Molecular Docking for Drug Discovery: Machine-Learning Approaches for Native Pose Prediction of Protein-Ligand Complexes

Hossam M. Ashtawy and Nihar R. Mahapatra(✉)

Department of Electrical and Computer Engineering,
Michigan State University, East Lansing, MI 48824, USA
{ashtawy,nrm}@egr.msu.edu

Abstract. Molecular docking is a widely-employed method in structure-based drug design. An essential component of molecular docking programs is a scoring function (SF) that can be used to identify the most stable binding pose of a ligand, when bound to a receptor protein, from among a large set of candidate poses. Despite intense efforts in developing conventional SFs, which are either force-field based, knowledge-based, or empirical, their limited docking power (or ability to successfully identify the correct pose) has been a major impediment to cost-effective drug discovery. Therefore, in this work, we explore a range of novel SFs employing different machine-learning (ML) approaches in conjunction with physicochemical and geometrical features characterizing protein-ligand complexes to predict the native or near-native pose of a ligand docked to a receptor protein's binding site. We assess the docking accuracies of these new ML SFs as well as those of conventional SFs in the context of the 2007 PDBbind benchmark datasets on both diverse and homogeneous (protein-family-specific) test sets. We find that the best performing ML SF has a success rate of 80 % in identifying poses that are within 1 Å root-mean-square deviation from the native poses of 65 different protein families. This is in comparison to a success rate of only 70 % achieved by the best conventional SF, ASP, employed in the commercial docking software GOLD. We also observed steady gains in the performance of the proposed ML SFs as the training set size was increased by considering more protein-ligand complexes and/or more computationally-generated poses for each complex.

1 Introduction

1.1 Background

Bringing a new drug to the market is a complex process that costs hundreds of millions of dollars and spans over ten years of research, development, and testing. A fairly big portion of this hefty budget and long time-line is spent in the early stages of drug design that involves two main steps: first, the enzyme, receptor, or other protein responsible for a disease of interest is identified; second, a small

E. Formenti et al. (Eds.): CIBB 2013, LNBI 8452, pp. 15–32, 2014.
DOI: 10.1007/978-3-319-09042-9_2

molecule or *ligand* is found or designed that will bind to the target protein, modulate its behavior, and provide therapeutic benefit to the patient. Typically, *high-throughput screening* (HTS) facilities with automated devices and robots are used to synthesize and screen ligands against a target protein. However, due to the large number of ligands that need to be screened, HTS is not fast and cost-effective enough as a lead identification method in the initial phases of drug discovery [1]. Therefore, computational methods referred to as *virtual screening* are employed to complement HTS by narrowing down the number of ligands to be physically screened. In virtual screening, information such as structure and physicochemical properties of a ligand, protein, or both, are used to estimate both *binding pose* and/or *binding affinity*, which represents the strength of association between the ligand and its receptor protein. The most popular approach to predicting the correct binding pose and binding affinity (BA) in virtual screening is *structure-based* in which physicochemical interactions between a ligand and receptor are deduced from the 3D structures of both molecules. This *in silico* method is also known as *protein-based* as opposed to the alternative approach, *ligand-based*, in which only ligands that are biochemically similar to the ones known to bind to the target are screened.

In this work, our focus will be on protein-based drug design, wherein ligands are placed into the active site of the receptor. The 3D structure of a ligand, when bound to a protein, is known as *ligand active conformation*. *Binding mode* refers to the orientation of a ligand relative to the target and the protein-ligand conformation in the bound state. A binding pose is simply a candidate binding mode. In molecular *docking*, a large number of binding poses are computationally generated and then evaluated using a *scoring function (SF)*, which is a mathematical or predictive model that produces a score representing binding stability of the pose. The outcome of the docking run, therefore, is a ligand's top pose ranked according to its predicted binding score as shown in Fig. 1. Typically, this docking and scoring step is performed iteratively over a database containing thousands to millions of ligand candidates. After predicting their binding poses, another scoring round is performed to rank ligands according to their predicted binding free energies. The top-ranked ligand, considered the most promising drug candidate, is synthesized and physically screened using HTS.

The most important steps in the docking process are scoring ligands' conformations at their respective binding sites and ranking ligands against each other. These core steps affect the outcome of the entire drug search campaign. That is because predictions of scoring functions determine which binding orientation/conformation is deemed the best, which ligand from a database is considered likely to be the most effective drug, and the estimated binding affinity (BA). Correspondingly, three main capabilities that a reliable scoring function should have are: (i) the ability to identify the correct binding mode of a ligand from among a set of (computationally-generated) poses, (ii) the ability to correctly rank a given set of ligands, with known binding modes when bound to the same protein, and, finally, (iii) the ability to produce binding scores that are (linearly) correlated to the experimentally-determined binding affinities of

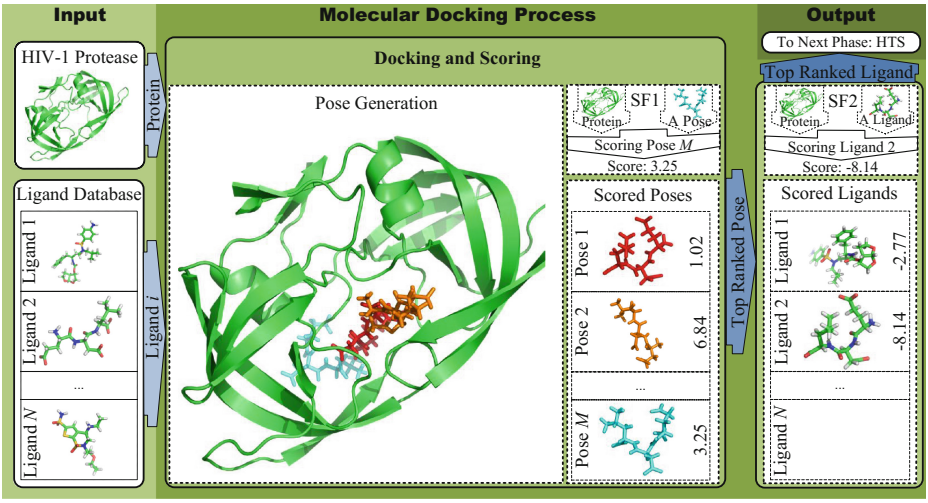

Fig. 1. Protein-ligand docking and ranking workflow.

protein-ligand complexes with known 3D structures. These three performance attributes were referred to by Cheng et al. as *docking power*, *ranking power*, and *scoring power*, respectively [2]. We refer to the corresponding problems as *ligand docking*, *ligand ranking*, and *ligand scoring problems*. In practice and in all existing work, a single general SF is trained to predict protein-ligand BA and then used in both the ligand docking and ranking stages to identify the top pose and ligand, respectively. In this work, we propose docking-specialized machine-learning SFs capable of predicting native poses more accurately than the conventional BA-based SFs. These native-pose prediction models are used as SF1 in Fig. 1. As for the second scoring round, designated by SF2 in Fig. 1, in previous work we built accurate machine-learning SFs to score and rank ligands against each other using their predicted binding affinities [3,4].

