J Comput Aided Mol Des (2012) 26:725–735
DOI 10.1007/s10822-012-9549-y

# Lead Finder docking and virtual screening evaluation with Astex and DUD test sets

**Fedor N. Novikov · Viktor S. Stroylov · Alexey A. Zeifman · Oleg V. Stroganov · Val Kulkov · Ghermes G. Chilov**

Received: 15 September 2011 / Accepted: 23 January 2012 / Published online: 9 May 2012
© Springer Science+Business Media B.V. 2012

**Abstract** Lead Finder is a molecular docking software. Sampling uses an original implementation of the genetic algorithm that involves a number of additional optimization procedures. Lead Finder's scoring functions employ a set of semi-empiric molecular mechanics functionals that have been parameterized independently for docking, binding energy predictions and rank-ordering for virtual screening. Sampling and scoring both utilize a staged approach, moving from fast but less accurate algorithm versions to computationally more intensive but more accurate versions. Lead Finder includes tools for the preparation of full atom protein and ligand models. In this exercise, Lead Finder achieved 72.9% docking success rate on the Astex test set when the original author-prepared full atom models were used, and 74.1% success rate when the structures were prepared by Lead Finder. The major cause of docking failures were scoring errors resulting from the use of imperfect solvation models. In many cases, docking errors could be corrected by the proper protonation and the use of correct cyclic conformations of ligands. In virtual screening experiments on the DUD test set the early enrichment factor of several tens was achieved on average. However, the area under the ROC curve ("AUC ROC") ranged from 0.70 to 0.74 depending on the screening protocol used, and the separation from the null model was not perfect—0.12–0.15 units of AUC ROC. We assume that effective virtual screening in the whole range of enrichment curve and not just at the early enrichment stages requires more accurate solvation modeling and accounting for the protein backbone flexibility.

**Keywords** Lead Finder · Docking · Virtual screening · Benchmarks · Astex · DUD

**Electronic supplementary material** The online version of this article (doi:10.1007/s10822-012-9549-y) contains supplementary material, which is available to authorized users.

F. N. Novikov · V. S. Stroylov · O. V. Stroganov · G. G. Chilov (✉)
MolTech Ltd, Leninskie gory, 1/75A,
119992 Moscow, Russia
e-mail: ghermes@moltech.ru

A. A. Zeifman · O. V. Stroganov · G. G. Chilov
N.D.Zelinsky Institute of Organic Chemistry,
Leninsky pr-t, 47, 119991 Moscow, Russia

V. Kulkov
BioMolTech Corp, 226 York Mills Rd, Toronto,
ON M2L 1L1, Canada

## Introduction

Although the development of molecular docking methods has been underway for a few decades, the accuracy of modern methods is still far from satisfactory. Therefore, the development of theory and practical application has not lost its significance. Several criteria are typically used to evaluate accuracy of a docking method. First, it is the accuracy of a putative ligand's pose in a protein–ligand complex [1, 2]. Second, it is the accuracy of prediction of ligand binding affinity [3–6]. Third, it is the method accuracy in virtual screening where active ligands must be separated from inactive ones [5, 7–9]. These criteria are obviously interrelated. If a method fails to dock a ligand correctly, it is unlikely that the same method will succeed in classifying a ligand as active or inactive. However, the determination of a correct ligand pose is not enough. In order to rank-order ligands by their activity, an accurate scoring function is necessary. Conversely, finding a correct docking pose is impossible without accurate scoring of a pose and extensive sampling of the conformational space.

Therefore, the development of a molecular docking method focuses on both sampling and scoring issues.

In the development of Lead Finder software as a practical application of molecular docking methods we necessarily had to concentrate on the theoretical aspects of sampling and scoring [10]. Lead Finder's approach to sampling combines the original implementation of genetic algorithm with multilevel local optimizations. Key features of our genetic algorithm are: (1) niching (clusterization of ligand conformations during each round of evolution), which keeps diversity of the population and prevents algorithm from premature convergence; and (2) removal of non-evolving individuals (even favorable ones) from calculations, which gives space for other individuals to evolve and saves computational time. Energy calculations are performed using the original semi-empiric molecular mechanics functional. Lead Finder introduces three specialized scoring functions designed to rank predicted ligand poses, estimate the binding energy of docked ligand poses and rank compounds in virtual screening experiments. A series of computational implementations of scoring functions (ranging from faster and less accurate to slower but more precise) are applied during the docking process depending on its maturity.

The main objective of the current work was to evaluate the weaknesses of Lead Finder's sampling and scoring approaches by participating in the docking and virtual screening contest, which used well-known Astex [6] and DUD [7, 8] test sets. What was important—in our computational experiment we used protein and ligand structures prepared by the organizers of the contest, which enabled transparent interpretation and comparison of the results. Also, we used a number of additional protocols for docking set up in order to assess how the quality of input data affect the results. Ideally, modern docking software must evaluate the quality of input data and automatically correct errors in the input data.

The conclusions from our study are generally quite obvious. They can be summarized as follows: accurate docking and virtual screening requires the ability to account for protein flexibility, which is a sampling task. However, sampling of the extended conformational space necessarily requires accurate scoring. The evaluation of results from the Astex test set demonstrates that scoring methods employed by Lead Finder require more accurate scoring of the solvation models and intermolecular interactions.

## Methods

### Preparation of protein and ligand models

The Astex test set for docking [6] comprises 85 protein–ligand complexes selected from the PDB by the following main criteria: (a) good quality of a crystal structure; (b) pharmaceutical relevance of proteins; (c) drug-likness of ligands; (d) diversity of a protein composition of the test set.

