Evaluation of Docking Performance: Comparative Data on Docking Algorithms

Maria Kontoyianni,* Laura M. McClellan, and Glenn S. Sokol

Computer Assisted Drug Discovery, Johnson & Johnson Pharmaceutical Research & Development, LLC, Welsh and McKean Roads, P.O. Box 776, Spring House, Pennsylvania 19477

Received June 19, 2003

Docking molecules into their respective 3D macromolecular targets is a widely used method for lead optimization. However, the best known docking algorithms often fail to position the ligand in an orientation close to the experimental binding mode. It was reported recently that consensus scoring enhances the hit rates in a virtual screening experiment. This methodology focused on the top-ranked pose, with the underlying assumption that the orientation/conformation of the docked compound is the most accurate. In an effort to eliminate the scoring function bias, and assess the ability of the docking algorithms to provide solutions similar to the crystallographic modes, we investigated the most known docking programs and evaluated all of the resultant poses. We present the results of an extensive computational study in which five docking programs (FlexX, DOCK, GOLD, LigandFit, Glide) were investigated against 14 protein families (69 targets). Our findings show that some algorithms perform consistently better than others, and a correspondence between the nature of the active site and the best docking algorithm can be found.

Introduction

Structure-based drug design methods utilize knowledge of the three-dimensional structure of a receptor complexed with a lead molecule in an attempt to optimize the bound ligand or a series of congeneric molecules. Docking plays an important role in this process by placing a molecule into the active site of the target macromolecule in a noncovalent fashion. In that light, docking can be viewed as a search or optimization method which, given the degree of conformational flexibility at the macromolecular level, can be a very challenging problem.[1−5] Regardless of the many difficulties, structure-based drug design methods have become valuable tools in the design of new chemical entities by attempting to predict and explain their binding modes, when the active site is known. The increasing number of X-ray, NMR, and model-built structures of receptors and enzymes have made the above methodologies quite useful in the pharmaceutical industry.

Docking consists of two parts, namely, the accurate prediction of the orientation (pose) of the bioactive conformation into the binding pocket, and the estimation of the tightness of target−ligand interactions (scoring). Several approaches have been employed in an attempt to solve the docking problem. Almost all current docking programs perform flexible ligand docking, but treat the receptor as rigid, with the exception of GOLD, which applies some limited flexibility to the active site side chains.[6] The methods these programs are based upon vary from incremental construction approaches, such as FlexX,[7] to shape-based algorithms (i.e., DOCK[8]), genetic algorithms (GOLD[6]), systematic search techniques (Glide[9]), and Monte Carlo simulations (LigandFit[10]).[11]

The number of target−ligand complexes that these programs have reportedly been validated against varies as well. FlexX was verified on a set of 19 protein−ligand complexes in the original paper,[7] with a subsequent evaluation of a larger set of 200 complexes,[12] while GOLD was validated on 100 complexes.[6] Most recently, an even larger dataset of 305 protein−ligand complexes was used to evaluate GOLD.[13] Glide may have been rigorously assessed internally, since it is an industrial rather than an academic algorithm; however, no published reports have appeared to date. LigandFit was reported recently for 19 protein−ligand complexes,[10] while DOCK has been verified on several targets over the years;[14−21] however, we are not aware of a comprehensive report in regard to its overall performance.

It was reported recently that docking programs are able to predict experimental poses with deviations averaging from 1.5 to 2 Å rms[11,22] However, this has not been our experience with the available docking programs. Furthermore, in recent years a number of papers exploring the performance of the docking programs in database searching have been published.[11,23−25] These reports investigated the accuracy of the scoring functions, after docking is completed, thus making the underlying assumption that the docking procedure is successful in identifying the experimental pose accurately.[26] Whether or not a docking program will affect the hit rates of an in silico screening methodology remains to be answered. However, prior to answering the above question, one needs to know how accurate the best-known docking programs are in finding experimental solutions for ligand−target complexes in a comprehensive and comparative fashion. Thus, a reference study comparing the strengths and limitations of docking programs is still missing.

Consequently, we decided to explore whether docking programs do indeed find experimental solutions for target−ligand complexes. We also wished to explore

* To whom correspondence should be addressed. Phone: (215) 628-5236. Fax: (215) 628-4985. E-mail: mkontoyi@prdus.jnj.com.

10.1021/jm0302997 CCC: $27.50
Published on Web 12/31/2003

which docking algorithms perform best for specific receptor families, and what the ranking is of the experimental binding mode of the ligand, when a docking program is able to reproduce it. We have performed an extensive computational study in which five docking programs (FlexX, DOCK, LigandFit, Glide, and GOLD) were investigated against 14 protein families (69 targets). We report here the results of these algorithms in terms of their ability to predict the binding modes of the respective protein-bound ligands relative to the experimental poses, their comparative performance against the same protein families, and what the ranking of the crystallographically observed binding modes is, if identified by a program.

