Evaluation of the Performance of Four Molecular Docking Programs on a Diverse Set of Protein-Ligand Complexes

XUN LI, YAN LI, TIEJUN CHENG, ZHIHAI LIU, RENXIAO WANG
*State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic Chemistry,
Chinese Academy of Sciences, Shanghai, People's Republic of China*

Received 6 July 2009; Revised 3 December 2009; Accepted 7 December 2009
DOI 10.1002/jcc.21498
Published online 1 February 2010 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Many molecular docking programs are available nowadays, and thus it is of great practical value to evaluate and compare their performance. We have conducted an extensive evaluation of four popular commercial molecular docking programs, including Glide, GOLD, LigandFit, and Surflex. Our test set consists of 195 protein-ligand complexes with high-resolution crystal structures (resolution $\leq$2.5 Å) and reliable binding data [dissociation constant (K_d) or inhibition constant (K_i)], which are selected from the PDBbind database with an emphasis on diversity. The top-ranked solutions produced by these programs are compared to the native ligand binding poses observed in crystal structures. Glide and GOLD demonstrate better accuracy than the other two on the entire test set. Their results are also less sensitive to the starting structures for docking. Comparison of the results produced by these programs at three different computation levels reveal that their accuracy are not always proportional to CPU cost as one may expect. The binding scores of the top-ranked solutions produced by these programs are in low to moderate correlations with experimentally measured binding data. Further analyses on the outcomes of these programs on three suites of subsets of protein-ligand complexes indicate that these programs are less capable to handle really flexible ligands and relatively flat binding sites, and they have different preferences to hydrophilic/hydrophobic binding sites. Our evaluation can help other researchers to make reasonable choices among available molecular docking programs. It is also valuable for program developers to improve their methods further.

© 2010 Wiley Periodicals, Inc. J Comput Chem 31: 2109–2125, 2010

Key words: molecular docking; GLIDE; GOLD; ligandfit; surflex; PDBbind

Introduction

Molecular docking is the very essential method in structure-based drug design as well as other studies for modeling the binding mode of a ligand relative to a receptor when they form a stable complex.[1–7] They all have a basic design that integrates a sampling algorithm, which serves as a searching engine to explore the orientational and conformational space of the given ligand, and at least one scoring method, which serves as a ranking engine to evaluate all sampled binding poses. Since 1980s, a good number of molecular docking programs have been developed, many of which are either implemented in commercial software or distributed by academic groups. It is reasonable to expect that the performance of these programs is not at the same level, or at least they should have different advantages or preferences. It is thus desired by both the users and the developers of molecular docking programs to evaluate the performance of these programs preferably in a comparative manner.

Indeed, a number of comparative studies of docking/scoring methods have been published in literature in recent years (refs. 8–21, Table 1). They can be roughly divided into two categories by their evaluation strategies. Studies in the first category focus on the abilities of docking programs of producing the correct binding mode of a ligand to its biological target (usually a protein molecule). For this purpose, they rely on some protein-ligand complexes with known three-dimensional structures, typically selected from the Protein Data Bank (PDB).[22] Each molecular docking program is applied to these complex structures, and then its performance is evaluated by how well it can reproduce the observed binding modes with certain criteria. Some of these

Additional Supporting Information may be found in the online version of this article.

Correspondence to: R. Wang; e-mail: wangrx@mail.sioc.ac.cn

Contract/grant sponsor: Chinese National Natural Science Foundation; contract/grant numbers: 20772149, 90813006

Contract/grant sponsor: Chinese Ministry of Science and Technology; contract/grant numbers: 2006AA02Z337, 2009ZX09501-002

Contract/grant sponsor: Science and Technology Commission of Shanghai Municipality; contract/grant number: 074319113

Table 1. Summary of Some Previous Comparative Evaluations of Molecular Docking Programs.

Study	Test set	Subset classification criteria	Docking solutions for evaluation	Performance rank
Category I. Evaluations through molecular docking trials				
Bissantz et al.[16]	10 complexes of thymidine kinase and 10 complexes of estrogen receptor	None	Top-ranked solution	GOLD > FlexX ≈ DOCK
Bursulaya et al.[8]	37 protein-ligand complexes formed by 11 different proteins	Protein family	Top-ranked solution	ICM > GOLD > AutoDock > FlexX > DOCK
Kellenberger et al.[9]	100 diverse protein-ligand complexes from an in-house collection	Protein family Steric and electrostatic features of the ligand and of the protein cavity	Top-ranked solution; Docking solution with minimal RMSD	GLIDE > GOLD > Surflex
Perola et al.[10]	150 diverse protein-ligand complexes from the Vertex dataset	Flexibility of the ligand Predominant nature of the interactions between ligand and receptor Degree of solvent exposure of the binding pocket	Top-ranked solution; Docking solution with minimal RMSD	GLIDE > GOLD > ICM
Kontoyianni et al.[11]	69 protein-ligand complexes formed by 14 pharmaceutical targets	Protein structure resolution Number of rotatable bonds in ligand Polarity of active sites of the targets	Top-ranked solution; Docking solution with minimal RMSD	GOLD > GLIDE > others
Xing et al.[17]	Factor Xa (PDB entry 1FAX)	None	All the docking solutions	FlexX, D-Score, G-Score, ChemScore > PMF
Hu et al.[18]	40 protein-ligand complexes formed by metalloproteinase	None	Top-ranked solution Docking solution with minimal RMSD Top-ranked solution with good/fair “zinc binding group” geometry	GOLD > AutoDock > DrugScore > FlexX > DOCK
Cummings et al.[12]	31 protein-ligand complexes formed by 5 pharmaceutical targets	Protein family	Top-ranked solution	GLIDE or GOLD
Chen et al.[13]	164 diverse protein-ligand complexes selected from an in-house collection	Protein family Protein structure resolution Number of rotatable bonds on ligand	Top-ranked solution	GLIDE > GOLD > FlexX
Warren et al.[14]	136 protein-ligand complexes formed by 8 pharmaceutical targets	Protein family	Top-ranked docking solutions Docking solutions with minimal RMSD	Target-dependent
Onodera et al.[15]	116 diverse protein-ligand complexes selected from the GOLD test set	Protein family The size of binding site	Top-ranked docking solutions from 1000 repeated runs	GOLD > AutoDock > DOCK
This study (2009)	195 diverse protein-ligand complexes formed by 65 types of proteins selected from the PDBbind database	Total number of the rotatable bonds on ligand Buried percentage of the solvent-accessible surface area of ligand A “hydrophobic index” of the binding pocket	Top-ranked docking solutions Docking solutions with minimal RMSD Binding scores	GLIDE ≈ GOLD >Surflex > LigandFit

Table 1. (Continued)

Study	Test set	Performance rank
Category II. Evaluations through virtual screening trials		
Bursulaya et al.[8]	37 complexes formed by 11 different proteins 37 known ligands in these complexes + 10000 random compounds in ACD	ICM > GOLD > AutoDock > FlexX > DOCK
Kontoyianni et al.[19]	6 pharmaceutical target proteins A certain number of active compounds + 996 random compounds in ACD and MDDR	LigandFit/LigScore1 ≈ LigandFit/GOLD > FlexX/FlexX, GLIDE/LigScore1, LigandFit/PMF, LigandFit/LigScore2, DOCK/PMF on all targets
Zhou et al.[20]	18 pharmaceutical targets A certain number of active compounds for each target + 1000 random drug-like compounds from Schrödinger	GLIDE > GOLD > DOCK
Cross et al.[21]	40 diverse sets of protein targets Active and decoy compounds from the Directory of Useful Decoys (DUD)	GLIDE, Surflex > DOCK, FlexX, ICM, PhDOCK

studies attempt to evaluate the general performance of the molecular docking programs by employing test sets consisting of diverse protein-ligand complexes, such as Bursulaya et al.,[8] Kellenberger et al.,[9] Perola et al.,[10] Kontoyianni et al.,[11] Cummings et al.,[12] Chen et al.,[13] Warren et al.,[14] and Onodera et al.[15] Some other studies employed test sets containing particular types of proteins. For example, Bissantz et al.[16] used 10 complexes of thymidine kinase and 10 complexes of estrogen receptor, Xing et al.[17] focused on Factor Xa, and Hu et al.[18] employed 40 complexes formed by metalloproteinase to evaluate docking programs. This type of studies have more practical values for the researchers who are interested in these particular proteins.

Studies in the second category attempt to evaluate molecular docking programs through virtual screening trials, such as the studies by Bursulaya et al.,[8] Kontoyianni et al.,[19] Zhou et al.,[20] and Cross et al.[21] Virtual screening has been proved by many successful applications as a cost-effective strategy for the discovery of new hits.[23–25] The basic idea of virtual screening is to dock a whole library of molecules onto a given molecular target, rank them by predicted binding scores, and select only the most promising candidates for subsequent experimental tests. Accordingly, this type of evaluations typically employs a random library of molecules, for example, selected from the Available Chemical Directory or the ZINC database,[26] and mix it with some known binders to the selected target proteins. The performance of each molecular docking program under test is then judged by the enrichment factor observed in virtual screening trials, *i.e.* the percentage of known binders found among the top-ranked molecules.

Our evaluation of molecular docking programs reported here belongs to the first category, i.e., the performance of each program is judged primarily by its ability of reproducing the known binding modes of some given ligands. This is the very essential quality of a molecular docking program. As a matter of fact, most today's molecular docking programs are developed and validated merely for this purpose, although they can be applied to virtual screening jobs later on. As demonstrated in previous studies as well as ours, the performance of today's molecular docking programs is not always satisfactory even in this basic aspect. A virtual screening trial is certainly a more complex scenario. In virtual screening, one normally relies on predicted binding scores to identify the promising binders. Obviously, the binding score of a given ligand molecule cannot be correct if its predicted binding mode is not correct. As virtual screening aims at differentiating binders from nonbinders, another interesting question here is what the correct binding scores should be for nonbinders. In addition, outcomes of a virtual screening trial, such as enrichment factor, are also dependent on other factors besides the intrinsic quality of the applied molecular docking program, including the contents of the compound library as well as the characteristics of target proteins. Thus, conclusions given by this type of studies are context-dependent and sometimes even conflicting. For example, Warren et al.[14] reported that GOLD produced higher enrichment factors than Glide in a virtual screening trial against Factor Xa; whereas Chen et al.[13] reported that Glide outperformed GOLD against the same target in a similar virtual screening trial. Although some sorts of benchmarks for virtual screening trials have already been proposed, such as the Directory of Useful Decoys,[27] the methods adopted by those studies still need the acceptance by a wider audience.

