Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine

Ajay N. Jain[†]

UCSF Cancer Research Institute and Comprehensive Cancer Center, University of California, San Francisco, California 94143-0128

Received September 17, 2002

Surflex is a fully automatic flexible molecular docking algorithm that combines the scoring function from the Hammerhead docking system with a search engine that relies on a surface-based molecular similarity method as a means to rapidly generate suitable putative poses for molecular fragments. Results are presented evaluating reliability and accuracy of dockings compared with crystallographic experimental results on 81 protein/ligand pairs of substantial structural diversity. In over 80% of the complexes, Surflex's highest scoring docked pose was within 2.5 Å root-mean-square deviation (rmsd), with over 90% of the complexes having one of the top ranked poses within 2.5 Å rmsd. Results are also presented assessing Surflex's utility as a screening tool on two protein targets (thymidine kinase and estrogen receptor) using data sets on which competing methods were run. Performance of Surflex was significantly better, with true positive rates of greater than 80% at false positive rates of less than 1%. Docking time was roughly linear in number of rotatable bonds, beginning with a few seconds for rigid molecules and adding approximately 10 s per rotatable bond.

Introduction

Discovery of novel lead compounds through virtual screening of chemical databases against protein structures is well established.[1] Many methods have been published, varying primarily two components: scoring functions[2−8] and search methods[9−12] (for a more complete review, see Bissantz et al.[13]). The primary criteria for evaluating docking strategies are docking accuracy, scoring accuracy, screening utility, and speed. These criteria tend to overlap. Docking accuracy reflects an algorithm's ability to discover a conformation and alignment (pose) of a ligand relative to a cognate protein that is close to that experimentally observed and to recognize the pose as correct. Recognition of a pose as correct embeds one aspect of scoring accuracy: a scoring function must ideally rank a correct pose of a molecule higher than an incorrect one. The second aspect of scoring accuracy is the ability to correctly predict the rank order of binding affinities of ligands to a particular protein. Scoring accuracy strongly influences screening utility, which measures the ability of a docking algorithm to detect true ligands of a protein within a background of random ligands not thought to bind the protein. Very low false positive rates are required, since the size of libraries to be computationally screened commonly exceeds 100 000 compounds. Computational speed is critical in the application of a docking algorithm to a screening problem.

Surflex is a new docking methodology that combines Hammerhead's empirical scoring function[2] with a molecular similarity method (morphological similarity)[14] to generate putative poses of ligand fragments. It implements an incremental construction search approach, as in Hammerhead,[9] but also implements a new fragment assembly methodology that is both faster and more accurate. This new fragment assembly method is loosely related to genetic algorithm approaches,[11] but it is deterministic.

Results are presented evaluating reliability and accuracy of dockings compared with crystallographic experimental results on 81 protein/ligand pairs of substantial structural diversity, beginning from randomized initial conformations and alignments of minimized ligands. Results are also presented assessing Surflex's utility as a screening tool on two protein targets (thymidine kinase and estrogen receptor) using the same data sets as in a recent paper that compared several docking and scoring algorithms.[13]

Performance of Surflex in terms of docking accuracy was comparable to the best available methods. Performance of Surflex in terms of screening utility was significantly better than that of competing methods, with true positive rates of greater than 80% at false positive rates of less than 1%, representing a 5- to 10-fold improvement. Docking time was competitive with the fastest methods and was roughly linear in number of rotatable bonds, beginning with a few seconds for rigid molecules and adding approximately 10 s per rotatable bond on a Pentium III based 933 MHz desktop hardware running Windows 2000 Professional.

Surflex is available free of charge to academic researchers for noncommercial use (see http://jainlab.ucsf.edu for details on obtaining the software).

Methods

Three data sets were employed in assessing Surflex docking performance relative to competing methods. In

[†] Phone: (415) 502−7242. Fax: (415) 502-3179. E-mail: ajain@cc.ucsf.edu. Address for regular postal service: UCSF Cancer Center, Box 0128, San Francisco, CA 94143-0128. Address for express mail: UCSF Cancer Center, 2340 Sutter Street, #S336, San Francisco, CA 94115.

10.1021/jm020406h CCC: $25.00
Published on Web 01/21/2003

what follows, first the data sets are described, then the algorithms, and last the precise procedures and parameters used.

Data Sets. The GOLD docking program[11] has been extensively tested on data sets of 100 protein/ligand complexes from the PDB (reported in Jones et al.[11]) and 34 additional complexes, both of which have been made publicly available (http://www.ccdc.cam.ac.uk/prods/gold/value.html). Of these 134 complexes, 81 meet the following criteria: (1) 15 or fewer rotatable bonds; (2) no covalent attachments between ligand and protein; (3) ligands with no obvious errors in structure. The reasons for these criteria are as follows. First, Surflex has been designed primarily as a screening tool of small-molecule libraries, and over 80% of ligands from commercial small-molecule screening libraries have 15 or fewer rotatable bonds. Second, Surflex's scoring function was developed strictly on noncovalent complexes, and the utility of screening hits that are reactive is generally thought to be minimal. Third, rather than "fixing" ligands, the complexes in which ligands had obviously incorrect structures were eliminated. Modifying the structures would have entailed generating starting poses based on newly minimized structures. Interpretation of the direct subset of acceptable structures has the benefit of using precisely the same ligands, proteins, and configurations used in the initial study, allowing for direct comparisons of results. These 81 complexes were used to evaluate Surflex's docking and pose recognition accuracy (see Results and Discussion) and are available via the author's web site (http://jainlab.ucsf.edu). They were used unmodified. For each structure, the minimized native ligand was used to generate 10 random conformations and alignments from which to perform dockings. Only bonds outside the ring systems were randomized, with the ring conformations used as found in the minimized ligands. These random initial poses are part of the available data set. This data set is referred to as the "81 complex set."

The data sets from the comparative paper of Bissantz et al.[13] were used to test Surflex's screening utility. These included protein structures for HSV-1 thymidine kinase (1KIM) and estrogen receptor alpha (3ERT), 10 known ligands of TK in arbitrary initial poses, 10 known ligands of ERα in arbitrary initial poses, and 990 randomly chosen nonreactive organic molecules from the ACD ranging from 0 to 41 rotatable bonds. The data sets were used without modification. These data sets are referred to as the "screening set".

Computational Methods. Surflex employs an idealized active site ligand (called a protomol, as described previously[15]) as a target to generate putative poses of molecules or molecular fragments. These putative poses are scored using the Hammerhead scoring function,[2] which also serves as an objective function for local optimization of poses. Flexible docking proceeds either by incremental construction from high-scoring fragments as in Hammerhead[9] or by a crossover procedure that combines pieces of poses from intact molecules, which will be described in detail. Since the scoring function, pose generation procedure, and protomol representation have been described elsewhere, they will be described only briefly here with indications of modifications to the algorithms. The software is available for academic, noncommercial research free of charge, so details of software operation will not be replicated here (see http://jainlab.ucsf.edu to obtain the software).

The following describes the overall high-level procedure, with details provided below. There are two phases to the algorithm.

1. Protomol Generation. An idealized binding site ligand is generated from the protein structure. This is done once for a particular protein.

1.1. Input: (a) protein structure including hydrogens, (b) list of residues to identify the protein site or a ligand structure within the protein binding site, used solely to identify residues proximal to the binding site.

1.2. Output: a protomol (mol file) that serves as a target to which putative ligands or ligand fragments are aligned on the basis of molecular similarity.

1.3. Procedure: three different types of molecular fragments are placed into the protein binding site in multiple positions and are optimized for interaction to the protein. High-scoring nonredundant fragments collectively form the protomol (see below for details).

2. Docking. Ligands are docked into the protein to optimize the value of the scoring function.

2.1. Input: (a) protein structure, (b) protomol, (c) ligand or ligands.

2.2. Output: the optimized poses of docked ligands along with corresponding scores.

2.3. Procedure (for each putative ligand) is the following.

2.3.1. Input ligand is fragmented, resulting in 1−10 molecular fragments, each of which may have some rotatable bonds.

2.3.2. Each fragment is conformationally searched.

2.3.3. Each conformation of each fragment is aligned to the protomol to yield poses that maximize molecular similarity to the protomol.

2.3.4. The aligned fragments are scored and pruned on the basis of the scoring function and the degree of protein interpenetration.

2.3.5. One of two procedures is used to construct full molecules from the aligned fragments: (a) an incremental construction approach as in Hammerhead or (b) a new whole molecule approach that is described in detail below. Results are presented with the whole molecule approach, which is faster than the incremental construction approach.

2.3.6. The best scoring poses are subjected to gradient-based optimization of conformation and alignment, and the top scoring poses are returned along with their scores.

Details of the scoring function, protomol generation process, and the docking search algorithm follow.

