PROTEINS: Structure, Function, and Bioinformatics 56:235–249 (2004)

A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance

Emanuele Perola,* W. Patrick Walters, and Paul S. Charifson
Vertex Pharmaceuticals Incorporated, Cambridge, Massachusetts

ABSTRACT **A thorough evaluation of some of the most advanced docking and scoring methods currently available is described, and guidelines for the choice of an appropriate protocol for docking and virtual screening are defined. The generation of a large and highly curated test set of pharmaceutically relevant protein–ligand complexes with known binding affinities is described, and three highly regarded docking programs (Glide, GOLD, and ICM) are evaluated on the same set with respect to their ability to reproduce crystallographic binding orientations. Glide correctly identified the crystallographic pose within 2.0 Å in 61% of the cases, versus 48% for GOLD and 45% for ICM. In general Glide appears to perform most consistently with respect to diversity of binding sites and ligand flexibility, while the performance of ICM and GOLD is more binding site–dependent and it is significantly poorer when binding is predominantly driven by hydrophobic interactions. The results also show that energy minimization and reranking of the top N poses can be an effective means to overcome some of the limitations of a given docking function. The same docking programs are evaluated in conjunction with three different scoring functions for their ability to discriminate actives from inactives in virtual screening. The evaluation, performed on three different systems (HIV-1 protease, IMPDH, and p38 MAP kinase), confirms that the relative performance of different docking and scoring methods is to some extent binding site–dependent. GlideScore appears to be an effective scoring function for database screening, with consistent performance across several types of binding sites, while ChemScore appears to be most useful in sterically demanding sites since it is more forgiving of repulsive interactions. Energy minimization of docked poses can significantly improve the enrichments in systems with sterically demanding binding sites. Overall Glide appears to be a safe general choice for docking, while the choice of the best scoring tool remains to a larger extent system-dependent and should be evaluated on a case-by-case basis. Proteins 2004;56:235–249.** © **2004 Wiley-Liss, Inc.**

Key words: ICM; Glide; GOLD; ChemScore; Glide Score; OPLS-AA; virtual screening; test set; enrichment factors; drug-like

INTRODUCTION

A large number of docking programs have been developed in the last 20 years based on a variety of search algorithms.[1–3] The use of such programs in conjunction with one or more scoring functions to evaluate and rank-order potential ligands from chemical collections is now one of the paradigms of virtual screening. While several successful applications of this methodology have been described in recent publications,[4,5] there are certainly many issues surrounding docking and ranking of docked structures that can still be improved. In this article we address some of these issues and try to determine how to best apply the tools that are presently available in a drug design context.

The primary question all docking programs try to address is what combination of orientation and conformation (pose) is the most favorable relative to all the other combinations sampled. When applied to screening, the process also requires a comparison of the best pose (or top few poses) of a given ligand with those of the other ligands such that a final ranking or ordering can be obtained. For the purposes of this article, we will refer to the function used in evaluating different poses of a given ligand as a docking function and any function(s) used in either refinement/reranking of docked ligand poses or for comparing different ligands as a scoring function. There is no requirement that the docking function and the scoring function be the same, although this has most often been the case in the early years. In practice, several studies have shown that rescoring docking poses with an alternative function can have a favorable impact both on pose selection and on rank-ordering in a virtual screening context.[6–12] It must also be emphasized, however, that the docking function must evaluate a large number of solutions (numbers of poses ranging from 10^4–10^5 are typical). Even if one saves a small set of top-ranking poses and rescores them with a different function, an initial bias will have been introduced by the docking function.

It has been suggested that a reasonable compromise between speed and accuracy might include using a simplified function (e.g., an empirical or knowledge-based function) as a docking function to save a set of viable poses and then use a more rigorous (e.g., energy-based) function for the final pose selection/ranking of ligands.[10,12] In this respect, it has been observed that certain types of scoring functions

*Correspondence to: Emanuele Perola, Vertex Pharmaceuticals Incorporated, 130 Waverly Street, Cambridge, MA 02139. E-mail: emanuele_perola@vrtx.com

Received 4 November 2003; Accepted 22 December 2003

Published online 28 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20088

tend to give better predictions in certain types of binding sites.[7,8,11] This dependence should be properly weighted when a secondary function for rescoring is selected. It has also been shown that combining the best results from two or more scoring functions can lead to a considerable reduction in false positives (consensus scoring).[9]

In addition to the general methodology, the performance of the individual tools for docking and scoring has been assessed in several studies. An analysis of the recent literature seems to indicate that DOCK,[13] FlexX,[14] and GOLD[15] are the most widely used docking programs.[5,16–21] Direct and indirect comparisons have shown that GOLD consistently outperforms the other two in terms of average docking accuracy on a variety of systems.[13,15,17,22,23] Among recently developed programs, ICM[24] and Glide[25] have been reported to achieve a high degree of accuracy.[24,26,27] While both programs have performed well on internal validation sets at Vertex, a large-scale comparison involving these and the previous programs has not yet been reported. In terms of scoring, the empirical function ChemScore[28,29] is the most widely used scoring function for virtual screening, and it has been shown to outperform most of the others in comparative studies.[9,30] GlideScore, recently developed at Schrodinger, Inc., using ChemScore as the initial template, has been specifically designed to maximize enrichment in database screening, and it is claimed to be an effective tool in its ability to discriminate between active and inactive compounds on a variety of systems.[27]

In this work, we address three main topics: the importance of test set selection, the appropriate choice of an accurate docking tool, and the performance of various docking/scoring combinations in virtual screening. The first part of this study describes a rigorous attempt at generating a database of pharmaceutically relevant protein–ligand complexes, which we view as a prerequisite for the evaluation of docking and scoring tools dedicated to drug discovery. The need for the generation of larger test sets, selected with consistent criteria and highly refined, has been highlighted by two recent publications.[31,32] A careful analysis of the recently reported test sets shows that, while the diversity and pharmaceutical relevance of the protein structures are satisfactory, the same cannot be said with respect to the ligands. Structural classes that are of less relevance to drug discovery programs (peptides, sugars, nucleotides) are still over-represented, with a high degree of redundancy, and the molecular weight of the ligands generally ranges from 100 to 1000, far beyond the range of interest for a drug discovery program. In general the reported test sets only contain a small percentage of truly drug-like ligands. Since such test sets are used in the evaluation/calibration of tools for drug design, it is important that the complexes be representative of what is relevant to the process. If the ultimate objective is to predict binding of drug-like molecules to pharmaceutically relevant proteins, complexes between such partners should clearly be emphasized. Following this premise we generated a new test set of complexes of known binding affinity, geared toward drug-like ligands and suitable for a variety of tasks: evaluation of docking programs and existing scoring functions, development and calibration of new scoring functions, and analysis of various aspects of protein–ligand binding.

In the second part of this study, we compare the well-established GOLD program with the recent additions ICM and Glide for their ability to reproduce crystallographic binding orientations. Critical features evaluated include the effect of energy minimization on the top scoring poses and the impact of the nature of the binding site on the accuracy of each program.

Finally, we analyze the performance of the three docking programs above in conjunction with three different scoring functions with respect to their abilities to maximize enrichments in database screening. The well-established empirical function ChemScore is compared with the recent addition GlideScore and with the OPLS-AA force field interaction energy.[33] Three targets with different binding site features were used in these calculations: HIV-1 protease, IMPDH, and p38 MAP kinase. The objective of this part of the study was to establish a protocol that efficiently combines the best available tools to maximize the outcome of a virtual screening.

METHODS

Complex Selection

A set of over 200 protein–ligand complexes was initially selected from the Protein Data Bank (PDB) and from the Vertex structure collection according to the following criteria:

General:

- binding constant (Ki or Kd) available
- noncovalent binding between ligand and protein
- crystallographic resolution < 3.0 Å

Ligands:

- molecular weight between 200 and 600
- 1 to 12 rotatable bonds
- drug/lead-like
- structurally diverse

Proteins:

- multiple classes
- diverse within classes
- relevant to drug discovery

The cutoffs for molecular weight and number of rotatable bonds reflect the distribution reported for the orally delivered drugs listed in the *Physicians' Desk Reference* (PDR).[34] The initial selection was pruned based on a number of additional criteria. In order to prioritize structures that are of higher pharmaceutical relevance, we excluded complexes involving ligand or protein classes that are less likely to be the focus of a modern drug discovery program. In particular we removed all the complexes with sugar-containing ligands (e.g., 4hmg), steroids (e.g., 1a27), and macrocycles (e.g., 1mmq), as well as complexes of heme-containing proteins (e.g., 1phg). On the same basis, complexes with ligands containing atoms other than C, N, O, S, F, Cl, Br, and H were also excluded (e.g., 1tha). We then removed structures with severe clashes between protein and ligand atoms (e.g., 1dth), or between protein or ligand atoms and water molecules involved in binding (e.g., 1c4y). Such structures may be poorly refined in the binding region and therefore unsuitable for the evaluation of docking programs and scoring functions. We also excluded complexes with potential ambiguities in the binding region, including structures with crystallographically related protein units involved in ligand binding (e.g., 1bm7) and structures with uncertain protonation state in the binding site (e.g., 1k4g). Finally, we removed complexes with uncommon features, which introduce additional complications without adding any specific value to the test set. Two examples of such instances are complexes in which ligand binding is mediated by a complex network of water molecules (e.g., 1jqe) and complexes with unconventional amino acid residues in the binding site (e.g., 1hlf).

