Comparison of Automated Docking Programs as Virtual Screening Tools

Maxwell D. Cummings,*,† Renee L. DesJarlais,† Alan C. Gibbs,‡ Venkatraman Mohan,‡ and Edward P. Jaeger†

Johnson & Johnson Pharmaceutical Research & Development, Eagleview Corporate Center, 665 Stockton Drive, Exton, Pennsylvania 19341, and Johnson & Johnson Pharmaceutical Research & Development, Cedarbrook Corporate Center, 8 Clarke Drive, Cranbury, New Jersey 08512

Received March 11, 2004

The performance of several commercially available docking programs is compared in the context of virtual screening. Five different protein targets are used, each with several known ligands. The simulated screening deck comprised 1000 molecules from a cleansed version of the MDL drug data report and 49 known ligands. For many of the known ligands, crystal structures of the relevant protein−ligand complexes were available. We attempted to run experiments with each docking method that were as similar as possible. For a given docking method, hit rates were improved versus what would be expected for random selection for most protein targets. However, the ability to prioritize known ligands on the basis of docking poses that resemble known crystal structures is both method- and target-dependent.

Introduction

The practice of testing large numbers of molecules for activity in a model system that is representative of a human disease, known as screening, is well-established in the pharmaceutical industry. High-throughput screening technologies allow for the testing of thousands to millions of molecules for activity against a new target system as part of the drug discovery process. The goal and challenge of screening as applied to pharmaceutical lead discovery is to select a small number of candidate lead molecules from a large and diverse collection of molecules. The screen often employs an assay designed to detect molecules that bind specifically to a target protein. Molecules of interest will be those that show relatively high activity in the screening method applied. Further prioritization of lead molecules based on the assessment of various molecular characteristics, ADME properties, or selectivity profiles is often carried out immediately after confirmation of the initial screening result.

One goal of computational chemistry is to predict the binding interactions of molecules. An aspect of this ambitious goal that is of particular interest to the pharmaceutical industry is the modeling of the binding of small molecules to proteins, commonly referred to as docking, because this has direct application in the processes of drug discovery and drug design. Over the past 25 years, since the initial descriptions of automated computer-based docking programs,[1,2] much progress has been made in the field of docking-based virtual high-throughput screening. For example, the seminal program DOCK, first described in 1982 by Kuntz and colleagues,[2] continues to find application and evolve as a docking-based virtual screening tool.[3−19] Many other docking programs have been reported as well, and the reader is referred to several recent reviews for perspectives on the state-of-the-art.[20−22]

An experimental high-throughput screening system that relies on a protein−ligand binding interaction can be used to examine a collection of molecules for those that interact with the target protein. Docking-based screening is an in silico analogy of a protein binding or competition experiment. By application of some configurational (and, possibly, conformational) search strategy to small organic molecules in a defined region of the target protein, the direction of which is guided by some score function, a collection of molecules is evaluated for virtual binding to the target protein. In both in vitro and in silico protocols, each ligand is ultimately assigned a score. After thus searching through a large collection of potential ligands, those molecules with the best score are interesting. Ideally, the results obtained with the virtual screening method would be identical to those obtained with an experimental screening method.

A standard procedure adopted when applying experimental screening methods to pharmaceutical lead discovery and development is to test one or more control compounds, which have some well-established activity, or lack thereof, with respect to the target system under study.[23] Initially this can serve as a test to validate the idea that the screen under study will serve to identify new ligands with the desired activity, and when applied repetitively throughout the duration of a screening campaign serves to ensure that the screening method continues to detect molecules that exhibit the desired activity. Such controls also provide information on the normal variation of the observed experimental results. Docking-based virtual screening methods have been evaluated by exploring their ability to prioritize (i.e., rank well) known active compounds that have been seeded into a collection of inactive (either known or presumed) compounds.[22,24−27] In virtual screening validation, success is defined as the ability to enrich some relatively small fraction of best-scored ligands with respect to the proportion of seeded known actives

* Author to whom correspondence should be addressed. Phone: 610-458-5264 ext. 6581. Fax: 610-458-8249. E-mail: mcumming@prdus.jnj.com.

† Johnson & Johnson Pharmaceutical Research & Development, Eagleview Corporate Center.

‡ Johnson & Johnson Pharmaceutical Research & Development, Cedarbrook Corporate Center.

10.1021/jm049798d CCC: $30.25
Published on Web 01/20/2005

therein. When the suitability of a docking-based virtual screening method is tested as a general tool for pharmaceutical lead discovery, additional parameters of interest include the ability to reproduce known binding modes, level of success with a variety of different target proteins and different binding interactions, the ability to treat and succeed with a wide diversity of potential ligands, and the CPU time required per ligand.

We became interested in putting in place a docking-based virtual screening system as a tool to apply to the problem of selecting target-specific screening subsets from our compound collection. Several commercially available docking programs seemed applicable to the problem of virtual high-throughput screening, and we sought to determine which one(s) seemed most appropriate for our purpose. Most published descriptions of new or established docking programs provide an informative level of detail on how the programs work and how they perform with some test systems. Thus, it is often relatively straightforward to determine that a particular program may or may not be well-suited to a particular docking-based problem or a particular protein target. Choosing one of several programs for general use is much more challenging, because of the difficulty of comparing the results of nonequivalent method validation tests. Although different score functions have been compared as secondary (re-)scoring tools in docking-based virtual screening,[24,25,27] there are relatively few reports describing the direct comparison of different docking tools applied to identical virtual screening problems.[24,25,28−30] Furthermore, for our purposes, previously published reports were limited with respect to the programs compared or the number of targets tested. We were interested in comparing independently validated docking programs against a common set of virtual screening problems, while simulating as much as possible the conditions of real virtual screens carried out in the context of pharmaceutical drug discovery. To identify a generally applicable docking tool, we concluded that we would need to run our own comparative tests of the docking methods. Here we describe our comparison of four different docking programs as tools for virtual high-throughput screening.

Methods

Target Proteins. Five target proteins were used in our tests: human immunodeficiency virus protease (HIV-Pr), protein tyrosine phosphatase 1b (PTP1b), thrombin, urokinase plasminogen activator (uPA), and the human homologue of the mouse double minute 2 oncoprotein (HDM2). HIV-Pr is an aspartyl protease that is the protein target of several current AIDS therapeutic agents.[31] This enzyme binds relatively hydrophobic peptides, with the involvement of multiple intermolecular hydrogen bonds. Synthetic inhibitors also tend to have large hydrophobic substituents. Protein tyrosine phosphatase 1b (PTP1b) is a signal transducing enzyme that recognizes and dephosphorylates phosphorylated tyrosine residues on intracellular proteins. PTP1b is under investigation as a potential target for chemotherapeutic intervention in both type II diabetes and obesity.[32] Because PTP1b recognizes charged tyrosyl-phosphate groups, it is not surprising that small negatively charged ligands are found to interact strongly with PTP1b. Thrombin is a trypsin-like serine protease that is part of the blood clotting cascade. Compounds that inhibit thrombin are expected to have anticoagulant effects.[33] Thrombin and its complexes with inhibitors are well-studied. The Protein Data Bank[34] (PDB) contains approximately 150 thrombin structures, and there are many more proprietary structures within pharmaceutical companies. The active site includes a subsite that binds basic groups (S1) and a subsite that binds an aromatic group (S4). In addition to S1 interactions, there are several other important hydrogen bonds as might be expected for a protease. Urokinase plasminogen activator (uPA) is a trypsin-like serine protease associated with tumor invasion and metastasis.[35,36] There are approximately 20 uPA structures available from the PDB and likely many more proprietary structures within pharmaceutical companies. The active site includes an S1 subsite that binds basic groups but is otherwise rather flat. Many inhibitors take advantage of a hydrophobic subsite termed S1′ that is not part of the binding site for peptide substrates.[37] HDM2 has a large, nonenzymatic hydrophobic binding site that binds the p53 protein. Bound ligands, both natural and synthetic, form relatively few intermolecular hydrogen bonds with this target protein. This protein is under investigation as a possible target for chemotherapeutic intervention in various cancers. The PDB identification codes for the HIV-Pr, PTP1b, and thrombin structures that we used as docking targets are 1HVR,[38] 1C84,[39] and 1QBV,[40] respectively. For uPA and HDM2, unpublished structures were used as docking targets.

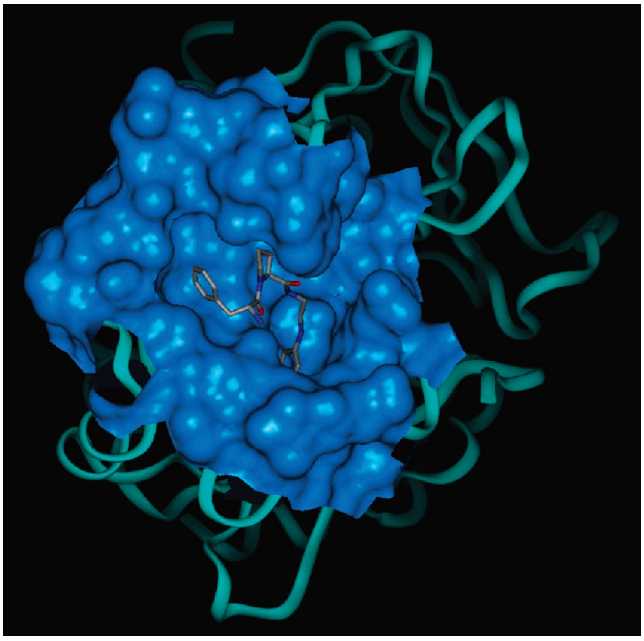

Figure 1. Example of a generously large binding site to be searched during a docking experiment. This figure shows the binding region defined with DOCKVISION when screening thrombin in our docking study. The thrombin structure shown here is PDB entry 1QBV.