The DUD test set for virtual screening [7, 8] comprises 40 protein targets with 81 active ligand (on average) and 36-fold excess of decoy ligands for each protein. The principle for selecting particular decoy ligands for a particular protein was to chose ligands without structural similiraity to the active ones, but with similar distribution of physicochemical properties (Mw, LogP, numbers of H-bond donors and acceptors, number of freely rotatable bonds). The latter was performed in order to obtain virtual screening results unbiased with respect to physicochemical properties of the active ligands (coupled to particular protein's active site). In addition to the standard set of DUD ligands we also benchmarked virtual screening performance using the so-called Wombat set of ligands, which was available for 13 out of 40 DUD proteins [11]. The Wombat ligands were selected using slightly different from the DUD protocol: the emphasize was put on ligand diversity, size (smaller size was preferred), affinity (lower affinity was preferred). Also the Wombat set contained more active ligands.

The full-atom models of protein and ligand structures of Astex and DUD test sets were prepared by the organizers of the current contest with the aim of correcting various errors present in pdb files and making results of different participants comparable between each other.

The basic protocol of docking and virtual screening experiments with Lead Finder used full atom models offered by the docking competition organizers. No changes were introduced to the original full atom models.

The additional protocol assumed a fully automatic preparation of full atom protein models from the positions of heavy atoms, provided by the docking competition organizers. The automatic preparation was performed by the Build Model module that is part of the Lead Finder software package [12]. The Build Model module implements the TSAR algorithm recently developed by ourselves [13] to compute pKa of protein residues and optimize protonation states at a given pH. Only the optimization of ionization states and positions of functional protons was performed; the positions of heavy atoms were left intact.

We used the ligand models provided by the docking competition organizers. The Build Model module is not presently designed to perform accurate protonation of ligands.

### Docking and virtual screening

In order to assess docking success rate on the Astex test set, 20 separate docking experiments were performed for each ligand. In each experiment, the docking success was

**Table 1** The average docking success rate obtained on top-scoring poses and, in parentheses, on the closest-to-PDB poses among the top-20 poses

| RMSD, Å | Intact test-set | Automatically prepared test-set | Manually corrected structures |
|---|---|---|---|
| 1.5 Å | 63.5 (82.4) | 67.1 (87.1) | 81.2 (89.4) |
| 2.0 Å | 72.9 (90.6) | 74.1 (91.8) | 90.6 (94.1) |
| 2.5 Å | 81.2 (91.8) | 80.6 (91.8) | 92.9 (94.1) |
| 3.0 Å | 84.8 (91.8) | 82.5 (91.8) | 92.9 (94.1) |

measured by the fitness of the RMSD value into a certain interval, such as up to 2.0 Å. If the RMSD value fit within a given interval 10 or more times out of 20, the docking was deemed successful. Table 1 summarizes the results, and Table S1 of the Supplementary Information provides data for each structure. We used default program settings in both docking and virtual screening experiments with Lead Finder [14].

### Structural filtration of docked ligand poses

To perform a coarse evaluation of the accuracy of prediction of docked poses in docking experiments and the accuracy of rank-ordering in virtual screening experiments, we used the structural filtration protocol [15]. The structural filtration is based on the idea what active ligands form similar interaction patterns with the active site of a protein. As such, we identified interaction patterns formed by active ligands for each protein in the DUD test set. The selection of such interactions was performed manually by the visual analysis of PDB structures containing active ligands. The structural filters contained both the required interactions (logical operator: "AND") and the optional interactions (logical operators: "OR"). The full listing of structural filters we developed and used with the DUD test set is provided in Table S2 of the Supplementary Materials. The evaluation of structural filtration criteria was performed by the Structure Filter module that is part of the Lead Finder software package [16].

### Rank-ordering of ligands in virtual screening experiments

Rank-ordering of ligands in virtual screening experiments was performed by two methods. In the first method, the VS score generated by Lead Finder was used to rank-order docked ligands. In the second method, the rank-ordered ligands were subjected to additional rank-ordering by the structural filtration: the ligands that did not satisfy the structural filtration criteria were moved to the bottom of the list [15, 17].

### Calculation of enrichment values in virtual screening

We adhered to the method provided by the docking competition organizers in the calculation of enrichment values in virtual screening experiments. For a given fraction of screened set we calculated the area under the ROC curve, thus obtaining ESX values where X is the fraction of the screened library. To facilitate the interpretation of results, we additionally calculated the relative ESX% values that represented the percentage of area under the ROC curve for a screened library relative to the area under the ideal ROC curve for a given fraction of the screened library. Such relative ESX% values conveniently illustrate the difference between the theoretical maximum enrichment and the obtained results.

In addition to the described enrichment characteristics, a null hypothesis test was performed for each protein. For this purpose the active and decoy ligands of one protein were docked to the other (predefined) protein. Enrichment parameters obtained in this experiment were regarded as a null model enrichment. The difference between the normal and null hypothesis enrichment values, as well as closeness of the null hypothesis enrichment value to 0,5 were the indicators of virual screening accuracy. The correspondence between "true" targets and their null hypothesis comparators was provided by the organizers of the current contest.

### Evaluation of docking success in virtual screening

The structural filtration method described above was used not only for ligand rank-ordering in virtual screening experiments, but also for a coarse evaluation of docking success of active ligands for each protein target in the DUD test set. If a docked ligand pose satisfied the structural filtration criteria, then docking was deemed successful. We performed no additional visual analysis of the correctness of a docked ligand pose. It is important to note that while this evaluation method should not be used in the accurate evaluation of the docking success rate, it is useful in the qualitative assessment of how closely the docked ligand pose resembles the correct one. We used such assessment in the interpretation of enrichment values in virtual screening experiments.