Scoring Function. The scoring function was tuned to predict the binding affinities of 34 protein/ligand complexes, with its output being represented in units of $-\log(K_d)$[2]. The range of binding affinities in the training set ranged from 10^{-3} to 10^{-14} and represented a broad variety of functional classes. The parametrization of the function effectively models the noncovalent interactions of organic ligands with proteins, including proteins with bound metal ions in their active sites. The function is continuous and piecewise differentiable with respect to ligand pose, which is important for the gradient-based optimization procedures employed. The

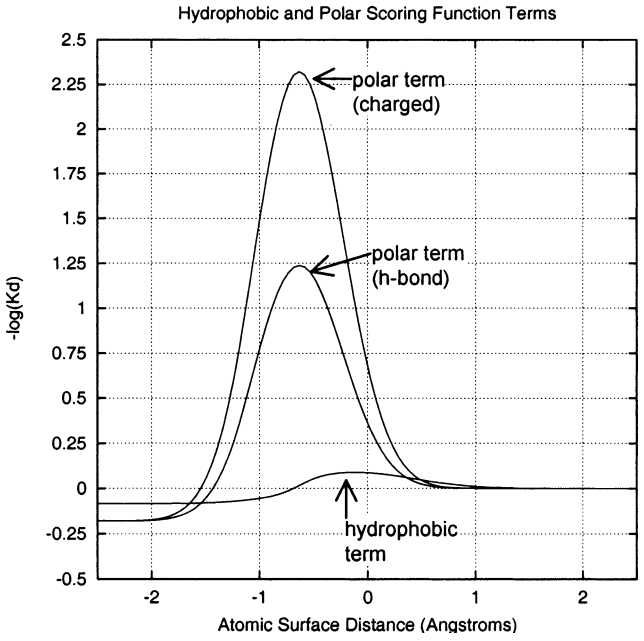

Figure 1. Hydrophobic and polar terms of the scoring function. The hydrophobic term peaks at approximately 0.1 units with a slight surface interpenetration. The polar term for an ideal hydrogen bond peaks at 1.25 units, and a charged interaction (tertiary amine proton (+1.0) to a charged carboxylate oxygen (−0.5)) peaks at about 2.3 units.

terms, in rough order of significance, are hydrophobic complementarity, polar complementarity, entropic terms, and solvation terms. The full scoring function is the sum of each of these terms.

The dominant terms are the hydrophobic contact term and a polar contact term that has a directional component and is scaled by formal charges on the protein and ligand atoms. These functional terms are parameterized on the basis of distances between van der Waals surfaces, with negative values indicating interpenetration. Each atom on the protein and ligand is labeled as being nonpolar (e.g., the H of a C−H) or polar (e.g., the H of an N−H or the O of a C═O), and polar atoms are also assigned a formal charge if present. Figure 1 shows plots of the hydrophobic term, the polar term for a hydrogen bond with no formal charge, and an interaction between the proton of a charged tertiary amine with an ideally oriented charged carboxylate. Note that formal charge for resonant structures such as carboxylates is spread on the heteroatoms for negative charges and across the hydrogens attached to positively charged heteroatoms. The hydrophobic term (bottom curve) yields approximately 0.1 units per ideal hydrophobic atom/atom contact. A perfect hydrogen bond yields about 1.2 units (middle curve) and has a peak corresponding to 1.97 Å from the center of a donor proton to the center of an acceptor oxygen (learned entirely on the basis of the empirical data and corresponding quite closely to the expected value range). The charged interaction peaks at about 2.3 units (top curve). Despite the large difference in the value of an individual hydrophobic versus polar contact, the hydrophobic term accounts for a *larger* total proportion of binding energy on average. There are many more hydrophobic contacts than ideal polar contacts in a typical protein/ligand interaction.

Apart from the hydrophobic and polar terms, the remaining terms that have a significant impact on ligand scores include the entropic term and the solvation term. The entropic term includes a penalty that is linear in the number of rotatable bonds in the ligand, intended to model the entropic cost of fixation of these bonds, and a term that is linearly related to the log of the molecular weight of the ligand, intended to model the loss of translation and rotational entropy of the ligand. The solvation terms are linearly related to a count of the number of missed opportunities for appropriate polar contacts at the ligand/protein interface. However, neither the solvation term nor any of the terms intended to guard against improper clashes received much weight in the training. This was due to the fact that no negative data were employed; only ligands with their cognate proteins were used in parameter estimation, so there was essentially no data from which to induce such penalty terms.

Protomol Generation. The protomol docking targets differ from those previously reported by making use of slightly different molecular fragments. Surflex's protomols utilize CH_4, C═O, and N−H molecular fragments (Hammerhead used single H atoms instead of CH_4 molecules). The molecular fragments are placed in the protein active site based on identification of empty 1 Å voxels that are between residues on the protein that have been marked to identify the active site. Lines connecting all pairs of marked residues are traversed, and voxel scores are incremented as they are traversed. After a 3D Gaussian smoothing, voxels that score above a threshold are used as starting points for the placement of molecular fragments. The three types of molecular fragments are each placed in each voxel and are locally optimized using the scoring function. The polar fragments are placed in 36 different orientations. High-scoring fragments are retained, with redundant fragments being eliminated. Figure 2 shows the protomol generated for streptavidin based on identification of the protein residues containing any atoms whose surface was within 2.0 Å of any atom of the native ligand biotin. The protomol generated mimics the interactions made by biotin with streptavidin and identifies some contacts that are not made by the native ligand. The entire process for a typical protein takes less than 1 min.

Docking Search Algorithm. a. Molecular Alignment. Surflex utilizes the morphological similarity function and fast pose generation techniques described previously[14] to generate putative alignments of molecules or molecular fragments to the protomol. Briefly, morphological similarity is defined as a Gaussian function of the differences in molecular surface distances of two molecules at weighted observation points on a uniform grid. The surface distances computed include both distances to the nearest atomic surface and distances to donor and acceptor surfaces. Rapid generation of molecular alignments that maximize the similarity function is possible because the molecular observations are local and are not dependent on the absolute coordinate frame. So, two unaligned molecules or molecular fragments that have some degree of similarity will have some corresponding set of observers that are seeing the same things. Optimization of the similarity of two unaligned molecules is performed by finding sets of

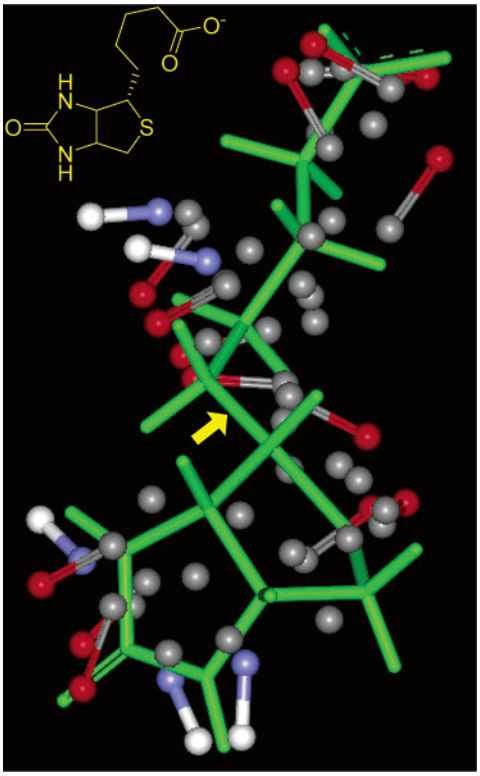

Figure 2. Protomol for streptavidin (1stp) compared with the native pose of biotin (green). The protomol consists of CH_4 (hydrogens not shown), N−H, and C═O molecular fragments. The protomol was generated solely on the basis of protein structure. The location of the protomol was indicated by marking the residues of the protein proximal to the native ligand. Each of the interactions made by biotin with streptavidin is well represented by a matching molecular fragment. The indicated bond is broken by Surflex to make fragments of biotin for docking.

observers of each molecule that form triangles of the same size, where each pair of corresponding points in the triangles are observing similar features. The transformation that yields a superposition of the triangles will tend to yield high-scoring superpositions of the molecules. In Surflex's docking search algorithm, poses of molecular fragments that tend to maximize similarity to protomols are used as input to the scoring function and are subject to thresholds on protein interpenetration and local optimization based on the gradient of the scoring function.

b. Molecular Fragmentation. Molecular flexibility in docking is addressed by molecular fragmentation. Molecules are fragmented by breaking non-ring rotatable bonds. Each such break eliminates a bond for conformational search and eliminates the need to cross the conformations of the two fragments. So, a molecule with seven rotatable bonds, where each bond is sampled at six rotameric positions, is reduced from 6^7 (>250000) conformations to $6^3 + 6^3$ (432) conformations, a reduction by nearly 3 orders of magnitude. In practice, a heuristic set of rules are employed in conformational sampling, where two, three, or six rotamers are used for each bond (e.g., three for SP3−SP3 bonds). Also, a maximum number of conformations per fragment can be specified (default 20), and the algorithm selects the most different conformations based on the root-mean-square deviation (rmsd). Following completion of fragmentation and conformational search (and fast internal clash relaxation), the resulting molecular fragments are aligned to the protomol. For biotin, there are two such fragments (the molecule breaks at the bond indicated by the arrow in Figure 2) with a total of 21 conformations (1 for one fragment and 20 for the other) to be aligned. There are two alternative procedures for molecular construction from fragments: incremental construction and "whole molecule". They also differ slightly in how the fragments are aligned. However, both procedures use the same fragmentation process.

c. Incremental Construction. In the incremental construction mode, all molecular fragments are aligned to maximize similarity to the protomol. The highest scoring (by the docking scoring function) are used as "heads", which are locally optimized for fit to the protein. From these, a directed alignment of the "tail" (next molecular fragment) occurs by aligning each conformation of the appropriate fragment on the basis of similarity to the protomol but subject to the constraint that the alignments generated must place the connector atom proximal to where it must be to make a connection to the head. This procedure is highly analogous to that of Hammerhead, but it relies on similarity to the protomol instead of direct atom matching in order to generate putative alignments. The process of incremental construction, while being reliable and quite fast compared with many flexible docking methods, has one primary weakness. It makes a very strong independence assumption that maximizing the similarity of potentially very small molecular fragments to the protomol will tend to generate good poses. For particularly flexible molecules, this may not be a good assumption. A very small fragment may have a very different optimal similarity based alignment than the same fragment in the context of the remainder of the molecule.

d. Whole Molecule Algorithm. Surflex also implements a new approach to circumvent the strong independence assumption above. In this "whole" molecule docking approach, after the molecule is fragmented, the "dead" pieces are still carried along with the "live" piece in conformational search. However, only the live piece corresponding to the fragment is searched. So, at the end of the fragmentation and search process, the same number of total conformations exists as in the incremental construction procedure. The difference is that in generating putative poses to the protomol, the "dead" molecular pieces in their arbitrary initial conformation are carried along with the searched fragment in the molecular similarity computation. So, the pool of docked poses to the protein contains alignments of initial conformations of the ligand, which have been locally conformationally sampled on the basis of the fragmentation process. The poses that have the most inappropriate protein interpenetration are quickly eliminated.

