Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations

Caterina Bissantz,[‡] Gerd Folkers, and Didier Rognan*[,‡]

Department of Applied Biosciences, ETH Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received July 28, 2000

Three different database docking programs (Dock, FlexX, Gold) have been used in combination with seven scoring functions (Chemscore, Dock, FlexX, Fresno, Gold, Pmf, Score) to assess the accuracy of virtual screening methods against two protein targets (thymidine kinase, estrogen receptor) of known three-dimensional structure. For both targets, it was generally possible to discriminate about 7 out of 10 true hits from a random database of 990 ligands. The use of consensus lists common to two or three scoring functions clearly enhances hit rates among the top 5% scorers from 10% (single scoring) to 25−40% (double scoring) and up to 65−70% (triple scoring). However, in all tested cases, no clear relationships could be found between docking and ranking accuracies. Moreover, predicting the absolute binding free energy of true hits was not possible whatever docking accuracy was achieved and scoring function used. As the best docking/consensus scoring combination varies with the selected target and the physicochemistry of target-ligand interactions, we propose a two-step protocol for screening large databases: (i) screening of a reduced dataset containing a few known ligands for deriving the optimal docking/consensus scoring scheme, (ii) applying the latter parameters to the screening of the entire database.

Introduction

Virtual screening of chemical databases is now a well-established method for finding new leads, provided that a three-dimensional structure of the target is known.[1] When used prior to experimental screening, it can be considered as a powerful computational filter for reducing the size of a chemical library that will be further experimentally tested. As the number of pharmaceutical targets is predicted to dramatically increase in the coming years[2] with the sequencing of the human genome, virtual screening methods will undoubtely play a major role in pharmacogenomics by finding the very first leads of new targets, especially in cases of orphan receptors[3] for which no information on potential ligands is known.

Any virtual screening method has to face two critical issues: docking and scoring. In a first step, the target-bound conformation and orientation (from now on called *pose*) of screened ligands should be predicted with the best possible accuracy. Several docking programs are now available that generally are able to predict known protein-bound ligand poses with averaged accuracies of about 1.5−2 Å.[4] As flexible docking clearly outperforms rigid body matches,[5] most of the current docking programs consider the ligand as a flexible molecule. Flexible docking is generally based on one of the following methods: fast shape matching (Dock,[6] Eudock[7]), incremental construction (FlexX,[8] Hammerhead[9]), Tabu search (Pro_Leads,[10] SFDock[11]), genetic algorithms (Gold,[12] AutoDock3.0,[13] Gambler[14]), evolutionary programming,[15] simulated annealing (AutoDock2.4[16]), Monte Carlo simulations (MCDock,[17] QXP[18]), and distance geometry (Dockit[19]). Only those able to dock a flexible ligand within a reasonable time scale (100−200 s) are suited for virtual screening purposes. Once the ligand has been docked, it should be scored according to the tightness of target−ligand interactions. Again, several scoring methods have been described in the past decade.[20] If one excludes computationally expensive conformational sampling-based methods (free energy perturbations,[21] linear interaction energies approximations[22]) which are the most accurate but unsuitable to database screening, they are basically based on either force-field methods (Dock,[6] Gold[12]), empirical free energy scoring functions (Ludi[23], Chemscore,[24] Score,[25] Fresno,[26] FlexX,[8] Plp[15]), or knowledge-based potential of mean force (Pmf,[27] Drugscore[28]). All of them have been validated for various test sets of high-resolution protein−ligand X-ray structures and are generally able to predict absolute binding free energies within 7−10 kJ/mol. However, it is presently unknown which docking/scoring combinations will provide the best results in terms of hit rates.

With few exceptions,[7,9,29−31] most reported database docking attempts regarding either methodological[32−36] or application aspects[37−41] have been addressed using the pioneer Dock program.[6] Thus, a reference study comparing the merits of several database docking programs is still missing. Furthermore, consensus scoring from two or three independent scoring lists has recently been shown to outperform single scoring.[14] Which scoring function(s) should then be used for ranking potential hits and is there any relationship between docking and ranking accuracies?

To answer these questions, we compared the combined use of three popular database docking algorithms

* To whom correspondence should be addressed. Phone: +41-1-635 60 36. Fax: +41-1-635 68 84. E-mail: didier@pharma.ethz.ch.

‡ New address: Laboratoire de Pharmacochimie de la Communication Cellulaire, UMR CNRS−ULP 7081, 74 route du Rhin, BP 24, F-67401 Illkirch, France.

10.1021/jm001044l CCC: $19.00
Published on Web 11/22/2000

(Dock, FlexX, Gold) with seven scoring functions (Dock, FlexX, Gold, Pmf, Chemscore, Fresno, Score) for screening a 1000-compound library against two different protein targets, thymidine kinase (TK) and the ligand-binding domain of the estrogen receptor α subtype (ERα). A specific database comprising 990 random and 10 known ligands was specifically created for each target. Thus, results of the virtual screening will be examined in terms of (i) docking accuracy (rmsd to known solutions), (ii) scoring accuracy (prediction of the absolute binding free energy), (iii) consensus versus single scoring, (iv) discrimination of active from random compounds, and (v) hit rates and enrichment factors among the top scorers.

Computational Methods

Preparation of Three-Dimensional Databases. The Advanced Chemical Directory (ACD v.2000-1, Molecular Design Limited, San Leandro) was first filtered in order to eliminate chemical reagents,[1,42] inorganic compounds, and molecules with unsuitable molecular weights (lower than 250, higher than 500). Out of the 75 000 remaining molecules, 990 were randomly chosen and their three-dimensional coordinates generated using Corina.[43] Hydrogen atoms and Gasteiger−Marsili atomic charges[44] were then added using a Sybyl (TRIPOS Inc., St. Louis, MO) SPL macro, and the final coordinates stored in a multi mol2 file. For each test case (TK, ER), a set of 10 known ligands was prepared using the above-described procedure starting from a IsisDraw (MDL) 2D structure. Special caution was given to the protonation state of ionizable groups (amines, amidines, carboylic acids) of all 1000 ligands assumed to be ionized at a physiological pH of 7.4. These new stuctures were then appended to the random database to create two final libraries of 1000 molecules: a TK library containing 10 TK ligands and an ER library containing 10 ER antagonists.

Preparation of Protein Coordinates and Definition of Active Sites. Reference protein coordinates used for docking were taken from the X-ray structure of TK in complex with deoxythymidine (pdb entry: 1kim, monomer A)[45] and of ER in complex with 4-hydroxy-tamoxifene (pdb entry: 3ert).[46] Although alternative rotameric states exists for a few TK side chains depending on the bound ligand, we felt that choosing the crystal coordinates of TK in complex with its natural substrate (dT) was a reasonable choice since the latter active site is opened enough to accommodate a broad variety of ligands. Bound ligand and cofactor atoms were then removed. As the coordinates of bound water molecules depend on the type of ligand in the active site (purine vs pyrimidine compounds), all water molecules were removed from the TK binding site. Hydrogen atoms were added when necessary (computing Dock grid energies) using standard Sybyl geometries. For each protein target, the active site was defined as the collection of amino acids enclosed within a 6.5 Å radius sphere centered on the bound ligand. It comprised 16 and 34 amino acids for TK and ER, respectively.

