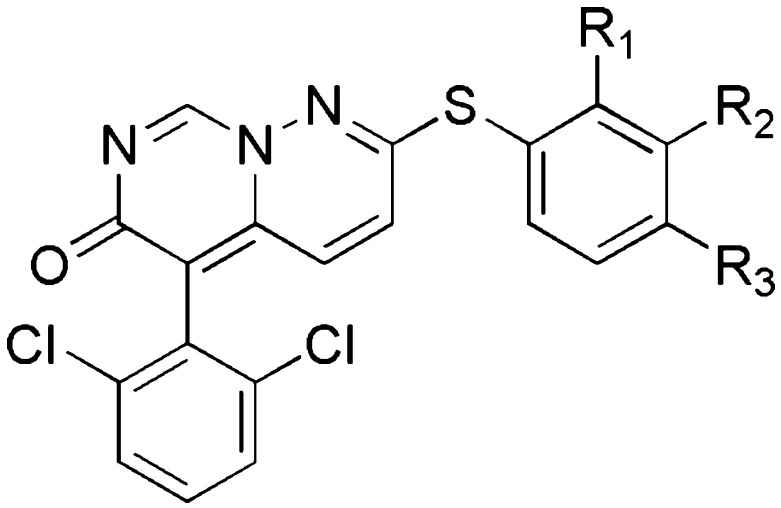
Journal of Medicinal Chemistry

Article

Evaluating the Molecular Mechanics Poisson–Boltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to p38 MAP Kinase

David A. Pearlman

J. Med. Chem., **2005**, 48 (24), 7796-7807• DOI: 10.1021/jm050306m • Publication Date (Web): 05 November 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 13 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Journal of Medicinal Chemistry is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Evaluating the Molecular Mechanics Poisson−Boltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to p38 MAP Kinase

David A. Pearlman†

150 Jason Street, Arlington, Massachusetts 02476

Received April 5, 2005

The recently described molecular mechanics Poisson−Boltzmann surface area (MM-PBSA) method for calculating free energies is applied to a congeneric series of 16 ligands to p38 MAP kinase whose binding constants span approximately 2 orders of magnitude. These compounds have previously been used to test and compare other free energy calculation methods, including thermodynamic integration (TI), OWFEG, ChemScore, PLPScore, and Dock Energy Score. We find that the MM-PBSA performs relatively poorly for this set of ligands, yielding results much inferior to those from TI or OWFEG, inferior to Dock Energy Score, and not appreciably better than ChemScore or PLPScore but at an appreciably larger computational cost than any of these other methods. This suggests that one should be selective in applying the MM-PBSA method and that for systems that are amenable to other free energy approaches, these other approaches may be preferred. We also examine the single simulation approximation for MM-PBSA, whereby the required ligand and protein trajectories are extracted from a single MD simulation rather than two separate MD runs. This assumption, sometimes used to speed the MM-PBSA calculation, is found to yield significantly inferior results with only a moderate net percentage reduction in total simulation time.

Introduction

In the early 1980s, the first of what would prove to be a small torrent of papers devoted to macromolecular free energy calculations appeared in the literature.[1−3] Early publications described the approach, which had been known for many years[4] but which had just been made practicable by galloping advances in computer speed and affordability. Within a year or two, publications appeared that seemed to indicate that these so-called free energy perturbation (FEP) or thermodynamic integration (TI) calculations could easily and reliably be used to determine the free energy difference between molecules or molecular states.[5−7] A handful of very high profile papers demonstrated amazingly good agreement between theoretical prediction and experiment and moved interest in these methods beyond the confines of the modeling community and into the scientific mainstream. Modern macromolecular computational chemistry (spearheaded less than a decade earlier with the publication of the first molecular dynamics simulations) had, it seemed, come of age.

Alas, reality intervened. When free energy methods were taken up by laboratories that had not traditionally concerned themselves with modeling endeavors, they discovered that they were unable to reproduce the promised close agreement between theory and experiment when the methods were applied to their favorite systems. Subsequently, more careful applications of the free energy approaches were carried out in various groups, and it was quickly discovered that much of the agreement claimed in early landmark mainstream papers was probably fortuitous.[8−15] The reality of the situation was that these simulations were much more difficult than originally thought to carry out and that appreciably more computer simulation time was required to achieve reliable results. In fact, the amount of simulation required to reliably determine a free energy value for even a simple system was more than was generally available until the 1990s, when the needs of the simulations and the amount of computing power available to the average laboratory finally started to converge. Now, nearly 2 decades hence, we are finally at a point where free energy simulations can be reliably carried out, though still only for carefully selected systems and with a significant investment of computer resources. The democratic "free energy calculation in every lab's pot" promised by the early conscience raising papers in the field never materialized and most likely never will.

If the initial oversell of free energy calculations was not enough to spoil mainstream interest in these approaches, surely the changes in other facets of drug design did not help. Twenty years ago (at the dawn of the free energy era) drug design consisted of identifying a lead from a modest screening library, then optimizing that lead at the chemistry bench. Today, virtual and combinatorial screening libraries that are hundreds or thousands of times bigger are available and bench optimization is performed in conjunction with high-throughput combinatorial synthesis about a targeted scaffold.[16−18] The keywords for today's approaches are big and fast: Large screening libraries are screened quickly, which speeds the identification of leads, which are then often subjected to faster optimization through combinatorial synthesis. Traditional FEP and TI (inherently slow even with very fast computers) are simply not complementary with these types of approaches. Meticulous bench chemistry is still required near the

† Phone: 617-899-6634. Fax: 270-712-7396. E-mail: Science@arlingtonmass.com.

