Review

Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database

DARIUSZ PLEWCZYNSKI,[1]* MICHAŁ ŁAŹNIEWSKI,[1,2]* RAFAŁ AUGUSTYNIAK,[1,2] KRZYSZTOF GINALSKI[1]
[1]*Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw,*
Pawinskiego 5a Street, 02-106 Warsaw, Poland
[2]*Department of Physical Chemistry, Faculty of Pharmacy, Medical University of Warsaw,*
Banacha 1 Street, 02-097 Warsaw, Poland

Received 13 May 2010; Revised 30 June 2010; Accepted 1 July 2010
DOI 10.1002/jcc.21643
Published online 1 September 2010 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Docking is one of the most commonly used techniques in drug design. It is used for both identifying correct poses of a ligand in the binding site of a protein as well as for the estimation of the strength of protein–ligand interaction. Because millions of compounds must be screened, before a suitable target for biological testing can be identified, all calculations should be done in a reasonable time frame. Thus, all programs currently in use exploit empirically based algorithms, avoiding systematic search of the conformational space. Similarly, the scoring is done using simple equations, which makes it possible to speed up the entire process. Therefore, docking results have to be verified by subsequent *in vitro* studies. The purpose of our work was to evaluate seven popular docking programs (Surflex, LigandFit, Glide, GOLD, FlexX, eHiTS, and AutoDock) on the extensive dataset composed of 1300 protein–ligands complexes from PDBbind 2007 database, where experimentally measured binding affinity values were also available. We compared independently the ability of proper posing [according to Root mean square deviation (or Root mean square distance) of predicted conformations versus the corresponding native one] and scoring (by calculating the correlation between docking score and ligand binding strength). To our knowledge, it is the first large-scale docking evaluation that covers both aspects of docking programs, that is, predicting ligand conformation and calculating the strength of its binding. More than 1000 protein–ligand pairs cover a wide range of different protein families and inhibitor classes. Our results clearly showed that the ligand binding conformation could be identified in most cases by using the existing software, yet we still observed the lack of universal scoring function for all types of molecules and protein families.

© 2010 Wiley Periodicals, Inc. J Comput Chem 32: 742–755, 2011

Key words: protein–ligand docking; PDBbind database; molecular recognition; software evaluation; scoring functions

Introduction

Finding a molecule that can potentially bind to a target protein is essential in the drug discovery process. Nevertheless, using only experimentally based techniques, makes it an expensive and time-consuming task. This is one of the reasons why computational methods were introduced. They are used primarily in virtual high-throughput screening of large molecular libraries (like ligand.info[1]) to identify new bioactive compounds (lead identification). Secondly, modifications of the molecule structure are possible, to better fit the pharmacological purpose (lead optimization). Because the crystallography and multidimensional nuclear magnetic resonance (NMR)[2] provide a wealth of structural information about various biological targets, collected in protein data bank (PDB) database,[3] structure-based techniques in drug

Additional Supporting Information may be found in the online version of this article.

Correspondence to: D. Plewczynski; e-mail: darman@icm.edu.pl

*These authors contributed equally

Contract/grant sponsor: National Institute of Health; contract/grant numbers: 1R01GM081680-01

Contract/grant sponsor: Polish Ministry of Science and Higher Education; contract/grant numbers: N301 159735, N301 159435, N301 246436

Contract/grant sponsor: Interdisciplinary Center for Mathematical and Computational Modeling (ICM) at Warsaw University; contract/grant numbers: G14-6, G30-2

Contract/grant sponsor: EMBO Installation

Contract/grant sponsor: Foundation for Polish Science (Focus)

design gained more ground. Among them is the docking of small organic compounds to given protein targets using known ligand and receptors structures[4–9] (proteins from different structural and functional classes during docking are often referred to as receptors, although they are not real receptors in strict biological sense). Other methods like protein–protein[10–12] or ligand–nucleic acid[10,13] docking are also reported. Still the most popular programs fit a small chemical molecule to a protein and consider only ligand flexibility with the receptor treated as rigid, because of its size, complexity, and high-computational costs. Techniques currently used to simulate protein flexibility (like docking to protein ensemble,[14] using rotamer libraries or molecular dynamics (MD) simulation[15]) will probably be replaced gradually by new approaches that allow for full protein movements.

During the last two decades, a large variety of over 60 different docking programs have been proposed for both commercial and academic use (DOCK,[16] AutoDock,[17] FlexX,[18] Surflex,[19] GOLD,[20] ICM,[21] Glide,[22] Cdocker,[23] LigandFit,[24] MCDock,[25] and many others). Although they exploit different strategies in the ligand placement, all of them can be categorized into four broad categories: stochastic Monte Carlo, fragment-based, evolutionary-based, and the shape complementary methods. None of those programs use a systematical search to fully explore all degrees of freedom in the ligand molecule because of the enormous computational cost of such a procedure.[5] If a molecule is placed in a cubic active site of 10^3 Å^3, and energy evaluation is performed every 10°, the change of the angle between the ligand and the protein, as well as a rigid movement every 0.5 Å for a molecule with four rotatable bonds—there are 6×10^{14} conformations to be sampled. It would take roughly 20,000 years to probe all of them when computing 1000 conformations per second. Although avoiding systematic search and using simplistic scoring considerably shorten the amount of time needed for docking, this may often lead to significant errors.

In this article, we evaluated the performance of seven docking programs to predict the correct three-dimensional structures of complexes, and simultaneously seven scoring functions provided by those programs to calculate the docking score for the proposed poses. As new versions of programs are frequently released, docking software must be evaluated by the community almost every year. Some excellent works providing important benchmarks were published before 2009[26–28] and their summary is included in the paper by Moitessier et al.[5] Although analyzing those studies, several observations can be made. First of all, a benchmarking dataset is usually limited to several dozen of protein–ligand complexes.[29] However, those sets are usually randomized with the protein selected from various protein folds, their number is insufficient to cover the full diversity of known polypeptide structures. Second, only a limited number of programs are tested, usually from 2 up to 5, with no particular software being successful in all performed tests. Third, once a newer version of the previously presented program is released, the authors frequently provide their own evaluation. Because of their deep knowledge of the program algorithm, those tests usually result in much higher docking accuracy than those performed by the docking community. Independent authors in most cases use default settings in benchmarking. Finally, the tests carried out by different groups on the same dataset can give different results, despite the use of similar methodology and identical docking software.[8] Fortunately, similar results are also reported. For example, a recent evaluation performed by Cheng et al.[30] on 11 scoring functions on 195 complexes selected from the PDBbind 2007 database confirms previous results obtained by Wang et al.[31] Both authors identify X-score as the most successful function with the Spearman correlation around 0.66. Those two papers focus on the evaluation of scoring functions, however, neglecting the ability of individual programs to predict correct binding poses. In most cases, to overcome certain drawbacks of the individual scoring functions, the consensus approach is proposed. The combination of different scoring functions with assigned weights is becoming an increasingly popular computational technique, yet working only for specific protein families.[32,33]

The aim of our studies was to evaluate the commonly used docking programs and their scoring abilities. Programs were chosen based on their popularity and their implementation in molecular modeling packages, namely, Sybyl, Discovery Studio, and Maestro. At least one representative of each class of docking algorithms was selected. The total number of seven programs were extensively tested on 1300 complexes, which to our knowledge comprises the most populated testing dataset to be published as of 2010. Knowing the key role of docking software in drug design, we focused here on both aspects of docking programs, i.e., the ability to recreate the ligand binding conformation, and second, the proper measurement of binding strength of the molecules. Additionally, we checked the influence of the ligand starting conformation on final docking results. To do so, we created different input structures of the ligand using popular Corina[34] and Omega2[35] software. Additionally, we explored the proper posing and scoring, when different benchmarking subsets are created based on some physicochemical properties of ligands. The results of investigations presented here represents in our opinion the most extensive and at the same time the most detailed evaluation of the docking software performance.

