität Berlin

2. Gutachter: Prof. Dr. Andrew Torda, Universität Hamburg

Disputation am 02. Februar 2016

Acknowledgments

While working on this thesis, I received support and help from many people. I would like to thank my advisor, Prof. Dr. Ernst-Walter Knapp, for his continuous support of my work. I would like to thank my colleagues for helpful discussions and a great work environment. In particular, I want to thank Dr. Myong-Ho Chae for a fruitful cooperation. I would like to thank my family and friends for their spiritual support.

This work was funded by the International Research Training Group (IRTG) on "Genomics and Systems Biology of Molecular Networks" (GRK1360, Deutsche Forschungsgemeinschaft (DFG)).

Contents

List of Publications

This cumulative PhD thesis is based on the following publications:

- M.H. Chae, F. Krull, S. Lorenzen, E.W. Knapp
  Predicting Protein Complex Geometries with a Neural Network
  Proteins: Struct., Funct., Bioinf., 78 (2010) 1026–1039

- O. Demir-Kavuk, F. Krull, M.H. Chae, E.W. Knapp
  Predicting Protein Complex Geometries with Linear Scoring Functions
  Genome Inform., 24 (2010) 21-30

- M.H. Chae, F. Krull, E.W. Knapp
  Optimized Distance-Dependent Atom-Pair-Based Potential DOOP for Protein Structure Prediction
  Proteins: Struct., Funct., Bioinf., 83 (2015) 881–890

- F. Krull, G. Korff, N. Elghobashi-Meinhardt, E.W. Knapp
  ProPairs: A Data Set for Protein–Protein Docking
  J. Chem. Inf. Model., 55 (2015) 1495–1507

Additionally, while working on my PhD thesis I contributed to the following publications:

- A. Guerler, S. Lorenzen, F. Krull, E.W. Knapp
  Sampling Geometries of Protein-Protein Complexes
  Genome Inform., 20 (2008) 260-269

- A. Robertazzi, F. Krull, E.W. Knapp, P. Gamez
  Recent Advances in Anion–π Interactions
  CrystEngComm, 13 (2011) 3293-3300

1 Introduction

Protein-protein interactions are known to play a key role in many functions of living cells. They govern signal transduction and regulate metabolic processes. Anomalies in protein-protein interactions can lead to diseases like Alzheimer's and cancer [1]. Therefore, the study of protein-protein interactions is of great scientific and medical interest.

The existence of an interaction between two proteins can be detected by experimental techniques, such as yeast two-hybrid screening [2]. Computational methods, like phylogenetic profiling [3], have also been developed that aim to predict interactions. Many such protein-protein interactions are important therapeutic targets and are therefore studied in various ways. For example, protein-protein interactions can be analyzed in order to design new drugs and to understand diseases [4]. However, for such tasks knowledge about the structures of the underlying protein-protein complex is required. With experimental methods, such as X-ray crystallography [5], nuclear magnetic resonance (NMR) [6] and electron microscopy (EM) [**?**], the structure of proteins and protein-complexes can be determined.

Proteins interact through contact surfaces (interfaces) which consist of residues belonging to two or more different polypeptide chains. Protein-protein interactions can be classified as (i) obligate, where all partners of the interaction are required in order to form a stable structure; as (ii) permanent, where the binding partners have relatively long-lasting, stable interactions; and (iii) transient, where the binding partners have non-permanent or short-lived interactions. Usually, protein-protein complexes belonging to class (ii) or (iii) are heterocomplexes where the binding partners differ [7] (examples shown in Fig. 1 on page 16a).

Often the structures of the individual partners of a protein-protein interaction are available, while, for various reasons, the structure of the protein-protein complex remains unknown. These reasons are listed as follows. One reason is that resolving the structure of any protein or protein-protein complex generally requires resources such as expert knowledge and time. Additionally, many protein complexes disqualify for X-ray crys-

tallography because they are too unstable to crystallize, which is especially true when the interaction between the binding partners is weak. According to Vaynberg et al. [8], solving structures of weak protein-protein interactions lags far behind the progress on strong interactions. Combinatorial problems also exist: the individual proteins may have multiple interaction partners, thus the number of structures that emerge from the interaction of individual proteins can be much larger than the number of structures of the individual proteins themselves. Furthermore, when the partners of an interaction are mutated at the interface, the effect on the structure of the protein-protein complex is generally stronger than on the individual structures of the binding partners. Currently, the number of structures of interacting proteins obtained by experiments is far behind the number of detected protein-protein interactions [9]. Thus, predicting the structure of a protein-protein complex computationally from the individual protein structures is an important task in life science today.

1.1 Protein-Protein Docking Algorithms

The problem of having only the individual (unbound) structures of interacting proteins while not knowing their spacial configuration in the bound state is addressed by protein-protein docking algorithms. These procedures aim to predict the structure of a protein-protein complex from the structures of the two binding partners in the unbound state.

Docking algorithms usually consist of two components: a sampling algorithm and a scoring function. The first component, the sampling algorithm, uses the two unbound structures and produces a set of candidate docking structures (decoys), which potentially may resemble the protein-protein complex. For a successful docking prediction, this set is required to contain at least one decoy that is close to the native structure. A scoring function is used to discriminate near-native decoys from others, which is the second component of a docking algorithm.

Sampling algorithms can be classified into rigid docking and flexible docking. In rigid docking, the algorithm takes the binding partners and reorients them such that they

constitute a potentially correct binding mode, while the inner geometry of each binding partner is kept fixed. However, studies have shown that unbound protein structures often undergo significant changes in backbone conformation when they form the docking complex [10]. In such cases, rigid docking approaches may have difficulties finding the correct structure of the protein-protein complex. Flexible docking algorithms, in contrast, alter the backbone geometry of the binding partners to produce candidate structures for a protein-protein complex. In principle, this approach allows the docking algorithm to find the correct binding mode, even when the binding partners undergo large conformational changes upon docking. However, flexible docking algorithms have a larger search space compared to rigid docking algorithms [11], making them more expensive to compute and more likely to produce false geometries. In both flexible and rigid docking algorithms, a priory information, like the knowledge of the binding site on one or both binding partners, can reduce the search space, thus increasing the proportion of near-native structures among all generated decoys. Many rigid sampling approaches are based on fast Fourier transform (FFT) [12, 13, 14, 15, 16], as proposed by Katchalski-Katzir et al. [17]. Given two structures, each consisting of n atoms, the Katchalski-Katzir algorithm is able to compute the translation of one structure, such that the resulting complex geometry has the best surface complementarity. Modifications of the algorithm also consider electrostatics, solvation energy and atom potentials. The time complexity of the algorithm is $\mathcal{O}(n \cdot \log n)$ where a naive approach would require $\mathcal{O}(n^3)$.

To discriminate the near-native decoys that are generated by a sampling algorithm from all the other geometries, scoring functions are used. Typically, a scoring function ranks decoys with the near-native decoys expected being top-ranked. Such scoring functions can employ a large variety of informations, such as physical force fields, experimental binding energies, shape complementarity of the assumed binding sites and atom- or residue-pair potentials. So called soft scoring functions are tolerant moderate defects of input decoy structures, which is especially important for rigid docking, where the decoys generated from the unbound structures do not take into account conformational changes of the binding partners [19]. After scoring, clustering of the protein complex geometries and

prediction of energy funnels can additionally help to identify the near-native structures [20].

1.2 Machine Learning in Protein Docking

Supervised learning is a branch of machine learning. It can be employed on a set of training data in conjunction with desired response variables in order to derive a function that is able to predict response variables for new input data. The procedure of supervised learning involves several elements: the representation of the input objects, a training set of input objects and a learning algorithm. Typically, an input object is described by a feature vector $x \in \mathbb{R}^d$ that contains d features. The training set is used by the learning algorithm to deduce generalized information about new input objects. Such training sets are required to be representative in order to be effective. From the variety of existing machine learning algorithms, including artificial neural networks [21, 22], support vector machines [23] and random forests [24], each with individual strengths and weaknesses, the choice of the algorithm has to be made with respect to the training data and its representation. In addition to the training set, a prediction set of input objects is used to evaluate the accuracy of the learned function.

In protein-protein docking approaches, a scoring function that is used to rank decoys, is often defined as the combination of weighted interaction terms. Obtaining optimized weights for such scoring functions is a typical task for machine learning algorithms. Each decoy serves as an input object and its feature vector contains the values of the individual interaction terms. Examples are given in the following. The scoring function of Palma et al. [25] combines four weighted interaction terms: surface matching, side chain contacts, electrostatics and solvation energy. The weights were optimized on a training set by an artificial neural network. Bordner et al. [26] use random forests to parameterize their scoring function which consists of residue propensities, evolutionary conservation and shape complementarity. Fink et al. [27] attempt to extract a probability-like score from

a support vector machine that combines electrostatic energy, Van der Waals energy and knowledge based pair-potentials.

When applying supervised learning, there are two major sources of error that need to be considered: under-fitting and over-fitting. Under-fitting occurs when the learning algorithm fails to detect relevant characteristics within the training data. As a consequence, the algorithm computes inaccurate outputs for the training data. Typically, such scenarios are caused by inaccurate or oversimplified representations of the input objects and can be improved by choosing different or extended models. In the case of over-fitting, the other major type of error in supervised learning, the output values for the training set show good accuracy while the output values for the prediction set are inaccurate. This scenario can be caused by models of the input data that have a high complexity, so the learning algorithm gives importance to very specific characteristics of each input object in the training set that are not present in the prediction data. To overcome such situations, the complexity of the model can be lowered. Thus, given a fixed training set, choosing the complexity of the model is a trade-off between the risk of under-fitting (low complexity) and over-fitting (high complexity). Besides lowering the model complexity, another technique to avoid over-fitting is to use a training set that is more representative for any input data, typically by enlarging smaller training sets. Alternatively, the accuracy of the prediction can be improved by a validation set which is monitored to stop the learning algorithm as soon as the prediction for the validation set decreases in performance.

1.3 Data Sets for Protein-Protein Docking

This section explains the propose of data sets in protein-protein docking and gives an overview on existing data sets.

