Improving Docking Validation

Maria I. Zavodszky[1] and Leslie A. Kuhn[*,1,2]

[1]*Department of Biochemistry and Molecular Biology and* [2]*Quantitative Biology and Modeling Initiative, 502C Biochemistry Building, Michigan State University, East Lansing, Michigan, 48824-1319*

[*]To whom correspondence should be addressed: Tel. (517) 353-8745. Fax: (517) 353-9334. E-mail: KuhnL@msu.edu. URLs: http://www.bch.msu.edu/labs/kuhn and http://QBMI.msu.edu

Abstract

Recent comparative studies indicate that no one docking method currently outperforms others across diverse protein targets. However, validation studies are also subject to error, and the results can be misleading. Here, the results of a prominent docking validation study are analyzed, showing that the way the software is run and the results are measured can significantly affect the conclusions. Furthermore, testing docking methods for their ability to redock ligands, or to identify known ligands when screening against a protein target whose structure is pre-conformed to one of the ligands, is insufficient to measure their performance in the real-world setting of identifying and predicting the binding modes of substantially novel ligands. Finally, we address how comparing the sampling and scoring accuracies of a docking tool can lead to substantial improvements in docking, by guiding the choice of a scoring method that robustly detects the correct ligand binding mode for a variety of targets.

Introduction

Docking validation. Numerous validation studies have assessed docking quality, ranging from a limited number of test systems and known ligands to comprehensive studies[1-13]. A thorough overview of docking and scoring methods has also been published recently[14]. Most authors agree that no current tool has consistently superior performance across a range of systems. Some find that consensus scoring enhances screening results[1,3,12], while others believe this is an artifact[13]. Some claim that improving scoring is the solution to the problem of differentiating true ligands from false positives[4,5], while others have found that better sampling, leading to more accurate dockings for ligands, is needed to improve their scores relative to non-ligands[8,10,15]. Clearly, enhancements in both scoring and sampling, including modeling flexibility of the protein, and validation across a diversity of targets, are needed to yield consistent accuracy. One group's analysis of docking validation[16] underscores the importance of the diversity and experimental quality of the test set, analysis of the statistical significance of results, careful attention to the initial molecular geometry, and other details of running docking software for ensuring the results are correct and generalize well.

Goals of this study. Given the inconsistent performance of methods across protein families, many researchers in pharmaceutical discovery carry out comparative studies of docking and scoring on their targets, to discover which method or combination of methods works best. Here we present a case study comparing the results of the SLIDE docking and high-throughput screening software[15] run by two groups on the same systems: redocking a series of 50 ligands into their target protein structures and screening to identify 10 known ligands of thymidine kinase mixed into a database of 1,000 drug-like compounds. The following ideas are discussed in light of the results:

- *Docking quality can depend significantly on how the input is prepared and the software parameters are set.* Flexibility in how the software is run can be a strength or a weakness, depending on who is using the software.

- *Even an intuitive metric of success, such as the number of known ligands retrieved as a function of the percentage of the database screened, can have serious shortcomings.*

- *Redocking does not necessarily provide an accurate assessment of how well the software will perform when screening for new ligands.* For this real-world case, assessing the ability of the docking program to model or accommodate conformational change in the protein, at least at the side-chain level, is crucial.

- *The quality of sampling the correct binding mode for a ligand, versus identifying that binding mode by correctly scoring or ranking it, should be considered separately when evaluating a docking package.* Often the sampling performed by a docking program can be coupled with an independent scoring method to provide superior results.

Materials and Methods

As the starting point for this case study, we refer to the work of Kellenberger et al.[6] which compares the ability of seven docking tools, including SLIDE, to redock a range of ligands into their pre-conformed protein targets, and to identify 10 known thymidine kinase ligands as top-scoring ligand candidates when mixed into a set of 1,000 drug-like compounds.

Redocking of 50 ligands into their targets. Each docking tool was used with its built-in scoring function. 50 complexes (Table 1) were selected (the first 30 and last 20, to ensure no bias) from the list used by Kellenberger et al.[6].

Screening of 10 thymidine kinase ligands mixed with a database of 1,000 drug-like compounds. Ligand enrichment was assessed by identifying the top-scoring conformation and orientation of each ligand candidate. The 10 known thymidine kinase (TK) ligands were identified based on their 2D structures and references cited in Bissantz et al.[1] and extracted from the PDB structures 1e2k, 1e2m, 1e2n, 1e2p, 1ki2, 1ki3, 1ki6, 1ki7, 1kim, and 2ki5. The 1,000 drug-like molecules were downloaded from the website of Dr. Didier Rognan at: `http://bioinfo-pharma.u-strasbg.fr/download/random1000.mol2`

The protein target for screening was Protein Data Bank (PDB) structure 1kim (`http://www.rcsb.org/pdb`). This structure was selected to enable comparing our results with those of Kellenberger et al.[6]. However, this is not an optimal structure for unbiased screening, since it is the structure of TK in complex with one of the 10 known ligands. Therefore, the apo (ligand-free active-site) structure of thymidine kinase, PDB entry 1e2h, was also used as an unbiased target to test the ability to accommodate side-chain conformational change in the active site.

Target preparation for SLIDE screening. Given a protein target structure, SLIDE calculates the optimal positions of protein hydrogen atoms for hydrogen-bond calculations. For consistency, all hydrogen atoms, water molecules, and metal ions included in the redocking and TK targets were removed from the PDB files for use with SLIDE v. 2.3.

