PROTEINS: Structure, Function, and Bioinformatics 57:225–242 (2004)

Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy

Esther Kellenberger, Jordi Rodrigo, Pascal Muller, and Didier Rognan*
Bioinformatics Group, Laboratoire de Pharmacochimie de la Communication Cellulaire, CNRS UMR7081 Illkirch, France

ABSTRACT **Eight docking programs (DOCK, FLEXX, FRED, GLIDE, GOLD, SLIDE, SURFLEX, and QXP) that can be used for either single-ligand docking or database screening have been compared for their propensity to recover the X-ray pose of 100 small-molecular-weight ligands, and for their capacity to discriminate known inhibitors of an enzyme (thymidine kinase) from randomly chosen "drug-like" molecules. Interestingly, both properties are found to be correlated, since the tools showing the best docking accuracy (GLIDE, GOLD, and SURFLEX) are also the most successful in ranking known inhibitors in a virtual screening experiment. Moreover, the current study pinpoints some physicochemical descriptors of either the ligand or its cognate protein-binding site that generally lead to docking/scoring inaccuracies. Proteins 2004;57:225–242.**

Key words: docking; scoring; virtual screening; PDB; drug design

INTRODUCTION

Docking small-molecular-weight ligands to therapeutically relevant macromolecules has become a major computational method for predicting protein–ligand interactions and guide lead optimization.[1] From the pioneering work of Kuntz et al.,[2] numerous docking programs based on very different physicochemical approximations have been reported.[3,4] Since any docking tool needs to combine a docking engine with a fast-scoring function, the recent literature is full of benchmarks addressing three possible issues: (1) the capability of a docking algorithm to reproduce the X-ray pose of selected small-molecular-weight ligands[5–24]; (2) the propensity of fast-scoring functions to predict binding free energies from the best-scored pose[10,24–34]; and (3) the discrimination of known binders from randomly chosen molecules in virtual screening experiments.[19,20,23–26,29,35–37] However, analyzing all these data for a comparative analysis of available docking tools is very difficult. First, many tools are not available. Second, independent studies assessing the relative performance of docking algorithms/scoring functions are still rare[26,35,36] and focus on the use of few methods. Third, the quality judgment may vary depending on the examined properties (quality of the top-ranked pose, quality of all plausible poses, binding free energy prediction, and virtual screening utility). Fourth, most docking programs assume approximation levels that can vary considerably[3] and lead, for example, to very inhomogeneous docking paces ranging from few seconds to few hours. Last, many docking programs have been calibrated and validated on small protein–ligand data sets. Detailed benchmarks (>100 Protein Data Bank (PDB)–ligand complexes) are only reported for a few docking tools.[12,14,16,18,24,38]

The purpose of the current study is to provide independent benchmarks for widely used docking programs. As it would be almost impossible to consider all of them, docking programs were selected based on three criteria: (1) availability, (2) use of conventional file formats as input (e.g., pdb, sdf, mol2), and (3) easy application to virtual screening (database docking). Eight tools were finally selected (DOCK,[15] FLEXX,[39] FRED (Open Eye Scientific Software; Santa Fe, NM), GLIDE (L. Schrödinger, Portland, OR), GOLD,[24] SLIDE,[19] SURFLEX,[20] and QXP[8]) and examined under comparable conditions (see Material and Methods section) for their ability to reproduce the X-ray pose of 100 small-molecular-weight ligands[18] and their capability to discriminate by protein-based virtual screening true inhibitors of an enzyme of known X-ray structure (thymidine kinase) from randomly chosen "drug-like" molecules. For achieving a fair comparison, all programs were here considered using settings allowing fast screening (<4 min/ligand).

MATERIAL AND METHODS

Docking

Setting up a data set of 100 protein–ligand complexes

The crystal structure of 100 protein–ligand complexes[18] from the PDB[40] were used to generate a separate set of coordinates for the whole protein, its ligand, and the corresponding active site. Unless specified below, the input conformation of the ligand was directly extracted from the X-ray structure. No energy minimization of the ligand was performed. The protein active site was defined as the collection of amino acids for which at least one atom is nearer than 6.5 Å to any nonhydrogen atom of the bound ligand. Important metal ions and cofactors were included

*Correspondence to: Didier Rognan, Bioinformatics Group, Laboratoire de Pharmacochimie de la Communication Cellulaire, CNRS UMR7081, F-67400 Illkirch. E-mail: didier.rognan@pharma.u-strasbg.fr

Received 18 September 2003; Accepted 10 February 2004

Published online 10 August 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20149

in binding sites. Except for 5 entries (1aaq, 1dbj, 1dwd, 1lna, and 4phv) all crystallographic water molecules were removed from the active site. Hydrogen atoms were added using SYBYL 6.9 (TRIPOS Associates; St. Louis, MO) standard geometries. In order to accommodate the input requirement of the 8 docking programs used, coordinates were saved in several formats; protein files were stored in pdb,[40] mol2 (TRIPOS Associates), and mae (L. Schrödinger) formats; ligand files were saved in pdb, mol2, mae, and sd (MDL Information Systems; San Leandro, CA) formats. File format conversions were achieved using UNITY4.4 (TRIPOS Associates), except for the mae file format, which was obtained using Schrödinger's Maestro interface.

DOCK4.0, FLEXX1.11, and GOLD2.0 docking

DOCK, FLEXX, and GOLD calculations were performed as previously described.[18] For each ligand, the 30 best scored poses were kept.

FRED1.1 docking

FRED requires a set of input conformers for each ligand. The conformers were generated by OMEGA (Open Eye Scientific Software) and stored in a single binary file. Modifications applied to the default settings of OMEGA were the following: The maximum number of output conformers was set to 500 (GP_NUM_OUTPUT_CONFS); the upper bound relative to the global minimum was set to 3 kcal/mol (GP_ENERGY WINDOW); and the root-mean-square deviation (RMSD) value below which two conformations are considered to be the similar was set to 0.8 Å (GP_RMS_CUTOFF). In addition, the maximum number of rotatable bonds in the molecule was raised to 25 (GP_MAX_ROTORS) in order to generate conformers for all ligands of our data set (the most flexible ligand contains 24 rotatable bonds as defined by OMEGA).

FRED docking roughly consists of 2 steps: shape fitting and optimization. During shape fitting, the ligand is placed into a 0.5-Å-resolution grid box encompassing all active-site atoms (including hydrogens) using a smooth Gaussian potential.[21] A series of three optimization filters is then processed, which consists of (1) refining the position of hydroxyl hydrogen atoms of the ligand, (2) rigid body optimization, and (3) optimization of the ligand pose in the dihedral angle space. In the optimization step, 4 scoring functions are available: Gaussian shape scoring,[21] ChemScore,[41] PLP,[5] and ScreenScore.[29] Preliminary docking trials induced us to select ChemScore for the 3 optimization filters.

GLIDE2.0 docking

GLIDE calculations were performed with Impact version v2.0 (L. Schrödinger). The grid generation step requires mae input files of both ligand and active site, including hydrogen atoms. The protein charged groups that were neither located in the ligand-binding pocket nor involved in salt bridges were neutralized using the Schrödinger pprep script. The center of the grid enclosing box was defined by the center of the bound ligand as described in the original PDB entry. The enclosing box dimensions, which are automatically deduced from the ligand size, fit the entire active site. For the docking step, the size of bounding box for placing the ligand center was set to 12 Å. A scaling factor of 0.9 was applied to van der Waals radii of ligand atoms.[42] No further modifications were applied to the default settings. The GlideScore scoring function was used to select 30 poses for each ligand.

SLIDE2.0 docking

SLIDE looks for chemical and geometrical similarity between the ligand and a binding-site template that defines points for favorable interactions with the protein surface atoms.[19,43] It handles the mol2 file of the ligand and the PDB file of the target active site (without hydrogens). Note that none of the active-site residues were truncated, thereby allowing side-chain rotations for induced-fit modeling. Unbiased construction of the template in the dense mode of interaction points generation (Grid_spacing 0.5 Å, Hbonding_point_density dense and Clustering_threshold = 3 Å) yields about 100–150 points. The SlideScore empirical function[19] was used to generate 30 solutions for each docked ligand.

SURFLEX1.1 docking

The SURFLEX docking algorithm[20] uses an idealized active site called a protomol.[6,44] The protomol was built from the hydrogen-containing protein mol2 file. The construction was based on protein residues that constitute the active site (see above definition) using parameters tuned to produced a small and buried docking target (proto_thresh = 0.5 and proto_bloat = 0). Docking of the ligand in sd format was run using the "whole molecule" approach, a maximum of 100 conformations per fragment in each stage of the incremental construction process, and default settings for all other parameters. Thirty poses were finally saved for each ligand.

QXP docking

Docking was carried out using the MCDOCK conformational searching/energy minimization procedure of QXP (flo01.11S[8]). Protein and ligand PDB coordinates were used for single docking. Between 4 and 7 amino acids (depending on the protein) lining the active site were marked to define the binding cavity. Free movements of both ligand- and protein-marked residues were allowed. The ligand was subjected to 30 cycles of Monte Carlo conformational searching and energy minimization. For each ligand, the 25 lowest energy conformations were finally saved.

RMSD calculations

The RMSDs have been calculated over heavy atoms of 2 different conformations of a same molecule. In order not to overestimate RMSD values, symmetry operators have been included in the calculation routine to interchange equivalent atoms belonging to symmetrical groups (e.g., carboxylate or phosphate oxygens).[18] From here on, the "best pose" is defined as the docking solution that is the

nearest to the experimental binding mode, whereas the "top pose" is defined as the docking solution that is ranked first.

Virtual Screening

Compounds library

The library comprises 10 known inhibitors of the HSV-1 thymidine kinase (TK), as well as 990 "drug-like" molecules, as previously described.[26] Three-dimensional (3D) input coordinates of the 1000 ligands were generated from a 2D sd file using Concord (TRIPOS Associates) and stored in mol2 format. For GLIDE, the compound library was converted to Schrödinger's mae format using Maestro without energy refinement. For QXP, the multimol2 file was converted into a 3D sd format using UNITY (TRIPOS Associates). Last, a multiconformer OMEGA library was produced for FRED screening, as previously described.

Protein target

Protein target coordinates were extracted from the PDB entry corresponding to the crystal structure of the TK enzyme in complex with deoxythymidine (PDB code: 1kim). As previously described,[26] all water molecules were removed from the active site. The active site (see above definition) contains 16 amino acids. File modifications necessary for the 8 docking programs were performed as described in the previous section.