Since the scoring function is based on atom/atom pairwise interactions, it is possible to generate a score for any fragment of a docked pose. The poses that remain after elimination of the worst interpenetrations are scored on the basis of their individual fragments, with the best of these being subject to local optimization. Figure 3 illustrates this situation with two poses of biotin relative to its binding site. In this case, the ring

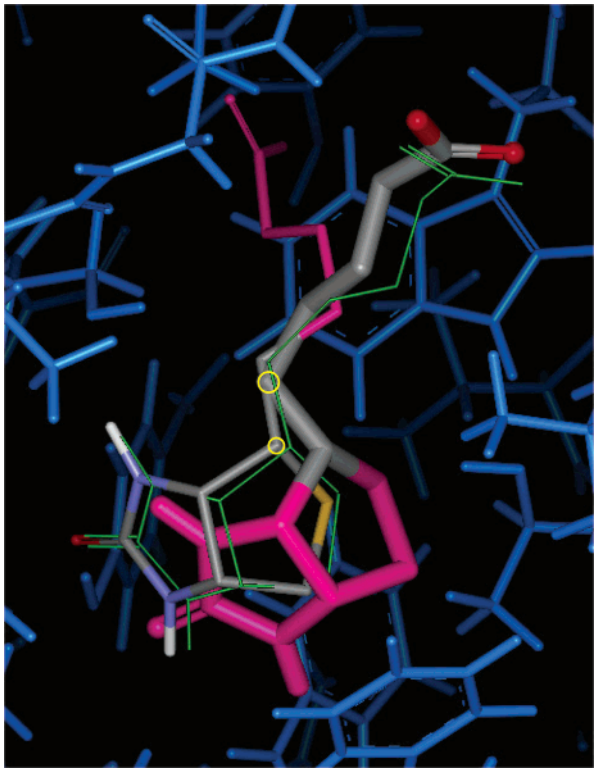

Figure 3. Biotin during the docking process to streptavidin (blue): (thin sticks) biotin's ring system in a high-scoring, well-docked configuration (atom color) with the carboxylate tail (magenta) extending into the protein; (thick sticks) biotin's carboxylate tail in a high-scoring, well-docked configuration (atom color) with the ring system (magenta) deviating from ideal; (green lines) the result of merging the two well-docked fragments at the atoms indicated by yellow circles. The merged pose closely follows the parent fragments' original configurations.

system of biotin on one pose (thin sticks) is close to being correct, with the tail section containing the carboxylate being close to correct in the other pose (thick sticks). The atoms indicated in magenta belong to the "dead" piece of their respective poses. The goal is to identify the high-scoring "live" pieces (shown in atom color in Figure 3) and to merge them while minimizing the final deviation from the original poses, which yielded the high scores to begin with. From the collection of docked poses, a recursive search is performed for high-scoring fragments from different posed conformers that have acceptable mutual geometry to be merged by enforcing the broken bond between the fragments. In Figure 3, the atoms on either side of the fragmentation bond in the two separate fragments are indicated with circles. The difference between the length of the fragmentation bond and the distance between the atoms on either side of the bond, but in different poses, is computed. Those pose pairs where the difference is low (default of 0.5 Å) become candidates for merging (in Figure 3, the difference is 0.36 Å). Further pruning is done on the basis of the degree of overlap of the atoms in the fragments to be joined. The first pass of this procedure yields a list of fragment pairs potentially suitable for merging. If the molecule has two fragments, the procedure proceeds to merging the fragment pairs; otherwise, it recursively enumerates additional fragments to be merged onto the growing molecules.

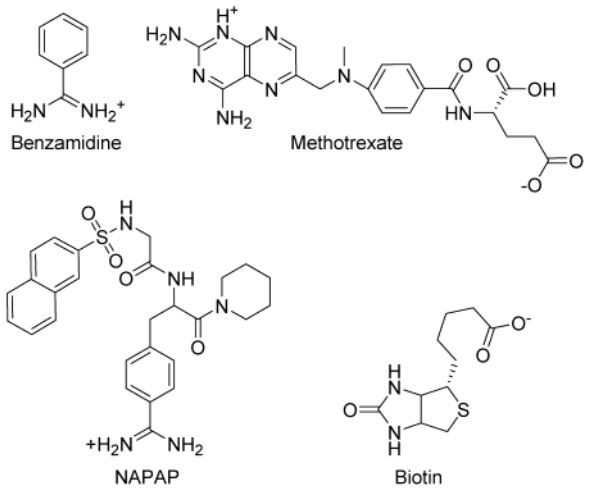

Figure 4. Ligands of trypsin (3PTB), DHFR (4DFR), thrombin (1DWD), and streptavidin (1STP) (left to right, top to bottom, respectively) used in tuning Surflex for docking accuracy. The ligands are shown in the protonation states used.

The recursive search yields a list of whole molecules, each consisting of fragments chosen from different docked poses of the ligand and each of which scores well in total over all fragments. Each fragment bond is enforced one at a time. After each constraint is applied, the new pose is quickly optimized to minimize the rmsd from the atomic positions of the original fragments. Then a fast local optimization with respect to the protein is applied using the scoring function. In Figure 3, the pose of biotin that corresponds to this stage is shown in green lines. Note that the merged pose corresponds very closely to the original poses of the separate fragments but now correctly respects the bond constraints. The whole molecules resulting from the procedure are pruned on the basis of the docking score and are subjected to further gradient-based score optimization. The procedure ultimately returns the 10 best scoring poses. In each stage of search and construction, there are limits on the number of partial solutions retained by the algorithm. At no point does the number of simultaneously considered posed fragments exceed a few thousand, even in the case of highly flexible molecules. Generally, the alignment generation process and the fragment enumeration and merging process take roughly similar amounts of time and account for the bulk of the computational cost. The molecular merging process is similar to the technique of Pitman et al.[16] The primary difference is that in Surflex, the hierarchy of optimizations following the enforcement of a bond constraint varies rotatable bonds as well as alignment parameters.

e. Search Algorithm Tuning. Tuning of the algorithm was performed on the ligands and cognate proteins indicated in Figure 4. These were the same complexes used in tuning Hammerhead[9] and represent a range of flexibilities (1−10 rotatable bonds) and binding affinities (10^{-6}−10^{-13}). The goal of the tuning process was to ensure sufficient search depth to reliably discover poses close to correct, beginning from random initial conformations and alignments. Presuming a reliable scoring function, the dockings would then be accurate. No effort was made to develop a parameterization for more rapid screening of large databases

of molecules, where sufficiently accurate scoring is more important than accurate docking by rmsd, and considerable speedup is possible toward that end.

Computational Procedures. The protomol generation and docking algorithms described above are completed in two procedural steps. For both data sets, the same docking procedure and parameters were used in all runs. For the protomol generation, the same procedure and parameters were used within all complexes of the 81 complex set, but there was a minor difference in treatment of the two proteins in the screening set due to the lack of a pose for the two native ligands in the latter experiments.

For the 81 complex set, protomol construction was based on protein residues proximal to the native ligand and on parameter settings to produce a small and buried docking target (Surflex parameters: −proto_thresh 0.5 and −proto_bloat 0). For the screening set, protomols were generated for TK and ER using default parameters based on identification of the residues proximal to the native ligands in 1KIM and 3ERT, respectively. Docking was run with the new whole molecule approach (Surflex parameter: −whole) and with default settings for all other parameters.

Each docking of a putative ligand returned up to 10 scored poses, with the score consisting of a nominal affinity score in units of $-\log(K_d)$ as in Hammerhead.[2] A second value, representing the degree of protein interpenetration, was also reported. This negative value, called "bump", has arbitrary units. For choosing poses with the best score from multiple dockings of the same ligand (in the 81 complex set), the score and bump were combined into a single value, scaling the bump value down by a factor of 5.0. In the tables of experimental results, the protein penetration value reported is the reported bump value divided by 5.0. Where results are reported on the best pose by rmsd from multiple dockings for the 81 complexes, all dockings returned by Surflex (maximum of 10 for each) were considered. For all other results on both data sets, only the top ranked pose returned by Surflex was used.