Because so many factors may be relevant, a fair comparison of different molecular docking programs is actually difficult as properly pointed out by Cole et al.[28] Our opinion is that an evaluation of the essential qualities of molecular docking programs on a well-established benchmark will produce less misleading information. The primary aim of our study is to evaluate the general performance of molecular docking programs. For this purpose, we choose to base our evaluation on a test set of diverse protein-ligand complexes. The test set employed in our study consists of 195 protein-ligand complexes with high-resolution crystal structures (resolution <2.5 Å) and reliable binding data [dissociation constant (K_d) or inhibition constant (K_i)] selected from the PDBbind database.[29,30] In addition, this test set is constructed with a control on redundancy so that our evaluation results are in principle not biased towards any particular type of protein. The molecular docking programs evaluated in our study include four popular commercial programs, i.e., Glide[31–33] (version 4.5) in the Schrödinger Suite (version 2007, Schrödinger LLC),[34] GOLD[35–38] (version 3.2, CCDC Software), LigandFit[39,40] (version 2.3) in the Discovery Studio software (version 2.0, Accelrys Software)[41] and Surflex[42–44] (version 2.0) in the Sybyl software (version 7.3, Tripos).[45] To the best of our knowledge, no one has evaluated these four molecular docking programs in a head-to-head comparison. In our study, influence of different input structures, postoptimization, and CPU time cost on the final docking accuracy of these programs is examined systematically. The outcomes of our study may help other researchers to make reasonable choices among available molecular docking programs in practice. It also provides valuable clues for the developers of molecular docking programs to further improve their methods.

Materials and Methods

Docking Programs Under Evaluation

Four popular commercial docking programs were evaluated in our study. Technical aspects of each program relevant to our study are briefly described below.

Glide (Version 4.5)

Glide was originally developed by Friesner et al.[31,32] and has become a popular option for molecular docking later on.[46,47] It is available now as a module in the Schrödinger software suite, which is released by the Schrödinger LLC. Before a docking job can be started, this program needs a set of grids to be generated with different types of fields representing geometries and properties of the binding site on the given receptor. During docking, an exhaustive sampling in the torsional space of the ligand molecule is performed to generate ligand binding poses. The docking process consists of four major stages. At the first two stages, the program employs a series of hierarchical filters to search for possible locations of the ligand on the grids prepared *in prior* and generate plausible ligand binding poses through a coarse screening. The initial filters examine the steric complementarity of the ligand to the defined binding site, and evaluates ligand-receptor interactions with GlideScore,[31] an expanded version of the ChemScore scoring function.[48,49] At the third stage, the ligand binding poses selected by the initial screening are minimized *in situ* with the OPLS-AA force field.[50] At the last stage, a composite score, namely Emodel, is used to rank the resulting ligand binding poses and select the ones to report by considering GlideScore, nonbonded interaction energies, and internal steric energies of the generated ligand binding poses.

GOLD (Version 3.2)

GOLD was originally developed by Jones et al.[35] It is now commercially released by the Cambridge Crystallographic Data Center. GOLD relies on a genetic algorithm (GA) to explore the conformational space of the ligand. It also allows the consideration of the conformational flexibility of several selected amino acid residues on the protein. Given the three-dimensional structures of protein and ligand, an initial population of ligand binding poses is generated randomly. Each individual of the population is assigned a fitness score based on its predicted binding affinity. Three scoring functions, namely GoldScore,[35–37] ChemScore,[48,49] and ASP,[51] are implemented in GOLD for this purpose. All individuals are ranked according to their fitness scores, and the entire population is iteratively optimized via mutation, crossover, and migration operations. The GA sampling procedure is controlled by a number of parameters, such as "Population Size," "Selection Pressure," "Number of Operations," "Number of Islands," "Niche Size," and "Operator Weights: Migrate, Mutate, Crossover."

LigandFit (Version 2.3)

LigandFit was originally developed by Venkatachalam et al.[39] Now it is implemented in the Discovery Studio software (version 2.0, Accelrys Software) as a molecular docking module. A docking procedure by LigandFit consists of two major steps. The first step is to employ a cavity detection algorithm to identify potential binding sites. The second step is to fit the given ligand to a specified binding site through a Monte Carlo conformational sampling procedure, which is guided by a shape-based filter to generate binding poses matching the binding site. Then, rigid body energy minimization is performed on candidate binding poses using the DockScore energy function.[39] Ligand binding poses can be further evaluated by external scoring functions, including LigScore1,[52] LigScore2,[52] PLP1,[53] PLP2,[54] Jain,[55] PMF,[56] PMF04,[57] and the LUDI scoring function.[58,59]

Surflex (Version 2.0)

Surflex was originally developed by Jain.[42] The molecular docking procedure by Surflex also can be divided into two major steps. At the first step, an ideal ligand fitting to the binding site, called "protomol," is generated.[60] For this purpose, three different types of molecular fragments, i.e., hydrogen bond donor group, hydrogen bond acceptor group, and hydrophobic group, are placed into the binding site on the protein, and their positions/orientations are optimized for forming maximal interactions with the protein. Top-scored fragments are then assembled to form the protomol. At the second step, an incremental algorithm is employed to find the optimal binding poses of the given ligand. The ligand is firstly broken into fragments. The conformations of each fragment are explored, and they are aligned to the corresponding regions on the protomol. The aligned fragments are evaluated by their steric complementarity to the binding site as well as binding scores. Complete molecules are generated either by incremental construction from top-scored fragments or a crossover operation combining intact molecules. The final binding poses are subjected to *in situ* conformational optimization. Their binding scores are computed by the Jain scoring function,[55,61] and the top-scored binding poses will be output.

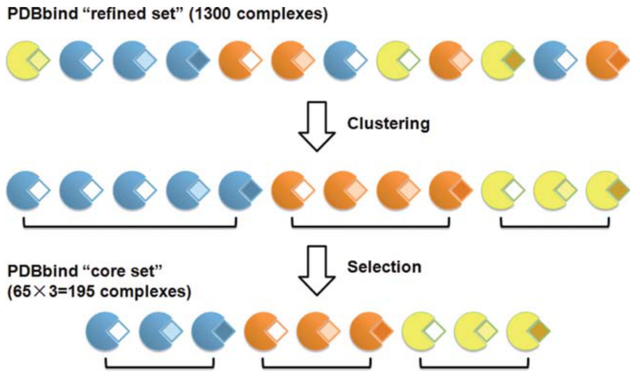

Figure 1. Selection of the test set. (Each icon represents a certain protein-ligand complex. Different types of proteins are indicated in different colors. Darkness of the ligand indicates the level of its binding affinity.)

Preparation of the Test Set

Contents of the test set used in this study are identical to that used in our recent study of scoring functions.[62] For the sake of readers' convenience, compilation of this test set is repeated briefly here. This test set was selected from the PDBbind database[29,30] through systematic mining. The PDBbind database was created to collect the experimentally measured binding data of the protein-ligand complexes deposited in the PDB.[22] It is now actively maintained by our group and is publicly accessible at http://www.pdbbind.org.cn/. The PDBbind version 2007, which consists of binding data of over 3100 protein-ligand complexes, was considered in this study. Not all of them, however, are "healthy" enough for validating molecular docking programs. We therefore applied a number of filters to select the qualified ones among them. Those filters are summarized briefly as following.

1. Concerns on the quality of structures. Only the protein-ligand complexes which structures are determined through crystal diffraction were considered. Each qualified complex structure should have an overall resolution better than or equal to 2.5 Å. In addition, both the structures of the protein and the ligand need to be complete.

2. Concerns on the quality of binding data. Only the protein-ligand complexes with known K_d or K_i were considered. In addition, both the protein and the ligand used in binding assay have to match exactly the ones used in structure determination.

3. Concerns on the components of complexes. Only noncovalently bound protein-ligand complexes were considered. Each qualified complex should be formed by one protein molecule and one ligand molecule in a binary manner. In other words, there should not be multiple ligands bound in close vicinity at a common binding site. The ligand molecule must not contain any uncommon elements, such as Be, B, Si, and metal atoms. Its molecular weight should not exceed 1000. Oligo-peptides (up to 9 residues) and oligo-nucleotides (up to three residues) were also considered as valid small-molecule ligands.

The outcome of the above selection is the so-called "refined set" of the PDBbind database. The refined set in PDBbind version 2007 consists of 1300 protein-ligand complexes. It provides a good starting point for selecting the test set used in this study. Firstly, the refined set was grouped into clusters by sequence similarity computed by BLAST. A similarity cutoff of 90% was applied in clustering. Each resulting cluster typically consisted of complexes formed by one particular type of protein. A total of 65 clusters in the refined set were found to contain at least four protein-ligand complexes. In each cluster, the one with the highest binding affinity, the one with the lowest binding affinity, and the one with a binding affinity closet to the mean value were selected as the representatives of this cluster. These three complexes are termed as the "topper", the "lower", and the "middler" in this manuscript for the sake of convenience. As result, a total of 65 × 3 = 195 protein-ligand complexes were selected, which was termed by us as the "core set" of the PDBbind database. A graphical illustration of the above process is given in Figure 1. A full list of the protein-ligand complexes included in our test set is given in the Supporting Information (Tables S1 and S2).

