New Scoring Functions for Virtual Screening from Molecular Dynamics Simulations with a Quantum-Refined Force-Field (QRFF-MD). Application to Cyclin-Dependent Kinase 2

Ph. Ferrara,*,[†] A. Curioni,[§] E. Vangrevelinghe,[†] T. Meyer,[‡] T. Mordasini,[§] W. Andreoni,[§] P. Acklin,[†] and E. Jacoby[†]

Novartis Institutes for BioMedical Research, Discovery Technologies, CH-4002 Basel, Switzerland, Novartis Institutes for BioMedical Research, Oncology Research, CH-4002 Basel, Switzerland, and IBM Research, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland

Received July 15, 2005

A recently introduced new methodology based on ultrashort (50−100 ps) molecular dynamics simulations with a quantum-refined force-field (QRFF-MD) is here evaluated in its ability both to predict protein−ligand binding affinities and to discriminate active compounds from inactive ones. Physically based scoring functions are derived from this approach, and their performance is compared to that of several standard knowledge-based scoring functions. About 40 inhibitors of cyclin-dependent kinase 2 (CDK2) representing a broad chemical diversity were considered. The QRFF-MD method achieves a correlation coefficient, R^2, of 0.55, which is significantly better than that obtained by a number of traditional approaches in virtual screening but only slightly better than that obtained by consensus scoring ($R^2 = 0.50$). Compounds from the Available Chemical Directory, along with the known active compounds, were docked into the ATP binding site of CDK2 using the program Glide, and the 650 ligands from the top scored poses were considered for a QRFF-MD analysis. Combined with structural information extracted from the simulations, the QRFF-MD methodology results in similar enrichment of known actives compared to consensus scoring. Moreover, a new scoring function is introduced that combines a QRFF-MD based scoring function with consensus scoring, which results in substantial improvement on the enrichment profile.

1. INTRODUCTION

Computer-aided structure-based drug design (SBDD) methods aim at predicting the binding mode of a ligand in the binding site of a protein or any molecular target and at obtaining an estimate of the binding affinity. These methods involve two computational steps, docking and scoring. First, multiple protein−ligand configurations, called *poses*, are generated using a docking program. Several of these programs have the ability to generate poses close to the native structure for many targets. Then, a scoring function is used to compute the affinity between the receptor and the ligand for each pose. One of the requirements that a scoring function should satisfy is that it must be sufficiently fast to be used in high throughput docking (HTD). Therefore, SBDD methods usually make use of a single pose to calculate the binding energy. The fixed pose approximation represents an important barrier in the development of a more accurate scoring function and may partly explain the low correlation with the experimentally determined affinities[1] and the moderate enrichment that current HTD programs obtain in virtual screening applications.[2−5]

The most rigorous computational techniques for binding affinity calculation are the free energy perturbation and thermodynamic integration methods. These methods, which employ molecular dynamics (MD) or Monte Carlo (MC) simulations, are well suited to compute the binding energies of a series of congeneric ligands.[6,7] The drawbacks are that they are computationally very intensive and not practical for ligands that are structurally very different. A variety of approximations to these approaches have been developed and those that require only simulations (MD or MC) at the two endpoints of the binding process are particularly attractive, since the computational cost is drastically reduced. These methods, which include the linear interaction energy (LIE) approach,[8] the extended linear response (ELR) model[9] and the molecular mechanics Poisson−Boltzmann surface area (MMPB/SA) method,[10] have been successfully validated in a variety of applications.[11−17] In these studies, success was defined as the ability to rank related ligands with respect to their experimental binding affinity. However, in a virtual screening application, the foremost goal is to identify true hits in a database of mainly nonbinders, and therefore in this case it is more relevant to calculate enrichment. Apart from one study, which evaluated the MMPB/SA approach in its ability to discriminate actives from inactives,[18] end-point free energy models, such as the LIE or the MMPB/SA ones, have not yet been assessed in their ability to enrich a virtual screening hit list with known active compounds.

Recently, a novel methodology, which is based on a two-step MD simulation with quantum-refined force-fields (QRFF-MD), has been proposed to compute binding affinities.[19] This

* Corresponding author phone: ++41 (61) 324 6591; fax: ++41 (61) 324 3357; e-mail: philippe.ferrara@novartis.com.

[†] Novartis Institutes for BioMedical Research, Discovery Technologies.

[‡] Novartis Institutes for BioMedical Research, Oncology Research.

[§] IBM Research, Zurich Research Laboratory.

10.1021/ci050289+ CCC: $33.50 © 2006 American Chemical Society
Published on Web 11/01/2005

method goes beyond the fixed pose approximation by generating a weighted ensemble of protein−ligand configurations prior to estimate the binding affinity. This sampling is achieved by performing classical MD simulations in explicit solvent using the GROMOS96 force-field refined so as to better represent the electrostatics of the ligand. This refinement is the novelty of the QRFF-MD methodology and is obtained in a two-step procedure. The ligand coordinates are first extracted from the configuration of the complex. Using only the ligand conformation, i.e., without the protein, and starting from the GROMOS96 parametrization, the ligand atomic charges are refined on the basis of ab initio calculations of the electrostatic potential, subject to specific restraints in order to preserve the consistency of the original parametrization (see below). The protein−ligand complex is then placed in a box of water molecules and a first MD simulation of 50 ps is carried out using the refined ligand charges. After 50 ps, a second refinement of the ligand charges is made prior to another 50 ps MD simulation. In this case, the ligand conformation corresponding to the final snapshot after 50 ps is used with the charges obtained after the first refinement. Interaction energies, which include a van der Waals and a Coulombic contribution, are calculated from averaging over the configurations generated by the MD simulations. These interaction energies are used as an estimate of protein−ligand affinities. In the QRFF-MD methodology, the unbound states are not explicitly considered. Using the unbound states may increase the “structural noise”, because the intramolecular energies do not cancel with this approach. Furthermore, the simulations for the unbound states are usually started from the bound state, and convergence is therefore slower than for the simulation for the bound state. Solvation is considered in the MD simulations, in the sense that the simulations are carried out in explicit solvent. Solvation is not considered in the QRFF-MD energies, since these energies are made up of a van der Waals and a Coulombic interaction term. The relatively fast procedure (100 ps) is intended to provide a compromise between accuracy and speed. The details are explained in ref 19, where application was made to the prediction of the affinity of a small set of HIV-1 protease inhibitors. Clearly, the final target of the two-step QRFF-MD approach is high-throughput screening. However, while its speed performance depends on the hardware available, it is mandatory to first evaluate its accuracy, and this can easily be done in a low-throughput mode.