Advances in the field of protein-protein docking heavily depend ons size and quality of data sets of protein-protein complexes with experimentally determined structure. Often scoring functions are optimized by supervised learning which, as described in section 1.2, requires a training set. In order to benchmark scoring functions a prediction set is required. For a prediction scenario one typically needs the structure of the binding partners in both, unbound and bound state. The unbound state serves as input for the scoring function. The predicted output can subsequently be compared with the correct solution which is the bound state.

The Protein Data Bank (PDB) [28] is the major source for structures of biological macromolecules. Currently, the PDB contains over 100,000 protein structures, including the structures of many protein-protein complexes, and the number is constantly growing. However, the PDB does not provide an automatic way to identify protein-protein complexes among all other structures. Therefore, numerous data sets that consist of geometries of protein complexes have been compiled from the PDB and published over many years to be studied by the protein-protein docking community. An overview is given in Table 1. Aside from data sets that focus on protein complexes, other sets provide (i) decoys [29, 30], (ii) unbound structures generated from the bound structures [31], (iii) modeled unbound structures [32] and protein–protein binding affinities [33].

One of the most widely used data sets for protein-protein docking is provided by Weng and coworkers [38, 18, 34]. The authors first obtain a list of potential complexes from the PDB by applying an automated pre-filter, that involves a minimum length for the polypeptide chains and a minimum resolution for the structure. From these candidate structures the authors exclude large molecular assemblies, as they consider them to be unrealistic for docking [38]. The complexes are then categorized by hand as either transient or obligate and the obligate ones are discarded. To remove redundancy between

	release year	last update (09/2015)	complexes [a]	unbound structures [b]	superposed unbound [c]	multi-chain interfaces [d]	automatic [e]	redundancy detection [f]
Protein-Protein Docking Benchmark [34]	2003	07/2012	176	164+ 12	yes	yes	no	SCOP
DOCKGROUND BOUND [35]	2006	07/2012	3,170	no	no	no	yes	SCOP+ seq. al.[g]
DOCKGROUND UNBOUND [36]	2007	02/2009	233	99+ 134	no	yes	no	SCOP+ seq. al.[g]
Huang et al. [37]	2007	2007	851	no	no	no	no	homology scores
Score_set [30]	2014	2014	15	yes	yes	yes	no	-
ProPairs (section 2.4)	2015	2015	2,409	932+ 1,477	yes	yes	yes	interface seq. al.[h]

Table 1: Overview of five data sets from the literature used for developing protein-protein docking algorithms. [a] The number of protein-protein complexes contained in the set. [b] The number of complexes with two unbound structures (before "+") and with one unbound structure (after "+"). [c] Indicator if the set provides residue-aligned superposed unbound structures. [d] Indicator if the complexes may contain more than two polypeptide chains. [e] Indicator if the set was compiled automatically. [f] Strategy to detect redundant entries within the data set. [g] Sequence alignment. [h] Interface sequence identity by sequence alignment.

the remaining complexes, the SCOP database [39] is used to classify the proteins into a families and from each family only one representative complex is kept in the final set. Finally, only those complexes are kept for which corresponding unbound structures are found by sequence alignments. The unbound structures are required to have less than three missing residues at the binding site and to have the same cofactors at the binding

site as the bound counterpart. Updates to the protein-protein docking benchmark sets have been made every three to four years with the most recent set consisting of 176 complex structures. In Fig. 1 the structures of four complexes of the protein-protein docking benchmark 3.0 [18] are illustrated in bound and unbound state.

The authors of DOCKGROUND [35, 36] used an automatic approach and also a manually selected data set. The automatic method generates a list of protein structures that involve two polypeptide chains that form an interface without any cofactors nearby. Heuristic methods are applied that attempt to automatically identify non-obligate protein-protein complexes. For all complexes that were considered non-obligated, an algorithm identifies binding partners that are not interwoven. These early-stage classifications, which—according to the authors—do not work perfectly, tend to produce false positives [35]. Thus, the resulting data set is likely to include protein-protein complexes of binding partners that are interwoven or obligate. The SCOP database and sequence alignments are used to remove redundancies between the complexes. With this approach, 1,460 representative protein-protein complexes have been identified in 2006 [35] and the web page lists 3,170 in 2015.[1]

With DOCKGROUND UNBOUND a smaller, representative set has been released [36]. Protein-protein complexes consisting of more than two chains are considered. Also interfaces with small cofactors nearby are accepted for the final data set. For all the complexes, unbound structures are searched by sequence based methods, and if not found, simulated by using the bound structure as a basis. In contrast to the automatic method, pairs of complexes that have similar sequences but different binding modes, are kept for the final, non-redundant data set which contained 523 protein-protein complexes in 2007 and 233 protein-protein complexes in September 2015[2] (99 complexes with two unbound structures and 134 with one unbound structure). Hwang et al. [18] note that DOCKGROUND does not provide the structures of the unbound proteins residue-aligned and superposed to their bound counterparts, which requires non-trivial manual effort.

Huang et al. [37] presented a manually compiled data set, consisting of 851 protein-protein

[1] http://dockground.compbio.ku.edu/BOUND/auto_selected_new.php
[2] http://dockground.bioinformatics.ku.edu/UNBOUND/manual_selected_new.php

complexes. The set was used to derive distance dependent atom-pair potentials for protein-protein docking. Unbound structures were not provided. To remove redundancy, the authors calculated homology scores between the complexes and eliminated similar pairs by discarding the structure with the lower resolution.

Score_set by Lensink et al. [30] is a benchmark set for testing scoring functions. The complexes were taken from 15 published CAPRI [40] targets. The set's purpose lies in providing a variety of docking decoys generated by different sampling approaches, which explains its relatively small number of complexes.

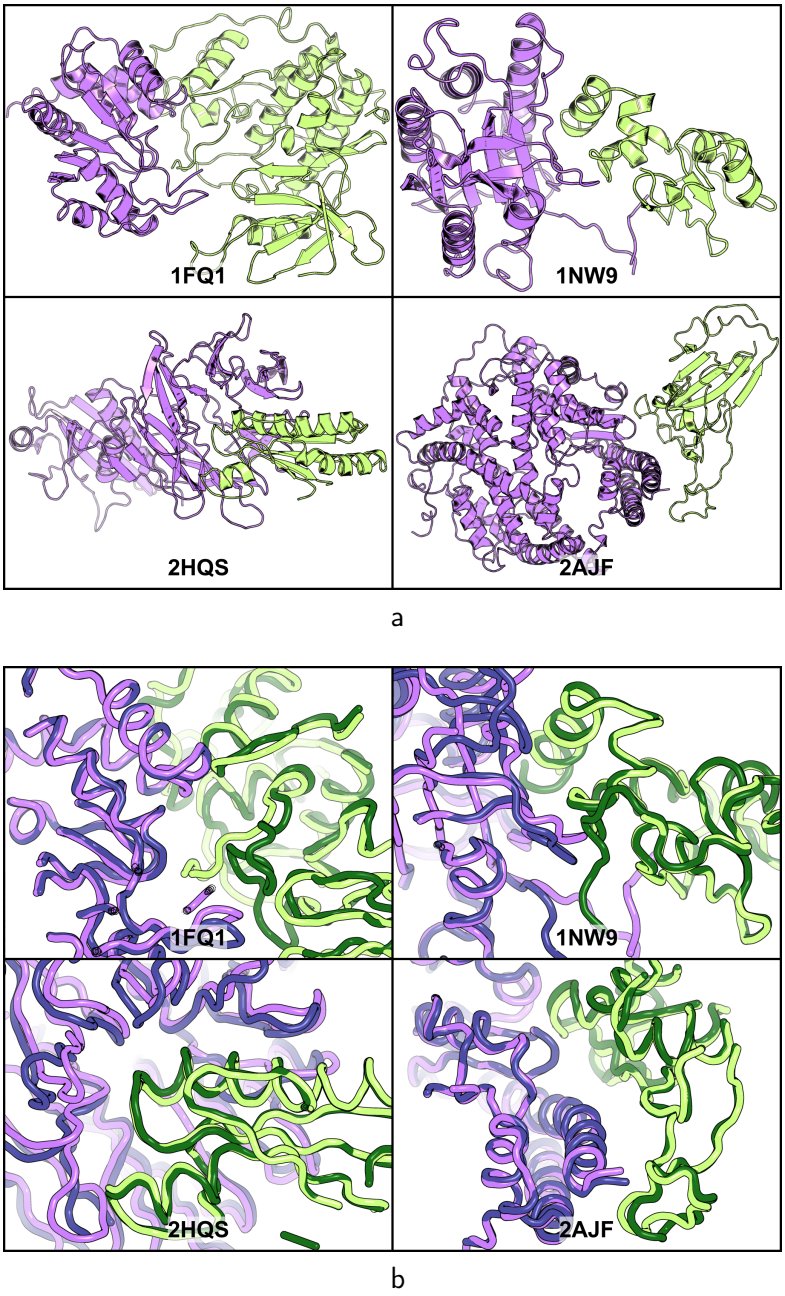


Figure 1: **a**: The bound structure of four docking complexes is shown in a cartoon representation. Each complex consists of two binding partners colored in purple and light green. **b**: Parts of the protein-protein interfaces are illustrated for the same four complexes in a ribbon representation. Each binding partner is shown in bound state (purple and light green) and in unbound state (blue and dark green) superposed to the bound structure. The interface RMSDs (iRMSD) are 1FQ1: 3.41 Å, 1NW9: 1.97 Å, 2HQS: 1.14 Å, 2AJF: 0.65 Å. All structures and iRMSDs were taken from the protein-protein docking benchmark 3.0 [18].

1.4 Aim of This Work

The aim of this study is to contribute to the improvement of protein-protein docking algorithms. One key component of a protein docking algorithm is its scoring function which is used to recognize the near-native structures within a set of candidate structures (decoys). This study investigates how to optimize a scoring function with the goal of competing with state-of-the-art scoring functions. The optimization of the scoring function is carried out by supervised learning and thus influenced by several decisions that are studied in this work: (i) the selection of a machine learning algorithm; (ii) the compilation of a training set of protein complexes, which is required for supervised learning; (iii) the generation of decoys from these complexes; and (iv) a suitable representation for those decoys that is optimized for the machine learning algorithm.