Each ligand was included only once, thus avoiding common redundancies like methotrexate bound to different versions of dihydrofolate reductase or the same ligand bound to two closely related proteins. The purpose was to avoid repetitions of almost identical sets of interactions, thus maximizing the diversity of the interactions represented in the test set. These criteria reflect our intention to include the maximum amount of structural information on systems that are of high interest in a structure-based drug design context, and exclude those that are only rarely considered. The final selection included 100 complexes from the PDB and 50 complexes from the Vertex structure collection. The PDB codes of the 100 complexes selected from the PDB are reported in Table I, along with crystallographic resolutions, dissociation constants, expressed as pKi [$-\mathrm{Log}_{10}(\mathrm{Ki})$], and additional descriptors that will be discussed in the following section. The test set includes 63 different proteins from a variety of classes, including proteases, kinases, nuclear receptors, phosphatases, oxidoreductases, isomerases, and lyases. Kinases (43 complexes) and proteases (42 complexes) are the most widely represented, which reflects their high relevance in modern drug discovery and the fact that these classes more than others have been the focus of structure-based drug design efforts, resulting in the generation of a large amount of structural information. The kinase subset includes 12 different proteins with representatives of tyrosine kinases, serine/threonine kinases, and nucleotide kinases, while the protease subset includes 14 different proteins with representatives of serine proteases, aspartyl proteases, and metalloproteases, thus ensuring diversity within these classes. The overall set includes 24 metalloprotein complexes, all of them with a zinc ion in the active site. Several examples of approved drugs in complex with their targets are also included (e.g., Agenerase/HIV protease, Aricept/acetylcholinesterase, Lisinopril/Angiotensin converting enzyme).

Complex Preparation

Each of the PDB files of the 150 complexes was processed according to the following protocol: the ligand was extracted, bond orders and correct protonation state were assigned upon visual inspection, and the structure was saved to an SD file. If a cofactor was present, the same procedure was applied, and a separate SD file was generated. After removal of the ligand, a “clean” protein file was generated by removing subunits not involved in ligand binding and far from the active site, solvent, counterions, and other small molecules located away from the binding site. Metal ions and tightly bound water molecules in the ligand binding site were preserved, and the protein structure was saved to a PDB file. Hydrogen atoms were then added to the protein, and the structures of protein, ligand, and cofactor were combined in a single Macromodel file. The active site was visually inspected and the appropriate corrections were made for tautomeric states of histidine residues, orientations of hydroxyl groups, and protonation state of basic and acidic residues. The hydrogen atoms were minimized for 1000 steps with Macromodel in OPLS-AA force field, with all nonhydrogen atoms constrained to their original positions. Protein (with cofactor if present) and ligand with optimized hydrogen positions were finally saved to separate files.

Docking Studies

The test set of complexes described above was used in the evaluation. Each ligand was docked back into the corresponding binding site, and the accuracy of each prediction was assessed on the basis of the root-mean-square deviation (RMSD) between the coordinates of the heavy atoms of the ligand in the top docking pose and those in the crystal structure. The following paragraphs describe the search algorithm and scoring methods used in the three programs. For each program, details of the calculations performed in this study are provided.

ICM (MolSoft LLC). The Internal Coordinate Mechanics (ICM) program is based on a stochastic algorithm that relies on global optimization of the entire flexible ligand in the receptor field (flexible ligand/grid receptor approach).[24] Global optimization is performed in the binding site such that both the intramolecular ligand energy and the ligand–receptor interaction energy are optimized. The program combines large-scale random moves of several types with gradient local minimization and a history mechanism that both expels from the unwanted minima and promotes the discovery of new minima. The random moves include pseudo-Brownian moves, optimally biased moves of groups of torsions, and single torsion changes. The energy calculations are based on the ECEPP/3 force field,[35] with Merck molecular force field (MMFF) partial charges. Five potential maps (electrostatic, hydrogen bond, hydrophobic, van der Waals attaractive and repulsive) are calculated for the receptor. The location of the receptor binding pocket can be specified by the user or selected by the cavity detection module implemented in the program.

TABLE I. Composition and Properties of PDB Fraction of Test Set

Code	Res	PKI	MW	WT	HA	RB	HB	HB/HA	BF	Code	Res	PKI	MW	WT	HA	RB	HB	HB/HA	BF
13gs	1.90	4.62	398.4	—	28	7	2	0.07	0.73	1jsv	1.96	5.70	265.3	—	18	3	4	0.22	0.83
1a42	2.25	9.89	383.5	—	23	7	6	0.26	0.79	1k1j	2.20	7.68	522.6	—	37	9	3	0.08	0.69
1a4k	2.40	8.00	429.4	—	31	6	1	0.03	0.65	1k22	1.93	8.40	429.5	—	31	9	7	0.23	0.84
1a8t	2.55	5.80	452.5	—	34	7	2	0.06	0.76	1k7e	2.30	2.92	232.2	2	17	4	6	0.35	1.00
1afq	1.80	6.21	385.5	2	28	9	4	0.14	0.71	1k7f	1.90	3.32	274.3	2	20	5	5	0.25	0.94
1aoe	1.60	9.66	269.4	—	20	3	4	0.20	0.90	1kv1	2.50	5.94	306.8	—	21	3	3	0.14	0.93
1atl	1.80	6.28	325.4	—	22	8	3	0.14	0.75	1kv2	2.80	10.00	527.7	—	39	8	3	0.08	0.93
1azm	2.00	6.14	222.2	—	13	2	3	0.23	0.88	1l2s	1.94	4.59	317.8	1	19	4	7	0.37	0.73
1bnw	2.25	9.08	338.4	—	19	5	4	0.21	0.74	1l8g	2.50	6.22	466.4	1	31	6	7	0.23	0.72
1bqo	2.30	7.74	513.6	—	34	7	5	0.15	0.73	1lqd	2.70	8.05	424.5	1	32	6	3	0.09	0.79
1br6	2.30	3.22	312.3	—	23	4	4	0.17	0.79	1m48	1.95	5.09	446.5	—	33	8	3	0.09	0.60
1cet	2.05	2.89	319.9	—	22	8	1	0.05	0.56	1mmb	2.10	9.22	477.6	—	32	12	8	0.25	0.69
1cim	2.10	9.55	296.4	—	17	1	6	0.35	0.82	1mnc	2.10	9.00	349.4	—	25	9	8	0.32	0.72
1d3p	2.10	5.11	543.7	1	39	11	1	0.03	0.73	1mq5	2.10	9.00	537.9	—	34	6	1	0.03	0.78
1d4p	2.07	6.30	360.5	—	27	4	3	0.11	0.86	1mq6	2.10	11.15	566.8	1	36	8	1	0.03	0.75
1d6v	2.00	6.17	381.5	1	28	7	1	0.04	0.70	1nhu	2.00	5.66	496.3	—	33	8	2	0.06	0.65
1dib	2.70	7.74	471.4	—	34	7	4	0.12	0.86	1nhv	2.90	5.66	550.4	—	37	8	2	0.05	0.55
1dlr	2.30	9.18	325.4	1	24	4	4	0.17	0.93	1o86	2.00	9.57	405.5	—	29	12	8	0.28	0.77
1efy	2.20	8.22	267.3	—	20	3	2	0.10	0.84	1ohr	2.10	8.70	567.8	1	40	10	6	0.15	0.91
1ela	1.80	6.35	456.5	—	32	10	4	0.13	0.69	1ppc	1.80	6.16	521.6	1	37	9	6	0.16	0.67
1etr	2.20	7.41	508.6	—	35	9	7	0.20	0.84	1pph	1.90	6.22	428.5	1	30	7	7	0.23	0.68
1ett	2.50	6.19	428.5	—	30	7	4	0.13	0.84	1qbu	1.80	10.24	596.7	—	43	10	6	0.14	0.86
1eve	2.50	8.48	379.5	—	28	6	0	0.00	0.84	1qhi	1.90	7.30	299.3	3	22	5	6	0.27	1.00
1exa	1.59	6.30	399.4	—	29	4	4	0.14	1.00	1ql9	2.30	5.36	499.0	2	34	4	1	0.03	0.75
1ezq	2.20	9.05	458.6	—	34	10	6	0.18	0.78	1qpe	2.00	8.40	301.8	—	21	2	3	0.14	0.87
1f0r	2.10	7.66	453.5	—	31	5	2	0.06	0.76	1r09	2.90	4.90	284.4	1	21	3	1	0.05	0.95
1f0t	1.80	6.00	445.5	—	30	6	5	0.17	0.71	1syn	2.00	9.05	500.5	—	37	8	2	0.05	0.85
1f4e	1.90	2.96	269.3	—	18	3	2	0.11	0.81	1thl	1.70	6.42	476.6	—	35	11	7	0.20	0.65
1f4f	2.00	4.62	428.4	—	29	9	1	0.03	0.79	1uvs	2.80	5.40	465.6	—	32	10	2	0.06	0.80
1f4g	1.75	6.48	499.5	—	34	12	5	0.15	0.82	1uvt	2.50	7.64	383.5	—	27	8	2	0.07	0.86
1fcx	1.47	7.12	388.5	—	29	3	2	0.07	1.00	1ydr	2.20	5.52	291.4	1	20	2	2	0.10	0.92
1fcz	1.38	9.22	362.4	—	27	5	2	0.07	1.00	1yds	2.20	5.92	265.3	2	18	5	3	0.17	0.93
1fjs	1.92	9.70	524.5	—	38	9	4	0.11	0.74	1ydt	2.30	7.32	446.4	—	27	9	2	0.07	0.93
1fkg	2.00	8.00	449.6	—	33	10	2	0.06	0.66	2cgr	2.20	7.27	384.4	—	29	7	4	0.14	0.80
1fm6	2.10	7.33	357.4	—	25	7	3	0.12	0.94	2csn	2.50	4.41	285.7	—	18	4	0	0.00	0.85
1fm9	2.10	8.82	546.6	—	41	12	4	0.10	0.95	2pcp	2.20	8.70	243.4	—	18	2	1	0.06	0.95
1frb	1.70	7.77	419.4	—	29	5	3	0.10	0.94	2qwi	2.00	8.40	341.4	—	24	6	10	0.42	0.91
1g4o	1.96	8.68	290.3	—	20	4	4	0.20	0.68	3cpa	2.00	4.00	238.2	—	17	5	7	0.41	0.96
1gwx	2.50	7.30	581.9	—	38	11	3	0.08	0.96	3erk	2.10	5.12	338.4	—	25	3	3	0.12	0.81
1h1p	2.10	4.92	247.3	—	18	3	3	0.17	0.91	3ert	1.90	9.60	387.5	—	29	9	2	0.07	0.87
1h1s	2.00	8.22	402.5	—	28	6	5	0.18	0.85	3std	1.65	11.11	364.4	1	28	6	2	0.07	1.00
1h9u	2.70	8.52	363.5	—	27	3	3	0.11	0.99	3tmn	1.70	5.90	303.4	—	22	6	7	0.32	0.70
1hdq	2.30	5.82	224.2	—	16	4	5	0.31	0.89	4dfr	1.70	8.62	454.5	1	33	9	8	0.24	0.76
1hfc	1.56	8.15	349.4	—	25	9	8	0.32	0.70	4std	2.15	10.33	338.2	2	20	3	3	0.15	1.00
1hpv	1.90	9.22	505.6	1	35	12	5	0.14	0.95	5std	1.95	10.49	375.4	2	28	5	2	0.07	1.00
1htf	2.20	8.09	574.7	1	41	12	6	0.15	0.75	5tln	2.30	6.37	323.3	—	23	8	7	0.30	0.76
1i7z	2.30	6.40	303.3	2	22	5	2	0.09	0.94	7dfr	2.50	4.96	441.4	—	32	9	6	0.19	0.81
1i8z	1.93	9.82	471.6	—	30	5	5	0.17	0.73	7est	1.80	7.60	441.4	—	30	9	3	0.10	0.69
1if7	1.98	10.52	371.4	—	26	6	4	0.15	0.67	830c	1.60	9.28	425.9	—	28	6	5	0.18	0.81
1ly7	2.00	6.19	244.3	—	16	5	8	0.50	0.96	966c	1.90	7.64	391.4	—	27	6	5	0.19	0.80