Results and Discussion

Special attention was given to the targets chosen for this study. The criteria used were (1) varied crystallographic resolution of chosen targets, (2) wide spectrum of receptor families, (3) metal presence in the binding pocket, (4) range of active site topologies and water accessibility, (5) varied flexibility of receptor-bound ligands, and (6) activities of bound ligands varying from the low micromolar to nanomolar range. We also chose targets with more than one bound ligand per crystal structure, to explore the ability of the programs to handle various conformations of the same receptor, and eliminate potential failures due to their inability in dealing with induced fit. Consequently, we believe that our dataset is comprehensive, large enough, and demanding. The PDB identification codes of the proteins studied in this work are in the first column of Table 1. Furthermore, the ionization state of the ligands we were attempting to dock was of particular concern to us. All carboxylic acids were deprotonated, amines were positively charged, phosphonates were partially deprotonated, and guanidiniums were positively charged.[11]

Because there is no reference study that we could use as a guide to set the adjustable parameters required by each program to run, we performed diagnostic calculations for each docking algorithm with a range of settings and evaluated the results. Given that we were investigating single-ligand dockings, our objective was accuracy and exhaustive search for all plausible target−ligand complexes, rather than speed. For all targets, the active sites were defined within a 12 Å radius from the bound ligand or an amino acid central to the binding pocket, unless otherwise noted below or in the computational methods. In a previous work on FlexX, the active sites were defined within 6.5 Å from a ligand atom;[12] however, we chose a more enlarged definition of the binding pockets because, first, this exercise should not cater to a particular algorithm for one and, second, we were concerned that smaller sites might not be able to accommodate larger ligands. Waters and metals not involved in binding were removed from the protein. For consistency purposes, we chose to increase the number of return poses to 60 for all programs, although in some cases, such as DOCK, the calculations did not result in as many poses.

In FlexX, we used formal charges, which were assigned by the program, and default parameters otherwise. In LigandFit, the protein model was generated after the ligand and solvent removal. The definition of the binding site was based on the docked ligand. Therefore, those grid points inside the protein, which lay within a consistent distance from the ligand atoms and were not occupied by the ligand, formed the site. Stochastic conformational searching was applied to the ligands with a higher than the default number of Monte Carlo search steps, to ensure extensive conformational sampling. Charges were assigned via the CFF1.01 force field,[27,28] rather than the charge equilibration method.[29] Because the difference in the docking score between the two methods was very slight, we did not use the charge equilibration method to calculate charges.[30] Ligands were scored with the PLP1,[31,32] PLP2,[33] Ludi,[34] PMF,[35] and Ligscore scoring functions, the latter using CFF in turn. For Ligscore a higher grid extension of 12.0 Å was utilized to lower the time used for scoring. Regarding DOCK 4.01, partial Gasteiger−Marsilli charges were calculated. To increase accuracy, we increased the number of maximum orientations to 500 for the anchor fragment, and to 25 configurations per cycle for growth of the ligand. In GOLD, we performed 50 genetic algorithm runs, as opposed to the default 10. Glide presented a bigger challenge since it is the newest code; thus, we leaned toward using the default settings for the most part. However, we used 0.90 to scale the vdW radii of the nonpolar ligand atoms, which may be a little less permissive than the standard 0.80 but not by much, and kept 60 poses per ligand in the end. We also discovered that it is crucial for Glide to have the correct ionization state, because the "atomtype" functionality is not able to add hydrogens; therefore, a complete ligand structure is required. Using the utility script glide_sort, we reordered the poses on the basis of the Emodel scoring function, which was then used to analyze our results. [9] Emodel is a weighted combination of E_{CvdW}, GlideScore, and the strain energy of the ligand, with E_{CvdW} referring to the ligand−receptor Coulomb−van der Waals interaction energy.

To assess the docking accuracy of each algorithm, we used two approaches. All solutions were inspected visually and evaluated on the basis of the rms deviation of the predicted pose from the experimental bound orientation. With visual inspection as a guide, we classified the solutions in a subjective manner similar to the one used in the GOLD paper.[6] We first selected the poses that were the closest to the respective binding modes, and subsequently assigned a classification (close, active site, inaccurate) on the basis of the definitions that follow. A "close" classification corresponds to solutions that reproduce the bioactive conformation and most, if not all, of the important binding interactions by positioning the functional groups of the ligand in proximity to the active site residues. Thus, this category includes predictions which may not be fully superimposable onto the crystallographic binding modes, but show a rather minimal displacement of one group per ligand. The "active site" category includes predictions where the binding mode is reproduced by and large but a few groups are pointing toward active site residues that are different from the corresponding crystallographically observed pose. Finally, the assignment "inaccurate" means that the algorithm was not able to reproduce the binding mode or that it failed to position the ligand in the binding site. Figures 1−3 show

Table 1. Summary of Docking Predictions[a]