10.1021/jm050306m CCC: $30.25
Published on Web 11/05/2005

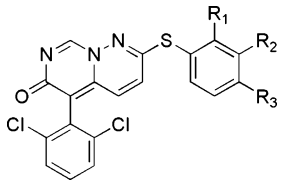


Figure 1. Scaffold class for all p38 variants examined. All ligands in the series are defined by differences at positions R_1, R_2, and R_3, as given in Table 1.

end of the process, and this is where TI or FEP can be useful, but these approaches have been somewhat marginalized in the high-throughput world of commercial drug design.

This has led to the search for faster and more generally applicable methods to calculate free energies. Recently, several approaches have caught the fancy of the modeling world.[19−27] Perhaps most intriguing of these is the molecular mechanics Poisson−Boltzmann surface area (MM-PBSA) method,[24−26] wherein the absolute free energy of a system is *estimated* from a combination of molecular mechanics energy, a Poisson−Boltzmann estimate of the electrostatic free energy,[28] an estimate of the solvation free energy determined from the exposed surface area, and an estimate of the entropy of the molecule derived from a normal modes calculation. In principle, this approach is much more widely applicable than FEP or TI because it can be applied to any species (FEP and TI, which calculate free energy *differences* between two molecules, are practically limited molecules that are relatively similar[29]). Depending on how the calculations are performed, it may also be faster to apply. It also shares something else with FEP and TI: Initial publications have indicated that the approach works surprisingly well on test systems.[24,30−36] However, there is as yet very little published data directly comparing the predictions of MM-PBSA with those of other, more established methods on the same system. Therefore, in this paper we attempt to apply the MM-PBSA method to a test set of data that we have previously used to test a variety of mostly better-known modeling approaches.[37] Methods already tested against this set are among the best in current use:[38] TI,[29] one window free energy grid (OWFEG),[23] ChemScore,[39,40] piecewise linear potential (PLP),[41] and Dock Energy Score.[42,43] The test set consists of 16 congeneric ligands that bind to the active site of p38 MAP kinase protein,[44] with an experimental range of roughly 3 orders of magnitude in IC_{50}. In simple number of compounds and in terms of demonstrated difficulty for other approximate methods, this is a more rigorous test bed than often appears in model simulations.[45]

Methods

The p38 MAP kinase test bed consists of the 330-residue protein plus a set of 16 congeneric ligands differing in the attachments at positions R_1, R_2, and R_3 (Figure 1 and Table 1). The small groups R_1, R_2, and R_3 are variously H, Cl, F, Br, CH_3, OH, or NH_2 and are listed in Table 1. The experimental IC_{50} values for these ligands range from 36 to 1900 nM, roughly 2 orders of magnitude. The set of ligands was chosen both to be representative of the types of changes one considers during the process of drug design (once a scaffold has been chosen) and to span a significant range in IC_{50}. In addition, the IC_{50} values were measured in a consistent fashion[37] and the relative binding efficacies are not intuitively predictable. Finally, and perhaps most critically, *scoring ligands against the p38 MAP kinase protein has* proven to be a difficult task in earlier work, clearly differentiating among the various methods applied.[37,45]

Table 1. Ligands for Which Binding Energies Were Determined[a]

sequence no.	R_1	R_2	R_3	pIC_{50}
1	H	H	H	6.602
2	H	H	F	7.000
3	H	H	CH_3	5.854
4	H	Cl	Cl	6.097
5	H	CH_3	H	5.854
6	H	CH_3	CH_3	5.721
7	H	F	H	6.347
8	CH_3	H	H	6.699
9	H	Cl	F	6.301
10	H	Cl	H	6.553
11	CH_3	H	Cl	6.745
12	Br	H	H	6.602
13	CH_3	H	CH_3	6.577
14	OH	H	H	6.444
15	NH_2	H	F	6.658
16	Cl	H	F	7.444

[a] Refer to Figure 1 for the definitions of the R_1, R_2, and R_3 substitutents. Data are from ref 37.

The importance of this last characteristic of the training set cannot be overemphasized. We have observed that data sets used to characterize scoring functions frequently fall into two categories: those that most scoring functions do well against and those that most scoring functions fail against.[46] Good performance against a test data set does not necessarily validate a scoring function. This is a necessary but not sufficient criterion. One must also demonstrate that the data set is selective for better scoring functions or, conversely, that there are reasonable scoring functions that fail (or at least do appreciably more poorly) for the same data set. Unfortunately, many new scoring methods are only validated against data sets of unknown selectivity. This includes some of the promising MM-PBSA publications. The 16 p38 MAP kinase ligand set we employ in this paper has been validated in previous work to be selective.

The free energy of binding for a protein P and ligand L can be calculated as

$$\Delta G_{\text{bind}} = \langle G_{\text{P.L}} \rangle - \langle G_{\text{P}} \rangle - \langle G_{\text{L}} \rangle \quad (1)$$

where $\langle G_{\text{P.L}} \rangle$, $\langle G_{\text{P}} \rangle$, and $\langle G_{\text{L}} \rangle$ are the free energies of the protein + ligand complex, the protein, and the ligand, respectively, averaged over a set of snapshots taken to represent the ensemble of available states. This set of snapshots is generated using molecular dynamics (MD).