## Results and discussion

### Docking success rate

The results of docking experiments on the Astex test set are summarized in Table 1. The accuracy of docking was evaluated by the percentage of correctly docked poses

falling within a certain range of RMSD in Å. Since the deviation of individual RMSD values from the average value for 20 separate docking experiments was significant, the average RMSD value was considered to be not a useful criterion of docking accuracy. However, we did provide the average docking success rate for each protein structure of the test-set in Table S1 Supporting Information for a reference.

As seen from Table 1, the docking success rate was essentially the same for the original test set and the test set we prepared by utilizing our automatic structure preparation tools. The success rate was 72.9 and 74.1% respectively with the cut-off point of 2.0 Å. Such result was rather surprising to us since the original test set had been carefully examined and manually proofed by the docking competition organizers while our full atom protein models had been automatically prepared by software with no human intervention. One possible explanation is that the automatic preparation of full atom structure, including the calculation of ionization states by TSAR, our new graph-theoretical algorithm recently published [13], worked quite well. The other possible explanation is that the existing docking software, including Lead Finder, are not very sensitive to alterations of a full atom protein model. In this respect it would be appropriate to further investigate the impact of the protonation protocols on the docking success rate and when it becomes a significant factor, as was done in a recent work [18]. In this exercise we studied the factor of protein and ligand structure set up by way of increasing of the quality input models through manual examination and correction of subtle details in protein and ligand structure. It turned out that manual correction of structure models increased the docking success rate to 90.6%, as evidenced in Table 2. Assuming that such manually curated protein and ligand models were essentially free from errors, this docking success rate represented Lead Finder's intrinsic docking success rate. Below we attempt to analyze both the errors that can be fixed by a careful preparation of protein and ligand models and the intrinsic docking failures that are caused by scoring and sampling errors.

## Analysis of docking failures

We found that the percentage of scoring errors exceeded the percentage of sampling errors by the factor of 6. Scoring errors were found in 24% of the ligand structures where the top-scoring pose did not match the reference ligand pose but there were some not top-scoring poses that were close to the reference ligand pose. Sampling errors were found in 4% of the ligand structures where none of the top scoring poses matched the reference ligand pose. As such, it appears that scoring errors were the major source of docking failures (at least for the self-docking task).

Let us now examine the docking failures in detail. The docking failures can be classified as "preparatory" failures that can be fixed by a careful preparation of protein and ligand models, and "fundamental" failures that stem from the deficiencies of scoring or sampling. The results of our analysis are summarized in Table 2. Out of 14 preparatory failures on the Astex test set, 6 were caused by the incorrect preparation of ligand models and the remaining 8 were caused by the incorrect preparation of protein models. Further analysis of the incorrectly prepared ligand models revealed that the errors were caused by the suboptimal protonation and the suboptimal choice of ring conformations. Clearly, the protonation errors can be fixed by careful sampling of possible protonation states and accurate scoring of the free energy of such protonation states. As for the ring conformations, choosing the correct conformation appeared to be a significant issue despite the fact that Lead Finder could perform sampling of the 5- and 6-membered ring conformations. Lead Finder could not perform sampling of the 7-membered ring conformation that were present in the Astex test set.

Of the other preparatory failures, 5 could be fixed by the correction of the orientation of polar hydrogen atoms in protein model. In those cases (Table 2) the suboptimal orientation of hydrogens led to the formation of wrong H-bonds between those hydrogens and the ligand structure. In each of those cases the formation of wrong H-bonds could be prevented by the correction of the orientation of polar hydrogens. One may conclude that explicit accounting for the mobility of functional hydrogen atoms in protein models is necessary to avoid problems of this nature. However, we found that in all of those cases scoring, and not sampling, was the major source of errors since the correct ligand poses were found but they did not receive the top score. Paradoxically, we cannot conclude that explicit accounting for the mobility of functional hydrogens in protein models is necessary to avoid preparatory failures. It appears that we need a more sophisticated H-bond energy scoring method.

Another 3 preparatory failures could be fixed by not removing the water molecule from protein model where that water molecule was instrumental for the correct positioning of a ligand. For instance, in accordance with the model preparation protocol all water molecules were removed from 1mmv, including the water molecule bound to the heme group. Consequently, a ligand structure attempted to take the vacant place in order to coordinate with the iron ion. Had the water molecule remained in the protein structure, the ligand would have been docked correctly.

Thus, out of the total 22 docking failures, the 14 preparatory failures could have been prevented by a more accurate preparation of the protein and ligand models.