The performance of supervised learning heavily depends on the amount and quality of training data. Current data sets for protein-protein docking are outdated, very small, hard to access or lacking quality. Therefore, a special focus of this study lies on the problem how current data sets can be improved, being beneficial for scoring functions for protein-protein docking. In addition to the optimization of scoring functions, the work on large, high-quality data sets also assists research in other areas of protein-protein docking, including sophisticated benchmarking of docking algorithms and comprehensive analyses of protein-protein interactions. Therefore, the presentation of the improved data set should provide accessibility and annotation that is useful beyond the requirements on a training set.

2 Publications

2.1 Optimizing a Scoring Function for Protein Docking with a Neural Network

M.H. Chae, F. Krull, S. Lorenzen and E.W. Knapp, Predicting Protein Complex Geometries with a Neural Network, Proteins: Struct., Funct., Bioinf., 78 (2010) 1026–1039

DOI: http://dx.doi.org/10.1002/prot.22626

Own contribution:

- Development of research question
- Selection of training data
- Generation of decoys
- Development of software tools
- Preparation of manuscript

In this publication we aim to developed a scoring function capable of ranking candidate structures (decoys) of protein-protein complexes so that the decoys close to the near-native structure are top-ranking. This scoring function was derived from a set of experimental data using an artificial neural network (ANN).

The prediction capability of an ANN depends heavily on both its topology and on the input data that the network is trained on. Here, we used a data set of 191 protein docking complexes (48 complexes from Hwang et al. [18] and 143 complexes from Huang et al. [37]). In Fig. 2a, four complexes are illustrated. While many other methods use only the native structure during the training stage [41, 42, 37], we proposed a new approach, which uses a large number of near-native decoys. For each of the complexes in

the training set, we generated decoys by starting from the geometries of the two binding partners in bound state and applying small random rotations and translations to one of them. Decoys with small interfaces or large atom clashes were discarded. A total of 2,000 near-native decoys per complex were generated, consisting of ten sets. Each set $j \in [1, \ldots, 10]$ contains 200 decoys and each of these decoys has an interface RMSD (iRMSD) to the native geometry within the interval $[(j-1) \cdot 0.6, j \cdot 0.6]$ Å. Examples of these training decoys are given in Fig. 2b.

A suitable representation for the protein complexes geometries is required such that the ANN is able to use only the representation in order to make meaningful predictions. For this purpose, we use atom-pairs, where the two atoms belong to different proteins of the complex. Depending on the type of the atoms and the distance between the atoms, we assign each atom-pair to a specific class. For each decoy that we want to make a prediction for, we count the frequencies of atom-pair classes between the two proteins in that specific geometry. We use 20 atom types from the literature [37] and two additional polar hydrogen atom types, which results in $\frac{22^2}{2} + \frac{22}{2} = 253$ classes of unordered atom-pairs. Different distance classes (bins) are used, all of which consider only atom-pairs with distances up to 6.5 Å. A total of 253·(*number of distance bins*) input neurons are used. We designed the ANN as a feed-forward neural network with one input neuron for each of the atom-pair classes (Fig. 3a). We also introduced additional input neurons which serve as protein identity neurons. During training, we assign each complex of the training set to such a protein identity neuron. When we present a complex geometry to the ANN, all protein identity neurons are set to 0 except for the neuron that belongs to the complex, which is set to 1. During training, this protein identity neuron allows normalization between the different interfaces. The desired output of the network depends on the distance of a decoy to the native geometry as follows

$$g(i\text{RMSD}) = 1 - \frac{1}{1 + (i\text{RMSD}/d)^2}\,, \tag{1}$$

where we set $d = 8$ Å. The ANN was trained by using the back-propagation learning

algorithm. To avoid over-fitting we reduced the training set from 191 to 185 complexes and use the remaining six complexes as a validation set. The prediction capability for the validation set was monitored and training was stopped as soon as the prediction did not show any improvement.

After training, the prediction power of the trained neural network was measured on a prediction set of 65 complexes taken from DB2.0 [38]. For each of those complexes, 54.000 decoys were generated by ZDOCK [43]. The scoring function was used to rank the decoys and to obtain success rates. The success rate is defined as the fraction of protein complexes with at least one *HIT* within the first n predictions (or top-ranking decoys), where a *HIT* is a decoy with an iRMSD below 2.5 Å. Ideally, a scoring function has a high success rate even for small n.

We observed that varying the hidden layer size between 0 and 4 neurons does not significantly affect the success rate. Therefore, we continued further tests without a hidden layer. In this topology, the output of the ANN can be described by the function:

$$score(D) = \sum_{i \leq j}^{i,j,r} a_{ijr}(D) \cdot w_{ijr} \,, \tag{2}$$

where the number of atoms-pairs of decoy D that lie within the distance bin r and belong to the atom types i and j is given by $a_{ijr}(D)$. The corresponding weights are given by w_{ijr}. Due to the presence of protein identity neurons, Eq. (2) has an additional term during training.

We tested different numbers of distance bins, and found that the success rates increase with more distance bins but saturate at 8 distance bins.

The predictive power of the ANN has been compared to the state-of-the-art scoring function ZDOCK [43] and ZRANK [44]. As shown in Fig. 3b the ANN performs better than ZDOCK, regarding the success rate, but worse than ZRANK when more than 50 predictions are allowed. If only the 50 top-ranking decoys or less are considered, the predictive power of the ANN outperforms both ZDOCK and ZRANK. Thus, using

distance dependent atom-pair potentials that are learned by an ANN on a training set consisting of near-native decoys yields a powerful scoring function that performs equally or even better than current scoring functions from the literature.

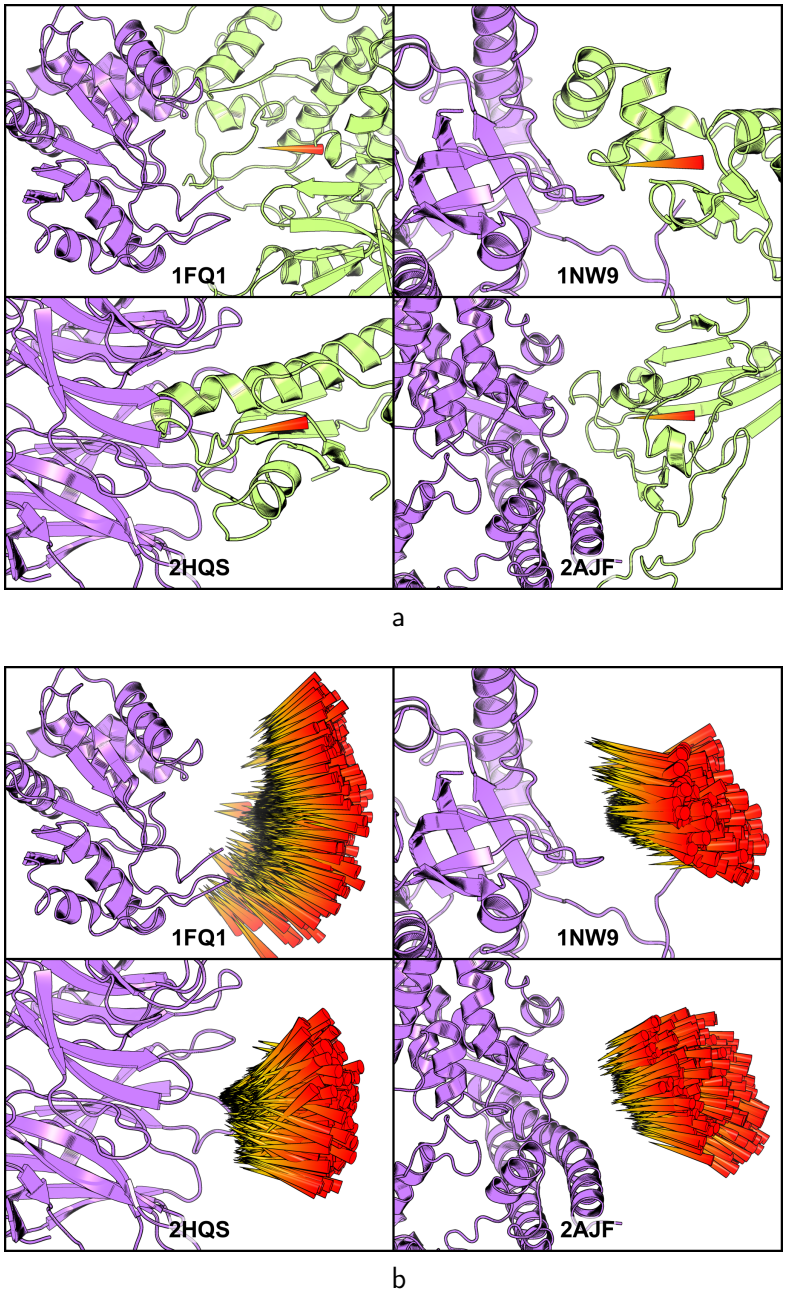


Figure 2: **a**: Four docking complexes, each with two binding partners (purple and green), of the training set are shown in a cartoon representation. Positions and orientations of the binding partners shown in green are additionally represented by a cone, with its base (red) lying in the proteins center of mass and its apex (orange) pointing to the other binding partners center of mass. **b**: The 2,000 near-native geometries (decoys) that were used for training are illustrated for the same four complexes. One binding partner (purple) is kept fixed in its native orientation while the other binding partner (represented by a cone) has 2,000 new orientations.