Materials and Methods

Docking Software

Seven docking programs were used in our benchmark and can be categorized into four distinctive categories based on the algorithm used to generate the active conformation. A fragment-based incremental method is represented by Surflex (ver. 2.2),[19] eHiTS (ver. 9.0)[36] (SimBioSys), and FlexX (ver. 2.2.1)[18] (BioSolveIt). In this approach, a ligand is split into fragments which are docked independently and then their molecule structure is recreated typically in an incremental way. The evolutionary methods are used in GOLD (ver. 3.2)[20] (CCDC) and AutoDock (ver. 4.2.1)[17] (The Scripps Research Institute). These two programs use genetic algorithms to perform the conformational search. Force field-based methods, like Glide (ver. 4.5)[22] (Schrodinger), implement Monte Carlo based engine. Finally, the shape complementarity methods, like LigandFit (ver. 2.3)[24] (Accelrys Software), exploits grids to fit the shape of a ligand into an active site of the target combined with Monte Carlo sampling (for a block diagram presenting how those algorithms work see Supporting Information Fig. S1)

GOLD

The program was developed by Jones et al.[20] and uses a genetic algorithm, with the adopted island model, to generate the ligands conformers in an active site. Four scoring functions implemented in this software are force field-based GoldScore and piecewise linear potential (PLP), empirically based ChemScore and knowledge-based Astex statistical potential (ASP).[37] Various levels of accuracy can be chosen. In our test, it was set to 100%, i.e., around 30,000 genetic algorithms (GA) operations could be performed during docking. Conformations were scored using only GoldScore function. The active site was chosen based on native ligand placement in the considered complex. In our evaluation, we used version 3.2 of the program, instead of newest 4.1. However, in our opinion, no new features regarding docking using only GoldScore functions were added, thus results obtained using those two versions of the program should be virtually identical.

AutoDock

Similarly to GOLD, AutoDock uses a genetic algorithm to generate the poses of the ligand inside a protein active site. Developed by Morris[17] it utilizes the Lamarckian version of GA, where the changes in conformations adopted by molecules after *in situ* optimization are used as a make up for offspring poses. In our tests the population size was 150, with the number of generations set to 27,000. The elitism parameter was chosen to 1, i.e. only one best fitted conformation was transferred from the parental to the offspring conformation without any change. Similar to GOLD, an active site was selected based on the position of native ligand structure in the active site.

FlexX

Docking engine of FlexX works as in all other fragment-based tools. The choice of the ligand base fragment was the key step because it makes the ligand core responsible for principal interactions with a target protein. The torsion angle database[38] was used to generate different poses of a fragment that was considered as rigid in further steps. Subsequently, the selected fragment was placed in the active site of the protein and alignment procedure attempted to establish favorable interactions. Once a single fragment was docked and all steric distortions were removed, the interaction energy was estimated with Böhm's algorithm.[39] This procedure was repeated for other fragments of the ligand, reconstructing them in an incremental manner. FlexX was docked with standard parameters allowing the program to automatically choose the core fragment and ring flexibility turned off. The ligand torsion angle model was taken from MIMUBA.[40]

Surflex

Surflex analyzes the protein active site to recreate possible contacts between a protein and a ligand. The idealized ligand structure for a specific protein is called a protomol, because it describes all active site contacts. Three different types of molecular fragments were used to create it: CH_4, C=O, and N—H. Then the ligand was fragmented into 1 up to 10 molecular fragments and each of them could have several rotatable bonds. Then, they were placed into the active site and underwent further conformational search to maximize the molecular similarity to the protomol. Finally, the ligand was recreated in an incremental way from the core fragments chosen by the program during fragment docking run. In our test the, protomol was generated based on ligand placement with threshold values set to 0.5. No additional starting conformations were generated and ring flexibility was not considered. Ten poses were saved to output the file per each ligand, the same as for other docking programs.

LigandFit

LigandFit program was developed by Venkatachalam et al.[24] Its major steps of docking include: identifying an active site, creating its grid using the cavity detection algorithm, and fitting a given ligand to a specified binding site through the Monte Carlo conformational sampling procedure and matching the ligand to grid points. Finally, those conformations undergo the rigid body energy minimization using the DockScore energy function. For each given complex, the binding site for docking was defined using the native binding pose of the ligand. The Monte Carlo parameters depended on the number of rotatable bonds in docked molecules. If there were two rotatable bonds the maximal number of trials to perform (per number of torsions), and the number of consecutive failed trials was 1000 and 240, respectively. If the molecule had 10 of those bonds the number increased to 20,000 and 5000, respectively. The number of poses that undergo rigid minimization was set to 100.

Glide

The docking process consists of four major steps. During the first 2 stages, the program uses a series of filters to search for possible locations of the ligand in the active site, and then to generate the best ligand binding poses through a coarse screening. The filter examines steric complementarity of the ligand to the protein and evaluates various ligand–protein interactions with the Glide Score function, which is a modified version of the ChemScore scoring function. Next, the ligand binding poses selected by the initial screening are minimized *in situ* with the OPLS-AA force field.[41] Finally, the composite score is used to rank the resulting ligand binding poses and select the ones to report by considering GlideScore. In our tests, we used the standard precision (SP) mode, as in others only modifications of GlideScore are made, not effecting docking algorithms. It should be noted that the number of atoms in the ligand cannot exceed 200 and the maximal number of rotatable bond is equal to 35.

eHiTS

The eHiTS program, unlike other fragment-based docking programs, does not use the incremental model, where the ligand is expanded by adding new fragments to the core. Instead, it attempts to find a global optimum based on individually docked fragments. The procedure consists of several steps. At the beginning, the grid is created inside the active site of the protein with equal spacing of 0.5 Å. Then the ligand is divided into a number of rigid fragments connected by flexible linkers, each of the fragments is independently docked to every possible place in the active site using geometrical criteria. For each docked fragment, the scoring is performed, but at this early point no fragments are discarded even if they do not fit well into the active site. Next, the hyper-graph detection algorithm matches compatible rigid fragments. Based on these results, the best-suited combination of fragments is chosen for further studies. Then the algorithm attempts to recreate the ligand structure by connecting ligand fragments with flexible linkers. Finally, the energy minimization is performed in the given active site of the receptor. Both torsion angles changes and rigid translation of the ligand are possible. eHiTS is the only program in our test that uses knowledge-based scoring functions to predict the ligand binding affinity. In our test, the number of cavity modes was between 200 and 600 points. The number of candidates to be selected as the cavity node center was no less than 200. The limit on the number of LigNode position in the cavity was 250 and the number of poses considered in each step of GraphMatch was 600.

Scoring Functions

During the docking procedure, a large number of poses is generated, thus the fast and reliable function that can estimate the strength of the interaction between the protein and the ligand is required. It is also crucial to select those conformations that are close to the native structure, so that contacts between the ligand and the protein are recreated with the same geometry as in the crystal. Scoring functions express the geometric matching of the two interacting molecules and the strength of this interaction, based on the physicochemical parameters of the system. The main complication of those functions is the estimation of the binding energy as the sum of used terms. Thus, a significant dependence between the size of the ligand and its score can be observed. In fact, large molecules that are able to create many more specific interactions like hydrogen bonds, usually obtain higher docking score.

More than 30 different scoring functions have been developed until 2009,[32,42–45] and they can be grouped in three major categories: force-field based methods,[43] empirical methods,[46] and knowledge-based (statistical) methods.[47] A short description of each class is provided below. In this work, we evaluated default scoring functions of the tested docking programs. They are: GoldScore for GOLD, LigScore for LigandFit, GlideScore in SP for Glide, Surflex score, eHiTS score, AutoDock score, and FlexX score. Description of each scoring functions categories as well as more detailed information about function used in this test is provided with Supporting Information.

Benchmarking Dataset

To perform effective redocking procedure, several conditions have to be fulfilled. The structure of a target protein solved experimentally at the atomic resolution should be found, or three-dimensional high-quality structural model based on only its sequence of amino acids should be prepared. In many cases, the protein structure is extracted from X-ray crystal of a protein–ligand complex, because it allows one to omit protein structural changes during the process of binding the same ligand. Additionally, the position of the active site is easy to determine. The collection of 1300 protein–ligand complexes from refined PDBbind 2007[48,49] was used as a test set in this evaluation (description of the refined set is given in the Supporting Information Table S1). The complete version of the database contains more than 3100 complexes, for which apart from structural data also data concerning ligands activity is provided, but the 'refined' set is ideal for our purpose. This is because there are several criteria that each complex must fulfill to be accepted for the refined set. Below, we present a brief description of those criteria.

1. Resolution of the protein–ligand complex must be below 2.5 Å. Previous studies by Jones et al.[20] showed that when using poor resolution structures, more incorrect conformations of ligands were generated. Proteins with chain breaks and unsolved region were also excluded. Also no structures solved using NMR were chosen for the refined set.
2. Activity of the complexes should be given as either pK_i (an inhibition constant) or pK_d (a dissociation constant).
3. Complexes with ligands containing other than standard atom types (like Be or Si) and those that are covalently bound with protein were excluded from the refined set. Moreover, ligand mass should not exceed 1000 amu. A complex was rejected if the distance between its ligand and the protein heavy atoms was closer than 2 Å.
4. The complex should have only one ligand in the active site.

Proteins extracted from crystal structures undergo the following preparation steps.