	subjective results				
PDB code	LigandFit (rank)	DOCK (rank)	FlexX (rank)	Glide (rank)	GOLD (rank)
			Thermolysin		
2tmn	active site (1)	active site (15/17)	inaccurate (1)	close (1)	close (30)
3tmn	active site (1)	inaccurate	inaccurate (39)	active site (36)	close (1)
4tmn	active site (58)	inaccurate	inaccurate (1)	close (6)	close (39)
5tmn	inaccurate (58)	inaccurate	inaccurate (3)	close (2)	close (31)
5tln	inaccurate (59)	active site (23/23)	close (20)	close (27)	active site (30)
			Carbonic Anhydrase		
1bnt	active site (7)	active site (24)	inaccurate (5)	active site (1)	active site (34)
1bnm	inaccurate (1)	inaccurate	close (27)	close (48)	active site (1)
1bn1	active site (4)	inaccurate	close (17)	active site (40)	close (1)
1i9l	active site (47)	active site (53)	active site (19)	close (1)	inaccurate
1i9n	active site (49)	active site (32)	active site (59)	close (30)	active site (1)
			Stromelysin		
1hy7	active site (25)	active site (36/53)	close (9)	close (22)	active site (30)
1g49	active site (4)	active site (7/60)	close (4)	close (37)	close (27)
1g4k	active site (9)	inaccurate	inaccurate (3)	close (1)	inaccurate
1d5j	inaccurate (59)	inaccurate	inaccurate (39)	active site (38)	active site (29)
1d8f	inaccurate (40)	active site (10)	active site (38)	active site (36)	inaccurate (1)
1d8m	close (5)	active site	close (13)	close (1)	inaccurate (1)
1biw	active site (10)	inaccurate	inaccurate (9)	close (1)	close (14)
1bqo	active site (1)	inaccurate	active site (1)	active site (1)	active site (20)
1sln	inaccurate (55)	active site (1)	inaccurate (51)	active site (60)	active site (11)
			Aspartate Carbamoyltransferase		
1d09	active site (39)	active site (1)	close (1)	close (1)	close (11)
4at1	active site (44)	active site (19)	active site (4)	inaccurate (60)	active site (19)
			Dihydrofolate Reductase		
1aoe	close (2)	inaccurate	active site (8)	close (2)	close (7)
1boz	active site (8)	inaccurate	inaccurate (3)	close (1)	close (12)
1daj	active site (2)	inaccurate	inaccurate (12)	active site (9)	close (50)
1dg5	active site (6)	inaccurate	active site (2)	active site (48)	close (36)
1dg7	close (3)	inaccurate	close (19)	close (2)	close (12)
1hfp	active site (24)	inaccurate	active site (5)	active site (27)	close (15)
1ia1	active site (15)	inaccurate	close (29)	close (1)	close (10)
3dfr	active site (5)	active site (15)	close (1)	inaccurate (34)	close (12)
4dfr	active site (1)	active site (12)	close (1)	close (1)	close (32)
			Thymidine Kinase		
3vtk	close (55)	active site (52)	inaccurate (6)	close (1)	close (44)
2ki5	close (11)	inaccurate	close (1)	close (27)	inaccurate
1vtk	close (1)	inaccurate	close (1)	close (1)	close (39)
1ki8	close (6)	active site (36/36)	inaccurate	close (10)	close (42)
1ki4	close (26)	inaccurate	inaccurate	close (1)	close (10)
1kim	close (1)	inaccurate	active site (52)	close (1)	inaccurate
			HIV-1 Protease		
2bpx	close (5)	inaccurate	active site (53)	inaccurate (1)	close (27)
2bpz	inaccurate (12)	inaccurate	active site (10)	active site (53)	close (19)
2bpw	active site (4)	inaccurate	inaccurate (22)	close (7)	close (19)
1tcx	close (1)	inaccurate	inaccurate (16)	close (1)	close (18)
1upj	close (11)	close (27)	inaccurate (7)	active site (23)	inaccurate
1sbg	close (4)	inaccurate	inaccurate (1)	close (1)	close (4)
			COX-2		
6cox	close (15)	inaccurate	active site (7)	close (2)	close (29)
3pgh	close (2)	active site (27)	close (8)	close (4)	close (25)
4cox	close (5)	inaccurate	close (6)	close (8)	active site (11)
			CDK-2		
1dm2	inaccurate (3)	close (34/52)	active site (11)	active site (2)	active site (50)
1aq1	close (5)	close (30)	close (1)	close (17)	close (28)
1ckp	active site (1)	active site	close (6)	close (1)	close (12)
1di8	active site (6)	active site	close (2)	close (34)	inaccurate
1fvt	active site (52)	inaccurate	close (1)	active site (2)	close (9)
1fvv	active site (26)	close (40)	close (33)	close (6)	close (15)
			FGFR-1		
1agw	active site (58)	active site (39)	close (6)	active site (2)	close (1)
1fgi	active site (2)	inaccurate (44)	active site (1)	active site (1)	active site (1)
2fgi	active site (5)	inaccurate	close (8)	active site (32)	close (1)
			Reverse Transcriptase		
1rt1	close (22)	inaccurate	close (20)	close (4)	close (36)
1rt5	close (1)	inaccurate	inaccurate	active site (1)	close (27)
1rt7	active site (37)	inaccurate	inaccurate	active site (1)	active site (50)

Table 1. Continued

PDB code	subjective results				
	LigandFit (rank)	DOCK (rank)	FlexX (rank)	Glide (rank)	GOLD (rank)
			PPAR-γ		
1fm6	close (7)	inaccurate	active site (1)	inaccurate (3)	close (8)
1fm9	active site (21)	inaccurate	active site (34)	inaccurate (1)	close (49)
			TACE TNF-α Converting Enzyme		
1bkc	active site (50)	inaccurate	inaccurate	active site (1)	close (47)
			Neuraminidase		
1a4g	close (1)	active site (12)	wrong (21)	active site (23)	close (1)
1nsd	close (16)	inaccurate	close (1)	close (3)	close (1)
2bat	inaccurate (2)	inaccurate	active site (11)	active site (32)	close (18)
2qwk	active site (42)	inaccurate	active site (48)	close (7)	close (1)
1inf	active site (44)	inaccurate	close (24)	active site (1)	close (11)
1a4q	active site (21)	inaccurate	active site (1)	active site (19)	active site (10)
1b9s	inaccurate (10)	inaccurate	inaccurate (1)	inaccurate	close (1)
1b9t	close (1)	inaccurate	close (2)	close (11)	close (1)
1b9v	close (4)	inaccurate	inaccurate	close (1)	close (1)

[a] The numbers in parentheses represent the rank numbers of the respective poses.