Dock4.01 Docking. First, a Connolly surface of each protein's active site was created using a 1.4 Å probe radius and further used to generate a set of 31 and 35 overlapping spheres for TK and ER, respectively. To compute interaction energies, a three-dimensional grid of 0.35 Å resolution was centered on both protein active sites. Final grids containing 165 672 (dimension: 17.7 × 20.1 × 18.5 Å) and 260 975 points (dimension: 22.1 × 18.7 × 25.0 Å) were obtained for TK and ER, respectively. Energy scoring grids were obtained using an all atom model and a distance-dependent dielectric function ($\epsilon = 4r$) with a 10 Å cutoff. Amber95 atomic charges[47] were assigned to all protein atoms. Database molecules were then docked into the protein active site by matching sphere centers with ligand atoms.[6] A flexible docking of all molecules (peripheral search and torsion drive) with subsequent energy minimization was performed as follows: (i) automatic selection and matching of an anchor fragment within a maximum of 100 orientations, (ii) iterative growing of the ligand using at least 20 conformations (peripheral seeds) for seeding the next growing stage with assignment of energy-favored torsion angles, and (iii) simultaneous relaxation of the base fragments as well as of all peripheral segments and final relaxation of the entire molecule. Orientations/conformations were relaxed (energy score only) in 100 cycles of 100 simplex minimizations to a convergence of 0.1 kcal/mol. The top solution corresponding to the best Dock energy score for each ligand was then stored into a single multi mol2 file.

FlexX1.8 Docking. Unless specified in the Results section, standard parameters of the FlexX program[8] as implemented in the 6.62 release of the SYBYL package were used for iterative growing and subsequent scoring of FlexX poses. Only the top solution was retained and further stored in a single mol2 file.

Gold1.1 Docking. For each of the 10 independent genetic algorithm (GA) runs, a maximum number of 1000 GA operations was performed on a single population of 50 individuals. Operator weights for crossover, mutation, and migration in the entry box were set to 100 100 and 0, respectively. To allow poor nonbonded contacts at the start of each GA run, the maximum distance between hydrogen donors and fitting points was set to 5 Å, and nonbonded van der Waals energies were cut off at a value equal to 10 k_{ij} (well depth of the van der Waals energy for the atom pair i, j). To further speed up the calculation, the GA docking was stopped when the top three solutions were within 1.5 Å rmsd of each other.

Consensus Scoring. All ligands for which a docking solution had been found were rescored using the CScore module of Sybyl6.62 comprising the following scoring functions: ChemScore,[24] Dock, FlexX, Gold, and Pmf.[27] It should be noted that FlexX scores calculated either from FlexX or Cscore are very similar (r^2 about 0.96 for both sets of 1000 complexes). Original Dock4.0 and Gold scores differ from that calculated by Sybyl and thus cannot be compared. Therefore, the Dock, FlexX, and Gold scores proposed by Sybyl were discarded when the corresponding scoring function was coupled to the docking procedure. Two additional scoring functions, Fresno[25] and Score,[26] were used as part of in-house SPL scripts. For rescoring docked poses, atomic coordinates of the target protein were unchanged, and no relaxation of the bound ligand was performed.

Results and Discussion

Virtual Screening of TK Substrates. Thymidine kinase (TK) was chosen as a hard test because of several potential limitations: (i) its binding cavity is accessible to water, (ii) some side chains of the active site adopt rotameric states depending on the bound ligand (induced fit), (iii) the participation of water molecules in mediating ligand binding differs upon the chemical series to which the ligand belongs (purine vs pyrimidine-like substrates, see Figure 1), and (iv) most ligands bind with rather low binding constants (micromolar range). Out of the set of 10 known TK ligands (Figure 1), only two display a submicromolar binding constant (dT and idu) whereas all others bind to TK with micromolar binding constants (from 1.5 to 200 μM; Dr. L. Scapozza, ETH Zürich, personal communication).

Initial docking attempts with various parameter sets for each docking program were first undertaken to determine the best compromise between docking accuracy (rmsd to the TK-bound X-ray structure of each ligand) and speed. Using Dock4.01, flexible docking with rigid-body energy minimization was clearly the method of choice, when compared to rigid docking (with or without further minimization, data not shown). Using a grid resolution of 0.35 Å (alternative resolutions of

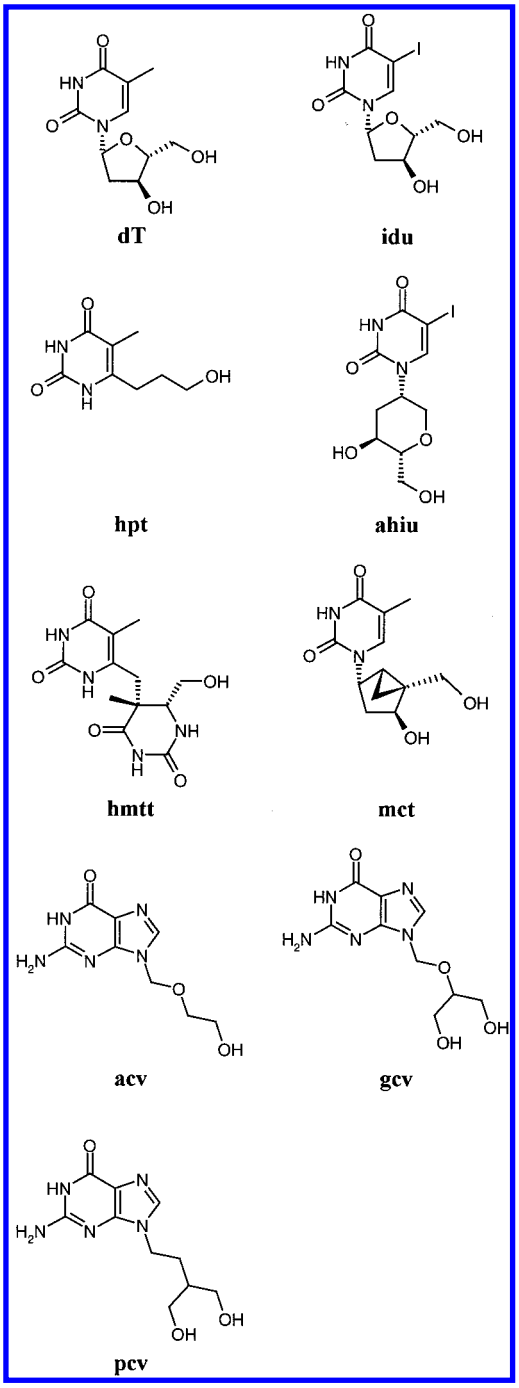

Figure 1. HSV-1 thymidine kinase ligands. The abbreviations are as follows: dT, deoxythymidine; idu, 5-iododeoxyuridine; hpt, 6-(3-hydrody-propyl-thymine); ahiu, 5-iodouracil anhydrohexitol nucleoside; mct, (North)-methanocarba-thymidine; hmtt, (6-[6-hydroxymethy-5-methyl-2,4-dioxo-hexahydro-pyrimidin-5-yl-methyl]-5-methyl-1*H*-pyrimidin-2,4-dione; acv, aciclovir; gcv, ganciclovir; pcv, penciclovir. The structure of one ligand (dhbt) is currently not publicly available.