Res, crystallographic resolution; PKI, $-\log_{10}K_i$; MW, molecular weight of the ligand; WT, number of structural water molecules retained in the binding pocket; HA, number of heavy atoms of the ligand; RB, number of rotors of the ligand (amide bonds not counted as rotors); HB, number of hydrogen bonds between protein and ligand in the complex (metal coordination included); HB/HA, degree of hydrogen bonding; BF, fraction of the solvent-accessible surface area of the ligand that is buried upon binding.

In the present work, the binding pocket of the receptor was defined using the crystallographic coordinates of the ligand as a reference. For each complex, the ligand input structure was generated with Corina[36] (Molecular Networks GmbH), and the protein structure, prepared as described in the previous section, was used as a receptor input structure. The Monte Carlo (MC) docking runs were performed using an MC thoroughness setting of 3, which controls the length of the run, and the top 20 poses were generated. Subsequent energy minimization of the ICM-generated poses was performed with Macromodel (v. 8.1) using both MMFF[37–39] and OPLS-AA[33] force fields, with flexible ligand and rigid receptor. Conjugate gradient minimization was performed for 1000 steps. The strain energy of the minimized ligand poses was calculated with a two-step procedure: restrained minimization of the ligand geometry (half-width of flat bottom restraint = 0.5 Å, force constant = 500 kcal/mol/Å) to convergence (0.01 kJ/Å/mol) followed by removal of the constraints and full minimization until convergence (0.01 kJ/Å/mol) into the closest local minimum.[9] The refined poses were reranked based on the calculated interaction energy (van der Waals and electrostatic) minus the strain energy of the ligand conformation.

Glide (Schrodinger, Inc.). The Glide (Grid-Based Ligand Docking With Energetics) algorithm[27] approximates a systematic search of positions, orientations, and conformations of the ligand in the receptor binding site using a series of hierarchical filters. The shape and properties of the receptor are represented on a grid by several different sets of fields that provide progressively more accurate scoring of the ligand pose. The fields are computed prior to docking. The binding site is defined by a rectangular box confining the translations of the mass center of the ligand. A set of initial ligand conformations is generated through exhaustive search of the torsional minima, and the conformers are clustered in a combinatorial fashion. Each cluster, characterized by a common conformation of the core and an exhaustive set of side-chain conformations, is docked as a single object in the first stage. The search begins with a rough positioning and scoring phase that significantly narrows the search space and reduces the number of poses to be further considered to a few hundred. In the following stage, the selected poses are minimized on precomputed OPLS-AA van der Waals and electrostatic grids for the receptor. In the final stage, the 5–10 lowest-energy poses obtained in this fashion are subjected to a Monte Carlo procedure in which nearby torsional minima are examined, and the orientation of peripheral groups of the ligand is refined. The minimized poses are then rescored using the GlideScore function, which is a more sophisticated version of ChemScore[28] with force field–based components and additional terms accounting for solvation and repulsive interactions. The choice of the best pose is made using a model energy score (Emodel) that combines the energy grid score, GlideScore, and the internal strain of the ligand.

In the present work, the binding region was defined by a 12 Å × 12 Å × 12 Å box centered on the mass center of the crystallographic ligand to confine the mass center of the docked ligand. Protein and ligand input structures were prepared as described in the previous section. No scaling factors were applied to the van der Waals radii. Default settings were used for all the remaining parameters. The top 20 docking poses were energy-minimized with Macromodel using both the OPLS-AA and MMFF force fields, and reranked as described in the previous section.

GOLD (Cambridge Crystallographic Data Centre). The GOLD (Genetic Optimization for Ligand Docking) program uses a genetic algorithm (GA) to explore the full range of ligand conformational flexibility and the rotational flexibility of selected receptor hydrogens.[15,23] The mechanism for ligand placement is based on fitting points. The program adds fitting points to hydrogen-bonding groups on protein and ligand, and maps acceptor points in the ligand on donor points in the protein and vice versa. Additionally, GOLD generates hydrophobic fitting points in the protein cavity onto which ligand CH groups are mapped. The genetic algorithm optimizes flexible ligand dihedrals, ligand ring geometries, dihedrals of protein OH and NH_3^+ groups, and the mappings of the fitting points. The docking poses are ranked based on a molecular mechanics–like scoring function, which includes a hydrogen-bond term, a 4-8 intermolecular van der Waals term, and a 6-12 intramolecular van der Waals term for the internal energy of the ligand.

In the present work, the binding site was defined as a spherical region of 10 Å radius centered on the mass center of the crystallographic ligand. Protein and ligand input structures were prepared as described above. Default GA settings number 4[23] were used for all calculations, with the exception that 20 GA runs were performed instead of 10. The top 20 docking poses were energy minimized with Macromodel in both OPLS-AA and MMFF force fields and reranked as described above.