Results and Discussion

Surflex was evaluated for docking accuracy, scoring accuracy, screening utility, and speed. Docking accuracy and the pose recognition aspect of scoring accuracy were assessed on 81 diverse protein/ligand complexes (see Methods). The second aspect of scoring accuracy (within-protein ranking of ligands) was assessed in the context of screening utility on two proteins, each with 10 true ligands and 990 random ligands. Speed was assessed in all cases. For all of the results reported, a single parameter set for Surflex was used in the docking runs and no modification of the proteins or ligands was made from the original authors' data sets (see Methods for details).

Docking Accuracy and Correct Pose Recognition. Table 1 summarizes Surflex performance on the 81 complex data set. The table is broken into five blocks of columns. The first contains data about each protein's cognate ligand: number of rotatable bonds, rmsd (Å) of the pose corresponding to the minimized ligand gradient optimized for docking score, and the resulting score ($-\log(K_d)$). The second reports the mean docking time for 10 random initial conformations and alignments of the native ligand. The third block has data based on the best docked pose by rmsd over all 10 dockings: rms deviation, score, and "pen." (penalty for inappropriate interpenetration of protein). The fourth has the same data but for the *best scoring* pose over all 10 dockings. In the choice of the best scoring pose, the sum of the docking score and penetration was used. The fifth contains data on the proportion of good runs with respect to screening and identifying a correct pose as the top-ranked. A good screening run was defined as one in which the ligand scored well enough for screening (greater than 1 log unit less than the optimized native pose or 5.0 ($-\log(K_d)$), whichever was lower). A good docking accuracy was one in which the top scoring pose was within 2.5 Å rms deviation from experimental results.

In 76/81 cases (94%), Surflex returned a pose among the top poses for all dockings within 2.5 Å rmsd (72/81 or 89% within 1.5 Å). This is a measurement of how thorough the search procedure is and to what extent the scoring function is able to recognize good dockings, since Surflex returns just 10 poses for each docking. Figure 5 (left) compares these results to those of GOLD,[11] which employed 20 independent runs (versus 10 for Surflex). The Surflex results were slightly better, with half as many solutions having rmsd of greater than 2.0 Å. In practice, one is more concerned with whether a *single* docking from a random initial pose will yield a correct pose as its top ranked candidate. Table 1 also reports the proportion of 10 dockings for each protein that produced a correct pose as the top ranked based on score. In 43/81 (53%) of cases, the proportion of such dockings was 80% or greater. In 16/81 cases (20%), the proportion was 20% or fewer. With a single docking from a random initial pose, the chances were nearly 70%, on average, of finding a pose that was close to correct.

The issue of pose recognition is partially addressed by the discussion above. From the essentially infinite space of docked poses, the pool of poses from independent dockings contained a pose within 2.5 Å rmsd of the experimental result 94% of the time. However, the question of whether a docker can reliably identify a close to correct pose as scoring the *highest* among predicted poses is also very important. Figure 5 (right) is a plot of the results for Surflex and for GOLD for rms deviation of best scoring pose for each complex. The results were quite comparable. For Surflex, of the 76 proteins where a good docking was generated in any of the runs, the best scoring pose over all runs was within 2.5 Å 65 times (86%) and within 1.5 Å 50 times (66%).

There are two subtleties that are important to note. First, the pose selected as best on the basis of score includes both the reported score and the contribution of the interpenetration term. The latter term can be high even in the case of a correctly docked ligand, beginning from a minimized conformation. Often, the coordinates of a crystal structure are not compatible with a low-penetration configuration of a ligand structure that respects bond angles and lengths. Second, the search depth has a strong effect on the results. In the case where an excellent pose by rmsd exists, it may still not score nearly as well as a nearby pose with equivalent rmsd. So, there were 11 of 76 cases where the best

Table 1. Results for Surflex on 81 Protein/Ligand Pairs

	optimized minimized native ligand				best pose by rms over all 10 runs			best pose by score over all 10 runs			proportion of good runs	
PDB code	nrot	rmsd	score	mean time, s	rmsd	score	pen.	rmsd	score	pen.	screen	rms
6abp	4	0.28	6.6	9.8	0.14	7.1	−0.2	0.28	8.6	−0.3	0.8	0.9
1abe	4	0.30	6.7	10.8	0.18	6.6	−0.5	0.27	9.0	−0.3	0.8	1
1tng	1	0.33	5.4	3.8	0.20	5.0	0.0	0.22	5.1	0.0	0.8	1
1lst	5	0.24	12.6	27.0	0.23	12.5	−0.5	0.33	12.4	−0.3	1	1
1tnl	1	0.47	4.0	4.2	0.23	4.1	−1.0	2.26	4.0	−0.2	1	1
1wap	3	0.32	10.0	26.4	0.24	9.9	−0.2	0.30	9.9	−0.2	1	1
1lah	4	0.28	12.9	10.2	0.25	12.3	−0.3	0.30	12.7	−0.3	1	1
1ukz	6	0.41	10.9	49.2	0.25	10.1	−1.8	0.77	11.6	−1.6	1	0.7
1aha	0	0.38	6.3	2.4	0.26	6.1	0.0	0.37	6.4	0.0	1	1
2gbp	6	0.26	10.2	15.9	0.27	8.1	−0.4	0.63	8.1	−0.2	1	1
1dbb	1	0.56	7.6	9.9	0.28	6.2	−0.2	0.54	7.2	−0.3	1	1
2ada	6	0.29	15.3	47.6	0.29	14.3	−0.1	0.32	14.6	−0.1	1	1
1dr1	4	0.57	6.3	25.4	0.29	5.5	−0.2	1.25	7.3	−0.6	1	1
2ctc	4	0.39	9.4	8.2	0.32	6.8	−1.1	0.38	8.6	−0.4	0.8	1
1coy	1	0.66	7.0	7.4	0.32	6.7	−1.1	0.54	7.7	−0.5	1	0.9
3aah	3	0.44	14.3	27.4	0.33	14.9	−0.5	0.68	16.1	−0.4	1	1
7tim	5	0.42	3.4	8.4	0.34	5.4	−1.3	1.20	5.4	−0.5	1	1
1hsl	3	0.50	8.9	12.0	0.35	8.2	−1.7	0.51	8.8	−0.6	1	1
1srj	4	0.37	7.7	29.1	0.35	8.9	−1.0	0.39	9.2	−1.0	1	1
2sim	10	0.32	10.7	52.3	0.35	10.3	0.0	1.10	11.3	−0.1	1	1
3tpi	7	0.45	10.0	42.1	0.37	9.0	−0.4	0.52	10.3	−0.2	1	1
3ptb	1	0.60	6.3	4.4	0.37	4.5	−0.6	0.54	6.6	0.0	1	1
1ldm	1	0.51	7.3	2.9	0.38	6.5	−0.2	0.44	7.6	−0.2	0.8	1
1hdy	0	0.65	3.4	1.2	0.41	2.7	−2.6	0.66	3.4	−0.4	0.9	1
1phg	3	0.35	6.8	10.8	0.41	6.1	−3.5	4.44	6.1	−2.6	0.9	0.2
2phh	2	0.44	6.4	5.3	0.41	6.3	−0.2	0.44	6.3	−0.1	0.4	0.6
2cht	3	0.40	7.9	11.1	0.42	7.1	−0.7	0.42	7.4	−0.8	0.8	0.9
1mdr	3	0.64	7.9	8.5	0.45	5.3	−1.6	0.68	7.9	−0.3	1	1
1stp	5	0.51	11.4	28.3	0.46	11.4	−0.7	0.51	11.7	−0.3	1	1
1ack	3	0.36	3.9	9.9	0.50	3.2	−0.8	1.18	4.9	−0.5	1	0.9
1frp	8	0.41	9.4	41.6	0.50	9.8	−0.6	0.75	10.5	−0.5	1	0.8
1cbs	5	0.35	6.3	49.0	0.52	6.5	−0.6	1.77	7.3	−0.9	1	1
4cts	3	0.63	7.8	6.6	0.53	7.9	−0.4	2.20	7.9	0.0	0.8	0.9
1hyt	5	0.70	6.5	9.3	0.53	5.7	−0.5	0.55	6.0	0.0	0.2	0.4
1lcp	3	0.92	4.2	7.6	0.54	3.2	−1.3	2.01	5.3	−0.3	1	1
1rob	6	0.59	4.4	26.7	0.56	4.2	−1.0	0.82	4.5	−0.1	0.9	0.4
1dbj	1	0.62	6.6	6.8	0.57	4.9	−0.1	0.88	5.7	−0.2	0.3	0.4
1ulb	0	0.61	4.7	2.3	0.58	4.7	−0.3	0.77	5.5	−0.1	0.8	0.8
2ak3	6	0.62	8.5	48.1	0.58	9.0	−1.0	0.60	9.5	−1.1	0.7	0.6
1mrg	0	0.71	5.6	2.0	0.60	5.6	−0.1	0.70	5.6	−0.1	0.9	1
1lna	9	0.57	8.8	44.7	0.60	8.7	−0.3	0.88	9.2	0.0	1	0.8
1aco	4	0.62	12.0	8.4	0.63	9.2	−2.9	3.39	9.3	−0.5	1	0.7
1fki	0	0.69	6.6	6.7	0.64	6.6	−0.4	0.70	6.8	−0.4	0.6	0.8
1com	4	1.05	8.0	8.1	0.64	4.7	−0.5	0.86	7.4	−0.2	1	0.9
1tmn	14	0.53	14.5	171.2	0.65	13.6	−1.5	1.30	12.8	−1.1	1	0.6
3cpa	7	0.70	6.7	40.7	0.66	6.6	−1.7	1.90	8.7	−1.3	1	0.5
2dbl	6	0.63	8.5	64.6	0.66	8.7	−1.3	0.81	9.1	−1.0	1	0.9
1cbx	5	0.45	12.7	10.4	0.70	9.0	−0.4	0.70	9.0	−0.4	1	1
1mrk	5	0.58	5.3	34.1	0.75	8.0	−0.1	0.85	8.5	−0.1	1	1
6rsa	3	0.80	5.3	19.2	0.75	5.3	−0.6	0.78	5.3	−0.6	0.8	0.8
1trk	8	0.57	9.7	66.0	0.78	9.5	−1.8	1.22	12.6	−1.6	1	0.8
1fen	4	0.35	6.9	113.6	0.79	7.5	−1.0	1.18	9.4	−0.5	1	1
2lgs	4	0.83	7.7	7.4	0.79	4.9	−0.2	1.22	5.9	−0.2	0.5	0.7
1acj	0	0.35	5.2	3.7	0.81	4.7	−0.4	3.89	6.5	−0.3	1	0.2
1bma	14	0.67	6.6	100.4	0.86	6.1	−0.3	1.00	6.1	−0.2	0.4	0.5
2cgr	8	0.57	10.9	61.3	0.89	10.2	−1.2	1.63	10.2	−1.1	1	0.2
1eap	11	0.63	7.0	72.5	0.92	7.1	−3.2	4.89	10.0	−2.9	1	0.1
1dwd	11	0.49	9.1	113.1	0.93	8.7	−2.9	1.68	8.7	−1.7	1	0.4
8gch	9	0.68	7.9	69.9	0.96	5.3	−1.3	4.51	7.4	−1.2	0.9	0.2
1atl	11	0.65	7.5	65.5	1.05	7.1	−1.5	7.01	8.9	−1.0	0.9	0.3
1bbp	11	0.81	13.4	257.1	1.06	10.1	−1.0	1.07	13.2	−1.3	1	0.2
2r07	8	0.55	6.6	91.3	1.09	7.8	−0.7	1.35	8.6	−0.5	1	0.9
1baf	7	1.11	5.5	64.7	1.10	5.2	−0.9	6.52	7.9	−0.8	0.5	0.3
4dfr	10	0.56	11.6	74.2	1.24	8.4	−2.3	1.60	10.8	−1.7	1	0.6
1tni	4	1.48	4.8	9.6	1.33	3.6	−2.1	2.97	3.9	−0.1	0.9	0.4
1acm	7	0.74	9.6	31.6	1.35	6.0	−0.3	1.43	7.3	−0.5	0.8	0.2
1lpm	8	1.02	3.3	63.2	1.44	5.3	−2.7	1.87	6.4	−1.2	1	0.1
1hdc	6	0.60	3.9	104.3	1.47	6.6	−0.9	1.80	9.1	−0.6	1	0.7
1tka	8	0.47	5.4	63.7	1.49	7.6	−2.3	1.96	11.0	−1.3	1	0.6
2cmd	6	0.39	10.2	12.3	1.49	5.1	−2.9	1.60	7.7	−0.4	0.9	0.4
1epb	5	0.49	5.7	72.5	1.52	7.2	−2.2	2.87	8.5	−2.0	1	0.3
1fkg	11	0.82	6.6	108.2	1.52	4.7	−1.0	1.81	5.9	−1.0	0.2	0.2
3hvt	1	0.91	5.7	5.6	1.61	5.2	−0.3	1.64	5.4	−0.3	0.5	0.4
1hri	9	0.97	5.3	74.8	1.96	6.3	−0.7	1.98	7.1	−1.0	1	0.4
1lic	15	0.67	3.8	155.6	2.19	5.2	−1.1	3.46	7.1	−1.0	1	0
1snc	6	1.08	6.5	42.0	2.44	5.8	−2.0	4.92	6.4	−1.2	1	0
1etr	10	0.57	7.1	111.2	3.01	7.6	−3.6	4.05	9.5	−2.4	1	0
1glq	15	0.45	12.6	132.5	3.49	6.6	−1.2	5.68	8.7	−0.9	0.7	0
6rnt	7	0.55	5.3	44.1	4.68	3.4	−1.1	7.03	6.0	−1.9	0.3	0
1rds	11	0.86	12.2	94.4	4.79	6.7	−2.0	9.83	7.6	−0.8	0.6	0
1nco	11	0.57	10.6	151.9	6.69	9.5	−6.4	8.26	10.8	−3.2	1	0