1.2 Related Work

Most SFs in use today can be categorized as either *force-field-based* [5], *empirical* [6], or *knowledge-based* [7] SFs. Despite intense efforts into these conventional scoring schemes, several recent studies report that the docking power of existing SFs is quite limited. Cheng and co-workers recently conducted an extensive test of sixteen SFs from these three categories that are either employed in mainstream commercial docking tools and/or have been developed in academia [2]. The main test set used in their study consisted of 195 diverse protein-ligand complexes and four other protein-specific test sets. In order to assess the docking power of all SFs, they generated 100-pose decoy sets for each protein-ligand complex in the main test set. They defined the *docking power* of an SF as its rate of success in identifying binding poses that are within a certain root-mean-square

deviation (RMSD) from the native pose over all complexes. Using this criteria, three SFs were found to have a relatively higher level of accuracy when their docking abilities were judged in three different experiments. These SFs are ASP [8] in the GOLD [9] docking software, PLP1 [10] in Discovery Studio [11], and the stand-alone SF DrugScore [12].

In this work, we will compare our novel ML SFs against these three and the other thirteen SFs considered by Cheng et al. [2]. They used the four popular docking programs LigandFit [13], GOLD, Surflex [14], and FlexX [15] to generate diverse sets of decoy poses. Each of these tools employs different conformational search algorithms for best poses. Namely, LigandFit relies on a shape-directed algorithm, GOLD uses a genetic algorithm, Surflex is guided by a molecular-similarity based algorithm, and FlexX employs an incremental construction algorithm as a search engine [2]. They then combined the generated poses of each program and selected a subset of 100 decoys according to a systematic clustering procedure that will be explained in more detail in Sect. 2.3. The intention behind using four different docking algorithms was to explore the conformational space as thoroughly as possible and to avoid a potential sampling bias of this space if only one program were to be used.

In previous work, we have presented BA-based ML models for the ligand scoring and ranking problems [3,4]. However, the focus of this work is on the ligand docking problem and we present docking-specialized ML SFs in which we consider a more diverse collection of features and an explicit modeling of RMSD of binding poses, which dramatically improve docking performance.

1.3 Key Contributions

Various nonparametric ML methods inspired from statistical learning theory are examined in this work to model the unknown function that maps structural and physicochemical information of a protein-ligand complex to a corresponding distance to the native pose (in terms of RMSD value). Ours is the first work to perform a comprehensive assessment of the docking accuracies of conventional and machine-learning (ML) SFs across both diverse and homogeneous (protein-family-specific) test sets using a common diverse set of features across the ML SFs. We show that the best ML SF has a success rate of $\sim$80 % compared to $\sim$70 % for the best conventional SF when the goal is to find poses within RMSD of 1 Å from the native ones for 195 different protein-ligand complexes. Such a significant improvement ($>$ 14 %) in docking power will lead to better quality drug hits and ultimately help reduce costs associated with drug discovery.

We seek to advance structure-based drug design by designing SFs that significantly improve upon the protein-ligand modeling performance of conventional SFs. Our approach is to couple the modeling power of flexible machine learning algorithms with training datasets comprising hundreds of protein-ligand complexes with native poses of known high-resolution 3D crystal structures and experimentally-determined binding affinities. In addition, we computationally generate a large number of decoy poses and utilize their RMSD values from the

native pose and a variety of features characterizing each complex. We will compare the docking accuracies of several ML and existing conventional SFs of all three types, force-field, empirical, and knowledge-based, on diverse and independent test sets. Further, we assess the impact of training set size on the docking performance of the conventional BA-based SFs and the proposed RMSD-based models.

The remainder of the paper is organized as follows. Section 2 presents the compound database used for the comparative assessment of SFs (Sect. 2.1), the physicochemical features extracted to characterize the compounds (Sect. 2.2), the procedure for decoy generation and formation of training and test datasets (Sect. 2.3), and conventional SFs (Sect. 2.4) and the ML methods (Sect. 2.5) that we employ. Next, in Sect. 3, we present results comparing the docking powers of conventional and ML SFs on diverse (Sect. 3.2) and homogeneous (Sect. 3.3) test sets, and analyze how they are impacted by training set size (Sect. 3.4). Finally, we close with concluding remarks in Sect. 4.

2 Materials and Methods

2.1 Compound Database

We used the 2007 version of PDBbind [16], the same complex database that Cheng et al. used as a benchmark in their recent comparative assessment of sixteen popular conventional SFs [2]. PDBbind is a selective compilation of the Protein Data Bank (PDB) database [17]. Both databases are publicly accessible and regularly updated. The PDB is periodically mined and only complexes that are suitable for drug discovery are filtered into the PDBbind database. In PDBbind, a number of filters are imposed to obtain high-quality protein-ligand complexes with both experimentally-determined BA and three-dimensional structure from PDB [2]. A total of 1300 protein-ligand complexes are compiled into a *refined set* after applying rigorous and systematic filtering criteria. The PDBbind curators compiled another list out of the refined set. It is called the *core set* and is mainly intended to be used for benchmarking docking and scoring systems. The core set is composed of diverse protein families and diverse binding affinities. BLAST [18] was employed to cluster the refined set based on protein sequence similarity with a 90 % cutoff. From each resultant cluster, three protein-ligand complexes were selected to be its representatives in the core set. A cluster must fulfill the following criteria to be admitted into the core set: (i) it has at least four members and (ii) the BA of the highest-affinity complex must be at least 100-fold of that of the complex with the lowest one. The representatives were then chosen based on their BA rank: the complex having the highest rank, the middle one, and the one with the lowest rank. The approach of constructing the core set guarantees unbiased, reliable, and biochemically rich test set of complexes. In order to be consistent with the comparative framework used to assess the sixteen conventional SFs mentioned above [2], we too consider the 2007 version of PDBbind which consists of a 1300-complex refined set and a 195-complex core set (with 65 clusters).