Cofactors to be included in the binding site during docking were handled as rigid parts of the protein. Cofactor nitrogen and oxygen atoms with the capacity to form hydrogen bonds were relabeled with one of the following atom names in the PDB file for the target: NDD (nitrogen acting as a hydrogen-bond donor), NAA (nitrogen acting as acceptor), NNN (nitrogen acting as a donor and/or acceptor), ODD (oxygen acting as a donor), OAA (oxygen acting as an acceptor),

or ONN (oxygen acting as a donor and/or acceptor). Complexes including binding-site cofactors for the redocking study are listed in Table 3.

Defining input conformations for ligands and ligand candidates. A critical point is the handling of 2D and 3D conformers as input to SLIDE. If only 2D structure files are available, an initial, low-energy 3D conformer is generated for each ligand candidate using a tool such as Corina (v. 3.0, J. Gasteiger, Erlangen, Germany; http://www2.chemie.uni-erlangen.de/software/corina). Given a 3D conformer, SLIDE will dock the ligand candidate and model limited ligand and protein side-chain flexibility, which suffices if the molecule is in a low-energy, near-bioactive conformation. However, SLIDE does not exhaustively search for alternative, low-energy ligand conformations. Therefore, for any SLIDE docking involving significantly flexible molecules not necessarily near the bioactive conformation, it is important to sample and provide low-energy 3D conformations as input to SLIDE. We recommend using Omega (http://www.eyesopen.com/products/applications/omega.html; Open Eye Software, Santa Fe, NM). When available, we also recommend including the Cambridge Structural Database (http://www.ccdc.cam.ac.uk/products/csd) crystal structure for the ligand, as a known low-energy structure.

Standard mode of running SLIDE. SLIDE v. 2.3 was used with the default parameter settings for both the redocking and screening studies[15]. A step-by-step guide on how to run SLIDE in an unbiased mode, including the commands, default parameters, and ligand and protein preparation steps, may be found in the Supplementary Material.

Database for the redocking study. Using 50 of the complexes evaluated by Kellenberger et al.[6], we found that the following PDB entry replacements were necessary: 1ack had been superceded in the PDB by entry 2ack. Only two ligand interaction points (one hydrophobic and

one donor/acceptor) were automatically computed for the very small ligand in 2ack, and three interaction points are required to dock a ligand. Using molecular graphics inspection of the ligand alone, three additional hydrophobic carbon atom centers were added to the ligand interaction points file. The PDB entry for 1drl, listed as a DHFR structure in the Kellenberger et al. study, does not exist. We assumed that PDB entry 1dr1 (ending in one instead of L) was used. 1ebp is not a retinoic acid binding protein, as stated in their paper, whereas 1epb is, so 1epb was used. 6abp was not included, because it has two sugar isomers with partial occupancies in the x-ray structure, and no information is available about which isomer was used by Kellenberger et al. to detect the correct docking[6]. The final list of structures for redocking is given in Table 1.

Structural superposition for analyzing screening results. The thymidine kinase crystallographic complexes containing the 10 known ligands (1e2k, 1e2m, 1e2n, 1e2p, 1ki2, 1ki3, 1ki6, 1ki7, 1kim, and 2ki5) were superimposed to enable calculating the docking RMSD for each of the ligands, using the following residues' backbone atoms in the target structure (PDB 1kim): A47 -A69, A77 -A147, A154-A263, and A280-A374.

Results

Redocking 50 ligands into their targets. To measure redocking accuracy for the 50 complexes (Table 1), two metrics were used: the RMSD of the best ligand docking relative to its crystallographic orientation (measuring how closely the docking method samples the correct ligand orientation), and the RMSD of the best-scoring docking (measuring the ability of the scoring method to detect the most correct ligand orientation).

SLIDE docking accuracy. The docking accuracy results for the 50 complexes, using the protocol given above, are shown in Fig. 1 (trace shown as stars) alongside results published by

Kellenberger et al.[6] for SLIDE on 100 complexes (including our 50) and the results from other docking methods in their analysis. SLIDE run according to the defined protocol performed significantly better than in the Kellenberger et al. study, with 75% of the 50 ligands docking to within 2.0 Å RMSD of the crystal complex positions, and 56% docking to within 1.0 Å RMSD. This places SLIDE among the top performers – Surflex[17], GOLD[18], Glide (Schrödinger, LCC), and QXP[19] - whereas SLIDE run by Kellenberger et al. performed among the worst, apparently due to differences in parameter settings and protein and ligand input preparation for the SLIDE runs.

Scoring accuracy. For the same 50 complexes, scoring accuracy, assessed by selecting the top-scoring docking for each complex, is summarized in Fig. 2 alongside the results of Kellenberger et al. SLIDE scoring performance using the above protocol is in the middle of the docking methods: somewhat better than QXP though not as good as Surflex and GOLD, but about 10% better than the results reported by Kellenberger et al., in terms of the percentage of ligands having best-scoring dockings within 2.0 Å RMSD of the crystallographic position. The decrease in performance relative to the results in Fig. 1 indicates that the quality of scoring in v. 2.3 of SLIDE was moderate whereas its sampling accuracy was high. Employing the stand-alone scoring function DrugScore[20] to select the top SLIDE docking for each ligand improves the scoring accuracy considerably (Fig 2, trace with diamonds). The combined SLIDE-DrugScore approach ties with GOLD as the top-performing method in terms of the percentage of ligands docking within 1-2 Å RMSD.

Enrichment accuracy for TK ligands. The ability to select true ligands from a large number of decoys depends on the quality of sampling and scoring in docking, as well as the ability to accommodate side-chain flexibility in the binding site. This was assessed by

identifying the top-scoring conformation and orientation of each ligand candidate in a database of 10 known thymidine kinase (TK) ligands mixed with 1,000 drug-like molecules[1]. Input conformer generation by Omega for the 1,010 molecules resulted in 80,094 conformers for screening and docking by SLIDE. The docking time per conformer was 3 seconds on average.