0.2 and 0.5 Å were also investigated), limiting the number of orientations of the base fragment to 100 (instead of 200 or 500) as well as using 20 conformations for seeding the next growing stage (instead of 10 or 50) provided the best averaged results (data not shown). Only the Dock energy value was computed as it is generally the most robust among the three possible scores (energy score, chemical score, contact score).[14,28,36] Using FlexX, the type of charges used (formal vs partial Gasteiger charges) as well as the introduction of water

Table 1. Rms Deviations (non hydrogen atoms, in Å) of Docked TK Ligands (top solution of each docking tool) from the X-ray Pose[a]

	docking method		
ligand	Dock	FlexX	Gold
dT[b]	0.82	0.78	0.72
idu[c]	9.33	1.03	0.77
ahiu[d]	1.16	0.88	0.63
dhbt[e]	2.02	3.65	0.93
hpt[f]	1.02	4.18	0.49
hmtt[g]	9.62	13.30	2.33
mct[h]	7.56	1.11	1.19
acv[i]	3.08	2.71	2.74
gcv[j]	3.01	6.07	3.11
pcv[k]	4.10	5.96	3.01

[a] X-ray pose means here ligand coordinates merged into the reference protein structure (pdb code: 1kim) after fitting protein backbone atoms. [b] pdb code:1kim. [c] pdb code: 1ki7. [d] pdb code 1ki6. [e] pdb code:1e2p. [f] pdb code:1e2m. [g] pdb code:1e2n. [h] pdb code: 1e2k. [i] pdb code: 2ki5. [j] pdb code: 1ki2. [k] pdb code: 1ki3. Abbreviations are identical to that in Figure 1.

particles during the docking did not significantly affect the docking accuracy. Thefore, standard parameters were also used throughout this study. Last, speed requirements prompted us to utilize only the library screening settings of the Gold software (see Computational Methods). Thus, flexible docking could be perfomed at a pace of about 50−100 s/molecule when using all three docking programs on a standard workstation (SGI Indigo2, R10K processor).

Using above-cited parameters, we first addressed the docking accuray of each tool. The 10 TK ligands have been cocrystallized with TK.[45,48,49] Because the protein coordinates slightly vary with the bound ligand and only one set of protein coordinates was used for docking (pdb code: 1kim), the X-ray pose of each ligand was merged into the reference protein coordinates for comparing X-ray and docked poses. Out the three docking tools used, Gold clearly provides the best docking accuracy. Sixty percent of the known substrates were docked with less than 1.2 Å rms deviation, and even the worst docked ligand (ganciclovir) was orientated with a rmsd of only 3.1 Å (Table 1, Figure 3). Surprisingly, Dock as well as FlexX were not able to find a reliable solution for at least three TK ligands (idu, hmtt, and mct for Dock; hmtt, ganciclovir, and penciclovir for FlexX). For all these ligands, small conformational changes around a few side chains (Gln125 for purine compounds, Tyr132 for iodinated pyrimidines, Arg222 for hpt, dhbt, and mct) are observed in the corresponding X-ray structures when compared to the reference protein coordinates. However, as evidenced by the good performance of the Gold docking tool, the selected protein coordinates do not impair the binding of TK ligands in the active site.

We next looked at the ranking (position in the scoring list) of each TK ligand proposed by the seven scoring functions from the three independent docking poses (Figure 4). FlexX and Pmf scores were found to be the most robust on average. Chemscore and Fresno, two related scoring functions, provided very disappointing rankings from Dock and Gold poses. Score rankings were also rather poor when used in combination with FlexX and Gold poses. It seems rather difficult to explain the performance of each scoring functions with regard to the specific terms they enclose. Scoring functions with rather similar terms (notably a strong directional H-bonding term, e.g., FlexX, Chemscore, Fresno) do not perform equally well. The good perfor-