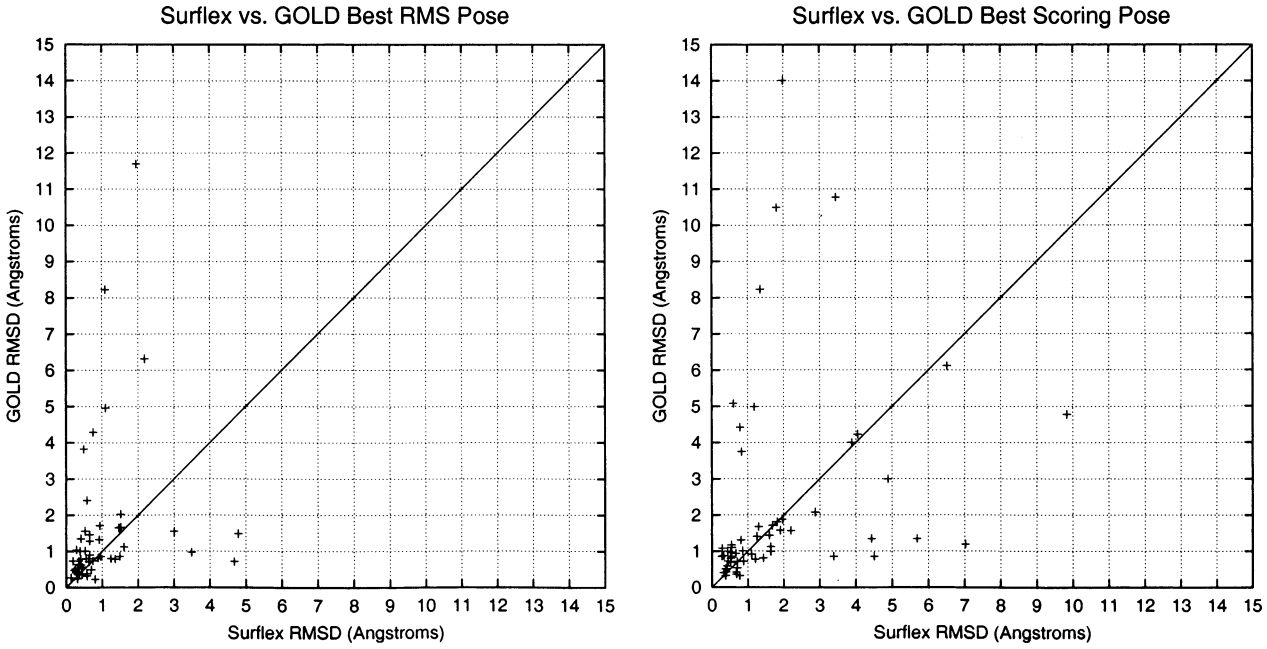

Figure 5. Plots of Surflex versus GOLD performance on 58 protein−ligand complexes tested in common: (left plot) rmsd values of best docked pose from 10 random starting conformations (Surflex) and 20 random initial seeds (GOLD); (right plot) rmsd values of highest scoring poses over the same dockings.

scoring pose was beyond 2.5 Å rmsd and scored better than the best pose by rmsd. However, 5 of these 11 scored *worse* than the score that resulted from gradient optimization of the minimized ligand. In these cases, additional search depth would have yielded a low rmsd pose as the best scoring pose. Of the 6 remaining cases (where the best score exceeded the optimized minimized ligand's score), only 2 scored significantly better than all poses discovered that were within 2.5 Å rmsd, again suggesting that increased search depth would have revealed a maximal scoring pose within an acceptable rmsd. Overall, it appears that the scoring function appears to recognize the correct binding mode reliably.

Figure 6 shows representative examples of dockings of varying levels of accuracy. High-accuracy docking was defined as rmsd less than 0.7 Å for the best scoring pose of a complex. This accounted for 33% of the 81 complexes. The ligand of 3tpi (7 rotatable bonds) was representative of this group, with all atoms of the ligand correctly placed. The ligand of 1tmn (14 rotatable bonds) was representative of a good accuracy docking ($0.7 \leq$ rmsd < 1.5), which accounted for 29% of the complexes. All parts of the ligand having significant interactions with the protein were very well docked. In particular, the interaction between the carboxylate and the protein's catalytic zinc ion is well predicted. The indole ring system was inverted in the docked pose, but this moiety was largely solvent-exposed in the crystal structure. The ligand of 1hri (9 rotatable bonds) is representative of acceptable accuracy, defined as $1.5 \leq$ rmsd < 2.5 and accounting for 18% of the complexes. There are no polar contacts between the protein and the native ligand to guide the docking. The envelopes of the docked pose and the native pose are very similar, with the docked pose accounting for the hydrophobic contacts between the ligand and protein. The representative of the poorly docked cases (rmsd > 5.0, 7% of complexes) was 1atl, with 11 rotatable bonds. The ligand in the docked pose shown in Figure 6 scored higher than the score of the experimentally determined ligand pose. In this case, Surflex incorrectly rotated four moieties out of the conjugated system. Here, Surflex's heuristic rules for conformational search failed. The Surflex scoring function does not count intramolecular ligand nonbonded contacts toward a ligand's docking score, and this also contributed to the problem. In the docked pose, the ligand was flipped, but it still retained the correct contact with the glutamine residue. The driving force behind the flip appeared to be the salt bridge between the primary amine of the ligand and the aspartic acid of the protein at left. Note, however, that Surflex yielded a pose within 2.5 Å rmsd that scored just 0.5 log units lower than the pose shown.