2.2 Compound Characterization

For each protein-ligand complex, we extracted physicochemical features used in the empirical SFs X-Score [6] (a set of 6 features denoted by X) and AffiScore [19] (a set of 30 features denoted by A) and calculated by GOLD [9] (a set of 14 features denoted by G), and geometrical features used in the ML SF RF-Score [20] (a 36-feature set denoted by R). The software packages that calculate X-Score, AffiScore (from SLIDE), and RF-Score features were available to us in an open-source form from their authors and a full list of these features are provided in the appendix of [4]. The GOLD docking suite provides a utility that calculates a set of general descriptors for both molecules. The set includes some common ligand molecular properties such as: molecular weight, number of rotatable bonds, number of hydrogen bonds, solvent exposed descriptors, etc. Protein-specific features are also calculated that account for the number of polar, acceptor, and donatable atoms buried in the binding pocket. As a complex, two protein-ligand interaction features are calculated which are the number of ligand atoms forming H-bonds and the number of ligand atoms that clash with protein atoms. The full set of these features can be easily accessed and calculated via the *Descriptors* menu in GOLD.

2.3 Decoy Generation and Formation of Training and Test Sets

The training dataset derived from the 2007 refined set is referred to as the *primary training set* (1105 complexes) and we denote it by Pr. It is composed of the 1300 refined-set complexes of 2007, excluding those 195 complexes present in the core set of the same year's version. The proteins of both these sets form complexes with ligands that were observed bound to them during 3D structure identification. These ligands are commonly known as native ligands and the conformation in which they were found at their respective binding sites are referred to as true or native poses. In order to assess the docking power of SFs in distinguishing true poses from random ones, a decoy set was generated for each protein-ligand complex in Pr and Cr. We utilize the decoy set produced for the core set Cr by Cheng et al. [2] using four popular docking tools: LigandFit in Discovery Studio, Surflex in SYBYL, FlexX in SYBYL (currently in LeadIT [21]), and GOLD. From each tool, a diverse set of binding poses was generated by controlling docking parameters as described in [2]. This process generated a total of ∼2000 poses for each protein-ligand complex from the four docking protocols combined. Binding poses that are more than 10 Å away, in terms of RMSD (root-mean-square deviation), from the native pose are discarded. The remaining poses are then grouped into ten 1 Å bins based on their RMSD values from the native binding pose. Binding poses within each bin were further clustered into ten clusters based on their similarities [2]. From each such subcluster, the pose with the lowest noncovalent interaction energy with the protein was selected as a representative of that cluster and the remaining poses in that cluster were discarded. Therefore, at the end of this process, decoy sets consisting of (10 bins $\times$ 10 representatives =) 100 diverse poses were generated for each protein-ligand

complex. Since we have access to the original *Cr* decoy set, we used it as is and we followed the same procedure to generate the decoy set for the training data *Pr*. Since we did not have access to Discovery Studio software, we did not use LigandFit protocol for the training data. In order to keep the size of the training set reasonable, we generated 50 decoys for each protein-ligand complex instead of 100 as it is the case for *Cr* complexes. Due to geometrical constraints during decoy generation, the final number of resultant decoys for some complexes does not add up exactly to 50 for *Pr* and 100 for *Cr*. It should be noted that the decoys in the training set are completely independent of those in the test set since both datasets share no ligands from which these decoys are generated.

We develop two types of ML SFs in this work. The first type are trained to predict binding affinities (BAs) and use these scores to distinguish promising poses from less promising ones. The second set involves building SFs to predict RMSD values explicitly. As it will be shown later, this novel approach has a superior accuracy over conventional BA-based prediction. Accordingly, two versions of training and test data sets are created. The first version uses BA as the dependent variable (Y = BA) and the size of *Pr* remains fixed at 1,105 while *Cr* includes 16,554 complexes because it consists of native poses and a decoy set for each pose. The dependent variable of the second version is RMSD (Y = RMSD) and because both training and test sets consist of native and decoy poses, the size of *Pr* expands to 39,085 while *Cr* still retains the 16,554 complex conformations.

For all protein-ligand complexes, for both native poses and computationally-generated decoys, we extracted X, A, R, and G features. By considering all fifteen combinations of these four types of features (i.e., X, A, R, G, $X \cup A$, $X \cup R$, $X \cup G$, $A \cup R$, $A \cup G$, $R \cup G$, $X \cup A \cup R$, $X \cup A \cup G$, $X \cup R \cup G$, $A \cup R \cup G$, and $X \cup A \cup R \cup G$), we generated ($15 \times 2 =$) 30 versions of the *Pr* and *Cr* data sets, which we distinguish by using the notation Pr_F^Y and Cr_F^Y to denote that the data set is characterized by the feature set F and its dependent variable is Y. For instance, Pr_{XR}^{BA} denotes the version of *Pr* comprising the set of features $X \cup R$ (referred to simply as XR) and experimentally-determined BA data for complexes in the *Pr* dataset.

2.4 Conventional Scoring Functions

A total of sixteen popular conventional SFs are compared to ML SFs in this study. The sixteen functions are either used in mainstream commercial docking tools and/or have been developed in academia. The functions were recently compared against each other in a study conducted by Cheng et al. [2]. This set includes five SFs in the Discovery Studio software [11]: LigScore, PLP, PMF, Jain, and LUDI. Five SFs in SYBYL software [22]: D-Score, PMF-Score, G-Score, ChemScore, and F-Score. GOLD software [9] contributes three SFs: GoldScore, ChemScore, and ASP. GlideScore in the Schrödinger software [23]. Besides, two standalone scoring functions developed in academia are also assessed, namely, DrugScore [12] and X-Score [6]. Some of the SFs have several options or versions, these include LigScore (LigScore1 and LigScore2), PLP (PLP1 and PLP2), and LUDI (LUDI1, LUDI2,

and LUDI3) in Discovery Studio; GlideScore (GlideScore-SP and GlideScore-XP) in the Schrödinger software; DrugScore (DrugScore-PDB and DrugScore-CSD); and X-Score (HPScore, HMScore, and HSScore). For brevity, we only report the version and/or option that yields the best performance on the PDBbind benchmark that was considered by Cheng et al.

