Comparison of Shape-Matching and Docking as Virtual Screening Tools

Paul C. D. Hawkins,* A. Geoffrey Skillman, and Anthony Nicholls

OpenEye Scientific Software, 3600 Cerrillos Road, Suite 1107, Santa Fe, New Mexico 87507

Received March 22, 2006

Ligand docking is a widely used approach in virtual screening. In recent years a large number of publications have appeared in which docking tools are compared and evaluated for their effectiveness in virtual screening against a wide variety of protein targets. These studies have shown that the effectiveness of docking in virtual screening is highly variable due to a large number of possible confounding factors. Another class of method that has shown promise in virtual screening is the shape-based, ligand-centric approach. Several direct comparisons of docking with the shape-based tool ROCS have been conducted using data sets from some of these recent docking publications. The results show that a shape-based, ligand-centric approach is more consistent than, and often superior to, the protein-centric approach taken by docking.

Introduction

In recent years virtual screening (VS) has become an important part of the armamentarium of modern drug discovery. Much of the drive to use virtual screening has arisen from increased pressure to put compounds into the development pipeline and to reduce the costs of getting suitable compounds to this point. Given the well-known costs involved in experimental high-throughput screening (HTS),[1] virtual screening has been increasingly applied to reduce the number of compounds going into experimental HTS. For the purposes of this discussion we define virtual screening as ranking molecules by descending order of likelihood of relevant biological activity, regardless of how that ranking is performed.[2] Given that the time and costs associated with HTS can be reduced by correctly applied virtual screening, much effort has been expended in identifying VS approaches that assign low ranks to most of the inactive compounds while assigning high ranks to most or all of the active compounds.

Structure-based virtual screening approaches fall into two main classes: those based on protein coordinates and those based on ligand coordinates. When a protein−ligand cocrystal is available in an industrial project setting, virtual screening using docking has become the method of choice.[3] This may reflect the preponderance of recent published work on virtual screening via docking and scoring. In addition, there are now several "distributed processing" initiatives, i.e., virtual screening projects using spare cycles on personal computers, based on docking (FightAIDS@Home, ScreenSaver/LifeSaver, D2OL). Some notable exceptions to this trend have focused on pharmacophore or QSAR-based approaches, often, though not exclusively, in systems where structural information on the protein−ligand complex is not available.[4−7] For recent reviews on structure-based virtual screening, see the articles by Lyne[8] and Jain.[9]

Docking can be divided into two parts: the correct positioning of the correct conformer of a ligand in the context of a binding site (posing) and its successful recognition/scoring by a scoring function (scoring). Such an approach is an attempt to simultaneously solve several difficult problems and often results in inconsistent performance.[10] The reasons for this are manifold and include (i) inadequate treatment of electrostatics, electronic polarization, aqueous desolvation, and ionic influences, (ii) lack of accounting for entropy changes in the protein and the ligand on binding, (iii) insufficient ligand conformer sampling, (iv) inadequate sampling and weighting of proton positions (tautomers, rotamers) and charge states (ionization) of both protein and ligand, (v) the assumption of a rigid protein, and (vi) the well-known deficiencies in the scoring functions used to rank the docked molecules.[10,11]

A few publications on structure-based virtual screening have been prospective[12] and have clearly demonstrated the strengths and the limitations of the technique. The work of Jenkins[12a] illustrates the importance of consensus methods in ranking compounds and the weaknesses of a single docking/scoring combination, as the performance of any one docking engine/scoring function was never found to be consistently optimal. The work of Forino,[12b] while providing disappointingly low hit rates, clearly shows that docking can be of use in identifying compounds that are selective for the target of interest over related proteins. In some cases prospective docking studies have been performed that have provided a direct comparison with experimental approaches,[13,14] while others have used docking in concert with ligand-based approaches.[1,15]

Most of the structure-based virtual screening papers, however, are retrospective evaluations or comparisons of two or more docking tools, using a small number of known binders to the target of interest (actives) placed into a database of compounds that are presumed to be inactive (decoys). The performance of the tools under examination is then quantitated by some metric related to the ranking of the actives versus the decoys. Since a number of these publications provide the data sets used in the evaluation as Supporting Information, the potential arises to compare directly the performance of these structure-based approaches to techniques based purely on the ligand. Accordingly we present data directly comparing the performance of structure-based methods and a shape-based, ligand-centric method (ROCS) on the same data sets, allowing the most direct and meaningful comparisons to be made. In this work we rank molecules on the basis of their similarity to a known active molecule in three-dimensional shape space, using atom-centered Gaussian functions to allow rapid maximization of molecular overlap.[16] A more extensive presentation on shape as a metric property for molecular comparison is contained in Haigh et al.[17] For a comprehensive listing of tools to align or overlay molecules based on a wide variety of metrics, see the work of

* To whom correspondence should be addressed. Phone: 505-473-7385, extension 65. Fax: 505-473-0833. E-mail: phawkins@eyesopen.com.

10.1021/jm0603365 CCC: $37.00
Published on Web 12/13/2006

Melani et al.[18] Lemmen and Lengauer[19] have extensively reviewed the use of ligand-based tools in virtual screening.

The comparisons in this paper focus on data provided in publications from Sanofi-Aventis,[20] Johnson & Johnson,[21] the Jain lab,[22] the Villoutreix lab,[23] and the Rognan lab.[24] In the work from Sanofi-Aventis docking was performed into homology models, and in the other publications docking was performed into experimental crystal structures. The studies from Sanofi-Aventis, Johnson & Johnson, the Villoutreix lab, and the Rognan lab have been replicated in their entirety with a shape-based approach, whereas in the case of the Jain data, only those cases in which at least 10 actives were provided for a given target are replicated. This is to ensure that the results have some statistical meaning. We note that there are operational difficulties in comparing applications for virtual screening in a statistically meaningful way, especially when there are only a handful of active compounds in the database being searched.[25]

The data presented herein were obtained using the same starting point as that for docking: a protein−ligand cocrystal. In docking, the ligand is extracted from the complex and discarded and then attention is focused entirely on the volume in the protein active site revealed by removal of the ligand. For shape-based similarity, however, it is the protein that is discarded and attention focused on identifying compounds that best match the volume and disposition of functional groups in the ligand. In this study the query ligand was used in its experimental conformation whenever that is available.