With the MM-PBSA approach, the free energy for any single snapshot structure, G_{mol}, is given as[24]

$$G_{\text{mol}} = V_{\text{MM}} + G_{\text{solv}} - TS_{\text{solute}} \quad (2)$$

V_{MM} is the total molecular mechanical energy in the gas phase. G_{solv} is the solvation free energy, itself the sum of electrostatic and nonpolar contributions:

$$G_{\text{solv}} = G_{\text{elec}} + G_{\text{nonpolar}} \quad (3)$$

S_{solute} is the entropy of the solute. The essence of the MM-PBSA method lies in how the various contributions are estimated. V_{MM} is calculated for the unsolvated molecule using the standard molecular mechanics force field[47]

$$V_{amber} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}(1 + \cos(n\phi - \gamma)) + \sum_{i<j}\left\{\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right] + \frac{q_i q_j}{\epsilon R_{ij}}\right\} \quad (4)$$

without applying any cutoff to nonbonded interactions and subject to the same moving atom restraints ("belly") as the MD simulation used to generate the snapshot. Here, we use the Sander module of the Amber program[48] along with the force field parameters described below to perform the V_{MM}. The electrostatic contribution to the free energy is calculated using a finite difference solution to the Poisson−Boltzmann (PB) equation.[49,50] Delphi II[51] is used to determine the PB energy. Dielectric constants of 1.0 and 80.0 are used for the interior and exterior of the molecule, respectively. A cubic lattice is used, and the largest dimension of the cubic lattice is 80% filled by the longest dimension of the molecule. The Poisson−Boltzmann equation was iteratively solved using 1000 linear steps of finite difference, with a Coulombic boundary condition. The dielectric boundary is defined using a 1.4 Å probe water on the atomic surface. The Parse[52] set of radii were used for atoms of the molecule plus radii of 1.55, 1.748, and 2.02 Å, respectively, for F, Cl, and Br, which are the radii for these ions in the Amber Parm99 force field reduced by 0.2 Å (the approximate difference between radii in Parm99 and Parse radii). For consistency, the atomic charges used are the same as those used in our previous studies of this test system, which are Cornell et al. charges[47] for the protein and ESP-fitted ab initio charges (3-21g*/6-31g* MK charges[53]) for the inhibitors.

The nonbonded contribution is estimated from the solvent accessible surface area (SA), using the algorithm of Sanner[54] and the relationship

$$G_{nonpolar} = \gamma\,\mathrm{SA} + b \quad (5)$$

with γ taken to be 0.004 52 kcal/Å^2 and b taken to be 0.92 kcal/mol,[52] as is standard in the MM-PBSA work that has been published. Finally, the solute entropy S for each species is estimated using a normal mode calculation[55] for a snapshot of the structure minimized in a vacuum with a distance-dependent dielectric of $\epsilon = 4r$ to a root-mean-square gradient of less than 10^{-4} kcal mol^{-1} Å^{-1}. The nmode module of the Amber package is used to perform this part of the calculation.

As per eq 1, we need to determine the free energies for all members of a set of states that represent the ensemble of configurations available to the molecule. These values are then averaged to give an effective free energy. To generate these states, MD was run using the Sander module of the Amber 5 program.[48] All force field parameters used to run MD were identical to those used in our previous thermodynamic integration free energy work:[37] protein and water nonbonded parameters are from the Cornell et al. force field,[47] as are nonbonded parameters for the inhibitors; ESP-fitted ab initio charges are used for the inhibitors (see above); and internal parameters for the inhibitors are assigned according to the Charmm Parm22 force field.[56]

For all MD simulations, a 2 fs time step was used and SHAKE restraints[57] were applied to all bonds not including a hydrogen. A 16 Å cap of TIP3P[58] water was added to the system centered on the ligand, and during simulations this cap was restrained with a 1 kcal/mol half-harmonic restraint at the outer boundary. A moving belly consisting of all atoms within 12 Å of the inhibitor was defined. During the simulations, all waters, all atoms of the ligand, and any atoms of the protein within the belly were free to move. These conditions were again chosen to be identical to those used in our previous free energy studies on this system.[45] It was important to keep this part of the protocol as similar as possible to that of the previous study because the purpose of this study was to determine how well the current method performs compared to the method applied earlier.

In total, four different simulations were performed for each of the 16 protein + ligand complexes. In the first set, 200 ps of equilibration were followed by 1 ns of sampling, and no position restraints were placed on the moving protein or ligand atoms. In the second set, 200 ps of equilibration were followed by 1 ns of sampling and 0.5 kcal mol^{-1} Å^{-1} harmonic position restraints were placed on the positions of the moving protein atoms. In the third set, the ending coordinates of the first simulation were equilibrated for another 200 ps (making the effective equilibration 1.4 ns) and this was followed by 5 ns of sampling with no position restraints on the moving protein. In the fourth set, the ending coordinates of the second simulation were equilibrated for another 200 ps (effective equilibrium 1.4 ns) and this was followed by 5 ns of sampling with 0.5 kcal mol^{-1} Å^{-1} harmonic restraints on the moving atoms of the protein.