2.5 Machine Learning Methods

We utilize a total of six regression techniques in our study: multiple linear regression (MLR), multivariate adaptive regression splines (MARS), k-nearest neighbors (kNN), support vector machines (SVM), random forests (RF), and boosted regression trees (BRT) [24]. These techniques are implemented in the following R language packages that we use [25]: the package *stats* readily available in R for MLR, *earth* for MARS [26], *kknn* for kNN [27], *e1071* for SVM [28], *randomForest* for RF [29], and *gbm* for BRT [30]. These methods benefit from some form of parameter tuning prior to their use in prediction. The optimal parameter values we use to build our models resulted from a grid search associated with 10-fold cross validation over the training set Pr and are provided in [4]. These values are obtained based on Pr_F^{BA} for any given feature set F; optimizing based on Pr_F^{RMSD} yielded similar parameter values, therefore, for brevity, we do not include them here. For every machine-learning method, we will be using these values to build ML SFs in the subsequent experiments.

3 Results and Discussion

3.1 Evaluation of Scoring Functions

In contrast to our earlier work in improving and examining scoring and ranking accuracies of different families of SFs [3,4], this study is devoted to enhancing and comparing SFs in terms of their docking powers. Docking power measures the ability of an SF to distinguish a promising binding mode from a less promising one. Typically, generated conformations are ranked in non-ascending order according to their predicted binding affinity (BA). Ligand poses that are very close to the experimentally-determined ones should be ranked high. Closeness is measured in terms of RMSD (in Å) from the true binding pose. Generally, in docking, a pose whose RMSD is within 2 Å from the true pose is considered a success or a hit.

In this work, we use comparison criteria similar to those used by Cheng et al. to compare the docking accuracies of sixteen popular conventional SFs. Doing so ensures fair comparison of ML SFs to those examined in that study in which each SF was assessed in terms of its ability to find the pose that is closest to the native one. More specifically, docking ability is expressed in terms of a success rate statistic S that accounts for the percentage of times an SF is able to find a pose whose RMSD is within a predefined cutoff value C $\mathring{A}$ by only considering the N topmost poses ranked by their predicted scores. Since success rates for

various C (e.g., 0, 1, 2, and 3 Å) and N (e.g., 1, 2, 3, and 5) values are reported in this study, we use the notation S_C^N to distinguish between these different statistics. For example, S_1^2 is the percentage of protein-ligand complexes whose either one of the two best scoring poses are within 1 Å from the true pose of a given complex. It should be noted that S_0^1 is the most stringent docking measure in which an SF is considered successful only if the best scoring pose is the native pose. By the same token and based on the C and N values listed earlier, the least strict docking performance statistic is S_3^5 in which an SF is considered successful if at least one of the five best scoring poses is within 3 Å from the true pose.

3.2 ML vs. Conventional Approaches on a Diverse Test Set

After building six ML SFs, we compare their docking performance to the sixteen conventional SFs on the core test *Cr* that comprises thousands of protein-ligand complex conformations corresponding to 195 different native poses in 65 diverse protein families. As mentioned earlier, we conducted two experiments. In the first, BA values predicted using the conventional and ML SFs were used to rank poses in a non-ascending order for each complex in *Cr*. In the other experiment, RMSD-based ML models directly predicted RMSD values that are used to rank in non-descending order the poses for the given complex.

By examining the true RMSD values of the best N scoring ligands using the two prediction approaches, success rates of SFs are shown in Fig. 2. Panels (a) and (b) in the figure show the success rates S_1^1, S_2^1, and S_3^1 for all 22 SFs. The SFs, as in the other panels, are sorted in non-ascending order from the most stringent docking test statistic value to the least stringent one. In the top two panels, for example, success rates are ranked based on S_1^1, then S_2^1 in case of a tie in S_1^1, and finally S_3^1 if two or more SFs tie in S_2^1. In both BA- and RMSD-based scoring, we find that the 22 SFs vary significantly in their docking performance. The top three BA-based SFs, GOLD::ASP, DS::PLP1, and DrugScorePDB::PairSurf, have success rates of more than 60 % in terms of S_1^1 measure. That is in comparison to the BA-based ML SFs, the best of which has an S_1^1 value barely exceeding 50 % (Fig. 2(a)). On the other hand, the other six ML SFs that directly predict RMSD values achieve success rates of over 70 % as shown in Fig. 2(b). The top performing of these ML SFs, MARS::XARG, has a success rate of ∼80 %. This is a significant improvement (>14 %) over the best conventional SF, the empirical GOLD::ASP, whose S_1^1 value is ∼70 %. Similar conclusions can also be made for the less stringent docking performance measures S_2^1 and S_3^1 in which the RMSD cut-off constraint is relaxed to 2 Å and 3 Å, respectively.

The success rates plotted in the top two panels (Fig. 2 (a) and (b)) are reported when native poses are included in the decoy sets. Panels (c) and (d) of the same figure show the impact of removing the native poses on docking success rates of all SFs. It is clear that the performance of almost all SFs does not radically decrease by examining the difference in their S_2^1 statistics which ranges from 0 to ∼5 %. This, as it was noted by Cheng et al. [2], is due to the fact that