In the present study, we evaluated the QRFF-MD methodology and derived scoring functions in their ability to rank about 40 known inhibitors of CDK2 and to select these actives out of a database of around 600 ligands, which are considered to be inactive. Protein kinases are currently a source of many important targets of interest to the pharmaceutical industry. Given the abundance of structural information, which includes several high-resolution crystal structures of this enzyme complexed with a variety of diverse inhibitors, CDK2 represents an ideal target for SBDD.[20−23] We selected a set of 38 active compounds, for which experimental data of activities are available and for which the binding modes have been elucidated by either crystallography or modeling (Figure 1 and Table 1). Here, we used this method to re-rank a HTD hit list generated by the docking program Glide.[24] After parametrization, the 650 protein−ligand complexes in this hit list and the known active compounds were subjected to an MD simulation for 100 ps. Average ligand−protein and ligand−water interaction energies were extracted from the simulations in order to calculate the binding energies. Finally, we present a comparison of the QRFF-MD results with those obtained by the LIE and ELR models, consensus scoring, and a variety of scoring functions commonly used in docking tools. Since MD simulations of more than 600 protein−ligand complexes have not been carried out by others, our work goes beyond the studies based on end-point free energy models that have been published so far.

Table 1. IC_{50} of the 38 CDK2 Inhibitors Considered in This Work

ligand	IC_{50} (μm)	ligand	IC_{50} (μm)
1	0.007	20	0.0057
2	0.001	21	0.0089
3	0.03	22	1
4	0.096	23	0.66
5	25	24	0.025
6	1	25	0.33
7	1.6	26	0.01
8	1.6	27	2.2
9	0.006	28	7
10	1	29	0.07
11	0.07	30	0.7
12	0.06	31	0.01
13	13	32	0.017
14	10	33	0.28
15	0.048	34	10
16	2	35	0.68
17	0.035	36	0.48
18	0.01	37	0.129
19	0.56	38	0.058

2. METHODS

2.1. The QRFF-MD Method. 2.1.1. Calculation of the Ligand Partial Charges. GROMOS atomic charges were assigned to the ligands by identifying building blocks (Figure 2). A charge of zero was assigned to the functional groups that do not have a GROMOS parametrization. After energy minimization of the protein−ligand complex in water, we extracted the ligand coordinates to calculate the ligand partial charges using density functional theory. For each ligand, the ab initio molecular electrostatic potential (MEP) was calculated on a three-dimensional grid in the density-functional theory approach (DFT) using the Car-Parrinello molecular dynamics program (CPMD).[25] This ab initio MEP was then fitted to a distribution of atom-centered charges with the use of a harmonic potential that restrains the fitted charges to the unrefined GROMOS ones.[19] These calculations required nearly 24 h on a 32-processor p655 IBM system for 650 ligands.

2.1.2. Topology Files for Low-Throughput Screening. A list of about 40 building blocks was identified, for which no GROMOS parametrization is available. Therefore we first generated the topology files for all the ligands using the program PRODRG,[26] which are meant to be used in conjunction with the GROMOS87 force-field, extracted the parameters that do not exist in the GROMOS96 force-field, and added them to the GROMOS96 parameter file. In total, 22 bonds and 4 angles were identified. The topology files generated by PRODRG were also used to compile a list of

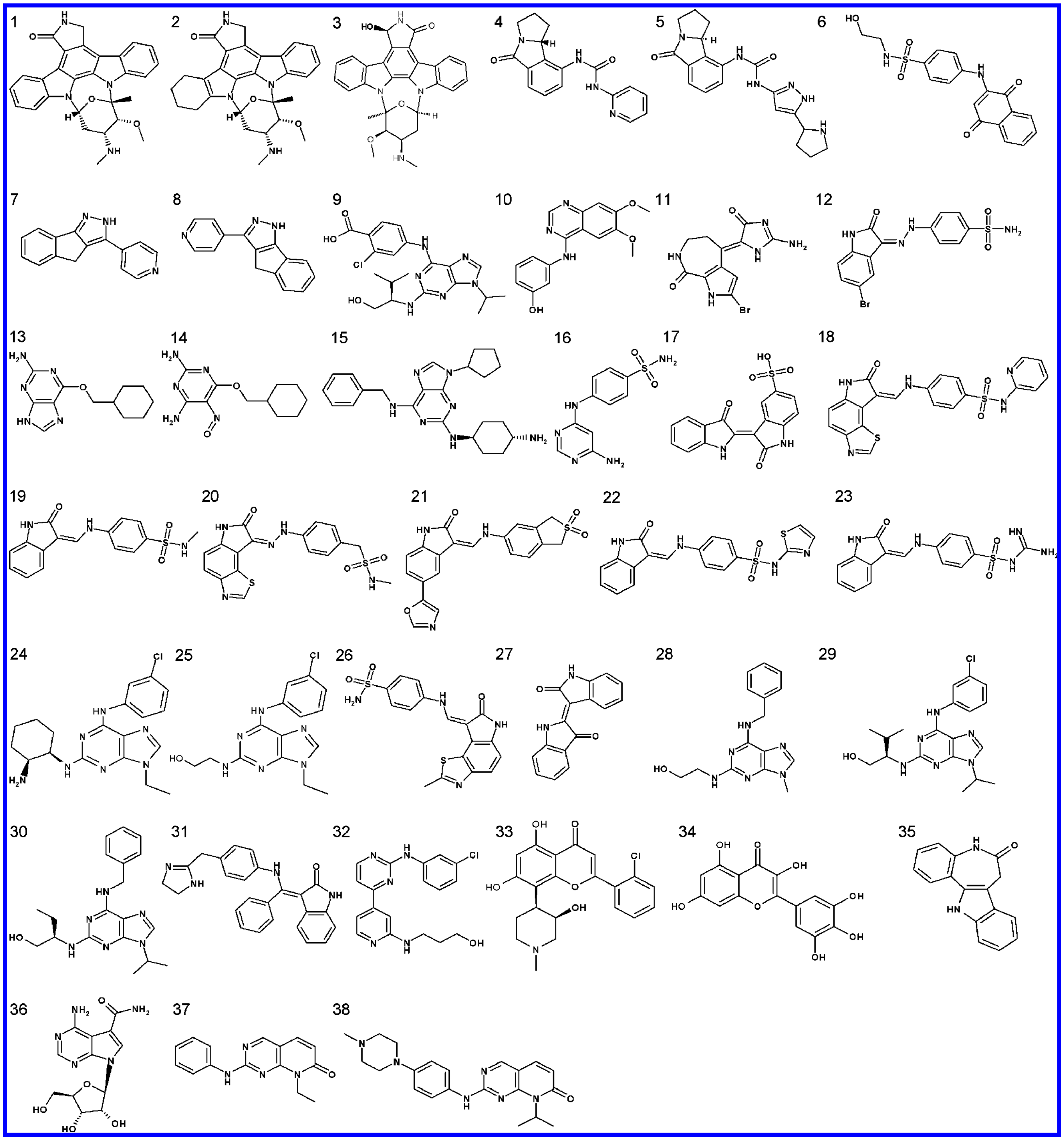

Figure 1. CDK2 inhibitors considered in this work.