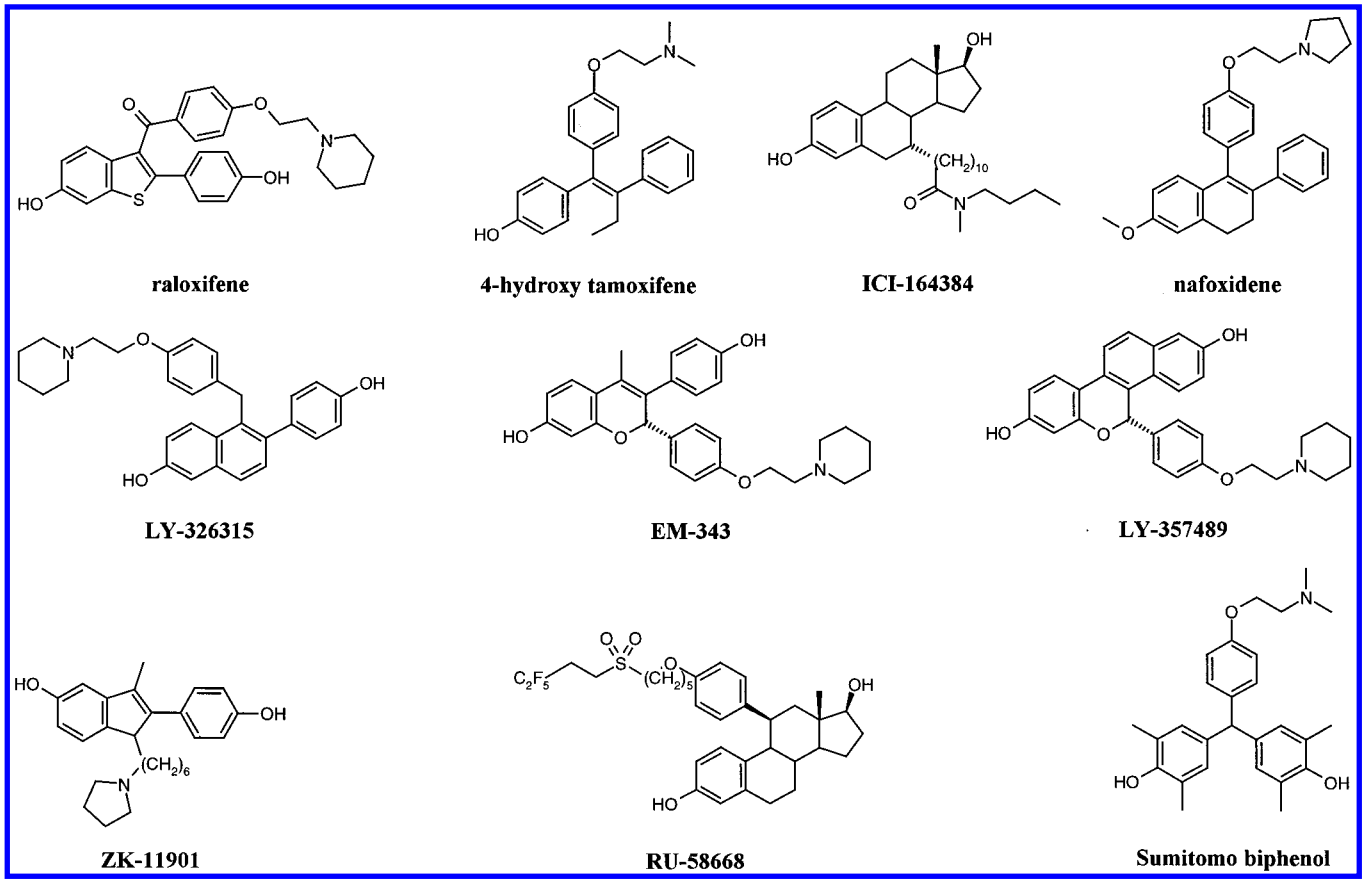

Figure 2. Estrogen receptor (ERα) antagonists.

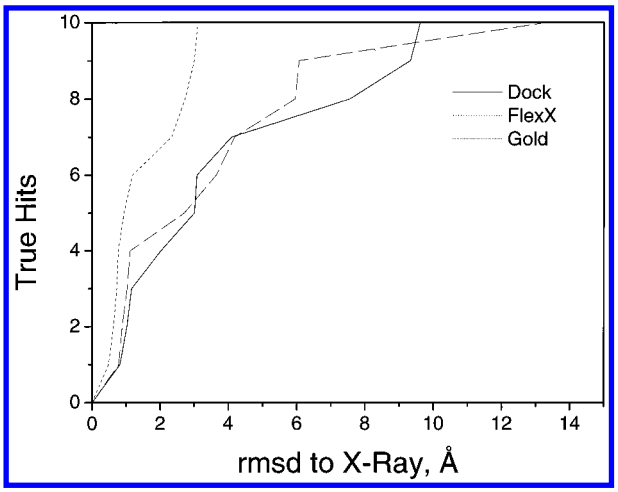

Figure 3. Rms deviations (in Å) of docked TK true hits from their X-ray pose. Docked structures were fitted to the protein-bound X-ray structure of each true hit merged into the reference protein coordinates (pdb code 1kim).

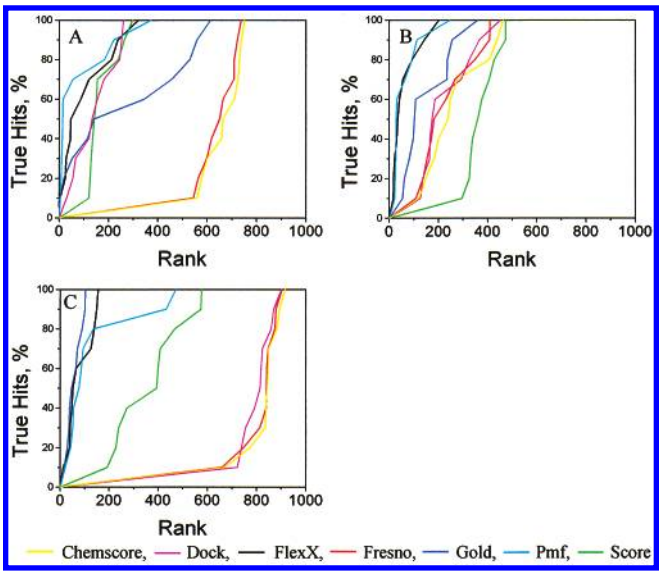

Figure 4. Ranking of TK ligands using a combination of three docking programs and seven scoring functions: (A) Dock docking, (B) FlexX docking, (C) Gold docking.

mance of the Pmf scoring function suggests that at least the true ligands interact through canonical interactions. Altogether, the most homogeneous results were observed using FlexX as docking tool. However, the comparison of the best docking/scoring scheme is clearly in favor of Gold docking and Gold scoring (Figure 5). It was the only combination for which 100% of true hits were found among the top 10% scorers. FlexX/Pmf and Dock/Pmf protocols provided reliable rankings for 70% of true hits but failed in ranking the remaining 30%.

To see whether ranking failures might be linked to docking inaccuracies, we plotted rms deviations from X-ray poses versus ranking obtained for each of the 10 ligands screened by the best three docking/scoring combinations (Figure 6). Excepting three ligands (idu, hmtt, mct) for which Dock clearly failed to find a reliable pose (rmsd above 7 Å) and could therefore not obviously be ranked among the top scorers, no relationships could be found between docking and ranking accuracies (Figure 6). For example, the natural substrate (deoxythymidine) was well docked and ranked by the three best combinations (Dock/Pmf, FlexX/Pmf, Gold/Gold).

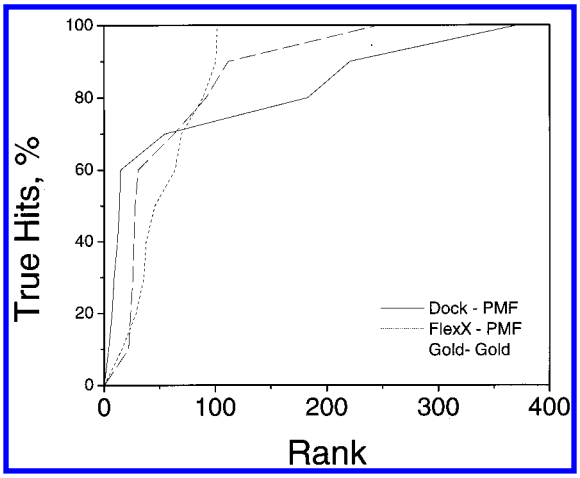

Figure 5. Comparison of the three docking methods each with its best performing scoring function (TK ligands).