**Table 2** Docking failures, classified by their sources of errors

| Cause of docking failure | PDB id | Docking or scoring error | Comment |
|---|---|---|---|
| *Preparatory errors* | | | |
| Incorrect ligand protonation | 1jd0 | Scoring | Ligand is more likely to bind in its de-protonated form. The test set contained the protonated form |
| | 1l2 s | Scoring | Ligand is more likely to bind in its protonated form. The test set contained the de-protonated form |
| Incorrect ligand conformation | 1xoz | Scoring | Non-pdb conformation of unsaturated ring of a ligand |
| | 1sj0 | Scoring | Non-pdb conformation of unsaturated ring of a ligand |
| | 1s3v | Sampling | Non-pdb conformation of unsaturated ring of a ligand |
| | 1mzc | Scoring | Non-pdb conformation of unsaturated ring of a ligand |
| Optimization of H-bond network in the binding site | 1yvf | Sampling | The orientation of S556 hydroxyl had to be altered to prevent the formation of incorrect H-bond with ligand |
| | 1ygc | Scoring | H41 residue had to be protonated; the orientation of S195 hydroxyl had to be altered |
| | 1p2y | Scoring | The orientation of T556 hydroxyl had to be altered to prevent the formation of incorrect H-bond with ligand |
| | 1oq5 | Scoring | The orientation of Y7 hydroxyl had to be altered to prevent the formation of incorrect H-bond with ligand |
| | 1ia1 | Scoring | The orientation of S61 hydroxyl had to be altered to prevent the formation of incorrect H-bond with ligand |
| Accounting for explicit crystallographic water molecules | 1n2v | Scoring | Water molecule 628 was left intact |
| | 1mmv | Scoring | Water molecule 1002 was left intact |
| | 1xm6 | Scoring | Water molecule 1007 was left intact |
| *Fundamental errors* | | | |
| Model of solvation | 1meh | Scoring | Incorrect position of the charged ligand group, partially exposed into the solvent |
| | 1hvy | Scoring | Incorrect position of the charged ligand group, partially exposed into the solvent |
| | 1gm8 | Scoring | Incorrect position of the charged ligand group, partially exposed into the solvent |
| | 1g9v | Scoring | Incorrect position of the charged ligand group, partially exposed into the solvent |
| Pair-wise interactions | 1sq5 | Scoring | Insufficient accuracy in accounting for the H-bond and electrostatic energy |
| | 1n2j | Scoring | Insufficient accuracy in accounting for the H-bond and electrostatic energy |
| Ligand force field | 2br1 | Scoring | Insufficient accuracy in the ligand force field |
| | 1y6b | Sampling | Insufficient accuracy in the ligand force field |

Obviously, docking software must be able to automatically perform such model preparation. The expert's time is better spent elsewhere.

The other 8 errors on the Astex test set were the fundamental errors, arising from the deficiencies in scoring and sampling methods currently employed by Lead Finder. Those errors could not be fixed by the manipulation and optimization of the protein and ligand models. There are various possible sources of deficiencies, such as the scoring function terms and the particulars of the sampling algorithm. A detailed analysis of such particulars requires a separate study. In this study, we attempt to qualitatively evaluate and outline the possible sources of deficiencies.

First, there is an inaccuracy in the estimation of the energy of a surface-exposed polar or charged ligand group with both protein and solvent. Lead Finder often overscores the conformations where the surface-exposed ligand group is attached to the protein while it is in fact exposed into the solvent. Such errors were found in 4 cases (Table 2) where incorrect positioning of the ligand's terminal group resulted in the increase of the RMSD value above the threshold of 2.0 Å. The elimination of problems of this nature requires a revision of the solvation model that is presently used by Lead Finder. It is unclear though that the problem can be solved by the incremental revisions to the current implicit solvation model since the solvation terms in our

current scoring function have already been extensively fine-tuned. It appears that some fundamental improvement of the solvation models is required. Likely, the use of explicit solvation models will be necessary.

In 2 other cases (structures 1sq5 and 1n2j, Table 2) Lead Finder could not correctly choose between the ligand poses possessing more or less equivalent sets of H-bonds and electrostatic interactions. We tend to think that an enhanced accounting of the H-bond energy parameters may solve this problem as the present model used by Lead Finder is not sophisticated. It is possible that the use of explicit force field (OPLSAA, MMFF94 and so on) may help to resolve the issue. Finally, we observed similar errors in structures 2br1 and 1y6b that were caused by inaccuracies in the ligand force fields. In those cases the incorrect ligand poses did not receive sufficient energy penalties.

Since the organizers of the exercise indicated that 26 structures had problems in their crystallographic models, such as: alternate side chain conformations in the active center, ligand crystal packing interactions, poor density of the active center residues, and alternate ligand conformations, the docking success rate on such complexes deserves a special comment. One might expect that errors in the crystallographic structure would affect the docking success rate. However, we found that in the overwhelming majority of cases docking was successful (Table 3). The docking failures were in fact caused by the reasons described earlier (Table 2) and they were not related to the errors in the crystallographic structure. It is possible that the Astex test set is not representative enough for a conclusion to be drawn in regard to the impact of the errors in a crystallographic structure on the docking success rate. Another more representative test set may have to be developed in order to investigate such impact.

Before we conclude the discussion of docking failures, we would like to comment on the results we obtained with structures 1gpk, 1hvy and 1s3v. Those structures of the Astex test set had errors in the ligands structures—namely, the ligands in those structures had incorrect stereochemistry. In case of 1gpk, docking was successful despite the error in its ligand structure. In case of 1s3v, docking was not successful but it became successful once we corrected the ligand's stereochemistry. In case of 1hvy, docking was unsuccessful and the reason of the failure was not the ligand stereochemistry but an error in scoring the solvation of the polar/charged surface-exposed ligand group, as discussed above.

**Table 3** The docking success rate for structures that had errors in their crystallographic models

| | Successfully docked 'as is' | Docked after manual tweaks | Not docked successfully |
|---|---|---|---|
| Alternate side chain conformations | 1ia1 1kzk 1n46 1opk 1q4 g 1r1 h 1t46 1vcj 1z95 | 1s3v | 1y6b |
| Crystal packing interactions | 1gkc 1jje 1r55 1t40 1v48 2bsm | – | 1hvy |
| Poor density | 1j3j 1u1c 1v4 s | – | 1meh |
| Alternate ligand conformations | 1tz8 1ig3 1sg0 | – | – |

In essence, our participation in the docking exercise allowed us to arrive to the following conclusions. First, it is clear that more sophisticated approaches are necessary for handling protein and ligand structures on input. In particular, sampling of the protonated states of a ligand is necessary when its input structure does not or may not contain correct protonation states. A more sophisticated way of sampling of macro-rings and unsaturated rings in a ligand structure is also necessary. We should be able to account for the conserved water molecules, especially when they are clearly unlikely to be displaced by a ligand upon its binding to protein.