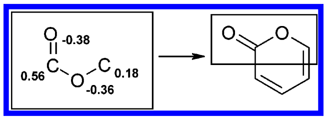

Figure 2. Definition of a building block for an ester. Left panel: the GROMOS partial charges are shown. Right panel: example of an ester in a six-membered planar ring.

pairs and triplets of atom types, which have more than one set of parameters assigned. This case can be exemplified in Figure 2, where the parameters for the angle between the two oxygens and the carbon depend on the location of these atoms, i.e., different values would be assigned if the ester group was in a five-membered ring. Around 20 pairs and 100 triplets of atoms were identified in this way. With all these ingredients at hand, it was straightforward to build topology files for all the compounds compatible with the GROMOS96 force-field.

2.2. Application to CDK2. Database for High-Throughput Docking. The 6465 compounds that belong to the Available Chemical Directory[27] and that are similar to compounds registered in the Novartis Corporate Database were selected and expressed as an sdf file. This data set contains 7 active compounds with experimentally known binding modes and affinities. An active compound is defined here as a compound with an IC_{50} below 10 μm. Thirty-nine diverse compounds, of which 35 are actives, were taken from the literature and added to the data set (Figure 1 and Table 1). For each compound, 3D coordinates were generated using Corina.[28] The compound structures were then ionized assuming a pH of 7.0 with the ionizer module supplied by Schrödinger.[30] Compounds were finally subject to a minimization based on the OPLS-AA force-field.

Preparation of the Binding Site Model. The structure of the protein target of interest, cyclin-dependent kinase 2 (PDB id # 1AQ1), with its bound inhibitor (staurosporine) was taken from the Protein Data Bank.[29] All water molecules were removed, and the protein structure was prepared according to the standard procedure described in the First Discovery manual.[30]

Docking using Glide. The default input parameters implemented in Maestro were used for the generation of the command files for the docking of the small-molecule database against CDK2. For the generation of the scoring grid, the van der Waals radii of the nonpolar protein atoms were not scaled. To limit the number of false positive compounds, a pharmacophoric constraint (hydrogen bond with the NH of Leu83, since all potent CDK2 inhibitors form this hydrogen bond) was used in the docking experiment. The best pose for each compound, as evaluated by the Glide scoring function, was written out.

Postprocessing of HTD Data. After the HTD experiment, the successfully docked compounds were sorted into a ranked list based upon the Glide scores. The compounds in the top 10% (646) of the ranked list were used as starting conformations for the MD simulations. [The ranked list does not contain all the compounds of the initial database, since Glide could not find an acceptable docking pose for a few compounds.] Due to the inaccuracy of the Glide scoring function, 11 active compounds were not ranked in the top 10%, and therefore the final hit list contains 31 active compounds (5% of the data set). Furthermore, 89% of the 646 compounds were not experimentally tested and therefore were supposed to be inactive. We assessed the performance of the scoring functions that are implemented in the CScore module of Tripos.[31] This module contains versions of the DOCK (D-Score),[32] PMF (PMF-Score),[33] Gold (G-Score),[34] ChemScore (ChemScore),[35] and FlexX (F-Score)[36] scoring functions. The consensus score of a pose, Z, was defined as $Z = \sum z_i$, where z_i is the normal deviate score z for a pose with a scoring function i. The sum is taken over the five scoring functions of CScore and z is defined as $z = (x - \bar{x})/\sigma$, where x is the raw score, $\bar{x}$ is the average raw score for all the 646 poses, and σ is the standard deviation of the raw scores for all the 646 poses. For each scoring function, the poses of 646 ligands in the hit list were ranked according to their scores. These ranked lists were then used to generate the enrichment and Receiver Operating Characteristic (ROC) curves.[37] These two curves are not equivalent. A ROC curve describes the tradeoff between sensitivity and specificity. Sensitivity is defined as the ability of the model to detect true positives, while specificity is its ability to avoid false negatives. The area below a ROC curve can be used to quantify the enrichment. A ROC value greater than 0.9 is considered excellent, and a value below 0.6 represents no enrichment.

MD Simulation Protocols. The simulations were performed in explicit solvent using the GROMOS biomolecular simulation package together with the GROMOS96 force-field (version 43A1).[38] Each protein−ligand complex and each ligand in its free state were placed at the center of periodic truncated octahedron box, which was filled with SPC water molecules.[39] The minimum distance between the solute and the box wall is 10 Å. To relax the configuration of the solvent, a steepest descent minimization was carried out in which the solute atoms were restrained to their initial positions using a harmonic potential. The system was equilibrated by performing a 10 ps simulation at constant volume with the solute positions restrained, followed by 10 ps at constant pressure. The restraints were then removed, and the system was gradually heated to 300 K by performing 5 ps simulations at 50, 100, 200, and 300 K. At 300 K, an additional 40 ps simulation at constant temperature and pressure was carried out. After this equilibration phase, the production run was started. A Berendsen thermostat with a coupling decay time of 0.1 ps was applied to maintain the temperature of the system at a constant value of 298 K.[40] The pressure was kept constant by linear coupling to a reference pressure of 1 atm, with a coupling decay time of 0.5 ps. Periodic boundary conditions were applied. Nonbonded interactions were calculated using a twin-range cutoff. Short-range interactions (within 9 Å) were evaluated every time step, while longer-range interactions (within 14 Å) were evaluated every 10 time steps and kept constant between updates. A reaction field correction was used to approximate electrostatic interactions outer the 14 Å cutoff. The SHAKE algorithm was used to fix the length of the covalent bonds where hydrogen atoms are involved. A time step of 1 fs was used to integrate the equation of motions, and structures were saved each 1 ps. Simulations were carried out for 50 ps. Afterward, the ligand partial charges were again subjected to the refinement procedure, using the ligand conformation corresponding to the final snapshot of the first phase. An additional 50 ps simulation was performed using these newly refined charges. Including heating and equilibration, the 100 ps simulations required nearly 6 weeks on a 32-processor p655 IBM system for 650 ligands.

LIE and ELR Models. The linear interaction energy (LIE) approach[8] has shown promising results in several studies.[11−13] In the original method, the binding free energy (ΔG_{bind}) for a protein−ligand complex was calculated using the difference in the average interaction energies between the unbound ligand in solution and the bound ligand with protein and water:

$$\Delta G_{\mathrm{bind}} = \alpha\langle\Delta V_{\mathrm{vdW}}\rangle + \beta\langle\Delta V_{\mathrm{elec}}\rangle + \gamma \qquad (1)$$

In eq 1, the ensemble averages, which are obtained using either molecular dynamics or Monte Carlo simulations, represent the difference in van der Waals and Coulombic interaction energies in the bound and unbound states. The coefficients α, β, and γ are determined by fitting the simulation results to experimental affinities. The LIE model can be extended to include other descriptors, which led to the extended linear response (ELR) expression:[14,16]

$$\Delta G_{\mathrm{bind}} = \sum c_i \varphi_i + \gamma \qquad (2)$$

Here, the φ_i's represent physically meaningful quantities relevant to protein−ligand binding, and the c_i's are the