1. Hydrogen atoms were added with the protonation state simulated to pH = 7. Therefore, aspartate and glutamate amino acids were negatively charged, histidine was neutral and arginine and lysine amino acids were positively charged.
2. The terminal carboxyl groups were deprotonated, whereas amine groups were protonated.
3. Atoms and bonds types were inspected using Sybyl software, yet no geometry optimization was performed.
4. After performing initial tests, we decided to remove all water molecules and metals ions form the pdb input files for the purpose of our studies, since no significant change in docking accuracy was observed (more details concerning the status of the metal ion are given in the supplementary materials Table S2)).

Ligand Preparation

In our work five distinctive cases were taken into consideration. First, Redocking where the input structure of a selected ligand was identical with that obtained using crystallographic methods. In principle, it is the easiest task for docking engines, because the correct conformation of the ligand is given directly to

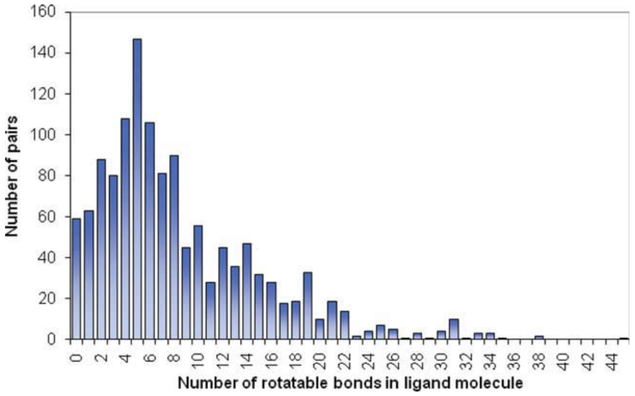

Figure 1. Histogram presenting distribution of ligands with specific number of rotatable bonds in PDBbind database. PDBbind database is diverse when it comes to distribution of rotatable bonds in ligands. As can be seen, most of them have between four and six rotatable bonds. Still the contribution of more rigid molecules with less than two bonds and more flexible ones with more than 15 bonds is quite significant. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

programs as the input. The second case called Corina one was implemented using Corina (ver. 3.4)[34] software to generate single lowest energy conformations. The same procedure was repeated using the Omega2 (ver. 2.3.2)[35] tool creating Omega one case. Those three cases address the question: has a ligand starting conformation any effect on docking results, as in the previously published results, docking X-ray native conformation usually resulted in a more accurate posing of the ligand. We also wanted to find out what would happen if we increase the number of input structures from 1 up to 10. To answer this problem of ensemble docking, we used once again Corina and Omega2 to generate 10 low-energy conformations analyzing two additional Corina ten and Omega ten cases. The total number of input conformations was chosen as a compromise between computational cost and quality of the obtained structures. We observed that docking 100 low-energy conformations did not improve the quality of docking for any method under consideration (unpublished results). In those four experimental setups, the converted 2D representation of the given ligand in SMILE[50] format was used as the input for prediction of the three-dimensional ligand structure.

Evaluation Methods

Performing the benchmark of docking software is a challenging task not only because there is a great number of programs available but also because of the countless data that needs to be processed. Another difficulty consists in applying different methods to interpret the results. Several research groups decided to use visual inspection of the obtained ligand conformations[51] to judge the docking tools performance. However, the empirical approach makes it difficult to compare their results with those obtained by others. That is why the majority of researchers decide to choose root mean square deviation (or root mean square distance) (RMSD) as the main parameter describing docking accuracy. It may not be ideal, but it is the only widely acceptable and reliable value, easy to recreate by others. Moreover, given the single input ligand many solutions are proposed by docking programs. Therefore, in our study, we decided to check not only the conformation with the highest docking score (called top score) but also the one that is the closest to the native structure (best pose). To evaluate the programs, the mean RMSD of those two poses obtained on the entire testing set was calculated. However, the average value of the RMSD can be heavily influenced by some "off the scale" results, thus we also calculated the number of successfully docked pairs. It is defined as the ratio of pairs for which top score or best pose conformations are below the given threshold in comparison with all evaluated pairs. In our case, the RMSD must be below 2 Å, which is the widely accepted[4,27] accuracy in docking predictions.

Another goal was to assess the quality of scoring functions predictions. There are two basic ways to identify the quality of those functions. We decided to compare the experimental values of binding affinity with the docking scores by calculating both Pearson and Spearman rank correlations. The character of our study does not allow us to calculate the second possible measurement of scoring functions quality, namely, the enrichment factors, as performing this type of experiment for all proteins in our test with sufficient randomized ligand database would be impossible within limits of our computational resources.

Biological and Chemical Diversity

Our test set composed of 1300 protein–ligand complexes shows that both protein and ligand properties are greatly diversified. The number of ligand rotatable bonds varies from 0 to 45, as can be seen in Figure 1, and its activity spans over 10 orders of magnitude, which is presented in Figure 2. Moreover, protein sequence demonstrates high redundancy as more than 400 protein clusters can be obtained when clustering the entire database at 90% sequence identity with the cd-hit tool.[52] Beside sequential diversity, also functional diversity can be observed, with different classes of enzymes and receptors participating in the test.

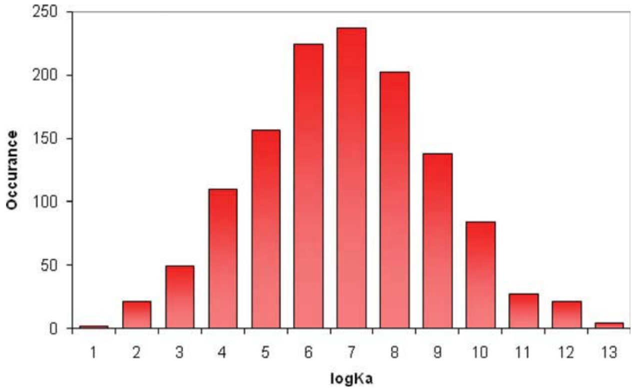

Figure 2. Distribution of ligands binding affinities in PDBbind database. Occurrence of binding affinities in PDBbind database shows that most of the ligands are bound to their protein target with moderate strength. However, in database, very strong binders can be observed with log K_a value over 10 as well as molecules whose binding to protein is quite weak with log K_a below 3. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Diversity in ligands structures on the primary set leads us to create various smaller benchmarking subgroups. First, the subset based on ligands rotatable bonds was created. Usually, the worst results in docking are obtained for large and more flexible molecules, due to the fact that conformational space increases exponentially with every new bond. Docking algorithms try to speed up the searching process by using different methods, for example, by applying filters, or using stochastic approaches like genetic algorithms. Yet, in the majority of evaluations, larger ligands perform significantly worse than smaller, more rigid molecules.[53] To confirm those previous observations that were usually reported on a limited number of complexes, we created two subsets called small and large. Ligands were qualified to the first category if they had no more than five rotatable bonds. Those two groups contained 651 and 649, respectively, protein–ligand complexes.

The other important distinctive features of ligands are their hydrophobic and hydrophilic characteristics. These are related to many important aspects of ligand behavior, mostly the ability to create hydrogen bonding with protein as well as forming interactions with hydrophobic cavern in the active site. The ability to cross the cell membrane by a molecule is a consequence of those properties. Therefore, we divided the entire dataset based on log P value (octanol/water partition coefficient), and calculated it using CLOGP program[54] that is the part of TRIPOS Sybyl software. We identified 645 hydrophilic and 655 hydrophobic ligands. The selected threshold, i.e., the value deciding where to classify each molecule, was set as 0, describing a molecule for which the concentrations in water and octanol are equal. In both cases, the proportions between small and large molecules within both hydrophobic and hydrophilic subsets were similar.

Third, we divided the ligands from PDBbind into three groups according to the strength of binding to the corresponding protein target. The first group ("strong") contained the ligands for which the concentration necessary to inhibit the protein activity was lower than 45 nM, the second group ("medium") which had their pK_a between 45 nM and 3.6 μM, and finally the inhibitors ("weak") for which the concentration of the compound to inhibit protein was greater than 3.6 μM. For those three groups, we calculated how many small and large molecules fall to each category to check if the results of the benchmarking procedure are based purely on the ligand binding strength and not on its size. The results showed that in the case of strong dataset there were 271 large ligands and 159 small ones, for medium dataset there were 213 large ligands and 222 small ones, whereas for weak dataset there were 165 large ligands and 270 small ones, respectively. That is why we additionally divided each of the previously created sets depending on the size of the ligands and in this way six independent categories were finally created.