TABLE II. Composition of Test Sets Used in Enrichment Studies and Cutoffs Implemented for Rotatable Bonds (RB) and Molecular Weights (MW)

Target	RB Cutoff	MW Cutoff	No. Actives
HIV-1 protease	12	600	206
IMPDH	8	500	142
p38	8	500	247

Simulated Virtual Screenings

Three targets with known high-resolution crystal structure were used in this study: HIV-1 protease, inosine monophosphate dehydrogenase (IMPDH), and p38 MAP kinase. Simulated virtual screening was performed on each target using test sets of 10,000 compounds, with N actives selected from Vertex research programs and 10,000 − N decoys selected from commercial databases. The experimental Ki's of the active compounds range from low nanomolar to high micromolar, with a few subnanomolar ligands included for p38. The selection of decoys was biased toward drug-like molecules using filters for functional groups and cutoffs for molecular weight and number of rotatable bonds. Composition of test sets and cutoffs applied are summarized in Table II. Importantly, active compounds and decoys were selected with a similar distribution of molecular weight, in order to minimize the effects of the notorious tendency of most scoring functions to favor larger molecules. Each test set was docked into the target crystal structures with the ICM, Glide, and GOLD programs, according to the procedures described in the previous sections. Energy-minimization was performed on the

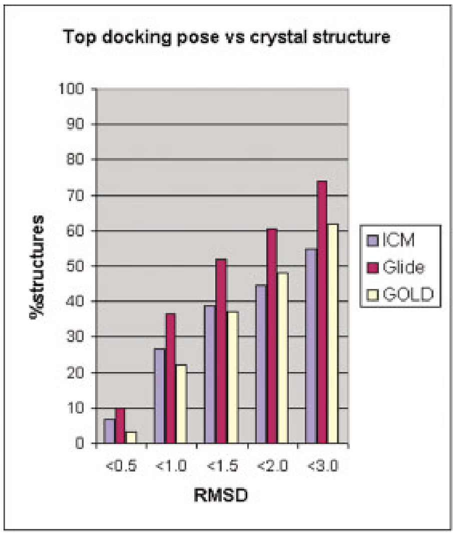

Fig. 1. Distribution of the RMSDs between the top-ranked docking poses and the corresponding crystal structures. The RMSDs were calculated on the coordinates of the heavy atoms of the ligands. *x* axis: RMSD cutoffs; *y* axis: percentage of top-ranked docking poses within a given RMSD cutoff from the crystallographic pose.

docking poses with Macromodel employing the OPLS-AA force field as described in the Docking section. Both nonminimized and minimized docking poses were rescored with ChemScore, GlideScore, and OPLS-AA interaction energy, the latter corrected by the strain energy of the ligand. The enrichment factors obtained with the three scoring methods were finally calculated on each set of poses, thus assessing the performance of the various docking/scoring combinations, as well as the impact of energy minimization.

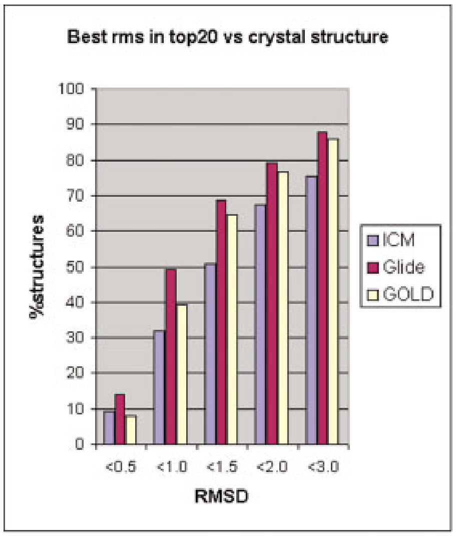

Fig. 2. Distribution of the RMSDs between the closest of top 20 docking poses (lowest deviation) and the corresponding crystal structure for each complex. *x* axis: RMSD cutoffs; *y* axis: percentage of closest docking poses within a given RMSD cutoff from the crystallographic pose.

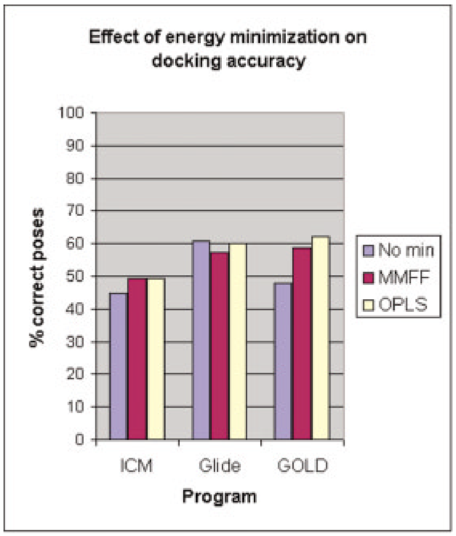

Fig. 3. Percentage of top-ranked docking poses within 2.0 Å from the experimental structure before and after minimization and reranking of the top 20 poses. *x* axis: pose generation method; *y* axis: percentage of top-ranked docking poses within the 2.0 Å cutoff.

RESULTS AND DISCUSSION

1. Evaluation of Docking Programs

General performance

The results of this study clearly identified Glide as the most accurate of the three docking programs examined, with 61% of the top-ranking poses within 2.0 Å of the corresponding crystal structure. Both GOLD and ICM also performed reasonably well, with 48% and 45% of top-ranking poses meeting the same criterion, respectively. The percentages of top-ranked solutions within a defined

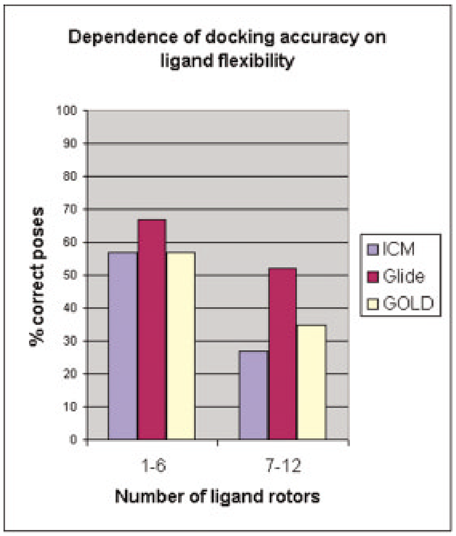

Fig. 4. Performance of the three docking programs on complexes with lower and higher ligand flexibility. *x* axis: range of ligand flexibility; *y* axis: percentage of top-ranked docking poses within 2.0 Å from the corresponding crystal structures.

RMSD from the experimentally determined structure are reported in Figure 1.

Analysis of the top 20 docking solutions produced by each program shows that GOLD was generally as effective as Glide in sampling the correct pose and placing it in the top 20. When the top 20 docking poses were compared to the corresponding crystal structure, the percentages of best poses (lowest RMSD) within 2.0 Å from the experimental structure were 79% for Glide and 77% for GOLD (see Fig. 2). Based on this observation, the GOLD algorithm appears to be equally efficient in terms of sampling, but the Glide docking function seems more accurate than the GOLD fitness function in ranking the sampled poses. The ICM algorithm appears to perform less well than the other two in terms of sampling, while the ICM docking function is better at ranking poses than the GOLD function but not as accurate as the Glide function.

It is important to point out that, in terms of thoroughness of sampling, the default settings for Glide are clearly defined by Schrodinger, Inc. and extensively validated on many test systems, while for GOLD there are four different sets of GA parameters defined as default and corresponding to different degrees of thoroughness and CPU consumption. ICM allows the user to specify the degree of thoroughness as well; however, literature and documentation do not provide a strong indication as to what settings to use on a routine basis. In order to perform the study in an objective manner we used settings that correspond to similar computing times. The docking studies described in this work averaged 1–3 min/molecule depending on processor speed on Linux (from 900 Mhz Intel Pentium III to 2.4 Ghz Intel Pentium IV) for all three programs. Both Glide and ICM require a precalculated set of grid potentials, with average computing times of 30–60 min per protein for Glide and 5–10 min for ICM.

Impact of energy minimization

The effect of energy minimization of the top 20 poses and reranking of the minimized poses was investigated using two different force fields. Comparison of the percentages of top-ranked structures within 2.0 Å of the crystal structure before and after energy minimization, reported in Figure 3, shows that minimization and reranking did not affect the accuracy of the Glide poses when the OPLS-AA force field was employed, while there was a slight decrease in performance relative to the unminimized poses when MMFF was used. Minimization and reranking marginally improved the accuracy of the ICM poses (from 45% to 49% with either force field), while there was a more significant improvement on the GOLD poses, especially when OPLS-AA was used (from 48% to 62%). The performance of GOLD equaled that of Glide when this additional procedure was applied. The effect of minimization is consistent with the features of the three docking programs examined. In Glide, minimization on an OPLS-AA potential energy grid is already performed in the final stages of docking. It is therefore not surprising that additional refinement with an all-atom minimization using the same force field does not result in an increase of the docking accuracy.

Energy-minimization is also performed as part of the ICM search protocol, but with a different force field. The slight improvement observed upon minimization with either MMFF or OPLS-AA may suggest that either these two force fields provide a more accurate description of the protein–ligand interactions than the ECEPP/3 force field implemented in ICM, or simply that the minimization performed by ICM is not as thorough and the docking poses require further refinement. The significant improvement of the GOLD poses after minimization is consistent with the fact that there is no energy minimization involved at the docking stage in this program. Severe clashes between protein and ligand atoms are not uncommon in GOLD-generated poses, partly because of the softness of the repulsive term implemented in the fitness function, and further refinement in a more rigorous fashion appears to be highly beneficial in this respect. In terms of force fields, the performances of MMFF and OPLS-AA were similar, with OPLS-AA achieving a slightly better accuracy in two out of three cases, and equal accuracy in the third (see Fig. 3). This observation further validates the choice of OPLS-AA as the force field used by Glide, which is also the best performing docking program in this study. It is important to mention that the average computing time for the energy minimization step ranges from 30 to 60 s/pose with the settings used in this work. The cost/benefit ratio should therefore be carefully evaluated when this extra step is considered.

Correlations between active site features and docking accuracy

In order to assess where the difference between Glide and the other programs lies and on what kinds of systems each program performs best, the dependence of the docking accuracy on specific structural descriptors was analyzed. The complexes were classified in a binary or ternary fashion with respect to three structural features: flexibility of the ligand, predominant nature of the interactions between ligand and receptor, and degree of solvent exposure of the binding pocket. Statistical analysis of the docking accuracies was performed with regard to such features.