Binding Site Definition. Binding site definition is one powerful tool for biasing a docking search. For our comparison, we used large binding sites (e.g., Figure 1) to avoid the possibility of bias introduced from binding site differences across the different docking methods. Although we did not come near to searching the entire protein surface with any method for any target protein, we are confident that in most "real world" quests for lead molecules we would define somewhat smaller binding sites. Each of the docking methods we tested uses a different approach for binding site definition, and it was not always clear that we could duplicate binding site definitions (with respect to, e.g., total search volume, target protein residues comprising the search surface) with the different docking methods. Furthermore, the methods differ in how a molecule being docked is restrained to the defined binding site. These concerns were at least partially addressed by using large binding site definitions. In addition, we attempted to keep the location and volume of a given target site the same for all docking methods. Further details of how

we implemented this philosophy with each of the different methods are provided in the following sections (see also Figure 1).

Ligand Database. An important element of the testing and comparison that we have undertaken is the database of molecules to be screened. Verdonk and colleagues have recently presented an insightful discussion on approaches to be taken in constructing databases for validation of docking-based virtual screening methods.[41] Their work suggests that studies aimed at validating a particular docking method should include tests involving sets of known actives and decoys with similar properties to the actives (the extreme here is distinguishing actives from inactives in a congeneric series).

We desired a single database of molecules for virtual screening against all of our target proteins as this most closely mimics the virtual screening experiment one would conduct to select molecules for a physical screen. We constructed the database so that it spanned the range of physical properties of our actives, but we did not bias the database to have an identical distribution of properties as this would not likely be the case in a physical screen carried out in the context of pharmaceutical drug discovery.

Our approach to creating a test database is similar to methods outlined previously by other research groups.[24,42] A set of criteria[43] was applied to remove pharmaceutically undesirable molecules from v2002.1 of the MDL Drug Data Report (MDDR).[44] We randomly selected 1000 molecules from the remaining MDDR compounds to serve as the set of presumed inactive molecules for our docking-based virtual screening tests. To this, we added 49 known active molecules, yielding a database of 1049 molecules (10 actives each for PTP1b, thrombin, and uPA, 5 for HIV-Pr, and 14 for HDM2). The two-dimensional (2D) structures of the seeded actives for HIV-Pr, PTP1b, and thrombin are shown in Figure 2 (for proprietary reasons we are currently unable to disclose the HDM2 and uPA actives). Lipinski[45] described several molecular features of marketed drugs that have since become common parlance in discussions of "drug-likeness". Figure 3 compares the actives and decoys in our database, with respect to these properties. This analysis establishes that most of the actives fall within the ranges spanned by the decoys, providing some assurance that we have not introduced a simple ligand-based bias into our testing protocol.

A single two-dimensional SDF file was prepared comprising the complete database. Formal charges and hydrogens were added to the complete set. The assigned charge state was carefully checked for all of the molecules. The 2D-to-3D conversion was carried out with CONCORD[46] or CORINA.[47] We used CORINA for the active molecules as it maintained the specified stereochemistry. We established that conversion with CORINA[47] gave similar results to those obtained with CONCORD for a subset of our MDDR compounds. The stereochemistry of the conformer thus generated for each of the known active molecules was checked by visual inspection. This final 3D database was used for all of our docking-based virtual screening tests. Reagent filtering, atom-based formal charge assignment, hydrogen addition, and random selection were carried out with various components of the Directed-Diversity toolset.[48] All three-dimensional molecular coordinates, for both actives and inactives, were generated from an initial two-dimensional database.

The database comprising the three-dimensional coordinates of all the ligand molecules except those of the HDM2 and uPA actives is provided as part of the Supporting Information.

Reproduction of Known Binding Modes. Preliminary experiments were run for at least one of the target systems for each of the docking methods to gain an understanding of the conditions required to reproduce a known binding mode. These initial tests were successful and instructive, but we found that when the conditions that we ultimately decided upon were applied to the problem of database screening the results obtained for known crystal structures were not always consistent with our initial tests. The discrimination of different methods on this basis is an important finding of the present comparison study (see Results section below).

We examined the ability of each of the methods to reproduce known binding modes during the virtual screening experiments. For each target, we had crystal structures of target−ligand complexes for several of the active molecules that were seeded into the screening database. For each of these actives, the heavy atom root-mean-squared deviation (rmsd) of the relevant docking pose was calculated for each docking method. For this comparison, protein structures were superimposed by minimizing the rmsd between all backbone heavy atoms of the two relevant protein structures. Rmsd's were calculated with the auxiliary tool supplied with DOCKVISION[49,50] (RMSTOOL).

Consensus Analysis of Docking Results. Consensus analysis was performed for pairs of output lists from the different docking programs. No rescoring was carried out on the original docking lists. Here, we have adapted the consensus scoring approach[25] to select those results that are ranked well by two different docking programs. Molecules common to the two docking lists under consideration represent the consensus of those two lists. The recently described ConsDock program[51] is based on a similar premise, using docking results obtained with the three docking programs DOCK,[2,5] FLEXX,[52] and GOLD.[53]

DOCK. DOCK 4.0[2,5] was used in these studies. DOCK characterizes concavities on a protein surface using sets of spheres generated from a Connolly surface.[54] Sets of overlapping spheres represent contiguous sites of which several are typically described for a given protein. One sphere set is chosen as the targeted binding site, and the centers of these spheres are used as superposition targets for the atoms of the compound being docked. Compounds are docked piecewise starting with a rigid anchor fragment. Molecules are built up by adding the remaining fragments, using torsions specified in a user-editable parameter file.

For each target, the Connolly molecular surface[54] was calculated using a probe radius of 1.4 Å, and spheres were generated with the DOCK program SPHGEN.[2] All spheres with radii between 1.4 and 3.0 Å for uPA, PTP1b, and HIV-Pr, or 4.0 Å for HDM2 and thrombin, were kept for further consideration. The limit of the dot product between surface normals (dotlim) was 0.0. SPHGEN outputs the spheres in clusters that overlap each other. Clusters were examined for each target, and the cluster covering the known binding site was chosen. For HDM2 and thrombin, some spheres were deleted to keep the target site confined to the primary cavity and to be consistent with the volume of these target sites as defined for other programs. Compounds were docked allowing for ligand flexibility, using the grid-based energy scoring option for minimization after initial placement in the site. Protein structures were prepared for grid calculations by first ensuring correct atom assignments for Asn, Gln, and His residues and then adding hydrogens with INSIGHTII (version 2000.1; Accelrys, Inc.). The box for the scoring grid was defined such that all spheres were enclosed with an extra 5.0 Å added in each dimension. Scoring grids for contact and energy scores were calculated with a grid spacing of 0.3 Å. The bump check was set such that compounds with atoms closer than half the sum of the van der Waals radii of the respective atoms were rejected. The energy cutoff was 99.0 Å. A 6-12 Lennard-Jones van der Waals potential was used, with a distance-dependent dielectric constant of $4r$. The radii used were those in the vdw_cornell.defn set.

Ligand atoms were matched to receptor spheres using the anchor first search with the anchor size set to 10 atoms. The automatic matching option was used, and conformations were generated on the fly with the torsion drive option. The flex_drive.tbl file was modified to allow for a finer torsion search around certain dihedrals.

A complete set of input parameters is provided as part of the Supporting Information.

DOCKVISION. The database version of DOCKVISION RSDB was used for the tests described here; we used version

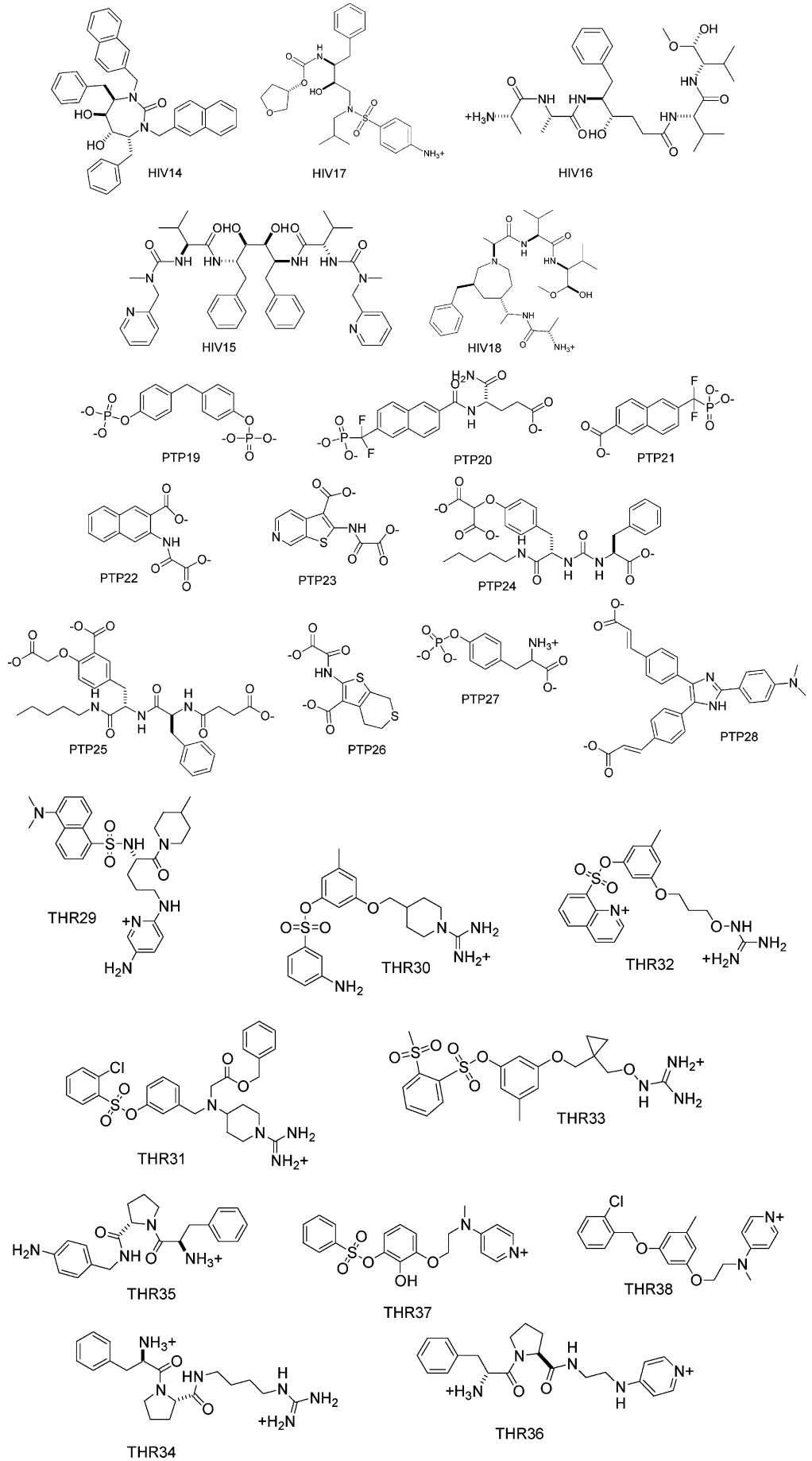

Figure 2. Structures of known actives for three of the docking targets. The seeded actives for HIV-Pr, PTP1b, and thrombin are shown.