Docking parameters

The docking parameters used were those specified above. It should be noted that the SURFLEX output (top-scoring poses) was postprocessed to filter out ligands showing a protein penetration value (pen score) above a defined threshold (-3 or -6, see text).

RESULTS

Eight docking programs, namely, DOCK, FLEXX, FRED, GLIDE, GOLD, SLIDE, SURFLEX, and QXP, have been extensively tested in order to evaluate their potency in drug discovery applications. Docking results are discussed in the light of the 3 major issues in the application of docking programs to virtual screening: docking accuracy, ranking accuracy, and speed. These criteria were assessed on a data set of 100 diverse protein–ligand complexes from the PDB. The ability of the scoring function to reliably rank potential hits was also evaluated by screening against the TK target a library containing 10 true TK inhibitors and 990 randomly chosen drug-like compounds.

Docking and Ranking Accuracy

Data set description

The X-ray structure of 100 selected protein–ligand complexes follow the 2 following criteria: high resolution (<3 Å) and well-defined atomic positions of the ligand (temperature factors below 40 Å^2). Thus, the X-ray pose is considered as the bioactive one and can be further used as the reference to evaluate computed conformations. A total of 94 proteins and 97 ligands compose the data set. The chemical diversity of ligands has been previously assessed[18] by the analysis of several physicochemical descriptors (molecular weight, number of rotatable bonds, number of H-bond donors and acceptors, MlogP, and polar surface area). By examining the descriptors, it appears that almost all ligands are "drug-like."[45] Ligands have a molecular weight ranging from 88 to 730 and occupy a protein cavity whose volume is comprised between 300 and 3510 Å^3. Plotting the molecular weight of the ligand versus the volume of the corresponding protein cavity clearly shows that these two features are mutually related [Fig. 1(A)]. Similarly, an obvious correlation is observed between ligand size and conformational freedom [Fig. 1(B)]. Both the size of the protein and the conformational flexibility of the ligand will be further used to classify binding sites and discuss docking performances in that respect (see below). Since docking is usually based on not only shape complementary but also favorable ligand–receptor interactions, docking results are also examined with regard to the polar surface area of the ligand. Special attention will be paid to hydrophobic ligands, as well as very polar ligands [Fig. 1(C)].

Docking accuracy

For the 8 docking tools, we identified the best pose out of 30 possible solutions. The ability to predict the correct binding of a ligand into its active site was thus evaluated by comparing the best pose and the experimentally determined solution. Docking was considered successful if the best pose was closer than a given threshold from the X-ray solution. Figure 2(A) shows that at a 1 Å RMSD cutoff, docking is successful for 61–63% of the cases using GLIDE, GOLD, or QXP. At this cutoff, FLEXX and SURFLEX only achieve successful docking in 48% and 54% of the cases, respectively. DOCK, FRED, and SLIDE perform significantly worse, exhibiting success rates of 29–38%. Similar discrepancies occur up to an RMSD threshold of 2 Å, which is usually considered the upper limit for drug discovery.[18] In a first group, GLIDE, GOLD, SURFLEX, and QXP placed about 80–90% of the ligands of the data set within 2.0 Å of the X-ray pose. FLEXX and FRED were considered successful in 66% and 62% of the cases, respectively. Last, DOCK and SLIDE performed poorly in locating only about 50% of the ligands within the 2 Å RMSD threshold.

Ranking accuracy

The performance of the 8 programs on the 100-complex data set in ranking the diverse poses of a single ligand in the target binding site was estimated by comparing the top-ranked pose with the experimentally determined solution [Fig. 2(B)]. Again, GLIDE, GOLD, and SURFLEX clearly outperformed DOCK, FRED, and SLIDE. FLEXX performance is comparable to that of the three former programs, whereas for QXP, performance is significantly worse in ranking than in docking. At a 2 Å RMSD cutoff, FLEXX, GLIDE, GOLD, and SURFLEX succeed in docking about 50–55% of the ligands, whereas the success rate of DOCK, FRED, SLIDE, and QXP does not exceed 40%. The latter inspection is, however, unable to identify the cause of ranking inaccuracy (insufficient conformational

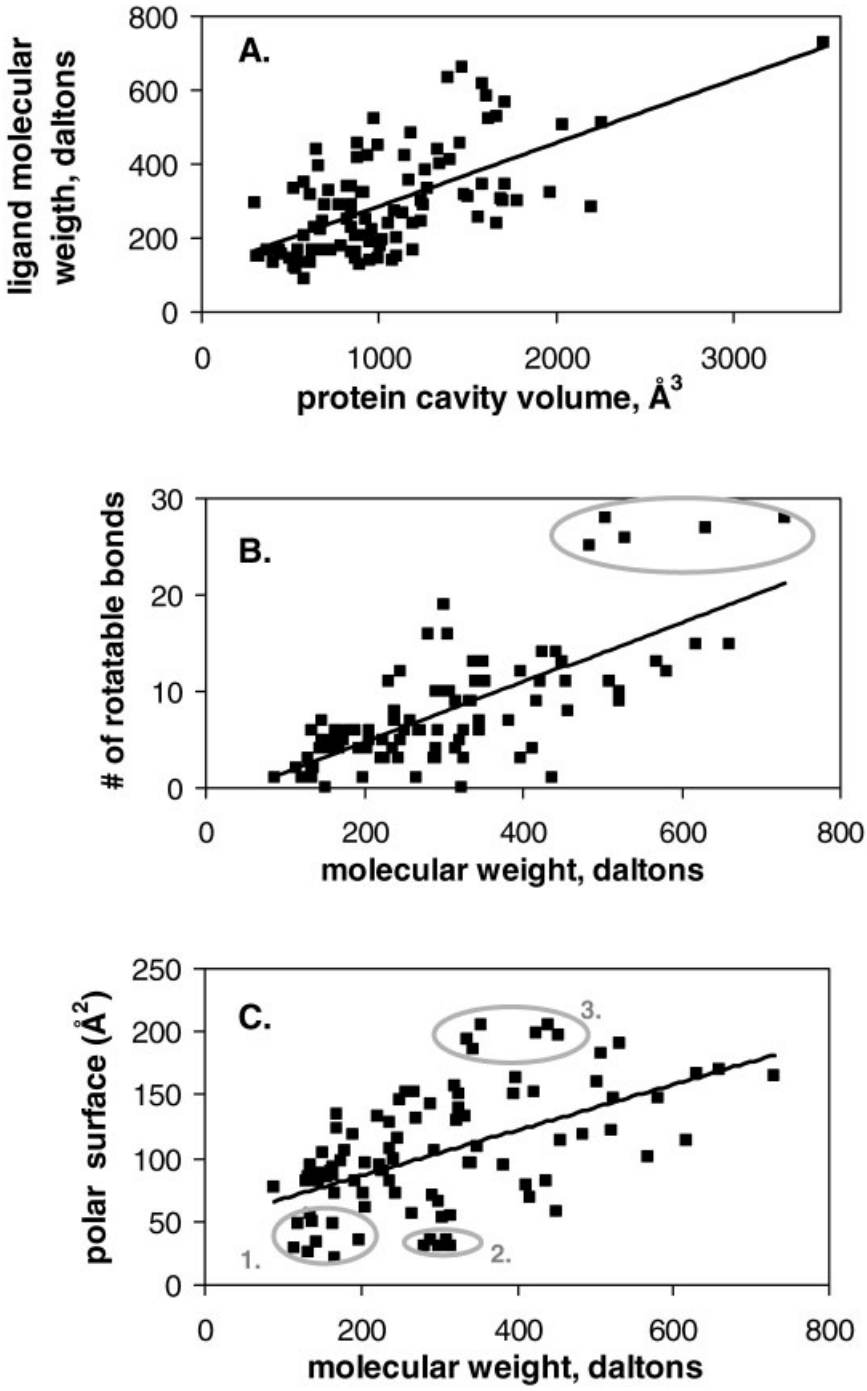


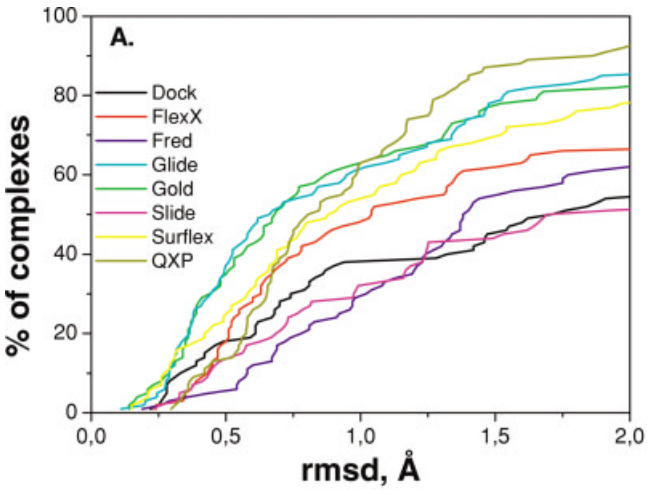
Fig. 1. Description of the PDB data set (100 high-resolution protein–ligand complexes). Linear fit curves are displayed on the three graphs. (**A**) Protein cavity volume versus ligand molecular weight. The volume of the binding site cavity was determined using SURFNET[46] with a grid resolution of 0.8 Å and radius for gap spheres between 1 and 4 Å. (**B**) Molecular weight versus number of rotatable bonds, counted according to SYBYL, of the ligands. Five highly flexible ligands (PDB codes: 1aaq, 1apt, 1eed, 1poc, and 1rne) are enclosed. (**C**) Molecular weight versus polar surface area of the ligands. Small hydrophobic, medium-sized hydrophobic, and very polar ligands are enclosed in classes 1, 2, and 3, respectively.

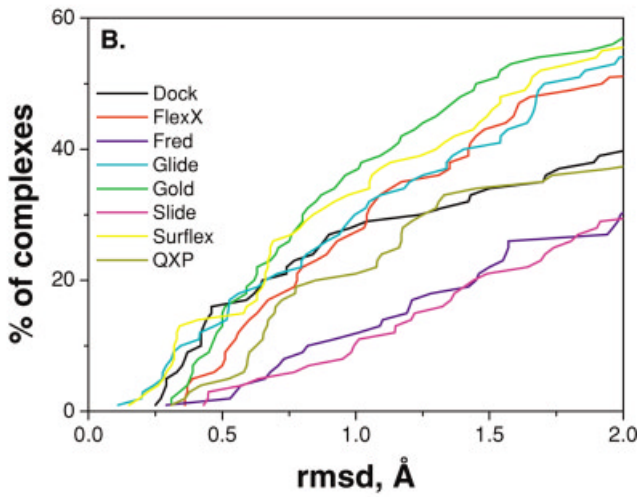
sampling, inaccurate scoring function). To identify true scoring failures, we next looked at the percentage of PDB entries for which docking predictions within 2 Å from the X-ray pose exist but are never ranked first [Fig. 2(C)]. True scoring errors were mainly observed using QXP, FRED, and to a lesser extent SLIDE, whose scoring functions have clear difficulties discriminating accurate from inexact poses. For other tools, only 25–30% of correctly predicted poses were inaccurately ranked [Fig. 2(C)].