The perfect result would be to identify the protein-bound conformations of the 10 TK ligands as the top-scoring dockings (considering only one docking per candidate) for the 1,010 compounds. The results (Fig. 3) show modest improvement in SLIDE performance using the standard protocol (trace shown as stars) relative to the results published by Kellenberger et al.[6]. However, the apparently modest results are an artifact of the way the data was plotted. In fact, SLIDE identified 6 of the 10 TK ligands within the 20 top-scoring compounds, while docking relatively few non-ligand compounds, with only 79 compounds docked out of the 1,010 candidates in the dataset. When plotted as percent of database coverage (Fig. 3), or top-scoring percent of all docked molecules, methods that dock more (false positive) compounds artificially look more successful, because the larger denominator means a smaller (more favorable looking) percentage value for the true positives relative to the case when only a few false positive compounds were docked. This is depicted in Fig. 4, for a case with the same number and ranking of true positives found within a total of either 100 or 500 dockings, the majority of which are false positives. For drug discovery, it is important that the top-scoring compounds are likely to be true hits, and that relatively few false positives are identified. If the true positives are instead plotted as a function of scoring rank (Fig. 5), it is clear that 6 of the known TK ligands are ranked within the top 20 SLIDE hits.

The screening against TK was also performed using the ligand-free, apo structure of TK (PDB code 1e2h) as a target, to show how SLIDE performs in the more difficult case of

screening against a protein structure that is not pre-conformed for one of its ligands. The ranks of true positives are similar for screening against the ligand-bound and apo conformations of TK (Fig. 5), due to the ability of SLIDE to model side-chain flexibility during docking (Fig. 6), except that one of the ten ligands did not dock into the apo structure (Fig. 5). Figure 6 shows that for the top-scoring dockings of one of the TK ligands, SLIDE models the generally small side-chain conformational changes throughout the active site that are needed to bind the ligand.

While it is reassuring to be able to identify the known TK ligands within the top 20 scoring compounds out of 1,010 screened, it should be noted that docking validation on any one protein, especially in a case like this where the true ligands are structurally related and do not show much chemical diversity, cannot be extrapolated as a general indicator of success. We recommend that similar validation be done for any protein target of particular interest, using the apo protein structure as a target to screen known ligands mixed with drug-like compounds.

Post-morten analysis of docking failures. A post-mortem analysis of docking across a range of targets can help guide future software improvements. For the redocking case using the SLIDE scoring function to detect the best docking for each ligand, 5 of the 50 ligands failed to dock, while 8 were not docked to within 2.0 Å RMSD. For 4 of these 8 cases (1acj, 1dbb, 2ctc, 4fab), the best-docking RMSD values were between 2.1-2.3 Å, which represent close-to-correct dockings. Table 2 summarizes the apparent causes of the other mis-docked or not-docked cases, the recommended solutions, and the final docking RMSD for the best-scoring ligand orientation when the solutions were implemented. The most common problem, for five out of nine ligands, is that the default parameters for template generation resulted in templates that were not dense enough to adequately describe the binding site. This problem can be solved by generating a denser template, either by decreasing the clustering cutoff of the hydrophobic

template points to sample hydrophobic regions more finely, or replacing the “sparse” option with “dense” if hydrogen-bonding regions are not sampled finely enough (for mostly polar ligands and binding sites). The only constraint is that the template should have no more than150 hydrogen-bonding and hydrophobic template points, total, to maintain computational efficiency. In a few cases, the default hydrophobic template representation proved inadequate when aromatic interactions dominated in binding. The problem can be alleviated by increasing the hydrophobic template point density via lowering the clustering threshold. These procedures may be implemented when validating the ability of software to correctly dock known ligands for a target, before performing unbiased screening. To more thoroughly address the representation of aromatic interactions, the favored separation and angular dependence of pi-pi and pi-cation interactions will be implemented in a future version of SLIDE, similarly to the knowledge-based template point placement that encodes favored angles and lengths for hydrogen bonds[15], which significantly improved docking and scoring between versions 1 and 2 of SLIDE.

Discussion

The docking validation case study presented here provides lessons to both the users and developers of docking tools:

- *Docking quality can depend significantly on how the input is prepared and the software parameters are set.* While it remains unclear what went wrong, the atypical results for SLIDE in the Kellenberger et al. study might be explained by something as simple as selecting the lowest scoring ligand docking rather than the highest. It could also be that the input parameters were set somewhat differently. In either case, when docking results are unexpectedly poor, it is often worthwhile to contact the software developers. Their expertise can more quickly identify the

source of the problem, and the software will improve through the process of finding and filling in the pitfalls that users encounter. Some docking and scoring tools, such as the scoring functions PLP[21] and ScreenScore[12] and the docking tool FlexX[22], have been implemented in more than one software package. These implementations can give significantly different results, sometimes due to hidden input preparation by the software package, and sometimes due to changes in how the software was coded. The author can also recommend an implementation or describe the differences between them.

There are two basic schools of thought in designing molecular modeling software: "plug-and-play" and "knowledge-based optimization". The plug-and-play school believes in software that runs robustly and with fairly uniform quality without parameters that can be set (or mis-set) by the user. This uniformity has the down side of making it difficult or impossible to adapt the method for targets with unusual properties (e.g., those with an unusually hydrophobic pocket) or to take advantage of prior knowledge about the target and its ligands (e.g., the presence of a cofactor that restricts the binding site and participates in ligand recognition). On the other hand, the knowledge-based optimization school is comprised of more experienced users who tune the software and encode their knowledge about the target to optimize performance. SLIDE aims to follow a middle ground, in that a robust, uniform parameter setting for template generation and docking runs has been documented and performs well across a range of cases. Feedback on how to improve the parameter settings for particular targets, particularly for better representing the binding site, can also be gained by analyzing failures in docking known ligands (e.g., Table 2), and by taking into account the output provided by SLIDE concerning the stage at which docking failed (e.g., docking score did not meet the default threshold).