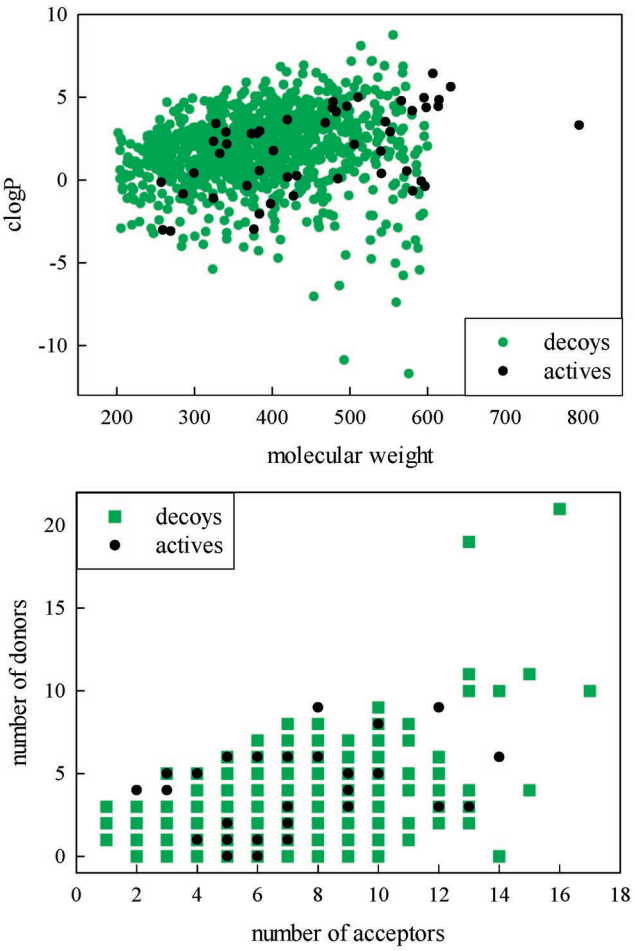

Figure 3. Properties of active and decoy molecules in the screening database.

2 of RSDB as distributed in the DOCKVISION 1.0.3 release.[49,50] RSDB docks each ligand to the target protein multiple times according to a rigid-body Monte Carlo/simulated annealing (MCSA) protocol, and the best docking found by this procedure is then further refined according to a flexible-ligand MCSA protocol. Scoring involves a function that counts intermolecular atom-based hydrogen bond donor−acceptor pairs in conjunction with a grid-based pseudo-van der Waals term.

The provided atom definition library file was used for all of our studies. Ligand atom typing and charge assignments are done on the fly by RSDB and are not user-controllable. The ligand database was converted from the original SDF file (see above) to the required PDB format with the included utility program DBCONVERT.

Explicit polar hydrogens were added with INSIGHTII (version 2000.1; Accelrys, Inc.). Active site polar hydrogen assignments were then visually inspected with particular concern applied to the ambiguous residues His/Asn/Gln, and appropriate adjustments were made when needed. Residue and neutral charge group libraries and the floating and scoring grids for the target proteins were prepared with included programs.

RSDB uses one or more "restraint spheres" for binding site definition. During the docking, the geometric center of the ligand being docked is constrained to remain within the restraint spheres for which the dimensions and Cartesian coordinates are specified by the user. For these docking studies, we used restraint spheres of 8 Å radii centered on the geometric center of a ligand bound at the relevant site for each of the target proteins of interest, as observed in crystal structures of the corresponding or related protein−ligand complexes. For example, the restraint sphere centers for the HIV-Pr and PTP1b tests coincided with the geometric centers of the bound ligands in PDB entries 1HVR[38] and 1C84,[39] respectively, and for thrombin the x, y, and z values of the restraint sphere center were within 0.6 Å of the center of the bound ligand in PDB entry 1QBV[40] (e.g., Figure 1). uPA and HDM2 were treated similarly, using unpublished crystal structures of protein−ligand complexes. For the tests described here, each ligand was subjected to 2000 rounds of multiple start MCSA.[49] The simulated annealing schedules were adjusted by us in our preliminary tests for each target.

A complete set of input parameters is provided as part of the Supporting Information.

GLIDE. GLIDE[55,56] version 2.5 release 12 was evaluated in our comparison. GLIDE follows a hierarchical protocol, going through a sequence of steps each of which serves to focus the ongoing search to good dockings. Ligand conformations are calculated internally. Protein−ligand configuration space is explored by the use of a grid of site points and rotations of the ligand about the site points. Relatively simple scoring is applied initially to reject unreasonable dockings; scoring becomes progressively more sophisticated as the docking search progresses.

Target protein protonation and tautomeric adjustments were performed with the included PPREP script. Hydrogen atoms were added within MAESTRO (version 5.1.016) prior to grid calculations using the "all-atom with no lone pair treatment". GLIDE constructs a grid that defines the ligand-binding site search region, which typically is centered on the ligand of a relevant protein−ligand complex structure. The grid encodes the Coulomb and van der Waals fields, and the interaction of the ligands within these fields is evaluated using linear interpolation formulas for a cubic box. The grid (ligand-binding pocket) is defined as an enclosing box surrounding a bounding box to which the ligand center is restricted during docking. The enclosing box uses a coarse interpolation, typically 3.2 Å, whereas the bounding box has a much finer interpolation, usually 0.4 Å. In our tests, the bounding boxes were 12 Å in all three dimensions, with the center of the box positioned on the center of a bound ligand in a relevant protein−ligand complex (see above), and the enclosing boxes ranged in size from 23.2 to 46.0 Å on a side.

GLIDE handles ligand flexibility by an exhaustive conformational search, augmented by a heuristic screen that eliminates unlikely receptor binding conformations. The exhaustive search examines conformational minima, on average three per rotatable bond, and eliminates unlikely conformers based on torsion energy. Dockings that progress to the later stages of a GLIDE run are evaluated with GlideScore 2.5 SP (standard precision Glide), which measures the interaction energy of each ligand with the Coulomb and van der Waals grids. GlideScore is an extension of the empirical ChemScore function.[57] A small number of the best poses are then minimized on precomputed OPLS-AA van der Waals and electrostatic grids. In our tests, the best scored pose of each molecule was saved for further analysis.

A complete set of input parameters is provided as part of the Supporting Information.

GOLD. Version 1.2 of the GOLD docking program[53] was evaluated in our comparison. GOLD employs a genetic algorithm (GA) search strategy wherein various molecular features of a protein−ligand complex are encoded as a chromosome; both ligand conformation and complex configuration are searched. Dockings are ranked based on an atom-based fitness function that includes terms describing protein−ligand hydrogen bonds, van der Waals interactions, and ligand internal energy.

The binding site was initially defined as all residues of the protein target within 6.5 Å of any ligand atom, based on appropriate reference crystal structures (see above). Subsequent automated cavity detection-based steps augment this initial user definition. Relevant hydrogen-bonding atoms within this surface were assigned using the provided SYBYL atom-typing scheme, and lone pairs were added with the

default geometry. Hydrogen atoms were added with SYBYL (Tripos, Inc.), as GOLD requires explicit definition of all hydrogens.

For the GA search, molecular features of the protein−ligand complex are encoded as a chromosome. A GA move operator is applied to parent chromosomes that are chosen randomly, with a fitness bias, from the existing population. A GA run comprised 100 000 genetic operations on an initial population of 100 members divided into five subpopulations, with migration between subpopulations allowed. GOLD performs a user-specifiable number of GA runs for each ligand, each of which starts from a different random ligand orientation. For our experiments, the number of GA runs per ligand was set to 3.

A complete set of input parameters is provided as part of the Supporting Information.

CPU Times. CPU time is considered to be a critical parameter in many docking studies. On the basis of small preliminary tests, we designed our docking experiments to consume roughly equivalent amounts of CPU time. On a single SGI R10000 processor, most of the docking experiments consumed between 1 and 3 CPU-days; one DOCK experiment and one DOCKVISION experiment consumed between 3 and 4 CPU days. In general, GLIDE tended to be the slowest program. For our purposes, we did not consider variation on the order of 2-fold to be a primary discriminator of performance.

Results

The main objective of our study was to compare the performance of the different docking methods as virtual screening tools when applied to the same problem set. In this context, the primary measure of performance is defined as the ability of the docking program to prioritize seeded active molecules specific for a target over the other molecules in the database. Ideally, the specific actives for a target would be ranked best. In practice, this goal is not achieved, and analysis therefore involves looking at the active molecules present in various fractions of the ranked list. Our perspective is that the fraction of interest is specific to a particular problem. Because our experimental screening approach is to routinely screen a significant fraction of our compound collection (5−50%) against a new protein target of interest, our analysis here spans the 2−50% range for the virtual screen. A second key measure of performance is the extent to which known binding modes are reproduced and ranked well in the docking experiments, and we explore this for all of the actives for which structures were available. Additionally, for a screening method to be of general utility, good performance against different targets is also desired, and we summarize this aspect of our study. We note again that our study involves no rescoring of the docking results with alternative score functions. In practice, rescoring of docking results represents an additional step and challenge to the analysis, because meaningful rescoring demands some local reoptimization of the docking poses being rescored according to each of the rescoring functions. Our study is restricted to a comparison of the performance of various integrated searching/scoring tools as they have been developed.