Methods

Data sets were obtained as Supporting Information to the publications[21,22] or directly from the authors.[20,23,24] Databases of conformers for the data sets were then created using OMEGA.[26] The ligand structures used as queries were extracted from experimental cocrystal structures (obtained from the PDB[27]) and processed using OEChem.[28] The ligand was then used in its crystallographic conformation as the query for ROCS.[29] In the work of Evers on G-protein-coupled receptors (GPCRs),[20] where no crystallographic conformations for the ligands are available, a single low-energy conformation for the query molecule was generated using OMEGA.

ROCS performs shape-based overlays of conformers of a candidate molecule to a query molecule in one or more conformations. The overlays can be performed very quickly based on a description of the molecules as atom-centered Gaussian functions. ROCS maximizes the rigid overlap of these Gaussian functions and thereby maximizes the shared volume between a query molecule and a single conformation of a database molecule.

In preliminary work the effects of various options available in ROCS were examined for their impact on the VS performance of ROCS. In default operation ROCS compares molecules based purely on their best shape overlap, quantitated by their shape Tanimoto. ROCS then ranks the database molecules based on their shape Tanimoto to the query molecule. It was quickly found that adding to the shape Tanimoto the score for the appropriate overlap of groups with like properties (donor, acceptor, hydrophobe, cation, anion, and ring), the so-called color score, and then ranking on this summed score improved virtual screening performance considerably. Donors and acceptors were defined according to Mills and Dean.[30] Cations and anions were defined according to an implicit pK_a model such that the same group (e.g., carboxyl group) had the same protonation state (e.g., ionized) regardless of the protonation state set in the input structure definition.

Given the marked improvement when using the color score, all our experiments were performed with ROCS in "color-optimization" mode. In this mode ROCS optimizes the molecular overlay to maximize both the shape overlap and the color overlap obtained by aligning groups with the same properties that are contained in the color force field file. This overlay is then subsequently scored using the sum of shape Tanimoto for the overlay and the color score (the so-called combo score). Customization or target-specific information can be incorporated by adding a term to the color force field file that rewards overlay of specific functional groups. For example, groups that might be required for tight binding to the protein in question might be given extra weight in the color file (e.g., amidines/guanidines for thrombin). For an application of this approach, see the results from the comparison with the work of Evers et al.[20] below.

There are a number of approaches to quantitating the success of a particular tool for virtual screening. The most often used, and simplest to calculate, is enrichment at a given percentage of the database screened. Enrichment (EF) is defined as

$$\mathrm{EF} = \frac{\mathrm{Hits}_{\mathrm{sampled}}{}^{x\%}}{N_{\mathrm{sampled}}{}^{x\%}} \frac{N_{\mathrm{total}}}{\mathrm{Hits}_{\mathrm{total}}} \quad (1)$$

where $\mathrm{Hits}_{\mathrm{sampled}}{}^{x\%}$ is the number of hits found at x% of the database screened, $N_{\mathrm{sampled}}{}^{x\%}$ is the number of compounds screened at x% of the database, $\mathrm{Hits}_{\mathrm{total}}$ is the number of actives in entire database, and N_{total} is the number of compounds in entire database. It can easily be seen that enrichment has a fixed maximum at any given percentage of the database screened. At 1%, the maximum is 100, at 2% the maximum is 50, and at 10% screened the maximum enrichment obtainable is 10.

Another technique that portrays success at fixed percentages through the database is the hit rate (HR), the percentage of the known hits that are contained in a set percentage of the database:[24]

$$\mathrm{HR} = \frac{\mathrm{Hits}_{\mathrm{sampled}}{}^{x\%}}{\mathrm{Hits}_{\mathrm{total}}} \times 100 \quad (2)$$

where $\mathrm{Hits}_{\mathrm{sampled}}{}^{x\%}$ and $\mathrm{Hits}_{\mathrm{total}}$ are defined as above.

These metrics both rely on cutoffs made at various points through the ranking and so can be sensitive to small changes in ranking. A variant of the enrichment statistic has been developed to avoid this sensitivity, the so-called robust initial enhancement (RIE) approach.[31] A measure that assesses virtual screening performance across the entire database and hence is not sensitive to small changes in ranking is the area under the receiver operator characteristic, or ROC. The ROC shows performance of a given tool when screening across the entire database is examined and not just at fixed, early points in the screen as enrichment, hit rate, and to some extent RIE do. The theoretically perfect performance of a virtual screening application gives the maximum area under a ROC curve (1.0), while random performance of a tool gives an area under the curve (AUC) of 0.5. AUC values of less than 0.5 imply a systematic ranking of decoys higher than the rankings of known actives. The ROC approach has been used extensively in the life sciences and social sciences arenas since its inception.[32,33] For a recent application of the ROC curve in virtual screening, see the work of Tribelleau.[34]

Results

In the following sections the performance of ROCS will be compared to the performance of a variety of docking tools. In

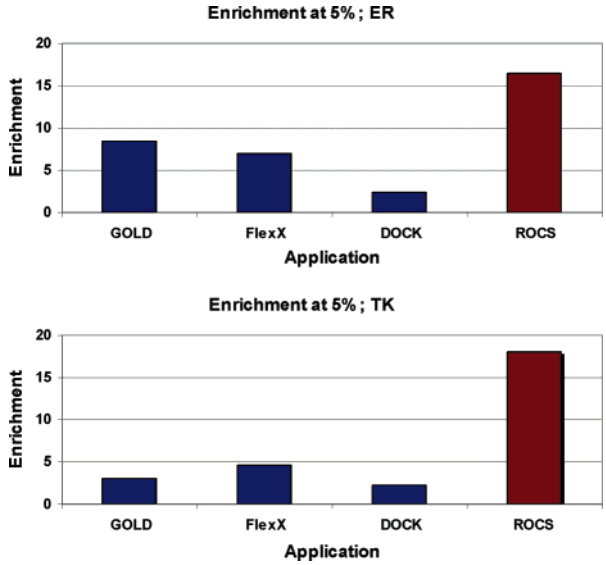

Figure 1. Hit rates for docking tools and ROCS against two targets, the estrogen receptor (ER) and thymidine kinase (TK). PDB codes for the docking targets are 3ERT (ER) and 1KIM (TK).

each experiment a shape and chemistry similarity based screen was performed with ROCS, based on similarity to a query molecule. In most cases this query molecule is the crystallographic ligand from the same cocomplex that was used in the docking experiments. In the case of the work on virtual screening against GPCRs, no crystal structure information is available, so the query molecule was selected in a different manner (vide infra). Tables showing the fingerprint similarities (Tanimoto coefficient based on MACCS keys) of the known active molecules being searched for to the query molecule can be found in the Supporting Information.