Docking/scoring performance as a function of ligand input coordinates for FRED, GLIDE, SLIDE, and QXP

DOCK, GOLD, and FLEXX are widely used docking programs whose performance has already been extensively tested.[7,15,18,24,39] Benchmarks recently became available for SURFLEX.[20] Its docking and pose recognition accuracy were evaluated on 81 protein–ligand complexes using 10 different ligand-input conformations. Results are comparable to those of the present study, suggesting that

the ligand-input conformation has little effect on docking success when using this program. Therefore, the current analysis only focuses on docking tools (FRED, GLIDE, SLIDE, and QXP) for which too few data are available to unambiguously assess docking performance as a function of ligand-input coordinates.





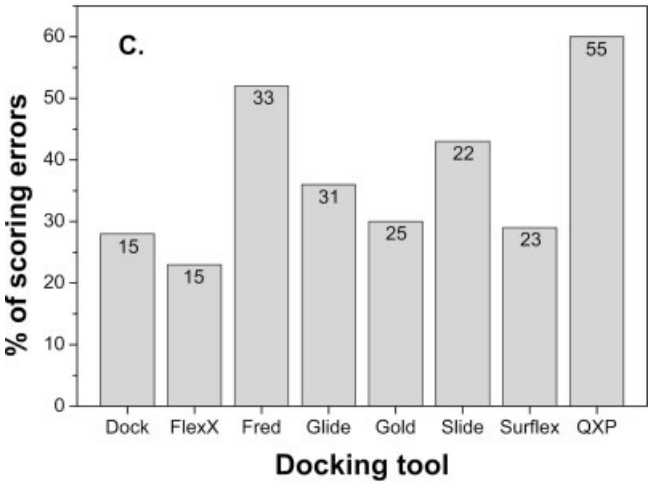


Fig. 2. Docking of 100 ligands to their cognate protein X-ray structure. Cumulative percentage of complexes as a function of the RMSD from the X-ray pose. (**A**) Docking accuracy: RMSD in Å of the best pose (nearest to the experimental binding mode) from the experimental solution. (**B**) Scoring accuracy: RMSD in Å of the top pose (best scored solution) from the experimental solution. Current plots have been obtained considering the X-ray pose as input conformation of the ligand to dock. Using an arbitrary input ligand conformation produces quite substantial different results for SLIDE and QXP (see Fig. 3). (**C**) True scoring failures: percentage of PDB entries for which docking predictions within 2 Å from the X-ray pose exist but are never ranked first are plotted for different docking programs. Numbers in bars indicate the absolute number of PDB entries for which a scoring failure has been detected.

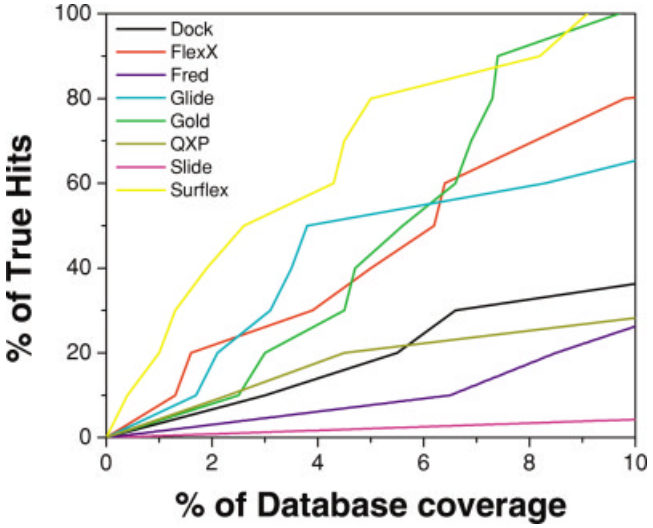


Fig. 4. Enrichment in thymidine kinase inhibitors. The cumulative percentage of known actives recovered by virtual screening is indicated as a function of the top-scoring fraction of the database selected for generating a hit list.

FRED rigid docking requires a preselection of suitable conformers for each ligand. The conformational ensemble was herein generated by OMEGA. OMEGA sampling has previously been shown to select a conformation similar to that of the X-ray input when using appropriate parameters[47] (a low-energy cutoff to discard high energy conformations, a low RMSD value below which two conformations are considered to be similar, and a maximum of 500–1000 output conformations). Indeed, using the X-ray structure of the ligand as input, OMEGA was able to propose at least one conformation closer than 2 Å from the protein-bound X-ray conformation for 99% of our 100 ligands. At lower thresholds (1.0 and 0.5 Å), the percentage of ligands for which at least one bioactive-like conformation has been generated decreases to 77% and 39%, respectively. To check whether OMEGA conformers are adequate for efficient FRED docking, 2 independent runs were carried out: (1) docking a conformational ensemble generated from X-ray coordinates, and (2) docking a conformation ensemble generated from a randomly defined conformation. Docking performances of both runs are quite comparable [Fig. 3(A,B)], suggesting that the OMEGA–FRED combination is insensitive to the input conformation of the ligand.

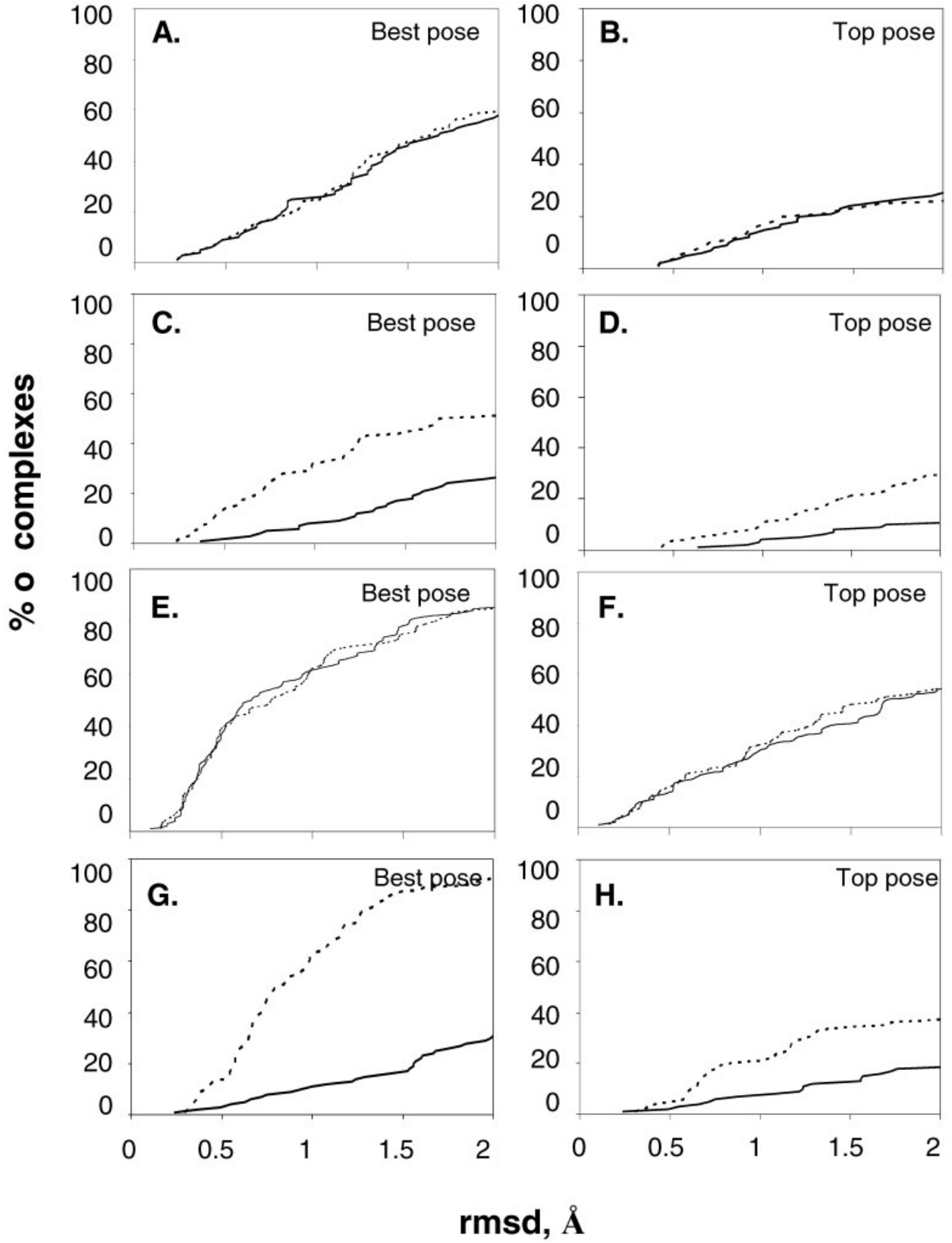


Fig. 3. Sensitivity of docking tools to ligand input coordinates. Cumulative percentage of complexes as a function of the RMSD from the X-ray pose for docking runs using as input either the X-ray pose (dotted line) or a randomly chosen ligand conformation obtained using the TRIPOS randconf script: (solid line) with FRED combined with OMEGA (**A,B**), SLIDE (**C,D**), GLIDE (**E,F**), and QXP (G,H). For QXP, the randomly chosen conformation was replaced by an altered X-ray pose (rotation by 180° along its main axis).

GLIDE and SLIDE sample conformational space of the ligand during docking. For this purpose, an incremental construction method is implemented in both docking algorithms. Nevertheless, the approaches used to treat positions and conformations of core regions and end "rotamer groups" differ. In GLIDE, core fragment poses are first clustered and then anchored into the binding site. Then, the position and orientation of end "rotamer groups" are selected by a series of topological and energetical filters. The best binding mode is finally identified by thoroughly exploring the active site using flexible ligand minimization and Monte Carlo sampling. In SLIDE,[19,43] molecular fragments are reassembled once the core fragment has been anchored in the active site, using the geometry of the input conformation. Rotatable bonds of the ligand and of protein side-chains are modified whether or not intermolecular bumps occur. The dependence of GLIDE and SLIDE docking performance on the ligand input coordinates was evaluated by comparing, for each program, results of two different runs: (1) starting from the X-ray conformation, and (2) starting from a randomly generated conformation. As shown in Figure 3(C, D), SLIDE was found to be very sensitive to the ligand input conformation, as its docking accuracy significantly decreases when starting from a randomly defined conformation. By contrast, changing the starting coordinates of the ligand does not alter the overall performance of GLIDE [Fig. 3(E,F)].