Prioritization of Known Actives. We begin with a summary analysis that compares the performance of the different methods averaged over all of the protein targets studied (Figure 4). This perspective indicates that when averaged over these targets all of the methods perform better than random selection. Interestingly, this coarse analysis also shows that all of the methods perform similarly when evaluated on the basis of enrichment in the top 10, 20, or 50% of the database. When 10% or less of the database is screened, the program GOLD finds significantly more of the active seeds than the other methods (discussed below).

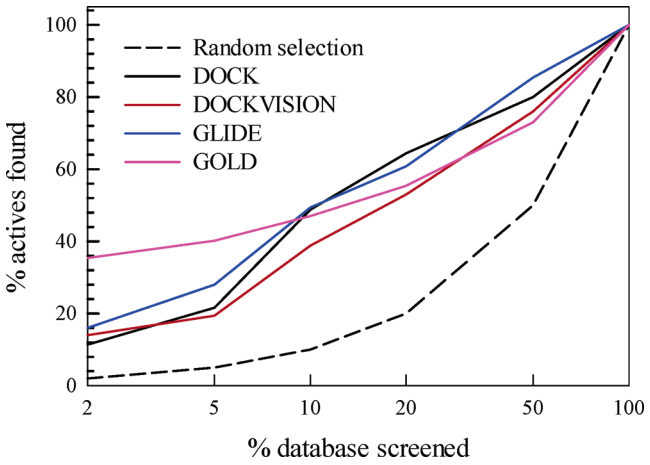

Figure 4. Summary of virtual screening results. For each of the docking methods, the percent of active seeds found was summed over the five target proteins and then divided by five.

From our perspective, the result of this relatively simple analysis is both encouraging and disheartening; it is encouraging that different docking methods all perform better than random selection, but at the same time disheartening to find that there is much room for improvement. The consistency of the performance of the different methods when enrichment is averaged over several targets suggests that these methods offer equally reasonable solutions to the docking problem when performance is defined as the ability to rank seeded active molecules well. Further analyses exploring the docking poses that give rise to the observed prioritizations and the performance for individual targets clearly distinguish the different docking methods.

The summary plot (Figure 4) obscures a key result that is more readily distinguished in the program-based plots (Figure 5). One of our goals is to put in place a docking-based virtual screening tool that we can routinely apply to as many of the new targets as possible for which experimental high-throughput screening will be undertaken. Therefore, we are interested in the generality of the tool's performance with respect to different protein targets. At the 2% level (2% of top-scoring molecules), GLIDE outperforms the other programs in this respect, being the only method that identifies known active molecules for four of the five protein targets (Figure 5, Table 1). Both DOCKVISION and GLIDE achieve this level of success when 5% of the top-scoring molecules are considered (Figure 5, Table 1). At the 10 and 20% levels, GLIDE was the only program that identified one or more of the known active molecules for each of the five target proteins. At these levels, all of the other programs identified one or more active seeds for four of the five target proteins (Figure 5, Table 1).

The outstanding performance of GOLD at levels below 10% of the top-scoring molecules (Figure 4) was noted above. Considering all five target proteins at the 2% level, 18 of the 105 molecules selected by GOLD were active seeds. This seems a rather striking result,

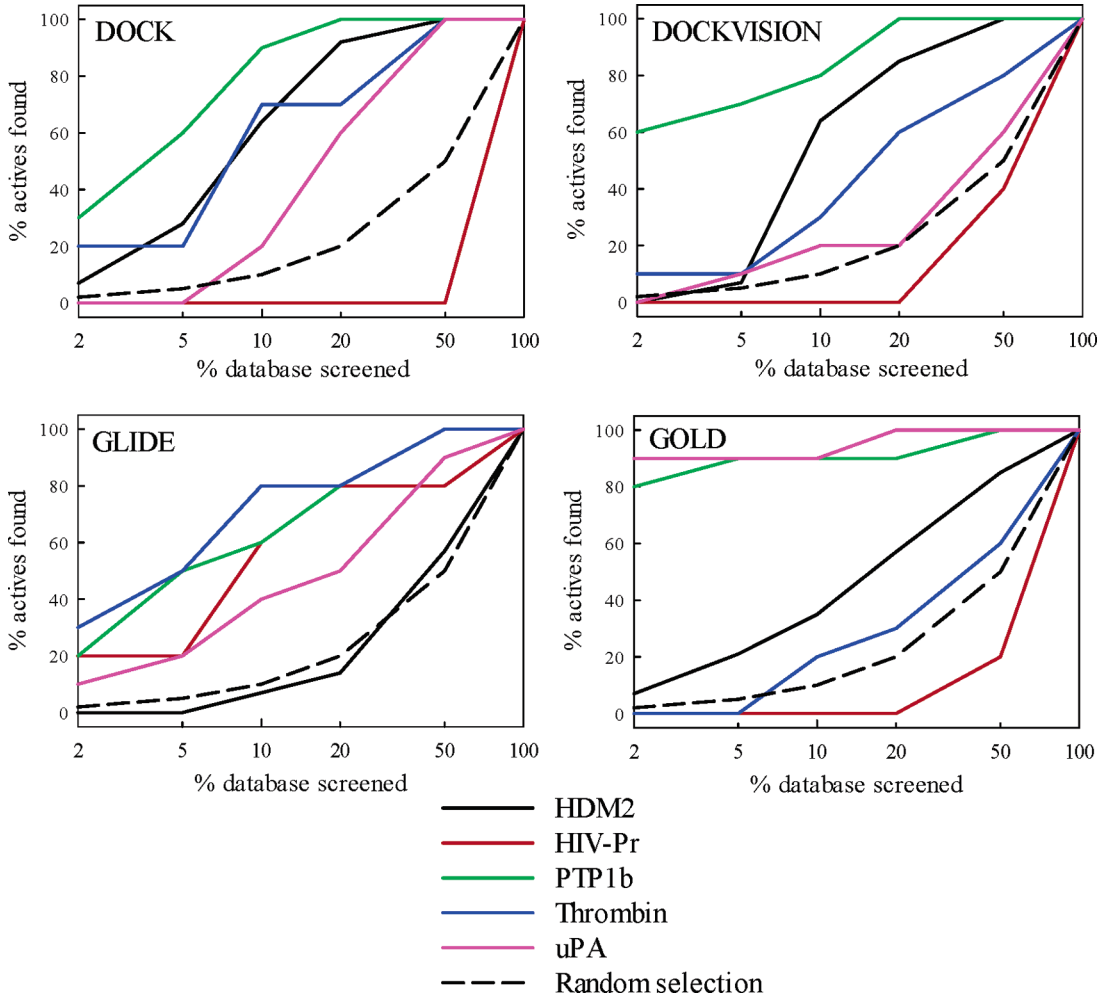

Figure 5. Program-based analysis of virtual screening results.

especially if one is restricted to experimental testing of relatively few molecules. The GOLD panel in Figure 5 shows that this is due to great success with PTP1b and uPA. At the 2% level, 17 of the possible 20 active seed molecules for these two targets were selected by GOLD. However, at lower database fractions, the performance of GOLD on the other three targets is quite poor, with two of the targets no better than random at fractions below 5% (Figure 5).

Comparison of the results obtained with each of the methods indicates that success is to some extent dependent on the target protein under study. The above-mentioned success of GOLD with both PTP1b and uPA provides an extreme example of this phenomenon. Figure 5 shows that across all of the methods PTP1b is consistently a high-performing target relative to the other proteins. Examining this same data organized by target (Figure 6) reinforces this point.

We have used the PTP1b actives as a test set for various perspectives on the virtual screening results we obtained. In the case of PTP1b, it is of interest to consider whether the good results observed were simply the result of prioritizing the compounds with large formal negative charge rather than being based on meaningful docking results. In our database of 1049 molecules, 25 molecules have formal charge, molecular weight, calculated log *P*, number of hydrogen bond donors, and number of hydrogen bond acceptors in the same range as the active PTP1b inhibitors. The ten PTP1b inhibitors all have formal charges of −2 or −3. This is a small number of decoys, but all methods rank at least some of the true inhibitors near the top of this abbreviated list. For DOCK, three of the top four are true actives; for DOCKVISION, six of the top eight are true actives; for GLIDE, five of the top six are true actives; for GOLD, nine of the top ten are true actives. This performance may be argued to be better than that for the full database. It seems clear that even the poorer performing methods are doing more than simply recognizing the correct charge when ranking the PTP1b inhibitors.

Consensus Scoring. Consensus scoring in the context of docking-based virtual screening has been shown to offer improved hit enrichment and decreased false positives by several groups.[24,25,27,58] In many cases, the improvements reported have been quite dramatic, making a compelling case for this approach. Here, we have taken an approach that might be termed "consensus docking"—we took intersections of the top 10% lists for each docking method. We have not rescored any of our original docking results. The parameters of interest include the total number of actives (true positives) found, the total number of negatives (false positives) found, and the number of targets for which hits were identified. Table 2 shows that while all of the individual methods identified active molecules for four or five of the targets at the 10% level, the consensus lists identified actives for three or four of the targets. Table 3

Table 1. Virtual Screening Results[a]

target	DOCK	DOCKVISION	GLIDE	GOLD
		2% Screened		
HDM2	1/14	0/14	0/14	1/14
HIV-Pr	0/5	0/5	1/5	0/5
PTP1B	3/10	6/10	2/10	8/10
thrombin	2/10	1/10	3/10	0/10
uPA	0/10	0/10	1/10	9/10
		5% Screened		
HDM2	4/14	1/14	0/14	3/14
HIV-Pr	0/5	0/5	1/5	0/5
PTP1B	6/10	7/10	5/10	9/10
thrombin	2/10	1/10	5/10	0/10
uPA	0/10	1/10	2/10	9/10
		10% Screened		
HDM2	9/14	9/14	1/14	5/14
HIV-Pr	0/5	0/5	3/5	0/5
PTP1B	9/10	8/10	6/10	9/10
thrombin	7/10	3/10	8/10	2/10
uPA	2/10	2/10	4/10	9/10
		20% Screened		
HDM2	13/14	12/14	2/14	8/14
HIV-Pr	0/5	0/5	4/5	0/5
PTP1B	10/10	10/10	8/10	9/10
thrombin	7/10	6/10	8/10	3/10
uPA	6/10	2/10	5/10	10/10
		50% Screened		
HDM2	14/14	14/14	8/14	12/14
HIV-Pr	0/5	2/5	4/5	1/5
PTP1B	10/10	10/10	10/10	10/10
thrombin	7/10	8/10	10/10	6/10
uPA	6/10	6/10	9/10	10/10

[a] In the fractions shown, the numerator represents the number of active seed molecules found, and the denominator represents the total number of active seed molecules for the specified target protein.