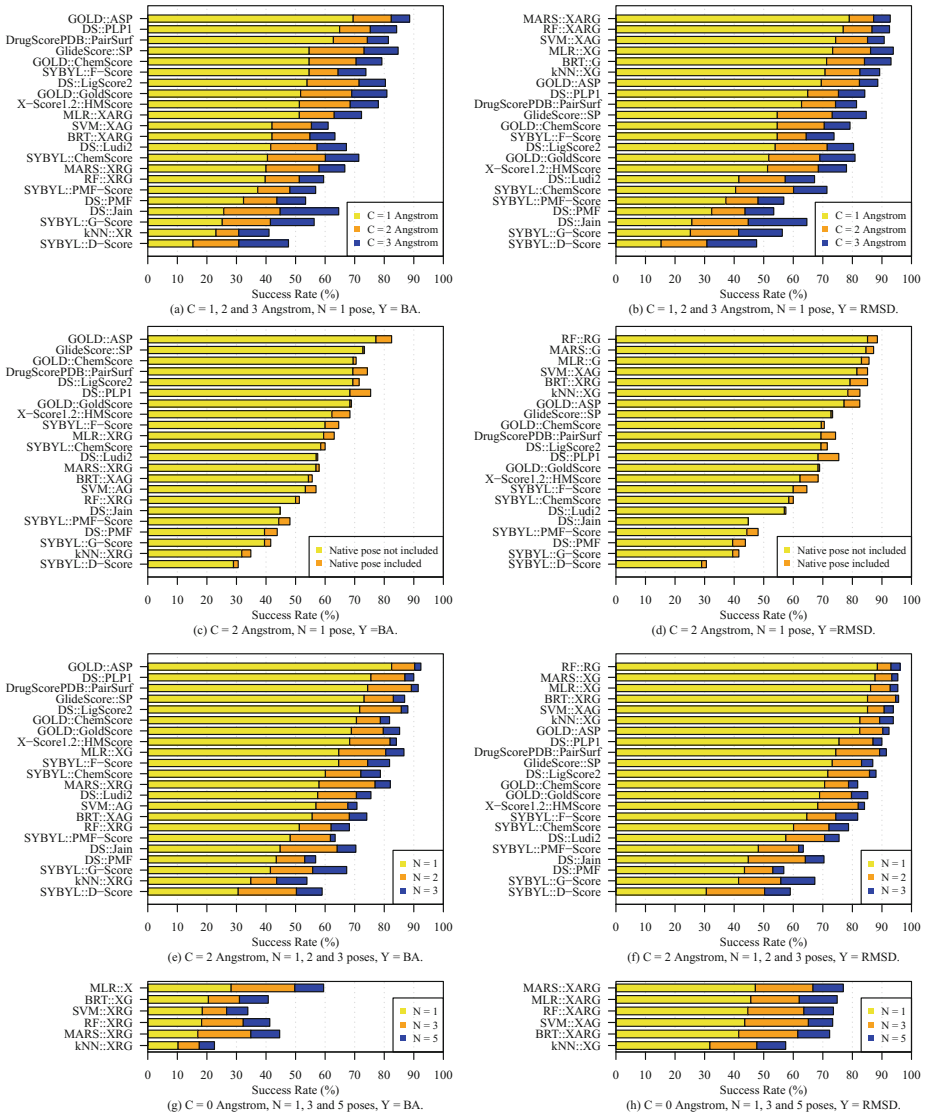

Fig. 2. Success rates of conventional and ML SFs in identifying binding poses that are closest to native ones. The results show these rates by examining the top N scoring ligands that lie within an RMSD cut-off of C Å from their respective native poses. Panels on the left show success rates when binding-affinity based (BA) scoring is used and the ones on the right show the same results when ML SFs predicted $RMSD$ values directly. Scoring of conventional SFs is BA-based in all cases and for comparison convenience we show their performance in the right panels as well.

some of the poses in the decoy sets are actually very close to the native ones. As a result, the impact of allowing native poses in the decoy sets is insignificant in most cases and therefore we include such poses in all other tests in the paper.

In reality, more than one pose is usually used from the outcomes of a docking run in the next stages of drug design for further experimentation. It is useful therefore to assess docking accuracy of SFs when more than one pose is considered (i.e., $N > 1$). Figure 2 (e) and (f) show the success rates of SFs when the RMSD values of the best 1, 2, and 3 scoring poses are examined. These rates correspond, respectively, to S_1^2, S_2^2, and S_3^2. The plots show a significant boost in performance for almost all SFs. By comparing S_1^2 to S_3^2, we observe a jump in accuracy from 82 % to 92 % for GOLD::ASP and from 87 % to 96 % for RF::RG that models RMSD values directly. Such results signify the importance of examining an ensemble of top scoring poses because there is a very good chance it contains relevant conformations and hence good drug candidates.

Upon developing RMSD-based ML scoring models, we noticed excellent improvement over their binding-affinity-based counterparts as shown in Fig. 2. We conducted an experiment to investigate whether they will maintain a similar level of accuracy when ML SFs are examined for their ability to pinpoint the native poses from their respective 100-pose decoy sets. The bottom two panels, (g) and (h), plot the success rates in terms of S_0^1, S_0^3, and S_0^5 for the six ML SFs. By examining the five best scoring poses, we notice that the top BA-based SF, MLR::X, was able to distinguish native binding poses in ∼60 % of the 195 decoy sets whereas the top RMSD-based SF, MARS::XARG, achieved a success rate of $S_0^5 = 77\,\%$ on the same protein-ligand complexes. It should be noted that both sets of ML SFs, the BA- and RMSD-based, were trained and tested on completely disjoint test sets. Therefore, this gap in performance is largely due to the explicit modeling of RMSD values and the corresponding abundance of training data which includes information from both native and computationally-generated poses.

3.3 ML vs. Conventional Approaches on Homogeneous Test Sets

In the previous section, performance of SFs was assessed on the diverse test set *Cr*. The core set consists of more than sixty different protein families each of which is related to a subset of protein families in *Pr*. That is, while the training and test set complexes were different (at least for all the ML SFs), proteins present in the core test set were also present in the training set, albeit bound to different ligands. A much more stringent test of SFs is their evaluation on a completely new protein, i.e., when test set complexes all feature a given protein—test set is homogeneous—and training set complexes do not feature that protein. To address this issue, four homogeneous test sets were constructed corresponding to the four most frequently occurring proteins in our data: HIV protease (112 complexes), trypsin (73), carbonic anhydrase (44), and thrombin (38). Each of these protein-specific test sets was formed by extracting complexes containing the protein from *Cr* (one cluster or three complexes) and *Pr* (remaining complexes). For each test set, we retrained BRT, RF, SVM, kNN, MARS, and MLR models

on the non-test-set complexes of *Pr*. Figure 3 shows the docking performance of resultant BA and RMSD-based ML scoring models on the four protein families. The plots clearly show that success rates of SFs are dependent on the protein family under investigation. It is easier for some SFs to distinguish good poses for HIV protease and thrombin than for carbonic anhydrase. The best performing SFs on HIV protease and thrombin complexes, MLR::XRG and MLR::XG, respectively, achieve success rates of over 95 % in terms of S_1^3 as shown in panels (b) and (n), whereas no SF exceeded 65 % in success rate in case of carbonic anhydrase as demonstrated in panels (i) and (j). Finding the native poses is even more challenging for all SFs, although we can notice that RMSD-based SFs outperform those models that rank poses using predicted BA. The exception to this is the SF MLR::XAR whose performance exceeds all RMSD-based ML models in terms of the success rate in reproducing native poses as illustrated in panels (c) and (d).