The Work of Bissantz. Bissantz et al. examined three well-known docking tools (FlexX, GOLD, and DOCK) as virtual screening tools against thymidine kinase (PDB code 1kim) and the estrogen receptor (PDB code 3ert).[24] The database being screened consisted of 990 decoy molecules chosen from the ACD, seeded with 10 TK actives and 10 ER actives. The study was initiated by identifying the appropriate scoring function(s) that most effectively discriminated the known actives from the decoys. With this knowledge in hand, the authors then performed the retrospective virtual screen and presented the results in terms of the enrichment and hit rate (vide supra) at 5% of the database screened. A comparison of enrichments between the docking tools and ROCS is illustrated in Figure 1. Note that the maximal enrichment that can be achieved at 5% of the database screened is 20. In both cases the hit rates from docking are obtained by a consensus score using two scoring functions that had been identified as effective against the target in question.

It should be noted that the performance of the docking tools is somewhat variable across the two targets. It has frequently been observed that the binding site of the estrogen receptor is highly hydrophobic, rigid, and sterically constrained, making it an easy target for docking. In contrast, the thymidine kinase binding site is hydrophilic with water molecules making bridging interactions between the ligand and the protein. There is also a flexible loop making up one side of the site, making it a challenging target for docking. Some or all of these differences between the sites may explain the reduced performance of GOLD and FlexX against TK (with GOLD showing the most significant decrease). ROCS, in contrast, shows consistently good performance on both targets when using the same ranking method in each case (the combo score).

The Work of Evers. Given the high importance of GPCRs as drug targets, where 9 of the top 20 best-selling prescription drugs in 2000 targeted GPCRs,[35] much attention has been focused on effective virtual screening methods for this target class. Docking has not seemed to be a promising tool for virtual screening against GPCRs, given the lack of any high-resolution structure of any human GPCR.

In an attempt to determine the effectiveness of building homology models of GPCRs and then performing docking, Evers et al. performed virtual screening by docking into a homology model of the α1-adrenoreceptor.[36] An extensive study was conducted to determine the scoring function that was most effective in discriminating actives from decoys after docking with GOLD. It was shown that a high proportion (5 of 9) of the scoring functions examined were unable to perform better than random selection at up to 10% of the database screened, while 2 of 9 were only little better than random at the same point. Both this study and the work of Bissantz[24] illustrate that a considerable amount of effort can be invested in identifying the appropriate scoring function(s) for a given combination of docking engine and target.

To test the performance of ROCS in GPCR virtual screening, we turned to related work from Evers that compared virtual screening performance against four biogenic amine binding GPCRs (α1A, 5HT2A, D2 and M1).[20] In this work several tools were utilized, including docking into homology models (with GOLD and FlexX-Pharm), pharmacophore tools (Catalyst), 3D similarity searching (FlexS), and a variety of 2D approaches. The virtual screening was conducted on a database consisting of 50 diverse active compounds and 950 diverse decoys, all selected from the MDDR.[37] Note that in the course of this work GOLD and FlexX-Pharm were used with constraints based both on the positions of protein atoms (side chains that are known to make contact with active ligands) and of ligand atoms (the protonatable nitrogen that is known to be essential for activity against these targets). To make the fairest comparison, we elected to try to mimic these types of constraints in ROCS. Lacking the homology models used in the paper, we could not utilize the protein-based positioning constraint. Rather, an additional term was added to the color force field file that rewarded placing protonatable nitrogens atop one another in the overlays, in an attempt to mimic the effect of the constraint in docking. Note, however, that the constraint in docking is an absolute constraint; failing to match it will result in a pose being rejected. In contrast, in ROCS an overlay not providing appropriate alignment of the protonatable nitrogens will not be rejected; it will simply not accrue the extra score. In the ROCS experiments the query molecule was the same molecule that was used in the FlexS portion of the study, and as mentioned above, a single conformation for these molecules was generated with OMEGA. Figure 2 shows comparisons of the performance of the 3D techniques in the paper with ROCS.

It can be seen that in two cases, 5HT2A and D2, ROCS with the amended color force field file performs very well compared to the other tools. For A1A, ROCS performs a little better than the other tools, while against M1, Catalyst or ROCS would be good choices.

For virtual screening against the A1A receptor, Evers provided data on the performance of GOLD with and without the protein-based and ligand-based constraints mentioned above. Unsurprisingly, GOLD's performance is significantly improved by the addition of these constraints (see Figure 3). The use of a color force field file in ROCS that rewards overlap of protonatable nitrogens (thereby mimicking one of the

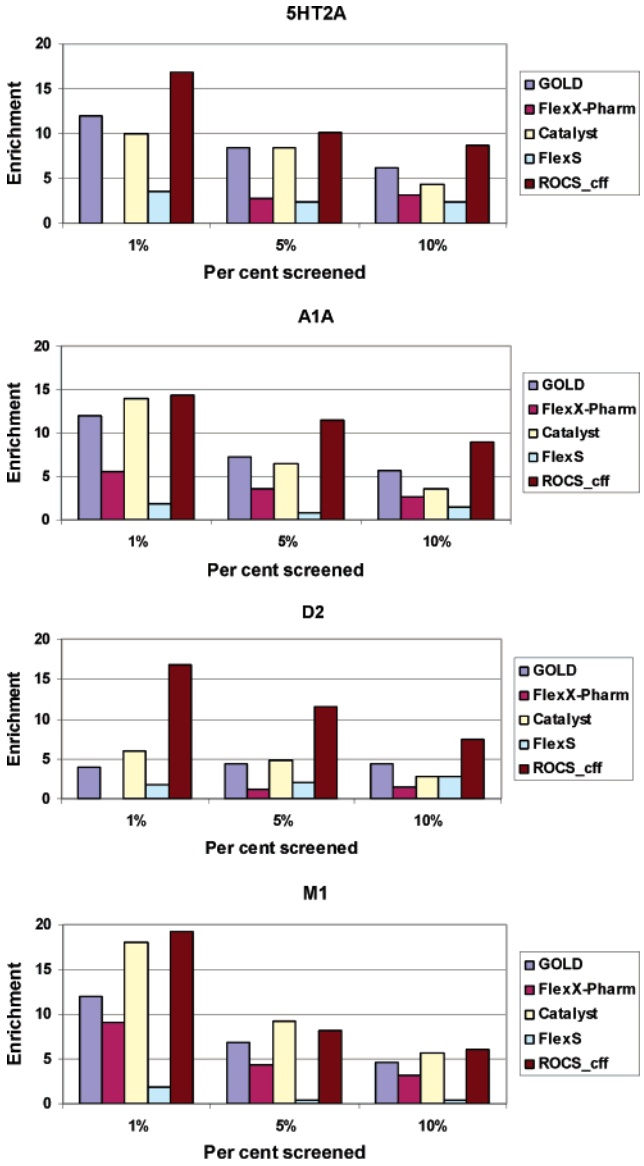

Figure 2. Performance of ROCS and other tools in virtual screening against four GPCRs. ROCS_cff denotes ROCS performance with the amended color force field file.