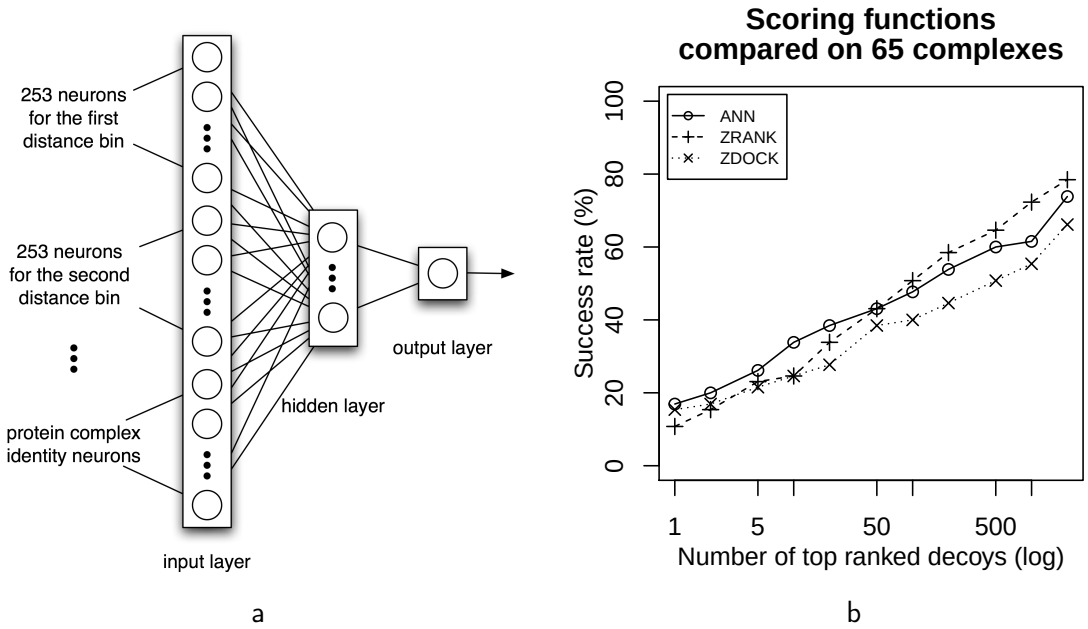


Figure 3: **a**: A schematic representation of the topology of the artificial neural network (ANN). For each distance bin there are $\frac{22^2}{2} + \frac{22}{2} = 253$ input neurons; one for each unordered pair of the 22 atom types. Additionally, one protein identity neuron is used for each complex of the training set. **b**: The success rates for varying n top ranked complex geometries (decoys) are plotted for a test set of 65 complexes. For each complex of the test set, 54,000 decoys were ranked by our ANN (using 8 distance bins and no hidden layer), by ZDOCK [43] and by ZRANK [44].

2.2 Optimizing a Scoring Function with a Linear Scoring Function

O. Demir-Kavuk, F. Krull, M.H. Chae, E.W. Knapp, Predicting Protein Complex Geometries with Linear Scoring Functions, Genome Inform., 24 (2010) 21-30

Own contribution:

- Development of research question
- Selection of training data
- Generation of decoys
- Development of software tools
- Analysis of results
- Preparation of manuscript

In this paper we investigated the sensitivity of our previous approach (section 2.1) to the underlying machine learning algorithm. Initially, we used an artificial neural network (ANN) to derive parameters of a scoring function from a training set of near-native complex geometries (decoys). In this study we replaced the ANN by a linear scoring function and we analyzed the resulting prediction power. To maintain comparability to our previous work (section 2.1), exactly the same training and prediction data was used. As before, the purpose of the scoring function resulting from the present study is to assign ranks to a large set of decoys such that decoys close to the native geometry can be found among the top scoring decoys.

As described previously, our training set consists of 191 protein complexes that were taken from the literature (48 complexes from Hwang et al. [18] and 143 complexes from Huang et al. [37]). For each decoy we used the same 2,000 near-native decoys that were generated as described earlier (section 2.1). To describe a decoy we count the number of atom-pairs that belong to a specific pair of atom type and a distance bin. We differentiate between 22 atom types that we introduced previously. In our previous work we discretized the

distances between atom-pairs in distance bins. We were able to show that it is sufficient to use eight distance bins to achieve maximum prediction power. Therefore, we describe a decoy k by a feature vector $\vec{x}_k$ containing $253 \cdot 8 = 2{,}024$ features.

To compute a score from a feature vector $\vec{x}_k$ of length d, we use a linear scoring function given by the formula

$$score(\vec{x}_k) = \vec{w}^t \cdot \vec{x}_k \,, \tag{3}$$

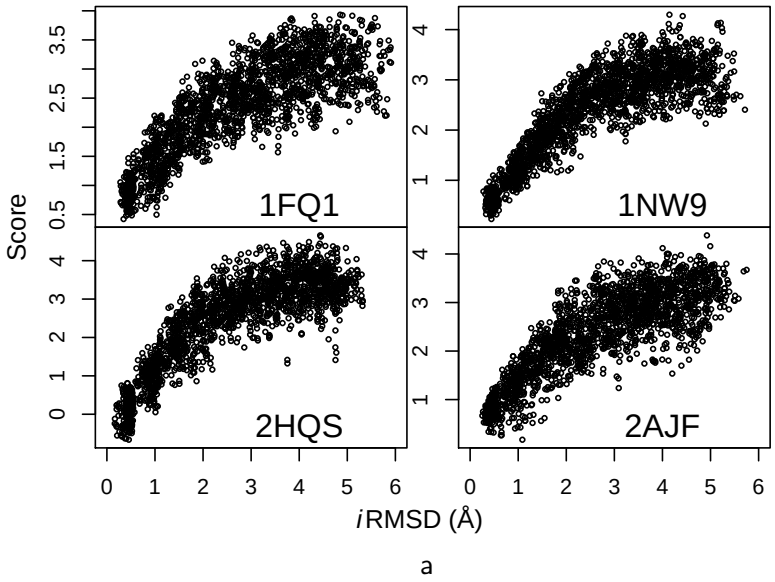
where $\vec{w} \in \mathbb{R}^d$ is a model parameter vector containing a weight for each feature. This model parameter vector is obtained from a training set of n decoys with known iRMSD, by minimizing an objective function which we define as

$$L(\vec{w}) = \sum_{k=0}^{n} \left[\mu_k \left\{\vec{w}^t \cdot \vec{x}_k - i\mathrm{RMSD}_k\right\}\right]^2 + \lambda \cdot \vec{w}^t \cdot \vec{w} \,. \tag{4}$$

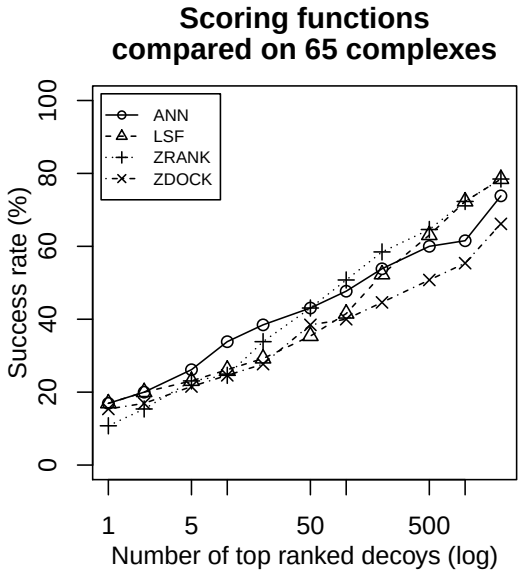
The Tikhonov regularization [45], which makes up the second term of the objective function, is used with $\lambda = 0.1$ to avoid overfitting by suppressing the least important features. We weight data points that have a small iRMSD by using the weights $\mu_k = (d - i\mathrm{RMSD}_k)^2$ with $d = 8$. In contrast to our previous work (section 2.1), with this approach we do not need a validation set during training. Therefore, we used the near-native decoys of all 191 protein complexes to derive the model parameter vector. Analogous to the protein identity neurons from our previous work, we extend the feature vector by 191 protein identity features; one for each complex of the training set. During training, for each decoy the identity feature corresponding to the complex is set to 1 while all others are set to 0. Using the protein identity features allows adjustment of the zero-point score for the particular protein complex.

After training the model parameter vector, we applied the linear scoring function Eq. (3) to the 2,000 near-native decoys of each protein complex from the training set. Examples are shown in Fig. 4a. For almost all complexes, we observed a well-established correlation between the score and the distance from the native geometry, such that an energy-funnel was modeled in the vicinity of the native protein complex geometry.

As described in section 2.1, we tested the prediction power of the linear scoring function Eq. (3) by analyzing the success rates for a prediction set of 65 protein complexes taken from Ref. [38], each with 54,000 decoys generated by ZDOCK [43]. We observed that, in terms of the predictive power, the linear scoring function seems to perform as well as the ANN (Fig. 4b). Interestingly, if the number of predictions is 1, the success rate is the same for the ANN and the linear scoring function. These observations demonstrate the robustness of the approach of deriving distance dependent atom-pair potentials from a training set of near-native decoys, as described in section 2.1 since it works equally well even when a different machine learning algorithm is used.



a



b

Figure 4: **a**: For four protein complexes of the training set, the iRMSD of all 2,000 near-native geometries (decoys) is plotted against the score resulting from the linear scoring function after minimization of the objective function. The lowest scores can be observed on decoys close to the native geometry (small iRMSD). In conclusion, the score mimics a funnel-like energy landscape for the interaction between the two proteins. **b**: The success rates for varying n top ranked decoys are plotted for a test set of 65 complexes. For each complex, 54,000 decoys were ranked by the linear scoring function (LSF), by the artificial neural network (ANN) from section 2.1, by ZDOCK [43] and by ZRANK [44].

2.3 Applying the Elements of the Protein Docking Approach to Protein Folding

M.H. Chae, F. Krull and E.W. Knapp, Optimized Distance-Dependent Atom-Pair-Based Potential DOOP For Protein Structure Prediction, Proteins: Struct., Funct., Bioinf., 83 (2015) 881–890

DOI: http://dx.doi.org/10.1002/prot.24782

Own contribution:

- Development of software tools
- Analysis of results
- Preparation of manuscript

Previously, we established a powerful scoring function using machine learning methods (section 2.1 and 2.2) in conjunction with near-native decoys to derive atom-pair potentials. In this publication we transferred this concept from the field of protein-protein docking to *de novo* protein structure prediction.

De novo protein structure prediction attempts to predict the tertiary structure of a protein from its amino acid sequence. By using experimental methods, the amino acid sequence of proteins is easier to obtain than their corresponding structure. However, the structure of a protein is important for many biological questions. A recent study [46] shows that about 1 % of the amino acid sequences available in the UniProtKB database [47] have also corresponding structures that are available within the Protein Data Bank (PDB) [28]. This situation creates a high demand for reliable prediction methods. Current methods still require large computational resources, but are able to predict the structure for small proteins with reasonable accuracy [48]. *De novo* protein structure prediction will also play an important role in protein-protein docking because one structure of the interacting proteins might be unknown. In such situations, protein structure prediction can serve as a preliminary step for a protein-protein docking algorithm [49].