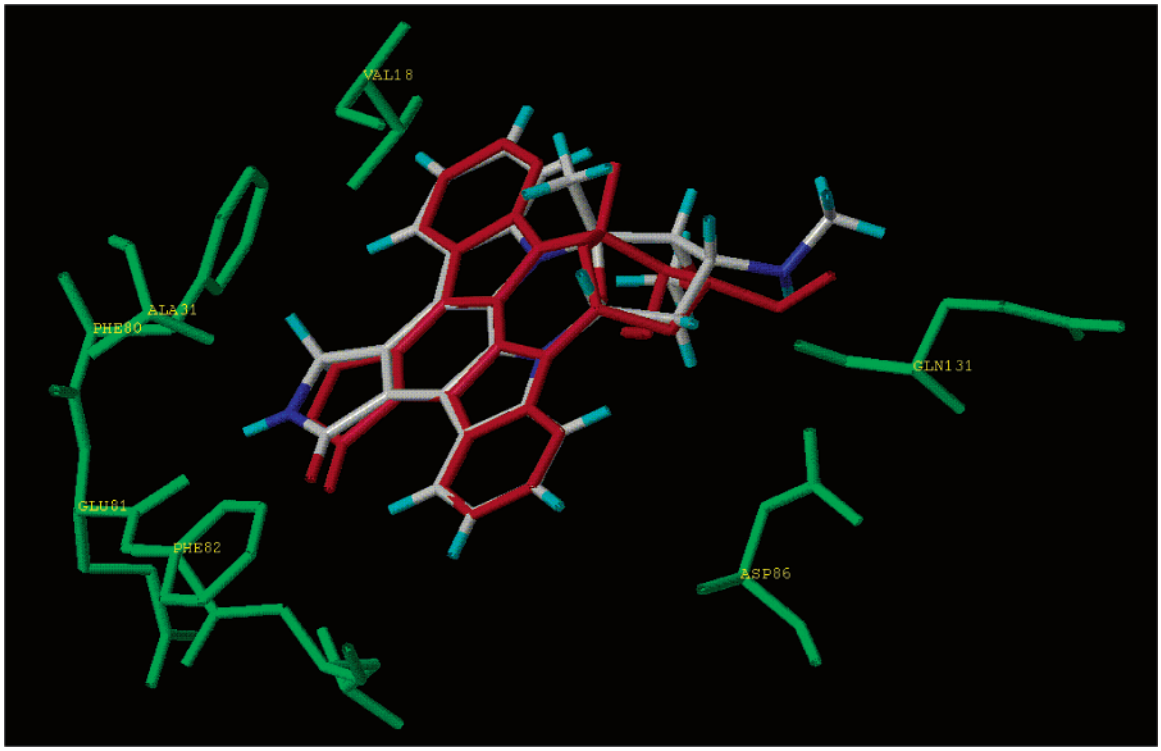

Figure 1. Example of a close prediction (target 1aq1).

representative examples of close, active site, and inaccurate predictions.

The results of our analyses reflecting these subjective assignments are shown in Table 1. Once the selection of the solutions that were the closest to the experimentally observed binding modes was made, we looked at the corresponding rank numbers, which are represented by the numbers in parentheses in Table 1. For inaccurate solutions, the one that would remotely resemble the observed binding pose is given. In the DOCK column, the second number corresponds to the total number of solutions per run found by the program; for all other docking tools the total number of returned poses was 60, as already discussed. Also, it should be noted that the LigandFit and Glide results are based on Ligscore and Emodel scoring functions, respectively.[9,10] If we ignore the numbers in parentheses for now, but instead pay attention to the frequency with which the close solutions appear in Table 1, it can be seen that very few of the Glide and GOLD solutions fall into the inaccurate category. This is more obvious in Table 2, which summarizes the results of Table 1. Table 2 shows how many times each program succeeds in finding poses close to the experimental binding modes, and how many of these solutions are in the active site or inaccurate. It can be seen that GOLD outperforms (47 correct poses out of the total 69 investigated) the other docking programs in being able to identify the experimental binding modes for the most part, followed by Glide (39 close predictions). The numbers in parentheses (Table 2) correspond to the total number of solutions ranked the highest by the respective programs, which in turn addresses the ranking accuracy of the docking algorithms under investigation, and will be discussed in the following paragraph. It should also be noted that for certain receptor families in Table 1, the