Coordinates of the complexes in our test set were all downloaded from the PDB. The original structural files from PDB were processed so that they could be readily utilized by the software tested in our study. Basically, each complex was split into a protein molecule, which consisted of a complete "biological unit", and a ligand molecule. Atomic types and bond types of the ligand molecule were automatically assigned by the I-interpret program,[63] and were manually inspected to make corrections when necessary. Hydrogen atoms were added to the protein and the ligand by using the SYBYL software. For the sake of convenience, both the protein and the ligand were set in a simple protonation scheme under neutral pH: all carboxylic acid and phosphonate groups were set in deprotonated forms; while all aliphatic amine and guanidino/amidino groups were set in protonated forms. The protein molecule was assigned the AMBER FF99 charges; while the ligand was assigned the Gasteiger-Hückel partial charges. Then, the processed protein structure was saved in a PDB-format file; while the processed ligand structure was saved in separate Mol2-format and SD-format files. Metal ions, if residing inside the binding pocket and coordinately bound to the ligand and the protein, were saved with the protein molecule. All water molecules included in the crystal structure were removed for the sake of convenience. No structural optimization was conducted on either the protein or the ligand to retain their original coordinates from PDB.

In addition to the binding pose observed in the crystal structure, a low-energy conformation of the ligand in each complex was also prepared. For this purpose, the "CONFORT" module in the SYBYL software was used to generate a low-energy conformation for each ligand. In our study, the key parameters for running CONFORT were set as: "Operation" = "Perform Global Minimization"; "Maximal number of concurrently searched rotors" = 50 (acyclic), 30 (per ring system), and 20 (per ring). The lowest-energy conformation output by CONFORT was adopted in our study. CONFORT failed to generated conformations for the ligands in PDB entries 1FKB and 1FKI as they contained macrocyclic structures. For these two ligand molecules, we manually translated and rotated their binding poses observed in crystal structures, and then performed structural optimization with SYBYL to obtain a low-energy conformation for each of them.

Table 2. Classification of the 195 Protein-Ligand Complexes in the Test Set.

Criterion for classification		Symbol	Number of complexes
Number of rotatable bonds on the ligand	0–2	A1	77
	3–5	A2	60
	≥6	A3	58
Buried percentage of the solvent-accessible surface area of the ligand upon binding	24–48%	B1	30
	49–75%	B2	129
	76–96%	B3	36
A hydrophobic index of the binding pocket on the protein	–18.91 to –8.11	C1	34
	–7.97 to 1.65	C2	131
	1.87–10.15	C3	30

Classification of the Test Set

We further divided the 195 complexes in the test set into subsets by three features relevant to protein-ligand binding (Table 2). Comparison of the outcomes of a certain docking program obtained on the entire test set and those obtained on certain subsets will hopefully provide a deeper understanding of its performance. The first feature was the total number of rotatable bonds on the ligand which was computed using the X-Score program.[64] Here, a rotatable bond is defined as an acyclic sp^3-sp^3 or sp^3-sp^2 single bond between two nonhydrogen atoms. Terminal groups, such as —CH_3, —NH_2, —OH, and —X (X = F, Cl, Br, I), whose rotation does not produce any new conformation of heavy atoms are not counted as rotors. The resulting subsets were named as A1, A2, and A3. They contain complexes with 0–2, 3–5, and ≥6 rotatable bonds on the ligand respectively so that each subset contained a roughly equal number of protein-ligand complexes (Table 2).

The second feature was the buried percentage of the solvent-accessible surface area (SASA) of the ligand upon binding. The Richards-Lee solvent-accessible surface of each ligand molecule was computed using an in-house computer program. The Richards-Lee solvent-accessible surface is essentially the complete trajectory of a probe rolling on the van der Waals surface of a given molecule.[65] In our computation, radius of the probe was set to 1.0 Å, and a dot density of 4 per Å^2 was used in the generation of surface dots. A surface dot on the ligand was considered to be buried upon binding if it was enclosed in the solvent-accessible surface of the protein molecule in the complex The buried percentage was determined after the whole set of surface dots were examined. The atomic radii used in our computation were cited from the classical work of Bondi as: C = 1.70 Å; N = 1.55 Å; O = 1.52 Å; P = 1.80 Å; S = 1.80 Å; F = 1.47 Å; Cl = 1.75 Å; Br = 1.85 Å; I = 1.98 Å; H = 1.00 Å.[66] These parameters, including the probe radius and dot density, were validated in our previous study of empirical solvation models based on SASA.[67]

The third feature was a “hydrophobic index” of the binding pocket on the protein. It was computed using an in-house computer program by summing up the fragmental log*D* value of each amino acid residue in direct contact with the bound ligand and then dividing the sum by the total number of such residues. An amino acid residue on the protein was considered to be in direct contact with the bound ligand if any heavy atom on its side chain was within 4.0 Å from any heavy atom on the ligand. The fragmental log*D* values of each type of amino acid residue were derived from a regression analysis of the experimentally measured log*D* values of more than 200 oligo-peptides by Tao et al.[68] (see the Supporting Information Table S4). Conceptually, a more positive fragmental log*D* indicates a more hydrophobic residue; while a more negative fragmental log*D* indicates a more hydrophilic residue. LogD values were employed instead of log*P* values to reflect the protonation states of amino acid residues under the physiological pH condition.

As for the latter two properties, we used the *Z*-Score of each feature instead of the absolute values in subset classification. For each given property, *Z*-Score of the *i*th protein-ligand complex in the test set was computed as:

$$Z\text{-Score}_i = \frac{f_i - \mu}{\sigma}$$

Here, f_i is the value of a certain property of this complex; while μ and σ are the mean value and the standard deviation (SD) of this property on the entire test set, respectively. The entire test set was then divided into three subsets containing complexes with *Z*-Scores falling in the range of $(-\infty, -1)$, $[-1, +1]$, and $(+1, +\infty)$, respectively. Conceptually, these three subsets consist of samples which are considerably lower than the average, around the average, and considerably higher than the average in terms of a given property. The subsets defined by using the buried percentage of the SASA of the ligand upon binding were named as B1, B2, and B3; while those defined by the hydrophobic index of the binding pocket were named as C1, C2, and C3 (Table 2).

Evaluation Methods

We evaluated all four docking programs in terms of their abilities of reproducing the native ligand binding poses in three aspects: (i) performing docking at three different levels of computational costs; (ii) using the native binding pose or a low-energy conformation of the given ligand as the input for subsequent docking; (iii) performing *in situ* postoptimization on docking solutions or not. The detailed settings for each program are described below.

Glide (Version 4.5)

For each given complex, receptor grids were generated using the binding sites defined by the native ligand. Limitations on atom number and rotatable bonds on the ligand were set to 200 and 35, respectively. All three computation levels provided by Glide, i.e., level 1 = high throughput virtual screening, level 2 = standard precision, and level 3 = extra precision (XP), were tested in our study. The maximal number of ligand binding poses to output in each docking job was set to 100. Whether to perform postoptimization on these binding poses was controlled by the “Perform postdocking minimization” option. Note that for the XP mode, postoptimization cannot be disabled.

GOLD (Version 3.2)

For each given complex, the binding site was defined by the native binding pose of the given ligand. According to the recommendations in the user manual of GOLD, the automatic ligand dependent settings were used in our study. The three computation levels were set by adjusting the "search efficiency" parameter to 50% for level 1, 100% (the default value) for level 2, and 200% (the allowed maximal value) for level 3. At each level, GOLD will automatically determine the necessary computational cost on each ligand by adjusting the parameters controlling the GA procedure (population size, selection pressure, niche size, and etc.). For example, when the "search efficiency" parameter is set to 100%, GOLD will attempt approximately 30,000 GA operations for a ligand with five rotatable bonds. The maximal number of ligand binding poses to output in each docking job was set to 100. Whether to perform postoptimization on these binding poses was controlled by the '*DO_SIMPLEX*' parameter.

LigandFit (Version 2.3)

For each given complex, the binding site for docking was defined by the native binding pose of the ligand. The three computation levels of LigandFit were set by adjusting the parameters controlling the Monte Carlo sampling process, including the maximal number of trials to perform (per number of torsions), and the number of consecutive failed trials to attempt before terminating the docking job for a given ligand. These two parameters were set as (1000, 250) for level 1, (3000, 750) for level 2, and (10,000, 2500) for level 3. The maximal allowed number of torsions on ligand was set to 100. The maximal number of ligand binding poses to output in each docking job was set to 100. Postoptimization on these binding poses could be performed by enabling the "Smart Minimizer" using the CHARMm force field.[69]

Surflex (Version 2.0)

For each given complex, the "protomol" was generated using the binding site defined by the native binding pose of the ligand. The three computation levels of Surflex were set by setting the "Additional Starting Conformations per Molecule" parameter to zero (the default value) for level 1, 5 for level 2, and 10 (a fairly large value mentioned in the user manual) for level 3. The maximal number of ligand binding poses to output was set to 100. Whether to perform postoptimization on these binding poses was controlled by the "PostDock Minimization" parameter.

The docking accuracy of each program was evaluated by how well it reproduced the native binding poses of the ligands observed in crystal structures, which was quantified by the root-mean-square deviation (RMSD) between the top-ranked docking solution and the native ligand binding pose, namely $\mathrm{RMSD_{top}}$, in each case. In molecular docking, few people would consider docking solutions with large deviations from the true binding pose. Therefore, we defined that when $\mathrm{RMSD_{top}} > 3$ Å, the docking solution for the given complex was unacceptable. The success rate of each docking program on the entire test set was thus counted as the percentage of the cases when $\mathrm{RMSD_{top}} \leq 3$ Å. To compare the overall docking accuracy of these programs in a statistical manner, the difference between the $\mathrm{RMSD_{top}}$ values produced by any two docking programs (*A* and *B*) on a given complex was computed as: $\Delta\mathrm{RMSD_{top}} = \mathrm{RMSD_{top,A}} - \mathrm{RMSD_{top,B}}$. The distribution of $\Delta\mathrm{RMSD_{top}}$ values on the entire test set was then analyzed with standard one-sample *T*-test to examine if the absolute value of $\Delta\mathrm{RMSD_{top}}$ was significantly larger than zero. If it is, a conclusion can be made on whether the performance of program *A* is better or worse than that of program *B*, depending on the sign of the mean value of $\Delta\mathrm{RMSD_{top}}$.