In terms of flexibility, it is well known that the accuracy of any docking program decreases with the number of rotatable bonds of the ligand. The size of the conformational space to be sampled increases exponentially with ligand flexibility, and the thoroughness of the sampling has to be partially sacrificed to keep the computing time within reasonable limits. Different algorithms use different methods to circumvent the problem and maximize the efficiency of the conformational sampling. In this study the test systems were divided in two groups: 87 complexes of ligands with 1–6 rotatable bonds and 63 complexes of ligands with 7–12 rotatable bonds. The results, illustrated in Figure 4, show that the loss of accuracy going from less flexible to more flexible ligands is relatively small for Glide (from 67% to 52% of correct solutions) and much more dramatic for GOLD and ICM, with the latter losing more than half of its predictive power. This indicates that the multistage systematic algorithm implemented in Glide results in a more extensive coverage of conformational space than both the genetic algorithm and the stochastic search implemented in GOLD and ICM when runs of similar timing are considered. This partially explains the relative performances observed on the complete test set.

In terms of interactions, hydrogen bonds and hydrophobic interactions are considered the main contributors to protein–ligand binding in the vast majority of complexes. In order to divide the complexes in our test set between hydrogen bond–driven and hydrophobic-driven, the number of hydrogen bonds between protein and ligand in each complex was determined. The degree of hydrogen bonding (DHB), defined here as the ratio between number of hydrogen bonds and number of heavy atoms in the ligand, was used to define the dominant contributor to binding for each complex. Complexes with a DHB of 0.15 or higher were classified as hydrogen bond–driven, while complexes with a DHB of 0.10 or lower were classified as hydrophobic-driven, with the remaining complexes in the intermediate category. Ligand–metal interactions were counted as hydrogen bonds for their similar nature. The results, illustrated in Figure 5, show that all programs perform best on complexes in which there is a relatively even balance between hydrogen bonding and hydrophobic interactions. Interestingly, for both ICM and GOLD, the docking accuracy decreases dramatically when binding is mainly driven by hydrophobic interactions, while Glide, which appears to be somewhat less sensitive to the nature of binding, performs better on hydrophobic-driven complexes than on hydrogen bond–driven complexes. The preference of GOLD for complexes rich in hydrogen bonds has been pointed out previously,[15] and it can be ascribed to the nature of the algorithm, in which the mapping of hydrogen bond fitting points plays a major role. In the case of ICM, this tendency has not been reported; one possible explanation is that in a Monte Carlo search, mostly characterized by low-energy moves, the presence of a set of hydrogen bonds may lock part of the molecule into its correct orientation during the search, thus allowing for a more efficient sampling of the rest of the molecule. For Glide, the difference in performance is less significant, and this consistency across active sites with various degrees of hydrophobicity/hydrophilicity is another reason for its better performance on the complete test set.

When interactions with metals were specifically considered, no difference in performance was observed among the three programs: on the 24 metal-containing complexes, Glide selected a solution within 2.0 Å of the experimental structure 9 times, while ICM and GOLD succeeded 8 times in the same subset. The success rate of the three programs on such systems was significantly poorer if compared to the overall performance, which points to the necessity of further progress in this area, especially considering the continued interest in zinc metalloproteins as drug discovery targets.

The third aspect analyzed in this context is the impact of the degree of burial of the binding pocket on the docking accuracy achieved with different search algorithms. It is generally the case that buried binding sites restrict the number of orientations, positions, and conformations accessible to putative binders, but at the same time, they require a finer sampling in order to achieve the proper set of interactions without clashes. On the other hand, solvent-exposed sites require more extensive sampling to cover all the accessible poses, but at the same time are more tolerant with respect to the combination of pose descriptors required to achieve the proper set of interactions. In this study, the binding sites of the test complexes were divided into three groups, with low, medium, or high degree of burial, and the docking results were dissected accordingly. In order to assign the complexes to each group, the solvent-accessible surface area of the crystallographic ligand was calculated in the presence and in the absence of the bound protein partner, and the fraction of buried ligand was determined for each complex. The degree of burial was defined as low if the fraction was 0.75 or lower, high if the fraction was 0.90 or higher, and medium for values in between. The analysis of the performances attained by the three programs on each class, summarized in Figure 6, shows that all of them achieve the highest degree of accuracy on complexes with buried binding pockets, and consistently lose accuracy with an increase in solvent exposure. Once again, Glide appears to be relatively less sensitive to the features of the binding pockets, while ICM shows the largest decay in performance going from buried to solvent-exposed pockets. These results indicate that all three search algorithms can explore an enclosed binding site much more efficiently than a relatively open one, and also points to the obvious observation that, in a more sterically constrained site, the best pose for a given ligand is more unequivocally defined by the shape of the site. As a consequence, the likelihood of generating multiple poses with similar score is much lower and the selection of the best pose is more straightforward. For the same reason, it is safe to say that, when docking compounds in a buried binding pocket, an efficient sampling process may be more important than an accurate scoring/ranking method, while in a solvent-exposed pocket, both aspects become equally important.

Analysis of problematic structures

In addition to the general trends observed, the results of this study highlight some limitations and shortcomings that are common to all docking programs examined. In 12 cases, none of the top 20 poses generated by any of the three programs was within 2.0 Å of the experimental structure. Most of these common failures can be ascribed to a combination of structural features that make it especially challenging for any docking program to identify the right solution. Four of the problematic complexes (1cet, 1k1j, 1nhu, 1nhv) are characterized by a dominance of hydrophobic interactions in solvent-exposed sites. In such cases, the shape of the pocket does not help to restrict the number of possible binding orientations, and the lack of a set of specific anchoring points for the ligand makes the selection of the best pose very challenging. Moreover, all four ligands are relatively flexible (8–9 rotatable bonds),

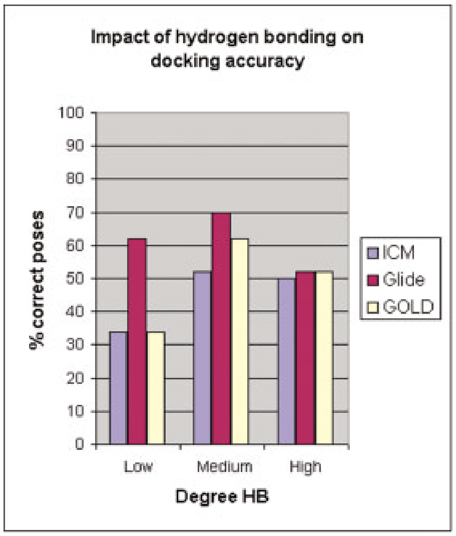

Fig. 5. Performance of the three docking programs on complexes with different degrees of hydrogen bonding between ligand and protein. The degree of hydrogen bonding (DHB) is defined as the ratio between the number of hydrogen bonds between ligand and protein and the number of heavy atoms in the ligand. *x* axis: DHB (low: DHB ≤ 0.10; medium: $0.10 <$ DHB < 0.15; High: DHB ≥ 0.15); *y*-axis: percentage of top-ranked docking poses within 2.0 Å from the corresponding crystal structures.

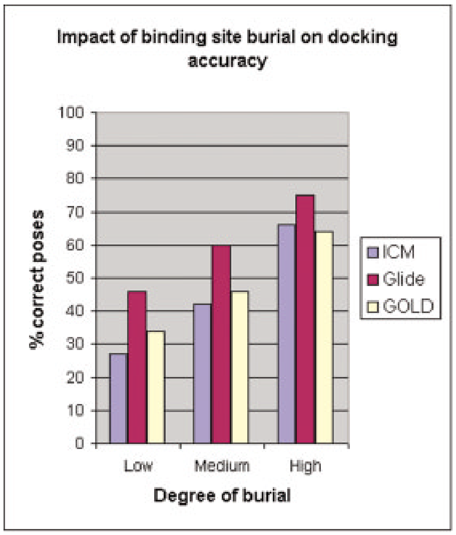

Fig. 6. Performance of the three docking programs on complexes with different degrees of binding site burial. The degree of burial is defined as the fraction of the solvent-accessible surface area of the ligand that becomes buried upon binding. *x* axis: degree of burial; *y* axis: percentage of top-ranked docking poses within 2.0 Å from the corresponding crystal structures.

which adds to the sampling problem. Three complexes (1qbu and two HIV-1 protease complexes from the Vertex collection) present highly flexible ligands in almost completely buried binding sites. In these cases the tightness of the binding pockets and the very specific conformational requirements for the ligands to achieve the correct pose call for a very thorough sampling process, which is very hard to attain within the boundaries of a limited computing time. Another aspect that is sometimes problematic in docking is the presence of charged functionalities in the ligand, because the desolvation energy required for such groups to become available for interaction with the protein is overlooked by most docking functions. In two of the failed complexes (1cet and 1i8z) there is a basic amine in the ligand that does not interact with any protein residue when the crystal complex is analyzed; docking functions tend to favor poses in which such groups form hydrogen bonds and/or salt bridges. Docking accuracy can also be impaired by the occurrence of unconventional interactions, not properly parameterized in the fitness functions of the programs employed. Two examples are hydrogen bonds between hydrogens of electron-poor aromatic rings and protein acceptors, as observed in one of the Vertex complexes, and hydrogen bonds between the imino form of anilino nitrogens and protein donors, as observed in 1jsv. Both complexes were not reproduced by any of the docking programs. Finally, there are cases in which the interactions between ligand and protein in the experimental pose are tighter than average and predominantly hydrophobic. Imperfect refinement of the crystal structure or the presence of legitimate short-range interactions can introduce apparent clashes that are not compensated by other obvious interactions in the crystallographic pose. When such poses are evaluated in the context of docking, they receive unfavorable scores since they are not properly treated by any known docking functions. This provides a partial explanation for the remaining failures (3ert and one of the kinase complexes from the Vertex collection).