To simulate ligand flexibility, QXP uses random Monte Carlo (MC) moves on dihedral space combined with energy minimization during the template fitting procedure.[8] In order to achieve a docking pace comparable to that of other tools, the number of MC cycles was herein restricted to 30. Starting from X-ray coordinates of the ligand, QXP predicts a top and best pose within 2 Å RMSD from the experimental pose in 37% and 92% of cases, respectively. However, a simple rotation of the X-ray pose by 180° along its main axis to generate alternative input coordinates yields much poorer results [Fig. 3(G, H)], demonstrating that QXP is very sensitive to the input conformation of the ligand.

Virtual Screening of a Compound Database

The virtual screening utility of 8 docking tools was tested by using the X-ray structure of HSV-1 TK as a target and a database of 990 randomly chosen drug-like molecules seeded with 10 known TK inhibitors. TK is a difficult target for protein-based virtual screening because its binding site is delimited by a very flexible loop amenable to induced-fit upon ligand binding.[48] Conformations of few side-chains vary depending on the bound ligand, and in some cases, water molecules are involved in protein–ligand interactions. Moreover, the affinity of most ligands for the target is low (in the micromolar range). The TK conformation used for screening was that of the crystal structure of the deoxythymidine-bound enzyme (PDB code: 1kim).

The percentage of true hits retrieved in increasing fractions of the starting database whose compounds have been ranked by the docking score is displayed in Figure 4 (see also Table I). Considering the top 5% of binders, programs may be classified as follows: (1) the very efficient one, SURFLEX (8 true hits among 50 compounds); (2) the good ones, namely, GLIDE, GOLD and FLEXX (5, 4, and 4 true hits, respectively); and (3) the less efficient ones, namely, QXP, DOCK, FRED, and SLIDE (2, 1, 0, and 0 true hits, respectively). Retrieving all TK inhibitors embedded in the full database would require the selection of the top 10% of scorers for both GOLD and SURFLEX, and even more for FLEXX and GLIDE (20–27% of top scorers, respectively). Using a hit list generated from the predicted top 5% of scorers, which seems reasonable with respect to recent virtual screening reports,[49,50] FLEXX, GLIDE, GOLD, and especially SURFLEX are thus the methods of choice for the current target.

Out of the 10 true TK ligands seeded in the library,[26] 7 are pyrimidine analogs, whereas 3 compounds share a larger purine scaffold. Among the 4 most successful docking tools, GOLD and FLEXX provide better ranks for pyrimidine derivatives. The opposite propensity is observed for GLIDE and to a small extent for SURFLEX. Addressing the question of docking accuracy, it appears that pyrimidines are generally better docked than purines, suggesting that scores assigned to larger molecules (here,

TABLE I. Description of Hit Lists generated by 8 Docking Tools on the Thymidine Kinase Example
A hit list is generated from the top-scoring compounds selected at a given threshold.

	Top 2.5 %		Top 5%		Top 10%	
	Hit Rate[a]	Yield[b]	Hit Rate	Yield	Hit Rate	Yield
DOCK	0	0	2	10	3	30
FLEXX	8	20	8	40	8	80
FRED	0	0	0	0	2	20
GLIDE	8	20	10	50	6	60
GOLD	4	10	8	40	10	100
SLIDE	0	0	0	0	0	0
SURFLEX[c]	16	40	16	80	10	100
QXP	0	0	4	20	2	20

[a]Hit rate: (AH/TH) × 100
[b]Yield: (AH/A) × 100, where TH is the total number of compounds in the hit list, AH the number of true hits in the hit list, and A the total number of true hits in the library.
[c]Figures reported for SURFLEX were obtained by using a protein penetration threshold value of −6.

TABLE II. Docking Times for 8 Programs (Single-Processor Docking Time in Seconds, on a 270 MHz SGI R12K Processor Running IRIX6.5)

Program	Average[a]	Minimum[b]	Maximum[c]
FRED[d]	18	0.1	193
DOCK	46	1	667
FLEXX	67	2	595
QXP	108	37	378
SLIDE	118	1	1743
SURFLEX	135	9	1460
GOLD	137	55	479
GLIDE	234	9	2825

[a]CPU time averaged over our data set of 100 PDB entries.
[b]Smallest observed CPU time in the PDB data set.
[c]Highest observed CPU time in the PDB data set.
[d]Includes CPU time required by OMEGA for generating conformers of the ligand to dock.

purines) by GLIDE and SURFLEX tend to be overestimated (data not shown).

Docking Times

Although we tried to set up docking parameters in order to achieve a fair comparison of docking programs, especially in terms of docking speed, analyzing the average docking times for our 100-ligand data set clearly indicates significant differences (Table II). The faster docking tool by far is FRED (mean docking time of 18 s). Thus, this tool is particularly attractive for ultrahigh-throughput docking (>1 million compounds). DOCK and FLEXX are also remarkably fast, since they are able to dock a ligand in less than 1 min. SLIDE, QXP, SURFLEX, and DOCK can be considered equally fast for the present data set (mean docking time about 2 min). However, even within this group, the range of observed CPU times varies considerably. GOLD achieves a remarkable narrow distribution of CPU docking times (55–479 s) and is much less sensitive to the ligand flexibility (which is the main factor influencing the docking pace) than SURFLEX (CPU times from 9 s to 1460 s), whose speed is much more dependent on the number of rotatable bonds of the ligand to dock. Last, GLIDE was significantly slower than other docking tools, with an average docking time of about 4 min.

Thus, the most accurate docking tools (SURFLEX, GLIDE, and GOLD) identified in the current study are also the slowest. It should be noted that some programs (e.g., DOCK, GLIDE) first require significant CPU time to compute grid energy values. These timings were not considered here, although precomputing grid energies is known to hasten scoring in the docking step.

DISCUSSION

Comparison With Published Benchmarks

To check whether the current settings of the described docking tools have dramatically altered their accuracy, we first compared on common PDB entries our results with previously described data. We already discussed in a previous report[18] that current docking data are very similar to existing benchmarks for DOCK, FLEXX, and GOLD. The comparison thus focuses on the other five tools.

Schulz-Gasch and Stahl[42] have analyzed 7 different targets using FRED (in combination with OMEGA), FLEXX, and GLIDE over a data set of 7528 noise compounds combined with data sets containing multiple active compounds (from 36 to 128). Since all screenings gave satisfactory hit rates, the authors concluded that both GLIDE and FRED are efficient docking tools, with FRED being especially attractive considering its high speed. Among the 7 tested proteins is thrombin (PDB code: 1dwd), which has also been used in the present study to assess docking and scoring accuracy. In the above-mentioned study, virtual screening against thrombin using FRED in combination with the ChemScore scoring function was able to recover 55% and 75% of known actives among the top 5% and 10% of scorers, respectively. Similar values were obtained when using GLIDE in combination with GlideScore. Though docking accuracy was not thoroughly investigated, the authors have verified that poses were "reasonably predicted." Results of our docking trials of NAPAP [*N*-α-(2-napthyl-sulfonyl-glyceryl)-para-aminoalanyl-piperidine] into α-thrombin using FRED and GLIDE are reported in Table III. Whatever the ligand input conformation, FRED gives satisfactory results, with RMSDs of docked poses from X-ray conformation between 1.76 and 2.14 Å. Using GLIDE, poses close to the experimentally determined solution could be found, yet the program had difficulties ranking one of them as a top-scoring solution, especially in terms of whether the ligand input conformation differs from the X-ray conformation.

Thrombin was also chosen as test case by SLIDE's authors[19] in docking known cocrystallized ligands into the apostructure of the enzyme and in virtual screening of a library containing 15,000 random compounds and known thrombin inhibitors. The thrombin-binding site, which consists of a narrow lipophilic pocket with an aspartic acid (Asp189) at its floor, was represented by 100–150 interacting points generated using an unbiased approach. Points within 5 Å of carboxylic oxygens of Asp189 were selected as key points, so that any docking must include a match to one of these points, in order to ensure that docked molecules will at least partially occupy the targeted site. Thirty-six of the 42 thrombin ligands were successfully docked into the apostructure, provided that both ligand and protein flexibility were modeled.[19] More especially, docking NAPAP (PDB code: 1dwd) generated a best pose, with a 0.44 Å RMSD from the X-ray solution. In the present study, docking of NAPAP into thrombin was performed using an unbiased template with no key points selected. It also gave satisfactory results when using X-ray coordinates as input (Table III). However, Slide could not correctly pose NAPAP starting from a random conformation. The virtual screening against thrombin published by Zavodszky et al.[19] produced high hit rates (about 64% of known ligands among the top-scoring 6.7% of molecules). These data have to be interpreted with care, since input coordinates for thrombin ligands were taken directly from thrombin–ligand complexes available in the PDB. Our

TABLE III. Docking of NAPAP Into α-Thrombin Using FRED, GLIDE, and SLIDE

OMEGA parameters	Schulz-Gash and Stahl[42]	Present Study	
GP_ENERGY WINDOW	5	3	
GP_RMS_CUTOFF	0.8	0.8	
GP_NUM_OUTPUT_CONFS	400	500	
FRED docking	**Schulz-Gash and Stahl[42]**	**Present Study**	
Input conformation	Random	X-ray	Random
Scoring function	ChemScore, PLP, or ScreenScore	ChemScore	ChemScore
RMSD top[a]	"Reasonably predicted"	2.14	2.14
RMSD best[b]		1.93	2.14
Omega best RMSD[c]		1.16	1.16
GLIDE docking	**Schulz-Gash and Stahl[42]**	**Present Study**	
Input conformation	Random	X-ray	Random
RMSD top	"Reasonably predicted"	2.54	8.55
RMSD best		1.14	1.08
SLIDE docking	**Zavodszky et al.[19]**	**Present Study**	
Input conformation	X-ray	X-ray	Random
RMSD top	n.a.[d]	1.49	5.61
RMSD best	0.44	1.10	5.61

[a]RMSD in Å of the top pose from the experimental solution.
[b]RMSD in Å of the best pose from the experimental solution.
[c]RMSD in Å of the best conformation generated by OMEGA from the experimental solution. The crystallographic water molecule was considered as part of the active site in FRED and GLIDE.
[d]Not available.
The PDB entry 1dwd was used, except for SLIDE docking by Zavodszky et al.,[19] where 1vrl coordinates have been used.