The sampling completeness of each docking program in each case was evaluated by the minimal RMSD among all docking solutions to the native binding pose, namely $\mathrm{RMSD_{min}}$. The overall performance of each docking program in this aspect was measured as the average $\mathrm{RMSD_{min}}$ value on the entire test set. All RMSD values were computed using the "Smart_RMS" tool included in the GOLD software. Only heavy atoms, i.e., nonhydrogen atoms, were considered in computation. Distributions of the resulting $\mathrm{RMSD_{min}}$ values on the entire test set were analyzed with two-sample *T*-test to examine if any significant difference exists between the results of any two docking programs.

As a supplementary evaluation method, we also considered the diffraction-component precision index (DPI). This concept was originally introduced by Cruickshank[70] as a precision index of the atomic coordinates obtained from crystal diffraction. DPI is computed with *R* or R_{free} factors through an approximation to the least-squares method. It has been used by some researchers as the theoretical limit for molecular docking. In our study, the same judgment by Goto et al.[71] was used: for each complex, if the RMSD value between a docking solution and the native binding pose is smaller than $2\sqrt{2}$ times of the DPI of the given complex structure, then the docking solution is believed to be comparable to the accuracy of experimental measurement. We computed the DPI values using the "pdb2dpi" program implemented in the autoBUSTER software.[72] Among the 166 protein-ligand complexes in our test set for which all the four docking programs were able to produce docking solutions, DPI values of 135 complexes were finally obtained (see the Supporting Information Table S3). The $\mathrm{RMSD_{top}}$ and $\mathrm{RMSD_{min}}$ values produced by each docking program on these complexes were then compared to the corresponding DPI to judge their quality.

In addition, the correlation between the binding scores of the top-ranked docking solutions produced by each docking program and the experimental binding constants of the protein-ligand complexes in the test set was computed. The Pearson correlation coefficient (*R*) and the SD between the known binding constants and the fitted values in regression were computed as:

$$R = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}}$$

$$\mathrm{SD} = \sqrt{\frac{\sum [y_i - (a + b \cdot x_i)]^2}{N - 1}}$$

Here, *N* is the total number of samples; x_i is the binding score computed by a certain docking program based on the top-ranked binding pose of the *i*th complex; y_i is the experimental binding constant of this complex; *a* is the intercept and *b* is the slope of the linear regression line between the binding scores and experimental binding constants. Binding constants were all given as the negative logarithm of K_d (or K_i) values.

All of the molecular docking jobs were completed on a Dell Precision 490 workstation (dual Intel Xeon 5140 processor with 4 GB memory) with a RedHat Linux operation system. The CPU time spent by each docking program for each given protein-ligand complex at each considered computation level was recorded. Every docking job was run on a "clean" CPU to ensure that the recorded CPU time was accurate. Note that Glide and Surflex require some necessary preparative steps, e.g., generating receptor grids and protomols, before a docking job can be submitted. The recorded CPU time for these two programs thus includes the time spent on these preparative steps.

Results and Discussion

On the Test Set

As the primary aim of our study is to evaluate the general performance of four popular docking programs, it is more appropriate to use a high-quality test set consisting of diverse protein-ligand complexes. The test sets employed in many previous studies,[8,9,11,12,14] are basically some random assemblies of protein-ligand complexes. In contrast, the test set used in our study was compiled through a systematic mining of the PDBbind database.[29,30] As the PDBbind database itself is based on the entire PDB, our test set can be considered as a miniature representative of the protein-ligand complexes in PDB.

The test set used in our study was compiled with special considerations on diversity. As described in the Materials and Methods section, a total of 65 families of protein-ligand complexes were selected from the entire PDBbind database through clustering by protein sequences. In other words, diversity on the protein side in our test set was defined by protein sequences. Our test set provides a broader coverage of different types of proteins than most test sets employed in previous studies (Table 1). Diversity on the ligand side in our test set was defined primarily by binding affinity. For this purpose, three complexes in each complex family, i.e., "topper," "middler," and "lower" defined by binding affinities, were selected to cover a maximal range of binding affinities. Distribution of the binding constants of all 195 protein-ligand complexes is shown in Figure 2. Binding constants of these complexes range from 1.40 to 13.96 (in logarithm units), spanning over 12 orders of magnitude. The mean values of the binding constants of the "toppers," the "middlers," and the "lowers" are 4.40 ± 1.60, 6.39 ± 1.55, and 8.34 ± 2.09, respectively. Note that diversity in binding affinity is in fact a direct consequence of the structural diversity of ligands. As additional proofs of the structural diversity of the 195 ligands in the test set, their molecular weights range widely from 103 to 974, and the numbers of the rotatable bonds on them range from 0 to 32 (see Figure 3).

Another important consideration is the redundancy of the test set. It has been demonstrated repeatedly that the performance of a molecular docking program is case-dependent, which seems to be decided primarily by the characteristics of the target proteins under study. Thus, it is essential to have a redundancy control on the contents of the test set to make fair evaluations of molecular docking programs. This feature was basically missing in most test sets employed in the previous comparative studies of docking programs. For example, the test set used by Chen et al.[13] was biased towards trypsin and thrombin; while the test set used by Kontoyianni et al.[11] was biased towards dihydrofolate reductase, exo-α-sialidase, and stromelysin. In contrast, our test set has a strict control on redundancy: each of the 65 types of proteins included in our test set has a uniform redundancy of three so that the results of our evaluation are in principle not biased towards any particular types of proteins.

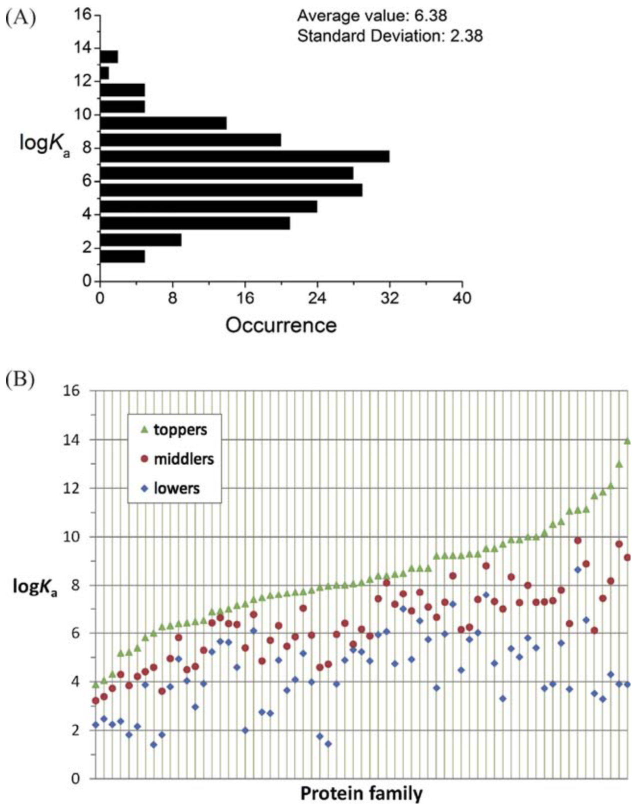

Figure 2. (A) Overall distribution of the experimentally measured binding constants of the 195 protein-ligand complexes in the test set. (B) Distribution of the binding constants in each protein family.

In addition, the size of our test set is larger than those employed in the previous studies of molecular docking programs of the same type. The size of the test set is of course important for deriving statistically significant results. More importantly, the protein-ligand complexes included in our test set all have relatively reliable structures and binding data, which should be attributed to our efforts on the development of the PDBbind database. Every complex structure in our test set has an overall resolution better than or equal to 2.5 Å; while in some previous studies, 3.0 Å was employed as the cutoff for the selection of complex structures.[9,10,13] Every complex has an experimentally measured K_d or K_i value rather than an IC_{50} value. Strictly speaking, IC_{50} values can be used only to compare the binding

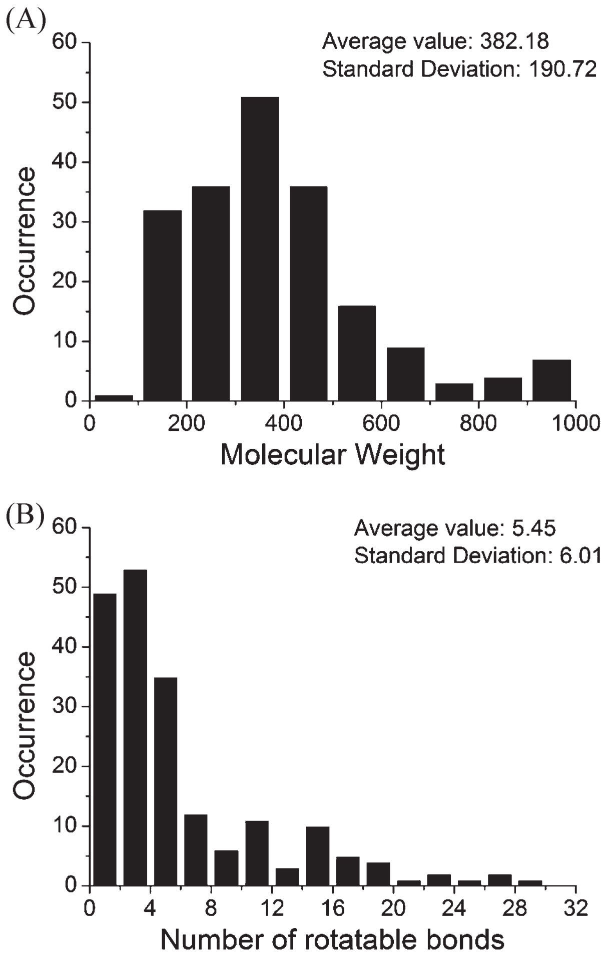

Figure 3. Distribution of (A) the molecular weights and (B) the number of rotatable bonds of the 195 ligand molecules in the test set.

affinities measured in the same assay and thus is not appropriate for a diverse data set like ours. Furthermore, the structures and the binding data of all protein-ligand complexes in our test set were manually examined more than once to ensure that (i) neither of the protein nor the ligand has missing fragments on their structures, and (ii) the complex seen in the crystal structure matches the one used in binding data measurement. All of these quality-control efforts have endowed our test set obvious technical advantages over those employed in previous similar studies.