Similar to protein-protein docking algorithms, *de novo* protein structure prediction generally involves two components consisting of a sampling algorithm and a scoring function. The sampling algorithm generates candidate structures (decoys). Analogous to flexible protein-protein docking (section 1.1), in protein structure prediction the sampling problem has a very large search space. Therefore, the sampling algorithm often uses either a low resolution model at its initial stages or some guidance by a scoring function. As in protein-protein docking the scoring function has to be able to identify the structures close to the native structure of a protein.

As in our previous work we used an artificial neural network (ANN) to derive atom-pair potentials. A training set consisting of 954 protein structures was compiled. Each of these proteins was successively partitioned into several receptor-ligand systems by removing residues at loops and β-turns. In total, 8,609 receptor-ligand systems were generated with this approach. Examples are shown in Fig. 5. For each of these receptor-ligand systems, 1,000 near-native decoys were generated analogous to the procedure described in section 2.1. The iRMSD of these 1,000 decoys show an even distribution within the interval $[0, 10]$ Å. A total of 10 iRMSD bins with a width of 1 Å were used, each containing 100 decoys. To describe such decoys, we used 32 atom types from the literature [50] and 14 distance bins resulting in $\left(\frac{32^2}{2} + \frac{32}{2}\right) \cdot 14 = 7{,}392$ distance dependent atom-pair classes. We designed the ANN as feed-forward neural network without hidden layers. In addition to the 7,392 input neurons for the distance dependent the atom-pair classes, we used 8,560 protein identity neurons, one for each receptor-ligand system of our training set. The remaining 49 receptor-ligand systems were used as a validation set to avoid overfitting of the ANN. The desired output of the network is given by Eq. (3) where we set $d = 2$ Å. To have the atom-pair potentials independent from the peptide bonds, we count only those pairs whose partners have a distance of at least six amino acids within the sequence of the polypeptide chain. After optimization of the parameters, the score of the ANN models a funnel-like energy landscape in the vicinity of the native geometry (Fig. 5).

In addition to the distance dependent atom-pair potentials $E_{\text{atom-pair}}$, the complete scor-

ing function, which we call the DOcking decoy-based Optimized Potential (DOOP), uses a torsion potential E_{tor}

$$E_{\text{DOOP}} = E_{\text{atom-pair}} + w_{\text{tor}} \cdot E_{\text{tor}} \,, \tag{5}$$

where w_{tor} is a weight parameter. The torsion potential of a polypeptide chain consisting of n amino acids is defined by the formula

$$E_{\text{tor}} = \sum_{i=2}^{n-1} tor(A_{i-1,} A_i, A_{i+1}, \varphi_i, \psi_i) \,, \tag{6}$$

where the function *tor* gives the torsion potential of a single residue, A_i denotes the residue type i; φ_i and ψ_i the dihedral angle. The individual torsion potentials are derived from statistical occurrences in a training set of 2,111 non-homologous proteins with pairwise sequence identities of less than 20 %. The weight w_{tor} of Eq. (5) was set to 0.1.

We applied the resulting scoring function to eight commonly used decoy sets and compared the results with other statistical potential scoring functions from the literature (Fig. 6). From a total of 168 targets, the DOOP scoring function correctly identified 151 native structures. The results demonstrate that the DOOP scoring function performs better than or as good as other state-of-the-art coring functions, which is especially true for the most challenging ROSETTA [53] decoy set. Moreover, the predictive power of the DOOP scoring function shows more consistency compared to other statistical potentials.

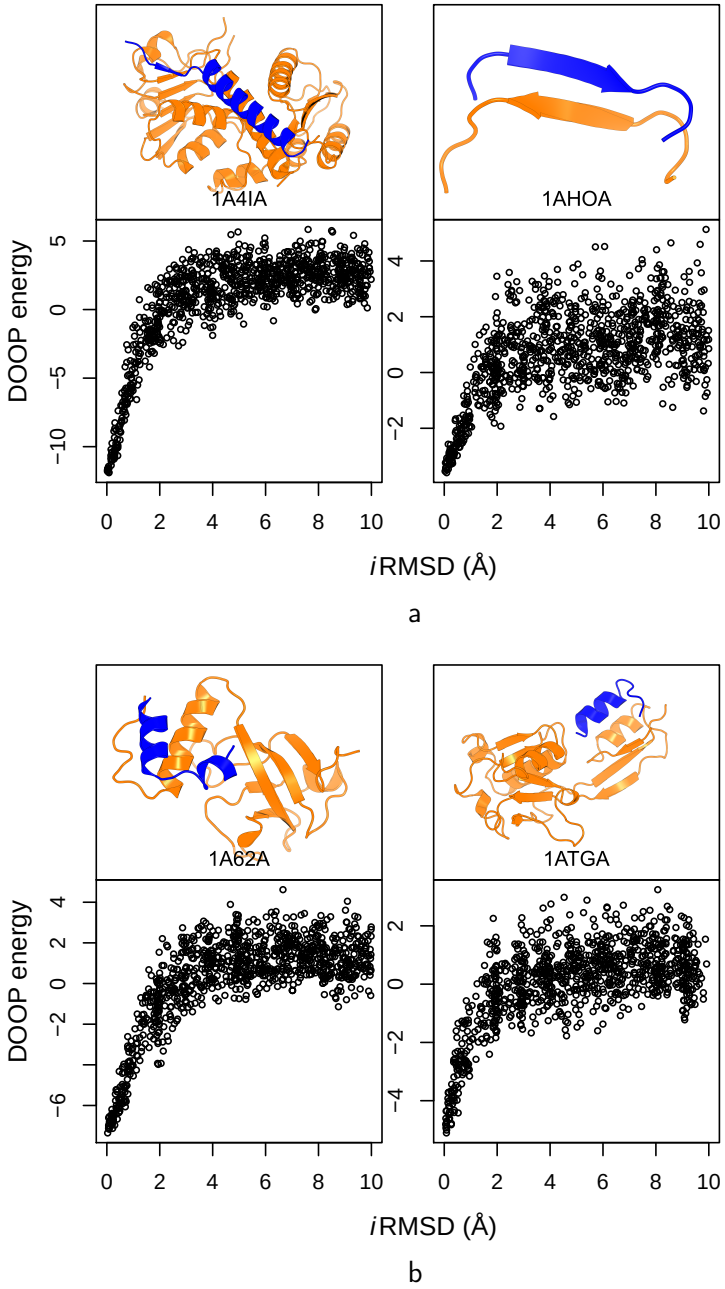


Figure 5: **a**, **b**: Four receptor-ligand systems of the training set are shown in the top row. The first four characters belong to the PDB code of the protein structure, while the fifth character identifies the chain id. The receptor is colored in orange and the ligand is colored in blue. For each receptor-ligand pair, a set of 1,000 near-native geometries (decoys) was generated. The second row shows the iRMSD vs. the score of the trained artificial neural network of the decoy sets belonging to the structure above.

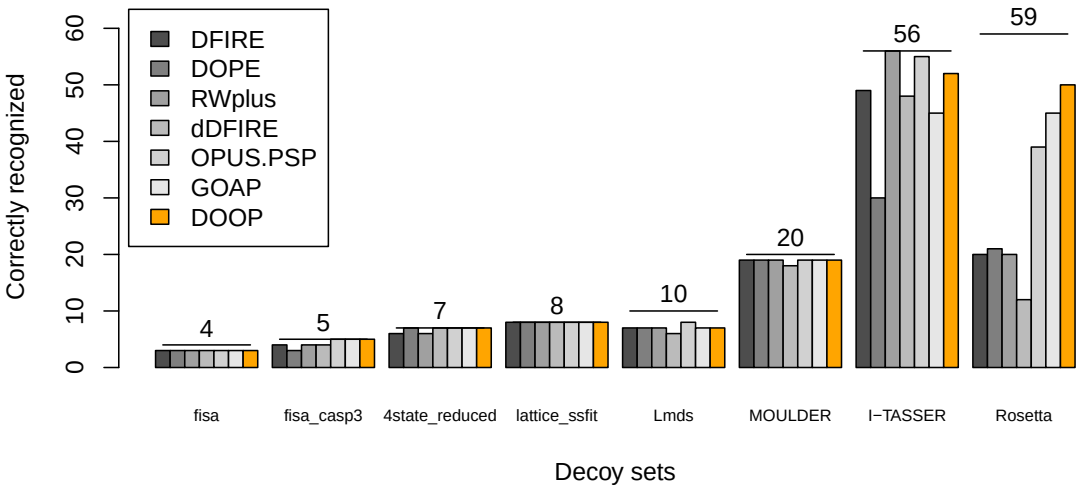


Figure 6: Comparison of results of the DOOP atom-pair potentials with six other scoring functions for eight decoy sets from the literature. For each decoy set, the total number of target structures is indicated by a horizontal line. Bars indicate the number of correctly identified targets by the individual scoring functions. Results for DFIRE, RWplus, dDFIRE, OPUS-PSP and GOAP were taken from Ref. [51] and results for DOPE were taken from Ref. [52].

2.4 Improving Protein-Protein Docking with New Data Sets for Training and Benchmarking

F. Krull, G. Korff, N. Elghobashi-Meinhardt, E.W. Knapp, ProPairs: A Data Set for Protein–Protein Docking, J. Chem. Inf. Model., 55 (2015) 1495–1507

DOI: http://dx.doi.org/10.1021/acs.jcim.5b00082

Own contribution:

- Development of research question
- Design and development of software
- Computation and analysis of results
- Development of web page
- Preparation of manuscript

In this publication we aim to enlarge the data set of protein-protein complexes that we used for training in sections 2.1 and 2.2.

Protein structures that consist of multiple polypeptide chains are potentially structures of protein-protein docking complexes. Many of such structures have been resolved and are available in the Protein Data Bank (PDB). However, most of them are obligate protein complexes, where all polypeptide chains are required in order to form a stable structure. In contrast, the binding partners that are relevant for protein-protein docking can also exist in unbound state. Based on the information in the PDB it is not directly clear which of the structures with multiple polypeptide chains are obligate and which ones are not.