Table 2. Consensus Scoring with Pairs of Docking Programs[a]

method	DOCK	DOCKVISION	GLIDE	GOLD
DOCK	4	3	4	4
DOCKVISION		4	4	4
GLIDE			5	3
GOLD				4

[a] Intersections were taken for the relevant top 10% lists. The numbers shown represent the number of target proteins for which active molecules occurred in the consensus. Numbers on the diagonal represent the results for the 10% list from that method.

provides a detailed summary of the consensus analysis results. The complete consensus lists for each target/method pair combination range in size from 7 to 35 members, with the number of active molecules present ranging from zero to eight. The consensus docking analysis results in a large decrease in the number of molecules selected for screening with a small decrease in true positives. This approach could be helpful in a situation where screening resources were extremely limited.

Reproduction of Known Binding Modes. The docking programs compared in our study have been shown to reproduce and correctly rank known binding modes for at least some test systems when used as single-ligand-docking tools.[2−5,14,49,50,53,55,56,59−61] Relevant protein−ligand complexes were available for 31 of the 49 actives we used, with each target protein having multiple ligand-bound structures. This information allowed us to compare the structural basis of the prioritization observed for these 31 actives.

Table 4 shows the docking ranks for all seeded actives with all four docking methods as well as the corresponding rmsd's for those actives with known protein−ligand structures. In general, a strong correlation between docking rank and rmsd is not readily discerned. Docking against HDM2 provides one useful example. When 5−10% of the database is screened, DOCK, DOCKVISION, and GOLD have some success with this target (Tables 1 and 4). For these three methods and this target (14 actives), there are a total of 42 docking poses to consider, and of the 14 actives for this target four have reference crystal structures, thus giving a total of 12 reference poses for this example. Of the 42 docking poses for actives, more than 50% (23 of 42) of the actives are ranked in the top 10%. Similarly, for the 12 docking poses that have a reference structure, six are ranked in the top 10% (Table 4). However, the rmsd's for these six well-ranked dockings range from 5 to 8.1 Å. The HDM2 compounds are well-ranked, but the docking poses do not resemble the known structures. For this example, docking success does not appear to be based on reproduction of known binding modes. The results obtained with PTP1b provide another striking example. Table 1 shows that all four docking methods perform well against this target when relatively small fractions of the database are screened. In some cases, well-ranked dockings obtained with DOCK, DOCKVISION, and GLIDE had low rmsd's, but others had inarguably high rmsd's (Table 4). The rmsd/rank correlation observed for GOLD with the two targets on which it performs so well, PTP1b and uPA, is more encouraging (Table 4).

Table 5 presents a target-by-target analysis of the actives with known structures that were ranked in the top 10% for each docking method. For this subset of

Table 3. Consensus Scoring with Pairs of Docking Programs[a]

target	number of actives	DOCK/ DOCKVISION	DOCK/ GLIDE	DOCK/ GOLD	DOCKVISION/ GLIDE	DOCKVISION/ GOLD	GLIDE/ GOLD
HDM2	14	5/32	1/24	3/19	1/19	3/8	0/17
HIV-Pr	5	0/21	0/35	0/32	0/21	0/23	0/24
PTP1b	10	7/27	5/29	8/31	5/25	7/22	5/24
thrombin	10	2/18	5/31	2/23	2/16	1/17	1/23
uPA	10	0/25	1/25	2/19	1/24	1/7	4/12
molecules for testing		123	154	124	105	77	100
positives		14	12	15	9	12	10
negatives		109	142	109	96	65	90

[a] Intersections were taken for the relevant top 10% lists. All method pairs are shown here. In the fraction shown, the numerator represents the number of active seed molecules in the intersection, and the denominator represents the total number of molecules in the intersection. "Molecules for testing" is the sum of the denominators for that method pair against all five targets; "positives" is the sum of the numerators for that method pair against all five targets; "negatives" is the number of false positives ((molecules for testing)−(positives)).

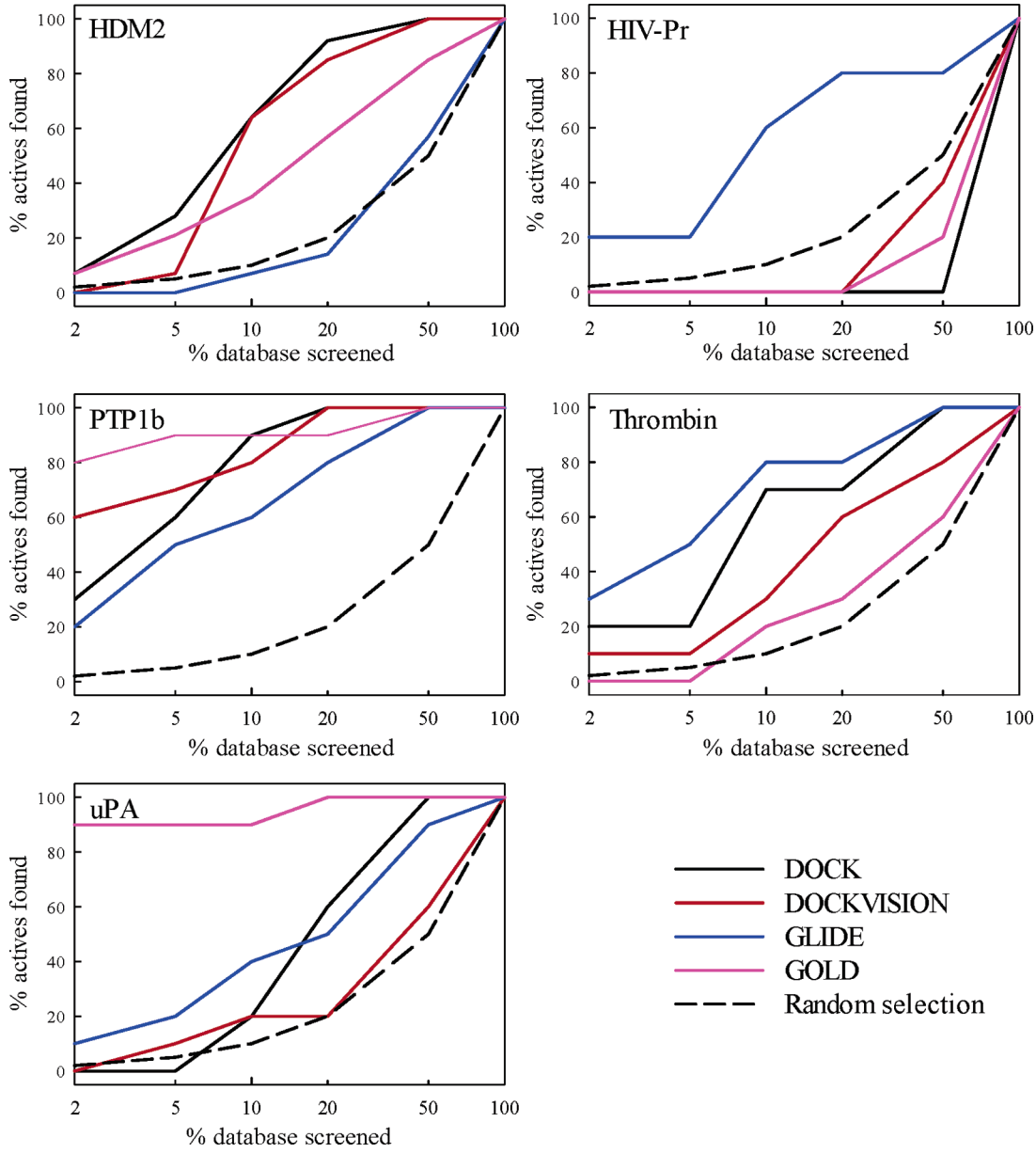

Figure 6. Target protein-based analysis of virtual screening results.

actives, it is clear that the basis of enrichment varies over the different programs and targets. All the docking programs have at least some success with PTP1b, with binding modes resembling the known crystal structures for at least half of the dockings identified in the top 10% (Table 5). Although all of the methods yield significant enrichment with one of either HIV-Pr or HDM2 (Figure 6), only that seen for the GLIDE/HIV-Pr pair is based on docking poses that resemble known binding modes (Table 5). Both GLIDE and GOLD identify binding modes that resemble the known structures for the majority of known actives that are ranked in the top 10%.

Discussion

Ligand Database. A large corporate screening collection is likely to be diverse with respect to the actives for any given protein target. The challenge of screening is to find the relevant trees in this forest. Verdonk and colleagues recently suggested[41] that tests of docking algorithms should use a database with a distribution of properties similar to the known actives. We agree that such tests are important for testing docking methods and in particular for addressing whether docking adds value beyond less time-consuming methods such as similarity or pharmacophore searching. In a typical screening situation, however, either few actives are known or one desires actives of a new chemical series in order to gain patentable compounds. In the first case, there may not be enough information to select a relevant subset based on physical properties or pharmacophoric features, and in the latter case such bias may exclude the novel actives that one is seeking. Our test, with a random set of decoys, the properties of which encompass those of the actives, and the use of a single database for all target proteins, is more representative of the docking screen that one would perform in advance of an experimental screen in pharmaceutical drug discovery.