Screening Utility. Table 1 reports one additional measure of scoring accuracy that affects screening utility: the proportion of the time that Surflex returned its highest scoring pose at a score that was big enough to be retained in a large screen. The threshold was set (somewhat arbitrarily) at the minimum of 5.0 units ($-\log(K_d)$) and 1 log unit less than the optimized score from the native minimized conformation. By this criterion, Surflex found a high enough scoring solution at least 80% of the time in 68/81 (84%) of the cases.

A more substantial test of screening utility was made on two proteins used by another group to quantitatively compare the performance of GOLD, Dock 4.0, and FlexX.[13] HSV-1 thymidine kinase (TK, PDB code 1kim) and estrogen receptor α (ER, PDB code 3ert) were used as targets in screens comprising 10 known ligands for each protein and 990 randomly chosen molecules (see Methods). The known ligands used for TK and ER are shown in Figures 7 and 8, respectively. For the TK case, experimental crystallographic data were also available for each of the 10 TK ligands. Table 2 summarizes the rmsd values for the 10 TK ligands. As above, Surflex performance was very similar to that of GOLD, but it far exceeded the performance of Dock and FlexX (note that in this case only a single run of Surflex was

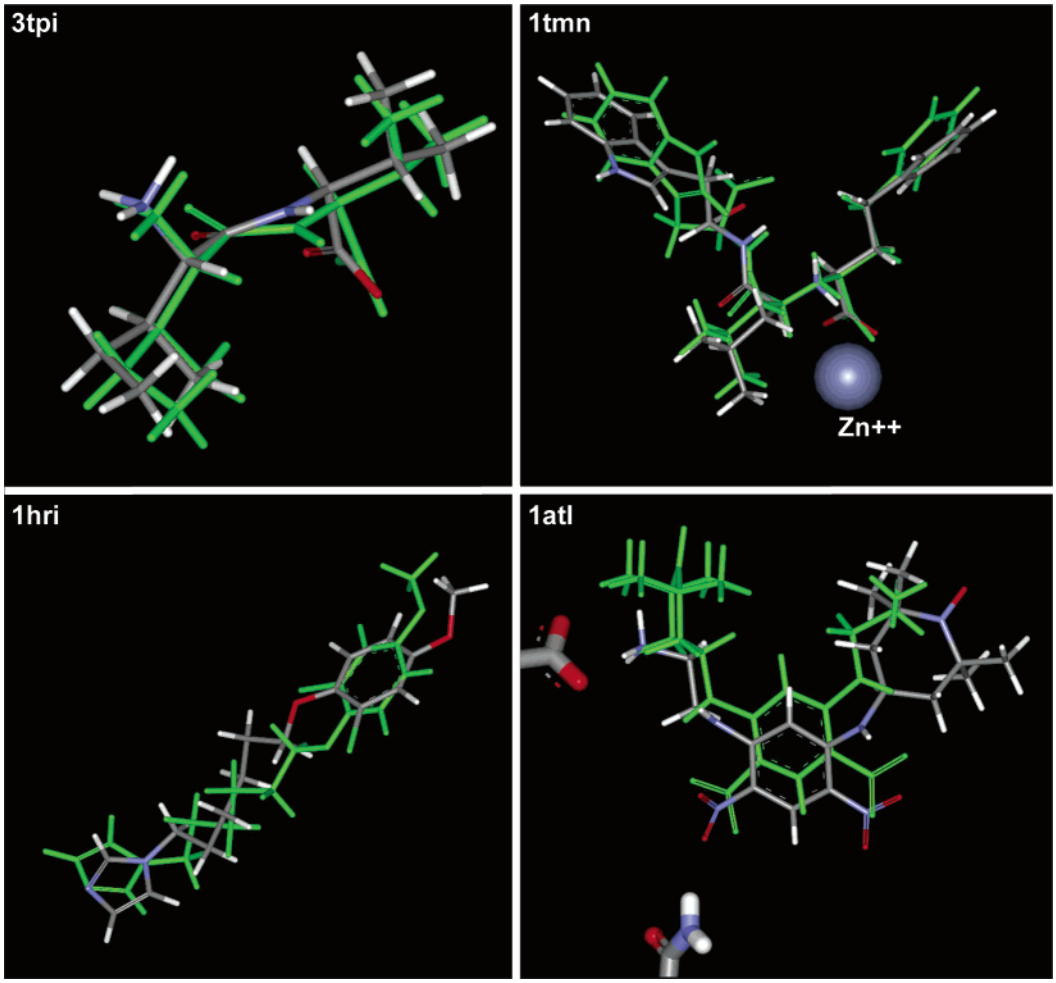

Figure 6. Representative examples of dockings of high accuracy (3tpi, 0.5 rmsd, 7 rotatable bonds), good accuracy (1tmn, 1.3 rmsd, 14 rotatable bonds), acceptable accuracy (1hri, 2.0 rmsd, 9 rotatable bonds), and poor accuracy (1atl, 7.0 rmsd, 11 rotatable bonds). Over 80% of the best scoring dockings had rmsd less than 2.5 (acceptable, good, or high accuracy).

performed, but the GOLD result was for the best scoring pose over 10 runs). The TK structure used was from a structure bound to thymidine. Consequently, the three pyrimidines were docked least accurately, since significant conformational adaptation occurs on the part of the protein on binding a pyrimidine versus a purine.

More important than docking accuracy for the purposes of screening is the ability of a docker to detect true positives against a background of random molecules. The false positive rate is exacerbated by the large size of screening libraries, which can commonly exceed 100 000 compounds. False positive rates of 5% to achieve a given true positive rate will yield 5000 inactive compounds to recover just a few active compounds. While this may enrich active compounds over an exhaustive screening approach, improvements in false positive rates enhance the efficiency of screening and reduce costs linearly with reduction in false positive rates.

In a screening experiment on a particular protein, interpretation of Surflex results requires a threshold on allowable protein penetration penalty. From the 81-complex set above, a reasonable threshold would be −3.0 (80/81 ligands meet this threshold). However, those complexes were all native ligands complexed to the proteins, and so this value might be optimistic in docking non-native ligands. In the TK case, a threshold of −3.0 allowed for all 10 positives to dock successfully, and results are presented for this threshold. In the ER case, a threshold of −3.0 allowed for 7/10 known ligands to dock successfully. A threshold of −6.0 was required to allow for 9/10 known ligands, and a threshold of −12.0 was required to allow all 10 ligands to dock successfully (results are presented for −6.0 and −12.0).

Table 3 summarizes the false positive rates of Surflex, DOCK, FlexX, and GOLD at true positive rates ranging from 80% to 100%. For Surflex on the TK library, the ranks of the known ligands were 1, 5−7, 10, 13, 15, 17, 36, and 40. So, just 9/990 random molecules were among the 8 highest scoring known ligands, corresponding to a 0.9% false positive rate. This was roughly 10-fold better than the best result of the competing methods (GOLD's result). For a true positive rate of 100%, the Surflex FP rate was 3.2%, with the GOLD rate moving to 9.3%. DOCK performs the weakest of the four methods, with a 23% FP rate for a TP rate of 80%. In the TK case, the binding interactions involve a number of polar contacts with a relatively small component of hydrophobic packing compared with the ER case, which follows.