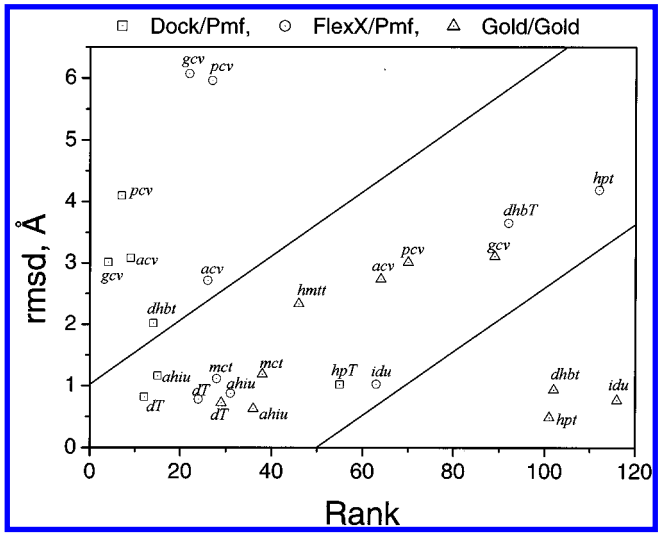

Figure 6. Ranking versus rms deviations from X-ray pose for TK ligands screened with the three best docking/scoring combinations. The lines delimit a confidence area in which docking and ranking accuracies are correlated.

However, penciclovir was poorly docked by Dock (rmsd = 4.1 Å) and FlexX (rmsd = 5.92 Å) but nevertheless well ranked using Pmf scoring (ranked 7th and 27th, respectively). As expected, a good docking does not guarantee an accurate scoring. Gold docked dhbt much better than penciclovir (rmsd of 0.93 and 3.01 Å, respectively) but ranked dhbt worse than penciclovir (102nd and 70th, respectively). Furthermore, filling the active site with water particles significantly improved the FlexX docking of dhbt (rmsd of 0.70 Å vs 3.65 Å for standard FlexX docking) but did not ameliorate FlexX ranking, either (24th versus 18th). One clear explanation of these discrepancies apart from well-known scoring function deficiencies (poor treatment of entropic components to the binding free energy, scoring of single conformations and not thermodynamic ensembles) is that the scoring of misdocked molecules (purine compounds in the present case) is generally overestimated for bigger ligands (Figure 6). Hence most of the molecules misdocked but properly ranked are purine compounds whereas molecules accurately docked but poorly ranked are all pyrimidine analogues (Figure 6).

Thus, it not surprising that no scoring function, whatever the docking method, was able to predict absolute binding free energies for the set of known TK ligands (Table 2). In most of the cases, the observed standard error in prediction was about 6−7 kJ/mol. Only the Gold/Gold combination gave a correlation with some statistical value (predictive $r^2 = 0.507$, standard error of prediction: 4.90 kJ/mol). This is the same combination which already gave the best ranking (Figure 4C). A simple reason explains the poor quantitative predictions observed in the present study. Water molecules have been retrieved from the TK active site. However, they mediate ligand binding, and their coordinates depend on the type of the bound ligand (purine vs pyrimidine compounds). It is interesting to notice that quantitative predictions are much better when two sets of protein coordinates are used with their corresponding crystal water molecules (purine-bound TK, pyrimidine bound TK; $r^2 = 0.98$, $n = 6$). Quantitative predictions being out of reach in the present study does not preclude the interest in virtual screening as a lead finding computational technique. Comparing the distribution of docking energies for active and random compounds clearly demonstrates that it is possible to partly discriminate true hits from random ligands (Figure 7). Using Dock and FlexX poses ranked by the Pmf method (the most accurate in both cases, Figure 4), there is still a significant overlap for at least 30% of true hits between active and random compounds (Figure 7A,B). Considering that the present target is really a very hard test, we can consider that the number of observed false positives (about 20%) is low enough for ensuring a successful computational screening of potential TK ligands even with Dock and FlexX docking tools. Using the best possible combination (Gold/Gold, Figure 7C), a good discrimination of active from random compounds could be reached with as few as 10% false negatives.

We next looked at hit rates (percentage of known true hits) among the top 5% scorers, which would correspond to a reasonable number of molecules (2500) to test experimentally if we would have screened a 50 000-compound virtual library. Confirmed hit rates from the best scoring function associated to each docking tool ranged between 10 and 12% for all three possible combinations (Figure 8). Using a consensus scoring out of two scoring functions, hit rates were raised significantly up to 24%. On average, the combined use of FlexX and Pmf scoring functions was found to be the most robust (Figure 8). Interestingly, the observed enriched hit rates were shown to be rather independent of the docking tool used (especially after single scoring), though Gold was shown to clearly outperform Dock and FlexX in terms of docking accuracy. We can then conclude, at least for the present case, that scoring is

Table 2. Prediction of Absolute Binding Free Energies from Docked Poses

	Dock pose		FlexX pose		Gold pose	
scoring function	r^2 [a]	s [b]	r^2	s	r^2	s
Chemscore	0.117	6.57	0.096	6.65	0.125	6.54
Dock	0.231	6.13	0.361	5.59	0.194	6.28
FlexX	0.061	6.77	0.040	6.85	0.199	6.26
Fresno	0.212	6.21	0.108	6.61	0.387	5.47
Gold	0.024	6.91	0.354	5.61	0.507	4.90
Pmf	0.039	6.85	0.041	6.85	0.018	6.93
Score	0.047	6.83	0.228	6.15	0.141	6.48

[a] Correlation coefficient after multiple linear regression. [b] Standard error, kJ/mol.

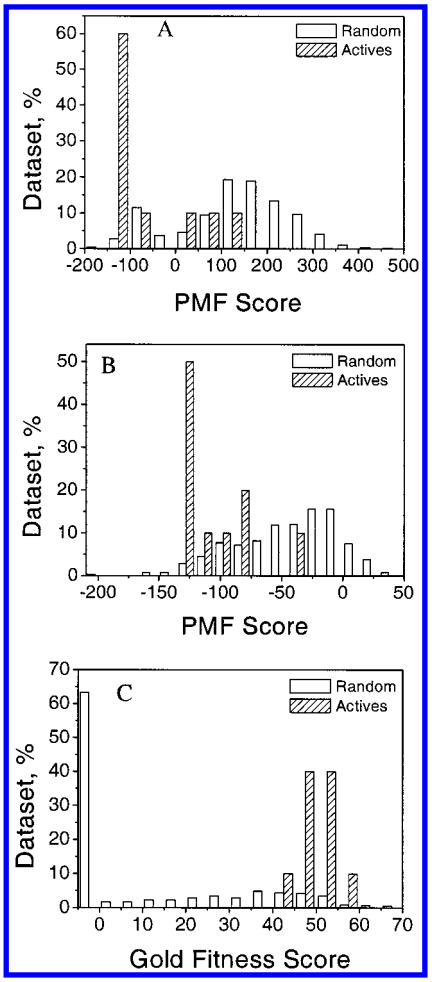

Figure 7. Discrimination of TK true hits from random ligands using the following docking/scoring combinations: (A) Dock/Pmf, (B) FlexX/Pmf, (C) Gold/Gold. Results are indicated as percentages of the total number of ligands for which a docking solution had been found (Dock: 10 true hits, 774 random ligands; FlexX: 10 true hits, 488 random ligands; Gold: 10 true hits, 927 random ligands). Please notice that Pmf scores of true hits should be as low as possible whereas Gold fitness scores should be as high as possible.