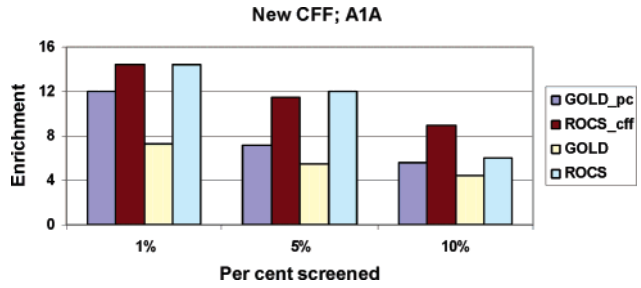

Figure 3. Comparison of GOLD and ROCS performance with and without constraints. GOLD_pc shows performance with constraints. GOLD shows GOLD default performance. ROCS_cff shows ROCS performance with an amended color file. ROCS shows ROCS default performance.

constraints used in GOLD and FlexX-Pharm) provides an improvement over the default file only at 10% of the database screened.

In a comparison of the data from Catalyst, it is worth noting that multiple active compounds (up to 20) were used to develop the pharmacophores used for searching rather than the single molecule used in the ROCS and FlexS portions of the study.

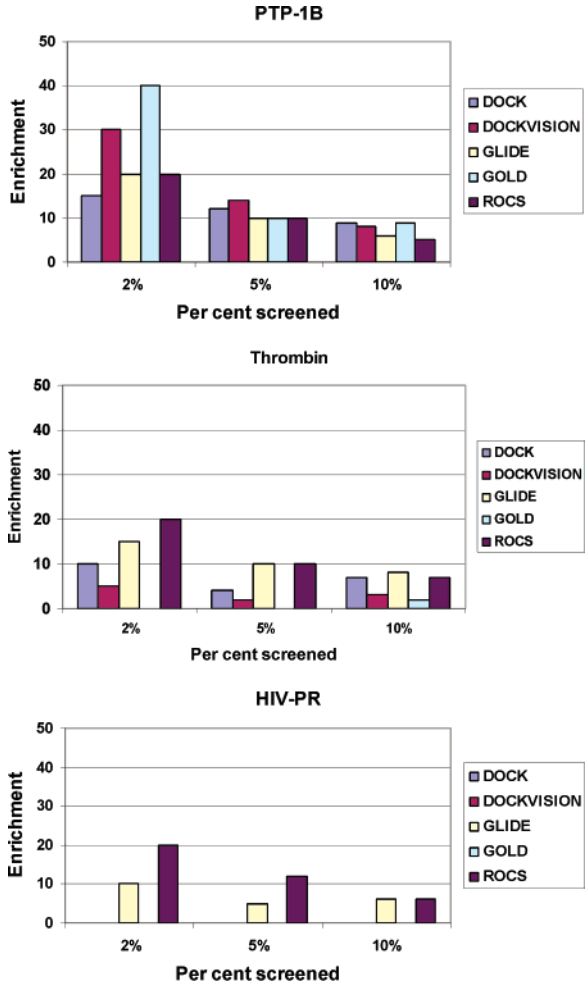

Figure 4. Comparison of enrichment at 2%, 5%, and 10% for docking tools and ROCS. The PDB codes for the crystal structures used in the docking are 1HVR (HIV-PR), 1C84 (PTP-1B), and 1QBV (thrombin).

On the basis of the much superior performance of Catalyst to FlexS, Evers et al. state, "Accordingly, the good performance [of 3D ligand-based methods] must be attributed to the fact that a wide range of structurally diverse active compounds for each target is available." However, given that overall ROCS outperforms Catalyst quite significantly in these examples, this conclusion must be questioned. ROCS has performed well when using only a single, low-energy conformation of a single molecule as a query.

The Work of Cummings. Cummings et al. examined four docking tools (DOCK, DOCKVISION, GLIDE, and GOLD) for use against three publicly available targets (HIV-1 protease, protein tyrosine phosphatase-1B, and thrombin).[21] In each case there are 10 active compounds placed into a background of 990 decoy compounds from the MDDR.[37]

The paper utilizes enrichment at three points through the database (2%, 5%, and 10%) as its performance metric. Figure 4 shows the comparison of ROCS to the docking tools investigated in the paper. Note that while some docking tools perform very well on certain targets (e.g., GOLD on PTB-1B, especially at the 2% point), there is no tool that is more consistent than ROCS, and only GLIDE shows performance of equivalent consistency. In the case of HIV-PR DOCK, DOCKVISION and GOLD were unable to identify a single active even in the top ranked 10% of the database, accordingly giving no enrichment at up to 10% of the database screened. The extreme difficulty that most of the docking tools have with HIV-1