Second, scoring presents a greater challenge in docking simulations than sampling, at least in the self-docking set up. In Astex test set docking task, scoring errors significantly outnumbered sampling errors. We used Lead Finder with its default configuration settings. Had we provided more computing resources to sampling by tweaking configuration parameters, this difference would have been even more pronounced. The majority of scoring errors stemmed from the deficiencies of the solvation model. The analysis of those errors suggests that the incremental development of implicit solvation models is unlikely to result in a breakthrough in docking accuracy. New models are necessary, likely the explicit solvation models that better reflect the physics of the solvation process.

In contrast to the solvation models that demand new scientific methods, the development of better force field models to describe the H-bond energy and the ligand internal energy appears to require mere meticulous technical work.

Lastly, the quality of crystallographic model of a protein did not significantly impact the docking success rate (Table 3). It is possible that the Astex test set was not diverse enough for such impact to be observed. It seems that the study of the impact of crystallographic model quality requires the development of a specialized test set.

## Virtual screening with DUD test-set

The results of virtual screening on the DUD test set are presented in Table 4. A quick look at the table reveals that the early enrichment by Lead Finder is in fact quite good.

**Table 4** The accuracy of virtual screening on the DUD test set

| | Intact DUD set | | | | | de novo prepared DUD set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ES0.1%[a] | ES1% | ES2% | ROC AUC | DROC AUC | ES0.1% | ES1% | ES2% | ROC AUC | DROC AUC |
| Average | 3.4 (1.6) | 10.6 (5.3) | 15.2 (9.1) | 0.735 (0.701) | 0.148 (0.115) | 4.3 (1.6) | 12.2 (6.4) | 16.9 (10.4) | 0.728 (0.683) | 0.133 (0.091) |
| SD | 4.2 | 10.6 | 13.4 | 0.147 | 0.170 | 6.7 | 12.7 | 15.6 | 0.152 | 0.192 |
| Median | 1.7 | 8.2 | 11.2 | 0.760 | 0.106 | 2.1 | 8.8 | 13.3 | 0.717 | 0.123 |
| Min | 0.0 | 0.4 | 0.7 | 0.393 | −0.211 | 0.1 | 0.5 | 1.1 | 0.420 | −0.245 |
| Max | 16.0 | 46.2 | 56.5 | 0.961 | 0.467 | 34.0 | 56.3 | 64.9 | 0.954 | 0.525 |

The data in parentheses are the screening results without the post-screening structural filtration

[a] The values of shown are the enrichment values relative to the ideal enrichment values. They are obtained by the division of the area under the curve for a given enrichment curve by the area under the ideal enrichment curve, at a given percentage of screened library (as described in methods)

The enrichment values of ES0.1, ES1 and ES2% for ligands selected at random were 0.05, 0.5 and 1.0% respectively. Those enrichment values were 10–60 times lower than the ones obtained with Lead Finder. Notably, the results were essentially the same when the intact test set models provided by the organizers were used and when the test set was prepared de novo using the protocol described in Methods above. This suggests the suitability of Lead Finder for an automatic preparation of protein models. It is also important to note that the enrichment values in the protocol where the ligands were rank-ordered only by their VS-score were notably lower compared to the protocol in which structural filtration was applied to move the ligands not satisfying the filtration criteria to the bottom of the rank-ordered list. In the experiments discussed below we use the rank-ordered list with the structural filtration subsequently applied as we believe it more closely follows our virtual screening methodology. The detailed list of structural filters applied to each structure of the DUD test set is presented in Table S2 of the Supplementary Materials.

While the early enrichment results looked good, ROC AUC—the integral parameter—did not impress us (Table 4). The significant variation of the ROC AUC value tells that some proteins follow an almost ideal enrichment curve while others do not, in some cases falling even below the random enrichment curve. We should also note that ROC AUC was not significantly different from the same parameter of a random model that was evaluated in the screening of active ligands on a non-cognate protein model and was shown to have been 0.5.

Before discussing the analysis of virtual screening results in details, let us introduce an additional parameter we used to evaluate screening effectiveness alongside the enrichment parameters. We used the structural filtration as described above to assess the success of docking of active ligands for each of the protein targets. Obviously, such assessment was qualitative in nature since the crystallographic structures were not available in the great majority of cases. Even so, such information was useful because it provided clues as to what caused virtual screening to be effective or ineffective. We assumed that an active ligand was docking correctly in virtual screening experiments if it satisfied the structural filtration criteria, and not correctly if it did not. Therefore, we could get a rough idea of how many active ligands were docked correctly, since there was no reason to believe that incorrectly docked active ligands would receive better scores than decoy ligands—especially when the decoys were pre-selected to be similar to active ligands in their physicochemical characteristics. The full set of structural filters for each protein target from the DUD test set is provided in Table S2 of Supplementary Materials.

The results of virtual screening on the protein classes are summarized in Table 5. The information on individual proteins is provided in Table S3 of the Supplementary Materials. In summary, applying the structural filtration criteria as described above, about half of the active ligands were docked correctly on the nuclear receptors on average. Visual inspection of the randomly selected ligands confirmed that the docked ligand poses were reasonable. The data in Table S3 indicates that there was no correlation between the ROC AUC and the early enrichment results, or the percentage of the correctly docked ligands. It appears that for each of the protein structures there was a certain number of successfully docked and rank-ordered ligands that secured good results at early enrichment stages, and that those early results did not correlate with the later results. For instance, in case of RXR almost all active ligands were similar in their nature and as a result, almost all of them were docked and rank-ordered correctly.