- *Even an intuitive metric of metric of success, such as the number of known ligands retrieved as a function of the percentage of the database screened, can have serious shortcomings.* It is important for docking metrics to penalize false positives as well as reward true positives with high rank. Thus, a metric such as "number of true ligands retrieved as a function of the percentage of the database screened (database coverage)" actually favors docking methods with high false-positive rates, whereas a similar metric, "number of true ligands retrieved as a function of rank in the database", clearly distinguishes docking methods with fewer false positives.

- *The quality of sampling the correct binding mode for a ligand, versus identifying that binding mode by correctly scoring or ranking it, should be considered separately when evaluating a docking package.* While it would be most convenient to have one modeling package that reliably handles docking and modeling induced conformational change, solvation, and scoring, no optimal package exists or is likely to exist, given the number of research groups focusing on different aspects of the problem. Thus, evaluating docking software with respect to its ability to handle conformational change, sample the correct binding mode, and correctly score or rank that binding mode relative to other orientations and ligands will suggest combinations of tools that can outperform any one docking program. The latter two steps were analyzed in the Kellenberger et al. study and indicated that the scoring function in SLIDE v. 2 was a limitation. This encouraged us to test SLIDE with other scoring functions, and its combination with DrugScore has proven to be highly successful (Fig. 2). This further motivated our developing a new scoring function within SLIDE, with similar accuracy to DrugScore and 35 times the speed, to enable use in high-throughput screening (available in SLIDE v. 2.4; manuscript in preparation).

- Redocking does not necessarily provide an accurate assessment of how well the software will perform when screening for new ligands. Beyond sampling and scoring accuracy and speed, the ability to perform well when docking into apo binding sites is also a major consideration for choosing a docking tool for real-world applications. Redocking presents the easiest possible case, when the ligand and protein are provided in their correct, bound conformations. Not only does this circumvent the ubiquitous need to adapt at least the side-chain conformations of the ligand and protein to form the correct complex, but also for most scoring functions it is far easier to accurately score the correct complex than to identify (as nearly correct) a complex that has one or more functional groups misplaced, as is typical in docking. Thus, validation studies need to move forward to address the quality of docking when starting with an unbiased (preferably ligand-free) protein structure and a low-energy, but not necessarily correct, conformation for each ligand and ligand candidate. Modeling protein main-chain conformational change upon complex formation and the positions of bridging water molecules present additional, very interesting challenges to surmount in order to attain accurate dockings for many systems of interest (e.g., protein kinases and protein-protein complexes).

Acknowledgments

We thank Open Eye Software for providing its tools for academic use, including Omega for conformer generation and QuACPAC for assignment of partial charges to ligands, and Professors Holger Gohlke and Gerhard Klebe for the use of DrugScore. Dr. Didier Rognan (CNRS) generously provided the set of 1,000 drug-like compounds for the screening study. This research was supported through NIH Partnerships for Novel Therapeutics grant AI53877.

Table 1. PDB codes of 50 complexes used to test docking and scoring accuracy

1aaq	1bbp	1die	2ctc	3hvt
1abe	1cbs	1dr1	2dbl	3ptb
1acj	1cbx	1dwd	2gbp	3tpi
2ack	1cil	1eap	2lgs	4cts
1acm	1com	1eed	2phh	4dfr
1aha	1coy	1epb	2plv	4fab
1apt	1cps	1etr	2r07	4phv
1atl	1dbb	1fkg	2sim	7tim
1azm	1dbj	1fki	4aah	8atc
1baf	1did	1frp	3cpa	8gch

Table 2. Sampling failures in docking, with recommended solutions

PDB code	RMSD with Default Protocol[a] (Å)	Reason for Failure	Adjustment to Protocol	RMSD with New Protocol[a] (Å)
1aha	No dockings	Wrong ligand protonation	Corrected ligand protonation	0.9
1atl	8.4	Hydrophobic template not dense enough	Calculated denser hydrophobic template using cluster threshold of 3.0 Å	1.0
1azm	No dockings	Template not dense enough; score cutoff is too high	Calculated larger template with "dense" option for H-bonding point density and 3.0 Å cluster threshold for hydrophobic template; set score cutoff to 0.	1.0
1baf	5.2	Only one protein-ligand H-bond; lack of aromatic template points	Calculated denser hydrophobic template using cluster threshold of 2.5 Å	2.3
2plv	20.2	Hydrophobic template not dense enough	Calculated denser hydrophobic template using cluster threshold of 3.0 Å	1.4
2r07	13.8	No protein-ligand H-bonds. Hydrophobic template not dense enough	Calculated denser hydrophobic template using cluster threshold of 3.0 Å	0.7
3hvt	No dockings	No protein-ligand H-bonds. Hydrophobic template not dense enough.	Calculated denser hydrophobic template using cluster threshold of 2.5 Å and minimal H-bond points	1.1
3ptb	No dockings	Score cutoff too high, for small ligand	Set score cut-off to 0	1.3
4cts	No dockings	Score cutoff too high	Set score cut-off to 0	1.2

[a] RMSD values are those of dockings closest to the X-ray orientation.

Table 3. PDB complexes with cofactors included in SLIDE dockings

PDB code	Cofactor ID	Cofactor Name
1coy	FAD	Flavin-adenine dinucleotide
1dr1	NAP	Nicotinamide-adenine-dinucleotide phosphate (NADP+)
1frp	AMP	Adenosine monophosphate
2phh	APR	Adenosine-5-diphosphoribose

Figure Legends

Figure 1. Docking accuracy presented as the percent of complexes attaining a given best RMSD value, using the docking of each ligand closest to its crystallographic binding mode. SLIDE results using the protocol presented here (curve with star symbols) are shown in comparison with the results of Kellenberger et al.[6] for SLIDE, DOCK, FlexX, Fred, GLIDE, GOLD, Surflex, and QXP.