Our last subset contained the complexes with proteins with co-crystallized short peptides or other protein-like molecules. To determine if the ligand can be classified to that category, we manually inspected all candidates. The molecule must have at least one peptide bond between alpha amino acids, sometimes not identical with those observed in living organisms. Structures with nonstandard atoms (all types except oxygen, hydrogen, carbon, and sulfur) were discarded, all except 14 complexes when phosphate atoms were present, and six complexes, when some fluorine atoms were found. In all cases, the presence of these atypical atoms was the result of cap treatment to prevent covalent bonding between the ligand and the receptor.

In our opinion creating the above subsets not only illustrates well the diversity of PDBbind database but also allowed us to fully evaluate the docking algorithms and to show in more detail their individual advantages and drawbacks.

Results and Discussion

In this section, we will present the results of our investigations on seven docking programs: AutoDock, eHiTS, FlexX, Glide, GOLD, LigandFit, and Surflex. First, we will discuss the ability of programs to predict the ligand binding pose. It is one of the crucial aspects of docking, as the better three-dimensional binding pose will be proposed, the contacts between the ligand and the receptor will be recreated in a more realistic way. Second, we will analyze the ability of the programs to correctly calculate *in vitro* binding affinity. Other capabilities of scoring functions, like the ability to correctly rank poses based on their RMSD value will also be mentioned.

Evaluation of Pose Prediction Capabilities of the Examined Programs

General Performances on the Entire Dataset

In this section, we will report the performances of seven docking programs on the entire dataset to show their general docking accuracy. In general, we can describe our test as the fivefold repeated redocking experiment, each time using different three-dimensional input ligand structure. Therefore, the repetition of experiments minimizes the errors that result from using stochastic approach when poses are generated. We will discuss here the influence of starting three-dimensional ligand conformations and the number of those initial conformations on final docking results.

The entire docking database proved to be a tough challenge for all the programs and in fact none of them produced output conformations for all 1300 pairs. The performance of Surflex, FlexX, LigandFit, eHiTS, and GOLD was reasonable, with no more than 30 failed complexes. Glide had problems with automation of the process of protein preparation (it additionally required the change of formats from pdb to mae), and with the limitation as concerns the number of rotatable bonds of the ligand, which was limited to 35, and the ligand size, which was limited to 200 atoms. Only 1170 (90% of the entire database) complexes overcame those restrains. AutoDock failed to dock nearly 90 pairs. Therefore, all statistics presented here for individual programs were calculated on complexes for which the results were obtained. The distribution of the mean RMSD value of top score conformations and percentage of successfully docked pairs, i.e., complexes for which top score RMSD was below 2 Å threshold, are shown in Figure 3.

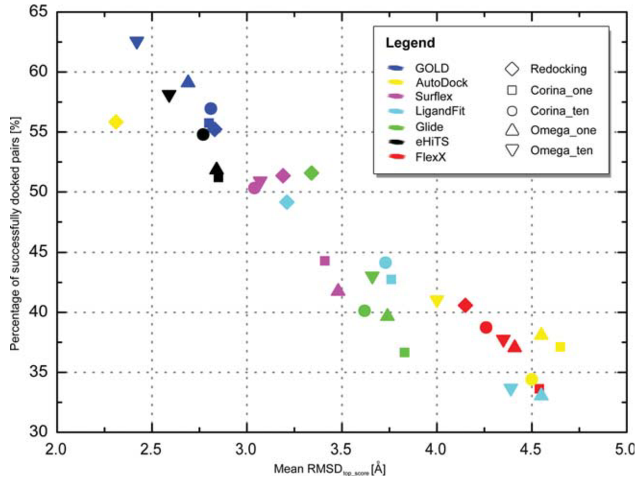

Figure 3. Docking accuracy of programs on 1300 protein–ligand complexes from refined set PDBbind 2007 database for top score conformations.

On the basis of those results, we can order programs in the following way: GOLD ~ eHiTS > Surflex > Glide > LigandFit > FlexX ~ AutoDock. The best programs have the average RMSD$_{\text{top score}}$ around 2.7 Å, and it increases to nearly 4.5 Å for the weakest FlexX. As expected, better results were observed for best pose conformations (Fig. 4). For those poses, the mean RMSD value was even below 2 Å for GOLD, eHiTS, and Surflex. For other programs, meaningful improvements were also observed in comparison with top score poses results. Nevertheless, the ability of correct posing by programs was measured rather by top score than best pose, which seems quite inadequate. The value of 3 Å for mean RMSD$_{\text{top score}}$ may result in a situation where important contacts between the ligand and the protein would be missed or at least their geometry significantly changed. Moreover, the percentage of pairs for which top score conformation is below 2 Å shows that even for the best programs the success rate is below 60%, and in some cases even below 40%. This means that for almost 600 different complexes most docking programs failed. Because no correct pose could be picked up in the first place, also the scoring function was unable to predict true binding affinities of ligands, as their contacts with the protein would be recreated incorrectly.

Although the results for best pose that are acceptable in the case of a few programs, yet this cannot change our negative opinion about the software, as the native pose is obviously not known before docking. It is virtually impossible to choose it manually, as millions of poses are generated during a typical virtual screening experiment. Moreover, top score conformations are rarely classified simultaneously as best pose, also their position is rather randomly distributed among all generated poses ordered by the docking score. It is not clear why there is such a significant difference between those two types of conformations. It could be the result of imperfection of the docking algorithm or the scoring function itself, which cannot correctly select the conformation of all generated conformations. It should also be emphasized that performing time consuming *in situ* optimization of top score conformation may not result in drastic improvement of pose RMSD, as it was pointed by Li et al.[53]

The Influence of Starting Conformations

Even more interesting results were observed when different starting conformations were used. It is common belief that if the input structure is similar to the native one, the better poses are predicted by docking programs.[27] To verify this opinion, we repeated our experiment using the following conformations: one identical with that from X-ray (case called redocking) and two others generated using popular software Corina and Omega2 (cases called Corina one and Omega one, respectively). Although those two were designed to recreate the three-dimensional structure of a ligand, however, not its bound conformation, sometimes output poses can be very similar to the conformation of a ligand in the protein complex. Figure 5 presents a histogram of RMSD values obtained for all conformers generated by all docking programs. The RMSD distribution for the generated poses is preserved for all cases regardless of the type of used input structure. Interestingly, most conformations fall into RMSD category between 0 Å and 2 Å, which is the positive tendency because the mean RMSD$_{\text{top score}}$ values on the entire set fall rather between 3 Å and 5 Å. This seems to support the opinion that programs are more accurate in probing search space than it was concluded previously. In the case of redocking, the majority of conformations have RMSD between 0 and 1 Å, but there were more conformations with much higher RMSD (of over 10 Å) than when Corina and Omega2 input conformations were used. Yet, in our opinion, this can be explained by the fact that an old version of eHiTS was used in that case, which failed to produce valid conformations. For eHiTS, when older 6.2 version of program was used mean RMSD$_{\text{top score}}$ exceeded 10 Å. That situation would place eHiTS as worst software in our work. However, for four other cases, i.e., Corina one, Corina ten, Omega one, and Omega ten, eHiTS 9.0 was used. New version introduces many new features one most important in our

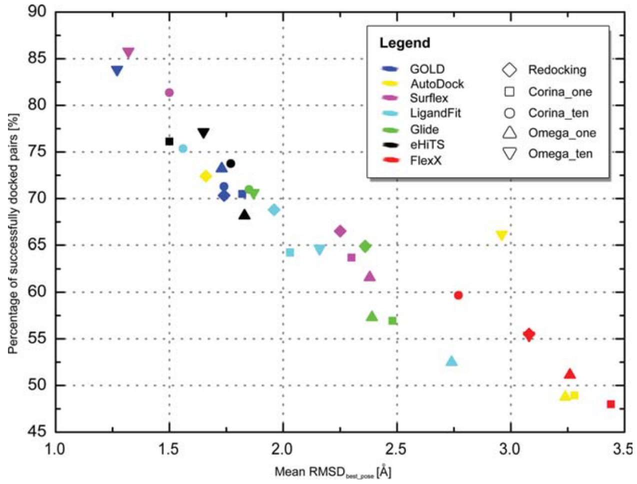

Figure 4. Docking accuracy of programs on 1300 protein–ligand complexes from refined set PDBbind 2007 database for best pose conformations.