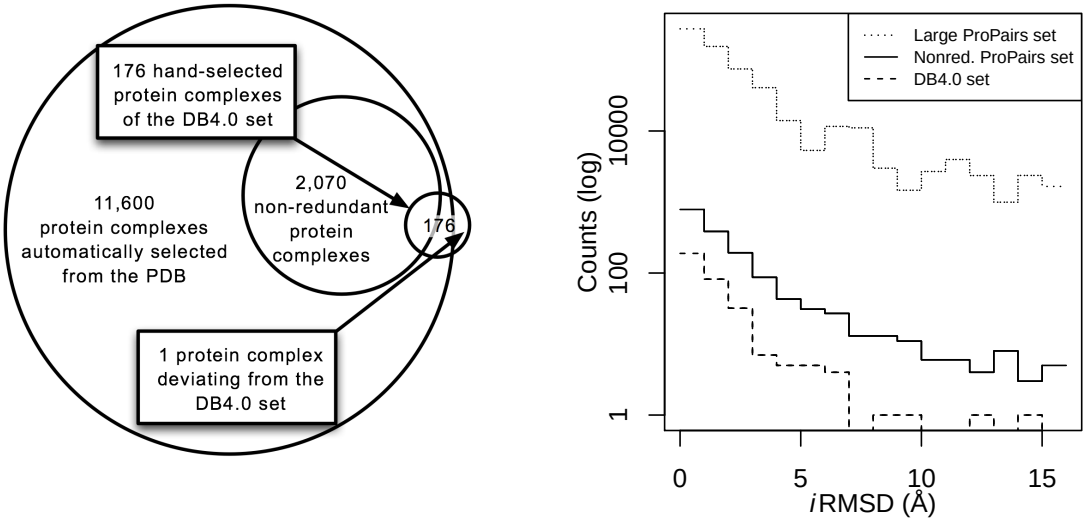
Several hand curated data sets have been compiled [34, 36, 37]. However, regarding the continuous growth of the PDB, keeping those data sets up-to-date is an important task

that has not been addressed. With the DOCKGROUND data set an automatic approach also exists, in which protein docking complexes are identified by an heuristic method.

With ProPairs we introduced a new method of compiling protein docking complexes within the PDB, by identifying the unbound structures of at least one of the binding partners. The underlying algorithm considers every structure of the PDB as a potential docking complex and tries to detect all the corresponding unbound structures. If at least one suitable unbound structure is found, the initial PDB structure is considered to be a legitimate protein docking complex. The decision as to whether an unbound structure is suitable or not relies on a defined set of rules that we have implemented in a software tool. For example, the unbound structure is required to have a high sequence identity in the interface region of the bound structure. Cofactors in the interface of the bound structure that have a biochemical function, are required to be also present in the unbound structure.

We applied our method to the PDB using its state of November 2013 and identified 11,600 interfaces. This large set was then further processed because it is important to remove redundancies. Redundancies are present because, due to similarity, one interface can be found multiple times within the same protein structure. We also want to avoid the biasing of structures that are over-represented in the PDB. To detect redundancy, we use a novel approach by computing the sequence identity of the interface to measure similarity between structures. The similarity is used to cluster all 11,600 interfaces and automatically select 2,070 representative protein docking complexes. The final nonredundant set also contains representative unbound structures that our method assigns by well defined rules. For 810 complexes, two unbound structures were identified and for the remaining 1,206 complexes, only one suitable unbound structure was assigned. With this approach, scanning the entire PDB has high computational costs. Therefore, our method uses efficient pre-filtering that reduces the set of candidate complex-unbound pairs without eliminating correct solutions.

We compared our data sets with the protein-protein docking benchmark 4.0 (DB4.0), which is one of the most widely used data sets for protein-protein docking algorithms.



a b

Figure 7: **a**: The overlap between the ProPairs data set and the DB4.0 set is schematically represented. With the exception of one case, the ProPairs data set contains all protein complexes of the DB4.0 set, while being significantly larger. **b**: For three data sets, the number of unbound proteins (logarithmic scale) are plotted as a function of their interface RMSD (iRMSD) with a bin width of 1 Å. For the small ProPairs set, 810 protein docking complexes are considered, each one possessing two unbound structures (solid line). Each unbound protein structure contributes independently to the distribution. In comparison, analogous data are shown for the DB4.0 data set for which we consider the unbound structures corresponding to the 175 complexes that are also contained in the large ProPairs data set of 11,600 protein docking complexes (dashed line). For the large ProPairs data set, all unbound structures are considered (dotted line).

We found that from the 176 complexes of DB4.0, 175 complexes are also identified with ProPairs (Fig. 7a), which demonstrates the reliability our automatic procedure. The one remaining complex (PDB id 1D6R) was not found by our approach because the biological assembly specified in the PDB is wrong.

We also analyzed the distribution of the unbound proteins' sequence lengths, interface sizes and iRMSDs in histograms. Overall, we observed for DB4.0 and the nonredundant ProPairs data set that the curves show a similar shape, despite the ProPairs data set being significantly larger. The distributions of the iRMSDs are shown in Fig. 7b. The results indicate that ProPairs generates comprehensive data sets that show similar characteristics

to smaller, hand-curated data sets. Unlike other approaches, ProPairs derives its data sets only from PDB structures and does not depend on other data bases. An overview of the nonredundant ProPairs data set in comparison with other data sets is given in Table 1 on page 13.

Along with the detailed description of our new method, we developed an interactive web page. Subsets of the protein docking complexes can be created by different selection criteria, inspected and downloaded. We also provide the ProPairs program, which can be set up in very few steps. It generates an updated data set along with a local web page displaying the data. The complete source code of the ProPairs program is released under an open source license, so users in the scientific community are also able to modify the automatic selection procedure. We gave public access to the interactive web page and the source code at http://propairs.github.io.

3 Discussion

In this section, it is demonstrated that our proposed method of generating data sets (section 2.4) is valuable for protein-protein prediction. Furthermore, it is showed that the resulting data and their presentation enables us to improve the results of our previous work (section 2.1 and section 2.2).

3.1 Motivating the Need for Automatic Procedures to Compile Data Sets

The number of structures available from the Protein Data Bank (PDB) [28] is continuously growing. In this section, it is investigated how this growth effects the number of protein complexes that can be identified with ProPairs (section 2.4). Therefore, our automatic procedure was applied to a recent state of the PDB and compared the resulting data set of protein docking complexes to a previous set.

	Nov. 2013	**Mar. 2015**	**Increase**
Large set	11,600	13,558	16.9 %
Nonred. set	2,070	2,409	14.1 %

Table 2: The number of interfaces that were found with the ProPairs method is shown for the large and the nonredundant set. Both sets were generated using the state of the Protein Data Bank (PDB) [28] from November 2013 and then generated again using PDB data from March 2015. The increase of the two sets during that time period is denoted in the third column.

The results show that the number of interfaces found in the large ProPairs set increased within 16 months by 16.9 % from 11,600 to 13,558 (Table 2). Within the same time, the number of complexes contained in the nonredundant set increased by 14.1 % from 2,070 to 2,409 docking complexes. Regarding the increasing amount of protein structures available from the PDB, exceeding 100,000 in 2015, it is obvious that the data cannot be processed manually. Thus, there is a strong need for automatic procedures such as our proposed method. With high time and cost efficiency, ProPairs also make short update cycles feasible, which are crucial due to the high growth rate of the underlying Protein Data Bank.

3.2 Demonstrating the Importance of Large Training Sets

Our method proposed in section 2.4 is able to provide training sets that are significantly larger than the ones we used in earlier studies (section 2.1 and 2.2). In this section, it is showed that these new data sets provide powerful training data for protein-protein docking algorithms.

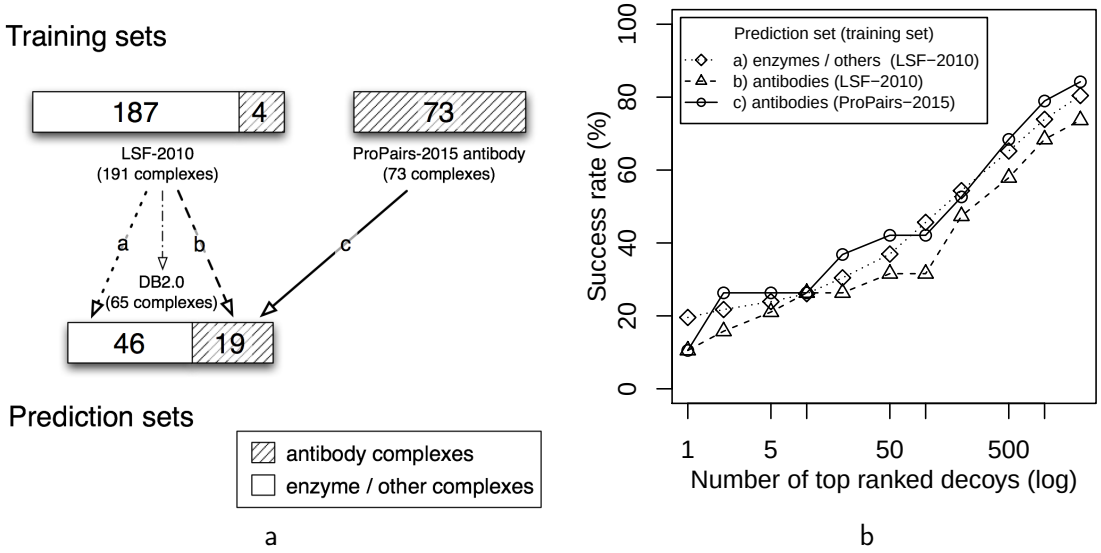


Figure 8: **a**: A schematic representation of training and prediction sets. In section 2.2, the scoring function was parameterized with a training set of 191 protein complexes (LSF-2010) and benchmarked with a prediction set of 65 protein complexes (DB2.0). This workflow is indicated by a dashed-dotted line. In this section, the prediction set of 65 protein complexes is separated into two subsets: 45 enzyme-inhibitor/other complexes and 19 antibody-antigen complexes. An additional training set (ProPairs-2015) was compiled with the ProPairs web page consisting of 73 antibody-antigen complexes only. The new combinations of training and prediction sets are investigated (lines a, b and c). **b**: The success rates for the three new combinations of training and prediction sets are plotted for varying n top ranked decoys (lines a, b and c). Compared to the set of enzyme-inhibitor/other complexes (dotted line, a), smaller success rates can be observed for the set of antibody complexes (dashed line, b), indicating that it is more challenging. However, the prediction power for the antibody complexes is increased when the training was carried out using the new ProPairs training set (solid line, c) and, for most n, exceeds the other predictions.