Summary of relevant findings

The Glide program is shown to have the highest degree of accuracy on a wide and diverse set of systems, which makes it the tool of choice in most cases. Energy minimization of multiple poses is a highly beneficial postprocessing step when docking is performed with GOLD, while the improvement on ICM poses is marginal. Minimization has no impact on the accuracy of the Glide-generated poses, and the combination GOLD docking/OPLS-AA minimization appears to be as reliable a predictor as Glide. The Glide program is more tolerant than both ICM and GOLD of the increase of ligand flexibility, which seems to point to a more effective conformational sampling method. Analogously, Glide appears to be less sensitive to variations in the polarity of the binding pocket, with a slight preference for complexes with prevalent hydrophobic character but a solid performance across the board. On the other hand, ICM and GOLD can be considered as reliable as Glide when operating on highly polar binding sites, where binding is strongly driven by hydrogen bonding. Comparatively, the ability of these same two programs to predict complexes where binding is driven by hydrophobic interactions is relatively poor. All three programs perform best on buried binding pockets, with a gradual decrease in performance at the increase of solvent exposure. In general, some systems remain a challenge for docking at the current stage, which suggests that there is still a margin for improvement on the existing methods. In particular, the inclusion of properly weighted solvation terms and a

TABLE III. HIV-1 Protease: Enrichment Factors Calculated on Top 3% of Ranking for All Pose Generation/ Scoring Combinations

	ChemScore	GlideScore	OPLS-AA
Glide	10.5	3.7	1.5
GOLD	8.3	1.1	0.8
ICM	9.9	2.4	0.3
Glide/min	9.4	3.6	6
GOLD/min	10.2	8.3	3.9
ICM/min	10.5	6.5	8.3

The rows contain the pose generation methods and the columns scoring methods.

TABLE IV. IMPDH: Enrichment Factors Calculated on Top 3% of Ranking for All Pose Generation/Scoring Combinations

	ChemScore	GlideScore	OPLS-AA
Glide	6.3	16.4	5.4
GOLD	3.8	2.3	0.9
ICM	4.7	14.1	1.4
Glide/min	11.3	17.6	12.4
GOLD/min	4	6.1	4.7
ICM/min	11.3	16.2	10.3

The rows contain the pose generation methods and the columns scoring methods.

TABLE V. P38 MAP Kinase: Enrichment Factors Calculated on Top 3% of Ranking for All Pose Generation/ Scoring Combinations

	ChemScore	GlideScore	OPLS-AA
Glide	5.7	8.8	5.8
GOLD	4.6	8.5	5.7
ICM	2.2	9	9
Glide/min	7.4	9	5.9
GOLD/min	3.8	12	6.9
ICM/min	4.6	10	6.9

The rows contain the pose generation methods and the columns scoring methods.

more effective representation of metal-mediated interactions in the fitness functions appear to be highly desirable.

2. Evaluation of Docking/Scoring Combinations for Virtual Screening

The objective of a virtual screening is to select a subset enriched in compounds with the desired activity relative to the entire collection. When the percentage of active compounds in the screening set is known or can be reliably estimated, the success can be described by the enrichment factor, which is the ratio between the percentage of active compounds in the selected subset and the percentage in the entire database. In a real-life virtual screening it is common practice to select the top portion of the ranked compounds for further evaluation, but the size of such portion is somewhat arbitrary, generally ranging from 1% to 10% of the entire ranking. The calculated enrichment factors are dependent upon the fraction of the ranking considered, and the relative enrichments achieved by two different methods may vary throughout the ranking. In order to provide a complete and unbiased account of the performance of each method, the results of the virtual screenings are presented here in two different formats. Tables III–V report the enrichment factors calculated on the three targets for each pose generation/scoring combination on the top 3% of the corresponding rankings. Figure 7 illustrates the performance of each method on the three targets throughout the top 30% of the corresponding rankings. Each panel represents the results of the calculations performed on one particular target with one particular program used at the docking stage. The *x* axis denotes the percentage of database sampled, or portion of the ranking examined, while the *y* axis denotes the percentage of active compounds correctly identified in that portion. Whenever the relative performances of different methods were different for different top portions of the ranking examined (e.g., method A achieved better enrichment than method B in the top 3%, worse enrichment in the top 10%), the enrichment in the top 3% will be used as the main indicator of performance in this report. The nature of the three active sites used in this study is significantly different, and the relative performances of the methods examined varied as a function of these differences. The results obtained on each target are summarized below.

HIV-1 protease

In HIV-1 protease the binding site is buried and predominantly hydrophobic, with an oblong shape suited for large and flexible ligands. The restrictive size and shape of this binding site make this enzyme a very challenging system for docking. Additionally, there is a conserved catalytic water molecule that is an integral part of this active site, contributing to the challenging nature of this system, since interactions with water are generally not handled accurately by most docking/scoring functions. As a demonstration of this, in the study described in the Docking section, all three programs performed poorly on the 9 HIV-1 protease complexes included in the test set.

In the virtual screening simulation on this system, ChemScore consistently achieved the best enrichment, regardless of the pose generation method (see Fig. 7, panels A–C, and Table III). The relative insensitivity of this function to repulsive interactions was probably beneficial in a system where a large amount of sampling would be necessary to generate an accurate docking result, and even otherwise correct docking poses may still contain severe clashes between protein and ligand atoms. The performance of ChemScore was largely unaffected by energy minimization of the docking poses, which is consistent with the fact that the attenuation of unfavorable interactions has limited impact on the scores. For this system, the combination ICM/ChemScore achieved the best enrichments, but the performances of Glide/ChemScore and GOLD/ChemScore were comparable.

Energy minimization of the docking poses dramatically improved the enrichments achieved by OPLS-AA and