Nevertheless, it is useful to examine the data and see if docking performs better or worse when comparing actives to a subset of the database having physical

Table 4. Docking Ranks and RMSD Values for Seeded Active Molecules in the Virtual Screening Database[a]

	DOCK		DOCKVISION		GLIDE		GOLD	
molecule	rmsd	rank	rmsd	rank	rmsd	rank	rmsd	rank
HDM00	6.2	77	8.1	78	6.3	511	3.5	201
HDM01	ns	241	ns	311	ns	778	ns	859
HDM02	ns	57	ns	65	ns	62	ns	378
HDM03	6.0	20	7.8	109	6.0	256	1.6	279
HDM04	5.0	44	8.9	140	5.7	407	6.1	128
HDM05	ns	125	ns	44	ns	374	ns	82
HDM06	ns	121	ns	66	ns	678	ns	34
HDM07	ns	91	ns	82	ns	623	ns	57
HDM08	ns	24	ns	143	ns	117	ns	36
HDM09	ns	145	ns	98	ns	281	ns	143
HDM10	ns	116	ns	76	ns	616	ns	741
HDM11	ns	49	ns	75	ns	710	ns	215
HDM12	ns	82	ns	246	ns	793	ns	19
HDM13	6.0	67	7.9	74	6.2	268	5.8	364
HIV14	7.5	1028	10.4	1015	7.0	2	7.0	961
HIV15	13.6	642	8.5	896	8.1	949	5.0	956
HIV16	6.2	931	8.9	465	3.0	103	4.7	560
HIV17	10.4	1044	8.9	687	2.4	113	5.5	227
HIV18	12.2	1041	9.1	289	3.0	100	5.3	950
PTP19	ns	4	ns	158	ns	420	ns	15
PTP20	ns	30	ns	128	ns	3	ns	2
PTP21	6.9	33	7.0	102	0.4	138	1.3	1
PTP22	6.1	76	1.1	11	1.2	91	1.6	10
PTP23	0.9	80	0.7	4	1.1	25	1.1	13
PTP24	10.8	24	10.0	6	3.0	1	7.5	471
PTP25	2.9	1	9.3	28	4.2	46	3.0	21
PTP26	11.5	110	1.3	19	1.9	36	1.8	17
PTP27	1.8	78	1.8	3	1.9	453	1.9	3
PTP28	ns	6	ns	18	ns	129	ns	32
THR29	6.0	377	9.9	318	4.2	37	6.3	956
THR30	8.6	75	7.1	167	0.8	16	7.6	134
THR31	ns	3	ns	555	ns	1	ns	837
THR32	3.4	102	7.1	193	2.3	79	nd	nd
THR33	9.3	300	9.2	64	1.9	60	nd	nd
THR34	8.6	16	9.8	14	1.4	32	4.2	468
THR35	7.1	63	9.6	164	0.8	5	1.1	65
THR36	8.0	78	5.2	60	7.8	319	7.8	90
THR37	8.8	57	5.2	480	8.0	220	2.9	446
THR38	6.5	242	7.6	896	1.7	67	2.9	224
UPA39	ns	112	ns	334	ns	363	ns	1
UPA40	4.0	367	7.4	639	5.7	60	2.8	7
UPA41	ns	170	ns	571	ns	47	ns	4
UPA42	6.7	78	5.8	276	6.6	97	1.1	2
UPA43	ns	80	ns	582	ns	162	ns	21
UPA44	ns	457	ns	561	ns	257	ns	10
UPA45	1.7	197	2.4	26	1.5	2	2.8	12
UPA46	14.6	431	7.6	74	4.9	780	1.7	130
UPA47	3.9	179	6.8	487	6.2	240	0.9	8
UPA48	3.6	281	6.3	320	6.2	225	1.0	17

[a] HDM00−HDM13, HDM2 actives; HIV14−HIV18, HIV-Pr actives; PTP19−PTP28, PTP1b actives; THR29−THR38, thrombin actives; UPA39−UPA48, uPA actives; ns, no structure available; nd, not docked because GOLD v1.2 did not recognize the oxyguanidine moiety (corrected in more recent versions of GOLD).

properties similar to those of the actives. We noted above that PTP1b consistently performed well in our comparison study, and we have used these actives to explore various conceivable extraneous (i.e., not docking related) causes for the observed results. When the database was reduced to a small subset of molecules with properties similar to those of the PTP1b actives, the PTP1b actives still ranked well when docked against PTP1b (Results). This suggests that prioritization of these molecules against this target was not based simply on 2D molecular descriptors. Cross-reactivity of PTP1b actives (prioritization of PTP1b actives when screened against non-PTP1b target proteins) was also examined to further explore the possibility of nonspecific or promiscuous selection of these molecules. It was recently reported[42] that a (proprietary and undefined) set of PTP1b actives were ranked well when docked against p38α (kinase) with GOLD, suggesting that these compounds may tend to promiscuity in docking-based virtual screening experiments. We examined our top 10% sets for all targets with all methods and found that with three of the four docking methods (DOCK, GLIDE, and GOLD) our PTP1b actives (Figure 2) were the least promiscuous of our five sets of actives (the PTP1b actives and two other sets were equally promiscuous with DOCKVISION; results not shown). Given the nature of the other binding sites used in this study, especially those of thrombin and uPA (Methods), this result may not be too surprising. Nonetheless, this analysis also supports the conclusion that the success observed with PTP1b makes sense in the context of docking and is not due to some simple and promiscuous selection of highly negatively charged molecules.

Analyses such as those outlined by Verdonk et al.[41] (selection biased by simple 2D descriptors of actives or decoys) and Vigers and Rizzi[42] (promiscuous selection of actives) are clearly important for the clarification of docking-based virtual screening results. In the present study, we have applied some of these criteria to an example dataset to show how such analyses can serve to validate virtual screening results.

Aspects of Docking Experiments. Binding site definition is one avenue for the incorporation of relevant experimental information (e.g., structural studies of related protein−ligand complexes, amino acid mutation studies, chemical modification, etc.) into the docking problem at hand. As noted above, we consider binding site definition to be a crucial component of a docking experiment. In general, the region of the protein accessible to the molecule during docking will dictate, to some extent, the molecules discovered or ranked well during the experiment. In structure-based drug design it can be useful to look beyond the limits of one or more bound ligands to probe new potential binding regions. This can

Table 5. Ranks and Goodness-of-Fit for Known Structures[a]

		DOCK			DOCKVISION			GLIDE			GOLD		
target	structures	top 10	≤2 Å	≤3 Å	top 10	≤2 Å	≤3 Å	top 10	≤2 Å	≤3 Å	top 10	≤2 Å	≤3 Å
HDM2	4	4	0	0	2	0	0	0	0	0	0	1	1
HIV-Pr	5	0	0	0	0	0	0	2	0	2	0	0	0
PTP1b	7	6	2	3	7	4	4	5	3	4	6	5	6
thrombin	9	6	0	0	3	0	0	7	5	6	2	1	1
uPA	6	1	0	0	2	0	1	3	1	1	5	3	5

[a] This table summarizes some of the data reported in Tables 1 and 4. Only dockings ranked in the top 10% for each target and each method are shown here. The structures column reports the number of known structures for this target, the top 10 column reports the number of known structures present in the top 10% of dockings for that target with that method, and the ≤2 and ≤3 Å columns report the number of known structure dockings ranked in the top 10% for which the rmsd meets these criteria, respectively.

lead to the discovery or design of new molecules or substituents that make productive binding interactions. For newly discovered and often poorly characterized proteins, relevant information may not be available, and it may therefore be prudent to define a relatively large binding site. Conversely, if the accessible space is very large, much of the search time in a docking experiment may be spent on searching unlikely binding modes. Binding-site definition is an important consideration for studies aimed at either validating a particular docking program or comparing results obtained with several different docking programs. An unbiased comparison of different docking methods requires careful consideration of binding site definition as implemented in each of the methods. We chose to address this aspect of the study by defining relatively large binding sites. Binding-site size was not systematically studied for each of the methods applied to each of the target proteins. However, in selected test cases, we established that changes in the size of the relatively large binding regions that we were using did not significantly alter the docking results we obtained, although docking times did increase with binding site expansion (results not shown).

In addition to simply restricting the total volume of the defined binding site, some of the methods we tested allow for the incorporation of additional information in the form of specific intermolecular constraints. For example, DOCK allows the user to mandate the presence of certain intermolecular interactions by specifying required atom types at specific positions or regions of the binding site,[14] and DOCKVISION provides for the definition of complex binding-site shapes by allowing the simultaneous use of multiple "include spheres", each of which defines a region of space accessible to the center of the molecule being docked.[50] Because these and other features were not implemented in all of the programs we compared, we chose to not use any of them. Full use of the feature sets of each of the programs might simply have resulted in comparison of the differences in the options available for each of the programs rather than the abilities common to all of the programs. To facilitate comparison on common ground, the docking methods we tested were used in a fairly naïve fashion. One might therefore argue that our approach to this aspect of the problem unrealistically burdens an already challenging problem. We were willing to pay this price toward our goal of running identical experiments with the different methods and, more importantly, to mimic the situation where little is known about a given target beyond its structure. This approach implies the caveat that for those methods that allow incorporation of additional information the results presented in the current study may represent a lower limit of anticipated success in other studies. Of course, we are not promoting the idea that available information should not be used when docking-based virtual screening is being applied to the discovery of new drug leads. On the contrary, it is our philosophy that if the goal is to discover new lead molecules, then a maximum of available information should be applied at this and other stages of the discovery process. However, many pharmaceutical drug discovery screens are currently run with little or no information about ligands that might translate to drug leads. Furthermore, imposing atom-based constraints on the docking experiment may limit the diversity of potential ligands evaluated to a subset of the total collection, which may be undesirable when maximal diversity of drug leads is sought. Finally, it is anticipated that in the future docking tools will increasingly be applied to relatively poorly characterized target structures (e.g., homology models, new proteins of unknown function), and in many of these cases, information relevant to docking constraints/restraints that could be imposed will be minimal. Understanding how the basic algorithms compare is critical to their appropriate application.