**Table 5** The accuracy of virtual screening on the DUD test set, by protein classes and individual proteins

| | Intact DUD set | | | | | | de novo prepared DUD set | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ligands docked, % | ES0.1% | ES1% | ES2% | ROC AUC | ΔROC AUC | Ligands docked, % | ES0.1% | ES1% | ES2% | ROC AUC | ΔROC AUC |
| *Folate enzymes* | | | | | | | | | | | | |
| Average | 54.4 | 10.3 | 31.9 | 41.1 | 0.878 | 0.151 | 59.9 | 5.1 | 25.1 | 31.8 | 0.863 | 0.131 |
| SD | 1.9 | 4.0 | 17.0 | 15.4 | 0.013 | 0.207 | 20.3 | 6.0 | 23.2 | 20.4 | 0.079 | 0.095 |
| Median | 54.4 | 10.3 | 31.9 | 41.1 | 0.868 | 0.151 | 59.9 | 5.1 | 25.1 | 31.8 | 0.863 | 0.131 |
| Min | 53.0 | 7.4 | 19.9 | 30.2 | 0.859 | 0.005 | 45.5 | 0.8 | 8.7 | 17.3 | 0.807 | 0.064 |
| Max | 55.8 | 13.1 | 43.9 | 52.0 | 0.877 | 0.298 | 74.2 | 9.3 | 41.5 | 46.2 | 0.920 | 0.199 |
| *Kinases* | | | | | | | | | | | | |
| Average | 33.5 | 2.4 | 6.7 | 9.7 | 0.660 | 0.124 | 34.2 | 2.1 | 5.5 | 8.1 | 0.623 | 0.086 |
| SD | 22.3 | 2.7 | 5.5 | 8.0 | 0.133 | 0.180 | 18.4 | 2.2 | 4.8 | 6.4 | 0.123 | 0.196 |
| Median | 29.0 | 0.9 | 5.3 | 6.8 | 0.669 | 0.191 | 34.8 | 1.4 | 2.8 | 4.2 | 0.630 | 0.127 |
| Min | 1.7 | 0.0 | 0.4 | 0.8 | 0.460 | −0.211 | 2.0 | 0.1 | 0.8 | 1.6 | 0.437 | −0.245 |
| Max | 73.1 | 7.0 | 17.1 | 22.4 | 0.837 | 0.313 | 55.8 | 6.3 | 13.4 | 17.3 | 0.778 | 0.313 |
| *Metalloenzymes* | | | | | | | | | | | | |
| Average | 36.5 | 1.8 | 7.8 | 10.3 | 0.670 | 0.177 | 43.2 | 8.6 | 11.8 | 13.4 | 0.677 | 0.145 |
| SD | 16.7 | 1.9 | 7.8 | 9.3 | 0.145 | 0.168 | 25.1 | 11.9 | 14.1 | 14.7 | 0.174 | 0.202 |
| Median | 34.3 | 1.4 | 6.0 | 8.2 | 0.637 | 0.145 | 45.4 | 4.4 | 7.7 | 9.4 | 0.639 | 0.090 |
| Min | 19.2 | 0.1 | 1.1 | 2.2 | 0.533 | 0.009 | 15.3 | 0.1 | 0.5 | 1.1 | 0.524 | −0.034 |
| Max | 58.3 | 4.3 | 18.0 | 22.7 | 0.873 | 0.409 | 66.7 | 25.6 | 31.3 | 33.6 | 0.906 | 0.435 |
| *NHR* | | | | | | | | | | | | |
| Average | 43.7 | 4.5 | 12.6 | 17.9 | 0.836 | 0.173 | 45.1 | 4.0 | 12.7 | 18.2 | 0.787 | 0.084 |
| SD | 26.5 | 5.3 | 10.2 | 12.2 | 0.085 | 0.193 | 25.5 | 3.6 | 10.8 | 14.7 | 0.119 | 0.214 |
| Median | 46.1 | 1.8 | 8.0 | 11.0 | 0.835 | 0.156 | 47.9 | 3.1 | 8.7 | 12.1 | 0.765 | 0.050 |
| Min | 8.0 | 0.3 | 5.0 | 9.8 | 0.716 | −0.028 | 8.0 | 0.2 | 1.8 | 3.0 | 0.647 | −0.158 |
| Max | 93.7 | 16.0 | 36.1 | 43.9 | 0.961 | 0.467 | 90.5 | 10.3 | 34.7 | 46.1 | 0.948 | 0.525 |
| *Other* | | | | | | | | | | | | |
| Average | 31.8 | 3.0 | 9.2 | 13.3 | 0.691 | 0.149 | 43.1 | 4.9 | 13.9 | 18.6 | 0.725 | 0.185 |
| SD | 21.7 | 4.4 | 11.8 | 14.8 | 0.153 | 0.172 | 19.9 | 8.8 | 15.8 | 19.2 | 0.159 | 0.196 |
| Median | 28.0 | 1.5 | 6.9 | 9.1 | 0.725 | 0.052 | 42.4 | 1.6 | 9.8 | 12.6 | 0.721 | 0.171 |
| Min | 0.0 | 0.1 | 0.5 | 0.7 | 0.393 | −0.055 | 17.4 | 0.1 | 0.5 | 1.1 | 0.420 | −0.096 |
| Max | 69.7 | 14.9 | 46.2 | 56.5 | 0.909 | 0.451 | 69.3 | 34.0 | 56.3 | 64.9 | 0.948 | 0.522 |
| *Serine proteases* | | | | | | | | | | | | |
| Average | 62.7 | 3.8 | 12.8 | 23.4 | 0.887 | 0.109 | 54.7 | 2.7 | 14.8 | 26.4 | 0.879 | 0.141 |
| SD | 6.2 | 4.3 | 2.9 | 6.0 | 0.048 | 0.202 | 15.6 | 2.2 | 8.6 | 11.2 | 0.072 | 0.224 |
| Median | 65.7 | 1.4 | 11.5 | 23.3 | 0.875 | 0.042 | 51.0 | 3.4 | 16.3 | 26.1 | 0.873 | 0.145 |
| Min | 55.6 | 1.2 | 10.8 | 17.4 | 0.846 | −0.051 | 41.3 | 0.2 | 5.6 | 15.3 | 0.811 | −0.085 |
| Max | 66.7 | 8.7 | 16.2 | 29.4 | 0.940 | 0.336 | 71.9 | 4.4 | 22.5 | 37.8 | 0.954 | 0.363 |