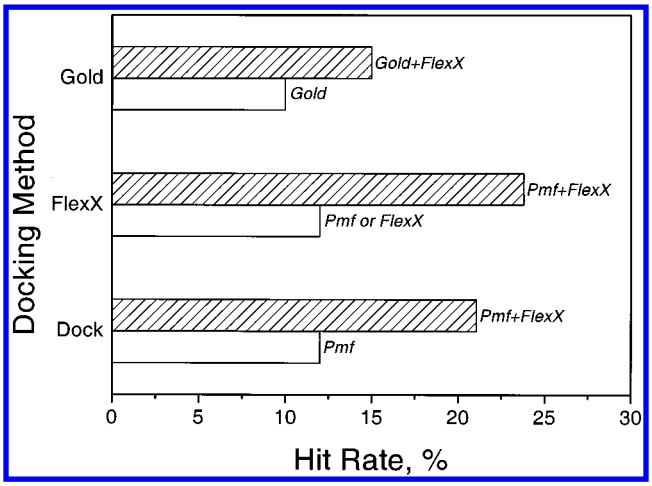

Figure 8. Hit rates (% of known TK ligands) among the top 5% scorers after single or consensus scoring.

more important than docking in database screening. It should be noted that not all possible consensus scoring combinations were studied here as many scoring functions (Fresno, Score, Dock, Chemscore) could not identify any true TK ligand in the top 5% scorers. It is thus of crucial importance to first determine on a reduced dataset which combination provides the best hit rates before screening a large chemical database. Unless additional scoring functions are directly implemented in docking programms,[50] one usually restrains rescoring to the top scorers identified by the native scoring tool. This means that database docking with Dock, FlexX, and Gold, and selection of the top 5% scorers, would have missed at least 9, 4, and 5 out of the 10 true hits, respectively. We would then advise the use of FlexX or Gold for searching a large library for potential TK ligands.

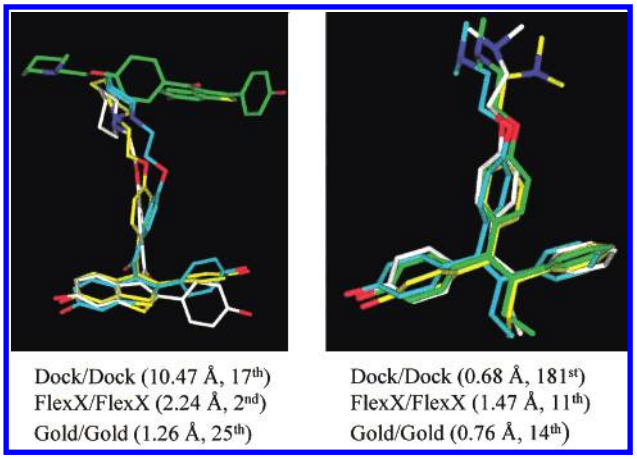

Figure 9. Overlay of X-ray and docked poses for two ERα antagonists, raloxifen (left panel) and 4-hydroxy-tamoxifen (right panel). Docked structures were fitted to the protein-bound X-ray structure merged into the reference protein coordinates (pdb code: 3ert). The number in parentheses indicates the rms deviation from the X-ray pose and the rank, respectively. The following color coding was used: X-ray pose, white carbon atoms; Dock pose, green carbon atoms; FlexX pose, yelow carbon atoms; Gold pose: cyan carbon atoms.

Virtual Screening of ERα Antagonists. We next tested our docking/scoring strategy using a target more suited for virtual screning. The ERα receptor was selected for several reasons: (i) its binding cavity is less opened to solvent than that of TK, (ii) the topology of its active site is less dependent on the nature of the bound ligand, (iii) all selected true hits bind in the low nanomolar range, and (iv) ERα has already been successfully used as virtual screening target using a Tabu search (TS)-based flexible docking method.[31]

Starting from the coordinates of the ERα receptor in complex with 4-hydroxy-tamoxifen (pdb code: 3ert), a 1000-compound database containing 10 known ERα antagonists (Figure 2) was docked and scored as previously described. The docking accuracy of each docking tool could only be assessed for the two ligands (raloxifen, 4-hydroxy-tamoxifen) for which a ERα-bound X-ray structure was available.[46,51] As for TK ligands, the best docking poses were obtained using the GA-based Gold algorithm (Figure 9). FlexX also performed well, whereas Dock failed to find a reliable pose for raloxifen. Again, no clear relationships between docking and ranking could be found. Although raloxifen was clearly misdocked by Dock, this ligand still belongs to the top scorers of the Dock energy list (Figure 9).

Rescoring all successfully docked ligands with six additional scoring functions showed clear docking/

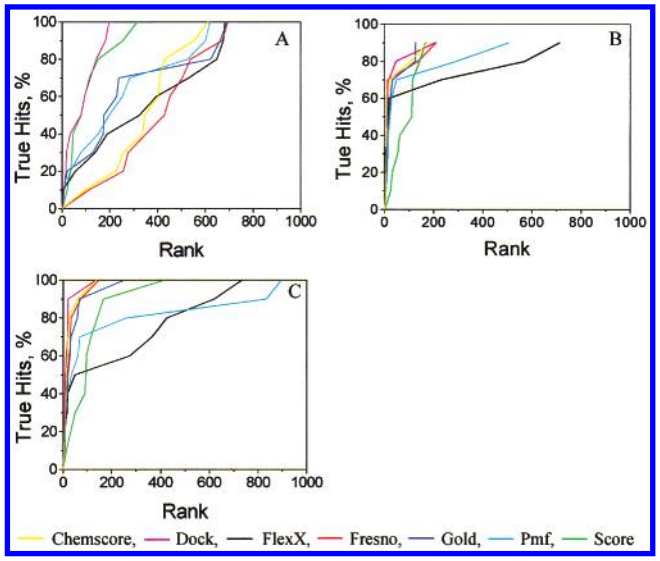

Figure 10. Cumulative ranking of ER antagonists using a combination of three docking programs and seven scoring functions: (A) Dock docking, (B) FlexX docking, (C) Gold docking.

ranking trends. Obtained rankings of true hits were generally much higher than those previously observed for TK ligands. Two out of the seven scoring functions performed well when ranking Dock poses (Figure 10A). The number of favored scoring functions was raised to five for FlexX and Gold poses (Figure 10B,C). Dock interaction energies provided the best ranking, whatever the docking program (Figure 10). Its force field based nonbonded van der Waals term mirrors rather well the apolar protein−ligand interactions. Surprisingly, the two most robust scoring functions for screening TK ligands (FlexX and Pmf) were among the poorest in the present case (Figure 10A−C). Dock and Score rankings were clearly the best when rescoring Dock poses (Figure 10A). Using FlexX poses, all scoring functions but FlexX and Pmf performed also well. In the latter cases, the same two ligands (ICI-164384, EM-343) were poorly ranked. The ICI compound was apparently well docked by FlexX in the apolar binding pocket, with the exception of the long acyclic side chain. The EM compound was clearly misdocked. By looking at the rigidified analogue (LY-357489) which was well docked, one could find a plausible explanation of the observed FlexX failure to dock the EM ligand. Its phenolic group has clearly rotated around the Csp2-C.ar bond with respect to the conformationally rigidified naphthol analogue (LY-357489) and could not be accommodated in the apolar ER pocket. For one ligand (RU-58668), no docking solution could be found using FlexX. It is interesting to notice that both of the misdocked compounds lack the canonical basic side chain found in most of ER antagonists. As for TK ligands, the Gold poses led to the best ranking of known ERα antagonists (Figure 10C). Using a Gold docking/Dock scoring scheme, 9 out of 10 true hits were ranked among the top 2% scorers (Figure 11).