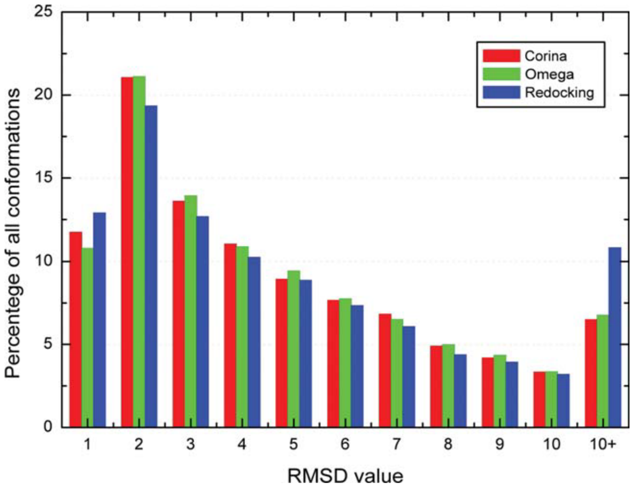

Figure 5. RMSD histogram of all conformations predicted by seven docking programs. To investigate the influence of using different starting conformations on docking algorithm accuracy, we used three different input structures. One of them was identical to that from X-ray (Redocking), and two others were generated using either Corina (Corina) or Omega2 (Omega) software. For all generated conformations by seven docking programs, we calculated RMSD to native ligand structure. This histogram presents the distribution of that RMSD for all conformations. No significant improvement can be observed when using X-ray conformations, so the programs seem to be unaffected by different starting conformations.

Table 1. Influence of Starting Conformation on Mean RMSD$_{\text{top score}}$ Value.

	Mean value	Corina	Omega	Redocking
eHiTS	2.80[a]	0.05	0.04	7.36[b]
FlexX	4.37	0.17	0.04	−0.22
Glide	3.64	0.19	0.10	−0.30
GOLD	2.77	0.03	−0.08	0.06
LigandFit	3.84	−0.08	0.71	−0.63
Surflex	3.36	0.05	0.12	−0.17
AutoDock	3.84	0.81	0.71	−1.53

We present here the mean RMSD$_{\text{top score}}$ value for all programs when different starting conformation were used. Mean value is obtained by calculating mean arithmetical value for those three cases, namely, Corina, Omega, and Redocking. Numbers in column Corina, Omega, and Redocking were calculated by subtracting RMSD value for particular column from mean value. Gray color indicates situation were number is greater than 0.3 value or smaller than −0.3.

[a] Calculated for Corina and Omega cases using 9.0 version of eHiTS program.

[b] Obtained using older 6.2 version of eHiTS program.

opinion new knowledge-based scoring function. New version of program is proven to be one among best in our evaluation. It seems that in fact using X-ray structure may provide better sampling of conformational space, yet it does not necessarily transfer into obvious boost of software overall performance. Analyzing the averaged deviations from the mean RMSD of the top score conformations for Corina, Omega2, and X-ray datasets (Table 1), it can be concluded that using the native conformation does not necessarily result in much higher accuracy. The deviations from the mean value are usually not greater than 0.3 for most programs. Only two of them seem to be strongly affected, namely, AutoDock and LigandFit. For the first 1, this dependence can be observed probably because of changing default parameters of the genetic algorithm. In the case of Corina and Omega2 runs, we decreased the number of allowed genetic operations to speed up the docking process. In the case of LigandFit, strong dependence on the quality of the input structure had already been reported by others.[53] The nature of the algorithm, which is based on shape complementarity between the ligand structure and the active site, would suggest strong dependence between the quality of input three-dimensional structure and the docking results.

Docking Ensemble of Ligand Conformations

The final experiment addressing the quality of input structure was to increase the number of docked conformations from a single 1 up to 10. We asked if the docking of such ensemble instead of the single lowest energy ligand conformation would produce better docking results. Additionally, we wanted to find out if the ensemble docking procedure, obviously allowing for

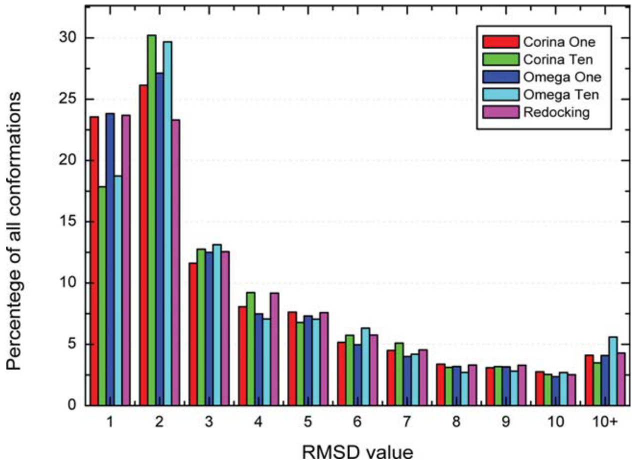

Figure 6. RMSD histogram of all conformations predicted by GOLD. To investigate the influence of ligands on docking ensemble, we once again used Corina and Omega2 programs to generate this time 10 different lowest energy conformations for all ligands in PDBbind database and those creating five distinctive cases. First Redocking with one ligand identical to native structure, Corina one and Omega one where the lowest energy conformation was used as an input, and finally Corina ten and Omega ten where the ensemble of structures was used using Corina and Omega software, respectively. As can be seen, no significant improvement on the quality of poses can be observed when more input conformations are used. The number of poses with acceptable RMSD between 0 Å and 2 Å is very similar. Thus, using the ensemble for docking with only one program seems to give no benefit.

Table 2. Results for Docking Programs Top Score Conformations on Various Subsets Created Based on Physicochemical Properties of Ligands.

	Small		Large		Hydrophilic		Hydrophobic		Proteins	
	RMSD	%	RMSD	%	RMSD	%	RMSD	%	RMSD	%
eHiTS	1.96	64.80	3.59	37.71	2.62	50.96	2.91	48.33	4.51	34.95
FlexX	3.04	49.03	5.64	26.04	3.73	43.76	5.07	31.40	6.12	32.71
Glide	2.91	49.58	4.39	34.49	3.61	43.78	3.81	40.73	4.03	46.50
GOLD	1.96	67.11	3.50	48.45	2.30	65.67	3.06	50.28	4.78	36.29
LigandFit	2.79	51.31	5.23	28.44	3.97	38.42	4.22	42.76	6.28	23.57
Surflex	2.71	53.94	3.95	41.28	2.88	51.29	3.61	44.17	4.06	46.86
AutoDock	2.11	61.05	5.43	27.46	3.70	50.43	5.48	37.00	8.99	7.77

Here we present results for docking programs on various subset created from PDBbind database based on various physicochemical properties of ligands. Two measurements are used: RMSD, which is mean "RMSD" value of all 1300 top score poses generated and "%," which informs about percentage of pairs for which top score conformation had RMSD below 2 Å. The created subsets are based on number of ligands rotatable bonds (small and large), hydrophobic properties (hydrophilic and hydrophobic), and separately for ligands that are small proteins, or peptides.

better searching over the conformational space by probing it from different starting points, is worth the extra computational time, as the docking takes 10 times more in this case. We used Corina and Omega2 software to create up to 10 conformations per each ligand. Surprisingly, even before docking for more than 80% of the complexes at least one conformation of the ligand itself generated by those programs was found to be very close to the native one with the RMSD value below 2 Å.

To estimate if the better conformational space searching was achieved, we created RMSD histograms for each program. Figure 6 presents the results obtained for GOLD. To easily compare the results of various testing cases, the column of histograms represented the percentage of conformation to fall for a specific RMSD category. It is clear that no significant improvement in quality of the generated poses can be observed. Most conformations have RMSD between 0 Å and 2 Å, the same as when using only a single structure as the input. Only a small decrease in the number of incorrectly generated conformations can be observed. Yet, in general following that procedure, all programs were able to further decrease both the mean RMSD top score and the percentage of successfully docked pairs. Unfortunately, the results for each docking case and those presented in Figure 3 in most cases differ only slightly and are not significant enough to change the overall programs performances. Only Surflex was gaining much from the increased number of input structures with more than 10% increase of successfully docked pairs. However, we are not sure what part of the algorithm is responsible for this change, especially because other fragment-based methods did not provide such boost in docking accuracy. More encouraging results for all the programs were observed for the best pose conformation (Fig. 4), where in most cases the mean RMSD values decreased dramatically in comparison with the cases when only a single input structure was used. This can be explained in two ways. One of the possible explanations is that using more structures results in better conformational space searching. Another explanation that seems very probable is that for each docking input we saved 10 output conformations, therefore, a 100 predicted conformations instead of 10 were available for each protein–ligand pair. This may result in preserving conformations that would be discarded when only 10 output conformations were saved, because of their low docking score. Nevertheless, best pose conformations, as mentioned before, cannot be used as the objective measurement of docking accuracy especially because their positions within the ordered list of poses are usually random. The analysis of top score suggests that the increased number of input structures does not produce better results even though conformational space is searched more extensively.