Figure 2. Scoring accuracy shown as the percent of complexes attaining a given best RMSD value relative to the crystallographic orientation, using the top-scoring docking of each ligand. SLIDE results from the protocol presented here (curve with stars) are shown in comparison with the results of Kellenberger et al.[6] for SLIDE, DOCK, FlexX, Fred, GLIDE, GOLD, Surflex, and QXP, as well as results from combining SLIDE and DrugScore.

Figure 3. Enrichment from screening a database of 10 known thymidine kinase inhibitors mixed with 1,000 random drug-like molecules. The cumulative percentage of known inhibitors recovered is plotted as a function of percentage of the database screened, where the database is ranked from the top-scoring compound (near 0%) to the worst-scoring compound (100%). SLIDE results from the protocol presented here (curve with stars) are shown in comparison with the results of Kellenberger et al.[6] for SLIDE, DOCK, FlexX, Fred, GLIDE, GOLD, Surflex, and QXP on the same dataset.

Figure 4. Enrichment metrics can be misleading. The two curves represent the enrichment values in two hypothetical screening experiments. Docking tool #1 returned 100 dockings (solid

line), while tool #2 returned 500 dockings (dashed line) from screening the same database against the same target. Both screening tools returned 10 true positives (known ligands) with scoring ranks 1, 2, 3, 7, 9, 12, 15, 16, 18, and 20. The enrichment provided by screening tool #2 (returning 500 dockings) seems to be better, according to this normalized metric (% of database coverage), even though the only difference with tool #1 is that it retrieved many more false positives.

Figure 5. A preferred metric for enrichment. When the SLIDE data in Fig. 3 is plotted as a function of scoring rank (instead of percent database coverage), it becomes clear that a majority of the known ligands appear among the top-scoring compounds. The curve with squares corresponds to screening against the bound structure of TK (PDB code 1kim) after removing the ligand from the binding site, while the curve with triangles represents the results obtained with the same protocol against the unbiased, apo structure of TK (PDB code 1e2h).

Figure 6. Induced fit modeled in the binding site of thymidine kinase. Side-chain conformational changes performed by SLIDE (side chains in red, docked ligands colored by atom types) for the top-scoring dockings of a known ligand (from PDB structure 1ki3) into the apo structure of TK (PDB code 1e2h) are compared to differences observed between the apo TK structure (in green) and the crystal structure of the protein-ligand complex (PDB code 1ki3, in white, with ligand colored by atom type except for the carbon atoms, which appear in white).

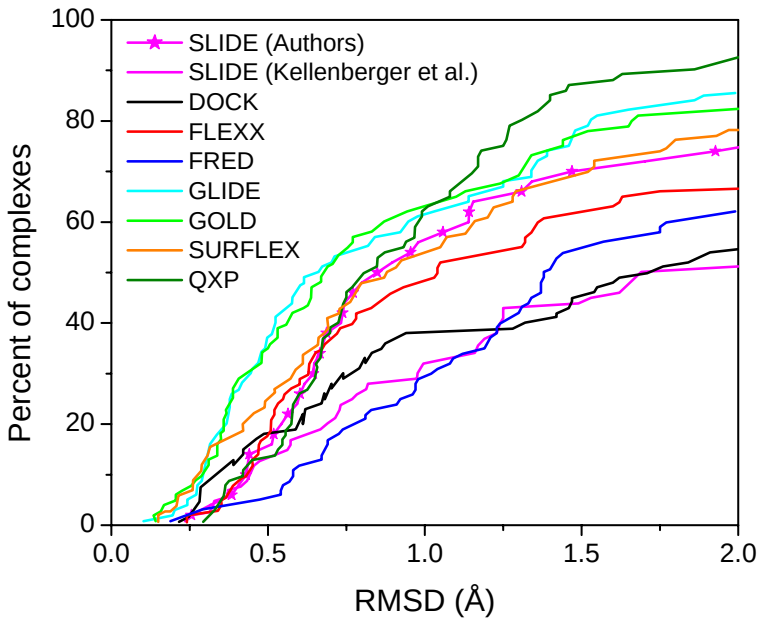
SLIDE (Authors)
SLIDE (Kellenberger et al.)
DOCK
FLEXX
FRED
GLIDE
GOLD
SURFLEX
QXP
Percent of complexes
RMSD (Å)
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0

Figure 1.

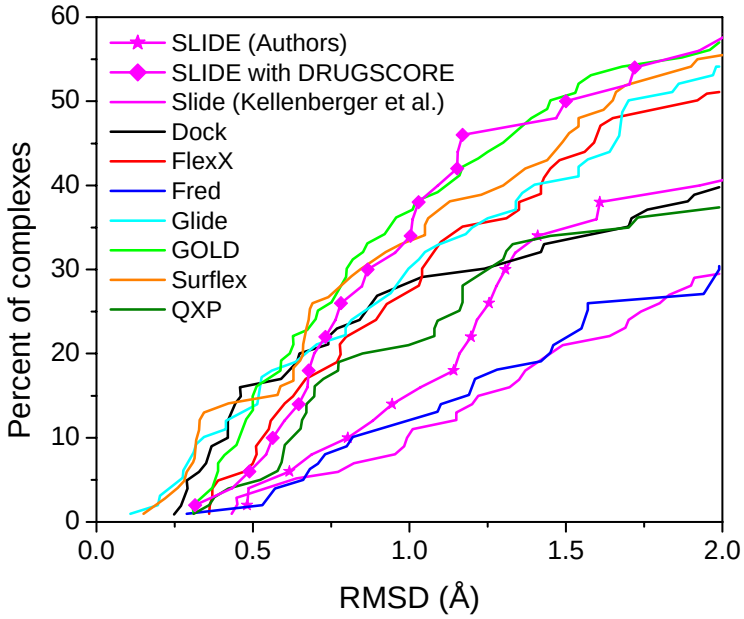
SLIDE (Authors)
SLIDE with DRUGSCORE
Slide (Kellenberger et al.)
Dock
FlexX
Fred
Glide
GOLD
Surflex
QXP
Percent of complexes
RMSD (Å)
0
10
20
30
40
50
60
0.0
0.5
1.0
1.5
2.0

Figure 2.