TABLE IV. Docking 4 Ligands Into Their Cognate Proteins Using QXP

	McMartin and Bohacek[8]			Present Study			
Input conformation	Top[a]	Best[b]	Best[c]	Top[d]	Best[c]	Top[d]	Best[c]
	"Distorted X-ray"			X-ray		Inverted X-ray[e]	
4phv	0.19	0.16	1.1	0.58	0.58	6.58	5.18
4dfr	0.74	0.30	0.96	0.42	0.42	9.59	8.53
1hfc	1.54	0.09	0.37	5.69	0.56	6.82	6.82
1stp	0.74	0.08	0.54	6.21	0.98	10.57	2.55

[a]RMSD in Å of the top pose from the energy-minimized X-ray solution.
[b]RMSD in Å of the best pose from the energy-minimized the X-ray solution.
[c]RMSD in Å of the best pose from the X-ray solution.
[d]RMSD in Å of the top pose from the X-ray solution.
[e]Rotation of the X-ray pose by 180° along its main axis.

data suggest that using random conformations for true actives is likely to dramatically decrease SLIDE's performance. It also hints that SLIDE would be more efficient in biased site point mode (i.e., making use of protein cavity–derived pharmacophoric points to screen out compounds that do not have an appropriate chemical functional group likely to intact with these key points of the active site). Such biases can, however, be applied only to well-defined protein cavities for which key amino acids have already been identified. It is therefore questionable whether SLIDE would be able to retrieve hits from a database for less defined active sites (e.g., orphan target).

QXP was evaluated using X-ray data for 12 protein–ligand complexes, 4 of which were common to our 100-complex data set.[8] Docking search was performed on randomly distorted ligands that were then subjected to 100 cycles of MC simulation (default value). Resulting RMSDs between retained docked poses and the X-ray structure were always smaller than 2 Å (Table V). In the present study, the number of MC cycles was limited to 30 to ensure fast screening. Unless the X-ray ligand coordinates were used as input, the conformational search was not long enough to allow the program to find a reliable solution for any of the tested examples (1hfc, 1stp, 4dfr, and 4phv).

SURFLEX docking benchmarks have recently been described for a set of 81 high-resolution protein–ligand complexes,[20] 67 of which are common to the data set investigated herein. Although different settings were used in both studies (docking of 10 energy-minimized random input conformations in the original study, docking a single nonminimized conformation in the current protocol), con-

TABLE V. Comparison of SURFLEX Benchmarks for Single-ligand and Database Docking

	Jain[20]		Present Study	
	Single-ligand docking[a]			
Success rate (best RMSD)[b]	92.4		84.8	
Success rate (best score)[c]	78.8		65.2	
	Virtual screening (TK dataset)[d]			
Penalty threshold[e]	−3	−6	−3	−6
True positives, %[f]	32	16	28	16
False negatives, %[g]	0.3	0.6	0.3	0.6

[a]Common set of 67 PDB entries.[18,20]
[b]Percentage of best poses closer than 2.0 Å from the X-ray pose.
[c]Percentage of top poses closer than 2.0 Å from the X-ray pose.
[d]Data set of 1000 compounds comprising 10 true inhibitors and 990 randomly chosen "drug-like" molecules.[26] Random molecules are presumed to be inactive although not experimentally tested for binding to TK.
[e]Protein penetration value.
[f]Percentage of actives in molecules selected by virtual screening (in the top 2.5% scorers).
[g]Percentage of actives in molecules eliminated by virtual screening (not present in the top 2.5% scorers).

sidering the best RMSD pose yields quite comparable success rates at a 2 Å RMSD threshold (Table V). More discrepancies occur if the best scored pose is considered. The better performance of SURFLEX in Jain's study[20] is mainly explained by a more exhaustive sampling of the ligand's conformational space. Since SURFLEX is a deterministic method, quite similar results are obtained on the TK screening example (Table V), with slight differences mainly explained by the use of different computer platforms (PC/Windows2000 vs SGI/IRIX). It should be noted that the use of a higher bump tolerance (protein penetration threshold equal to −6 instead of −3, as in Jain's original study) to allow the scoring of all TK inhibitors significantly decreased SURFLEX performance on this peculiar dataset (Table V). However, it still led to the best enrichment rates with regard to other docking tools examined in the current study (Fig. 4).

Docking Performance as a Function of Steric and Electrostatic Features of the Ligand and of the Protein Cavity

On the basis of the above conclusions, SLIDE and QXP performances have not been further analyzed, since parameter settings that would suit virtual screening purposes do not allow efficient docking. The top-ranked pose and the experimentally determined solution were compared with respect to the dimensions of the protein active site [Fig. 5(A), upper graph]. Four programs, namely, DOCK, FLEXX, GOLD, and SURFLEX, perform well on small binding sites (volumes below 700 Å^3), and a steady decrease of accuracy is observed upon increase of the protein cavity volume from 300 to 2000 Å^3. Conversely, FRED and GLIDE achieve best performance for medium-size binding sites (700–1000 Å^3 and 1000–1500 Å^3). Remarkably, FRED and SURFLEX still predict reliable ligand poses in very large binding sites (>2000 Å^3). More especially, SURFLEX is the only docking tool able to properly handle the location of CGP38560 into the 3500 Å^3 cavity of human renin (which corresponds to the largest binding site of the data set; PDB code: 1rne). Similar tendencies are observed by comparing the best pose to the X-ray solution [Fig. 5(A), lower graph]. It is noteworthy that GOLD and GLIDE also succeed in placing ligands in large protein cavities (>1500 Å^3), yet the programs are not able to properly rank solutions (see the differences between statistics generated from the best and the top pose). On average, the GLIDE empirical scoring function (GlideScore), as well as that of FLEXX and GOLD, does not efficiently rank poses into large binding pockets. A plausible explanation of this observation resides in the strong directional terms utilized by the latter scoring functions for describing H-bonding. Thus, the scoring function tends to overestimate buried ligand–protein electrostatic interactions, which leads to a poor docking of partially buried but flexible ligands.

Docking performances evaluated as a function of ligand flexibility [Fig. 5(B)] provide evidence for docking accuracy being inversely proportional to the conformational freedom of the ligand for DOCK, FLEXX and GOLD. This effect is much less pronounced for FRED, GLIDE, and SURFLEX. The latter tool is remarkably successful in docking highly flexible ligands [Fig. 5(B, C)].

To test whether docking performances depend on the shape of the active site, docking data were analyzed as a function of ligand burying. More than a third of the investigated active sites consist of closed cavities, enabling complete ligand burying. For 50 active sites, the protein-bound ligand buried surface ranges from 70% to 90%. In 11 complexes, cavities are more open and bound ligands are exposed to solvent. Considering the percentage of best poses docked closer than 2 Å from the X-ray solution using FLEXX, GLIDE, GOLD, and SURFLEX (Fig. 6), it appears that docking accuracy is generally improved upon experimental ligand burying increases from 38% to 100%. Note that the DOCK shape fitting protocol yields a good success rate for docking ligands in opened cavities. As opposed to the other 5 docking tools, FRED accuracy does not decrease when the ligand is partially buried in the active site (Fig. 6). It is likely that both the good quality of OMEGA-generated conformations and the soft Gaussian potential used to smoothen van der Waals interactions account for the latter observation.

As expected, the binding mode is mainly determined by the overall shape of the binding pocket for nonpolar ligands. As shown in Figure 7(A), DOCK, FRED, FLEXX, GLIDE, GOLD, and SURFLEX all perform well in docking small hydrophobic ligands. Upon increase of the size of hydrophobic ligands [Fig. 7(B)], GOLD and SURFLEX performances roughly remain unchanged, and FRED and GLIDE are still efficient in predicting the ligand placement (best pose), yet seldom succeed in ranking the best pose first. Last, DOCK, as well as FLEXX, undergoes a