Results for Physicochemical Subsets

The results discussed above show the general performance of docking software on the entire PDBbind database. Because of the diversity of ligands structures in the primary dataset, the number of subsets, based on physicochemical properties of the molecules, could be created as discussed above. Exploring features like ligand flexibility or hydrophobic potential may give a more detailed insight into the docking algorithm, pointing out desirable software to solve practical biological problems.

In the first test, we created two subsets: one consisting of small molecules, which were quite rigid (up to five rotatable bonds), and the other consisting of large molecules with more than five rotatable bonds. In the majority of previously published evaluations,[4] it was pointed out that the main problem that has to be addressed is to increase the successful rate for docking of flexible molecules, because the significant disproportion between flexible and rigid molecules is observed. Our results summarized in Tables 2 and 3 clearly support this conclusion, as the docking programs still lack that important ability. The mean $\mathrm{RMSD}_{\mathrm{top\ score}}$ was increased by almost 2 Å for large molecules, and the percentage of successfully docked pairs decreased in some cases by nearly 30%. Unfortunately, the success rate for the set of large molecules, for all programs, was always below 50%. It seems that flexible ligands still pose quite a challenge for docking software. Surprisingly, the program that suffered less from transition between those two subsets is the fragment-based

Table 3. Results for Docking Programs Best Pose Conformations on Various Subsets Created Based on Physicochemical Properties of Ligands.

	Small		Large		Hydrophilic		Hydrophobic		Proteins	
	RMSD	%	RMSD	%	RMSD	%	RMSD	%	RMSD	%
eHiTS	1.2	83.3	2.9	59.6	1.6	72.3	1.7	72.4	3.0	45.1
FlexX	2.0	68.2	4.3	39.6	2.6	60.7	3.7	47.3	4.9	40.6
Glide	1.74	73.01	2.66	54.80	2.0	66.9	2.3	61.5	2.6	61.8
GOLD	1.31	81.07	2.01	66.38	1.4	79.1	1.8	68.7	2.7	49.9
LigandFit	1.58	76.65	2.66	52.33	2.1	62.0	2.1	68.3	2.8	38.7
Surflex	1.50	78.66	2.42	64.82	1.71	74.58	2.04	69.04	2.82	62.84
AutoDock	1.43	77.58	3.92	37.23	2.59	63.74	3.93	49.77	8.63	12.04

Surflex, with only 12% change in the number of pairs for which top score conformation has RMSD below 2 Å. Incremental recreation of ligands from smaller fragments seems to be working for fragment-based programs, as also eHiTS results did not decrease so dramatically. We can also confirm that quality for Glide, which had already been reported by Perola et al.[4]

Inability of programs to predict correct poses for flexible ligands influences directly the docking of protein-like molecules (column peptide in Tables 2 and 3). Oligopeptides are usually large polymer-like structures, with a great number of rotatable bonds. On this specific subset, described above, dramatically low results can be observed. Some programs like AutoDock were completely missing true conformations for all given cases. Yet, Surflex and Glide, confirming previous observations, once again achieved the only acceptable accuracy.

The second test analyzed the impact of hydrophobic properties of ligands on docking results. Usually, a small number of hydrogen bonds can be created with the protein by hydrophobic molecules, and those are used as the leading feature, whereas conformations are predicted by docking programs. In fact, it is commonly accepted that docking hydrophobic molecules and predicting their activates using docking are quite a challenging task. The log *P* value calculated using CLOGP software available in Sybyl was chosen to approach this problem. The results presented in Tables 2 and 3 clearly confirm those observations. However, the drop in accuracy is not so dramatic at it might be suggested by the common opinion. Even in the case of LigandFit, the overall performance on lipophilic molecules is very similar to the lipophobic ones. Similarly, only a slight transition was observed for eHiTS and Glide. GOLD was more influenced, yet this is clearly the consequence of its algorithm that puts special emphasis on hydrogen bonds during docking. The lack of this crucial component reduced the algorithm effectiveness by almost 15%.

The final dataset was selected using the ligand binding strength to its corresponding protein target, with further dividing them by the flexibility of molecules (results in Tables 4 and 5). Each of three original categories—strong, medium, and weak—was therefore further divided into small and large subsets. Our primary objective for this procedure was to determine if there is a single preferable class of ligands, for example, small strongly bound ones. Surprisingly, the predictions of poses were similar for all strong, medium, and weak datasets. For Surflex, eHiTS, Glide, and FlexX, the mean RMSD values did not change much regardless of the molecule falling to small or large classes of ligands. The difference for those programs was less than 0.4 Å for small molecules, and not more than 0.8 Å for large ones. The larger changes could be observed in the percentage of successfully docked pairs, but the divergence was not more than 6% for the small subset and 8% for the large one. The most unaffected program of those four was again Surflex where the differences were even smaller. Interestingly, the results for the

Table 4. Results for Docking Programs Top Score Conformations Based on Strength of Ligand Binding to its Corresponding Protein.

	Strong				Medium				Weak			
	Small		Large		Small		Large		Small		Large	
	RMSD	%	RMSD	%	RMSD	%	RMSD	%	RMSD	%	RMSD	%
eHiTS	2.0	66.4	3.6	44.6	2.0	62.9	3.7	41.7	1.9	68.6	3.5	40.7
FlexX	3.4	48.6	6.1	23.5	3.0	50.4	5.6	27.9	3.0	46.9	5.3	24.1
Glide	3.0	50.4	4.7	33.2	2.9	47.8	4.0	35.5	3.2	44.6	4.8	25.8
GOLD	2.2	61.9	3.8	45.0	1.7	73.6	3.6	48.3	2.0	66.4	2.8	56.6
LigandFit	2.6	58.1	6.0	19.2	2.5	52.9	6.1	16.9	2.8	50.0	4.8	33.0
Surflex	2.5	57.6	4.2	38.9	2.5	53.9	3.9	40.2	2.5	52.3	3.8	39.1
AutoDock	2.7	52.1	7.4	16.2	2.6	49.5	6.7	19.2	2.5	54.5	5.0	29.2

Table 5. Results for Docking Programs Best Pose Conformations Based on Strength of Ligand Binding to its Corresponding Protein.

	Strong				Medium				Weak			
	Small		Large		Small		Large		Small		Large	
	RMSD	%	RMSD	%	RMSD	%	RMSD	%	RMSD	%	RMSD	%
eHiTS	1.2	88.3	2.1	64.2	1.2	85.6	2.3	60.0	1.2	85.7	2.1	63.3
FlexX	2.4	68.1	4.7	36.0	1.9	69.7	4.4	42.1	1.9	66.9	3.6	39.7
Glide	1.7	75.7	2.7	54.3	1.7	74.8	2.4	56.6	1.8	69.5	2.6	51.5
GOLD	1.4	79.6	2.1	65.1	1.1	85.3	2.1	65.9	1.4	80.3	1.7	72.5
LigandFit	1.3	85.0	2.9	44.6	1.4	80.9	3.1	43.5	1.6	73.9	2.4	57.1
Surflex	1.5	79.7	2.4	64.7	1.4	80.6	2.4	65.6	1.4	81.8	2.1	65.7
AutoDock	1.4	80.3	5.6	22.1	1.7	75.5	4.7	23.4	1.6	77.2	3.2	43.8

small subset of ligands in the case of strong, medium, and weak subsets were identical. GOLD and LigandFit follow a similar trend, i.e., for large ligands the weak ones are better predicted than the strong and medium ones, with almost 10% increase in docking accuracy and drop of the mean RMSD larger than 1 Å. Unfortunately, GOLD seems to prefer weak and medium binding ligands to the strongly bound ones, also for the small dataset. The results showed that for strong small ligands the drop in successfully docked pairs in comparison with medium small or weak small was nearly 10%. For LigandFit, this trend was much more preferable with small strong ligands achieving nearly 60% of docking accuracy, whereas medium and weak ones achieved only 50%. Summing up, the results in this final tests showed that the ligand binding strength did not differentiate docking software. Molecules with a similar number of rotatable bonds were predicted with the same accuracy regardless of their binding strength.

Evaluation of the Scoring Functions

Correlation with the Experimental Binding Affinities

The second important ability of docking programs is to predict the strength of binding the ligand to the protein target; preferably as *in vivo* to perform this challenging task, the scoring functions are used. As mentioned previously, many types of them were proposed, but still all of them use relatively simple equations to calculate a docking score. The main advantage of this approach is its speed, due to relative simplicity of the functions. However, when comparing the experimentally derived binding affinities with the calculated docking score, weak correlations were usually obtained.[31] This drawback is the reason why docking software is only a supporting tool in the drug design process. Our goal was to determine the current status of scoring functions that are distributed together with docking software. Recently, some promising results were reported in evaluation by Cheng et al.[30] on the dataset of 195 protein–ligand complexes. In that study, X-Score functions achieved high correlations close to 0.7 and correctly predicted three-dimensional conformations for more than 70% of the complexes. Similar percentage of the successfully predicted pairs was, however, achieved also by GoldScore function that resulted in relatively weak correlation with the experimental binding affinities close to 0.3. Thus, it seems that there is no direct transition from good "score prediction" to good "pose prediction."