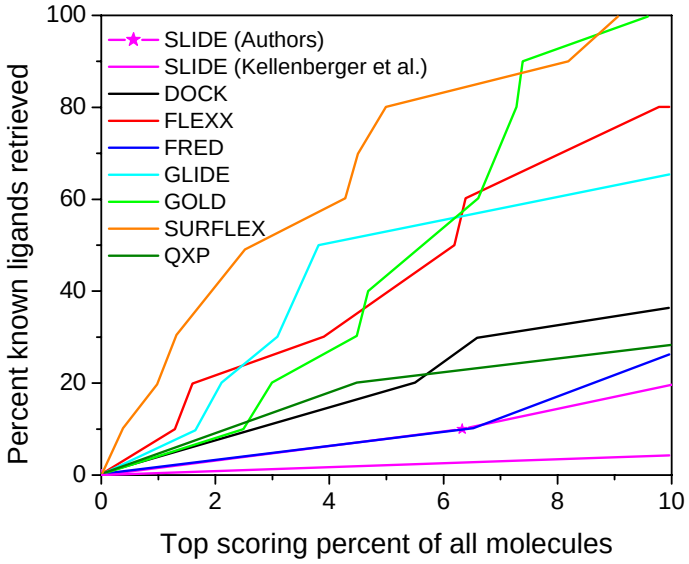
SLIDE (Authors)
SLIDE (Kellenberger et al.)
DOCK
FLEXX
FRED
GLIDE
GOLD
SURFLEX
QXP
Percent known ligands retrieved
Top scoring percent of all molecules
0
20
40
60
80
100
0
2
4
6
8
10

Figure 3.

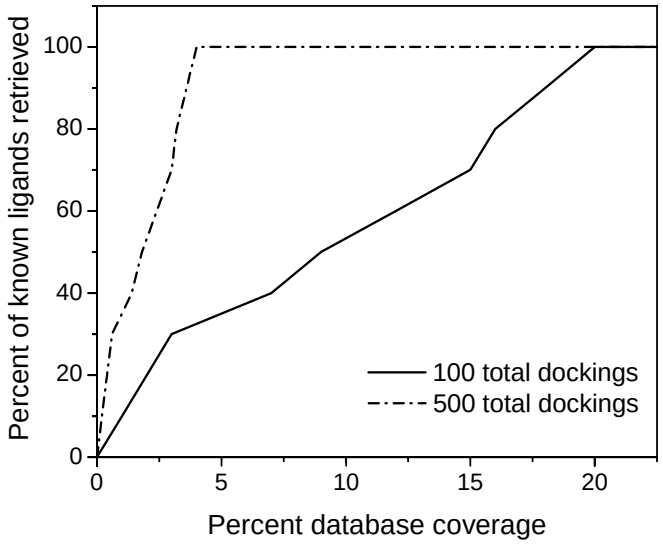
Percent of known ligands retrieved
100
80
60
40
20
0
0
5
10
15
20
Percent database coverage
100 total dockings
500 total dockings

Figure 4.

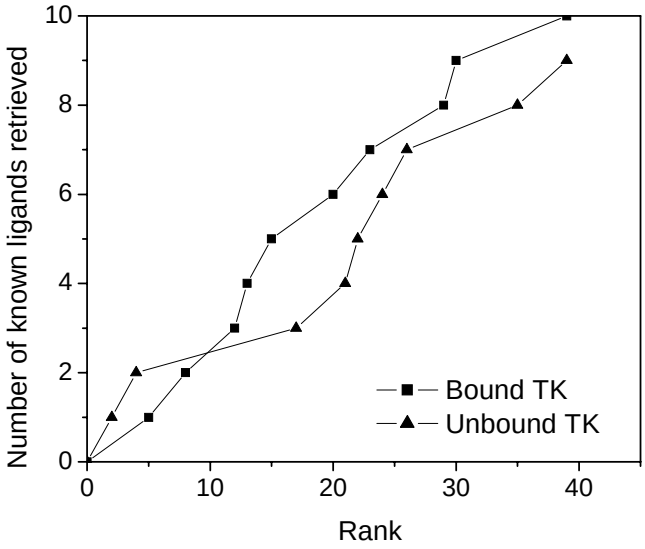
Number of known ligands retrieved
Rank
Bound TK
Unbound TK
0
2
4
6
8
10
0
10
20
30
40

Figure 5.