This paragraph focusses on the prediction of antibody-antigen complexes. Therefore, the result of section 2.2 was used as a starting point and the prediction set consisting of 65 protein complexes was separated into two classes: 45 enzyme-inhibitor/other complexes and 19 antibody-antigen complexes (lines a and b in Fig. 8a). As plotted in Fig. 8b, the linear scoring function that was discussed in section 2.2 performs better on the enzyme-inhibitor/other complexes (line a in Fig. 8a and b) than on the antibody-antigen complexes (line b in Fig. 8a and b), indicating that the antibody-antigen complexes constitute a more challenging prediction set.

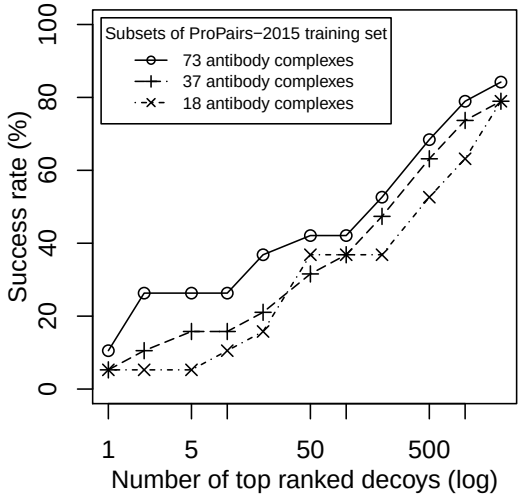


Figure 9: **a**: The success rates resulting from three different training sets are shown for varying n top ranked complex geometries (decoys). The prediction set consists of 19 antibody-antigen complexes. By selecting random complexes the ProPairs-2015 training set containing 73 antibody-antigen complexes was reduced to two smaller subsets containing 37 and 18 complexes. For almost all n, the success rates increase with the size of the training set.

From the ProPairs set of March 2015 (section 3.1), a training set consisting of 73 antibody complexes was compiled using our publicly available web page (http://propairs.github.io). With our procedure to detect similar interfaces (section 2.4), it was verified that no complex of the new training set had a similar interface to any complex of the prediction set.

The ProPairs set of 73 antibody complexes was used to derive parameters for a linear scoring function as described previously (workflow illustrated by line c in Fig. 8a). In Fig. 8b the resulting prediction power is compared to the results obtained with our previous training set of 191 complexes. The results show a superior prediction quality when using the new ProPairs training set of 73 antibody-antigen complexes (line c in Fig. 8b) compared with the data set from our previous work (line b in Fig. 8b).

To demonstrate the positive impact of larger training sets, the set of 73 antibody complexes was reduced to 37 and 18 by random selection. A comparison of the success rates, which result from training with these subsets, is shown in Fig. 9b. It can be observed that the success rates of the linear scoring functions increase with the size of the underlying training sets, which indicates the importance of comprehensive data sets for protein-protein docking such as ProPairs.

4 Conclusion and Outlook

In this study, supervised learning has been successfully applied to the protein-protein docking problem. A new method to derive a scoring function was established and notably, the concept of deriving atom-pair potentials by a machine learning algorithm from near-native decoys was introduced. This concept has been successfully carried out with two machine learning algorithms; first with an artificial neural network and then by linear scoring functions. As shown in this work, both versions of our concept resulted in scoring functions which performed equally well or even better than other state-of-the-art scoring functions for protein-protein docking.

With ProPairs a method has been developed to automatically identify new training data among experimental protein structures. This method proves to significantly enlarge the training set of protein-protein complexes that was used previously to derive atom-pair potentials. In this work it has been shown that machine learning algorithms benefit from this enlarged training set and resulting in further optimized atom-pair potentials. Generally, with larger training sets the risk of over-fitting is reduced. With enlarged training sets in combination with enlarged feature vectors the machine learning algorithm is more likely to detect additional generalities. Thus, a promising investigation would be to introduce additional features for the representation of docking decoys in order to enhance the scoring function.

The proposed ProPairs method, the resulting data sets and their representation on an interactive web page provide data for further studies, including benchmarking of docking algorithms, comprehensive analyses of protein-protein interactions or examinations of similar complexes for specific interactions. The ProPairs set provides the complete list of complexes considered to be similar for each identified protein-protein complex. For each identified unbound structure, the complete list of equivalent unbound structures is also provided.

An analysis carried out with the ProPairs method showed that the number of identified protein-protein complexes is rapidly growing. This observation supports the advantage

of an automatic identification of protein-protein complexes. The ProPairs program was released to the protein docking community and allows easy use. Consequently, ProPairs provides a sustainable way to allow the computation of up-to-date data sets also in the future.

In this work, many tasks required specific tools not covered by existing software libraries. Therefore, during this work software tools were developed consisting of reusable modules. Their code has been optimized with respect to execution time by using parallelization and hardware optimization while at the same time ensuring portability to various platforms. The portability of the library is additionally supported by its very few dependencies on third-party software. The library is available to the docking community under an open source license permitting its examination, usage, modification and extension [http://propairs.github.io].

Protein structure prediction may serve as a preliminary step to protein-protein docking when one structure of the unbound binding partners is unknown. It was successfully demonstrated that the concept of using machine learning in conjunction with near-native decoys to derive atom-pair potentials can be transferred from protein-protein docking to the field of protein structure prediction. The resulting atom-pair potentials for protein structure prediction outperformed many other statistical potentials from the literature. Thus, one can conclude that further advances in protein-docking are helpful to protein structure prediction and vice versa, especially when near-native decoys, atom-pair potentials and supervised learning are used in combination.

5 Summary

Protein-protein docking plays a central role in many biological processes, such as signal transduction and transport across membranes, and is therefore of great scientific interest. Methods exist determining if two proteins interact. For many scientific questions, notably pharmaceutical ones, knowledge about the structure of the underlying protein-protein complex is essential. Often the structures of individual protein molecules can be determined by experimental techniques, but the structural characterization of protein complexes of many molecules often remains a challenge.

This problem may be solved by computer-aided methods. Such approaches are called "protein-protein docking". They consist of methods that use the protein structure of the individual binding partners belonging to a protein-protein interaction pair as a starting point and compute the structure of the protein-protein complex. These docking algorithms usually consist of two components, one of which is a sampling algorithm that generates a set of promising structures (decoys) of the protein-protein complex. The other component is a so-called scoring function that aims to identify correct near-native protein complex structures among the generated decoys. Ideally, the scoring function assigns the best score to those structures that are closest to the native structure of the protein-protein complex.

In this study, scoring functions have been established with supervised learning. In particular, the concept of representing decoys by atom-pair potentials which are derived from near-native decoys was introduced. This concept has been successfully carried out with two machine learning algorithms; with an artificial neural network and with a linear scoring function. With both approaches a scoring function was derived that is able to compete and even outperform other state-of-the-art scoring functions from the literature.

The quality and quantity of training data play an essential role in supervised learning. Data sets from the literature consisting of protein-protein complexes turn out to be small or do not fulfill certain quality criteria. Therefore, in this study a method was developed to identify training data comprehensively and with well-defined, high quality

criteria. This method was implemented in a computer program such that data sets can be generated automatically. It has been shown that, the continuous growth of the Protein Data Bank [28]makes it necessary to provide a tool that can generate up-to-date data sets for protein docking in the future. Additionally, in this work it was demonstrated that the resulting training data notably improved the performance of the machine learning algorithm.

Finally, in this work we successfully transferred the concept of using atom-pair potentials with near-native decoys to the field of protein structure prediction. Protein structure prediction is carried out by methods that compute the structure of a protein from its amino acid sequence. Such methods can serve as a preliminary step in protein-protein docking whenever the structure of one of the two binding partners is unknown. In this study, it was shown that the concept of deriving atom-pair potentials from near-native decoys is also successfully applicable for state-of-the-art approaches in protein structure prediction.

6 Summary (German)

Protein-Protein-Interaktionen spielen eine zentrale Rolle in vielen biologischen Prozessen, wie Signaltransduktion und Transportfunktionen, und sind daher von großem wissenschaftlichen Interesse. Es existieren verschiedene Methoden, um festzustellen, ob eine Interaktion zwischen zwei Proteinen stattfindet. Beispielsweise für pharmazeutische Fragestellungen ist die räumlichen Struktur des zu Grunde liegenden Protein-Protein-Komplexes von entscheidender Bedeutung. Die Bestimmung der räumlichen Struktur von Proteinen ist mit experimentellen Methoden generell möglich, gestaltet sich jedoch für Protein-Protein-Komplexe deutlich schwieriger.

Über computergestützte Methoden versucht man dieses Problem mit geringem Aufwand zu lösen. Solche unter "Protein-Protein-Docking" zusammengefassten Verfahren gehen von der bekannten chemischen und räumlichen Struktur der Bindungspartner einer Protein-Protein-Interaktion aus und berechnen aus ihnen die Struktur des Protein-Protein-Komplexes. Zumeist bestehen Docking-Algorithmen aus zwei Komponenten. Eine Komponente generiert eine Menge aussichtsreicher Strukturen des Protein-Protein-Komplexes. Die andere Komponente, eine sogenannte "Scoring-Funktion", identifiziert unter all den aussichtsreichen Kandidaten die richtigen Strukturen. Dazu wird für jeden Kandidaten ein Zahlenwert (Score) berechnet. Idealerweise haben jene Kandidaten den höchsten Score, welche am nächsten zu der richtigen Lösung und somit am ähnlichsten zu dem nativen Protein-Protein-Komplex sind.