Table 2. Descriptors Considered in the ELR Analysis

symbol	description
$\Delta E_{\mathrm{LP,elec}}$	ligand−protein Coulombic interaction energy
$\Delta E_{\mathrm{LP,vdW}}$	ligand−protein van der Waals interaction energy
$\Delta E_{\mathrm{LW,elec}}$	ligand−water Coulombic interaction energy
$\Delta E_{\mathrm{LW,vdW}}$	ligand−water van der Waals interaction energy
$\Delta\Delta E_{\mathrm{elec}}$	$\Delta E_{\mathrm{LP,elec}} + \Delta E_{\mathrm{LW,elec}}$ (bound) $- \Delta E_{\mathrm{LW,elec}}$ (unbound)
$\Delta\Delta E_{\mathrm{vdW}}$	$\Delta E_{\mathrm{LP,vdW}} + \Delta E_{\mathrm{LW,vdW}}$ (bound) $- \Delta E_{\mathrm{LW,vdW}}$ (unbound)
ΔHB	change in number of hydrogen bonds for the ligand
#RB	number of rotatable bonds for the ligand

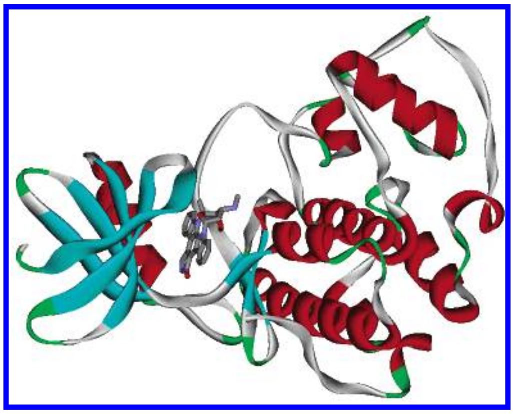

Figure 3. Overall structure of the CDK2−staurosporine complex.

corresponding weighting parameters. The ELR model can be seen as an extended empirical scoring function. In this work, the eight descriptors listed in Table 2 were considered.

Biological Results. Thirty compounds were selected and ordered from the Novartis chemical archives for biological testing in a CDK2 inhibitor assay. These compounds were chosen because they were ranked as top scorers by either consensus scoring or by a QRFF-MD-based scoring function (see below). The assay was performed as in ref 41.

3. RESULTS AND DISCUSSION

3.1. Prediction of Binding Affinities. We first assess the ability of the QRFF-MD methodology to rank experimentally determined binding affinities (Figure 4 and Table 3). To compare computed and experimental affinities, it is in principle necessary to use the ligand conformation in its native state. Therefore, each pose generated by Glide for the binding of the known inhibitors was visually inspected. When the pose was wrong, a "native" conformation was generated by superimposing the two binding sites. The ligand was then minimized in the rigid binding site of the protein. This procedure was applied for six ligands.

In Figure 4, we investigated the effect of performing a simulation by comparing the QRFF-MD results (4a) to those obtained by the force-field of the first-step refinement using only the initial configuration ("pose") of the simulation ("QRFF-pose") (4b). This in turn is compared with the results obtained when using the ligand charges assigned by the original GROMOS force-field ("GROMOS-pose" in Figure 4c). The QRFF-MD energies were extracted from the 2-step 100 ps simulations and averaged giving an equal weight to the configurations sampled over the whole simulation. Given that the force-fields used in the first and second 50 ps runs are different, this average should not be considered as rigorously correct but only as an approximation that is however a posteriori justified given the small changes found in the two steps (Note that the protocol in ref 19 was slightly different.). In both cases, we only consider the interaction energy between the ligand and the protein, which corresponds to the enthalpic contribution to the free energy of binding. For this particular set of inhibitors there are no experimental values available for the enthalpic contribution to the free energy of binding. Therefore, we cannot compare the calculated binding affinities with the experimental Δ*H*'s of binding. Apart from an outlier clearly visible on the left panel of Figure 4 (ligand 5), QRFF-MD and QRFF-pose yield very similar binding energies. The correlation value between the QRFF-MD and the QRFF-pose energies is 0.81 (0.89 without ligand 5). Furthermore, the QRFF-MD and QRFF-pose energies are reasonably correlated with the experimental binding affinities. Without ligand 5, the R^2 value is 0.55 for these two energy functions. The QRFF-MD results slightly differ when the energies are extracted from either the first or the second half of the simulations. The aforementioned correlation factors for the first 50 ps and last 50 ps are 0.57 and 0.51, respectively. Finally, the QRFF-MD energies obtained from the first 50 ps and the last 50 ps correlate with a R^2 of 0.93. This result shows that refining the ligand

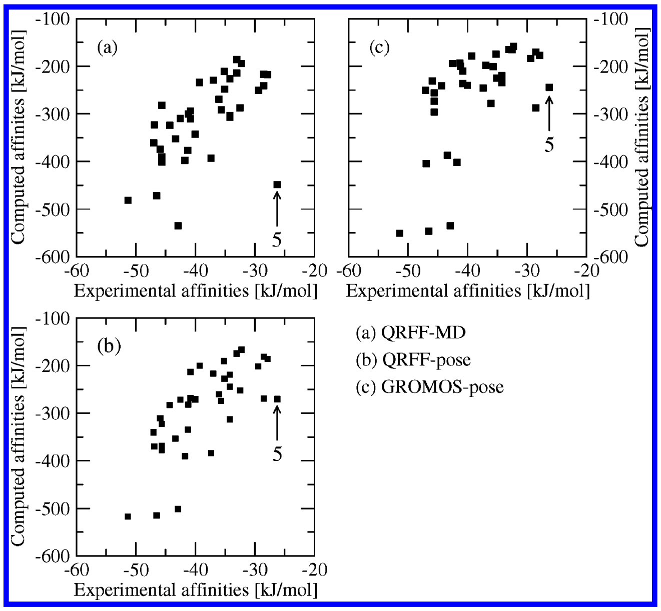

Figure 4. The computed vs the experimentally determined binding energies for (a) QRFF-MD, (b) QRFF-pose, and (c) GROMOS-pose are shown. The equation $\Delta G = RT\ln\mathrm{IC}_{50}$ at 298 K was used to calculate the experimental affinities.

Table 3. Correlation Factors (R^2) between Computed and Experimental Binding Energies for QRFF-MD and a Variety of Scoring Functions[a]

scoring function	R^2
QRFF-MD	0.55
QRFF-pose	0.55
D-Score	0.52
Consensus	0.50
G-Score	0.50
QRFF-MD (ligand−water)	0.40
GROMOS-pose	0.37
Glide	0.35
ChemScore	0.32
PMF-Score	0.19
F-Score	0.18

[a] The ligand number 5 was removed to compute the R^2 values.

charges after 50 ps does not significantly affect the ligand–protein interaction energies.