The results also indicate that multivariate linear regression models (MLR), which are basically empirical SFs, are the most accurate across the four families, whereas ensemble learning models, RF and BRT, unlike their good performance in Fig. 2, appear to be inferior compared to simpler models in Fig. 3. This can be attributed to the high rigidity of linear models compared to ensemble approaches. In other words, linear models are not as sensitive as ensemble techniques to the presence or absence of certain protein family in the data on which they are trained. On the other hand, RF- and BRT-based SFs are more flexible and adaptive to their training data that in some cases fail to generalize well enough to completely different test proteins as seen in Fig. 3. In practice, however, it has been observed that more than 92 % of today's drug targets are similar to known proteins in PDB [31], an archive of high quality complexes from which our training and test compounds originated. Therefore, if the goal of a docking run is to identify the most stable poses, it is important to consider sophisticated SFs (such as RF and BRT) calibrated with training sets containing some known binders to the target of interest. Simpler models, such as MLR and MARS, tend to be more accurate when docking to novel proteins that are not present in training data.

Sophisticated ML algorithms are not the only critical element in building a capable SF. Features to which they are fitted also play an important role as can be seen in Fig. 3. By comparing the right panels to the ones on the left, we can notice that X-Score features (X) are almost always present in BA-based SFs while those provided by GOLD (G) are used more to model RMSD explicitly. This implies that X-Score features are more accurate than other feature sets in predicting BA, while GOLD features are the best for estimating RMSD and hence poses close to the native one.

3.4 Impact of Training Set Size

An important factor influencing the accuracy of ML SFs is the size of the training dataset. In the case of BA-based ML SFs, training dataset size can be increased by training on a larger set of protein-ligand complexes with known

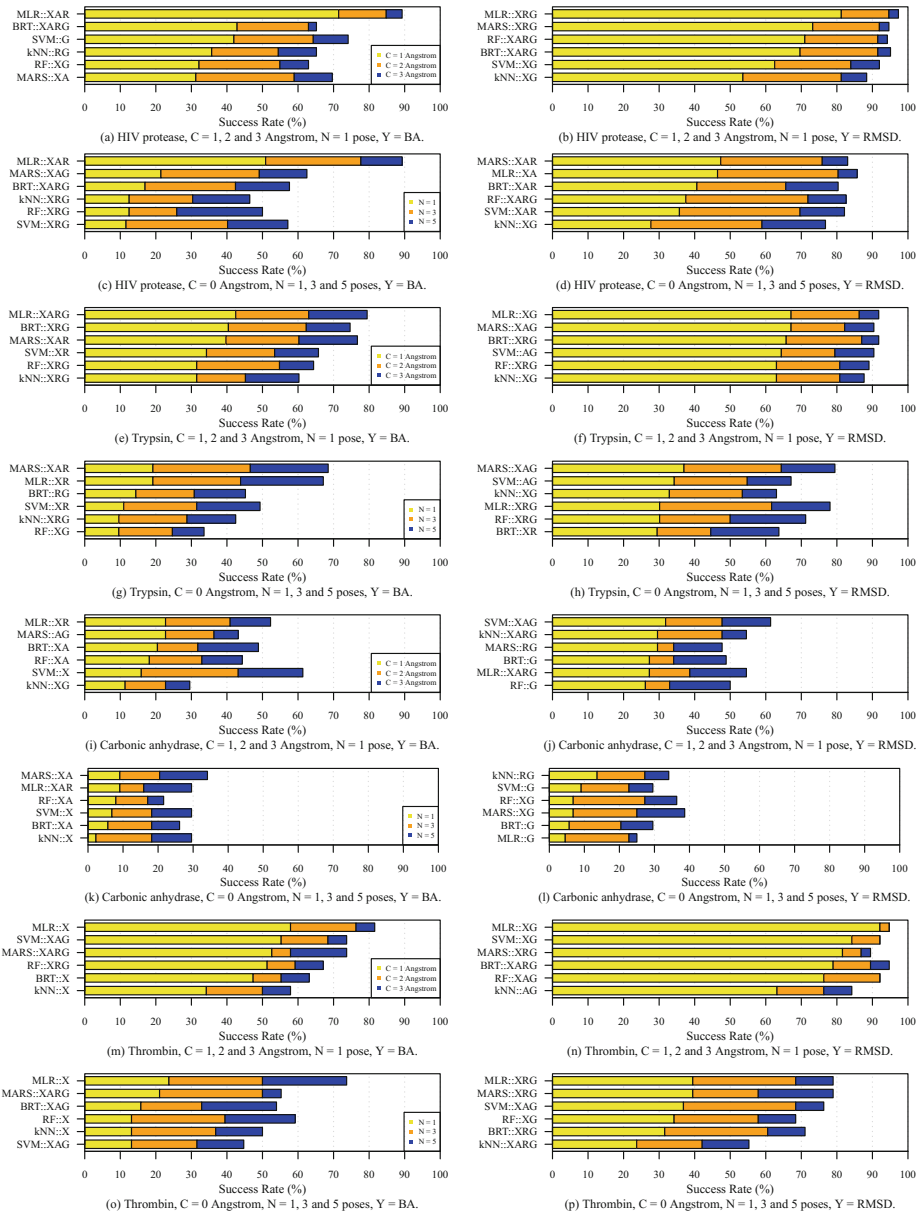

Fig. 3. Success rates of ML SFs in identifying binding poses that are closest to native ones observed in four protein families: HIV protease (a-d), trypsin (e-h), carbonic anhydrase (i-l), and thrombin (m-p). The results show these rates by examining the top N scoring ligands that lie within an RMSD cut-off of C Å from their respective native poses. Panels on the left show success rates when binding-affinity based (BA) scoring is used and the ones on the right show the same results when ML SFs predicted $RMSD$ values directly.

binding affinity values. In the case of RMSD-based SFs, on the other hand, training dataset size can be increased not only by considering a large number of protein-ligand complexes in the training set, but also by using a larger number of computationally-generated ligand poses per complex since each pose provides a new training record because it corresponds to a different combination of features and/or RMSD value. Unlike experimental binding affinity values, which have inherent noise and require additional resources to obtain, RMSD from the native conformation for a new ligand pose is computationally determined and is accurate.