In summary, we have compiled a high-quality diverse test set for the evaluation of molecular docking programs. It provides a solid basis for this study and could become an effective benchmark for other applicable studies. For example, we employed this data set in a comparative assessment of scoring functions recently.[62] As far as we know, the only data set compiled for the same purpose with a comparable quality to ours is the "Astex diverse set",[73] which was compiled at approximately the same time as ours. It consists of a total of 85 high-resolution protein-ligand complexes, which were also selected through clustering the protein-ligand complexes in the entire PDB with emphasis on diversity. A special concern of the Astex diverse set was that for each qualified complex, the protein had to be of pharmaceutical or agrochemical interests; while the ligand had to be marketed drugs, drug candidates on clinical trials, or at least "drug-like." This criterion was not adopted in the compilation of our test set. Whether a candidate compound can enter clinical trials or reach market is in fact determined by many other factors other than the binding to its primary biological target. Therefore, our opinion is that this concern is rather irrelevant to the study of protein-ligand binding events. Another notable difference between our test set and the Astex diverse set is that each type of protein is associated with three complexes in our test set instead of a single one as in the Astex diverse set. This feature allows the study of the binding of different ligands to the same protein in more details.

Performance on the Entire Test Set

In this section, the performance of the four docking programs under our study on the entire test set is described and discussed. Among these four docking programs, GOLD and Surflex had no problem in processing all of the 195 complexes in our test set; whereas Glide and LigandFit failed to process 25 and eight complexes, respectively, because of various reasons (see the Supporting Information Table S5). Consequently, all of the statistical data associated with Glide and LigandFit were computed based on the complexes which they were able to process. It needs to be emphasized that robustness is also an important feature for a successful docking program, although it will not be discussed further in the rest parts of this article.

Distributions of the RMSD values between the top-ranked docking solutions and the native ligand binding poses ($\mathrm{RMSD}_{\mathrm{top}}$) produced by each docking program at the highest computation level (level 3) plus postoptimization on the final docking solutions are given in Figure 4. One can see that when

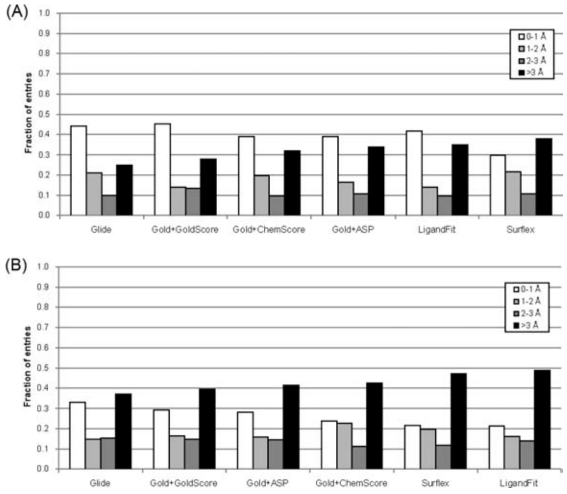

Figure 4. Distribution of the $\mathrm{RMSD}_{\mathrm{top}}$ values produced by four docking programs at the highest computation level (level 3) plus *in situ* optimization when (A) the native binding poses and (B) low-energy conformations were used as inputs, respectively. All docking programs are ranked by their success rates ($\mathrm{RMSD}_{\mathrm{top}} \leq 3$ Å) in a decreasing order.

Table 3. Mean Values of the $\Delta\mathrm{RMSD}_{top}$ Values (Å) Between Each Pair of Docking Programs.[a]

	Program *B*				
Program *A*	GOLD + GoldScore	GOLD + ChemScore	GOLD + ASP	LigandFit	Surflex
When native binding poses were used as docking inputs					
GLIDE	–0.10	–0.33	–0.42 (significant)[b]	–0.50 (significant)	–0.60 (significant)
GOLD + GoldScore	0	–0.23	–0.32	–0.40 (significant)	–0.50 (significant)
GOLD + ChemScore	–	0	–0.10	–0.18	–0.27
GOLD + ASP	–	–	0	–0.08	–0.18
LigandFit	–	–	–	0	–0.10
When random low-energy conformations were used as docking inputs					
GLIDE	0.02	–0.27	–0.22	–0.63 (significant)	–0.63 (significant)
GOLD + GoldScore	0	–0.28	–0.23	–0.64 (significant)	–0.64 (significant)
GOLD + ChemScore	–	0	0.05	–0.36 (significant)	–0.36 (significant)
GOLD + ASP	–	–	0	–0.41 (significant)	–0.41 (significant)
LigandFit	–	–	–	0	0.00

[a] RMSD_{top} (Å) values produced by each docking program considered in this statistical analysis were all obtained at the highest computation level plus *in situ* optimization on the final docking poses.

[b] One-sample *T*-test indicates that the mean value of $|\Delta\mathrm{RMSD}_{top}|$ is significantly larger than zero at the 95% confidence level. The complete results of this analysis are given in the Supporting Information (Table S7).

the native binding poses were used as docking inputs, these docking programs can be ranked by their success rates as: Glide > GOLD + GoldScore > GOLD + ChemScore ≈ GOLD + ASP ≈ LigandFit > Surflex. Notably, all docking programs except Surflex produced very good docking solutions (RMSD_{top} < 1.0 Å) in about 40% cases. Nevertheless, what is more likely in practice is that some low-energy conformations of the given ligand molecules are used as inputs for molecular docking. In such a scenario, the overall success rates of all docking programs under our test decrease by about 10%. It is encouraging to notice that some docking programs, such as Glide and GOLD, still maintain a success rate around 60% on the highly diverse test set used in this study. The docking programs under our test set now can be ranked as: Glide > GOLD + GoldScore ≈ GOLD + ChemScore ≈ GOLD + ASP > LigandFit ≈ Surflex. One can see the ranks of these docking programs do not alter much regardless the change in docking inputs. Thus, it is reasonable to conclude that the general performance of Glide and GOLD is better than LigandFit and Surflex in this test. The above qualitative judgment is supported by statistical analysis on the $\Delta\mathrm{RMSD}_{top}$ values between each pair of docking programs (Table 3). Our results indicate that when the native binding poses are used as docking inputs, the advantage of Glide and GOLD + GoldScore over LigandFit and Surflex is statistically significant at the 95% confidence level; when low-energy conformations are used as docking inputs, a more meaningful scenario in reality, the advantage of Glide and all three options of GOLD over LigandFit and Surflex is statistically significant.

The observed decrease in success rates when low-energy conformations are used as docking inputs indicates that choosing the right starting structures for docking is still important for obtaining correct final solutions even when all docking programs employ sophisticated sampling methods. Here, LigandFit seems to be particularly sensitive to starting structures as it is associated with the largest decrease in success rate after the docking inputs were switched to low-energy conformations. Obviously, the conformational sampling method implemented in LigandFit should account for this result. Among the previous studies summarized in Table 1, Kellenberger et al.[9] and Onodera et al.[15] also compared the outcomes produced by certain docking programs when native ligand binding poses versus low-energy conformations were used as docking inputs. They reported that Glide and GOLD were relatively robust in this aspect, which is consistent with our findings.

In our study, we also evaluated the completeness of conformational sampling by examining the minimal RMSD to the native binding pose among all docking solutions (RMSD_{min}) produced by each docking program (see Figure 5). Our results indicate that all four docking programs were able to generate binding poses close to the native ones. The average RMSD_{min} is generally lower than 2 Å and sometimes as low as 1 Å. The performance of all docking programs in this aspect was comparable no matter whether the native binding poses or the low-energy conformations were used as docking inputs, Indeed, two-sample independent *T*-tests on the distributions of RMSD_{min} values produced by all docking programs reveal that these distributions are not significantly different at the 95% confidence level in most cases (see the Supporting Information Table S9). We thus conclude that the docking programs under our test are more capable in terms of sampling completeness than docking accuracy. In particular, Surflex produced the smallest average RMSD_{min} value (1.3–1.5 Å) whereas the unacceptable rate (RMSD_{top} > 3.0 Å) produced by Surflex is the highest among all docking programs (see Figure 4). Therefore, the relatively poor performance of Surflex in identifying the correct binding poses can be attributed to the inaccurate ranking of docking solutions by its internal scoring function.

As described in the Materials and Methods section, the DPI was also considered in our study as an estimated theoretical

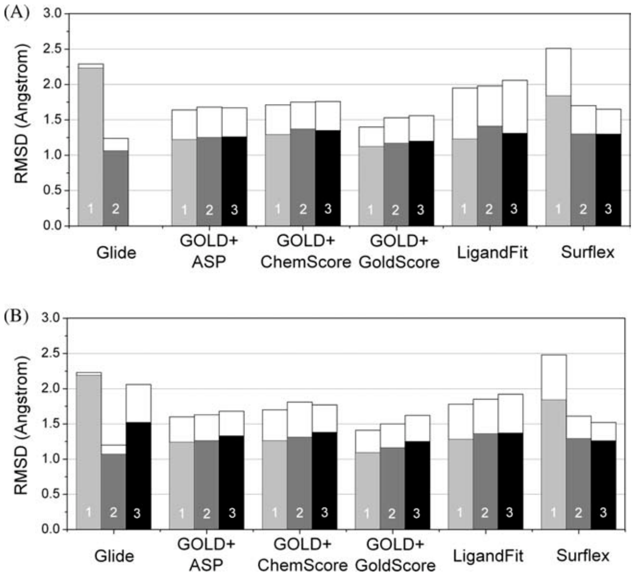

Figure 5. The average RMSD$_{min}$ values produced by four docking programs (A) without *in situ* optimization, and (B) with *in situ* optimization on the final docking solutions. The total height and the filled part of each column indicate the RMSD$_{top}$ values obtained when the random and the native binding poses were used as docking inputs, respectively. Numbers on the columns indicate different computation levels. Raw data for making this figure are given in the Supporting Information (Table S8).

limit for the accuracy of molecular docking. The RMSD$_{top}$ and RMSD$_{min}$ values produced by each docking program are compared to the corresponding DPI values (Table 4). One can see that all docking programs were able to produce $RMSD_{min} < 2\sqrt{2} \times$ DPI on a good number of protein-ligand complexes (30–50%) when the native binding poses were used as docking inputs. This rate, however, decreases considerably to 18–33% once the docking inputs were switched to low-energy conformations. In either case, the performance of all these docking programs is roughly at the same level. Considering RMSD$_{top}$ values, the performance of these docking programs is apparently less promising, indicating that identifying rather than generating the correct binding poses is still a challenging problem. The performance of these docking programs is more distinct in this aspect. All these observations are in qualitative agreement with what we have discussed above based on the distributions of RMSD values.