We notice that the origin of the failure of QRFF-MD to predict the binding affinity of ligand 5 is unclear. The simulation was performed with a proton on the secondary amine. Using the deprotonated state leads to very similar protein–ligand interaction energies (data not shown). It has been shown that compound 5 is about 100-fold more potent on CDK4 compared with CDK2.[42] The main difference between the two ATP binding sites is that CDK4 has additional space caused by the difference in the size of the side chain of residue 89 (CDK2:Lys, CDK4:Thr). Therefore, the loss of potency comes mainly from the repulsion between Lys89 and the pyrrolidine ring of ligand 5, which does not seem to be reproduced by the MD simulation.

Table 3 shows the correlation factors (R^2 value) between computed and experimental binding energies for QRFF-MD and a variety of scoring functions. QRFF-MD and QRFF-pose yield the highest correlation with a R^2 value of 0.55. Table 3 also shows that adding the ligand–water interaction energy to the ligand–protein energy leads to a significant decrease in performance ($R^2 = 0.40$ instead of $R^2 = 0.55$). It has been found in a study of 61 CDK2 inhibitors, based on a ELR analysis, that the van der Waals and Coulombic ligand–water interaction energies do not represent significant descriptors.[16] On the other hand, taking into account the ligand–water interaction energy was shown to improve the affinity prediction of the small set of HIV-1 protease inhibitors considered in the original application of this method.[19] The role of water in ligand binding needs to be further investigated, since it seems to be binding site dependent.

D-Score, G-Score, and consensus scoring give slightly lower R^2 values than QRFF-MD and QRFF-pose (between 0.50 and 0.52) (Table 3). We note that D-Score and G-Score are force-field based scoring functions. Finally, PMF-Score and F-Score yield low R^2 values (<0.20), which shows that a particular scoring function can perform very poorly and that consensus scoring function can circumvent this problem, at least in this case. It is not straightforward to estimate the uncertainties in the experimental affinities, but it is generally assumed that they are in the range of 1 order of magnitude (5.7 kJ/mol).

Other models were evaluated in their ability to predict binding energies from a simulation. A LIE analysis was carried out for the CDK2 inhibitors and the following parameters were obtained (see eq 1): $\alpha = 0.26$ ($p = 0.0009$), $\beta = 0.004$ ($p = 0.91$), and $\gamma = -14.98$ ($p = 0.027$). The p-values of the linear regression are indicated in parentheses, and the correlation value for this fit is $R^2 = 0.30$. The low R^2 value as well as the p-values show that the LIE model is not suitable in this case. A similar conclusion was found in a study on the binding of 40 HIV reverse transcriptase inhibitors.[14] As a result, a multiple linear regression analysis for the ELR model was performed using the descriptors listed in Table 2, and the following equation was obtained as optimal using two descriptors (see eq 2):

$$\Delta G_{\text{bind}}\,[\text{kJ/mol}] = 0.039\langle\Delta E_{\text{LP,elec}}\rangle + 0.093\langle\Delta E_{\text{LP,vdW}}\rangle - 16.94 \quad (3)$$

The correlation coefficient is 0.57, the leave-one-out cross-validated q^2 is 0.48, and the rms error is 4.15 kJ/mol. The R^2 value is marginally better than that obtained without fitting ($R^2 = 0.55$). Adding other descriptors did not lead to significantly better correlation values. This result suggests that the ELR approach is not suitable in this case, at least with the descriptors listed in Table 2. In a study of 61 CDK2 inhibitors based on Monte Carlo simulations in combination with the ELR method, the ligand–protein Coulombic and van der Waals interaction energies as well as the change in the total number of hydrogen bonds for the inhibitor upon binding emerged as the most significant descriptors.[16] The R^2 and the leave-one-out cross-validated q^2 were 0.76 and 0.72, respectively, significantly higher than in this work. The MC simulations for the protein–ligand complexes consisted of five cycles of annealing with a total sampling of 50 million configurations. Although it cannot be ruled out that our MD-based results are due to insufficient sampling, the critical difference here is that our data set of inhibitors contains a much larger diversity than that in ref 16.

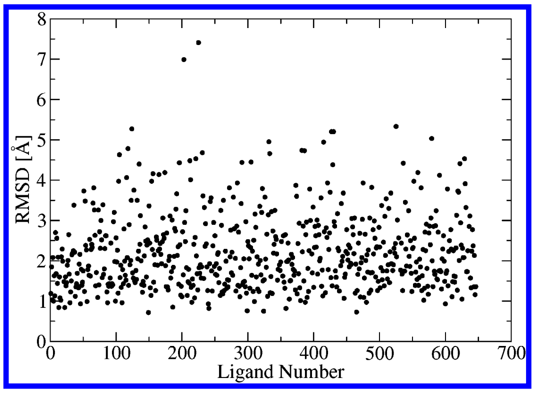

Figure 5. RMSD of the last configuration in the simulation from the docked one for the 646 ligands in the hit list.

3.2. Enrichment of Known Actives in Database Screening. Starting from the configuration obtained by Glide, MD simulations were carried out for the 646 ligands in the hit list. Figure 5 shows the root-mean-square deviation (RMSD) for the heavy atoms between the minimized docked configuration and that after 100 ps simulation. There is a significant spreading around an average value of 2.2 Å, which shows that major conformational changes can occur in a relatively short time scale. Large conformational changes are expected for the ligands whose binding modes in the docked conformation are wrong. The RMSD values for the active compounds range between 0.8 Å and 2.7 Å with an average value of 1.6 Å, which shows that all the actives remain in a configuration that is close to that obtained by Glide. The ligand with the largest RMSD (7.4 Å) between the configuration after 100 ps and that obtained by Glide is shown in Figure 6. The hydrogen bond formed by the amide nitrogen of LEU83 with the hydroxyl group of the ligand is lost during the simulation. The high RMSD value suggests that the unbound state is preferred over the bound state, which can be expected given the small size of the ligand. Nevertheless, only an IC_{50} determination could verify this conclusion.

We evaluated the ability of the QRFF-MD methodology to discriminate active compounds from inactive ones. Figure 7 shows the enrichment curves obtained by QRFF-MD, QRFF-pose, and consensus scoring. All the inhibitors shown in Figure 1 were considered to determine the enrichment curves. These three scoring functions perform much better

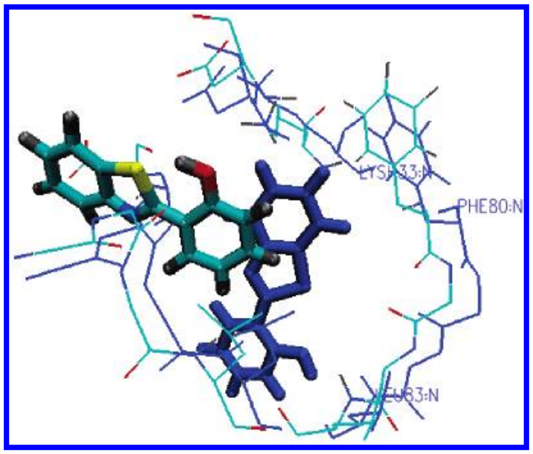

Figure 6. Structure of the ligand with the largest RMSD (7.4 Å) between the last configuration in the simulation (colored by atom type) and that obtained by Glide (colored in blue). The ligand (thick lines) as well as key residues in the binding site (thin lines) are shown.