We carried out three different experiments to determine: (i) the response of BA-based ML SFs to increasing number of training protein-ligand complexes, (ii) the response of RMSD-based ML SFs to increasing number of training protein-ligand complexes while the number of poses for each complex is fixed at 50, and (iii) the response of RMSD-based ML SFs to increasing number of computationally-generated poses while the number of protein-ligand complexes is fixed at 1105. In the first two experiments, we built 6 ML SFs, each of which was trained on a randomly sampled x % of the 1105 protein-ligand complexes in Pr, where $x = 10, 20, \ldots, 100$. The dependent variable in the first experiment is binding affinity (Y = BA), and the performance of these BA-based ML SFs is shown in Fig. 4(a) and partly in Fig. 4(d) (MLR::XARG). The set of RMSD values from the native pose is used as a dependent variable for ML SFs trained in the second experiment (Y = RMSD). For a given value of x, the number of conformations is fixed at 50 ligand poses for each protein-ligand complex. The docking accuracy of these RMSD-based ML models is shown in Fig. 4(b). In the third experiment, all 1105 complexes in Pr were used for training the RMSD-based ML SFs (i.e., Y = RMSD) with x randomly sampled poses considered per complex, where $x = 2, 6, 10, \ldots, 50$; results for this are reported in Fig. 4(c) and partly in Fig. 4(d) (MARS::XARG). In all three experiments, results reported are the average of 50 random runs in order to ensure all complexes and a variety of poses are equally represented. All training and test complexes in these experiments are characterized by the XARG ($=X \cup A \cup R \cup G$) features.

From Fig. 4(a), it is evident that increasing training dataset size has a positive impact on docking accuracy (measured in terms of S_1^1 success rate), although it is most appreciable in the case of MLR::XARG and MARS::XARG, two of the simpler models, MLR being linear and MARS being piecewise linear. The performance of the other models, which are all highly nonlinear, seems to saturate at 60 % of the maximum training dataset size used. The performance of all six models is quite modest, with MLR::XARG being the only one with docking success rate (slightly) in excess of 50 %. The explanation for these results is that binding affinity is not a very good response variable to learn for the docking problem because the models are trained only on native poses (for which binding affinity data is available) although they need to be able to distinguish between native and non-native poses during testing. This means that the training data is not particularly well suited for the task for which these models are used. An additional reason is that experimental binding affinity data, though useful,

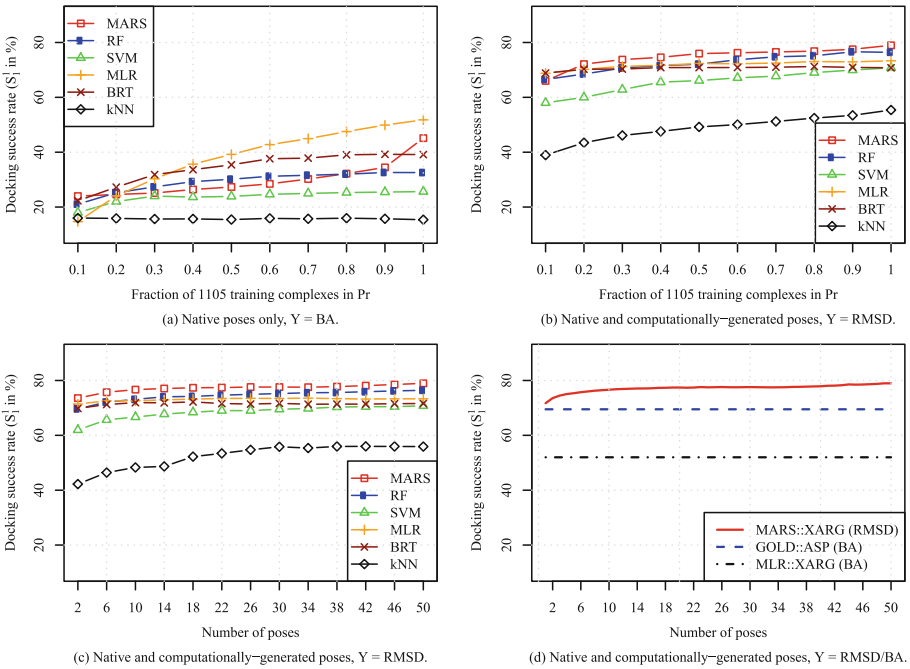

Fig. 4. Dependence of docking accuracy of ML scoring models on training set size when training complexes are selected randomly (without replacement) from *Pr* and the models are tested on *Cr*. The size of the training data was increased by including more protein-ligand complexes ((a) and (b)) or more computationally-generated poses for all complexes ((c) and (d)).

is inherently noisy. The flexible highly nonlinear models, RF, BRT, SVM, and *k*NN, are susceptible to this noise because the training dataset (arising only from native poses) is not particularly relevant to the test scenario (consisting of both native and non-native poses). Therefore, the more rigid MLR and MARS models fair better in this case.

When RMSD is used as the response variable, the training set consists of data from both native and non-native poses and hence is more relevant to the test scenario and the RMSD values, being computationally determined, are also accurate. Consequently, docking accuracy of all SFs improves dramatically compared to their BA-based counterparts as can be observed by comparing Fig. 4(a) to Fig. 4(b) and (c). We also notice that all SFs respond favorably to increasing training set size by either considering more training complexes (Fig. 4(b)) or more computationally-generated training poses (Fig. 4(c)). Even for the smallest training set sizes in Fig. 4(b) and (c), we notice that the docking accuracy of most RMSD-based SFs is about 70 % or more, which is far better than the roughly 50 % success rate for the largest training set size for the best BA-based SF MLR::XARG.