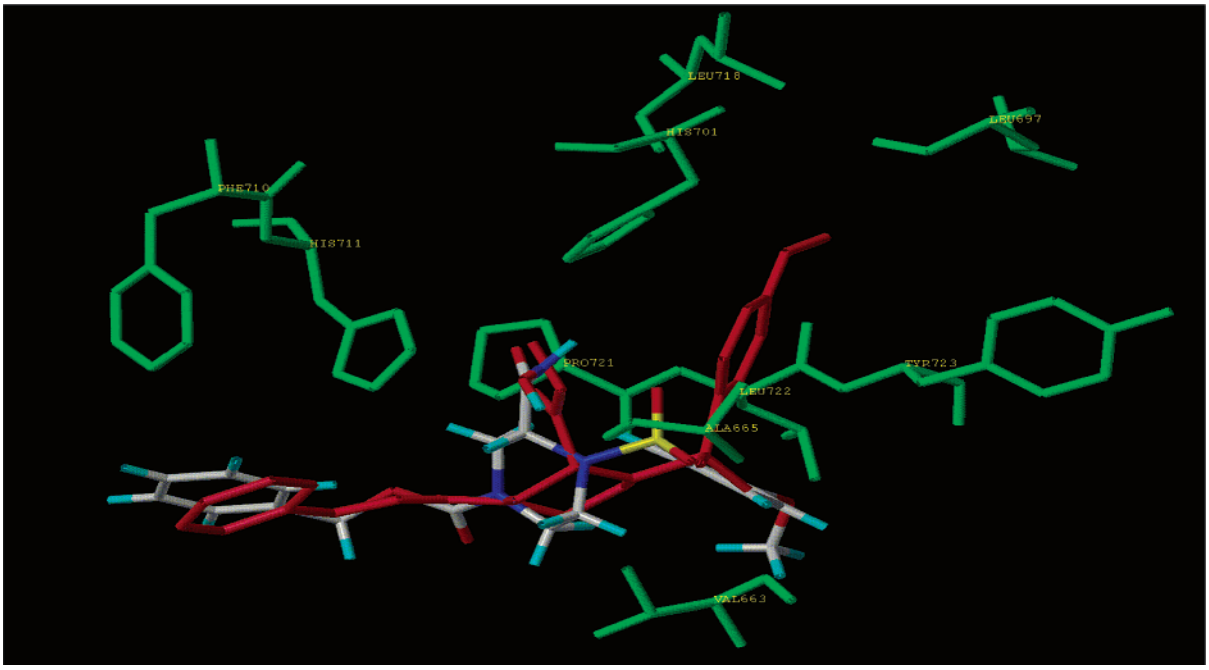

Figure 2. Example of an active site prediction (target 1d8f).

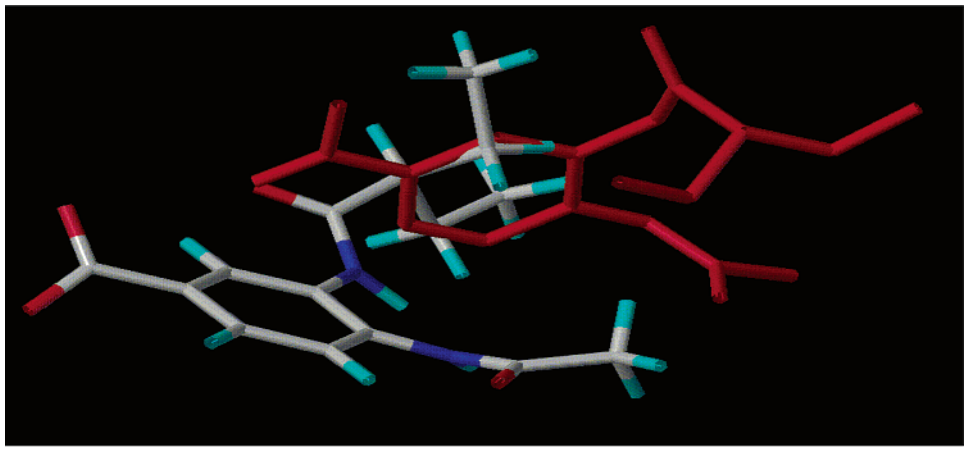

Figure 3. Example of an inaccurate prediction (target 1b9s).

Table 2. Summary of Docking Accuracy

	no. of results[a]		
program	close	active site	inaccurate
LigandFit	24 (7)	35	10
GOLD	47 (10)	13	8
DOCK	4 (0)	19	45
Glide	39 (17)	24	5
FlexX	26 (9)	18	25

[a] The values in parentheses indicate how many times each program finds the experimental binding mode as its top-scoring answer.

poses identified by some of the programs are either close only or a combination of close and active site (for example, see Glide on thymidine kinase and GOLD for neuraminidase). This suggests that some programs perform better for particular targets; thus, a correspondence between the nature of the active site and a docking algorithm may be implied. Because this study did not cover the whole spectrum of crystal ligand−target complexes reported for each protein family we investigated, it may be risky to draw conclusions relating families with programs. However, given that we studied 70% of the reported thymidine kinase complexes and 90% of the COX-2 crystal structures, we can say that Glide and LigandFit should be the programs of choice for targets of these families. Glide and FlexX are the best performers in CDK-2, on the basis of coverage of half of the reported crystal complexes for this family. Similarly, GOLD should be preferred for neuraminidases and thermolysins, with the first studied in its entirety, while 41% of the latter was investigated here. Glide is also excellent for the thermolysins under investigation.

Furthermore, we addressed the ranking accuracy of the docking algorithms, that is, whether the experimental binding modes of the ligands are found among the highest-ranked conformations by any of the programs. As already mentioned, the numbers in parentheses in Table 1 represent the ranking for each solution, which was the closest to the experimentally observed binding mode. It seems rather difficult to make claims that any of the algorithms identify the bound ligands among the highest-ranked poses. This becomes clearer in Table 2, which shows in parentheses how many correct answers

Table 3. Summary of rms Deviations

	LigandFit		GOLD		Glide		FlexX	
rms	total	close, act	total	close, act	total	close, act	total	close, act
≤0.5	0	0, 0	6	6, 0	3	3, 0	2	2, 0
0.5−1.0	10	10, 0	25	25, 0	18	18, 0	9	9, 0
1.0−1.5	10	8, 2	12	12, 0	12	12, 0	10	10, 0
1.5−2.0	6	3, 3	5	1, 4	7	1, 6	10	4, 6
2.0−3.5	21	0, 21	5	0, 5	14	0, 14	10	0, 10
>3.5	22	0, 5	16	0, 5	14	0, 1	28	0, 2