However, the situation with ER was different since not all ligand scaffolds could be correctly positioned in the same protein model, and the explicit accounting for the protein flexibility was necessary. We can also refer to the suboptimal protonation of ER ligands containing pyridine rings, however we chose not to alter these ligands to ensure no confounding factors are introduced. Docking of AR and GR ligands was substantially less successful, which was also confirmed by a visual analysis. It seems likely that the explicit accounting for the protein flexibility was also necessary in those cases. Lastly, we should note that the difference between the ROC AUC of a null model and those of some nuclear receptors was insignificant. This was the case for proteins resembling the decoy targets from the null model.

As for the kinases, docking of their active ligands was in general significantly less successful than for other classes of proteins. The same held true for the early enrichment factors and the integral parameter of ROC AUC. In our opinion, the main reason for that was the treatment of a protein as a rigid structure. The worst results were obtained on FGFR1 and PDGFRb where the protein structure models were of the lowest quality. Therefore it was not surprising to obtain better results on the null model simply because the null model structures were of a superior quality. Nevertheless, scoring errors were also present as evidenced in cases of SRC and EGFR where the active ligands were docked reasonably correctly, however the integral parameter was still far from ideal. The serine proteases were among the most successful examples. In that case, the active ligands had a positively charged group that was binding at a particular place within the active center. Therefore, a correct placement of that group was practically enough for a correct docking of the entire ligand. As a result, that particular protein feature ensured the good distinction from the null model.

Among the metalloproteases, good docking results were achieved on the COMT active ligands, as evidenced by the structural filtration and visual analysis criteria. In case of PDE5, the ligands attempted to form incorrect contacts with the metal ion in the active center and that led to incorrect docking results. An additional investigation revealed that an addition of a water molecule into the ligand's coordination sphere results into a striking improvement of docking, however we decided against providing such improved data since that would have introduced confounding factors. In case of ADA it seems that accounting for the protein flexibility was crucial for the correct docking since the docked poses in the rigid structure did not look reasonable.

In case of COX-1 and COX-2, the good separation of active ligands from the inactive ones can be explained by the fact that the active center of COX-1 is exposed into the solvent while the active center of COX-2 is relatively closed. In case of AcChE, the removal of all water molecules from the active center resulted in the incorrect docking of active ligands since they form H-bonds with the protein structure only via the water molecules. The docking

**Table 6** The accuracy of virtual screening on the wombat ligands of the DUD test set

| Target | Ligands docked, % | ES0.1% | ES1% | ES2% | ROC AUC | ΔROC AUC |
|---|---|---|---|---|---|---|
| ALR2 | 11.6 | 0.6 (0.1) | 3.5 (0.7) | 4.8 (1.8) | 0.618 (0.574) | 0.125 (0.059) |
| AR | 2.9 | 0.6 (0.3) | 3.7 (1.6) | 4.7 (3.4) | 0.542 (0.542) | −0.023 (0.047) |
| CDK2 | 27.6 | 5.3 (0.2) | 8.3 (6.3) | 10.6 (8.2) | 0.690 (0.661) | 0.085 (0.051) |
| COX2 | 29.6 | 1.6 (0.2) | 5.9 (2.5) | 9.5 (4.2) | 0.620 (0.589) | 0.188 (0.157) |
| EGFR | 24.8 | 0.5 (0.3) | 5.8 (1.9) | 9.9 (5.6) | 0.724 (0.664) | 0.109 (0.046) |
| ER agonist | 17.5 | 7.5 (0.3) | 13.0 (7.1) | 18.5 (12.8) | 0.846 (0.841) | 0.190 (0.190) |
| ER antagonist | 52.8 | 0.1 (0.0) | 0.6 (0.5) | 5.2 (1.3) | 0.703 (0.427) | 0.383 (0.251) |
| FXa | 60.8 | 0.2 (0.1) | 3.5 (1.1) | 7.8 (2.9) | 0.741 (0.653) | 0.072 (0.027) |
| HIVRT | 14.0 | 0.5 (0.3) | 2.0 (2.1) | 4.0 (5.6) | 0.621 (0.606) | 0.169 (0.155) |
| p38 | 22.4 | 1.0 (0.6) | 3.5 (4.8) | 4.3 (7.1) | 0.386 (0.387) | −0.005 (0.013) |
| PDE5 | 19.9 | 3.3 (0.9) | 7.6 (3.8) | 8.7 (5.0) | 0.430 (0.346) | 0.021 (−0.063) |
| PPAR | 51.5 | 0.0 (0.0) | 0.1 (0.0) | 0.2 (0.0) | 0.474 (0.106) | 0.322 (0.023) |
| Average | 28.0 | 1.8 (0.3) | 4.8 (2.7) | 7.4 (4.8) | 0.616 (0.533) | 0.136 (0.080) |
| SD | 18.0 | 2.4 (0.3) | 3.6 (2.3) | 4.7 (3.5) | 0.137 (0.191) | 0.124 (0.089) |
| Median | 23.6 | 0.6 (0.2) | 3.6 (2.0) | 6.5 (4.6) | 0.620 (0.581) | 0.117 (0.049) |
| Min | 2.9 | 0.0 (0.0) | 0.1 (0.0) | 0.2 (0.0) | 0.386 (0.106) | −0.023 (−0.063) |
| Max | 60.8 | 7.5 (0.9) | 13.0 (7.1) | 18.5 (12.8) | 0.846 (0.841) | 0.383 (0.251) |

The primary data are the screening results with post-screening structural filtration applied. The data in parentheses are the screening results without the post-screening structural filtration

results improved significantly when at least one water molecule was left within the active center. In our opinion, the explicit accounting for protein flexibility is essential to obtain good docking of active ligands in case of InhA. The same was true for other proteins, especially for HivRT and GPB. A simple change of the symmetric protein subunit from the pdb file improved docking reults for AmpC.