In Fig. 4(d), we compare the top performing RMSD SF, MARS::XARG, to the best BA-based SFs, GOLD::ASP and MLR::XARG, to show how docking performance can be improved by just increasing the number of computationally-generated poses, an important feature that RMSD-based SFs possess but which is lacking in their BA-based conventional counterparts. To increase the performance of these BA-based SFs to a comparable level, thousands of protein-ligand complexes with high-quality experimentally-determined binding affinity data need to be collected. Such a requirement is too expensive to meet in practice. Furthermore, RMSD-based SFs with the same training complexes will still likely outperform BA-based SFs.

4 Conclusion

We found that ML models trained to explicitly predict RMSD values significantly outperform all conventional SFs in almost all testing scenarios. The estimated RMSD values of such models have a correlation coefficient of 0.7 on average with the true RMSD values. On the other hand, predicted binding affinities have a correlation of as low as -0.2 with the measured RMSD values. This difference in correlation explains the wide gap in docking performance between the top SFs of the two approaches. The empirical SF GOLD::ASP, which is the best conventional model, achieved a success rate of 70 % in identifying a pose that lies within 1 Å from the native pose of 195 different complexes. On other hand, our top RMSD-based SF, MARS::XARG, has a success rate of ∼80 % on the same test set, which represents a significant improvement in docking performance. We also observed steady gains in the performance of RMSD-based ML SFs as the training set size was increased by considering more protein-ligand complexes and/or more computationally-generated ligand poses for each complex.

References

1. Lyne, P.D.: Structure-based virtual screening: an overview. Drug Discov. Today **7**(20), 1047–1055 (2002)
2. Cheng, T., Li, X., Li, Y., Liu, Z., Wang, R.: Comparative assessment of scoring functions on a diverse test set. J. Chem. Inf. Model. **49**(4), 1079–1093 (2009)
3. Ashtawy, H.M., Mahapatra, N.R.: A comparative assessment of conventional and machine-learning-based scoring functions in predicting binding affinities of protein-ligand complexes. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 627–630. IEEE (2011)
4. Ashtawy, H.M., Mahapatra, N.R.: A comparative assessment of ranking accuracies of conventional and machine-learning-based scoring functions for protein-ligand binding affinity prediction. IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB) **9**(5), 1301–1313 (2012)
5. Ewing, T., Makino, S., Skillman, A., Kuntz, I.: Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput. Aided Mol. Des. **15**(5), 411–428 (2001)

6. Wang, R., Lai, L., Wang, S.: Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J. Comput. Aided Mol. Des. **16**, 11–26 (2002). doi:10.1023/A:1016357811882
7. Gohlke, H., Hendlich, M., Klebe, G.: Knowledge-based scoring function to predict protein-ligand interactions. J. Mol. Biol. **295**(2), 337–356 (2000)
8. Mooij, W., Verdonk, M.: General and targeted statistical potentials for protein-ligand interactions. Proteins **61**(2), 272 (2005)
9. Jones, G., Willett, P., Glen, R., Leach, A., Taylor, R.: Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. **267**(3), 727–748 (1997)
10. Gehlhaar, D.K., Verkhivker, G.M., Rejto, P.A., Sherman, C.J., Fogel, D.R., Fogel, L.J., Freer, S.T.: Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. Chem. Biol. **2**(5), 317–324 (1995)
11. Inc., A.S.: The Discovery Studio Software, San Diego, CA (2001) (version 2.0)
12. Velec, H.F.G., Gohlke, H., Klebe, G.: DrugScore CSD - knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. J. Med. Chem. **48**(20), 6296–6303 (2005)
13. Venkatachalam, C., Jiang, X., Oldfield, T., Waldman, M.: LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J. Mol. Graph. Model. **21**(4), 289–307 (2003)
14. Jain, A.: Surflex-dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. J. Comput. Aided Mol. Des. **21**(5), 281–306 (2007)
15. Rarey, M., Kramer, B., Lengauer, T., Klebe, G.: A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. **261**(3), 470–489 (1996)
16. Wang, R., Fang, X., Lu, Y., Wang, S.: The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J. Med. Chem. **47**(12), 2977–2980 (2004)
17. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic Acids Res. **28**(1), 235–242 (2000)
18. Madden, T.: The BLAST sequence analysis tool. In: McEntyre, J., Ostell, J. (eds.) The NCBI Handbook. National Library of Medicine (US), National Center for Biotechnology Information, Bethesda (2002)
19. Schnecke, V., Kuhn, L.A.: Virtual screening with solvation and ligand-induced complementarity. In: Klebe, G. (ed.) Virtual Screening: An Alternative or Complement to High Throughput Screening?, pp. 171–190. Springer, Amsterdam (2002)
20. Ballester, P., Mitchell, J.: A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics **26**(9), 1169 (2010)
21. BioSolveIT.: LeadIT, St. Augustin, Germany (2012) (version 2.1)
22. Inc., T.: The SYBYL Software, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA (2006) (version 7.2)
23. Schrödinger, L.: The Schrödinger Software, New York (2005) (version 8.0)
24. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)
25. Team, R.D.C.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2010) ISBN 3-900051-07-0
26. Milborrow, S., Trevor, H., Tibshirani, R.: earth: Multivariate Adaptive Regression Spline Models (2010) (R package version 2.4-5)

27. Hechenbichler, K.S.K.: kknn: Weighted k-Nearest Neighbors (2010) (R package version 1.0-8)
28. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A.: e1071: Miscellaneous Functions of the Department of Statistics (e1071), TU Wien (2010) (R package version 1.5-24)
29. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
30. Ridgeway, G.: gbm: Generalized Boosted Regression Models (2010) (R package version 1.6-3.1)
31. Overington, J., Al-Lazikani, B., Hopkins, A.: How many drug targets are there? Nat. Rev. Drug Discovery **5**(12), 993–996 (2006)

http://www.springer.com/978-3-319-09041-2

Computational Intelligence Methods for Bioinformatics and Biostatistics
10th International Meeting, CIBB 2013, Nice, France, June 20–22, 2013, Revised Selected Papers
Formenti, E.; Tagliaferri, R.; Wit, E. (Eds.)
2014, XIII, 275 p. 99 illus., Softcover
ISBN: 978-3-319-09041-2