In structure-based drug design, docking programs are applied not only to predict the binding poses of some given ligands, but also to rank them by their binding scores. It is desirable that the given ligands are ranked in the correct order by their real binding affinities, a feature critical for the success of docking-based virtual screening. The correlation between the experimental binding constants and the binding scores of top-ranked docking solutions by all docking programs were also examined in our study (Table 5). One can see that none of the four docking programs really produced satisfactory results. Among them, Glide, GOLD + ASP, GOLD + ChemScore, and Surflex were able to produce Pearson correlation coefficients (R) higher than 0.40, a modest level of correlation, at their highest computation levels. In contrast, a stunning observation is that LigandFit basically failed to produce any correlation between its binding scores and experimental binding constants in all scenarios. The reason accounting for the disappointing performance of all docking programs on this test is two-folded: firstly, as revealed by the tests discussed above, these docking programs do not always produce the correct binding poses. Secondly, they may not produce satisfactory binding scores even given the correct binding poses, which has been demonstrated in some studies on scoring functions.[62,74–77] On the other hand, it has to be mentioned that the test set employed in our study is especially challenging as it consists of a remarkable diversity of protein-ligand complexes. Thus, it is rather encouraging to see that some docking program produced modest correlations on such a test set. In structure-based drug design practice, one works on a particular target protein in most cases. Modeling different ligands bound to the same protein is assumed to be less challenging, which have been proved by many successful applications reported in literature.

Docking Accuracy as a Function of Computational Cost

Molecular docking is basically a conformational sampling process. A more complete sampling is thus associated with a higher

Table 4. Performance of All Docking Programs on the 135 Protein-Ligand Complexes with Available DPI Values.[a]

	Number of complexes with RMSD$_{top}$ $<2\sqrt{2} \times$ DPI		Number of complexes with RMSD$_{min}$ $<2\sqrt{2} \times$ DPI	
Docking program	Native binding poses as docking inputs	Low-energy conformations as docking inputs	Native binding poses as docking inputs	Low-energy conformations as docking inputs
GLIDE	28	18	47	24
GOLD + GoldScore	28	17	53	34
GOLD + ChemScore	26	14	58	34
GOLD + ASP	25	14	68	45
LigandFit	27	9	43	30
Surflex	15	9	52	34

[a] All RMSD values were obtained at the highest computation level (level 3) plus *in situ* optimization of docking solutions.

Table 5. The Correlation between the Binding Scores of Top-Ranked Docking Solutions and the Experimental Binding Data.[a]

			Native binding poses as docking inputs		Low-energy conformations as docking inputs	
Docking program	Computation level	*N*	Without *in situ* optimization	With *in situ* optimization	Without *in situ* optimization	With *in situ* optimization
GLIDE	Level 1	170	0.176 (2.20)	0.181 (2.20)	0.191 (2.20)	0.205 (2.19)
	Level 2		0.247 (2.17)	0.247 (2.17)	0.234 (2.18)	0.261 (2.16)
	Level 3		N/A	0.505 (1.93)	N/A	0.478 (1.97)
GOLD + ASP	Level 1	195	0.479 (2.09)	0.474 (2.10)	0.472 (2.10)	0.475 (2.10)
	Level 2		0.476 (2.10)	0.488 (2.08)	0.481 (2.09)	0.483 (2.09)
	Level 3		0.478 (2.10)	0.490 (2.08)	0.467 (2.11)	0.476 (2.10)
GOLD + ChemScore	Level 1	195	0.431 (2.15)	0.434 (2.15)	0.399 (2.19)	0.416 (2.17)
	Level 2		0.443 (2.14)	0.472 (2.10)	0.422 (2.16)	0.418 (2.17)
	Level 3		0.445 (2.14)	0.461 (2.12)	0.416 (2.17)	0.432 (2.15)
GOLD + GoldScore	Level 1	195	0.283 (2.29)	0.299 (2.28)	0.306 (2.27)	0.317 (2.26)
	Level 2		0.298 (2.28)	0.309 (2.27)	0.313 (2.27)	0.322 (2.26)
	Level 3		0.301 (2.28)	0.309 (2.27)	0.306 (2.27)	0.320 (2.26)
LigandFit	Level 1	187	0.048 (2.38)	0.048 (2.38)	0.012 (2.38)	0.012 (2.38)
	Level 2		0.056 (2.37)	0.033 (2.38)	0.002 (2.38)	0.002 (2.38)
	Level 3		0.033 (2.38)	0.033 (2.38)	0.013 (2.38)	0.013 (2.38)
Surflex	Level 1	195	0.292 (2.28)	0.376 (2.21)	0.272 (2.30)	0.350 (2.24)
	Level 2		0.358 (2.23)	0.441 (2.14)	0.360 (2.23)	0.401 (2.19)
	Level 3		0.380 (2.21)	0.448 (2.13)	0.371 (2.22)	0.404 (2.18)

[a] The data outside brackets are the Pearson correlation coefficients; the data inside the brackets are standard deviations in correlation (in logK_a units). For readers' convenience, the cells are shaded when the correlation coefficients are above 0.400.

level of computational cost. Thus, one would expect that better docking accuracy can be achieved at a higher level of computation. Nevertheless, most previous studies evaluated docking programs with a single "optimal" setting[8,9,11,13,14] or a special setting so that all programs under test would spend approximately the same amount of time.[10,12] As described in the Materials and Methods section, we tested each docking program at three different computation levels. This practice allows us to examine the performance of each docking program as a function of computational cost, which is a notable feature of our study.

The success rates of all four docking programs at three computation levels are illustrated in Figure 6. Surprisingly, the correlation between success rates and computational costs is not obvious no matter whether the native binding poses or the low-energy conformations are used as docking inputs. A trivial correlation can be seen in the outcomes of Glide and Surflex, but there is basically no such correlation in the outcomes of GOLD and LigandFit. The same trend can also be observed in the correlation between the sampling completeness (RMSD$_{\mathrm{min}}$) of all four docking programs and their computational costs (see Figure 5). Our results indicate that for the docking programs under our test, it is somewhat naïve to expect an improved docking accuracy by increasing computational costs significantly.

After removing the entries for which Glide or LigandFit failed to produce docking solutions, the remaining 166 entries were used to investigate the average speeds of all four docking programs at three computation levels (see Figure 7). Our results show that Surflex is generally faster than the other three. In particular, Surflex spent <30 s at average to process one protein-ligand complex at the basic computation level (level 1) when postoptimization was disabled. For some unknown reasons, GOLD + GoldScore is the most time-consuming option. Examining the CPU time at three different computation levels, one can see that the CPU time consumed by GOLD and LigandFit are basically proportional to users' choices because they rely on exhaustive GA or Monte Carlo processes for conformational sampling. In contrast, the CPU time consumed by Surflex and Glide do not increase much at higher computation levels. It is probably because these two docking programs adopt hierarchical conformational sampling procedures and some built-in rules allowing early terminations. In practice, one should make an appropriate comprise between accuracy and computational costs by considering the results presented in Figures 5–7.

Postoptimization of docking solutions within the geometrical constraints of the binding site, i.e., *in situ* structural optimization, is also a common practice in molecular docking, which results in additional computational costs. Note that postoptimization can be conducted in two ways: one is to employ the protocol implemented in the given docking program itself like our study; the other is to employ an external program like what was in Perola's study.[10] The first approach is more likely to be chosen in reality for its convenience, although the potential advantage of the second approach is that the outcomes by all docking programs under test is processed by a uniformed method. Our results show that after postoptimization was enabled, running time of Surflex, GOLD, and Glide did not change much whereas

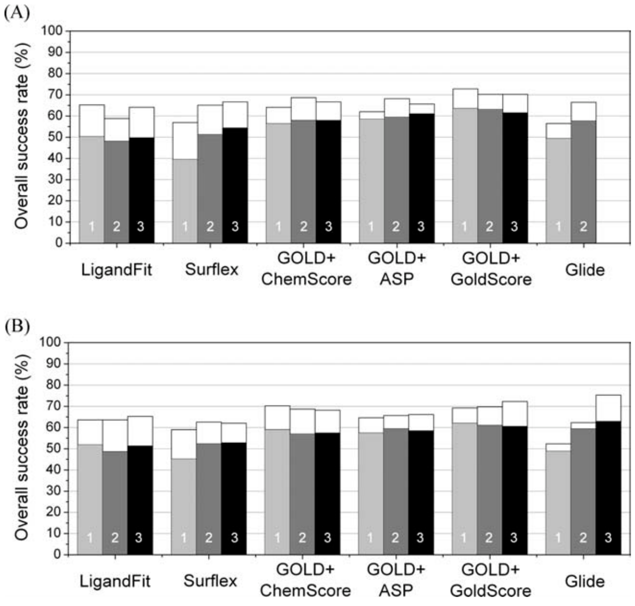

Figure 6. The overall success rates ($\mathrm{RMSD}_{\mathrm{top}} \leq 3$ Å) of four docking programs (A) without *in situ* optimization, and (B) with *in situ* optimization on the final docking solutions. The total height and the filled part of each column indicate the success rate obtained when the native and low-energy conformations were used as docking inputs, respectively. Numbers on the columns indicate different computation levels. All docking programs are ranked by their success rates at the highest computation level (level 3) in an increasing order. Raw data for making this figure are given in the Supporting Information (Table S6).

the one of LigandFit increased significantly (see Figure 7). This observation prompts that the *in situ* optimization algorithm in LigandFit perhaps needs to be more efficient. Comparing the results produced by all four docking programs when postoptimization was disabled or enabled, one can see with surprise that these two sets of results do not exhibit significant difference in terms of either sampling completeness (see Figure 5) or docking accuracy (see Figure 6). This observation suggests that all these docking programs are able to produce reasonable conformations in respect to internal strains or intermolecular contacts with proteins so that additional *in situ* optimizations do not make much difference. Therefore, postoptimization can be safely skipped in practice especially if it is time-consuming, such as in the case of LigandFit.