From the resulting trajectories, snapshots were extracted every 10 ps to use in calculating V_{MM} and G_{solv}, giving 100 snapshots for the 1 ns runs and 500 snapshots for the 5 ns runs. Solvent water was stripped off of these snapshots before they were used. It was not possible to use snapshots of the entire system when performing the normal modes analysis required to determine S_{solute} or $S_{complex}$ owing to the impossibly large memory requirements of such systems when performing the analysis. For this reason, an 8 Å sphere of atoms centered on the ligand was extracted for each snapshot and these were used when performing the normal modes analysis. In addition, owing to the substantially larger CPU requirements of the normal modes analysis (as well as the intrinsically very approximate nature of this term in the calculation), snapshots for normal modes analysis were extracted every 100 ps, yielding 10 snapshots for the 1 ns runs and 50 snapshots for the 5 ns runs.

In addition to the simulations already described for the protein + ligand (P + L) complexes, separate analogous simulations were run for the ligands (L) alone and for the unbound protein (P). The simulations of the

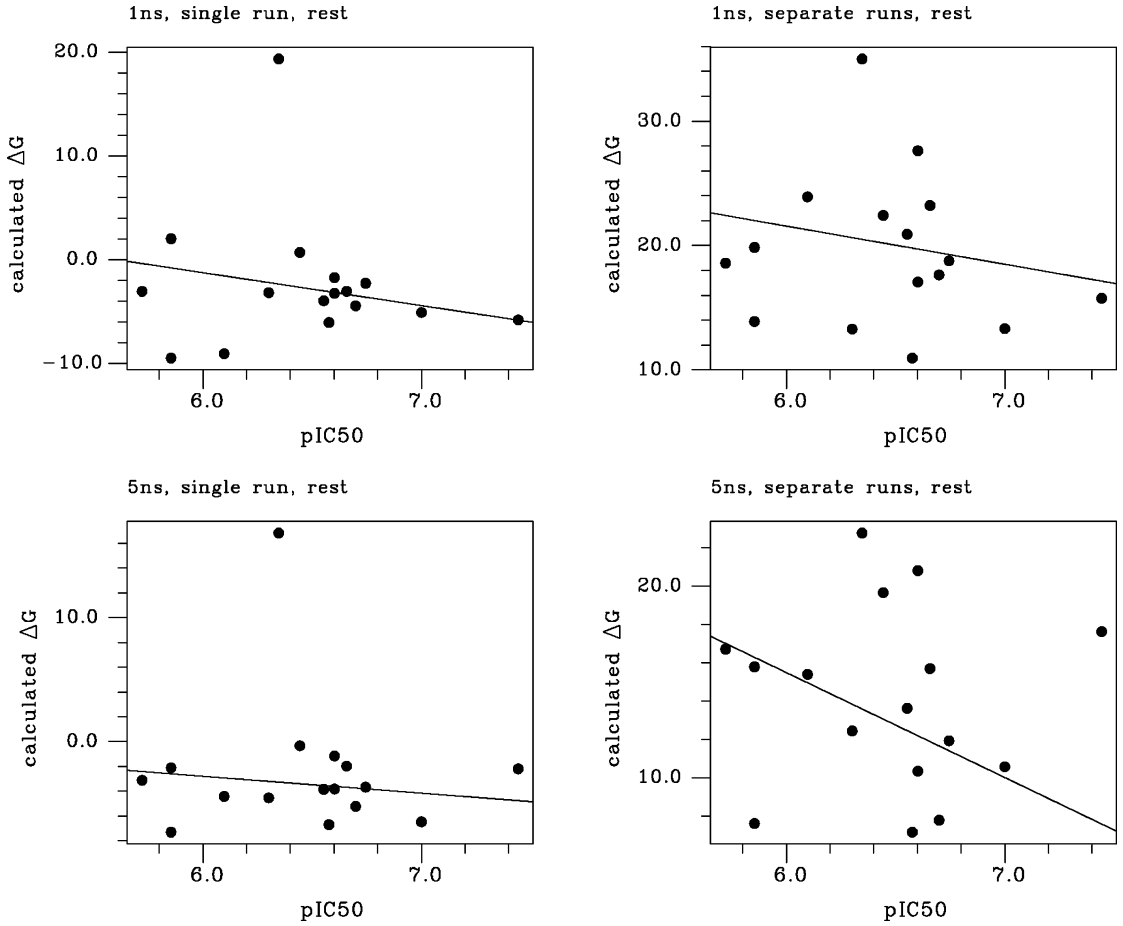


Figure 2. Free energy of binding ΔG for each p38 inhibitor as calculated using MM-PBSA versus the experimentally measured pIC_{50}. Results for the four protocols with a 0.5 kcal/mol positional restraint on the moving atoms of the protein are shown: (upper left) 1 ns sampling, snapshots for protein, protein + ligand, and ligand derived from a single MD run ("single run"); (upper right) 1 ns sampling, snapshots for protein, protein + ligand, and ligand derived from separate MD runs ("separate runs"); (lower left) 5 ns sampling, single run; (lower right) 5 ns sampling, separate runs. The line drawn in each plot is a least-squares fit to those points that suggest, on an empirical qualitative basis, a linear correlation. This line is shown for visual emphasis.

ligand alone were run in a periodic box of TIP3P water approximately 30 Å on a side at a constant pressure of 1 atm. The simulation conditions for the protein by itself were identical to those for the protein + ligand complex. These simulations allow us to compare two protocols for calculating the MM-PBSA energies: (1) all snapshots for P + L, P, and L are taken from a single P + L MD simulation (hereafter "single run"); (2) snapshots for the P + L are taken from the P + L MD simulation; those for the protein alone are taken from a simulation of the protein alone, and those for the ligand alone are taken from a simulation of the ligand by itself (hereafter "separate run"). The latter protocol is more rigorous, but the former protocol has been used in some earlier work[30,32,34] and is appealing because it requires fewer simulations.