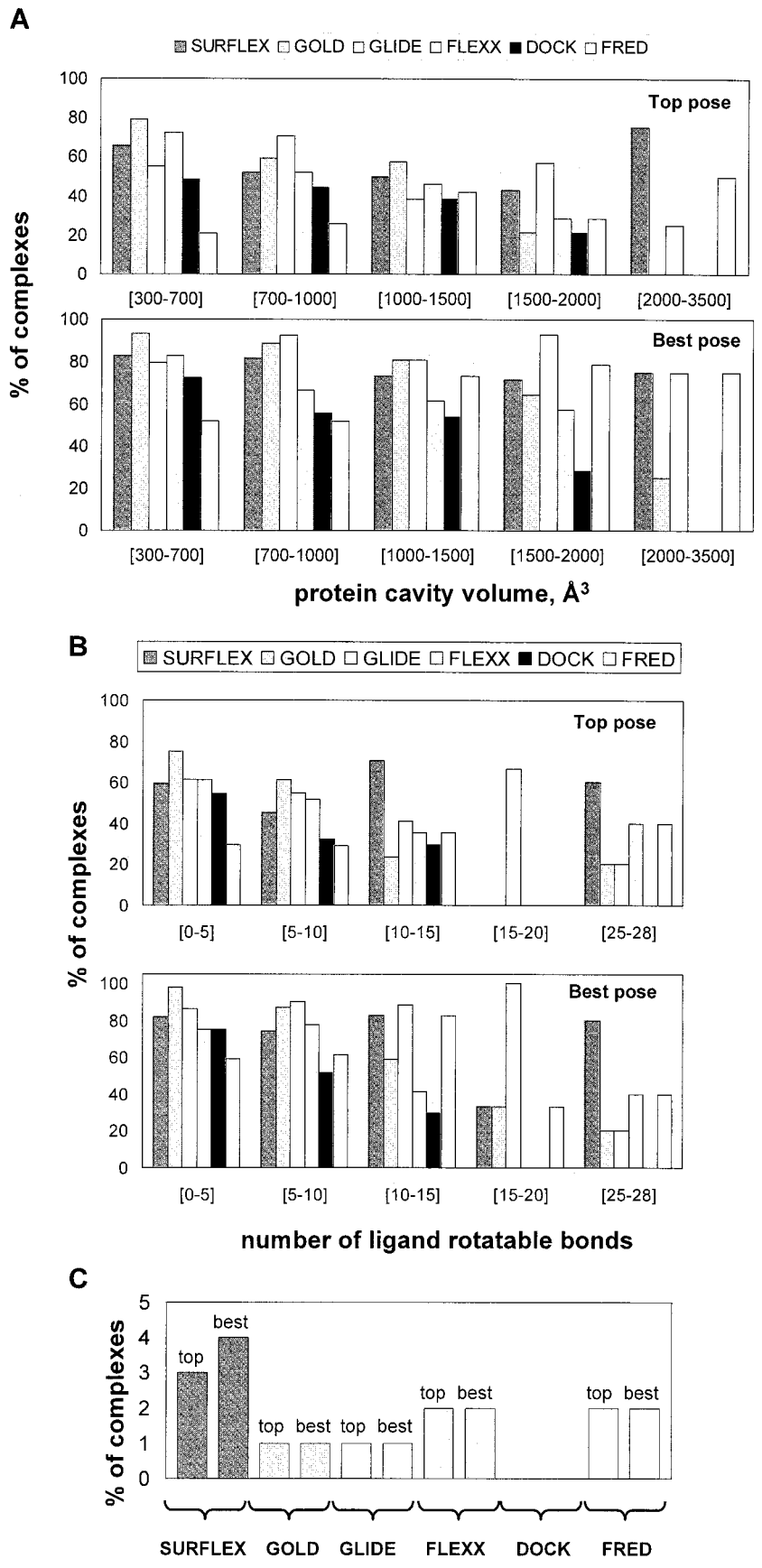


Fig. 5. Docking success rate as a function of physicochemical properties of protein–ligand complexes. (**A**) Percentage of successful docking at a 2 Å RMSD cutoff as a function of the volume of the protein cavity; 5 subsets corresponding to increasing volume intervals contain 29, 27, 26, 14, and 4 PDB entries, respectively. (**B**) Percentage of successful docking as a function ligand flexibility; 5 subsets corresponding to increasing ligand flexibility contain 44, 31, 17, 3, and 5 members, respectively. Rotatable bonds have been calculated using the TRIPOS Sybyl Line Notation (sln) representation. (**C**) Docking of 5 highly flexible ligands (1aaq, 1apt, 1eed, 1poc, and 1rne) with more than 25 rotatable bonds. For each program, results for top and best poses are displayed from left to right.

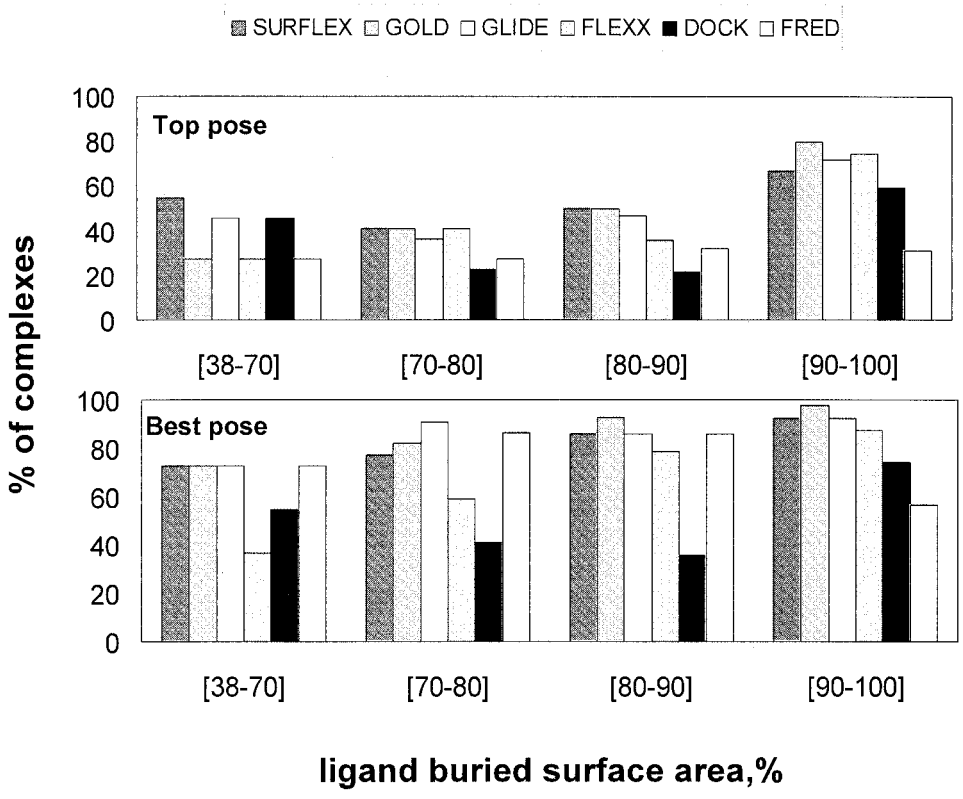


Fig. 6. Docking success rate as a function of ligand burying. Percentage of successful dockings at a fixed 2 Å RMSD cutoff for the top and the best poses are plotted for 4 ligand subsets showing increasing receptor-bound burial, as estimated by SAVOL. Four subsets showing increasing percentage of protein-bound ligand burial contain 11, 22, 28, and 39 members, respectively.

little decay of both docking and ranking accuracy. Hence, insofar as the conformational space to search is limited, shape complementarity is taken into account by all docking algorithms studied herein. Increasing the size and thus the conformational flexibility of the ligand leads to some docking errors, probably because too few conformations have been scanned in the allowed CPU time that was required (notably for FLEXX and DOCK), generating ranking discrepancies.

For very polar ligands [Fig. 7(C)] requiring strong hydrogen bonding to the receptor, DOCK shaped fitting leads, as expected, to a rather poor accuracy. By contrast, FRED, FLEXX, and SURFLEX docking gives satisfactory results, superior to those obtained with GLIDE and GOLD, which manage to achieve correct placement of hydrophilic ligands into binding pockets yet produce bad ranking of obtained poses [Fig. 7(C)]. The poorer behavior of GLIDE and GOLD for the latter set of ligands is somehow surprising; most programs use a scoring function containing an explicit H-bond term with a directional restraint. GlideScore, the scoring function utilized by GLIDE, is directly derived from ChemScore, the scoring function used by FRED in its optimization step. Nevertheless, both programs behave differently for the current set of highly polar ligands. We have not a single and clear explanation for this observation. As described in the next section, a partial explanation may come from the structure of some very hydrophilic ligands for which several H-bond donors/acceptors are symmetrically distributed and thus lead to inaccurate ranking of the near-native pose.

Systematic Docking Failures

About 10% of the 100 PDB entries led to systematic docking errors (Table VI), whatever the docking program. Failures were often due to insufficient conformational sampling for highly flexible ligands (e.g., 1rne, 1aaq, 1eed, 1glq, 2lic, and 1poc). Docking was also impaired by an open and shallow active site that binds to a partially buried ligand (e.g., 1mrc, 1glq). Some entries exhibit unusual binding mode and may be unsuitable for the purpose of validating docking algorithms. This is exemplified by 5 entries (1ghb, 8ghc, 1ive, 1lmo, and 2plv) for which significant clashes between protein and ligand atoms occur in the X-ray structure.[38] The example relative to D-xylose isomerase represents a more difficult case. Four entries in our data set (1xid, 1xie, 1did, and 1die) involve D-xylose isomerase. Cognate ligands are L-ascorbic acid, 1,5-dianhydrosorbitol, 5-dideoxy-2,5-imino-D-glucitol and 1-deoxy-nojirimycin in 1xid, 1xie, 1did, and 1die entries, respectively [Fig. 8(A)]. They all are characterized by a very high ratio between the number of H-bond donors/acceptors and the number of carbon atoms. No satisfactory poses were found for L-ascorbic acid, except using GOLD, whereas the 3 other molecules are correctly placed in the enzyme-binding site by most programs. In the X-ray structure of the 4 complexes, the ligand does not fully occupy the polar cavity of the protein but coordinates a Mn^{2+} ion (Mg^{2+} in the 1did entry). The metal ion constitutes a strong anchor in the binding site and thus facilitates

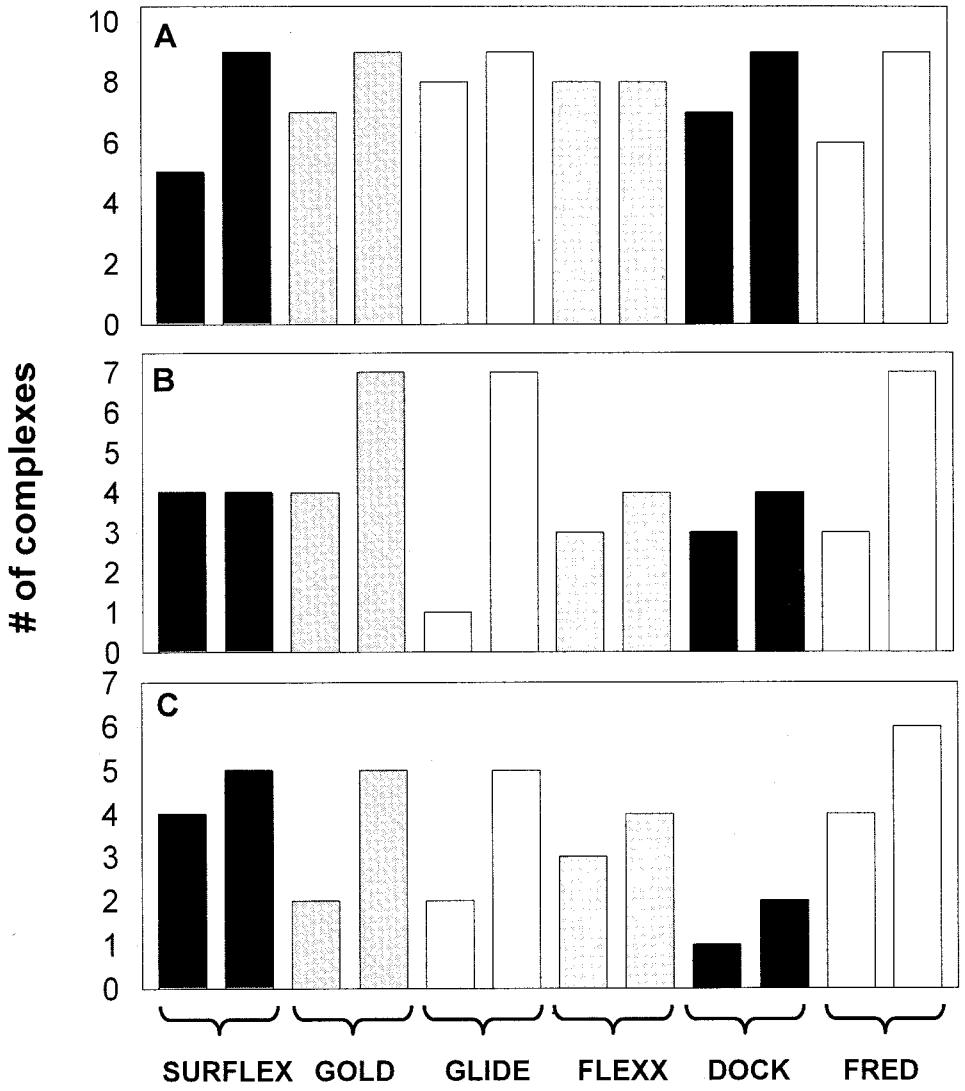


Fig. 7. Docking success rate as a function polar surface area of the ligand. Number of successful docking at a fixed 2 Å RMSD cutoff for the top (left) and the best (right) poses are plotted ligand subsets of increasing polarity (**A**, 10 small hydrophobic ligands; **B**, 7 medium-sized hydrophobic ligands; **C**, 7 very polar ligands). Subsets refer to groups 1, 2, and 3, as described in Figure 1(C).

the docking. Indeed, it explains why ligands in 1did and 1die entries are generally accurately docked. In 1xid and 1xid entries, the binding pocket selected for docking contains a second Mn^{2+} ion, whose coordination sphere may be completed. As a consequence, significantly different poses of the ligand coordinating either one or both metal ions resulted in favorable interactions with the binding site, more especially for the 5-membered ring of L-ascorbic acid [Fig. 8(B)].