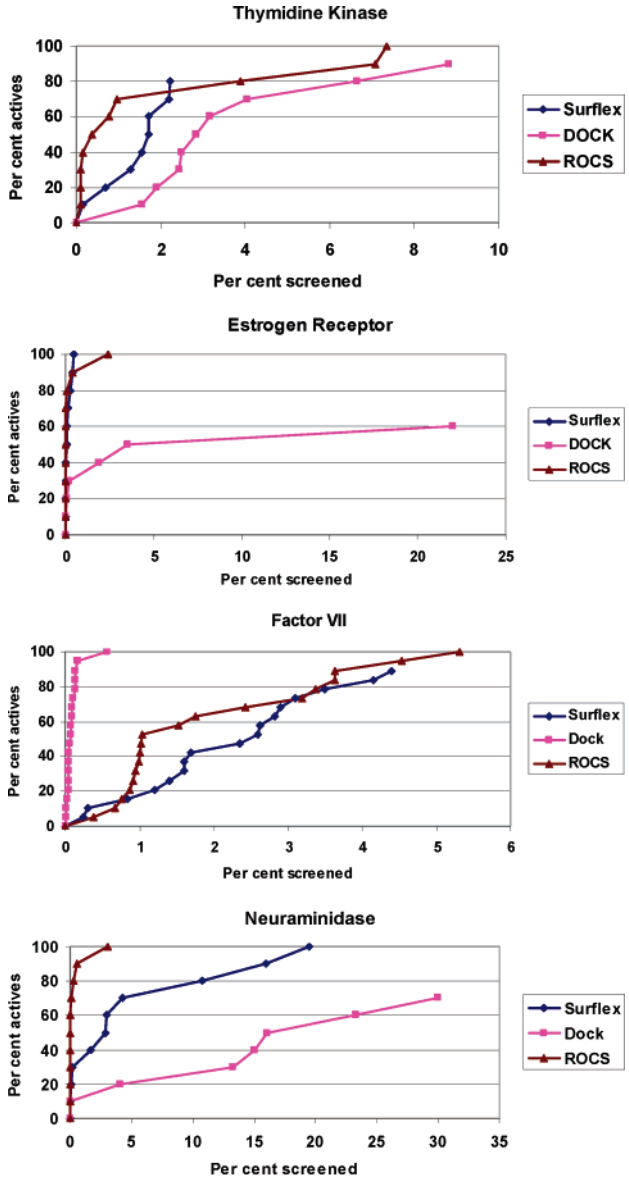

Figure 5. Plots for four targets comparing recovery of known actives with percentage of database screened. Targets and PDB codes are as follows; TK, thymidine kinase (1F4G); ER. estrogen receptor (3ERT); F7, factor VII (1DVA); NA, neuraminidase (1B9S).

protease most probably arises from the large size and high flexibility of many HIV-1 protease ligands.

In the cases illustrated so far the experiments have been on a relatively small scale (1000 molecules in total, between 10 and 50 actives). Now we discuss a larger scale experiment that was recently disclosed.

The Work of Miteva. In a large-scale virtual screening experiment Miteva et al. used a two-step procedure for docking, first using FRED[38] as a fast, shape-based filter to remove compounds that cannot fit into the target protein's binding site and then docking and ranking the remaining compounds with DOCK or Surflex. They examined four targets: the estrogen receptor, thymidine kinase, neuraminidase, and factor VII. The decoy compounds were selected by removing unsuitable compounds from the ACD, leaving 65 611 "druglike" compounds. The actives were taken from the PDB (10 each for ER, TK, and NA and 19 for factor VII), to give a total database size of 65 660.

A comparison of these docking approaches to ROCS is shown in Figure 5. The data are presented as a plot of percentage of the database screened versus percentage of the known actives identified. It should be noted that the sizes of the databases screened are different, as the database used for the ROCS study was not prefiltered by FRED, so that the entire database of over 65 000 compounds was screened. In the case of DOCK and Surflex the databases are considerably smaller (between 15 000 and 30 000 compounds depending on the target) due to the prefiltering performed by FRED.

In a note on timing in their paper, Miteva et al. observe that the average time per ligand when docking with Surflex is around 10 s, and with DOCK around 8 s, on a 1.5 GB RAM, 2.8 GHz Xeon processor. In this study performing ROCS overlays on the 65 660 compounds required an average over the four targets of 0.25 s per ligand. The average time to make conformers for this database was 2.1 s per molecule. These timings are for a 1 GB RAM, 1.0 GHz Pentium 3 processor. Clearly this protocol is significantly faster than the docking part of the approach documented by Miteva.

Note that in all cases ROCS identifies all the actives in only 8% of the database whereas the docking tools are unable to do this in all cases. Surflex fails to identify all the actives for TK and F7, while DOCK fails to identify all actives for TK, ER, and NA. Possibly some of the actives are too large to fit into the rigid active site representation used by these docking tools, and hence, they fail to give scores. DOCK can only identify all the actives in one case, factor VII, but in this case it does so with remarkable efficiency, ranking all the actives in the top 0.6% of the database. Overall in this comparison ROCS shows itself to be a more efficient tool than docking by identifying, on average, a larger proportion of the actives at a fixed proportion of the database.

Further examination of the ranked lists from ROCS allows us to determine what fraction of the database must be screened before at least one example of every chemotype of inhibitor is found. For neuraminidase this fraction is 0.014%, and for the estrogen receptor the fraction is 0.36%. The factor VII inhibitors lie essentially in one structural class, and the answer is obscured in the case of TK by compounds very similar or identical to the actives that are present in the decoy set. Some of the pitfalls in this study that were encountered due to the nature of the TK decoy compounds are elaborated in the Discussion.

Inspection of the 2D similarities of the query and active molecules (see Table 3 in the Supporting Information) shows that the success of the shape-based approach is not sensitive to the 2D similarities of the query and the active compounds. For example, ROCS performs very well against the estrogen receptor and neuraminidase, and yet the actives used show only moderate 2D similarity to the query. A more thorough investigation of the comparison between shape-based approaches and 2D approaches in QSAR and ranking is found in ref 39.

The Work of Jain. In the work of Pham and Jain[22] the performance of Surflex is illustrated using the area under the receiver operator characteristic curve (vide supra). In the original study 26 protein targets were investigated, while in this work, a subset of targets for which at least 10 active molecules were available was used (8 targets). Figure 6 shows a plot of the area under the ROC curve for these eight targets from Surflex-Dock and from ROCS

Figure 6 shows that the performance of Surflex-Dock and ROCS is usually quite similar. In the case of trypsin, Surflex-Dock is clearly the superior tool, while for thrombin and PARP, ROCS is a superior choice. In all other cases the difference is not significant. Possible reasons for the poor performance of ROCS against trypsin are outlined in the Discussion. The

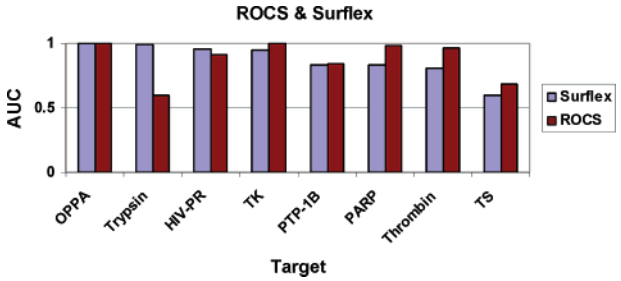

Figure 6. Comparison of the area under the ROC curve for eight targets using Surflex-Dock or ROCS: ROCS, result when using the crystallographic pose of the query. The PDB codes for the proteins are as follows: OPPA, 1B5J; trypsin, 1QBO; HIV protease, 1PRO; TK, 1KIM; PTP-1B, 1PTY; PARP, 2PAX; thrombin, 1C4V; TS, 1F4G.

approach to docking taken by Surflex (matching a ligand to a "proto-molecule" defined by the volume and features of the active site) is somewhat similar in conception to ROCS. It is therefore noteworthy that its results are so similar to those of ROCS.