Docking Accuracy on Subsets of Protein-Ligand Complexes

All results discussed above were obtained on the entire test set, which reflect the general performance of the molecular docking programs under our evaluation. As described in the Materials and Methods section, we also classified the entire test set into subsets using three properties relevant to protein-ligand binding. Each resulting subset thus consists of some protein-ligand complexes with distinctive features. Comparison of the results obtained on these subsets may provide more detailed information regarding the performance of each docking program.

In our first set of tests, the total number of rotatable bonds on the ligand, a rough indication of its conformational flexibility, is used as the criterion for defining subsets. Considering the success rates of each docking program at the highest computation level (Table 6), one can see that the performance of all four docking programs is apparently better on subsets A1 and A2 (rotatable bonds = 0–5) than on subset A3 (rotatable bonds >5). Even when the low-energy conformations are used as docking inputs, all docking programs are able to achieve success rates over 60% on subsets A1 and A2. While for subset A3, the success rates of all docking programs are below 50%. Thus, really flexible ligands remain challenging for today's docking programs. Among the four docking programs, the performance of Glide and GOLD on subset A3 was marginally better than the other two. This observation is basically in agreement with the findings of Perola (Glide > GOLD ≈ ICM),[10] Kontoyianni (GOLD > Glide > FlexX > LigandFit),[11] and Chen (Glide ≈ ICM > GOLD > FlexX).[13] It needs to be pointed out that Glide failed in 15 cases among the total 58 complexes in subset A3. It seems that the multiple-stage docking algorithm implemented in the current version of Glide has certain problems to handle really flexible ligands.

When the buried percentage of the SASA of the ligand is used as the criterion for defining subsets, one can see in Table 7 that the performance of all four docking programs clearly become better with the increase in this percentage (B3 > B2 > B1). At the highest computation level, all four programs

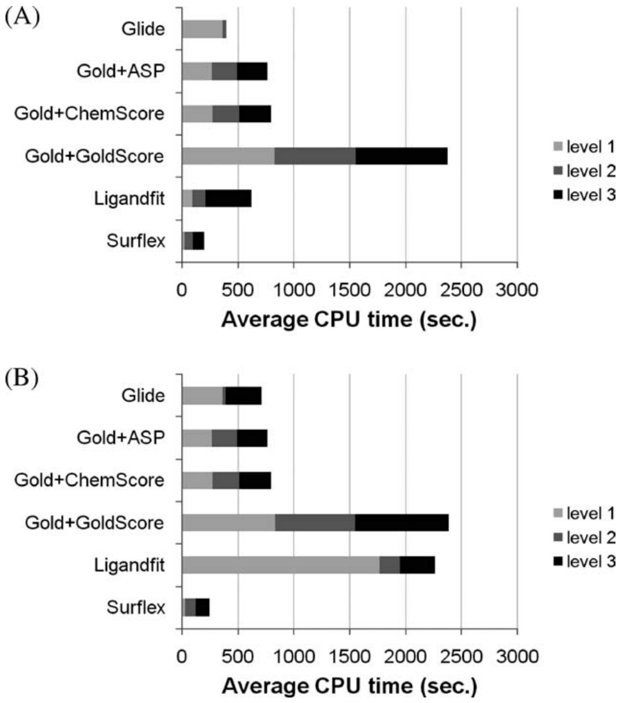

Figure 7. The average CPU time consumed on processing one protein-ligand complex by four docking programs when random binding poses were used as docking inputs. (A) No *in situ* optimization on final docking solutions; (B) *in situ* optimization on final docking solutions enabled.

Table 6. Docking Performance of All Four Programs on the Subsets Defined by the Number of Rotatable Bonds on the Ligands.[a]

			Success rate ($RMSD_{top} \leq 3$ Å)	
Docking program	Subset	*N*	Native binding poses as docking inputs	Low-energy conformations as docking inputs
GLIDE	A1	74	57 (77%)	52 (70%)
	A2	53	45 (85%)	34 (64%)
	A3	43	26 (60%)	21 (49%)
GOLD + ASP	A1	77	54 (70%)	53 (69%)
	A2	60	44 (73%)	36 (60%)
	A3	58	31 (53%)	25 (43%)
GOLD + ChemScore	A1	77	59 (77%)	51 (66%)
	A2	60	43 (72%)	38 (63%)
	A3	58	31 (53%)	23 (40%)
GOLD + GoldScore	A1	77	59 (77%)	54 (70%)
	A2	60	45 (75%)	38 (63%)
	A3	58	37 (64%)	26 (45%)
LigandFit	A1	76	48 (63%)	44 (58%)
	A2	60	36 (60%)	34 (57%)
	A3	51	38 (75%)	18 (35%)
Surflex	A1	77	52 (68%)	47 (61%)
	A2	60	42 (70%)	39 (65%)
	A3	58	27 (47%)	17 (29%)

[a] Data outside brackets are the total number of complexes with $RMSD_{top} \leq 3$ Å, and data inside the brackets are the corresponding percentages. All data were obtained at the highest computation level (level 3) plus *in situ* optimization. The complete results in this test are given in the Supporting Information (Table S10).

Table 7. Docking Performance of All Four Programs on the Subsets Defined by the Buried Percentage of the Solvent-Accessible Surface Areas of the Ligands.[a]

			Success rate ($RMSD_{top} \leq 3$ Å)	
Docking program	Subset	*N*	Native binding poses as docking inputs	Low-energy conformations as docking inputs
GLIDE	B1	25	11 (44%)	7 (28%)
	B2	115	91 (79%)	76 (66%)
	B3	30	26 (87%)	24 (80%)
GOLD + ASP	B1	30	10 (33%)	9 (30%)
	B2	129	85 (66%)	72 (56%)
	B3	36	34 (94%)	33 (92%)
GOLD + ChemScore	B1	30	11 (37%)	7 (23%)
	B2	129	89 (69%)	75 (58%)
	B3	36	33 (92%)	30 (83%)
GOLD + GoldScore	B1	30	8 (27%)	7 (23%)
	B2	129	99 (77%)	77 (60%)
	B3	36	34 (94%)	34 (94%)
LigandFit	B1	29	11 (38%)	7 (24%)
	B2	122	80 (66%)	58 (48%)
	B3	36	31 (86%)	31 (86%)
Surflex	B1	30	4 (13%)	2 (7%)
	B2	129	86 (67%)	71 (55%)
	B3	36	31 (86%)	30 (83%)

[a] Data outside brackets are the number of complexes with $RMSD_{top} \leq 3$ Å, and data inside the brackets are the corresponding percentages. All data were obtained at the highest computation level (level 3) plus *in situ* optimization. The complete results of this test are given in the Supporting Information (Table S11).

Table 8. Docking Performance of All Four Programs on the Subsets Defined by a Hydrophobic Index of the Binding Pocket On Protein.[a]

			Success rate (RMSD$_{top}$ ≤ 3 Å)	
Docking program	Subset	*N*	Native binding poses as docking inputs	Low-energy conformations as docking inputs
GLIDE	C1	27	23 (85%)	19 (70%)
	C2	118	84 (71%)	69 (58%)
	C3	25	21 (84%)	19 (76%)
GOLD + ASP	C1	34	21 (62%)	20 (59%)
	C2	131	85 (65%)	74 (56%)
	C3	30	24 (80%)	22 (73%)
GOLD + ChemScore	C1	34	21 (62%)	19 (56%)
	C2	131	88 (67%)	68 (52%)
	C3	30	24 (80%)	24 (80%)
GOLD + GoldScore	C1	34	27 (79%)	19 (56%)
	C2	131	90 (69%)	76 (58%)
	C3	30	22 (73%)	22 (73%)
LigandFit	C1	33	24 (73%)	18 (55%)
	C2	125	79 (63%)	60 (48%)
	C3	29	19 (66%)	18 (62%)
Surflex	C1	34	26 (76%)	20 (59%)
	C2	131	73 (56%)	62 (47%)
	C3	30	21 (70%)	22 (73%)

[a]Data outside brackets are the number of complexes with RMSD$_{top}$ ≤ 3 Å, and data inside the brackets are the corresponding percentages. All data were obtained at the highest computation level (level 3) plus *in situ* optimization. The complete results of this test are given in the Supporting Information (Table S12).

produced success rates over 80% on subset B3 (buried percentage = 76–96%) when the low-energy conformations were used as docking inputs; whereas the success rates on subset B1 (buried percentage = 24–48%) under the same condition were below 30%. Perola et al.[10] also found that Glide, GOLD, and ICM demonstrated their best accuracy on complexes with buried binding pockets, and consistently lost accuracy with an increase in solvent exposure. They stated 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..." We believe that this is the appropriate interpretation of the results observed in their study and ours. We however want to emphasize that for the protein-ligand complexes with somewhat exposed binding pockets, the inaccuracy of scoring functions is not the only factor accounting for the relatively poor performance of docking programs in such cases. As one can see in Table 7, all four docking programs actually produced fairly good success rates (66–79%) on subset B2 (buried percentage = 49–75%) when the native binding poses were used as docking inputs; while after the docking inputs were switched to low-energy conformations, the success rates fell in the range of (48–66%). This prompts that an effective conformational sampling method is still as important as an accurate scoring function for obtaining better docking results, which should receive adequate attention by the developers of molecular docking programs.

A "hydrophobic index" of the binding pocket on protein was used as the criterion for defining subsets in our third set of tests. One can see in Table 8 that all these docking programs can be divided into two groups by their performance on three subsets when the native binding poses were used as docking inputs: one group includes GLIDE, GOLD + GoldScore, LigandFit, and Surflex, which tend to perform better on relatively hydrophilic (subset C1) or relatively hydrophobic binding pockets (subset C3) than on "ambiguous" binding pockets (subset C2); the other group includes, GOLD + ChemScore and GOLD + ASP, which tend to perform better on relatively hydrophobic binding pockets (subset C3). Note that the influence of inadequate conformational sampling is largely eliminated when the native binding poses are used as docking inputs. Thus, it is reasonable to interpret that the intrinsic design of the scoring functions implemented in these programs accounts for the different performance of these docking programs mentioned above. As a matter of fact, the same trend can also be observed when the low-energy conformations are used as docking inputs. It seems that the elegant balance between polar interactions and hydrophobic effects occurring during protein-ligand binding is still challenging for scoring functions to model. The same challenge actually exists for other scoring methods too. Both Perola et al.[10] and Kontoyianni et al.[11] reported that GOLD tended to perform better on moderately or highly hydrophilic binding pockets; while Glide seemed to be insensitive to the nature of binding pockets. Here, our observations are somewhat different. Our analysis is based on a larger and more diverse test set than those two previous studies. The evaluation methods adopted by us are also more sophisticated. We hope that our results are more convincing because of these technical advantages.