Enrichment and Binding Bodes. A presumption of docking-based virtual screening is that favored dockings represent reasonable binding modes. The inclusion in the screening database of small molecules with known binding modes allows one to check for the reproduction of those known complexes in the docking results. Rmsd of non-hydrogen ligand atoms is routinely used to measure the difference between the docked and observed binding modes. Ideally, we would like to find that the known actives are ranked best in the docking output and that the complexes of the known actives are accurately reproduced—the relevant docking having a low rmsd when compared to the known complex. Our results clearly establish the importance of investigating the structural basis of the enrichment observed in docking-based virtual screening experiments. All of the docking methods evaluated here ranked highly one or more of the known actives for one or more of the targets on the basis of a docking pose that did not resemble the known binding mode. Given that our goal is structurally rational docking-based virtual screening results, simple comparison of enrichment levels attained with different methods may therefore be misleading when this problem occurs. Some consideration of the "sense" or plausibility of the docking poses is a crucial component of an evaluation or comparison of one or more docking-based virtual screening methods.

Simply checking rmsd values may also be misleading. Although a small rmsd unequivocally indicates similarity to a known binding mode, a large rmsd may not be indicative of a completely unreasonable docking. For development and validation of docking methods, it may be satisfactory to focus (almost) exclusively on minimizing rmsd and strive for success in all cases. For drug discovery purposes, however, it is worthwhile to keep in mind that docking-based virtual screening is at present an inexact science.[25] Useful information can be gleaned from docking studies that fail to reproduce or predict observed binding modes. This is not a new observation. In 1991, Shoichet and Kuntz[62] noted that in protein−protein docking studies with DOCK many false positives represented what seemed to be, upon visual inspection, reasonable protein−protein configurations and that discrimination was challenging with various score functions. A later study from these and other workers[16] involving virtual screening of small molecules in a search for thymidylate synthase inhibitors began from a docked structure that differed significantly from the later-solved crystal structure. The information derived from the initial result was expanded into a broader virtual screen involving similarity searching and DOCK-based virtual screening and ulti-

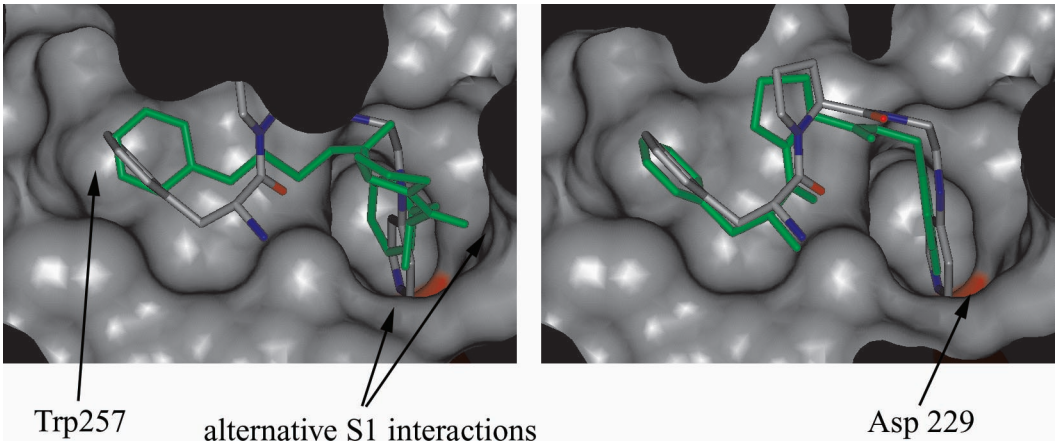

Figure 7. Docking poses for two similar thrombin inhibitors. The molecule in color-by-atom is the reference crystal structure for THR36 bound to thrombin (superimposed into the binding site of the docking target) in both panels. The purple molecule in the left panel is the GLIDE docking of THR36, whereas that in the right panel is the GLIDE docking for THR35 (see also Figure 2 and Table 4). The indole of Trp257 forms the back of the aryl binding pocket, and Asp 229 (carboxylate surface in red) forms the characteristic S1 interaction; Gly260 and Glu232 contribute to alternative hydrogen bonding interactions in the S1 region for the GLIDE docking of THR36 (see text).

mately led to the discovery of a more active compound for which the docking more closely resembled the observed crystal structure. More recently, Stouten and colleagues have reported the systematic application of a scoring scheme for docking based upon counting of specific intermolecular interactions.[63] Such methods of docking pose evaluation will be of particular use in the many screening cases where a structure of a relevant ligand−protein complex is not available.

Our analysis of rank and rmsd (Tables 4 and 5) for the structurally characterized known actives emphasizes the importance of evaluating rmsd's in addition to ranks or enrichment. The results show that in some cases known actives are ranked well when their docking pose does not resemble the known structure. Clearly, this is not ideal. Focusing on rank and rmsd for the top 10% of the dockings is enlightening. Consistent with discussion above, the PTP1b results are most encouraging. All methods select several actives, and in all cases at least half of the prioritized actives resemble the known structures (Table 5). This establishes that success can be achieved with all of the methods. Table 5 also shows that at this fraction of the database both the hit rates and the structural basis for the observed selections vary significantly across methods and targets (see also Table 1, Figures 4−6). With DOCK and DOCKVISION, the docking poses of most of the prioritized actives with known structures do not resemble the known structures when all five targets are considered. The results with GLIDE and GOLD are more promising (Table 5). For these methods, the majority of prioritized actives with known structures are based on docking poses that resemble the known structure (Table 5). GLIDE achieves this with four of the five targets, whereas GOLD maintains this level of performance with three of the targets. It is important to note that even in the best cases, however, the enrichment observed exceeds that expected strictly on the basis of reproduction of known structures (Table 5).

These results encourage more detailed examination of binding modes, and we provide a relevant example here. GLIDE performs relatively well with thrombin (Tables 1, 4, and 5). The left panel of Figure 7 compares the observed and GLIDE-docked structures of ligand THR36. For this pair, the rmsd is 7.8 Å due to complete reversal of the docking relative to the observed structure; the moiety that occupies S1 in the crystal structure occupies the aryl-binding pocket[40] in the docking and vice versa. On the basis of rmsd, this is a very poor docking, by any reasonable criterion. However, although the key interaction with the S1 carboxylate, characteristic of trypsin-like serine proteases, is not fulfilled, several other features of the docking make it reasonable (Figure 7). First, the two ligands occupy similar overall regions of the binding site. Second, the terminal aromatic moiety that occupies the aryl-binding pocket makes significant contact with the indole of Trp257. Finally, although the key interaction with Asp229 is not fulfilled for the docking, alternative hydrogen bonds in the S1 region are formed with Gly260···O and Glu232···N, and the aromatic ring that does occupy S1 is buried deeply in a slightly different region of the subsite than that of the ligand in the crystal structure. THR35 represents a fairly close analogue of THR36, and the right panel of Figure 7 clearly shows that the GLIDE docking of THR35 resembles the crystal structure of THR36 very closely. The rmsd for this docking with respect to its known structure (not shown, but the right panel of Figure 7 is an excellent surrogate) is 0.8 Å (Table 4). In this case, the docking reproduces all of the key interactions observed in the crystal structure, and the docking is ranked well (Table 4). The screening results are very different for these two similar compounds, with respect to reproducing the known binding modes. If only THR36 was present in the virtual screen, then the chemical class would likely be missed in the virtual screen. However, in cases such as this, tabulation of key interactions with the target protein may provide an alternative or additional scoring function and will certainly be useful in providing additional information regarding docked molecules.[2,53,59,60,62] Examples where the rank and rmsd are poorly correlated provide

instructive test sets for evaluation of rescoring schemes, allowing for detailed analysis of the effects of rescoring with different score functions.

Consensus Docking. Much work has been reported recently describing the rescoring of docking results with various score functions.[24,25,27,58,64] In the absence of a definitive score function, consensus scoring has been shown to be a powerful tool for postprocessing of docking-based virtual screening lists, yielding enhanced enrichment and reduction of false positives.[25] It is not clear, however, what the relationship is between the performance of a score function in a rescoring (or consensus scoring) context and that of the same function as the primary scoring function employed during a docking search with a particular docking search tool. The initial report describing consensus scoring[25] compared results obtained with two different primary docking/scoring tools, and the results thus obtained were rescored. Although this limited set of results was consistent with the utility of consensus scoring being independent of the primary score function used during docking, the authors noted that such a conclusion was premature and that a wider study was warranted. A study of rescoring schemes was not an objective in the present docking comparison. However, the current data set provides the basis for an expanded study of rescoring of results obtained with different docking tools, and these studies will be described in due course. In the present study, we adapted the consensus scoring approach to perform a consensus analysis on the results obtained with four different docking methods.

In contrast to previous reports describing *consensus scoring* where improvements have been significant, in the present study the benefit of *consensus docking* for the methods and targets we employed is marginal. Table 2 shows that four of the six consensus lists yield hits for four of the five target proteins. For these pairs, the total number of compounds for testing ranges from 77 to 154 (Table 3). This is a small improvement over the 2% level for single methods (cf. Table 1; at the 2% level 105 compounds are required for testing of the five target proteins). Overall, we observe slight improvement in the total number of actives for a given number of compounds selected (Tables 1 and 3).

Conclusion

Comparison of docking-based virtual screening methods is a challenging problem, and there are limitations to the present tests. Although we did study several different target proteins, we have not systematically explored chemical diversity within the active molecules for any of the targets. Ligand diversity in the context of docking-based virtual screening is an important area that remains to be studied, and Verdonk et al.'s recent investigation of databases for validation studies[41] is an important first step in this area. Our experimental design intentionally avoided exploitation of many of the features available in at least some of the docking programs, because we strove to run comparable experiments with the different docking tools. We expect that, at least in some cases, these omissions, as well as the large binding sites that we intentionally defined, may have limited the success we achieved with a particular docking tool.