Whatever the docking method, the virtual screning of ER ligands led to a clear discrimination of true hits from random ligands (Figure 12A−C). About 20% of false positives still remain using a Dock ranking/Dock docking combination (Figure 12A) but all true hits belong to the top scorers. Using FlexX poses, two known

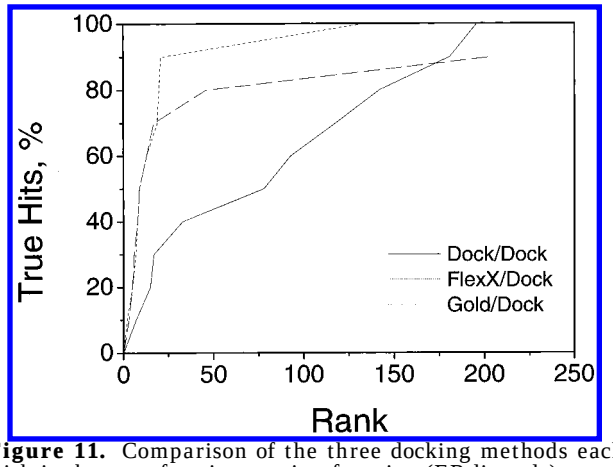

Figure 11. Comparison of the three docking methods each with its best performing scoring function (ER ligands).

ER ligands overlapped the random pool (Figure 12B) but very few random compounds (about 1.5%) got a high score. The distinction between active and random compounds is even better using Gold poses (Figure 12C) as the rate of false negatives and false positives were only 1.6 and 10%, respectively. Averaged hit rates among the top 5% scorers were higher and less dependent on the scoring function with FlexX and Gold poses than with Dock orientations (Figure 13). Averaged enrichement factors in true ER ligands among the top 5% scorers (factor 13) obtained with a single scoring function from FlexX or Gold poses were rather similar to that recently reported by a different tabu search-based method.[31] Using a consensus list from either two or three scoring functions led to much higher hit rates, up to 70%. This is a remarkable result if one considers that all seven scoring functions have been taken into account for averaging, and not only the top combination as previously shown for TK ligands.

Conclusions

For the two targets described herein, Gold poses were clearly shown to be the most suitable for virtual screening. Interestingly, standard parameters of the three docking tools performed rather well for both targets. However, scoring seems to predominate docking accuracy as no relationhip between docking and ranking could be found in the present study. As predicting experimental binding free energies with a high accuracy is still out of reach, it is of utmost importance to select a scoring function able do discriminate active from random compounds. Our results demonstrate that predicting which scoring function would perform the best is a very difficult task. Two scoring functions (Pmf, FlexX) that performed well in one case (TK screning) were found among the poorest in another case (ER screening). Although general rules cannot be drawn due to the limited number of virtual screening targets used in the current study, FlexX and Pmf scores tend to perform well for a highly polar active site (TK) whereas Dock scores were found to be the most reliable for the apolar ER active site. However, finding out a single scoring function to rank virtual hits is not a major concern as consensus scoring definitely outperforms single scoring whatever the target and the docking tool used. Thus, we propose a two steps protocol for optimiz-

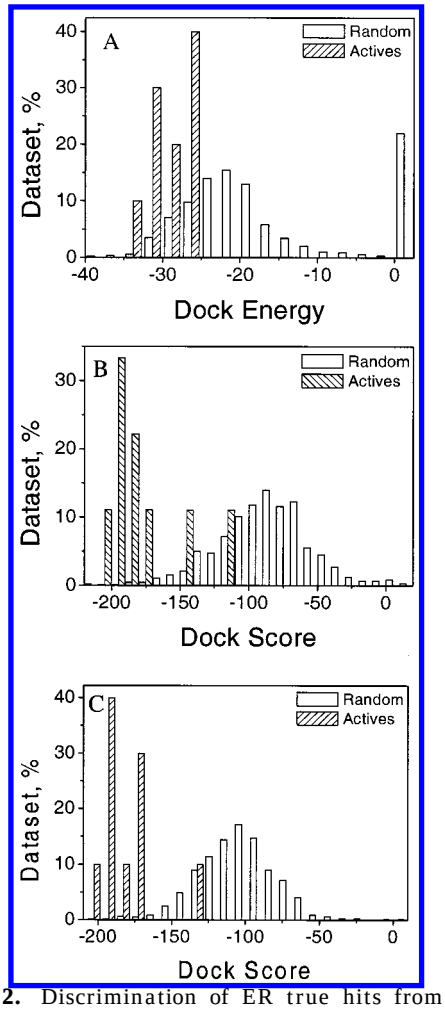

Figure 12. Discrimination of ER true hits from random ligands using the following docking/scoring combinations: (A) Dock/Dock, (B) FlexX/Dock, (C) Gold/Dock. Results are indicated as percentages of the total number of ligands for which a docking solution had been found (Dock: 10 true hits, 907 random ligands; FlexX: 9 true hits, 876 random ligands; Gold: 10 true hits, 926 random ligands). Dock energies correspond to interaction energies calculated by Dock4.0 whereas Dock scores have been computed by rescoring all poses with the SYBYL CScore module. Main differences between Dock energies and Dock scores are in the different dielectric functions used for computation (Dock energies: $4Dr_{ij}$; Dock score, Dr_{ij}).

ing the docking/scoring combination: (i) screening of a reduced dataset (1000 compounds) containing a few known true hits to find out the best docking/consensus scoring combination, and (ii) screening the full library with the optimized docking/scoring scheme.

This study shows that virtual screening of chemical databases is a powerful method for finding new hits and prioritizing ligand synthesis and experimental testing. Using consensus scoring, a reduced virtual dataset covering 0.5−1% of the full library, enriched by a factor 20 to 70, and containing 50 to 90% of all true hits could be designed, even for the difficult TK target where induced fit as well as water intercalation plays a significant role in ligand binding. However, because predicting the correct pose and, more importantly, the experimental binding free energy is much more difficult, it still cannot be considered as a general lead optimization method.

Acknowledgment. We sincerely thank TRIPOS GmbH (Münich, Germany) and the Cambridge Cristallographic Data Center (Cambridge, U.K.) for providing us with FlexX/Cscore and Gold releases.