In our study, we evaluated scoring functions using three quality parameters for the entire set of 1300 protein–ligand complexes. First, Pearson correlation was calculated between docking score and experimental binding affinities, as PDBbind collects those values for all complexes. Second, Spearman correlation coefficient was calculated that is determined based on rank obtained by the ligands rather than the actual values of the score. This situation is similar to virtual screening experiments, where crucial is the position of compounds within the ordered list rather than their relative activity. All correlations were calculated for both top score and best pose conformations. Finally, we compared all scores for each ligand conformations that were generated, usually 10 up to 100, with their RMSD values. In that way, we checked if the high-scored conformations have smaller RMSD value. We were interested to find out if the ability of the scoring function can explain the order of programs in the "pose prediction" test reported above. All data given by seven functions are collected in Table 6. Because five distinctive cases for input structure preparation were studied, the presented numbers are the average values of the results obtained on each of them. It was previously reported that there is a strong correlation between the obtained results for scoring functions, and the properties of the molecules.[31] Thus, subsets from the "pose prediction" study, grouping ligands based on different physicochemical properties, were also used in the "scoring prediction" test. Here, only Pearson correlation between the docking score and binding affinity was calculated.

Three distinct groups of scoring functions emerge from our study based on their performance. The first one is composed of functions implemented in eHiTS and in Surflex, which gave Pearson correlation 0.38 and 0.33, respectively. Moreover, for eHiTS scoring functions very high-Spearman correlation was obtained, proving good ranking ability of that function, which is the quality useful in virtual screening experiment. The results from the small and large subsets show that both functions, as well as all others in this test, prefer small, rather rigid molecules. Their Pearson correlations are much higher than for large

Table 6. Results for Scoring Functions Used in This Work.

	Top score						Best pose						
	Pearson correlation	Spearman correlation	Small	Large	Hydrophobic	Hydrophilic	Pearson correlation	Spearman correlation	Small	Large	Hydrophobic	Hydrophilic	$\langle r \rangle$
eHiTS	0.38	0.47	0.43	0.31	0.35	0.29	0.29	0.39	0.40	0.23	0.26	0.26	0.28
FlexX	0.10	0.06	0.11	0.01	0.13	0.30	0.09	0.07	0.12	0.02	0.11	0.31	0.14
Glide	0.25	0.26	0.37	0.09	0.30	0.28	0.23	0.23	0.36	0.10	0.24	0.24	0.23
GOLD	0.17	0.18	0.26	0.08	0.24	0.21	0.06	0.12	0.24	0.07	0.12	0.14	0.17
LigandFit	0.11	0.04	0.11	0.06	0.20	0.04	0.08	0.07	0.09	0.08	0.21	0.10	0.10
Surflex	0.33	0.34	0.32	0.18	0.33	0.34	0.22	0.31	0.28	0.12	0.22	0.32	0.19
AutoDock	0.25	0.27	0.50	0.15	0.27	0.21	0.19	0.20	0.50	0.14	0.20	0.20	0.32

Here, we present results for scoring functions used as defaults in seven programs we tested. As five distinctive cases were studied in pose prediction test, with different starting conformation we were able to calculate average performance of function in all this tests and those values are present in this table. Performance on entire test set for both top score and best pose conformations are put in Pearson and Spearman correlations column. Second, Pearson correlation was calculated in regard to specific type of ligands, namely, more rigid ones (small), more flexible (large), and for hydrophobic (hydrophobic) and hydrophilic (hydrophilic).

molecules. This can be explained by the fact that all functions calculate the final score as the sum of all contacts between the ligand and the receptor. Thus, higher scores are usually obtained for large ligands, which obviously can create more contacts with a protein target. This may lead to overestimation of the docking score of that type of ligands in comparison with smaller ones. Additionally, scoring functions do not include ADME properties of ligands, which is the crucial aspect of *in vivo* ligand activity. It is much harder for large molecules to cross cell membranes and get to its protein target, which can considerably decrease their real activity. Those reasons probably push the docking score even further from reality. Interestingly, hydrophobic properties of ligands seem to have a smaller effect on scores correlation than it might be expected. For most functions, the lipophilic ligands activity was predicted with similar accuracy as the hydrophilic one. It seems that adding various parameters describing hydrophobic contacts during scoring functions development is indeed a good strategy.

To sum up, the scoring functions can measure hydrophobic interactions, at least as well as for typical hydrophilic ones—hydrogen bonds and electrostatic contacts. Similar scoring quality does not result in better prediction of hydrophobic ligands conformations, as we have already pointed out. Nevertheless, the best functions in the presented evaluation achieved relatively weak correlations, and therefore, using either of them would probably lead to wrong conclusions regardless of ligands properties. The scores of conformations with the lowest RMSD value, which are *de facto* most similar to native, also correlate poorly with the experimental binding affinity. Moreover, it is even worse than that of the top score. This suggests that even if programs were able to recreate the ideal active site conformation, it would not have much effect on the predicted binding affinity, and in consequence on *in silico* drug development process.

The second group of scoring functions is the modified ChemScore function used in Glide in SP mode, GoldScore from GOLD, and AutoDock scoring function. They achieve correlation between 0.17 and 0.25 for the entire dataset. No significant improvement can be observed when only ranks are taken under consideration, which proves that those functions have only limited usefulness in real life virtual screening experiments.

The last group consists of LigScore and FlexX score that produced dramatically low overall results below 0.1, probably because both functions prefer specific types of molecules. LigandFit produced correlation at the level of 0.2 for hydrophobic molecules, even though the function was based only on van der Waals and polar interaction between the ligand and the protein. Thus, the entire scoring contribution for those specific ligands was based only on steric fitness to protein. FlexX is more puzzling because of its strong hydrophilic ligands preference, as the terms typical of other empirical functions were explored, but also hydrophobic contribution was part of the score.

Correlation with the Quality of Poses

Finally, we wanted to check if the functions could order conformations correctly based on their RMSD value to the native structure, therefore, higher score would be obtained for better poses. The higher was that correlation the better those programs should perform in the pose prediction test, because scoring functions were used not only to evaluate output conformations but also to guide algorithms during the docking process. For example, although genetic algorithms are used to generate new conformations, it is the scoring function score that decides which of them will be accepted for the next run, whereas the weak-scored ones are discarded. In the case when only a single output conformation was generated for a given protein–ligand complex, such a pair was excluded from this part of evaluation. This was done to avoid artificial improvement of results for such programs as for those pairs an ideal correlation would be calculated. The results are summarized in Table 4 in column named $\langle r \rangle$.

Unfortunately, there is no single function that was able to rank the generated poses with sufficient accuracy based on ligands RMSD. However, the position of programs in the pose prediction test is similar to the position of function in this experiment. AutoDock achieved the highest 0.32 correlation, although performing only moderately in the previous two "pose prediction" and "score prediction" tests. The best correlations were obtained in the redocking case (0.38 for detailed results for each case; see Supporting Information), confirming that after expanding sampling time AutoDock performance improves considerably. The second program was not surprisingly eHiTS, with averaged correlation close to 0.3, proving that with the increased value of that score, better conformations are usually produced. The worst correlation of FlexX and LigandFit functions might be the reason for their poor performance in previous "pose prediction" tests. Unfortunately, this paragraph points out another defect of functions that is their inability to choose from predicted conformations the closest to the native one.

Concluding, we confirm that scoring functions cannot recreate ligands true binding affinities and rank them according to the RMSD to the native structure. Scoring functions presented in this evaluation achieved similar, yet only moderate results. This is not surprising, because all of them belong to the same class of empirically based scoring functions. They explore similar parameters, like van der Waals repulsive hydrogen bonds and electrostatic terms. Although the authors of those functions claim to obtain much higher correlations for their functions on their training datasets, yet functions perform significantly poorer on our large and diverse group of protein–ligand complexes. The performance of eHiTS and AutoDock scoring functions in our test is also discouraging, as neither of them achieved the correlation between docking score and experimental binding affinity above 0.5. Even more, those functions are used as default ones in most of the popular modeling software packages used by many research groups, like Surflex that is now part of SYBYL platform. Although those programs allow one to use other functions like F-Score or ChemScore; however, they were evaluated[30] with similar conclusions as in our tests. Thus, using them instead of Surflex default scoring function would probably produce similar misleading results on the ligand binding affinity. On the other hand, in the evaluation of scoring functions on the representative set of PDBbind 2007 by Cheng et al.,[30] the distinct class of knowledge-based functions were tested with better results, which might suggest that they are presently the leading candidates for a more universal score. The weak RMSD/score correlations suggest that the additive nature of scoring functions and the quest to maximize preferable contacts, like hydrogen bonds, over other terms may result in less effective scoring and pose prediction.