(i.e., top-ranked poses) each program resulted in for the 69 complexes investigated. Although GOLD is more reliable in identifying the experimentally observed binding modes of the ligands, only 10 out of its 47 close solutions correspond to top-scoring poses. Our results do not seem to agree with the high rate of 71% that Jones et al. reported in their validation of the GOLD docking tool.[6] They reported that the top-scoring answers were correct for 71 out of the 100 complexes they tested. We are even more perplexed, because our genetic runs were longer than theirs. It may be that the scoring function used by GOLD does not perform as well for the targets studied herein. In contrast, we found that Glide performs better than GOLD in ranking the binding modes correctly. If we were a bit permissive and looked at the top six Glide answers, then 26 of the 39 correctly predicted binding modes are top scored, a rate of 67%. This appears to be our best scoring tool. To summarize, when Glide does well, it will most likely rank the observed pose as its top-scoring binding mode, while GOLD may not succeed at doing that, which is an observation made already by the developers of GOLD.[6]

As mentioned earlier, we looked at the results of the docking experiments statistically as well. Table 3 displays the rms deviations of the predicted poses from the corresponding observed binding modes for FlexX, GOLD, Glide, and LigandFit. DOCK was not included, because it did not perform as well as the other algorithms (see Table 1). The majority of the close solutions have rms deviations less than 1.5 Å for all programs. The active site poses fall mostly into the category of 1.5−3.5 rms. However, it can be seen that some of the active site solutions ranging from 1.5 to 2.0 rms were considered as close poses in the subjective analysis, if we compare the sums for close assignments in Tables 2 and 3. This finding is in accord with the classification used by Kramer et al., who considered results up to 2.0 Å rms as acceptable.[12] However, our criteria are more stringent than the ones used in the validation paper of GOLD, where the reported success rate of 71% represents answers with rms deviations of 3.0 Å or less.[6] This also sheds light on the discrepancy between our results and those of the developers of GOLD regarding GOLD's success rate in identifying the experimental binding modes as its top-ranked poses. Interestingly enough, almost 66% of the correct solutions found by GOLD are within an rms of less than 1.0 Å, while 54% of the correct Glide solutions fall in that range. The data in Table 3 seem to reiterate what we also observed through our subjective analysis; that is, GOLD is quite reliable in finding experimental binding modes.

In an attempt to understand what the factors are for the failure of the programs, we decided to analyze our results in terms of crystal structure resolution, nature of the active sites, such as polarity versus hydrophobicity, and complexity of the bound ligands. The results are displayed in Tables 3−5, respectively. Analysis of Table 3 indicates that all programs performed well, as long as the crystal structure resolution was 2.5 Å or better. However, it should be pointed out that LigandFit seems to be more sensitive, since a major part of its close poses fall into the most accurate X-ray resolution of less than 2.0 Å (see Table 3).

Table 4. Summary of Protein Structure Resolution and Distribution of Poses Accordingly

				close	
no. of targets	resolution	LigandFit	GOLD	Glide	FlexX
17	<2.0	12	12	14	9
38	≥2.0, ≤2.5	4	26	15	10
14	>2.5	8	9	10	7

Table 5. Ligand Complexity and Distribution of the Close Poses

no. of rotatable bonds	GOLD	LigandFit	Glide	FlexX
5	10	6	9	8
>5, ≤10	15	12	17	10
>10, ≤15	12	3	6	6
>15	10	3	7	2

In regard to complexity of the ligands, GOLD seems to be the least sensitive of all. In our hands, the studied docking programs perform best when the ligands have 10 or fewer rotatable bonds. An analysis of Table 4, however, shows that LigandFit to a greater extent than the other tools identifies 75% of its close dockings when the ligands have 10 or fewer rotatable bonds. For FlexX the rate is 69% for less than 10 rotatable bonds, which increases to 92% if the ligand has 15 or fewer degrees of freedom. Our findings are in agreement with previous analyses of FlexX.[12] In contrast, Glide is less sensitive with a success rate of 78% when the complexity is smaller than 15.

Table 5 gives an overview of the nature of the active sites of the 66 targets we investigated. Results were obtained using MOLCAD.[36] The lipophilic potential was mapped globally for all active sites, to bring the values on the same scale so that the reported surface area properties would be consistent. It can be seen that most targets are mildly hydrophilic. Stromelysin, HIV protease, and TACE are mostly hydrophilic, while only one of our targets, COX-2, is mainly hydrophobic. We looked at the results of Table 1 in conjunction with the findings of Table 5 to deduce whether polarity can be a cause for some of the failures of the programs. GOLD seems to perform very well with mildly or mostly hydrophilic targets. On the basis of the observation that GOLD fails with hydrophobic ligands,[6] which in turn could be extended to lipophilic binding pockets, our results do not surprise us. However, we did obtain successful

dockings with GOLD in the cases where there is some lipophilic character in the active site (i.e., thermolysins and PPAR-γ). Contrary to GOLD, Glide does not seem as discriminatory in regard to the nature of the polarity of the active sites. Both LigandFit and Glide performed well with COX-2, a target with a mainly hydrophobic binding pocket.