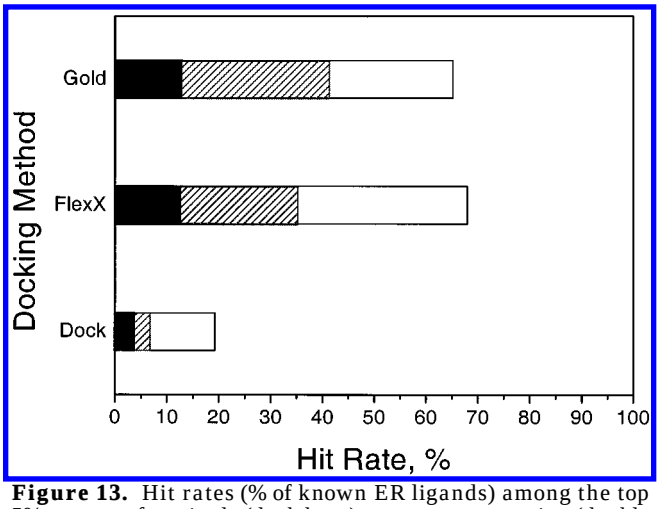

Figure 13. Hit rates (% of known ER ligands) among the top 5% scorers after single (dark bars) or consensus scoring (double scoring: pattern bars; triple scoring: white bars). Hit rates have been averaged over 7 (single scoring), 21 (double scoring), and 35 (triple scoring) possible ranking lists.

References

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening − an overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(2) Drews, J. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960−1964.
(3) Civelli, O. Functional genomics: the search for novel neurotransmitters and neuropeptides. *FEBS Lett.* **1998**, *430*, 55−58.
(4) Dixon, J. S. Evaluation of the CASP2 docking section. *Proteins* **1997**, *Suppl 1*, 198−204.
(5) Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: a way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 565−582.
(6) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(7) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **2000**, *43*, 401−408.
(8) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−89.
(9) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449−462.
(10) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998**, *33*, 367−382.
(11) Hou, T.; J., W.; Chen, L.; Xu, X. Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu search. *Protein Eng.* **1999**, *12*, 639−647.
(12) Jones, G.; Wilett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(13) Morris, G. M.; Goodsell, D. S.; Halliday, R.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(14) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. 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.
(15) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317−324.
(16) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated aneealing. *Proteins* **1990**, *8*, 195−202.
(17) Liu, M.; Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 435−451.

(18) McMartin, C.; Bohacek, R. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.
(19) Metaphorics LLC, Piemont, CA 94611.
(20) Tame, J. R. H. Scoring functions: A view from the bench. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 99−108.
(21) Kollman, P. A. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Acc. Chem. Res.* **1996**, *29*, 461−469.
(22) Aqvist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385−391.
(23) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(24) Eldridge, M.; Murray, C. W.; Auton, T. A.; Paolini, G. V.; Lee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(25) Rognan, D.; Laumoeller, S. L.; Holm, A.; Buus, S.; Tschinke, V. Predicting binding affinities of protein ligands from three-dimensional coordinates: Application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.* **1999**, *42*, 4650−4658.
(26) Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: a new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379−384.
(27) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(28) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(29) Böhm, H. J. On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 623−632.
(30) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: a system to select "quasi-flexible" ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153−174.
(31) Baxter, C.; Murray, C. W.; Waszkowycz, B.; Li, J.; Sykes, R. A.; Bone, R. G. A.; Perkins, T. D. J.; Wylie, W. New approach to molecular docking and ist application to virtual screening of chemical databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 254−262.
(32) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175−1189.
(33) Oshiro, C. M.; Kuntz, I. D. Characterization of receptors with a new negative image: use in molecular docking and lead optimization. *Proteins* **1998**, *30*, 321−336.
(34) Lorber, D. M.; Shoichet, B. K. Flexible ligand docking using conformational ensembles. *Protein Sci.* **1998**, *7*, 938−950.
(35) Godden, J. W.; Stahura, F.; Bajorath, J. Evaluation of docking strategies for virtual screening of compound databases: cAMP-dependent serine/threonine kinase as an example. *J. Mol. Graph. Model.* **1998**, *16*, 139−143.
(36) Knegtel, R.; Wagener, M. Efficacy and selectivity in flexible database scoring. *Proteins* **1999**, *37*, 334−345.
(37) Ring, C. S.; Sun, E.; McKerrow, J. H.; Lee, G. K.; Rosenthal, P. J.; Kuntz, I. D.; Cohen, F. E. Structure-based inhibitor design by using protein models for the development of antiparasitic agents *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3583−3587.
(38) Bodian, D. L.; Yamasaki, R. B.; Buswell, R. L.; Stearns, J. F.; White, J. M.; Kuntz, I. D. Inhibition of the fusion-inducing conformational change of influenza hemagglutinin by benzoquinones and hydroquinones. *Biochemistry* **1993**, *32*, 2967−2978.
(39) Filikov, A. V.; James, T. L. Structure-based design of ligands for protein basic domains: application to the HIV-1 Tat protein. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 229−240.
(40) Debnath, A. K.; Radigan, L.; Jiang, S. Structure-based identification of small molecule antiviral compounds targeted to the gp41 core structure of the human immunodeficiency virus type 1. *J. Med. Chem.* **1999**, *42*, 3203−3209.
(41) Hopkins, S. C.; Vale, R. D.; Kuntz, I. D. Inhibitors of kinesin activity from structure-based computer screening. *Biochemistry* **2000**, *39*, 2805−2814.
(42) Oprea, T. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(43) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.
(44) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(45) Champness, J. N.; Bennett, M. S.; Wien, F.; Visse, R.; Summers, W. C.; Herdewijn, P.; de Clerq, E.; Ostrowski, T.; Jarvest, R. L.; Sanderson, M. R. Exploring the active site of herpes simplex virus type-1 thymidine kinase by X-ray crystallography of complexes with aciclovir and other ligands. *Proteins* **1998**, *32*, 350−361.
(46) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95*, 927−937.
(47) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, J., K. M.; Ferguson, D. M.; Spellmeyer, D. M.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(48) Wild, K.; Bohner, T.; Folkers, G.; Schulz, G. E. The structures of thymidine kinase from herpes simplex virus type 1 in complex with substrates and a substrate analogue. *Protein Sci.* **1997**, *6*, 2097−2106.
(49) Prota, A.; Vogt, J.; Pilger, B.; Perozzo, R.; Wurth, C.; Marquez, V. E.; Russ, P.; Schulz, G. E.; Folkers, G.; Scapozza, L. Kinetics and crystal structure of the wild-type and the engineered Y101F mutant of Herpes simplex virus type 1 thymidine kinase interacting with (North)-methanocarba-thymidine. *Biochemistry* **2000**, *39*, 9597−9603.
(50) Muegge, I.; Martin, Y. C.; Hadjuk, P. J.; Fesik, S. W. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.* **1999**, *42*, 2498−2503.
(51) Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997**, *389*, 753−758.

JM001044L