Experimental Section

Conformer databases were generated with OMEGA 1.8.1,[26] using an energy window for acceptable conformers of 8 kcal/mol above the ground state and a rmsd cutoff of 0.8 Å (see Bostrom[40] for a discussion of the effect of parameters on OMEGA performance). The maximum number of rotatable bonds allowed in any molecule was 16 (except for the case of the OPPA data set from the Pham study where the maximum was set to 25), resulting in the loss of a very small fraction of compounds from some of the data sets. The conformer databases were searched with ROCS 2.1.1.[29] The query molecule for each ROCS run was in a single conformation, the conformation found in the protein−ligand X-ray structure in the PDB.[27] In the cases where no cocrystal structure was available, a single low-energy conformation was used as the query, generated with OMEGA (by setting the maxconfs parameter to 1). ROCS was run using a built-in color force field file (with the −chemff ImplicitMillsDean flag), and all overlays were optimized to maximize color overlap after the best shape overlay was located (using the −optchem flag). The hits were ranked on the basis of the sum of their shape Tanimoto and the normalized color score in this optimized overlay (using the −rankby combo flag). This sum is known as the combo score. The color force field file used in the GPCR studies was amended to add the following terms to reward overlap of protonatable nitrogens in the query molecule and candidate conformers:

```
DEFINE CATaventis [N;!$(*C=O);!$(a*(c)c);!$(*[+])]
TYPE gpcr
PATTERN gpcr [$CATaventis]
INTERACTION gpcr gpcr attractive gaussian weight=
                                        5.0 radius=1.0
```

Note that the standard weight of color interactions is 1.0, so the weight awarded to the overlap of these protonatable nitrogens is considerable. This is done to mimic the use of pharmacophoric protonatable nitrogens, where a molecule that cannot place the appropriate atom in the correct place is rejected by the docking program.

Discussion

Structure-based virtual screening is a well-known and widely applied technique in modern lead discovery, most commonly involving the procedure known as docking. Docking programs are typically used to predict one of three things, in order of decreasing difficulty: affinity, binding mode, and activity. There is general agreement[41] that docking programs cannot predict affinity to a degree that is useful. On the other hand, they have shown some utility in predicting binding modes and i activity, i.e., selection of active compounds from a larger set of inactives. In this study we have focused on the relative merits of docking and a shape-based, ligand-orientated method, ROCS, for the last of these functions, separating the binders from nonbinders. This study does not address the other two areas.

One criticism of such a comparison concerns explicit and implicit parametrization. Docking methods typically contain implicit parametrization via scoring functions. Because scoring functions are developed on systems typically not dissimilar to docking targets, there can be implicit parametrization toward the known. Many in the field do not see this as a problem; in fact, such "knowledge-based" potentials are very popular. Ligand methods are typically explicitly parametrized; i.e., a series of actives for a system are used to generate a query intended to locate and rank highly other actives. In this respect ligand-based methods fall midway between the pure 3D nature of docking and the connection table methods common in most QSAR. One of the advantages of the approach taken in this study is that the shape-based approach does not require multiple active compounds and, as such, requires no explicit parametrization. This work includes one exception, to favor protonatable nitrogens as ligands for amine binding GPCRs (vide infra) but only because in this case docking also included such an explicit parameter.

In order the minimize the possibility of local knowledge bias common in the field, we have extracted five data sets from the literature and replicated the docking experiment, using the combination of shape and chemical similarity as the ranking method. The results lead us to conclude that the shape-based approach can provide better performance than docking tools in more than half of the 21 systems examined. We also note that knowledge of the bioactive conformation of the query molecule is not necessary for the shape-based approach to give good performance.

The outcome of the experiments detailed above was unexpected. Shape similarity is not a profound technology, it merely aligns volumes and adds in a term for functional group similarity, and equal weight in the final score is given to both the shape and chemical similarity contributions. It is likely that extensive parametrization would show that assignment of different weights to the shape and color parts of the combo score could provide superior performance in virtual screening or other applications. However, this has not been done, and given the successes illustrated in this paper, it seems that the naive approach of assigning exactly equal weights to the shape and chemical similarity components of the combo score was justified. That such a straightforward approach should provide equivalent performance at the simplest task asked of docking (ranking) to more sophisticated methods with many years of investigation behind them requires comment. To fulfill its promise, docking needs to accurately predict protein−ligand interactions, something not yet possible. In place of these predictions are heuristic scoring functions that attempt to capture some of the essence of binding physics. That docking works at all is a triumph of such functions. However, scoring functions are also notorious in promoting false positives. It is not that such functions do not recognize active ligands and predict binding modes, but they cannot recognize inactive molecules. The necessity of including information on bad ligands, as well as good, in scoring function development is a major thrust of the work of the Pham and Jain.[22]

Shape-based approaches often suffer from the inverse issue, the problem of the false negative. The underlying assumption of shape-based methods is that compounds with shape and chemistry similar to those of a known active molecule have a high probability of also being active. Consequently active molecules with shapes different from that of the active used as a query could easily be missed. Accordingly, one could imagine docking finding radically different ligands, in size or shape that shape similarity is unable to capture, allowing the identification of novel ligands with unforeseen binding interactions. However the rigid protein assumption employed by all docking engines in this study (and other approximations mentioned above) often means that this promise is not fulfilled. We see an example of the size bias in shape-based approaches when examining trypsin in the Jain experiment, where Surflex-Dock gave much better performance than ROCS. Here, the ligand from the PDB structure 1QBO, used as the query, is relatively large, whereas a high proportion of the active ligands in this data set are quite small. The ROCS algorithm begins the shape overlay procedure by overlapping the centers of mass of the query and the database conformer and then aligning their principal moments of inertia. A consequence of this is that if two molecules of very different sizes, but possessing some of the same functional group(s), are aligned on the basis of the moments of inertia, then the functional groups will quite likely not be aligned at all, and the overlay solution may be trapped in a local minimum of the shape and color force hypersurface. Therefore, the combined shape and chemistry (color) score for the molecules will be low. However, such failures do not dominate the overall statistics and can be addressed by including alternative queries or using asymmetric measures of shape similarity rather than the symmetric Tanimoto measure used here.