Conclusions

The results presented here indicate that certain docking algorithms are more reliable in reproducing experimentally observed binding modes than others. Among all the programs studied, GOLD, followed by Glide, is clearly the most reliable in predicting accurate poses. It was also pleasing to observe trends relating docking accuracy with the nature of the active sites under investigation. Our results demonstrate that, for the given data set of 69 receptor−ligand complexes, one may be able to predict which docking tool will work the best depending upon the nature of the active site. However, we should bear in mind that this is a representative study, and therefore, our conclusions may be limited to the complexes studied here. Although the experimental binding modes of the ligands were not found among the highest-ranked conformations by any of the programs, we were particularly pleased to observe a higher success rate with Glide's ability to identify binding modes as its top-scoring poses. Although general rules cannot be drawn, it appears that there is a certain degree of confidence associated with this particular tool in being able to rank the poses more accurately than others. With the industry's need for virtual screening and the necessity for scoring functions, which can discriminate actives from inactives, this finding is quite encouraging. However, one needs to factor in not just accuracy and performance, but speed as well. Glide tends to be very slow (2−3 min per compound), while LigandFit is the fastest (10−18 s) among the codes presented in this study. We believe that the choice of a docking program is interrelated to the objective of the project. If single-ligand docking is performed, and the target belongs to one of the structural types we and others have reportedly investigated, the choice is perhaps more straightforward. Even if there is no study alluding to one code being more advantageous than another, we are confident enough that either Glide or GOLD will perform well. However, for library screening, and given that most corporate libraries consist of about 1 million compounds, we would recommend first using a fast code (i.e., LigandFit), followed by the more accurate ones on the hits identified in the first step.

This study shows that, despite its challenges, docking is a powerful tool. Although scoring is difficult, predicting the correct pose appears to be a very tangible exercise. Coupling this with our findings that Glide succeeds in identifying and ranking the observed binding modes at a 40% rate, we are hopeful that structure-based design is making noticeable advances.

Computational Methods

Preparation of Proteins and Ligands. All bound waters, ligands, and cofactors were removed from the proteins. Histidines and cysteines were mutated to HID and CYX, respectively, for DOCK calculations. Kollman charges were computed, when necessary (DOCK), using Sybyl. Hydrogen atoms were added subsequently. For each target, the active site was defined as a sphere with a radius of 12 Å from the bound ligand or an amino acid in the center of the site. Flexible ligand docking was performed in all calculations. Sixty solutions were reported from each docking exercise. Ligands were either extracted and distorted from their bound conformations or built within Sybyl. Ionizable groups (amines, carboxylic acids, phosphates, amidines) were assumed to be ionized at physiological pH. Ligands were assigned formal charges for the FlexX and Glide calculations, and partial Gasteiger−Marsilli charges for the DOCK runs.

Table 6. Polarity of Active Sites of the Targets

	area (%)		
PDB target	hydrophilic	mildly hydrophilic	lipophilic
thermolysin	0.00	65.00	35.00
carbonic anhydrase	0.00	89.80	10.20
stromelysin	82.33	17.66	0.00
aspartate carbamoyltransferase	19.24	80.76	0.00
DHFR	2.73	97.27	0.00
thymidine kinase	0.00	100.00	0.00
HIV	93.54	6.46	0.00
COX-2	0.00	7.31	92.69
CDK-2	0.00	100.00	0.00
FGFR-1	1.38	98.62	0.00
reverse transcriptase	51.16	48.84	0.00
PPAR-γ	0.09	82.64	17.27
TACE	64.46	35.54	0.00
neuraminidase	0.00	100.00	0.00

FlexX Docking. All default parameters, as implemented in the 6.72 release of Sybyl, were used. Cscore calculations were performed for ranking, and all 60 poses were inspected.

LigandFit Docking. LigandFit from Cerius2, consortium and 4.6 versions, within the Accelrys suite of programs was used for all runs. The active sites were defined using the docked ligands, and all calculations were performed with the CFF 1.01 force field. Conformations were generated with Monte Carlo simulations (10000 trials). Electrostatic energy was included in the calculation of the ligand internal energy. The default rigid body minimization parameters were used to dock the four orientations of each conformation into the site, followed by a 500-step final minimization of each docked ligand. Scoring was performed with PLP1, PLP2, PMF, Ludi, and Ligscore scoring functions. Hydrogen-bond and lipophilic contributions to the Ludi score were included in all calculations. Individual descriptors were checked for all Ligscore function calculations.

DOCK 4.01 Docking. A Connolly surface of each active site was generated using a 1.4 Å probe radius. A flexible docking was performed starting with a selection and matching of an anchor atom within a maximum of 500 orientations, followed by growth of the ligand with 25 configurations per cycle. The final step included relaxation of 100 simplex minimizations to a convergence of 0.2 kcal/mol.

Glide Docking. Distances from a grid point to the receptor surface were compared to distances from the ligand center to the ligand surface. Good matches were kept, followed by a clash test, subset scoring, greedy scoring, and final refinement of 5000 initial poses in the $X/Y/Z$ directions. The resultant 400 refined poses were kept, and then minimized with a distance-dependent dielectric constant, and 100 conjugate gradient steps. Final poses were scored with GlideScore with an inclusion of an energy score.

GOLD Docking. For each of the GA runs, a maximum number of 100000 operations were performed on a population of 100 individuals. Operator weights for crossover, mutation, and migration were set to 95, 95, and 10, respectively, which are the standard default settings recommended by the authors for careful work. The distance for hydrogen bonding was set to 4 Å, and the cutoff value for van der Waals was 2.5.