The calculations required to evaluate the MM-PBSA energy are not cheap. On an Intel Pentium 4 2.53 GHz PC, the MM-PBSA calculation, exclusive of the normal modes contribution, requires approximately 5.5 min of CPU time for each snapshot (P + L, P, and L). The normal modes evaluation required for each snapshot takes approximately 22 min. Total CPU time required to calculate each value of ΔG is determined by multiplying these snapshot requirements by the numbers of snapshots stored and evaluated, which is 100−500 for the non-normal-modes part of the calculation and 10−50 for the normal modes contribution. Thus, the total CPU requirement per ΔG calculated was roughly 13 h for each of the 1 ns runs and 65 h for each of the 5 ns runs. These total times are for each ligand, with the total times for all 16 ligands being 208 and 1040 h, respectively (and roughly 4 times these net amounts for the four different protocols tested herein). These calculation times do not include the CPU time required to generate the MD ensembles, which ranged from 160 to 234 CPU hours for each of the protocols run for 1 ns and from 800 to 1170 CPU hours for each of the protocols run for 5 ns. Thus, the total CPU time required per protocol ranged from 368 to 442 CPU hours (1 ns runs) and 1840−2210 CPU hours (5 ns runs).

Results

The results from the four restrained simulations are presented in Figure 2. Clockwise from upper left, these are the 1 ns "single run" simulation, the 1 ns "separate run" simulation, the 5 ns "separate run" simulation, and the 5 ns "single run" simulation. In each case, ΔG calculated using the MM-PBSA approach is plotted on the y axis against the experimental value of pIC_{50}. As can be seen, the results are quite poor for the runs using only 1 ns of sampling. When 5 ns of sampling is employed, the results get better, with appreciably better

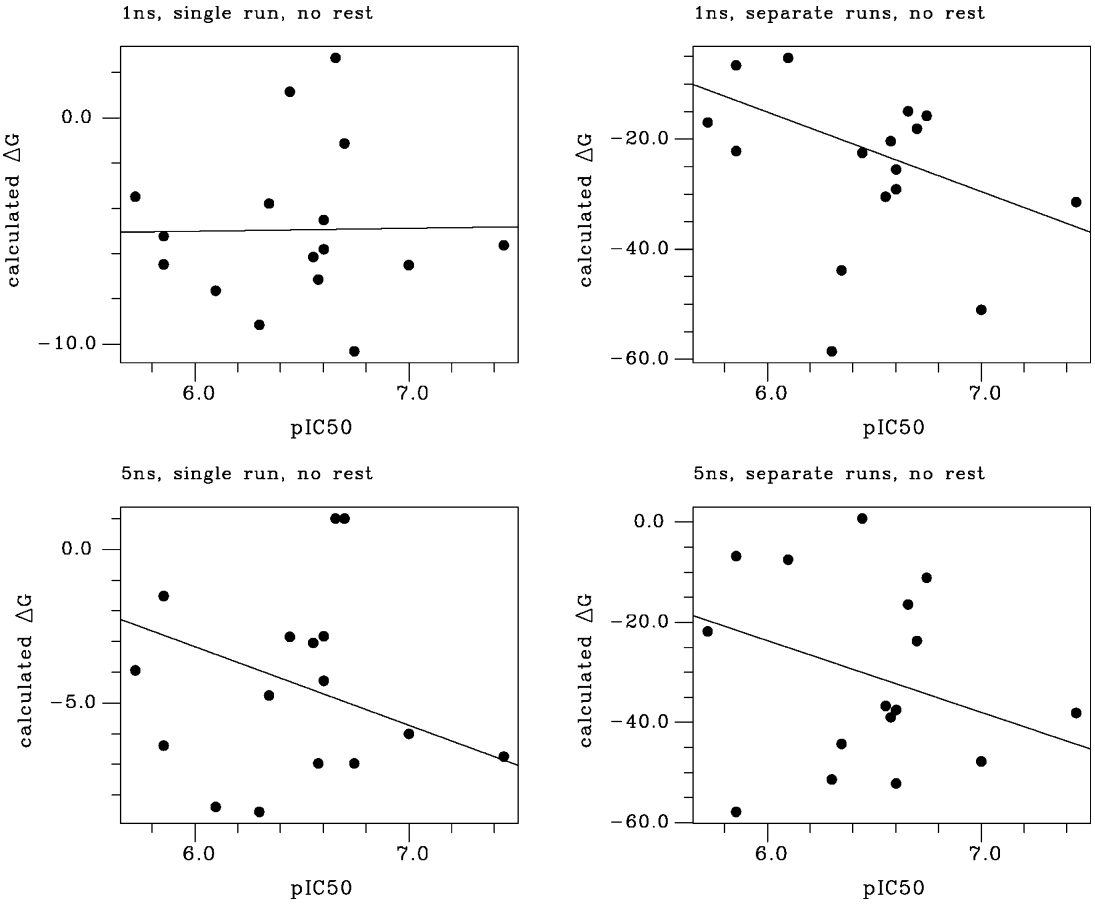


Figure 3. Free energy of binding ΔG for each p38 inhibitor as calculated using MM-PBSA versus the experimentally measured pIC_{50}. Results for the four protocols with no restraints on the moving atoms of the protein are shown: (upper left) 1 ns sampling, single run; (upper right) 1 ns sampling, separate runs; (lower left) 5 ns sampling, single run; (lower right) 5 ns sampling, separate runs. The line drawn in each plot is a least-squares fit to those points that suggest, on an empirical qualitative basis, a linear correlation. This line is shown for visual emphasis.