This observation does lead to one of the confounding questions of ligand-based design: Which compound(s) (and in which conformations) should be used as the query, and how should they be chosen? In our studies the ROCS approach to shape similarity has proven to be extremely robust to ligand choice. Only one molecule was used as the query (unlike the approach routinely used in pharmacophore tools where multiple active molecules are required). The shape-based approach routinely crosses boundaries in chemical space, as shown in a follow-up to a recent study by Warren et al.[10] A comprehensive comparison was made on 10 commercial docking programs using almost 1300 ligands from 21 chemical classes against 8 protein targets. Each target had crystal structures of ligands from multiple chemical classes. The authors kindly compared ROCS performance for the same benchmark, using the crystal structure of a ligand from each chemical class as the ROCS query (Martha Head, personal communication). For the protein with the most chemical classes of ligands, PPARD-δ, with ligands from five chemical classes, ROCS was able to cross all chemical class boundaries from every starting point. We do not expect shape similarity methods to find ligands of significantly different shapes or sizes using a single ligand query, but it has shown remarkable ability to discover novel chemistries within a shape class.

The issue of the choice of conformation for a query molecule is even more difficult. We had assumed the utility of shape similarity methods derived from the ligand providing a "negative image" of the active site, into which we could fit new ligands. This would suggest the need for a bioactive conformation of the ligand. However, in the Evers[20] experiment there are no bioactive conformations available. In this experiment using the lowest energy conformer of an active molecule proved to be an effective method for selection of the conformation to be used as a query. As such, whether an experimental conformation (which is arbitrary) or a low-energy conformer (which is also arbitrary in a different way) is used as a query has little effect on ROCS's performance. We also have evidence (P. C. D. Hawkins, unpublished results) that replacing the crystallographic conformation of a query molecule with a low-energy conformer from OMEGA has essentially no impact on ROCS's performance.

We believe the consistency of ROCS and the lack of any special parametrization are closely related. One of the hallmarks of overparametrization is fragility of results. The simple, relatively parameter-free, shape similarity approaches may help to avoid such fragility. The shape overlap is one parameter, and the sum of functional group comparisons (color) is another. These are combined equally to produce a similarity score. We suggest that the simplicity of this approach underlies the generality of application. In fact, in the Evers experiment adding a special-purpose parameter (a pharmacophore feature) had little effect on ROCS, whereas it had a dramatic effect on GOLD performance.

It is also worth noting that the nature of the decoys will have a profound effect on the performance metrics for all the tools. In the Miteva et al. study there are over 250 thymidine, cytosine, uridine, and adenosine analogues in the decoy set, while 10 compounds in the same sets of series were designated as actives. Many of these 250 or so "decoys" are certain to be active to some degree as TK inhibitors and thus fail to meet the commonly held criterion for a background or decoy compound, the criterion being that it is inactive against the target of interest. In our hands ROCS scored many of these "decoy" compounds very highly, and we may only assume that the other tools used in this study did so as well. Therefore, the plots for TK from this study certainly represent an underestimate of performance for all of the tools on this data set.

Docking has been shown to perform best when conducted in a holo enzyme complex due to the manifold conformational changes that often occur to an apo structure upon ligand binding (see Erickson et al.[42]). Docking often suffers a reduction in performance when conducted on apo structures versus holo structures.[43] However, the shape-based approach outlined here performs well when using an experimental conformation obtained from a holo structure or, as in the GPCR study, an arbitrary low-energy conformation of the query molecule. These two studies indicate that docking may sometimes benefit from ligand information, which is implicit in the holo structure, more than ligand-based methods do.

One of the criticisms of the shape-based, ligand-centric approach has been that it completely ignores protein information, even when such information exists. One modification to the standard shape similarity approach would be to postfilter results based on the in situ alignment and interactions with the protein. In a study from Wyeth[44] on virtual screening and lead-hopping with ROCS, overlays were obtained for candidate molecules to the query molecule in its crystallographic configuration. Given that these overlays were in the context of the structure of the target protein, a force-field cleanup (to eliminate candidate compounds that gave good overlays but also clashed with the protein) and energy analysis were applied to the ROCS overlays. This two-layered approach was found to give good results, resulting in the identification of an entirely new class of active compound against the target of interest. Work is underway to determine if this finding is true generally. While this would be exciting, we suspect that this approach may equally likely fall

foul of the rigid protein assumption that so often bedevils docking studies.

Another issue frequently raised in ligand-based studies is whether the performance of a 3D ligand-based tool is mostly due to trivial 2D similarity between the query molecule(s) and the active molecules being searched for. As mentioned above, the mean and maximum 2D similarities between the query molecule used in each of the experiments and the active molecules are tabulated in the Supporting Information. Inspection of the similarities between the queries and the active molecules used clearly shows that the success of the shape-based approach presented here is not due to close structural similarity between the query and the active molecules. Preliminary investigation into the comparison between 2D fingerprint and shape-based similarity shows that they are pleasingly complementary (data not shown).

Conclusion

In sum, direct comparisons between virtual screening results from a significant number of docking programs show that a shape-based ranking method (ROCS) performs at least as well as and often better than docking. In total, seven different docking programs were compared to ROCS across 21 different protein systems (15 unique proteins). Since exactly the same sets of active compounds and decoys were used in this study as the published docking studies, the conclusion that a shape-based approach is competitive seems warranted. ROCS provided superior performance even when a bioactive conformation of the ligand was not known. Given the success, speed, ease of use, predictability, and applicability of ranking using shape and chemical similarity, we suggest that this approach be given serious consideration in all projects where high-throughput virtual screening is warranted.