Conclusions

We have conducted a comparative evaluation of four popular commercial molecular docking programs, including Glide (version 4.5), GOLD (version 3.2), LigandFit (version 2.3) and Surflex (version 2.0), on a test set of 195 diverse protein-ligand complexes. This test set is compiled with special emphasis on diversity and redundancy. It surpasses those of the same kind employed in previous studies in terms of size and quality. All four docking programs are able to produce success rates above 60% ($\mathrm{RMSD_{top}} \leq 3$ Å) on the entire test set when the native binding poses are used as docking inputs. Among them, Glide and GOLD + GoldScore are able to maintain this level of accuracy when low-energy conformations are used for instead as docking inputs; whereas LigandFit and Surflex are more sensitive to this change. All four programs are actually able to generate binding poses close to native ones. Thus, the failure in finding the correct binding poses shall largely be attributed to the scoring functions implemented in these docking programs. This is also supported by the observation that the correlations between the binding scores of top-ranked docking solutions produced by these programs are only low to moderate. Apparently, scoring functions still need significant improvements or molecular docking programs perhaps need to consider alternative methods for evaluating binding poses. Additional analyses of the results produced by all four programs on some subsets of protein-ligand complexes indicate that they are generally less capable in handling really flexible ligands or relatively flat binding sites. GOLD + ChemScore and GOLD + ASP tend to perform better on relatively hydrophobic binding sites; while others do not have obvious preference.

All four docking programs are also tested at three different computation levels to evaluate their performance as a function of computational cost. Our results reveal that increasing the computation time on conformational sampling is somewhat helpful to Glide and Surflex but not very much to GOLD and LigandFit for improving docking accuracy. Another observation is whether to enable *in situ* optimization on final docking solutions or not does not make significant difference. The binding poses generated directly by these programs are already reasonable enough in terms of geometry and position. Thus, *in situ* optimization may not be necessary in practice especially when it is time-consuming.

Considering the overall accuracy in both binding pose prediction and binding affinity prediction, our conclusion is that Glide and GOLD have better performance than the other two. Nevertheless, it is relatively complicated to set up a docking job with Glide, and the robustness of this program is a practical concern as it failed on a number of protein-ligand complexes in our test. The problem with GOLD + GoldScore is that it is considerably slower than other docking programs. Therefore, GOLD users may want to consider the other two options, i.e. GOLD + ASP and GOLD + ChemScore, with minor loss in accuracy. Our results may help other researchers narrow their choices among available molecular docking programs down to a few promising ones. We also hope that a third-party evaluation like ours is valuable for the developers of molecular docking programs to improve their methods.

References

1. Kuntz, I. D. Science 1992, 257, 1078.
2. Colman, P. M. Curr Opin Struct Biol 1994, 4, 868.
3. Robertus, J. Nat Struct Biol 1994, 1, 352.
4. Blundell, T. L. Nature 1996, 384 (6604 Suppl.), 23.
5. Gane, P. J.; Dean, P. M. Curr Opin Struct Biol 2000, 10, 401.
6. van Dongen, M.; Weigelt, J.; Uppenberg, J.; Schultz, J.; Wikstrom, M. Drug Discov Today 2002, 7, 471.
7. Waszkowycz, B. Curr Opin Drug Discov Devel 2002, 5, 407.
8. Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L. J Comput Aided Mol Des 2003, 17, 755.
9. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Proteins 2004, 57, 225.
10. Perola, E.; Walters, W. P.; Charifson, P. S. Proteins 2004, 56, 235.
11. Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. J Med Chem 2004, 47, 558.
12. Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. J Med Chem 2005, 48, 962.
13. Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. J Chem Inf Model 2006, 46, 401.
14. Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. J Med Chem 2006, 49, 5912.
15. Onodera, K.; Satou, K.; Hirota, H. J Chem Inf Model 2007, 47, 1609.
16. Bissantz, C.; Folkers, G.; Rognan, D. J Med Chem 2000, 43, 4759.
17. Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. J Comput Aided Mol Des 2004, 18, 333.
18. Hu, X.; Balaz, S.; Shelver, W. H. J Mol Graph Model 2004, 22, 293.
19. Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. J Comput Chem 2005, 26, 11.
20. Zhou, Z. Y.; Felts, A. K.; Friesner, R. A.; Levy, R. M. J Chem Inf Model 2007, 47, 1599.
21. Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. J Chem Inf Model 2009, 49, 1455.
22. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
23. McInnes, C. Curr Opin Chem Biol 2007, 11, 494.
24. Shoichet, B. K. Nature 2004, 432, 862.
25. Lyne, P. D. Drug Discov Today 2002, 7, 1047.
26. Irwin, J. J.; Shoichet, B. K. J Chem Inf Model 2005, 45, 177.
27. Huang, N.; Shoichet, B. K.; Irwin, J. J. J Med Chem 2006, 49, 6789.
28. Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Proteins 2005, 60, 325.
29. Wang, R. X.; Fang, X. L.; Lu, Y. P.; Wang, S. M. J Med Chem 2004, 47, 2977.
30. Wang, R. X.; Fang, X. L.; Lu, Y. P.; Yang, C. Y.; Wang, S. M. J Med Chem 2005, 48, 4111.
31. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. J Med Chem 2004, 47, 1739.

32. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. J Med Chem 2004, 47, 1750.
33. Glide, Version 4.5; Schrödinger LLC: New York, 2007.
34. Schrödinger Suite, Version 2007; Schrödinger LLC: New York, 2007.
35. Jones, G.; Willett, P.; Glen, R. C. J Mol Biol 1995, 245, 43.
36. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. J Mol Biol 1997, 267, 727.
37. Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. J Med Chem 2005, 48, 6504.
38. GOLD, Version 3.2; CCDC Software Ltd.: Cambridge, 2007.
39. Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. J Mol Graph Model 2003, 21, 289.
40. LigandFit, Version 2.3; Accelrys Software Inc.: San Diego, CA, 2007.
41. Discovery Studio, Version 2.0; Accelrys Software Inc.: San Diego, CA, 2007.
42. Jain, A. N. J Med Chem 2003, 46, 499.
43. Jain, A. N. J Comput Aided Mol Des 2007, 21, 281.
44. Surflex, Version 2.0; BioPharmics, LLC: San Mateo, CA, 2007.
45. Sybyl, Version 7.3; Tripos Inc.: St. Louis, MO, 2007.
46. Krovat, E. M.; Steindl, T.; Langer, T. Curr Comput Aided Drug Des 2005, 1, 93.
47. Kirkpatrick, P. Nat Rev Drug Discov 2004, 3, 299.
48. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. J Comput Aided Mol Des 1997, 11, 425.
49. Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Protein Struct Funct Genet 1998, 33, 367.
50. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. J Am Chem Soc 1996, 118, 11225.
51. Mooij, W. T. M.; Verdonk, M. L. Proteins 2005, 61, 272.
52. Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. J Mol Graph Model 2005, 23, 395.
53. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Chem Biol 1995, 2, 317.
54. Gehlhaar, D. K.; Bouzida, D.; Rejto, P. A. In Rational Drug Design: Novel Methodology and Practical Applications; Parrill, A. L.; Reddy, M. R., Eds.; American Chemical Society: Washington, DC, 1999, p. 292–311.
55. Jain, A. N. J Comput Aided Mol Des 1996, 10, 427.
56. Muegge, I.; Martin, Y. C. J Med Chem 1999, 42, 791.
57. Muegge, I. J Med Chem 2006, 49, 5895.
58. Bohm, H. J. J Comput Aided Mol Des 1994, 8, 243.
59. Bohm, H. J. J Comput Aided Mol Des 1998, 12, 309.
60. Ruppert, J.; Welch, W.; Jain, A. N. Protein Sci 1997, 6, 524.
61. Welch, W.; Ruppert, J.; Jain, A. N. Chem Biol 1996, 3, 449.
62. Cheng, T. J.; Li, X.; Li, Y.; Liu, Z. H.; Wang, R. X. J Chem Inf Model 2009, 49, 1079.
63. Zhao, Y.; Cheng, T.; Wang, R. J Chem Inf Model 2007, 47, 1379.
64. Wang, R. X.; Lai, L. H.; Wang, S. M. J Comput Aided Mol Des 2002, 16, 11.
65. Lee, B.; Richards, F. M. J Mol Biol 1971, 55, 379.
66. Bondi, A. J Phys Chem 1964, 68, 441.
67. Wang, R. X.; Lin, F.; Xu, Y.; Cheng, T. J. J Mol Graph Model 2007, 26, 368.
68. Tao, P.; Wang, R. X.; Lai, L. H. J Mol Model 1999, 5, 189.
69. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
70. Cruickshank, D. W. J. Acta Crystallogr D 1999, 55, 583.
71. Goto, J.; Kataoka, R.; Hirayama, N. J Med Chem 2004, 47, 6804.
72. Roversi, P.; Blanc, E.; Vonrhein, C.; Evans, G.; Bricogne, G. Acta Crystallogr D 2000, 56, 1316.
73. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. J Med Chem 2007, 50, 726.
74. Wang, R. X.; Lu, Y. P.; Wang, S. M. J Med Chem 2003, 46, 2287.
75. Wang, R. X.; Lu, Y. P.; Fang, X. L.; Wang, S. M. J Chem Inf Comput Sci 2004, 44, 2114.
76. Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B. O.; Glen, R. C. Org Biomol Chem 2004, 2, 3267.
77. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. J Med Chem 2004, 47, 3032.