Conclusions

Summing up, our goal was to perform extensive large-scale docking software benchmark that has substantial improvement over previously published studies. First, seven programs were compared under the same conditions, providing valuable data that may be helpful to choose the optimal software. It is well known that evaluations performed by different groups, even on the same set of pairs and using the same software might result in dramatically different conclusions thus making it hard to choose the optimal program. The software that we benchmarked is typically part of popular molecular modeling platforms like Sybyl, Maestro, or Discovery Studio, making it very likely to be considered by various research groups. Second, the set of 1300 protein–ligand complexes allows us not only to compare performance on the diverse dataset but also to create various subsets based on physicochemical properties of the ligands. Moreover, the entire database was redocked five times to check the influence of different starting conformations on final docking results.

In our study, both pose prediction and scoring capabilities were measured. Two most successful programs were eHiTS and GOLD for which nearly 60% of the complexes have their top score conformations below 2 Å threshold. Programs like Glide, AutoDock, and Surflex achieved results around 50%. Interestingly, it made practically no difference if the starting conformation was or was not close to the native pose extracted from crystallographic data. Unfortunately, there is very weak correlation between the docking score and *in vitro* measured activity of ligands. A comparison of those two values provides the correlation of about 0.4 for the best program, namely, eHiTS. Other software provides even worst results, like AutoDock with the correlation close to 0.1. Functions also have poor ability to order the same compounds based on their RMSD to the native structure. Here, the performance was similar across all functions with the correlation value close to 0.5.

There was no single program that consistently outperformed all others. Good pose prediction did not always correlate with good scoring. Programs that use genetic algorithms seem to be the best choice for the pose prediction. Yet, due to the nature of the algorithm, docking takes much longer time (for GOLD almost 305 s/ligand) than other types of algorithms. Both GoldScore and AutoDock scoring are very weak in recreating experimental binding affinities.

Thus, can we trust docking programs? The answer must be given individually for two aspects of docking programs. In terms of pose prediction, we can say that GOLD and eHiTS performance is accurate enough but still there is ground for improvement. Nevertheless, 60% docking accuracy is a reasonable number. In the case of scoring functions, the answer must be negative, as virtually no correlations could be observed between the docking score and *in vitro* binding affinities. Using terms as they are now implemented in scoring functions would seem to be a good direction. Yet, the empirically derived functions have now reached the saturation of year-to-year improvement, because their performance has not changed dramatically in the last few years. The future direction should be either to use statistical approach based on increasing number of X-ray protein–ligand complexes, as can be determined from the results obtained by eHiTS scoring functions, or to develop completely new approaches in terms of predicting *in vivo* activity of the ligand.

We would like to point out that the best solution in future would be to combine various techniques into a single consensus approach that could benefit from using different approaches proposed by single docking methods. In the second part of our work, we will describe in detail the new algorithm, VoteDock for combining the results from multiple docking resources into a single and consistent prediction. VoteDock can substantially increase both the number of correctly predicted conformations (by at least 10%) and their scoring capabilities (by more than 0.2).

Acknowledgments

The authors thank Prof. Iwona Wawer, Ph.D., and Kamil Steczkiewicz for fruitful discussions and the help during the preparation of the manuscript.

References

1. von Grotthuss, M.; Koczyk, G.; Pas, J.; Wyrwicz, L. S.; Rychlewski, L. Comb Chem High Throughput Screen 2004, 7, 757.
2. Hus, J. C.; Marion, D.; Blackledge, M. J Mol Biol 2000, 298, 927.
3. Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. Acta Crystallogr Sect D 2002, 58(Part 6), 899.
4. Perola, E.; Walters, W. P.; Charifson, P. S. Proteins 2004, 56, 235.
5. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Br J Pharmacol 2008, 153 (Suppl 1), S7.
6. Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. J Chem Inf Model 2006, 46, 401.
7. Bissantz, C.; Folkers, G.; Rognan, D. J Med Chem 2000, 43, 4759.
8. Perola, E.; Walters, W. P.; Charifson, P. J Chem Inf Model 2007, 47, 251.
9. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Proteins 2004, 57, 225.
10. Vajda, S.; Kozakov, D. Curr Opin Struct Biol 2009, 19, 164.
11. Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson, H. J. Proteins 2008, 73, 271.
12. Ritchie, D. W. Curr Protein Pept Sci 2008, 9, 1.
13. Holt, P. A.; Chaires, J. B.; Trent, J. O. J Chem Inf Model 2008, 48, 1602.
14. Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. J Mol Biol 1997, 266, 424.
15. Carlsson, J.; Boukharta, L.; Aqvist, J. J Med Chem 2008, 51, 2648.
16. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. J Comput Aided Mol Des 2001, 15, 411.
17. Morris, M. J Comput Chem 1998, 19, 1639.
18. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. J Mol Biol 1996, 261, 470.
19. Jain, A. N. J Med Chem 2003, 46, 499.
20. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. J Mol Biol 1997, 267, 727.
21. Abagyan, R. A.; Totrov, M. M.; Kuznetsov, D. A. J Comput Chem 1994, 15, 488.
22. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. J Med Chem 2004, 47, 1739.
23. Wu, G.; Robertson, D. H.; Brooks, C. L., III; Vieth, M. J Comput Chem 2003, 24, 1549.
24. Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. J Mol Graph Model 2003, 21, 289.
25. Liu, M.; Wang, S. J Comput Aided Mol Des 1999, 13, 435.
26. Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., III. J Comput Aided Mol Des 2003, 17, 755.
27. Onodera, K.; Satou, K.; Hirota, H. J Chem Inf Model 2007, 47, 1609.
28. Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. J Med Chem 2004, 47, 558.
29. Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N. Proteins 2007, 69, 160.
30. Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. J Chem Inf Model 2009, 49, 1079.
31. Wang, R.; Lu, Y.; Fang, X.; Wang, S. J Chem Inf Comput Sci 2004, 44, 2114.
32. Teramoto, R.; Fukunishi, H. J Chem Inf Model 2008, 48, 288.
33. Bar-Haim, S.; Aharon, A.; Ben-Moshe, T.; Marantz, Y.; Senderowitz, H. J Chem Inf Model 2009, 49, 623.
34. Sadowski, J.; Schwab, C.; Gasteiger, J. CORINA, Version 3.4; Available at: http://www.mol-net.de.
35. Bostrom, J.; Greenwood, J. R.; Gottfries, J. J Mol Graph Model 2003, 21, 449.
36. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. J Mol Graph Model 2007, 26, 198.
37. Mooij, W. T.; Verdonk, M. L. Proteins 2005, 61, 272.
38. Klebe, G.; Mietzner, T. J Comput Aided Mol Des 1994, 8, 583.
39. Bohm, H. J. J Comput Aided Mol Des 1994, 8, 243.
40. Klebe, G.; Mietzner, T.; Weber, F. J Comput Aided Mol Des 1999, 13, 35.
41. Jorgensen, W. L.; Tirado-Rives, J. J Am Chem Soc 1988, 110, 1657.
42. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, B. S.; Johnson, A. P. Curr Protein Pept Sci 2006, 7, 421.
43. Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. J Comput Chem 1992, 13, 505.
44. Wang, R.; Lu, Y.; Wang, S. J Med Chem 2003, 46, 2287.
45. Perez, C.; Ortiz, A. R. J Med Chem 2001, 44, 3768.
46. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. J Comput Aided Mol Des 1997, 11, 425.
47. Gohlke, H.; Hendlich, M.; Klebe, G. J Mol Biol 2000, 295, 337.
48. Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. J Med Chem 2005, 48, 4111.
49. Wang, R.; Fang, X.; Lu, Y.; Wang, S. J Med Chem 2004, 47, 2977.
50. Weininger, D. J Chem Inf Comput Sci 1988, 28, 31.
51. Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. J Chem Inf Comput Sci 2004, 44, 871.
52. Li, W.; Godzik, A. Bioinformatics 2006, 22, 1658.
53. Li, X.; Li, Y.; Cheng, T.; Liu, Z.; Wang, R. J Comput Chem 2010, 31, 2109.
54. Machatha, S. G.; Yalkowsky, S. H. Int J Pharm 2005, 294, 185.