Acknowledgment. The authors who have made the data sets from their publications freely available to allow direct comparisons such as this are warmly thanked. Particular thanks go to Prof. Villoutriex for assistance in obtaining the data used in his study and to Prof. Jain for the interest he took in this work in its earlier stages. Drs. Martha Head and Gregory Warren (GSK) are thanked for their assistance with replicating their docking study with ROCS. Dr. Robert Tolbert is thanked for his invaluable assistance on the intricacies of Python to one of the authors (P.C.D.H.), and Dr. Roger Sayle is thanked for his trenchant comments.

Supporting Information Available: Tables 1−5 listing the results of Bissantz, Evers, Cummings, Miteva, and Pham. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kraemer, O.; Hazemann, I.; Podjarny, A. D.; Klebe, G. Virtual screening for inhibitors of aldose reductase. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 814−822.
(2) Hert, J.; et al. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information *J. Med. Chem.* **2005**, *48*, 7049−7056.
(3) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore^{CSD}-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296−6306. Kontoyanni, M; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11−19.
(4) Ahlstrom, M. M.; Ridderstrom, M.; Luthman, K.; Zamora, I. Virtual screening and scaffold hopping based on GRID molecular interaction fields. *J. Chem. Inf. Model.* **2005**, *45*, 1313−1320.
(5) Krovat, E. M.; Fruwirth, K. H.; Langer, T. Pharmacophore identification, in silico screening and virtual library design for inhibitors of factor Xa. *J. Chem. Inf. Model.* **2005**, *45*, 146−154.
(6) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−558.
(7) Abrahamian, E.; et al. Efficient generation, storage and manipulation of fully flexible pharmacophore multiplets and their use in 3-D similarity searching. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 458−466.
(8) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7*, 1042−1047.
(9) Jain, A. N. Virtual screening in lead discovery and optimization. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 396−404.
(10) Warren, G. L.; et al. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(11) (a) Marsden, P. M.; Puvenendrampillai, D.; Mitchell, J. B. O. Predicting protein−ligand binding affinities: a low scoring game? *Org. Biomol. Chem.* **2004**, *2*, 231−237. (b) Maiorov, V; Sheridan, R. P. Enhanced virtual screening by combined use of two docking methods: Getting the most on a limited budget. *J. Chem. Inf. Model.* **2005**, *45*, 1017−1024.
(12) (a) Jenkins, J. L.; Kao, R. Y. T.; Shapiro, R. Virtual screening to enrich hit lists from high-throughput screening: A case study on small-molecule inhibitors of angiogenin. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 81. (b) Forino, M.; et al. Virtual docking approaches to protein kinase B inhibition. *J. Med. Chem.* **2005**, *48*, 2278−2289.
(13) Pavia, A. M.; et al. Inhibitors of dihydropicolinate reductase, a key enzyme of the diaminopimelate pathway of *Mycobacterium tuberculosis. Biochim. Biophys. Acta* **2001**, *1545*, 67−75.
(14) Doman, T. N.; et al. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213−2219.
(15) Ward, R. A.; Perkins, T. D. J.; Stafford, J. Structure-based virtual screening for low molecular weight chemical starting points for dipetidyl peptidase IV inhibitors. *J. Med. Chem.* **2005**, *48*, 6991−6999.
(16) Grant, A. J.; Pickup, B. T. A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653−1659.
(17) Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small molecule shape-fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 673−680.
(18) Melani, F.; et al. Field interaction and geometrical overlap: A new simplex and experimental design based computational procedure for superposing small ligand molecules. *J. Med. Chem.* **2003**, *46*, 1359−1367.
(19) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules *J. Comput.-Aided Mol. Des.* **2000**, *14*, 199−205.
(20) Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. Virtual screening of biogenic amine-binding G-protein coupled receptors: Comparative evaluation of protein- and ligand-based virtual screening protocols. *J. Med. Chem.* **2005**, *48*, 5448−5460.
(21) Cummings, M. D.; et al. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48*, 962−971.
(22) Pham, T. A.; Jain, A. J. Parameter estimation for scoring protein−ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856−5868.
(23) Miteva, M. A.; Lee, W. H.; Montes, M. O.; Villoutriex, B. O. Fast structure-based virtual ligand screening combining FRED, DOCK and Surflex. *J. Med. Chem.* **2005**, *48*, 6012−6021.
(24) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring functions. *J. Med. Chem.* **2000**, *43*, 4759−4771.
(25) Cole, J. C.; et al. Comparing protein−ligand docking programs is difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 325−331.
(26) OMEGA, OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.
(27) Berman, H. M.; et al. The Protein Databank. *Nucleic Acids Res.* **2000**, *28*, 235. http://www.rcsb.org.
(28) OEChem Toolkit, OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.
(29) ROCS, OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.
(30) Mills, J. E. J.; Dean, P. M. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey *J. Comput.-Aided Mol. Des.* **1996**, *10*, 607−613.
(31) Sheridan, R. P.; et al. Protocols for bridging the peptide to non-peptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395−1403.
(32) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29−34.

(33) Baldi, P.; et al. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412−427.
(34) Tribelleau, N.; et al. Virtual screening workflow development guided by the "receiver operator characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2541.
(35) Renfrey, S.; Featherstone, J. Structural proteomics. *Nat. Rev. Drug Discovery* **2002**, *1*, 175−192.
(36) Evers, A.; Klabunde, T. Structure-based drug discovery using GPCR homology modeling: Successful virtual screening for antagonists of the Alpha1A adrenergic receptor. *J. Med. Chem.* **2005**, *48*, 1088−1099.
(37) MDDR. http://www.mdl.com/products/knowledge/drug_data_report/.
(38) McGann, M. R.; et al. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76−91.
(39) Nicholls, A. N.; MacCuish, N. E.; MacCuish, J. D. Variable selection and model validation for 2D and 3D descriptors *J. Comput.-Aided Mol. Des.* **2004**, *18*, 451−461.
(40) Bostrom, J; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449−457.
(41) Perola, E.; Walters, W. P.; Charifson, P. S. An Analysis of Critical Factors Affecting Docking and Scoring in Virtual Screening. In *Virtual Screening in Drug Discovery*, 1st ed.; Alvarez, J., Shoichet, B., Eds.; Taylor & Francis: Boca Raton, FL, 2005; pp 47−85.
(42) Erickson, J. A.; et al. Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47*, 45−53.
(43) McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46*, 2895−2904.
(44) Rush, T. S., III; Grant, A. J.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1494.

JM0603365