Relative to the virtual screening results on the *wombat* ligand set provided by the organizers of the exercise specifically for this publication, the main tendencies in enrichment were similar to those of the native DUD test set as shown in Table 6, although the screening effectiveness was notably lower. We may conclude that the need for explicit accounting for protein flexibility and better scoring is even more pronounced on this test set. Notably, the null model approached the randomly selected model on this test set. The likely reason for that, however, may be that not all targets from the wombat set had proper counterparts as a null model.

## Conclusion

This docking exercise provided us with a chance to formulate a few important conclusions. It is obvious that the progress in the development of docking methods, as seen from the currently observed docking success rate, the accuracy of binding energy calculation and the enrichment values in virtual screening, is not yet satisfactory despite many years of labor of many research groups. There is still some room for improvement available to those who choose to focus on the development of smarter approaches to handling protein and ligand structures in docking. Lead Finder software needs a more comprehensive method of sampling ligand ring conformations, a better accounting for the ionization states of both ligand and protein structures, accounting for the conserved water molecules and automatic identification and correction of inconsistencies and errors in the crystallographic structures. It is not all that is needed, however. The improvement in virtual screening necessitates more accurate docking of active ligands that in turn requires the explicit accounting for protein flexibility, including the backbone flexibility. Our coarse analysis of docking accuracy by measuring the number of ligands satisfying the structural filtration criteria was more or less in correlation with the enrichment results. Furthermore, once the correct ligand pose is found it has to receive the correct score against other poses of the same ligand, and against other ligands in virtual screening. The docking exercise on the Astex test set revealed that the main issue with scoring was the balance of protein and ligand solvation interactions. The intuition suggests that the use of explicit solvation models describing the interactions with the solvent from the standpoint of statistical mechanics may be necessary to solve this problem.

**Acknowledgments** The work was supported by the Foundation for assistance to small enterprises in the scientific area (Contract 8175p/7168).

## References

1. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50(4):726–741. doi:10.1021/jm061277y
2. Nissink JW, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R (2002) A new test set for validating predictions of protein-ligand interaction. Proteins 49(4):457–471. doi:10.1002/prot.10232
3. Smith RD, Dunbar JB, Ung PM, Esposito EX, Yang CY, Wang S, Carlson HA (2011) CSAR benchmark exercise of 2010: combined evaluation across all submitted scoring functions. J Chem Inf Model. doi:10.1021/ci200269q
4. Novikov FN, Zeifman AA, Stroganov OV, Stroylov VS, Kulkov V, Chilov GG (2011) CSAR scoring challenge reveals the need for new concepts in estimating protein-ligand binding affinity. J Chem Inf Model. doi:10.1021/ci200034y
5. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49(20):5912–5931. doi:10.1021/jm050362n
6. The CCDC/Astex Test Set. http://www.ccdc.cam.ac.uk/products/life_sciences/gold/validation/astex. Accessed 8 September, 2011
7. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801. doi:10.1021/jm0608356
8. DUD—A directory of useful decoys. http://dud.docking.org/ Accessed 14 September, 2011
9. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C (2009) Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. J Chem Inf Model 49(6):1455–1474. doi:10.1021/ci900056c
10. Stroganov OV, Novikov FN, Stroylov VS, Kulkov V, Chilov GG (2008) Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening. J Chem Inf Model 48(12):2371–2385. doi:10.1021/ci800166p
11. Good AC, Oprea TI (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? J Comput Aided Mol Des 22(3–4):169–178. doi:10.1007/s10822-007-9167-2
12. BuildModel, utility to prepare protein model for docking. http://moltech.ru/_downloads/download.php?dl=2 Accessed 8 September, 2011
13. Stroganov OV, Novikov FN, Zeifman AA, Stroylov VS, Chilov GG (2011) TSAR, a new graph-theoretical approach to computational modeling of protein side-chain flexibility: modeling of ionization properties of proteins. Proteins 79(9):2693–2710. doi:10.1002/prot.23099
14. Lead Finder user manual. http://www.moltech.ru/_downloads/leadfinder_usermanual.pdf. Accessed 8 September, 2011
15. Novikov FN, Stroylov VS, Stroganov OV, Chilov GG (2010) Improving performance of docking-based virtual screening by structural filtration. J Mol Model 16(7):1223–1230. doi:10.1007/s00894-009-0633-8

16. Lead Finder distributive. http://www.moltech.ru/_downloads/download.php?dl=1. Accessed 8 September, 2011
17. Novikov FN, Stroylov VS, Stroganov OV, Kulkov V, Chilov GG (2009) Developing novel approaches to improve binding energy estimation and virtual screening: a PARP case study. J Mol Model 15(11):1337–1347. doi:10.1007/s00894-009-0497-y
18. Plewczynski D, Lazniewski M, Augustyniak R, Ginalski K (2011) Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. J Comput Chem 32(4):742–755. doi:10.1002/jcc.21643