Even if a docking program could easily find an accurate pose for a ligand, accurate ranking sometimes remains an issue (Table VI). Analyzing poses close to the crystal structure but badly ranked suggests the following observations:

1. Poses are generally inaccurately evaluated when no or very few lipophilic interactions occur between the protein and its ligand. In 2cmd, 4cts, 1tdb, 1did, and 1imb PDB entries, ligands are very polar, and several of them can be viewed as pseudosymmetrical molecules with regard to H-bond acceptors/donors. Most scoring functions cannot distinguish the X-ray pose from predicted solutions.
2. Similarly, good poses are poorly ranked when the ligand makes no or very few electrostatic interactions (H-bonds or salt bridges) with the protein. In 1acj, 2r07, 1icn, and 1epb PDB entries, ligands are all very hydrophobic. In the 1igj example, none of the H-bond acceptor/donors of the ligand buried upon protein binding are actually involved in intermolecular hydrogen bonding. As the primary goal of most docking algorithms is to favor energetically stable poses for which intermolecular H-bonds are optimized, such entries are usually badly handled.
3. Scoring is also challenging for complexes whose X-ray

TABLE VI. PDB Entries Leading to Frequent Docking Failures

PDB entry	Protein	Ligand	Main Cause of Failure
Unsuccessful dockings[a]			
1aaq	HIV-1 protease	Hydroxyethylene isostere	Very flexible ligand
1rne	Renin	CGP38560	Very flexible ligand
1poc	Phospholipase A2	1-O-octyl-2-heptylphosphonyl-SN-glycero-3-phosphoenolamine	Very flexible ligand
1lic	Adipocyte lipid-binding protein	Hexadecanesulfonic acid	Very flexible ligand
1eed	Endothiapepsin	Cyclohexyl renin inhibitor PD125754	Very flexible ligand
1eap	Catalytic antibody 17E8	Phenyl [1-(1-succinylamino)pentyl] phosphonate	Flexible ligand with symmetrical distribution of apolar groups
1glq	Glutathione S-transferase yfyf	S-(p-nitrobenzyl) glutathione	Partially buried ligand
1mer	Immunoglobulin δ light chain	*N*-acetyl-L-His-D-Pro-Oh	Open and shallow active site
1ghb	δ-Chymotrypsin	*N*-acetyl D-Trp	Short protein–ligand intermolecular distances
1ive	Neuraminidase N2	4-(acetylamino)-3-aminobenzoic acid	Short protein–ligand intermolecular distances
1lmo	Mucopeptide *N*-acetylmuramylhydrolase	Di-*N*-acetylglucosamine	Short protein–ligand intermolecular distances
2plv	Poliovirus	Myristic acid	Short protein–ligand intermolecular distances
8gch	δ-Chymotrypsin	Gly-Ala-Trp	Short protein–ligand intermolecular distances
1xid	D-Xylose isomerase	L-ascorbic acid	Metal coordination mismatch (see Fig. 8)
Unsuccessful scoring[b]			
2cmd	Malate dehydrogenase	Citric acid	Pseudosymmetrical distribution of H-bond donors/acceptors in the ligand
4cts	Citrate synthase	Oxaloacetate ion	Pseudosymmetrical distribution of H-bond donors/acceptors in the ligand
1imb	Inositol monophosphatase	L-myo-inositol-1-phosphate	Pseudosymmetrical distribution of H-bond donors/acceptors in the ligand
1did	D-Xylose isomerase	2,5-dideoxy-2,5-imino-D-glucitol	Pseudosymmetrical distribution of H-bond donors/acceptors in the ligand
1tdb	Thymidylate synthase	5-fluoro-2′-deoxyuridine-5′-monophosphate	Solvent-accessible ligand
1acj	Acetylcholinesterase	Tacrine	Hydrophobic ligand
1icn	Fatty acid–binding protein	Oleate	Hydrophobic ligand
1ebp	Retinoic acid–binding protein	Retinoic acid	Hydrophobic ligand
1igj	Igg2A (κ) antibody Fab fragment	Digoxin	Hydrophobic ligand
2r07	Rhinovirus 14	Antiviral agent WIN VII	Hydrophobic ligand
1hfc	Collagenase	[(*N*-(2-hydroxymatemethylene-4-methyl-pentoyl) phenylalanyl)] Methylamine	Metal coordination mismatch
1hyt	Thermolysin	Benzylsuccinic acid	Metal coordination mismatch

[a]No more than 2 programs manage to successfully dock the ligand in its active site.
[b]Successful scoring for less than half of the programs that yield successful docking.
Only results produced by the 6 programs (DOCK, FLEXX, FRED, GLIDE, GOLD, and SURFLEX) were analyzed. Docking starting from X-ray input coordinates is considered as successful when the RMSD of best pose from the X-ray structure is below 2 Å. Similarly, scoring is successful when the RMSD of the top pose from the X-ray structure is below 2 Å.

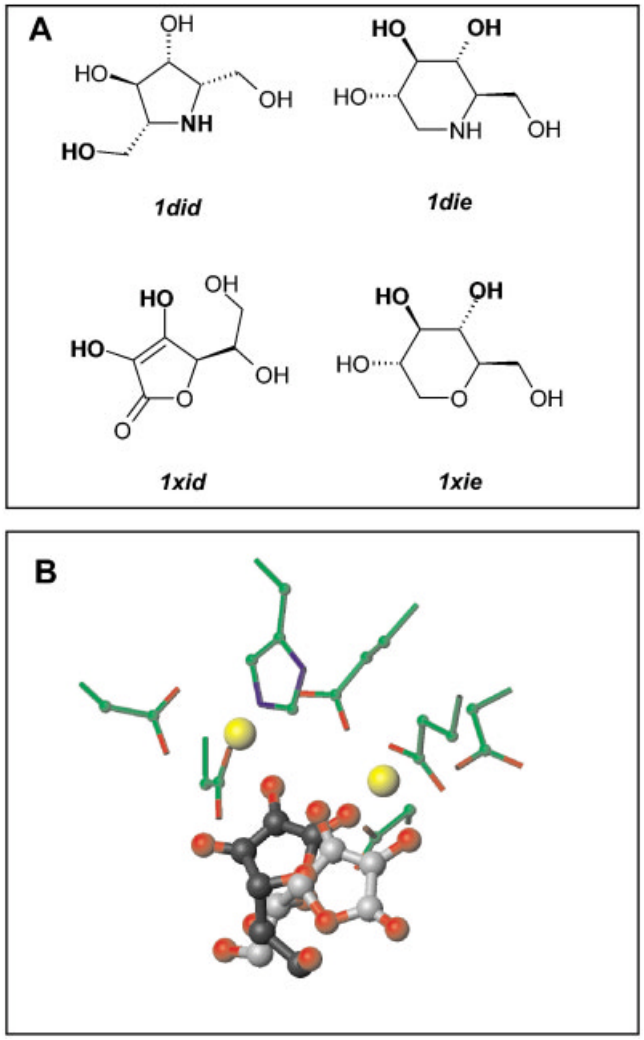


Fig. 8. Docking in D-xylose isomerase binding sites. (**A**) Ligands of the four entries in our test set. Atoms involved in metal coordination as observed in the X-ray structure of complexes are displayed in bold. (**B**) Overlay of native (white) and SURFLEX best pose (black) for L-ascorbic acid bound to 1xid (green sticks). For the sake of clarity, the ligand is displayed using ball and stick representation, whereas active site atoms are represented by sticks. The two Mn^{++} ions are displayed as yellow balls.

structure reveals a mismatch between hydrophobic/electrostatic potentials of the ligand and those of the active site. In this respect, the examples relative to the complexes between a zinc-containing enzyme and a ligand involved in the metal coordination are demonstrative.

Most programs manage to predict an accurate pose of the ligand, since the zinc ion prevents the placement of the ligand in irrelevant regions of the binding pocket. Though docking was generally accurate, ranking first the best pose often failed for 1azm, 1cil, 1hfc, and 1hyt zinc-containing PDB entries. Lipophilic moieties of the ligand facing hydrophilic parts of the active site and vice versa are observed in the three former PDB complexes and could partially explain the scoring failure. In addition, imperfections in the zinc coordination sphere may contribute to the bad score. In the 1hyt crystal structure, the ligand moiety involved in zinc coordination is a carboxylate but too close (less than 3 Å) to two other carboxylate groups of the protein to be well scored by an empirical scoring function. Since very close contacts between protein and ligand atoms bearing the same charge are penalized by most scoring functions, such poses are difficult to recover.

Main Strengths and Weaknesses of Investigated Docking Programs

Analyzing the above-described results enables us to propose prioritization schemes for specific methods depending on the docking context (Table VII).

We have tested two types of deterministic approaches: multiconfomer docking (FRED) and incremental construction (DOCK, FLEXX, GLIDE, SLIDE, and SURFLEX). Though FRED failed to retrieve known TK inhibitors in a virtual screening experiment, our docking data suggest that this program may be successful for other targets [e.g., proteins with large cavities; recall Fig. 5(A)]. Actually, most of docking failures with this program in our data set of 100 PDB complexes were observed for entries whose ligands are small, polar, and deeply buried in the protein cavity (e.g., 6abp, 1aha). Most likely, the smooth Gaussian function used during the shape-fitting step is not able to predict the location of such polar ligands, since it accounts only for shape in the primary docking step.