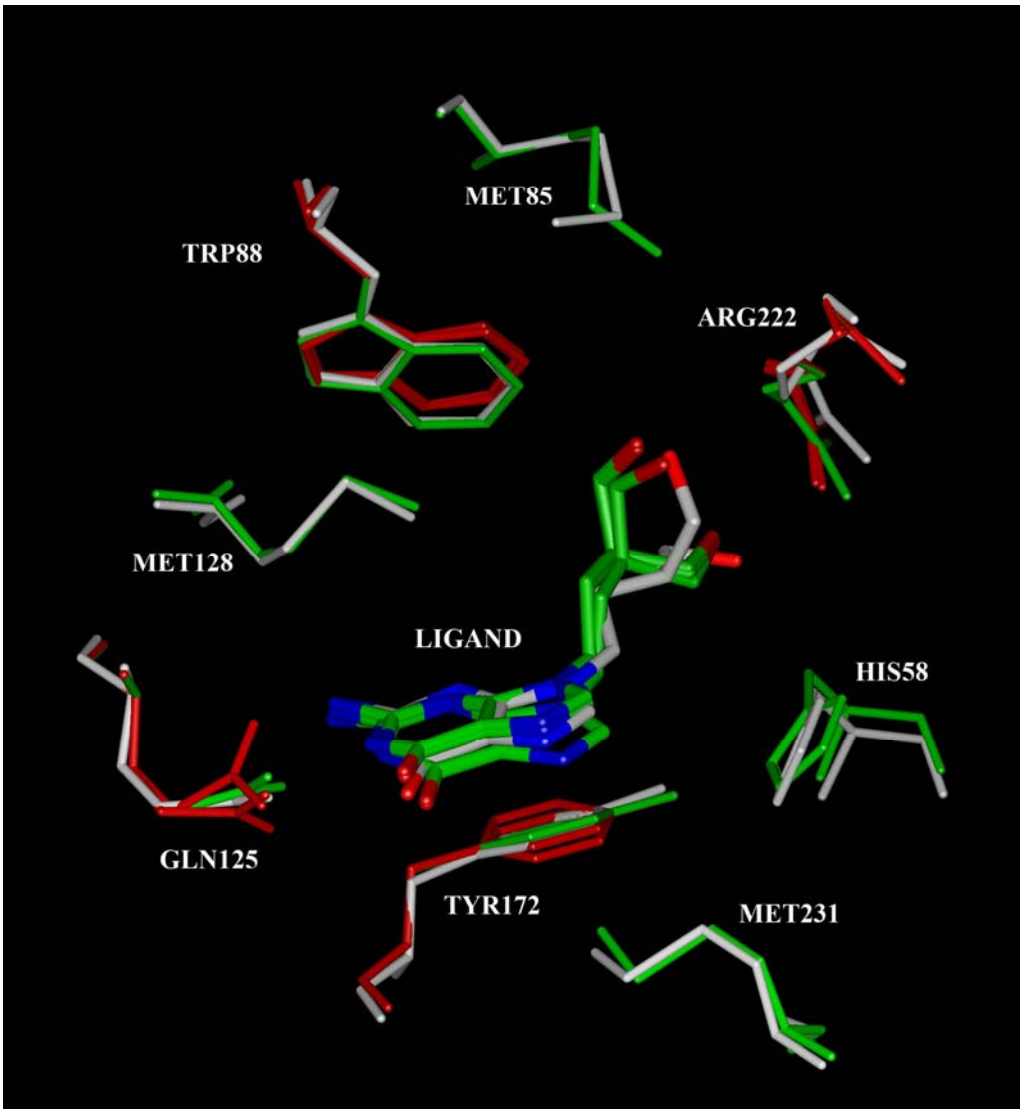
MET85
TRP88
ARG222
MET128
LIGAND
HIS58
GLN125
TYR172
MET231

Figure 6.

References

1. Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 2000;43(25):4759-4767.
2. Bursulaya BD, Totrov M, Abagyan R, Brooks CL, III. Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des 2003;17(11):755-763.
3. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB. Consensus scoring for ligand/protein interactions. J Mol Graph Model 2002;20(4):281-295.
4. Dixon JS. Evaluation of the CASP2 docking section. Proteins 1997;Suppl 1:198-204.
5. Ha S, Andreani R, Robbins A, Muegge I. Evaluation of docking/scoring approaches: a comparative study based on MMP3 inhibitors. J Comput Aided Mol Des 2000;14(5):435-448.
6. Kellenberger E, Rodrigo J, Muller P, Rognan D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 2004;57(2):225-242.
7. Kontoyianni M, McClellan LM, Sokol GS. Evaluation of docking performance: comparative data on docking algorithms. J Med Chem 2004;47(3):558-565.
8. Kontoyianni M, Sokol GS, McClellan LM. Evaluation of library ranking efficacy in virtual screening. J Comput Chem 2005;26(1):11-22.
9. Krovat EM, Langer T. Impact of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. J Chem Inf Comput Sci 2004;44(3):1123-1129.
10. Perola E, Walters WP, Charifson PS. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 2004;56(2):235-249.
11. Schulz-Gasch T, Stahl M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. J Mol Model (Online) 2003;9(1):47-57.
12. Stahl M, Rarey M. Detailed analysis of scoring functions for virtual screening. J Med Chem 2001;44(7):1035-1042.
13. Xing L, Hodgkin E, Liu Q, Sedlock D. Evaluation and application of multiple scoring functions for a virtual screening experiment. J Comput Aided Mol Des 2004;18(5):333-344.
14. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 2002;47(4):409-443.
15. Zavodszky MI, Sanschagrin PC, Korde RS, Kuhn LA. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. J Comput Aided Mol Des 2002;16(12):883-902.
16. Cole JC, Murray CW, Nissink JW, Taylor RD, Taylor R. Comparing protein-ligand docking programs is difficult. Proteins 2005;60(3):325-332.
17. Jain AN. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 2003;46(4):499-511.
18. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins 2003;52(4):609-623.
19. McMartin C, Bohacek RS. QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 1997;11(4):333-344.
20. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 2000;295(2):337-356.