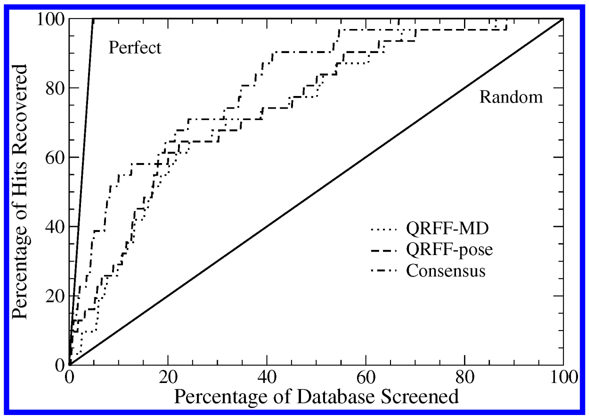

Figure 7. Enrichment curves for QRFF-MD, QRFF-pose, and consensus scoring. The curve "Perfect" is based on the assumption that all the compounds that have not been experimentally tested are inactive. All the inhibitors shown in Figure 1 were considered.

to pick up the active compounds than a random model. QRFF-MD and QRFF-pose yield similar results and slightly worse than those obtained by consensus scoring. This result agrees with a very recent study, which compared for five targets the enrichments obtained by MMPB/SA with those obtained using only the docked structure.[18] The use of dynamics improved the enrichment only for one target (p38 kinase), while for the other targets (COX-2, estrogen receptor, neuraminidase, and thrombin) the enrichments were comparable or even worse. The enrichment curve for QRFF-MD was obtained by averaging the ligand−protein interaction energies over the 100 ps simulations. Using either the first or the second half of the simulations leads to very similar enrichment curves. This result indicates that our simulations are long enough to yield averages with low fluctuations. Although we cannot exclude that much longer MD simulations would significantly improve the results, one should keep in mind that the higher computational requirement would be barely affordable, particularly in virtual screening applications.

Figure 8 shows the top scored ligand (MFCD00078944) found by QRFF. In the pose found by Glide, the amide nitrogen of LEU83 makes a hydrogen bond with the hydroxyl group of the ligand. We note that none of the active compounds is hydrogen-bonded with the NH of LEU83 by such a functional group. Although the RMSD between the docked pose and the configuration after 100 ps is only 1.6 Å, this hydrogen bond is lost at the end of the simulation, and instead the NH of LEU83 is hydrogen-bonded with a water molecule. The comparison with the staurosporin (Figure 3), which binds CDK2 with an IC_{50} of 7 nM, shows that MFCD00078944 is much more flexible. The latter and the former ligands contain 8 and 2 rotatable bonds, respectively. This result suggests that the ligand entropy change upon binding is not fully accounted for in the QRFF-MD methodology. Moreover, the MD simulations sample states of the ligand in a particular conformation of the protein. Other biologically relevant conformations of the protein may not be sampled in the simulations. This deficiency could be overcome by increasing the sampling of the conformational space of the bound and unbound states or, alternatively, by performing free energy calculations.[43,44]

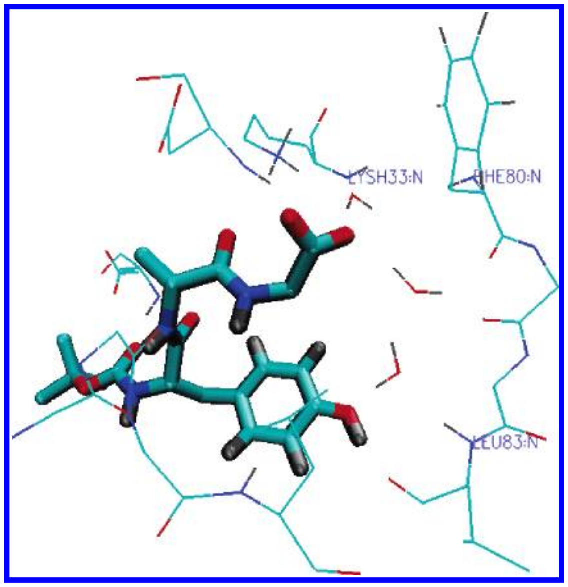

Figure 8. Structure after 100 ps of the lowest energy ligand found by QRFF-MD. The ligand (thick lines) as well as key residues in the binding site (thin lines) are shown. Three water molecules are also indicated.

In a docking study based on CDK2 and neuraminidase, it has been shown that enrichment is significantly higher for the actives whose binding modes are predicted correctly.[20] In our case, removing the five known ligands that have been docked incorrectly leads to slightly better enrichments (not shown). To do that, we compared the binding modes obtained by Glide to those elucidated by crystallography or modeling. However, this conclusion should be considered with caution, since it is based on a very small number of ligands.

We evaluated other models to calculate the enrichment curve from the simulations. In the best model that we obtained, the binding free energy is calculated as follows:

$$\Delta G_{\mathrm{bind}} = \langle \Delta E_{\mathrm{LP}} \rangle + \langle \Delta E_{\mathrm{LP}} - \Delta E_{\mathrm{LW}}(\mathrm{unbound}) \rangle + \langle \mathrm{RMSD}_{\mathrm{last}} \rangle \quad (4)$$

E_{LP} and E_{LW} represent the ligand−protein and ligand−water interaction energies, respectively. $\mathrm{RMSD}_{\mathrm{last}}$ stands for the RMSD between the configuration after 100 ps and the minimized docked one. This term penalizes the configurations with a high RMSD value, since the ligands with a wrong binding mode or a very weak binding affinity should

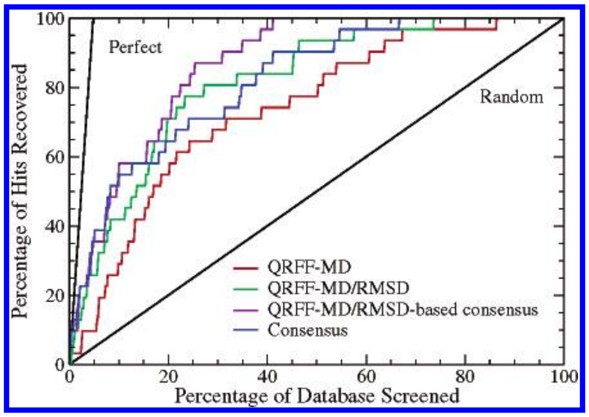

Figure 9. Enrichment curves for QRFF-MD, QRFF-MD/RMSD, QRFF-MD/RMSD-based consensus scoring and consensus scoring. QRFF-MD/RMSD-based consensus scoring is made up of QRFF-MD/RMSD and consensus scoring. The curve "Perfect" is based on the assumption that all the compounds that have not been experimentally tested are inactive. All the inhibitors shown in Figure 1 were considered.