In the ER case, with the aggressive penetration threshold (−6.0, allowing for 9/10 known ligands to dock), Surflex yielded a FP rate of 0.2% for a TP rate of 80%. The ranks of the 9 successful known ER ligands

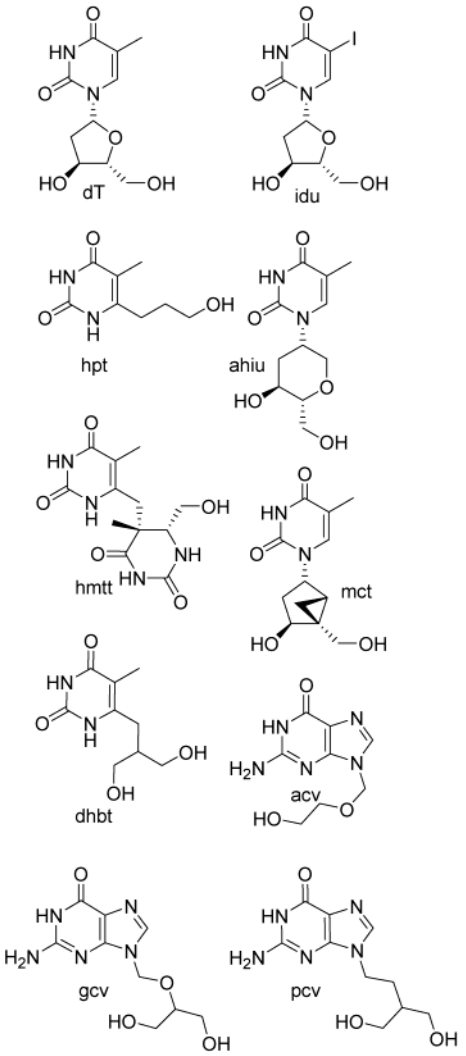

Figure 7. HSV-1 thymidine kinase ligands used as positive controls in screening. The abbreviations are as follows: dT, deoxythymidine; idu, 5-iododeoxyuridine; hpt, 6-(3-hydrodypropyl-thymine); ahiu, 5-iodouracil anhydrohexitol nucleoside; mct, (North) methanocarbathymidine; hmtt, (6-[6-hydroxymethy-5-methyl-2,4-dioxohexahydropyrimidin-5-yl-methyl]-5-methyl-1*H*-pyrimidin-2,4-dione; dhbt, 6-(3-hydroxy-2-hydroxymethylpropyl)-5-methyl-1*H*-pyrimidine-2,4-dione; acv, aciclovir; gcv, ganciclovir; pcv, penciclovir.

were 1−5, 7, 9−10, and 15. By use of a penetration threshold of −12.0, which accommodated all 10 ligands, the rate changed to 1.3%. Both the 0.2% rate and the 1.3% rate were significantly better than the best of the other three methods taken alone, which ranged from 5.3% for GOLD to 57.8% for FlexX. The combination of GOLD docking with DOCK scoring yielded poorer FP rates compared to Surflex under the more aggressive penetration threshold but were comparable at the 80% and 90% TP rates for the less aggressive threshold. However, for 100% coverage of the positives, Surflex yielded a 4-fold better FP rate than the hybrid GOLD/DOCK method. The issue of the penetration penalty deserves additional discussion. As noted above, a threshold of −3.0 was sufficient to admit dockings for 80/81 complexes in a diverse set of proteins. However, this threshold in the ER case rejected 3/10 known ER ligands. Of course, application of this more stringent threshold improved the FP rate at the cost of the additional false negative. The ranks of the seven surviving known ER ligands were 1−7 using the −3.0 penetration threshold, yielding a nominal false positive rate of 0%.

To probe this issue further and to address the question of docking accuracy to a hydrophobic bonding pocket that was not the result of cocrystallization with the target ligand, four ER ligands (minimized) whose bound structures were known were docked to the structure of ER bound to 4-hydroxytamoxifen (ligand abbreviation "tam", PDB code 3ERT). The bound poses of the four ligands were obtained by superimposing the coordinates of their respective complexes onto 3ERT based on the α carbons of the residues proximal to tam. Table 4 shows the results for these ligands plus tam for the top scoring pose, the top ranked pose with rmsd less than 2.0 Å, and the best pose (of the 10 returned) by rmsd. In four of five cases, the penetration penalty met the −3.0 threshold, so the initial 7/10 success rate is hardly changed (10/14 or 71%) with the additional four novel ligands (tam was in the original set of 10). In the case with the worst penetration penalty (chrys), docking the ligand to the native structure yielded an rmsd of 0.60 Å compared with 3.61 when docked to 3ERT. This ligand had a very accurately docked pose when docked to 3ERT that scored nominally higher than the top ranked pose (8.52 versus 8.03), but because of a significantly worse penetration penalty, it was ranked number 3. In all cases of docking to 3ERT, a pose within the top three yielded an rmsd less than 2.0 Å. Also, in all cases, the best pose by rmsd was very accurately docked (0.77 Å < rmsd < 1.34 Å). However, the docking of each of the four novel ligands to their native protein structures yielded better accuracies, and in the cases with significant penetration penalties (rx core and chrys), it also yielded much improved penetration values. For screening purposes, a penetration threshold of −3.0 appears to be appropriate provided that the tradeoff of some false negatives is warranted by the improved false positive rate. In practice, where possible, reliance on known positive ligands to quantify the expected false negative rate with various thresholds would be a better approach.

With respect to the scoring function, the ER and TK cases present very different binding pockets. In the ER case, the binding pocket is much more hydrophobic than in the TK case, and we see that the DOCK scoring function improved GOLD's docking performance. It appears that the scoring functions for GOLD, DOCK, and FlexX may have certain biases with respect to the types of proteins on which they perform well in terms of ranking the nominal binding affinities of ligands. The scoring function of Surflex may also have such limitations, but they were not revealed by these two data sets. With both TK and ER, Surflex yielded very low false positive rates compared to the other methods. If the results bear out on other proteins, assaying the top scoring 1000 out of 100 000 compounds (1%) should yield a large proportion of the true positives in the compound library.

Speed. Docking speed is a critical issue in screening large compound libraries and may be important even in a careful study of a small set of ligands. All of the dockings reported here were performed with the same

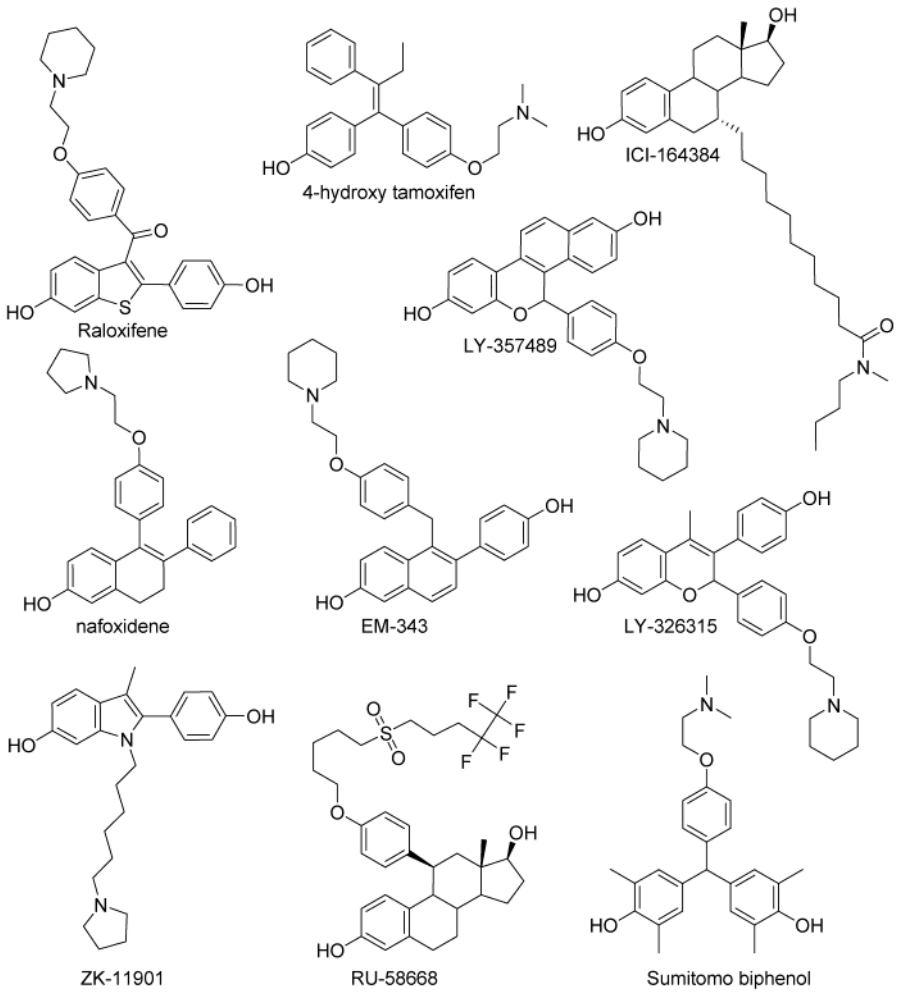

Figure 8. Estrogen receptor antagonists used as positive controls in screening.

Table 2. Thymidine Kinase Docking Accuracy[a]

	rmsd (Å) of best scoring poses			
ligand	Surflex	Dock	FlexX	GOLD
dT	0.74	0.82	0.78	0.72
ahiu	0.87	1.16	0.88	0.63
mct	0.87	7.56	1.11	1.19
dhbt	0.96	2.02	3.65	0.93
idu	1.05	9.33	1.03	0.77
hmtt	1.78	9.62	13.3	2.33
hpt	1.90	1.02	4.18	0.49
acv	3.51	3.08	2.71	2.74
gcv	3.54	3.01	6.07	3.11
pcv	3.84	4.1	5.96	3.01

[a] Data for Dock, FlexX and GOLD are taken from Bissantz et al.[13]

parameter set for Surflex. Figure 9 (left) shows a plot of mean docking time of the 81 ligands from the first data set versus number of rotatable bonds. Figure 9 (right) shows a plot of docking time of the 990 ligands for TK and ER versus number of rotatable bonds. The docking times for both cognate and random ligands were very similar. Docking time was roughly linear in number of rotatable bonds, beginning with a few seconds for rigid molecules and adding approximately 10 s per rotatable bond. The parametrization used for the docking runs was optimized for docking accuracy with cognate ligands. Consequently, Surflex returned dockings for all 990 of the TK random ligands and for 989 of the ER random ligands. Of course, many of these dockings scored very poorly or had impossibly high penetration values. Significant speedup for screening purposes is possible and was not attempted in these experiments.