predictive agreement for the separate run protocol. (A least-squares best fit line to the data is plotted in each case. Note that this line is fit only to those points that suggest, on an empirical basis, the best linear correlation; the line is shown for visual emphasis only.)

Results from the analogous set of simulations, in the case of no restraints on the system, are shown in Figure 3. By comparison to Figure 2, one can see that agreement for these simulations is overall relatively poorer. There is one exception: the 1 ns simulation using separate runs appears modestly predictive. However, given the fact that the 5 ns simulation using otherwise the same protocol yields considerably poorer results, it is difficult to place much faith in the 1 ns results. Previously, we observed similar behavior for this system when analyzing the results of TI calculations with no restraints.

As noted, we chose to apply the MM-PBSA method to this p38 test system in part because we had already applied a number of other techniques to this same data set.[37] In Figure 4 we present, for comparison, the results for the p38 ligand series using TI, OWFEG, and the widely used scoring potentials ChemScore, PLPscore, and Dock Energy Score. Chemscore, PLPscore, and Dock Energy Score are all simple and very fast methods. OWFEG is a grid-based scoring method where the value at each point on the grid is determined from the single-window free energy perturbation for growing a probe atom at that site. Generating the grid can be time-consuming, but once the grid is generated, any number of compounds can be quickly scored as quickly as with the other rapid methods. Therefore, in examining the results from these methods, one should consider MM-PBSA and TI to be very time-consuming and the remaining approaches to be fast.

Comparing these results visually to the best results we obtained using MM-PBSA (5 ns, separate run, restraints; bottom right-hand corner of Figure 2), we see that the MM-PBSA results are roughly comparable to those using PLPscore and ChemScore, a bit worse than Dock Energy Score, and surely inferior to those obtained using either TI or OWFEG.

To set the comparison on a more quantitative footing, we applied the predictive index (PI)[37] for the MM-PBSA method. The PI is calculated as follows:

$$\mathrm{PI} = \frac{\sum_{j>i}\sum_{i} w_{ij} C_{ij}}{\sum_{j>i}\sum_{i} w_{ij}} \tag{6}$$

with

$$w_{ij} = |E(j) - E(i)| \tag{7}$$

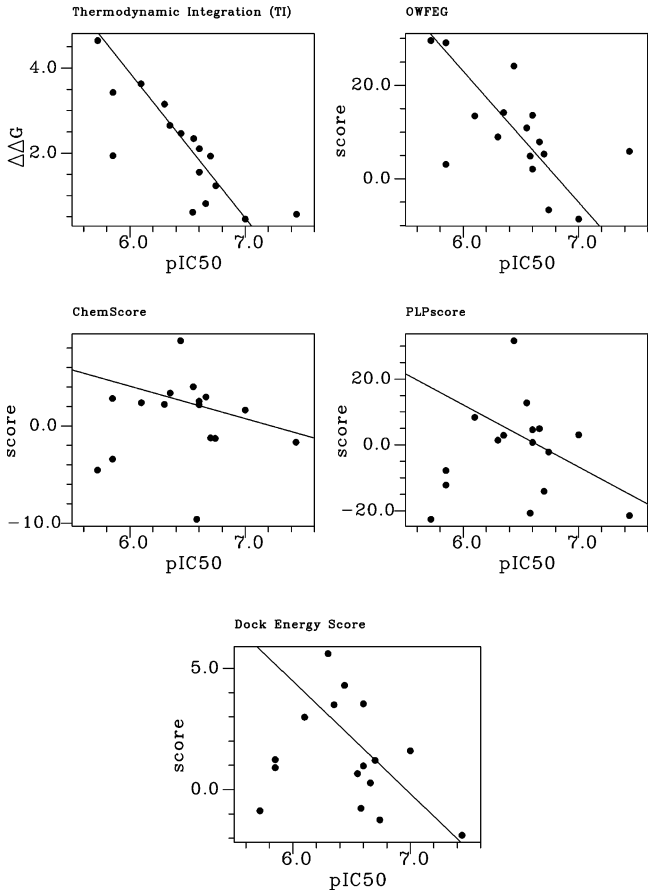


Figure 4. Predicted binding scores for the p38 ligands as calculated using five scoring schemes: TI, OWFEG, ChemScore, PLPScore, and Dock Energy Score. The line in each plot is a fit to those points that most strongly suggest, on an empirical qualitative basis, a linear correlation. This is intended as a visual cue and is not a least-squares fit to all the data.

and

$$C_{ij} = \begin{cases} 1 & \text{if } [E(j) - E(i)]/[P(j) - P(i)] < 0 \\ -1 & \text{if } [E(j) - E(i)]/[P(j) - P(i)] > 0 \\ 0 & \text{if } [P(j) - P(i)] = 0 \end{cases} \quad (8)$$

$P(i)$ is the calculated binding score assigned to ligand i by the model, and $E(i)$ is the corresponding experimental pK_i (or pIC_{50}). A PI of 1.0 would indicate that for any two molecules, the calculation correctly predicts which of those two molecules is, experimentally, lower in energy. A PI of −1.0 would indicate that the model incorrectly predicts the lower energy molecule in every case. A PI of 0.0 would be the expected result from completely random predictions.