Acknowledgment. We thank Drs. Adrea Mehl, Hege Beard, Hal Almond, and Mr. Jon Swanson for useful discussions during the course of this work, and Dr. Terry Lybrand for reading the manuscript and for his invaluable comments. Dr. Chuck Reynolds is also acknowledged for reading the manuscript. Finally, we thank Mrs. Nancy Walter for technical assistance with the manuscript.

References

(1) Morris, G. M.; Olson, A. J.; Goodsell, D. S. Protein-ligand docking. *Methods Princ. Med. Chem.* **2000**, *8*, 31−48.
(2) Mestres, J.; Knegtel, R. M. A. Similarity *versus* docking in 3D virtual screening. *Perspect. Drug Discovery Des.* **2000**, *20*, 191−207.
(3) Vieth, M.; Hirst, J. D.; Dominy, B. N.; Daigler, H.; Brooks, C. L., III. Assessing search strategies for flexible docking. *J. Comput. Chem.* **1998**, *19*, 1623−1631.
(4) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **1997**, *72*, 1047−1069.
(5) Monard, G.; Merz Jr. K. M. Combined Quantum Mechanical/ Molecular Mechanical Methodologies Applied to Biomolecular Systems. *Acc. Chem. Res.* **1999**, *32*, 904−911.
(6) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm to flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(7) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A. A fast flexible docking method using an incrememental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−89.
(8) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langidge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(9) Schrödinger, Portland, OR 97201.
(10) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; and Waldan, M. LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289−307.
(11) 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*, 4759−4767.
(12) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein−ligand docking. *Proteins* **1999**, *37*, 228−241.
(13) Nissink, J.; Willem, M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein−ligandinteraction. *Proteins: Struct., Funct., and Genet.* **2002**, *49*, 457−471.
(14) Hopkins, S. C.; Vale, R. D.; Kuntz, I. D. Inhibitors of kinesin activity from structure-based computer screening. *Biochemistry* **2000**, *39*, 2805−2814.
(15) Debnath, A. K.; Radigan, L.; Jiang, S. Structure-bases identification of small molecule antiviral compounds targeted to the gp41 core of the human immunodeficiency virus type. *J. Med. Chem.* **1999**, *42*, 3203−3209.
(16) Bodian, D. L.; Yamasaki, R. B.; Buswell, R. L.; Stearns, J. F.; White, J. M.; Kuntz, I. D. Inhibition of the fusion-inducing conformational change of influenza hemagglutinin by benzoquinones and hydroquinones. *Biochemistry* **1993**, *32*, 2967−2978.
(17) Shoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M. Structure-based discovery of inhibitors of thymidylate synthase. *Science* **1993**, *259*, 1445−50.
(18) Arorov, A. M.; Munagal, N. R.; Ortiz De Montellano, P. R.; Kuntz, I. D.; Wang, C. C. Rational design of selective submicromolar inhibitors of tritrichomonas foetus hypoxanthine-guanine-xanthine phosphoribosyltransferase. *Biochemistry* **2000**, *39*, 4684−91.
(19) Chen, Q.; Shafer, R. H.; Kuntz, I. D. Structure-based discovery of ligands targeted to the RNA double helix. *Biochemistry* **1997**, *36*, 11402−11407.
(20) Gschwend, D. A.; Sirawaraporn, W.; Santi, D. V.; Kuntz, I. D. Specificity in structure-based drug design: identification of a novel, selective inhibitor of Pneumocystis carinii dihydrofolate reductase. *Proteins* **1997**, *29*, 59−67.
(21) Checa, A.; Ortiz, A. R.; de Parscual-Teresa, B.; Gago, F. Assessment of solvation effects on calculated binding affinity differences: trypsin inhibition by flavonoids as a model system for congeneric series. *J. Med. Chem.* **1997**, *40*, 4136−4145.
(22) Dixon, J. S. Evaluation of the CASP2 docking section. *Proteins* **1997**, Suppl. 1, 198−204.
(23) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(24) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281−295.
(25) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(26) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.
(27) Hagler, A. T.; Ewig, C. S. On the use of quantum energy surfaces in the derivation of molecular force fields. *Comput. Phys. Commun.* **1994**, *84*, 131−155.
(28) Jwang, M.-J.; Stockfisch, T. P.; Hagler, A. T. Derivation of class II force fields. 2. Derivation and characterization of a class II force field, CFF93, for the alkyl functional group and alkane molecules. *J. Am. Chem. Soc.* **1994**, *116*, 2515−2525.
(29) Rappe, A. K.; Goddard, W. A., III. Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.* **1991**, *63*, 867−872.
(30) Adrea Mehl, private communication.
(31) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317−324.
(32) Gehlhaar, D. K.; Moeerder, K. E.; Zichi, D.; Sherman, C. J.; Ogden, R. C.; Freer, S. T. De novo design of enzyme inhibitors by monte carlo ligand generation. *J. Med. Chem.* **1995**, *38*, 466−472.
(33) Gehlhaar, D. K.; Bouzida, D.; Rejto, P. A. *Rational Drug Design: Novel Methodology and Practical Applications*; Parrill, L., Reddy, M. R., Eds.; ACS Symposium Series 719; American Chemical Society: Washington, DC, 1999; pp 292−311.
(34) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structures. *J. Comput.-Aided Mol. Des.* **1994**, *7*, 385−391.
(35) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(36) Connolly, M. L**.** Analytical molecular surface calculation *J. Appl. Crystallogr.* **1983**, *16*, 548−558.

JM0302997