Direct comparison of docking speed is somewhat problematic because of differences in hardware and methodology. The authors of GOLD reported a mean docking time of 14 600 s (total time for 20 runs), with a minimum of 3440 s on a set of 100 complexes (on SGI R4400 hardware).[11] They reported that two docking runs were sufficient to yield good answers in a large number of cases, which translates to 1460 s on average and 344 s minimum. A single run of Surflex yielded a good answer the majority of the time on the 81-complex data set (which shares 58 complexes with the published 100-complex GOLD set[11]). Surflex's mean time was 44 s over all 81, with the minimum being between 1 and 2 s. Clearly, even accounting for potential hardware differences, Surflex was much faster than GOLD with the settings and versions tested. However, speed optimization of GOLD has been an area that has received attention because the original validation set was published.

The more recent benchmarking work of Bissantz et al.[13] offers perhaps a more reasonable comparison, since multiple techniques were tested, including a more recent version of GOLD. For the ER and TK cases, Surflex's mean docking times for all 1000 molecules

Table 3. Comparative False Positive Rates in Screening[a]

	false positives from 990 random ligands, %									
	thymidine kinase				estrogen receptor					
TP, %	Surflex (pen. < 3)	DOCK	FlexX	GOLD	Surflex (pen. < 6)	Surflex (pen. < 12)	DOCK	FlexX	GOLD	GOLD/DOCK
80	0.9	23.4	8.8	8.3	0.2	1.3	13.3	57.8	5.3	1.2
90	2.8	25.5	13.3	9.1	0.7	1.6	17.4	70.9	8.3	1.5
100	3.2	27.0	19.4	9.3		2.9	18.9		23.4	12.1

[a] Data for Dock, FlexX, and GOLD are extrapolated from rank data and plots from Bissantz et al.[13]

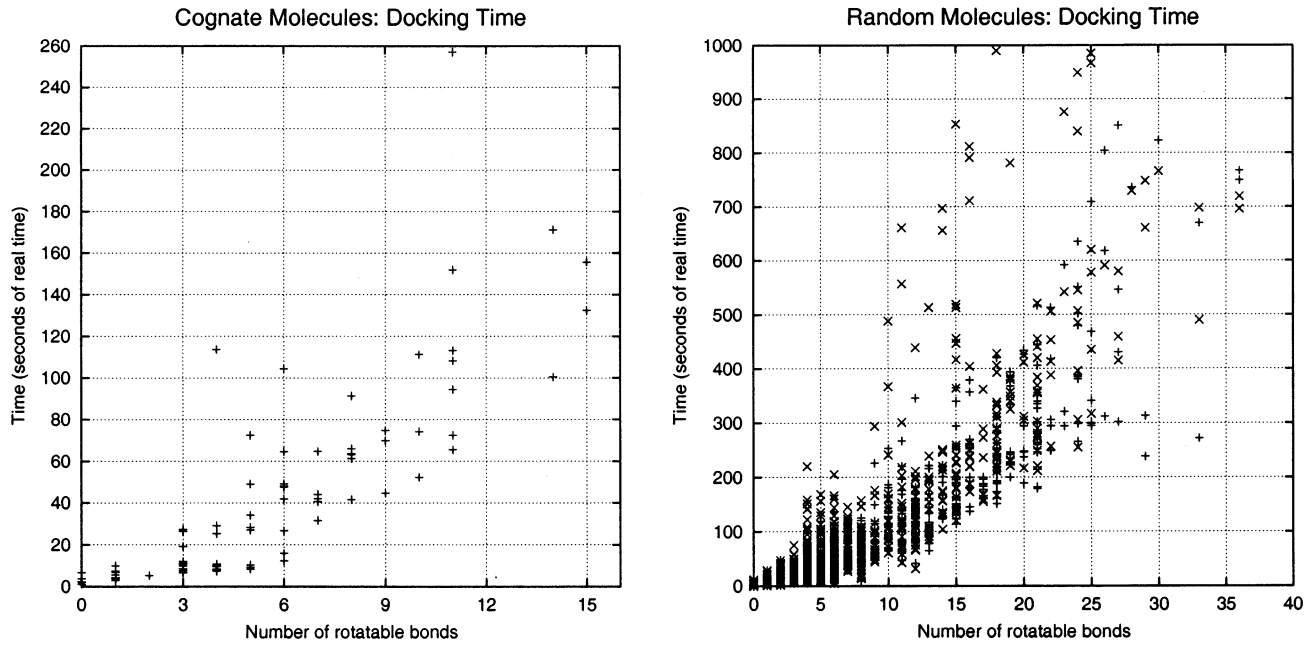

Figure 9. Surflex docking times versus number of rotatable bonds: (left plot) mean docking times for cognate ligands of 81 proteins; (right plot) docking times for 990 random molecules versus number of rotatable bonds for TK (+) and ER (×). Dockings were performed on a standard Windows 2000 Professional, 933 MHz Pentium III workstation.

Table 4. Estrogen Receptor Docking Accuracy[a]

structure:	3ERT	1GWR	1GWQ	3ERD	1L2I
ligand:	tam	rx-core	edl	des	chrys
docking target:	3ERT	3ERT	3ERT	3ERT	3ERT
	Top Ranked Pose				
rmsd	0.93	0.84	1.89	2.86	3.61
score	9.86	6.34	5.54	8.01	8.03
pen.	−0.45	−2.27	−0.63	−0.41	−4.28
	Top Pose with rmsd < 2.0				
rank	1	1	1	2	3
rmsd	0.93	0.84	1.89	1.74	0.898
score	9.86	6.34	5.54	6.85	8.52
pen.	−0.45	−2.27	−0.63	−0.55	−5.95
	Best Pose by rmsd				
rank	9	2	4	10	9
rmsd	0.826	0.83	1.34	1.11	0.77
score	9.03	6.25	5.18	6.53	7.03
pen.	−0.78	−2.2	−0.62	−2.49	−5.4
	Top Ranked Pose Docked to Native Structure				
rmsd	0.93	0.36	0.89	0.43	0.60
score	9.86	7.06	5.40	7.41	9.82
pen.	−0.45	−0.25	−0.51	−0.72	−1.18

[a] Values are rmsd (Å), score ($-\log(K_d)$), pen. (arbitrary units). Abbreviations: tam, 4-hydroxytamoxifen; rx-core, 2-(4-hydroxyphenyl)benzo[*b*]thiophen-6-ol); edl, 17β-estradiol; des, diethylstilbestrol; chrys, (*R*,*R*)-5,11-*cis*-diethyl-5,6,11,12-tetrahydrochrysene-2,8-diol.

were 124 and 93 s, respectively. The mean times for molecules with 20 or fewer rotatable bonds were 84 and 64 s, respectively. Unfortunately, detailed timing data were not published in the benchmarking study. The authors indicated that the docking pace was roughly between 50 and 100 s per molecule for FlexX, DOCK, and GOLD on an SGI Indigo2 R10K processor. It appears that under the parametrizations tested, Surflex had comparable docking times (note that GOLD was run with "library screening" settings that are more than 10-fold faster than the standard settings on which its docking accuracy has been validated).

Conclusions

Surflex represents an advance in flexible molecular docking. In comparison with the best methods in each category, Surflex is simultaneously as accurate in terms of rmsd of docked ligands, as fast in terms of docking speed, and *significantly* more accurate in terms of scoring to the extent that false positive rates are 5- to 10-fold lower for equivalent true positive rates compared to other methods. A diversity of protein active sites are tractable with a single approach using a single parametrization. However, there are areas for improvement in both its scoring function and search methodology. With respect to scoring, three improvements would be beneficial: (1) the scoring and penetration terms should be consolidated into a single score and the parameters should be re-estimated; (2) the scoring function parameters re-estimation should include explicit training on negative examples (nonbinding ligands), which should further reduce false positive rates; (3) the effect of nonbonded self-interactions within ligands should be

accounted for explicitly. With respect to search methodology, the most significant area for further development is in explicitly allowing a degree of protein flexibility (e.g., side chain movement). Apart from that, a number of incremental speed improvements are possible (e.g., more efficient gradient-based pose optimization) in addition to development of a specific parametrization for library screening.

Acknowledgment. The author is grateful to Didier Rognan for providing electronic data sets for validation.

References

(1) Walters, P. W.; Stahl, M. T.; Murcko, M. A. Virtual Screening—An Overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(2) Jain, A. N. Scoring noncovalent protein−ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427−440.
(3) Bohm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(4) 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.
(5) Rognan, D.; Lauemoller, S. L.; Holm, A.; Buus, S.; Tschinke, V. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.* **1999**, *42*, 4650−4658.
(6) Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379−384.
(7) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(8) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(9) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449−462.
(10) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.* **1996**, *9*, 1−5.
(11) 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.
(12) 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.
(13) 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.
(14) Jain, A. N. Morphological similarity: a 3D molecular similarity method correlated with protein−ligand recognition. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 199−213.
(15) Ruppert, J.; Welch, W.; Jain, A. N. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* **1997**, *6*, 524−533.
(16) Pitman, M. C.; Huber, W. K.; Horn, H.; Kramer, A.; Rice, J. E.; et al. FLASHFLOOD: a 3D field-based similarity search and alignment method for flexible molecules. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 587−612.

JM020406H