Table 4. Area under ROC Curves,[a] Percentage of Actives Recovered in the Top 20% of the Database Screened,[b] and Percentage of Database Screened at Which All the Actives Are Recovered[c] for the Scoring Functions Studied in This Work[d]

scoring function	area[a]	%[b]	%[c]
QRFF-MD/RMSD-based consensus	0.89	71.0	41.2
Consensus	0.84	64.5	66.7
QRFF-MD/RMSD	0.83	71.0	73.7
F-Score	0.83	71.0	95.2
PMF-Score	0.80	58.1	82.8
QRFF-pose	0.76	58.1	88.5
Glide	0.76	54.8	94.4
QRFF-MD	0.75	54.8	86.4
G-Score	0.75	54.8	96.4
QRFF-MD (ligand−water)	0.70	38.7	77.6
GROMOS-pose	0.69	38.7	94.4
ChemScore	0.66	54.8	99.4
D-Score	0.65	35.5	88.9

[d] Qualitative interpretation of the area under ROC curves is as follows: 0.0−0.6, fail; 0.6−0.7, poor; 0.7−0.8, fair; 0.8−0.9, good; 0.9−1.0, excellent. All the inhibitors shown in Figure 1 were considered.

in principle undergo a conformational change in the simulations. The first term is normalized between −1 and 0 and the two last ones between 0 and 1. We refer to these results as "QRFF-MD/RMSD". The enrichment curve obtained by QRFF-MD/RMSD is a significant improvement with respect to the QRFF-MD one (Figure 9). Furthermore, a "new" consensus scoring, which sums up the QRFF-MD/RMSD energies and consensus scoring (as defined previously) after normalizing both scoring functions between 0 and 1, performs better than "traditional" consensus scoring (Figure 9 and Table 4). We did not use any weighting parameter in this sum, since there is a priori no reason to favor either scoring function (QRFF-MD/RMSD or consensus scoring) if no information on active compounds is available.

Table 4 shows for the scoring functions studied in this work the values for the area under the ROC curve, the percentage of actives recovered in the top 20% of the database screened, and the percentage of database screened at which all the actives are recovered. QRFF-MD/RMSD-based consensus scoring yields the best result with good enrichment close to excellent and a ROC curve value of 0.89. This scoring function ranks all the actives in the top 40% of the database, which represents a significant improvement with respect to all the other scoring functions. As for the correlation values, adding the ligand−water interaction energies or using the ligand charges assigned by GROMOS lead to a significant decrease in performance. Among the scoring functions implemented in CScore, F-Score and PMF-Score perform best with good enrichment and ROC curve values of 0.83 and 0.80, respectively. It is worth noting that these two scoring functions yield very low correlation values (see Table 3). Furthermore, D-Score discriminates poorly active compounds from inactive ones but yields a reasonable correlation factor. This result shows that it can be misleading to assess the performance of a scoring function by calculating only correlation values. In virtual screening, the main issue is to select a number of weak inhibitors out of a database of mainly inactive compounds, and therefore enrichment factors are more relevant than correlation values.

It is instructive to compare the ligands that are ranked in the top 20% scored poses by the various scoring functions studied in this work. The highest overlap is found between QRFF-MD/RMSD-based consensus and consensus scoring, where 83% of the ligands belong to the top 20% of both scoring functions. More interestingly, significant overlaps (>70%) are found between QRFF-pose and QRFF-MD (79%), between QRFF-pose and GROMOS-pose (77%), and between QRFF-MD/RMSD and QRFF-MD (73%). Furthermore, there is only one active that belongs to the top 20% of QRFF-MD and not to the top 20% of consensus scoring (ligand 4). Finally, six actives belong to the top 20% of QRFF-MD/RMSD and not to the top 20% of consensus scoring (ligands 4, 16, 30, 35, 36, and 37).

Thirty compounds that belong to the top scorers based either on the QRFF-MD/RMSD model or on consensus scoring were selected for biological testing. In the assay, one of the selected compounds, a staurosporine aglycon, turned out to inhibit nearly 50% of the enzymatic activity of CDK2 at a concentration of 10 μm. The other compounds showed an inhibition below 20% at the same concentration.

4. CONCLUSIONS

We presented an assessment of a methodology based on molecular dynamics simulation with a quantum-refined forcefield in its ability to predict the binding affinities of 38 CDK2 inhibitors and to discriminate these ligands from a library of nearly 600 compounds that were considered to be inactive. Our data set of known ligands represents a broad chemical diversity, which makes the calculation of binding affinities particularly challenging. Moreover, the ATP binding site of CDK2 is very flexible, which makes it a difficult target. We compared these results to those obtained by consensus scoring and scoring functions commonly used in docking tools. We have found that the QRFF-MD and QRFF-pose methods yield very similar results, both in terms of correlation of binding scores with experimental affinities and enrichment values. For these two scoring functions, the correlation factor was 0.55 and the percentage of known actives retrieved at 20% of the database screened was around 55%, which corresponds to an 2.8-fold enrichment compared with random screening. As a basis of comparison, the enrichment factor obtained by the Glide scoring function is similar. We have also shown that refining the GROMOS ligand charges on the basis of ab initio calculations improves very significantly the results. This refinement represents a negligible fraction of the computational requirement of the QRFF-MD methodology and, therefore, can be used in drug design projects in combination with consensus scoring. This result does not contradict the fact that the second charge refinement (after 50 ps) does not significantly affect the ligand−protein interaction energies. Using structural information extracted from the simulations, we derived a QRFF-MD-based scoring function, called QRFF-MD/RMSD, that represents a significant improvement with respect to QRFF-MD. QRFF-MD/RMSD performs slightly better than consensus scoring to discriminate the known inhibitors from the inactive compounds.

Although more computationally expensive, the QRFF-MD/RMSD scoring function has a much better physical basis than consensus scoring, and therefore it is expected that the former method is more robust than the latter. Only an application of the QRFF-MD methodology to other targets could prove this assumption. Given the computational requirement of molecular dynamics simulation, more work is needed to make this approach useful in routine drug design projects. Combining the structural information from the simulations with better sampling techniques represents a possible extension of this research. The ever increasing computational power will also make this approach more attractive. Furthermore, our work focused on improving the electrostatics of the ligand, but the hydrophobic effect represents also a main contribution to the free energy of binding. It remains to be seen to which extent the current parametrization of the force-fields accurately describes the van der Waals interaction energies. Finally, the success of this method depends crucially on the quality of the poses obtained by the docking programs, and, therefore, any improvement in the reliability of the docking algorithms and scoring functions should also lead to an improvement in the performance of the QRFF-MD methodology. Very recently a docking algorithm, which replaces the fixed charges of the ligands by QM/MM (Quantum Mechanical/Molecular Mechanical) calculations in the protein environment, treating only the ligands as the quantum region, has been reported.[45] It was shown that this algorithm correctly docks ligands in cases where a fixed charge force-field fails, which suggests a possible way to develop more accurate docking methods.