Im Rahmen dieser Untersuchung wurden Scoring-Funktionen mit Hilfe von maschinellem Lernen erarbeitet. Dabei wurde das Konzept vorgestellt, Protein-Komplexe über Atompaar-Potentiale zu beschreiben und diese Potentiale ausschließlich von Struktur-Kandidaten mit hoher Ähnlichkeit zum nativen Protein-Protein-Komplex abzuleiten. Dieses Konzept wurde erfolgreich mit zwei Verfahren überwachten Lernens durchgeführt; mit einem künstlichen neuronalen Netz sowie mit einer linearen Bewertungsfunktion. Mit beiden Verfahren wurde eine Scoring-Funktion bestimmt, welche eine ähnlich hohe oder bessere Vorhersagekraft als andere aktuelle Scoring-Funktionen aufweist.

Entscheidende Faktoren für den Erfolg überwachten Lernens sind die Qualität und die Quantität der Trainingsdaten. Bereits publizierte Zusammensetzungen solcher Trainingsdaten, das heißt Strukturen von Protein-Protein-Komplexen, sind relativ klein oder weisen qualitative Mängel auf. In dieser Arbeit wurde daher ein Verfahren ausgearbeitet, solche Trainingsdaten umfassend und mit hohen, klar definierten Qualitätskriterien zu bestimmen. Dieses Qualitätskriterien wurden in einem Computerprogramm verwendet, welches automatisch einsetzbar ist. Es konnte gezeigt werden, dass diese automatische Methode aufgrund der stetig wachsenden Anzahl der Strukturen in der Protein Data Bank [28] von großer Wichtigkeit ist. Ebenso wurde in dieser Arbeit demonstriert, dass durch die resultierenden Trainingsdaten der Erfolg des maschinellen Lernens deutlich verbessert werden kann.

Abschließend wurden Erkenntnisse dieser Arbeit aus dem Bereich des Protein-Protein-Docking erfolgreich auf den Bereich der Proteinstrukturvorhersage angewendet. Zur Proteinstrukturvorhersage zählen Methoden, die die Struktur eines Proteins aus seiner Aminosäurensequenz bestimmen. Diese Verfahren kommen mitunter im Protein-Protein-Docking zum Einsatz und dienen dort als einleitender Schritt, wenn die Struktur einer der beiden Bindungspartner unbekannt ist. In dieser Arbeit wurde gezeigt, dass das Konzept, Atompaar-Potentiale von fast nativen Struktur-Kandidaten abzuleiten, sich ebenfalls in der Proteinstrukturvorhersage erfolgreich gegenüber anderen Methoden bewährt.

References

[1] Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. (2008) Principles of protein-protein interactions: what are the preferred ways for proteins to interact?. *Chem. Rev.,* **108**(4), 1225–1244.

[2] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA,* **98**(8), 4569–4574.

[3] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA,* **96**(8), 4285–4288.

[4] Ivanov, A. A., Khuri, F. R., and Fu, H. (2013) Targeting protein–protein interactions as an anticancer strategy. *Trends Pharmacol. Sci.,* **34**(7), 393–400.

[5] Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R., Wyckoff, H., and Phillips, D. C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature,* **181**(4610), 662–666.

[6] Clore, G. M. and Gronenborn, A. M. (1991) Structures of larger proteins in solution: three-and four-dimensional heteronuclear NMR spectroscopy. *Science,* **252**(5011), 1390–1399.

[7] Jones, S. and Thornton, J. M. (1996) Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA,* **93**(1), 13–20.

[8] Vaynberg, J. and Qin, J. (2006) Weak protein–protein interactions as probed by NMR spectroscopy. *Trends Biotechnol.,* **24**(1), 22–27.

[9] Moal, I. H., Torchala, M., Bates, P. A., and Fernández-Recio, J. (2013) The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinf.,* **14**(1), 286.

[10] Betts, M. J. and Sternberg, M. J. (1999) An analysis of conformational changes

on protein–protein association: implications for predictive docking. *Protein Eng.,* **12**(4), 271–283.

[11] Chaudhury, S. and Gray, J. J. (2008) Conformer selection and induced fit in flexible backbone protein–protein docking using computational and NMR ensembles. *J. Mol. Biol.,* **381**(4), 1068–1087.

[12] Vakser, I. A. and Aflalo, C. (1994) Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins: Struct., Funct., Bioinf.,* **20**(4), 320–329.

[13] Ten Eyck, L. F., Mandell, J., Roberts, V. A., and Pique, M. E. (1995) Surveying molecular interactions with DOT. In *Proceedings of the 1995 ACM/IEEE conference on Supercomputing* ACM p. 22.

[14] Gabb, H. A., Jackson, R. M., and Sternberg, M. J. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.,* **272**(1), 106–120.

[15] Kozakov, D., Brenke, R., Comeau, S. R., and Vajda, S. (2006) PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins: Struct., Funct., Bioinf.,* **65**(2), 392–406.

[16] Cheng, T. M.-K., Blundell, T. L., and Fernandez-Recio, J. (2007) pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins: Struct., Funct., Bioinf.,* **68**(2), 503–515.

[17] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA,* **89**(6), 2195–2199.

[18] Hwang, H., Pierce, B., Mintseris, J., Janin, J., and Weng, Z. (2008) Protein–protein docking benchmark version 3.0. *Proteins: Struct., Funct., Bioinf.,* **73**(3), 705–709.

[19] Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002) Principles of docking: An

overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Bioinf.,* **47**(4), 409–443.

[20] Vreven, T., Hwang, H., and Weng, Z. (2011) Integrating atom-based and residue-based scoring functions for protein–protein docking. *Protein Sci.,* **20**(9), 1576–1586.

[21] Werbos, P. (1974) Beyond regression: New tools for prediction and analysis in the behavioral sciences.

[22] McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1986) Parallel distributed processing. *Explorations in the microstructure of cognition,* **2**.

[23] Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine learning,* **20**(3), 273–297.

[24] Breiman, L. (2001) Random forests. *Machine learning,* **45**(1), 5–32.

[25] Palma, P. N., Krippahl, L., Wampler, J. E., and Moura, J. J. (2000) BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins: Struct., Funct., Bioinf.,* **39**(4), 372–384.

[26] Bordner, A. J. and Gorin, A. A. (2007) Protein docking using surface matching and supervised machine learning. *Proteins: Struct., Funct., Bioinf.,* **68**(2), 488–502.

[27] Fink, F., Hochrein, J., Wolowski, V., Merkl, R., and Gronwald, W. (2011) PROCOS: Computational analysis of protein–protein complexes. *J. Comput. Chem.,* **32**(12), 2575–2586.

[28] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The protein data bank. *Nucleic Acids Res.,* **28**(1), 235–242.

[29] Liu, S., Gao, Y., and Vakser, I. A. (2008) Dockground protein–protein docking decoy set. *Bioinformatics,* **24**(22), 2634–2635.

[30] Lensink, M. F. and Wodak, S. J. (2014) Score_set: A CAPRI benchmark for scoring protein complexes. *Proteins: Struct., Funct., Bioinf.,* **82**(11), 3163–3169.

[31] Anishchenko, I., Kundrotas, P. J., Tuzikov, A. V., and Vakser, I. A. (2015) Protein models docking benchmark 2. *Proteins: Struct., Funct., Bioinf.,* **83**(5), 891–897.

[32] Anishchenko, I., Kundrotas, P. J., Tuzikov, A. V., and Vakser, I. A. (2014) Protein models: The Grand Challenge of protein docking. *Proteins: Struct., Funct., Bioinf.,* **82**(2), 278–287.

[33] Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M., and Janin, J. (2011) A structure-based benchmark for protein–protein binding affinity. *Protein Sci.,* **20**(3), 482–491.

[34] Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010) Protein–protein docking benchmark version 4.0. *Proteins: Struct., Funct., Bioinf.,* **78**(15), 3111–3114.

[35] Douguet, D., Chen, H.-C., Tovchigrechko, A., and Vakser, I. A. (2006) Dockground resource for studying protein–protein interfaces. *Bioinformatics,* **22**(21), 2612–2618.

[36] Gao, Y., Douguet, D., Tovchigrechko, A., and Vakser, I. A. (2007) DOCKGROUND system of databases for protein recognition studies: Unbound structures for docking. *Proteins: Struct., Funct., Bioinf.,* **69**(4), 845–851.

[37] Huang, S.-Y. and Zou, X. (2008) An iterative knowledge-based scoring function for protein–protein recognition. *Proteins: Struct., Funct., Bioinf.,* **72**(2), 557–579.

[38] Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., and Weng, Z. (2005) Protein–protein docking benchmark 2.0: an update. *Proteins: Struct., Funct., Bioinf.,* **60**(2), 214–216.

[39] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.,* **247**(4), 536–540.

[40] Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I., and Wodak, S. J. (2003) CAPRI: a critical assessment of predicted interactions. *Proteins: Struct., Funct., Bioinf.,* **52**(1), 2–9.

[41] Sippl, M. J. (1995) Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.,* **5**(2), 229–235.

[42] Lu, H. and Skolnick, J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Struct., Funct., Bioinf.,* **44**(3), 223–232.

[43] Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R., and Weng, Z. (2007) Integrating statistical pair potentials into protein complex prediction. *Proteins: Struct., Funct., Bioinf.,* **69**(3), 511–520.

[44] Pierce, B. and Weng, Z. (2007) ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins: Struct., Funct., Bioinf.,* **67**(4), 1078–1086.

[45] Tychonoff, A. N. (1943) On the stability of inverse problems. *Doklady Akademii Nauk SSSR,* **39**(5), 195–198.

[46] Rigden, D. J. (2009) From protein structure to function with bioinformatics, Springer, .

[47] Magrane, M., UniProt Consortium, et al. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database,* **2011**, bar009.

[48] Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., and Voelz, V. A. (2007) The protein folding problem: when will it be solved?. *Curr. Opin. Struct. Biol.,* **17**(3), 342–346.

[49] Tovchigrechko, A., Wells, C. A., and Vakser, I. A. (2002) Docking of protein models. *Protein Sci.,* **11**(8), 1888–1896.

[50] Qiu, J. and Elber, R. (2005) Atomically detailed potentials to recognize native and approximate protein structures. *Proteins: Struct., Funct., Bioinf.,* **61**(1), 44–55.

[51] Zhang, J. and Zhang, Y. (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One,* **5**(10), e15386.

[52] Shen, M.-y. and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.,* **15**(11), 2507–2524.

[53] Tsai, C.-J., Lin, S. L., Wolfson, H. J., and Nussinov, R. (1996) A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.,* **260**(4), 604–620.