The goal of screening is to identify molecules for further study as drug leads. Although a reasonable success in screening is to detect one or more potential lead molecules in a collection of molecules, the ideal would seem to be the identification of all the molecules of potential interest. Because high-throughput screening is relatively error prone (when compared to low-throughput testing of small numbers of molecules), we expect that some will always be missed. Nevertheless, it seems reasonable to continue to strive for the detection of all the molecules of interest, in both experimental high-throughput screening and docking-based virtual screening. A docking method would ideally be applicable as a virtual screening tool for many different target proteins and with relatively large collections of small molecules that may contain multiple lead molecules of different chemotypes for some of the target proteins studied. Therefore, it is of interest to ascertain the ability of a given method to detect multiple structural classes of molecules of interest for a single target protein. This parameter is arguably as important as the ability to detect lead molecules for different target proteins (Results). Past studies have involved relatively large sets of active molecules, but in these cases, as in our own, systematic study of structural variation among the actives has not been described.[25,27] To further explore the generality of docking methods as virtual screening tools, it will be useful to run similar virtual screening tests using targets where larger sets of structurally diverse active molecules are known and where this aspect of the test system is rigorously investigated.

Ideally, we would like to see all of the actives ranked best in the docking output. The present study indicates that the results obtained with these docking tools and these docking targets are not ideal. Given this, what should we hope for? For a screening tool that will be applied to a variety of protein targets, some level of enrichment with as many targets as possible is a compelling criterion. In our studies, the docking program GLIDE gave the most consistent level of success across the targets (Figure 5). However, DOCK, DOCKVISION, and GOLD achieved greater success with some targets. In some cases, it may be possible to select a specific docking tool for a target protein when the relevant information about the target protein is available. Prioritization of molecules based on docking should be due to evaluation of reasonable binding modes, and when one or more relevant test structures of protein−ligand complexes are available then binding mode reasonableness is defined as binding mode similarity. If a docking method does a good job of prioritizing known actives but the generated binding modes do not resemble known binding modes, then it is hard to understand the basis of success and failure. Such a method seems less desirable than one that bases prioritization on more satisfactory docking poses. In our studies, GLIDE and GOLD produce and rank well a greater number of reasonable binding modes for the actives with known protein−ligand structures (Tables 4 and 5).

The present work shows that when we coarsely compare docking methods as virtual screening tools with several different target proteins, then overall the num-

ber of hits identified with different docking programs is similar. Finer analysis indicates that target-to-target variation of success levels and the ability to reproduce known binding modes of actives and to rank these well are both useful discriminators of docking programs. This is relevant to the general applicability of docking programs as screening tools.

Acknowledgment. The authors thank Dr. Roger Bone and Dr. Ray Salemme for their support of this work, and we also thank the developers of the programs evaluated here for their constructive input during the course of our study.

Supporting Information Available: Input and parameter files for each of the docking programs and the database of three-dimensional coordinates for all of the decoys and the actives for three of the target proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Wodak, S. J.; Janin, J. Computer analysis of protein−protein interaction. *J. Mol. Biol.* **1978**, *124*, 323−342.
(2) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(3) DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **1986**, *29*, 2149−2153.
(4) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D. et al. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* **1988**, *31*, 722−729.
(5) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.
(6) Good, A. C.; Ewing, T. J.; Gschwend, D. A.; Kuntz, I. D. New molecular shape descriptors: Application in database screening. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 1−12.
(7) Grootenhuis, P. D.; Kollman, P. A.; Seibel, G. L.; DesJarlais, R. L.; Kuntz, I. D. Computerized selection of potential DNA binding compounds. *Anticancer Drug Des.* **1990**, *5*, 237−242.
(8) Gschwend, D. A.; Kuntz, I. D. Orientational sampling and rigid-body minimization in molecular docking revisited: On-the-fly optimization and degeneracy removal. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 123−132.
(9) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424−440.
(10) Lamb, M. L.; Burdick, K. W.; Toba, S.; Young, M. M.; Skillman, A. G. et al. Design, docking, and evaluation of multiple libraries against multiple targets. *Proteins* **2001**, *42*, 296−318.
(11) Lewis, R. A.; Roe, D. C.; Huang, C.; Ferrin, T. E.; Langridge, R. et al. Automated site-directed drug design using molecular lattices. *J. Mol. Graph.* **1992**, *10*, 66−78, 106.
(12) Makino, S.; Ewing, T. J.; Kuntz, I. D. DREAM++: Flexible docking program for virtual combinatorial libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 513−532.
(13) Meng, E. C.; Gschwend, D. A.; Blaney, J. M.; Kuntz, I. D. Orientational sampling and rigid-body minimization in molecular docking. *Proteins* **1993**, *17*, 266−278.
(14) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng.* **1993**, *6*, 723−732.
(15) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins* **1999**, *34*, 4−16.
(16) 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*, 1445−1450.
(17) Strynadka, N. C.; Eisenstein, M.; Katchalski-Katzir, E.; Shoichet, B. K.; Kuntz, I. D. et al. Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. *Nat. Struct. Biol.* **1996**, *3*, 233−239.
(18) Sun, Y.; Ewing, T. J.; Skillman, A. G.; Kuntz, I. D. CombiDOCK: Structure-based combinatorial docking and library design. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 597−604.
(19) Wang, J.; Kollman, P. A.; Kuntz, I. D. Flexible ligand docking: A multistep strategy approach. *Proteins* **1999**, *36*, 1−19.
(20) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439−446.
(21) Muegge, I.; Rarey, M. Small Molecule Docking and Scoring. *Reviews in Computational Chemistry*; Wiley-VCH: New York, 2001; pp 1−60.
(22) Abagyan, R.; Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375−382.
(23) Macarron, R.; Hertzberg, R. P. Design and Implementation of High Throughput Screening Assays. *High Throughput Screening Methods and Protocols*; Human Press, Inc.: Totawa, NJ, 2002; pp 1−29.
(24) 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.
(25) 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.
(26) Gschwend, D. A.; Good, A. C.; Kuntz, I. D. Molecular docking towards drug discovery. *J. Mol. Recognit.* **1996**, *9*, 175−186.
(27) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(28) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based vistual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9*, 47−57.
(29) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242.
(30) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235−249.
(31) Wlodawer, A.; Vondrasek, J. Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 249−284.
(32) Zhang, Z. Protein tyrosine phosphatases: prospects for therapeutics. *Curr. Opin. Chem. Biol.* **2001**, *5*, 416−423.
(33) Fenton, J.; Ofosu, F.; Moon, D.; Maraganore, J. Thrombin structure and function: Why thrombin is the primary target for anti-thrombotics. *Blood Coagulation Fibrinolysis* **1991**, *2*, 69−75.
(34) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N. et al. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(35) Tkachuk, V.; Stepanova, V.; Little, P. J.; Bobik, A. Regulation and role of urokinase plasminogen activator in vascular remodelling. *Clin. Exp. Pharmacol. Physiol.* **1996**, *23*, 759−765.
(36) Andreasen, P. A.; Kjoller, L.; Christensen, L.; Duffy, M. J. The urokinase-type plasminogen activator system in cancer metastasis: A review. *Int. J. Cancer* **1997**, *72*, 1−22.
(37) Nienaber, V. L.; Davidson, D.; Edalji, R.; Giranda, V. L.; Klinghofer, V. et al. Structure-directed discovery of potent nonpeptidic inhibitors of human urokinase that access a novel binding subsite. *Structure Fold. Des.* **2000**, *8*, 553−563.
(38) Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y. et al. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* **1994**, *263*, 380−384.
(39) Andersen, H. S.; Iversen, L. F.; Jeppesen, C. B.; Branner, S.; Norris, K. et al. 2-(oxalylamino)-benzoic acid is a general, competitive inhibitor of protein-tyrosine phosphatases. *J. Biol. Chem.* **2000**, *275*, 7101−7108.
(40) Bone, R.; Lu, T.; Illig, C. R.; Soll, R. M.; Spurlino, J. C. Structural analysis of thrombin complexed with potent inhibitors incorporating a phenyl group as a peptide mimetic and aminopyridines as guanidine substitutes. *J. Med. Chem.* **1998**, *41*, 2068−2075.
(41) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W. et al. Virtual screening using protein−ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.
(42) Vigers, G. P.; Rizzi, J. P. Multiple active site corrections for docking and virtual screening. *J. Med. Chem.* **2004**, *47*, 80−89.
(43) Compounds containing atoms other than H, Li, B, C, N, O, F, Na, Mg, S, Cl, K, Ca, Br, and I were excluded. Molecular weight was retricted to the range 200−600. Compounds containing α-halo carbonyls, isocyanates, sulfonylhalides, acid halides, aldehydes, anhydrides, polyethers, 4 or more fused phenyl rings, 2 or more nitro groups, or polyfluorinated compounds were excluded.
(44) MDDR, http://www.mdl.com/products/knowledge/drug_data_report/index.jsp.
(45) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3−25.
(46) Pearlman, R. Rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Auto. News* **1987**, *2*, 1−6.

(47) Sadowski, J.; Rudolph, C.; Gasteiger, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.
(48) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial informatics in the post-genomics era. *Nature Rev. Drug Disc.* **2002**, *1*, 337−346.
(49) Hart, T. N.; Read, R. J. A multiple-start Monte Carlo docking method. *Proteins* **1992**, *13*, 206−222.
(50) Hart, T. N.; Ness, S. R.; Read, R. J. Critical evaluation of the research docking program for the CASP2 challenge. *Proteins* **1997**, *29*(S1), 205−209.
(51) Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein−ligand interactions. *Proteins* **2002**, *47*, 521−533.
(52) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(53) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(54) Connolly, M. L. The molecular surface package. *J. Mol. Graph.* **1993**, *11*, 139−141.
(55) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L. et al. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(56) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J. et al. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(57) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(58) 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.
(59) Cummings, M. D.; Hart, T. N.; Read, R. J. Monte Carlo docking with ubiquitin. *Protein Sci.* **1995**, *4*, 885−899.
(60) Cummings, M. D.; Hart, T. N.; Read, R. J. Fragment-based modeling of NAD binding to the catalytic subunits of diphtheria and pertussis toxins. *Proteins* **1998**, *31*, 282−298.
(61) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C. et al. A new test set for validating predictions of protein−ligand interaction. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 457−471.
(62) Shoichet, B. K.; Kuntz, I. D. Protein docking and complementarity. *J. Mol. Biol.* **1991**, *221*, 327−346.
(63) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y. et al. Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871−881.
(64) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

JM049798D