An analysis of the linear interaction energy and the extended linear response models have shown that these models were not suitable to explain the variance in the binding affinities. Therefore, these models may be limited to compute the binding energies of a series of congeneric ligands. Finally, the evaluation of the scoring functions implemented in CScore has shown large variations in performance. Furthermore, these scoring functions did not always perform consistently for predicting binding affinities (R^2 values) and for discriminating active compounds from inactive ones. In the literature, a scoring function is often validated by computing correlation factors. Our result stresses the importance of calculating enrichment curves, which is also more relevant from the point of view of virtual screening applications.

ACKNOWLEDGMENT

We thank Drs. Armin Widmer, Nicole Martin, Lionel Muller, Peter Fürst, and Doriano Fabbro for various support.

REFERENCES AND NOTES

(1) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.

(2) Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. Computational drug design accommodating receptor flexibility: The relaxed complex scheme. *J. Am. Chem. Soc.* **2002**, *124*, 5632−5633.

(3) Meagher, K. L.; Carlson, H. A. Incorporating protein flexibility in structure-based drug discovery: Using HIV-1 protease as a test case. *J. Am. Chem. Soc.* **2004**, *126*, 13276−13281.

(4) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47*, 45−55.

(5) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.* **2004**, *337*, 1161−1182.

(6) Price, D. J.; Jorgensen, W. L. Improved convergence of binding affinities with free energy perturbation: Application to nonpeptide ligands with pp60(src) SH2 domain. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 681−695.

(7) Oostenbrink, C.; van Gunsteren, W. F. Free energies of ligand binding for structurally diverse compounds. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6750−6754.

(8) Åqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. Ligand binding affinities from MD simulations. *Acc. Chem. Res.* **2002**, *35*, 358−365.

(9) Pierce, A. C.; Jorgensen, W. L. Estimation of binding affinities for selective thrombin inhibitors via Monte Carlo simulations. *J. Med. Chem.* **2001**, *44*, 1043−1050.

(10) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889−897.

(11) Hou, T. J.; Zhang, W.; Xu, X. J. Binding affinities for a series of selective inhibitors of gelatinase-A using molecular dynamics with a linear interaction energy approach. *J. Phys. Chem. B* **2001**, *105*, 5304−5315.

(12) Tounge, B. A.; Reynolds, C. H. Calculation of the binding affinity of beta-secretase inhibitors using the linear interaction energy method. *J. Med. Chem.* **2003**, *46*, 2074−2082.

(13) Almlof, M.; Brandsdal, B. O.; Åqvist, J. Binding affinity prediction with different force fields: Examination of the linear interaction energy method. *J. Comput. Chem.* **2004**, *25*, 1242−1254.

(14) Zhou, R. H.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L.; Levy, M. New linear interaction method for binding affinity calculations using a continuum solvent model. *J. Phys. Chem. B* **2001**, *105*, 10388−10397.

(15) Rizzo, R. C.; Udier-Blagovic, M.; Wang, D. P.; Watkins, E. K.; Smith, M. B. K.; Smith, R. H.; Tirado-Rives, J.; Jorgensen, W. L. Prediction of activity for nonnucleoside inhibitors with HIV-1 reverse transcriptase based on Monte Carlo simulations. *J. Med. Chem.* **2002**, *45*, 2970−2987.

(16) Tominaga, Y. J.; Jorgensen, W. L. General model for estimation of the inhibition of protein kinases using Monte Carlo simulations. *J. Med. Chem.* **2004**, *47*, 2534−2549.

(17) Wang, J. M.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc*. **2001**, *123*, 5221−5230.

(18) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.* **2005**, *48*, 4040−4048.

(19) Curioni, A.; Mordasini, T.; Andreoni, W. Enhancing the accuracy of docking: Molecular dynamics with quantum-refined force fields. *J. Comput.-Aided Mol. Des.* In press.

(20) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein−ligand docking: Avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.

(21) Muegge, I.; Enyedy, I. J. Virtual screening for kinase targets. *Curr. Med. Chem.* **2004**, *11*, 693−707.

(22) Chuaqui, C.; Deng, Z.; Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **2005**, *48*, 121−133.

(23) Sims, P. A.; Wong, C. F.; McCammon, J. A. A computational model of binding thermodynamics: The design of cyclin-dependent kinase 2 inhibitors. *J. Med. Chem.* **2003**, *46*, 3314−3325.

(24) 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. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(25) CPMD Copyright IBM Corp. 1990−2005, Copyright MPI für Festkörperforschung Stuttgart 1997−2001.

(26) Schüttelkopf, A. W.; van Aalten, D. M. F. PRODRG: A tool for high-throughput crystallography of protein−ligand complexes. *Acta Crystallogr.* **2004**, *D60*, 1355−1363.

(27) MDL Information Systems Inc., San Leandro, CA.

(28) Sadowski, J.; Rudolph C.; Gasteiger, J. Automatic Generation of 3D Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.

(29) Berman, H. M.; Westbrook, J.; Feng Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(30) Schrödinger Inc., Portland, OR.

(31) Tripos Inc., St. Louis, MO.

(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) Mügge, 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.

(34) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Mol. Biol.* **1995**, *245*, 43−53.

(35) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions. 1. 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.

(36) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(37) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; Morgan Kaufmann Publishers: New York, 1999.

(38) van Gunsteren, W. F. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Vdf Hochschulverlag AG an der ETH Zürich: Zürich, Switzerland, 1996.

(39) Berendsen, H. J. C. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; pp 331−342.

(40) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A. A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(41) Schöpfer, J.; Fretz, H.; Chaudhuri, B.; Muller, L.; Seeber, E.; Meijer, L.; Lozach, O.; Vangrevelinghe, E.; Furet, P. Structure-based design and synthesis of 2-benzylidene-benzofuran-3-ones as flavopiridol mimics. *J. Med. Chem.* **2002**, *45*, 1741−1747.

(42) Ikuta, M.; Kamata, K.; Fukasawa, K.; Honma, T.; Machida, T.; Hirai, H.; Suzuki-Takahashi, I.; Hayama, T.; Nishimura, S. Crystallographic approach to identification of cyclin-dependent kinase 4 (CDK4)-specific inhibitors by using CDK4 mimic CDK2 protein. *J. Biol. Chem.* **2001**, *276*, 27548−27554.

(43) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute binding free energies: A quantitative approach for their calculation. *J. Phys. Chem.* **2003**, *B 107*, 9535−9551.

(44) Bartels, C.; Widmer, A.; Ehrhardt, C. Absolute free energies of binding of peptide analogues to the HIV-1 protease from molecular dynamics simulations. *J. Comput. Chem.* **2005**, *26*, 1294−1305.

(45) Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comput. Chem.* **2005**, *26*, 915−931.

CI050289+