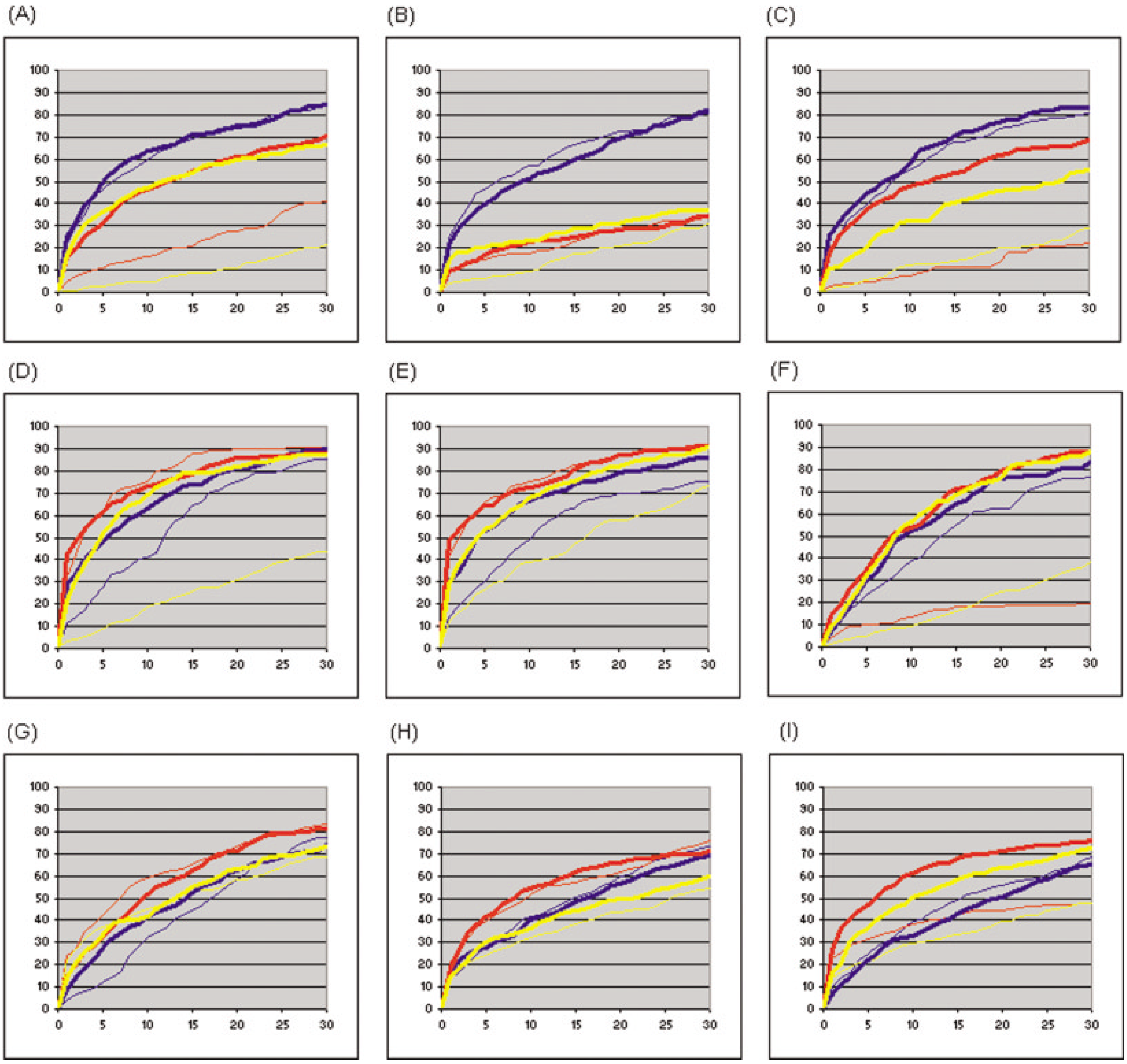

Fig. 7. Performance of different combinations pose generation/scoring method in the simulated virtual screenings on HIV-1 protease, IMPDH, and p38. Each panel illustrates the results obtained on one particular system using one particular docking program for pose generation. Target name and docking program utilized for pose generation are as follows. **A**: HIV-1 protease/ICM; **B**: HIV-1 protease/Glide; **C**: HIV-1 protease/GOLD; **D**: IMPDH/ICM; **E**: IMPDH/Glide; **F**: IMPDH/GOLD; **G**: p38/ICM; **H**: p38/Glide; **I**: p38/GOLD. The *x* axis shows the percentage of the global ranking considered, while the *y* axis shows the percentage of active compounds correctly identified in that fraction of the ranking. Each line color corresponds to a different scoring method: blue, ChemScore; red, GlideScore; yellow, OPLS-AA. Thin lines represent the results obtained on the unminimized docking poses; thick lines represent the results obtained on the energy-minimized docking poses.

GlideScore on both ICM and GOLD poses, while only OPLS-AA improved on Glide poses, and less significantly. Both GlideScore and OPLS-AA performed well on minimized ICM poses, while their performance on both unminimized and minimized Glide poses was surprisingly modest. In terms of pose generation, unminimized poses generated by the three programs achieved similar enrichments, with Glide poses slightly ahead of ICM and GOLD poses, while after minimization, ICM and GOLD poses produced better overall enrichments than Glide poses. This last observation is consistent with the differential benefits of postdocking minimization reported in the docking section for the three programs.

IMPDH

In IMPDH the binding site is relatively polar and predominantly solvent exposed, but it contains a narrow cavity at the bottom that often accommodates hydrophobic moieties of ligands sandwiched between the cofactor and protein residues. The tightness of such cavity requires a fine sampling to correctly place ligands, and makes docking challenging. The presence of a cofactor represents an additional challenge, since in many docking functions the interactions involving cofactors are not accurately parameterized. None of the three programs evaluated in this study were particularly effective at reproducing the four IMPDH complexes present in the docking test set. On this target, GlideScore consistently achieved the highest enrichment, regardless of the pose generation method, with the only exception being unminimized GOLD poses, on which ChemScore did better (see Fig. 7, panels D–F, and Table IV). In general, however, all of the scoring functions considered achieved modest enrichments on this particular set of poses. ChemScore consistently outperformed OPLS-AA on unminimized poses. Energy minimization improved the performance of ChemScore on ICM and Glide poses, and dramatically improved the performance of OPLS-AA on all sets of poses, raising it to the same level of ChemScore. Minimization dramatically improved the performance of GlideScore on GOLD poses, but only marginally improved the performance of GlideScore on ICM and Glide poses. In terms of pose generation methods, slightly better enrichments were achieved on Glide poses relative to ICM poses, and both docking methods clearly outperformed GOLD, especially in the absence of energy minimization.

The improvement observed in most cases upon minimization highlights the importance of refinement of the docking poses when the active site contains a narrow region important for binding, and more extensive sampling may be required. Refinement is more important when the docking protocol does not include a minimization component, as it is the case for GOLD. The great benefits of minimization on GOLD-generated poses confirm the observations reported in the docking section on this program.

p38 MAP kinase

The binding site of p38 is relatively buried and mostly hydrophobic, and it usually accommodates ligands through a combination of well-defined shape complementarity and hydrogen bonds with backbone amide groups. These features and the absence of narrow subpockets make of p38 a good system for docking, as confirmed by the fact that all the three docking program were quite accurate in reproducing the six p38 complexes present in the docking test set.

On this system, once again GlideScore showed the most consistent performance, only matched by OPLS-AA on ICM poses (see Fig 7, panels G–I and Table V). OPLS-AA performed relatively well on all three sets of unminimized poses, with enrichments equivalent to or better than ChemScore. Energy minimization did not significantly affect the enrichments on Glide and ICM poses regardless of the scoring method, while it marginally improved the enrichments achieved with GlideScore and OPLS-AA on GOLD poses. This is consistent with the observation that on this system the docking poses are generally correct even in the absence of minimization. Interestingly, the best enrichment was achieved in this case by the combination GOLD/GlideScore, while in general the performance of the three pose generation methods can be considered equivalent.

In order to explore all possibilities, the GOLD poses were also ranked on all three systems using the GOLD fitness function, and no significant enrichment was achieved in any of the three cases. ICM contains its own empirical function to rescore docked poses in a virtual screening context.[40] The score must be recalculated at the end of the docking stage, and it is relatively time-consuming (up to 1 min per pose on a Pentium III 1 Ghz). Since the function had not shown satisfactory results when previously tested on internal targets, we decided not to include it in this study.

Overall evaluation

The results obtained on the three systems confirm that, as recently stated,[7,8,30] a universal docking/scoring combination that outperforms all the others on every system does not exist. Nevertheless, this study suggests that some combinations do achieve more consistent performances, and it provides a set of useful guidelines on how to select and use the currently available tools. The three binding pockets used here did not differentiate the performances of the three docking programs as much as the diverse set of complexes used in the docking section. Careful analysis of the results indicates that comparable enrichments across different systems were achieved when Glide and ICM poses were rescored with different methods, with the Glide poses performing slightly better on tighter binding sites. Both Glide and ICM clearly outperformed GOLD in the most challenging systems, while GOLD poses achieved comparable enrichments only in the “easier” binding site of p38. The similar performances of Glide and ICM could be viewed as inconsistent with the results described in the docking section, but different aspects come into play when docking programs are engaged in virtual screening. A program must be able to fit real binders to an active site conformation that is not necessarily optimal for them, and at the same time minimize the occurrence of false positives (i.e., inactive compounds that are docked into the active site and favorably scored). Neither of these issues is present when the crystallographic ligand is docked back into the cognate active site, and the ways each program addresses them contribute to its ability to generate appropriate poses in a docking-based virtual screening.

Analysis of the poses generated by Glide and ICM in different systems shows that, when an ideal fit cannot be achieved, Glide tends to generate more strained ligand conformations in order to maximize the interactions with the protein (data not presented), while ICM tends to produce more stable ligand conformations at the expense of less optimal intermolecular contacts. The aggressive approach applied by Glide, which attempts to reconcile induced fit with the rigid receptor approach by softening some of the intramolecular repulsive interactions, increases the probability of finding active compounds but also the occurrence of false positives. The conservative and more rigorous approach applied in ICM entails a higher risk of missing actives but also minimizes the incidence of false positives. The interplay of these different tendencies, in conjunction with the intrinsic effectiveness of the search algorithms, the nature of the active site, and the affinity of the actual ligands in the screened collection, determines the relative performance of the two programs. The results obtained here indicate that the vigorous approach of Glide can be more successful on tight binding pockets, where the amount of sampling required to achieve the correct set of interactions with the correct conformation of the ligand may be quite large. In more spacious active sites, the conservative approach followed by ICM can be equally or more effective.

Energy minimization of the docked poses in OPLS-AA force field seems to significantly improve the enrichment on systems with tight binding pockets, while it has basically no impact on the enrichments achieved on the more spacious pocket of p38. This suggests that energy minimization should be common practice when docking-based virtual screening is performed on sterically demanding pockets, while in the absence of a clear classification of the binding pocket it should always be applied. Glide poses seem to generally benefit less from this postprocessing step, and this may be ascribed to the fact that the same force field is already implemented in the docking protocol, although in a less rigorous fashion (especially with respect to the intramolecular energy of the ligand).

In terms of scoring, GlideScore appears to be the most general scoring function tested, with the best performance across the board on IMPDH and p38. When the binding site is very tight and docking is likely to produce poses with severe intermolecular clashes, a function like ChemScore, more forgiving of repulsive interactions, can be more effective. An alternative approach, suggested by Glide developers, is to reduce the van der Waals radii of protein and ligand atoms, thus softening the repulsive terms in GlideScore along the lines discussed above.