In the case of deterministic algorithms using an incremental construction approach, SURFLEX and GLIDE generally outperformed DOCK, FLEXX, and SLIDE, whatever the application (single-ligand docking, virtual screening). In the 3 latter programs, core and end rotamer groups are treated separately during conformational search steps, whereas in the 2 former programs, end rotamer groups are carried along with the core fragment, thereby avoiding inappropriate placement of the core fragment and subsequent protein–ligand interpenetration at the peripheral fragments. Indeed, for DOCK and FLEXX, docking errors are generally related to ligand flexibility (number of rotatable bonds ≥ 10) and the resulting numerous possibilities of ligand fragmentation. It should be also noted that DOCK, which used a shape-based docking engine, also has difficulty finding an accurate pose for a ligands in a large and open binding cavity. Regarding the high docking speed of both programs, it is likely that the ligand flexibility issue can be partly addressed by a more exhaustive conformational sampling. GLIDE performances are in part due to the above-mentioned possibility of scaling down van der Waals radii of nonpolar atoms and the use of additional filtering steps that help to discriminate good from bad poses. After passing a series of steric and energy filters, ligands undergo grid-based energy minimization and MC sampling. SURFLEX robustness and reliability, notably for docking flexible ligands in large binding sites, may be explained by an efficient and original surface-based similarity match between ligand atoms and receptor hotspots.[20,44] The binding pocket is first characterized by a

TABLE VII. Strengths and Weaknesses of Docking Programs According to Physicochemical Properties of Protein–Ligand Complexes[a]

Program	Strengths	Weaknesses
DOCK	Small binding sites Opened cavities Small hydrophobic ligands	Flexible ligands Highly polar ligands
FLEXX	Small binding sites Small hydrophobic ligands	Very flexible ligands
FRED	Large binding sites Flexible ligands Small hydrophobic ligands High speed	Small, polar, buried ligands
GLIDE	Flexible ligands Small hydrophobic ligands	Ranking very polar ligands Low speed
GOLD	Small binding sites Small hydrophobic ligands	Ranking very polar ligands Ranking ligands in large cavities
SLIDE	Side-chain flexibility	Sensitivity to ligand input coordinates
SURFLEX	Large and opened cavities Small binding sites Very flexible ligands	Low speed for large ligands
QXP	Optimizing known binding modes	Sensitivity to ligand input coordinates

[a]Small polar ligands are generally well docked by most programs. Thus, this group of ligands is not described here.

set of probes that represent potential interactions with the protein; then both density of probes and surrounding cavities are analyzed to select the most favorable area for placing the core fragment of the ligand.

GOLD and QXP use stochastic methods (genetic algorithm and MC simulated annealing, respectively) to explore ligand positions and conformations. GOLD achieves accurate docking within a timeframe that suits virtual screening purposes. By contrast, the time necessary for QXP to achieve a thorough sampling of ligand poses and its high sensitivity to input coordinates of the ligand make it unusable for virtual screening applications.

CONCLUSIONS

Eight popular docking tools have been compared on common data sets for both docking and virtual screening accuracy. It is not our intention in the current study to propose a hierarchy of available docking programs but to notice advantages and drawbacks of selected tools in different contexts. As a matter of fact, we made the choice of examining docking tools from a virtual screening perspective, which means using settings compatible with fast docking. Hence, we believe that speed is nowadays an important aspect of computational drug discovery techniques, as the computational chemist has to cope with the throughput already reached by medicinal chemists and biologists. Thus, benchmarks reported herein may significantly differ from the peak performance that can be reached by a docking program under optimal conditions.

Many previous reports have pointed out that the docking performance of most docking tools is very much dependent on the target, and that predicting which one is the most suitable in a precise context is almost impossible.[26,29,32,35,42] We even reported discrepancies between docking and screening accuracies,[26] which is rather frustrating in a structure-based screening approach. For the first time, this issue has been simultaneously addressed by comparing different tools on a common protein–ligand data set (100 high-resolution PDB entries) that is large enough to be statistically relevant, and on a practical virtual screening application. Interestingly, the docking tools (FLEXX, GLIDE, GOLD, and SURFLEX) considered the most accurate in terms of docking (predicting the X-ray pose) were also the most successful in enriching a virtual hit list in known inhibitors after screening the X-ray structure of the cognate enzyme. This observation demonstrates that the preparation of our data sets, as well as the settings chosen for the different programs, is sufficiently unbiased to ensure a fair comparison and draw statistically reliable conclusions. However, it is important to note that good docking accuracy is necessary but not sufficient for accurate screening utility. Hence, among the most accurate docking tools considered herein (GLIDE, GOLD, and SURFLEX), there are still significant differences in their propensity to enrich a virtual hit list in true hits, as illustrated by our virtual screening test. However, we acknowledge that several virtual screening tests on different data sets and different targets would be necessary to fully address this issue. The current benchmarks can be used to prioritize

the selection of specific docking tools depending on the physicochemical properties of both the active site and the ligand(s) to be investigated.

NOTE-IN-PROOF

The application of the SLIDE docking tool in this work was carried out under conditions not currently recommended by its developers. However, additional calculations performed at the recommended conditions do not lead to significantly different results for SLIDE in comparison to the others evaluated herein.

ACKNOWLEDGMENTS

We wish to thank the Centre Informatique National de l'Enseignement Supérieur (CINES, Montpellier) for allocation of computing time and Ajay Jain for providing us electronic data sets for benchmark comparisons. Data sets (3D structures) and benchmarks (RMSDs, CPU times) described in this article are available for noncommercial academic research upon request to the authors.

REFERENCES

1. Brooijmans N, Kuntz ID. Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 2003;32:335–373.
2. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule–ligand interactions. J Mol Biol 1982;161:269–288.
3. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
4. Taylor RD, Jewsbury PJ, Essex JW. A review of protein–small molecule docking methods. J Comput Aided Mol Des 2002;16:151–166.
5. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel LJ, Freer ST. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. Chem Biol 1995;2:317–324.
6. Welch W, Ruppert J, Jain AN. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chem Biol 1996;3:449–462.
7. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
8. McMartin C, Bohacek RS. QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 1997;11:333–344.
9. Totrov M, Abagyan R. Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins 1997;1:215–220.
10. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD. Flexible docking using Tabu search and an empirical estimate of binding affinity. Proteins 1998;33:367–382.
11. Morris G, Goodsell D, Halliday R, Huey R, Hart W, Belew R, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 1998;19:1639–1662.
12. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. Proteins 1999;37:228–241.
13. David L, Luo R, Gilson MK. Ligand–receptor docking with the Mining Minima optimizer. J Comput Aided Mol Des 2001;15:157–171.
14. Diller DJ, Merz KM Jr. High throughput docking for library design and library prioritization. Proteins 2001;43:113–124.
15. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 2001;15:411–428.
16. Pang YP, Perola E, Xu K, Prendergast FG. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. J Comput Chem 2001;22:1750–1771.
17. Jackson RM. Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. J Comput Aided Mol Des 2002;16:43–57.
18. Paul N, Rognan D. ConsDock: A new program for the consensus analysis of protein–ligand interactions. Proteins 2002;47:521–533.
19. Zavodszky MI, Sanschagrin PC, Korde RS, Kuhn LA. Distilling the essential features of a protein surface for improving protein–ligand docking, scoring, and virtual screening. J Comput Aided Mol Des 2002;16:883–902.
20. Jain AN. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 2003;46:499–511.
21. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK. Gaussian docking functions. Biopolymers 2003;68:76–90.
22. Taylor RD, Jewsbury PJ, Essex JW. FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. J Comput Chem 2003;24:1637–1656.
23. Venkatachalam CM, Jiang X, Oldfield T, Waldman M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 2003;21:289–307.
24. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein–ligand docking using GOLD. Proteins 2003;52:609–623.
25. Muegge I, Martin YC, Hajduk PJ, Fesik SW. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. J Med Chem 1999;42:2498–2503.
26. 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.
27. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein–ligand interactions. J Mol Biol 2000;295:337–356.
28. Pearlman DA, Charifson PS. Are free energy calculations useful in practice?: a comparison with rapid scoring functions for the p38 MAP kinase protein system. J Med Chem 2001;44:3417–3423.
29. Stahl M, Rarey M. Detailed analysis of scoring functions for virtual screening. J Med Chem 2001;44:1035–1042.
30. Terp GE, Johansen BN, Christensen IT, Jorgensen FS. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein–ligand binding affinities. J Med Chem 2001;44:2333–2343.
31. Buzko OV, Bishop AC, Shokat KM. Modified AutoDock for accurate docking of protein kinase inhibitors. J Comput Aided Mol Des 2002;16:113–127.
32. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB. Consensus scoring for ligand/protein interactions. J Mol Graph Model 2002;20:281–295.
33. Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed Engl 2002;41:2644–2676.
34. Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 2003;46:2287–2303.
35. Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 1999;42:5100–5109.
36. Keseru GM. A virtual high throughput screen for high affinity cytochrome P450cam substrates: implications for in silico prediction of drug metabolism. J Comput Aided Mol Des 2001;15:649–657.
37. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 2002;45:2213–2221.
38. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein–ligand interaction. Proteins 2002;49:457–471.
39. 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.
40. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P,

Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 2002;58:899–907.
41. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. 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.
42. Schulz-Gasch T, Stahl M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. J Mol Model 2003;9:47–57.
43. Schnecke V, Swanson CA, Getzoff ED, Tainer JA, Kuhn LA. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. Proteins 1998;33:74–87.
44. Ruppert J, Welch W, Jain AN. Automatic identification and representation of protein binding sites for molecular docking. Protein Sci 1997;6:524–533.
45. Muegge I. Selection criteria for drug-like compounds. Med Res Rev 2003;23:302–321.
46. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13:323–330.
47. Bostrom J, Greenwood JR, Gottfries J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. J Mol Graph Model 2003;21:449–462.
48. Wurth C, Kessler U, Vogt J, Schulz GE, Folkers G, Scapozza L. The effect of substrate binding on the conformation and structural stability of *Herpes simplex* virus type 1 thymidine kinase. Protein Sci 2001;10:63–73.
49. Pickett SD, Sherborne BS, Wilkinson T, Bennett J, Borkakoti N, Broadhurst M, Hurst D, Kilford I, McKinnell M, Jones PS. Discovery of novel low molecular weight inhibitors of IMPDH via virtual needle screening. Bioorg Med Chem Lett 2003;13:1691–1694.
50. Wyss PC, Gerber P, Hartman PG, Hubschwerlen C, Locher H, Marty HP, Stahl M. Novel dihydrofolate reductase inhibitors: structure-based versus diversity-based library design and high-throughput synthesis and screening. J Med Chem 2003;46:2304–2312.