The predictive indices for each of the eight MM-PBSA calculations performed are given in Table 2, along with the PI values for the other methods applied to this same data set. As can be seen, the PI for the MM-PBSA method, for the best restrained case, is roughly in line with that for ChemScore, somewhat better than that for PLPscore, somewhat worse than that for Dock Energy Score, and appreciably worse than that for either TI or OWFEG. One point bears mentioning. The PI value for the 5 ns, separate runs, restrained MM-PBSA reflects a particular anomaly with this set: significantly off-line predictions for both the experimentally highest energy (worst) binder (the point at [5.8, 7.6] in the lower left-hand corner of the figure, sequence number 4 in Table 1) and for the experimentally second-lowest energy binder (the point at [7.4, 17.6] in the upper right-hand corner of the figure, sequence number 16 in Table 1). Both of these points will have a very large deleterious affect on the calculated value of PI because they result in the wrong prediction for any pair of points that includes either of them. The PI recalculated removing these two points is a respectable 0.51. It is further worth noting that the compounds represented by these outlier points are both among the four outlier compounds poorly predicted using TI (Figure 4). The poor behavior for these points in both the TI case and the MM-PBSA case implies that this may reflect a deficiency in the force field used for these calculations, given that the same force field was used for both. In a previous study, despite extensive efforts, we were not able to ascertain any procedural/methodological reasons for the anomalous predictions for these compounds using the TI method.[37]

Table 2. Predictive Indices (PI) for Various Scoring Methods

scoring method	PI
thermodynamic integration (TI)	0.85
OWFEG	0.56
ChemScore	0.04
PLPScore	−0.05
Dock Energy Score	0.25
MM-PBSA, 1 ns, single run, restrain protein	0.16
MM-PBSA, 1 ns, separate runs, restrain protein	0.18
MM-PBSA, 5 ns, single run, restrain protein	−0.04
MM-PBSA, 5 ns, separate runs, restrain protein	0.03
MM-PBSA, 1 ns, single run, no restraints	0.04
MM-PBSA, 1 ns, separate runs, no restraints	0.45
MM-PBSA, 5 ns, single run, no restraints	0.06
MM-PBSA, 5 ns, separate runs, no restraints	0.16
MM-PBSA, 5 ns, separate runs, restrain protein. Remove outlier points for ligands 4 and 16 (see text)	0.51

The Predictive Index (PI) is defined in Equations (6)−(8). A predictive index of 1 indicates perfect relative prediction for a series. A predictive index of 0 is essentially random prediction.

For the unrestrained 1 ns simulation with separate runs, we obtain a promising PI of 0.45, which places this protocol well ahead of ChemScore, PLPscore, and Dock Energy Score, though still worse than OWFEG or TI. However, two things temper our enthusiasm for this result. First, the PI for the longer (and presumably more accurate) 5 ns run with the same protocol is appreciably worse (0.163). In addition, looking at the results from this run (Figure 3, upper right) we see that the scores are only very crudely predictive. That is, the worst four binders are scattered randomly in a cluster at the upper left of the plot, the best binder is predicted to be reasonably low in energy (but not as low as three outliers at the bottom of the plot), and the remainder of the points are clustered randomly in the middle of the plot. In essence, the PI reflects an acceptable line to just three effective "points" (two clusters of points at the singleton at the right), with three additional points significantly off line. Combined with the fact that the fit does not tighten up (but instead worsens) with additional sampling, we are inclined to characterize the relatively high PI for this set as fortuitous.

A comparison of the 1 and 5 ns separate run sampling results in Figure 2 demonstrates rather clearly that the

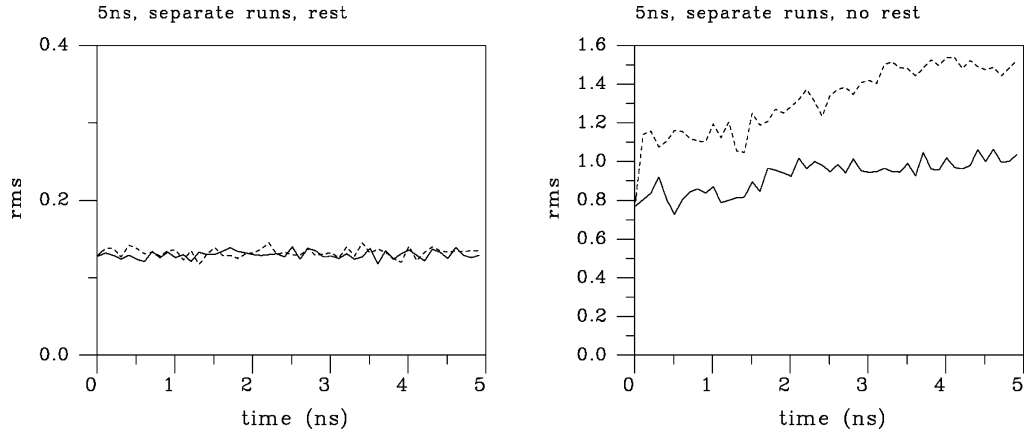


Figure 5. Root-mean-squared (rms) deviations from the initial coordinates for all of the heavy atoms of the protein, or of the protein + ligand complex, as a function of simulation time. Results are shown for the 5 ns simulations using separate simulations for the protein and for the protein + ligand complex. The solid lines represent the complex, and the dashed lines represent the protein: (left) run using restraints on atomic positions; (right) run without restraints on atomic positions. Results are for complex **16** in Table 1 but are similar for all complexes.