Overall, exclusive force field–based scoring, exemplified in this study by the OPLS-AA energy function, appears to be less reliable than empirical scoring when ranking different ligands, although reasonable enrichments can be achieved in some cases, particularly on minimized docking poses and spacious binding pockets. In general, testing different combinations on the target of choice and selecting the best performing one for the real screening is highly desirable whenever sufficient data are available. In the absence of it, the protocol should be selected based on the nature of the active site, and this study suggests a choice between Glide and ICM for the docking step, energy minimization in OPLS-AA, and a choice between GlideScore and ChemScore for rescoring, according to the criteria described above.

CONCLUSIONS

The field of docking and virtual screening is in continuous evolution, and a thorough assessment of the state of the art in the field is often incompatible with the time and resources available to most computational chemists. As a consequence, the choice of the methods to be used in real-life applications is often based on the long-term acquaintance with established methods rather than on a detailed comparison between earlier and more recent tools. This work provides an up-to-date evaluation of some of the most advanced docking and scoring methods, and analyzes the advancements achieved by recently developed tools. As a result of the assessment, criteria are defined to select the best protocol for docking and virtual screening on different systems.

As pointed out in the first part of this study, an appropriate test set for evaluation of docking/scoring methods dedicated to drug discovery should be representative of the systems that are normally considered in such context. Following this premise, the generation of a large and highly curated test set of pharmaceutically relevant protein–ligand complexes with known binding affinities is described. Details are provided on the portion of the set that is based on publicly available structures, thus making it available for others to use in similar studies.

The comparison among three highly regarded docking programs (ICM, Glide, and GOLD) for the ability to reproduce crystallographic binding orientations highlights the different impact of the binding site features on the accuracy of each program. While all three programs perform well and appear to be reasonable choices for docking in buried binding sites, Glide appears to perform most consistently with respect to diversity of binding sites, ligand flexibility, and overall sampling, resulting in the best overall docking accuracy. On the overall test set, Glide correctly identified the crystallographic pose 61% of the times within 2.0 Å, versus 48% for GOLD and 45% for ICM. The three programs perform equally well on complexes strongly driven by hydrogen bonding, but both ICM and GOLD perform poorly when hydrophobic interactions are predominant. The results also show that saving the top N scoring poses from any docking function followed by energy minimization and reranking can be an effective means to overcome some of the limitations of a given docking function. This extra step is particularly beneficial when applied to GOLD-generated poses. Since the additional computing cost is no longer prohibitive (e.g., 30–60 s per pose), it may make sense to minimize as a general postdocking step.

The evaluation of the same docking programs in conjunction with three different scoring functions in simulated virtual screenings confirms that variations in the nature of the binding sites have a different impact on different docking programs and scoring functions. GlideScore appears to be an effective scoring function for database screening, with consistent performance across several types of binding sites, while ChemScore appears to be most useful in sterically demanding sites, since it is more forgiving of repulsive interactions. Energy minimization of docked poses can significantly improve the enrichments in systems with sterically demanding binding sites, although Glide poses tend to benefit less than GOLD- or ICM-generated poses.

Overall, this study indicates that a certain degree of improvement has been recently achieved both in the docking and in the scoring methodology, and in both cases the technology developed for Glide appears to provide the most consistent benefits. While Glide appears to be a safe general choice for docking, the choice of the best scoring tool remains to a larger extent system-dependent. In general, it is always advisable to test different docking/scoring combinations on a given system of interest to select the best protocol, prior to the full database screen. It should be pointed out that in studies of the kind described in this article, there is always some degree of subjectivity with regard to choice of targets, data sets, and programs employed. However, we believe that the conclusions presented here are applicable to a wide range of systems.

In future perspective, the next large improvements in this field will hopefully be in scoring functions that more accurately describe the physics of binding and allow not only for good discrimination between actives and inactives, but also between closely related analogs. In this regard, factors such as protein flexibility and solvation need to be incorporated in a meaningful and efficient manner. Since we recognize some of the current limitations, we have tried to emphasize practices that can sometimes overcome some of them, such as saving and reranking poses, and using a variety of search/scoring combinations. To be fair, we should point out that it is common, in practice, to use graphics as part of the final round of selecting compounds in docking-based virtual screening. This can certainly remove some of the shortcomings of various scoring functions (although it tends to introduce other, subjective biases). One might imagine that a visualization tool that highlights both favorable and unfavorable contributions to binding based upon a variety of scoring functions could be quite useful. Additionally, the judicious use of a variety of postprocessing filters (e.g., Did we maintain a key hydrogen bond?) and/or the use of constraints during docking are commonly used to enhance the results of docking/virtual screening and should be used whenever possible.

REFERENCES

1. Muegge I, Rarey M. Small molecule docking and scoring. In: Lipkowitz KB, Boyd DB, editors. Reviews in computational chemistry. Vol. 17. New York: Wiley-VCH; 2001.
2. Taylor RD, Jewsbury PJ, Essex JW. A review of protein–small molecule docking methods. J Comput Aided Mol Des 2002;16:151–166.
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. Shoichet BK, McGovern SL, Wei B, Irwin JJ. Lead discovery using molecular docking. Curr Opin Chem Biol 2002;6:439–446.
5. Schneider G, Bohm HJ. Virtual screening and fast automated docking methods. Drug Discov Today 2002;7:64–70.
6. Perez C, Ortiz AR. Evaluation of docking functions for protein–ligand docking. J Med Chem 2001;44:3768–3785.
7. 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.
8. Stahl M, Rarey M. Detailed analysis of scoring functions for virtual screening. J Med Chem 2001;44:1035–1042.
9. 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.
10. Verkhiver GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW. Deciphering common failures in molecular docking of ligand–protein complexes. J Comput Aided Mol Des 2000;14:731–751.
11. Stahl M, Bohm HJ. Development of filter functions for protein–ligand docking. J Mol Graph Model 1998;16:121–132.
12. Charifson PS, Walters WP. Filtering databases and chemical libraries. J Comput Aided Mol Des 2002;16:311–323.
13. 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.
14. 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.
15. 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.
16. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB. Consensus scoring for ligand/protein interactions. J Mol Graph Model 2002;20:281–295.
17. Paul N, Rognan D. ConsDock: a new program for the consensus analysis of protein–ligand interactions. Proteins 2002;47:521–533.
18. Waszkowycz B. Structure-based approaches to drug design and virtual screening. Curr Opin Drug Discov Devel 2002;5:407–413.
19. Pickett SD, Sherborne BS, Wilkinson T, Bennett J, Borkakoti N, Broadhurst M, Hurst D, Kilford I, MCKinnell M, Jones PS. Discovery of a novel low molecular weight inhibitors of IMPDH via virtual needle screening. Bioorg Med Chem Lett 2003;13:1691–1694.
20. Jenkins JL, Kao RY, Shapiro R. Virtual screening to enrich hit lists from high-throughput screening: a case study on small-molecule inhibitors of angiogenin. Proteins 2003;50:81–93.
21. Vangrevelinghe E, Zimmermann K, Schoepfer J, Portmann R, Fabbro D, Furet P. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. J Med Chem 2003;46:2656–2662.
22. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. Proteins 1999;37:228–241.
23. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein–ligand docking using GOLD. Proteins 2003;52:609–623.
24. Totrov M, Abagyan R. Flexible protein–ligand docking by global energy optimization in internal coordinates. Proteins 1997;Suppl 1:215–220.
25. Glide 2.5. New York: Schrodinger; 2003.
26. Abagyan R, Totrov M. High-throughput docking for lead generation. Curr Opin Chem Biol 2001;5:375–382.
27. Available online at www.schrodinger.com/docs/2003_1/pdf/firstdiscovery/fd27_technical_notes.pdf
28. 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.
29. Murray CW, Auton TR, Eldridge MD. Empirical scoring functions: II. The testing of an empirical scoring function for the prediction of ligand–receptor binding affinities and the use of Bayesian regression to improve the quality of the model. J Comput Aided Mol Des 1998;12:503–519.
30. Schulz-Gasch T, Stahl M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. J Mol Model (Online) 2003;9:47–57.
31. Roche O, Kiyama R, Brooks CL III. Ligand–protein database: linking protein–ligand complex structures to binding data. J Med Chem 2001;44:3592–3598.
32. Nissink JW, 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.
33. Jorgensen WL, Maxwell D, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
34. Egan WJ, Walters WP, Murcko MA. Guiding molecules towards drug-likeness. Curr Opin Drug Discov Devel 2002;5:540–549.
35. Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. Energy parameters in polypeptides: 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J Phys Chem 1992;96:6472–6484.

36. Gasteiger J, Rudolph C, Sadowski J. Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedrom Comp Method 1990;3:537–547.
37. Halgren TA, Merck molecular force field: I. Basis, form, scope, paramaterization, and performance of MMFF94. J Comput Chem 1996;17:490–519.
38. Halgren TA. Merck molecular force field: II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem 1996;17:520–552.
39. Halgren TA. Merck molecular force field: III. Molecular geometries and vibrational frequences for MMFF94. J Comput Chem 1996;17:553–586.
40. Totrov M, Abagyan R. Derivation of sensitive discrimination potential for virtual ligand screening. In: RECOMB '99: Proceedings of the Third Annual International Conference on Computational Molecular Biology, Lyon, France, April 11–14, 1999; Istrail S, Pevzner P, Waterman M, eds. Association for Computing Machinery, New York; 1999.