21. Marrone TJ, Luty BA, Rose PW. Discovering high-affinity ligands from the computationally predicted structures and affinities of small molecules bound to a target: A virtual screening approach. Perspectives in Drug Discovery and Design 2000;20(1):209-230.
22. Rarey M, Kramer B, Lengauer T. Time-efficient docking of flexible ligands into active sites of proteins. Proc Int Conf Intell Syst Mol Biol 1995;3:300-308.

Supplementary Material

Recommended commands for Omega and Corina

1. *3D conformer generation*, if a low energy 3D conformer for the ligand candidate is not provided:

```
corina –i t=sdf -o t=mol2 -d wh,stergen,preserve,rc,rs,mc=10,names
<input_filename>.sdf <output_filename>.mol2
```

2. *Assigning partial charges and hydrogen atom positions in ligand candidates*, if not already provided:

```
molcharge -in <input_filename>.mol2 -out <output_filename>.mol2 –am1bcc
```

AM1BCC charges will not be assigned to molecules with unusual atom types, for example B, Co, etc. In this case AM1 charges can be assigned. The command for doing this is:

```
molcharge -in <input_filename>.mol2 -out <output_filename>.mol2 –am1
```

3. *Sampling 3D conformers of flexible ligands.* The maximum number of conformers to generate depends on the maximum number of rotatable bonds in the molecules being docked. Estimating three states for each single bond in a ligand candidate, the maximum number of conformers will be approximately 3^n, where n is the number of single bonds. The default maximum number of 400 conformers in Omega can be reset using the –maxconfs option. The v. 1.8.3b Omega command line used for processing the 1010 molecules in the ligand screening set:

```
omega -in <input_filename> -out <output_filename>.mol2
        -includeinput true -multioutputfiles false -warts true
        -rms 1.0 -fixcycle true -finalopt true -finalcut true
        -finalsheffield true -ewindow 7.5
```

Standard mode of running SLIDE

SLIDE v. 2.3 was used with the following commands and parameter settings for both the redocking and screening studies. SLIDE is available to academic and commercial groups; see http://www.bch.msu.edu/labs/kuhn under Software/SLIDE for details. The Quick Guide to SLIDE and manual (available at the same website) provide more information on the following commands plus utilities for analyzing SLIDE results.

1. *Organizing the ligand database in standard directories for screening, and calculating hydrophobic and hydrogen-bonding interaction points for the ligand candidates.* This step needs to be performed only once for a given database, and can be reused for other docking and screening runs.

```
setup_dbase <target> unbiased <database> <dbase_loc> <target>.pdb
```

2. *Preparing protein templates for redocking and screening.* The template, generated above the protein surface and based only on protein surface chemistry in the case of the unbiased templates used here, represents favored positions to position hydrophobic and hydrogen-bonding groups in a ligand candidate to interact with the protein. This template needs to be generated only once for a protein. Though not done here, SLIDE templates can also include information about known ligand binding interactions (using the biased template mode) or be tuned when visual assessment of the template or docking results suggests ways of improving the template (e.g., increasing hydrophobic point sampling when the site is very hydrophobic, or removing template points that appear outside the binding pocket). All hydrogen-bonding points in the 1kim template for TK were defined as key points, indicating that at least one of these hydrogen-bonding points should be matched by a ligand candidate during docking. Because we generally seek dockings that make at least one hydrogen bond, setting all hydrogen-bonding template points as key points is generally recommended and decreases the run time. However, the 50

redocking cases included highly nonpolar complexes, so the default mode of assigning all template points (including hydrophobic) as key points was used.

Ligand information was used only to define the bounding box for the binding site template. For TK screening, four ligands from PDB entries 1ki4, 1ki8, 1vtk, and 3vtk were used to define the bounding box; none of these was in the screening set of 10 known TK ligands. For the redocking study, in each of the 50 cases, the single ligand being docked was used to define the binding site box. This introduces minimal bias, because the ligand only defines the minimum and maximum x,y,z coordinates of the box, which is then expanded by 2 Å in each direction. Because these axes are typically not aligned with the ligand's own major and minor axes, this box has a far greater volume than the ligand itself. Furthermore, only a fragment of the ligand is required to fall within this box. The template generation command used for all runs, using the redocking case 1fkg as an example, was:

```
temp_gen -l 1fkg unbiased sparse 0.5 4.0 ligand_1fkg.mol2
[ligand_XXXX.mol2]
```

This defined an unbiased template with sparse hydrogen-bonding points, an initial grid spacing of 0.5 Å for hydrophobic point placement, and clustering of hydrophobic points only if they were within 4.0 Å of one another.

3. *Running SLIDE.* The following command invoked SLIDE with the previously defined ligand database, target, and template files, including all atoms within a radius of 9 Å from any template point for scoring the protein-ligand dockings.

```
run_slide <target> unbiased <database> 9.0
```

The default scoring function within SLIDE was used; this is a weighted sum of intermolecular hydrogen-bond and hydrophobic interaction terms. (Note that in v. 2.3 of SLIDE, scores that are

more positive are more favorable, though the v. 2.4 SLIDE scoring function uses the more common convention of more negative scores being better.)

The SLIDE parameters used for all runs were as follows; see the Quick Guide to SLIDE (at `http://www.bch.msu.edu/labs/kuhn` under Software - SLIDE) for more details:

```
DME_THRESHOLD:                          0.3
RMS_THRESHOLD:                          0.3
ANCHOR_TRANSLATION:                     0.3
ANCHOR_OVERLAP:                         0.3
SIDE_CHAIN_OVERLAP:                     0.3
INTRA_OVERLAP:                          0.1
INTERMEDIATELY_TOLERATED_OVERLAP:       2.0
FINALLY_TOLERATED_MAX_BUMP:             0.5
FINALLY_TOLERATED_OVERLAP:              2.0
SCORE_CUTOFF:                           20
MAX_TEMPLATE_TRIANGLES:                 1500000
```