results are considerably improved with the increased sampling. It is possible that further improvement could be attained by even more sampling. However, the high CPU cost of these calculations placed extension of these calculations to 10 or 20 ns beyond the scope of this study. It is worth noting that the protein + ligand TI calculations, which performed admirably for this series, used 420 ps of sampling in each "direction" (λ, 0 → 1 and λ, 1 → 0) for a net total of 840 ps of sampling. The free energy values in the two directions are averaged to give the net ΔG. The OWFEG grids, which also scored this set reasonably well, were based on 1 ns of sampling. The quality of the predictions from both TI and OWFEG free energy simulations depends critically on how successfully we have sampled the important conformational substates available to the molecules of interest. Thus, we can infer from the good TI and OWFEG results that the amount of sampling afforded these simulations ($\leq$1 ns) is sufficient to sample these substates. Since the MM-PBSA calculations presented here used the same force field and same simulation conditions (and were applied to the same ligands) as those used in the TI and OWFEG calculations, we expect that the 1 ns simulations used to evaluate the MM-PBSA energies should have been sufficiently long to sample the important conformational substates; missing important substates is not likely the reason for the poor MM-PBSA results.

It may be that inherent errors in calculating the MM-PBSA energy for a single snapshot are so large that greater sampling (and hence more snapshots) is required to average out the errors. The rms deviations from the starting structure for the protein + ligand complex and for the protein itself are shown as a function of time for the 5 ns runs in Figure 5. In these plots, the solid line represents the complex and the dashed line represents the protein by itself. Expectedly, the rms is stable throughout the run for the restrained simulation. For the unrestrained simulation, it appears that while the complex is relatively stable throughout the run (recall that the 5 ns simulation starts with coordinates already equilibrated for 1.4 ns), the unbound protein continues to shift relative to the starting coordinates until about 3 ns of simulation. However, MM-PBSA free energies obtained using just the last 2 ns of the simulation are effectively the same as those obtained using snapshots from the entire 5 ns of simulation (not shown).

Discussion

The intent of this study was to determine if the MM-PBSA protocol should be added to the list of methods that should be considered when attempting to predict the binding behavior for a series of ligands in a binding sight. From the results we have presented, the answer would appear to be "no". The method proved appreciably worse than either TI or OWFEG for the p38 series and worse than Dock Energy Score. Arguably, the method performs better than PLPscore or Dock Energy Score on this series, but that still places it a distant fourth in the rankings. This is poor performance indeed, when one considers the CPU investment MM-PBSA requires. By comparison, the exact free energy TI method requires less than $^1/_6$ to $^1/_7$ the amount of CPU time per ligand to run a "forward/reverse" cycle. OWFEG requires roughly 17 CPU hours total (to generate a single scoring grid, at which point scoring of any number of ligands takes only seconds) or less than $^1/_{100}$ the time of the MM-PBSA method for the 16 ligand series. ChemScore, PLPscore, and Dock Energy Score take about a minute (or less) to perform.

It is not clear if the performance of MM-PBSA would continue to move up in the ranks as the amount of sampling increases. It seems that the results for 5 ns of sampling are, in the restrained case, appreciably better than those for 1 ns of sampling, based on both empirical observation and PI values if one removes the two outlier points. But even if the method continues to improve with greater sampling, it would not be realistic to expect the predictions using this method to approach those of TI. In essence, the method is simply not cost-effective for this system. (It is important to note that the TI, OWFEG, and MM-PBSA simulations used the same force field parameters, which eliminates issues related to these parameters in comparing the efficacies of the various methods.) For the unrestrained protocol, we found that the results using the separate runs actually get worse as the sampling is increased. We think this reflects fortuitously good results at 1 ns and

demonstrates the importance of verifying what may appear to be good results with a single set of simulations. With respect to sampling, it is also instructive to examine the standard deviations in Table 3. Note the very large variations in the calculated V_{MM} potential and PB electrostatic energies for the runs where separate simulations were carried out for the P + L, P, and L components. Provided that the structures in each separate simulation are merely undergoing modest fluctuations about a well-defined mean, one might argue that the variance calculated from the snapshots (which are, effectively, arbitrarily grouped together as sets when the simulations are run separately) might overstate the uncertainty in the results. But nonetheless, this suggests that these simulations might need to be run a very long time to obtain truly converged results. Alternately, doing the MM-PBSA calculation based on a modest number of snapshots may not be advisable. Note that this is a protocol we have copied from all earlier MM-PBSA applications, and it is a protocol used because of the enormous costs associated with increasing the number of snapshots.

It has been suggested that one way to reduce the cost of MM-PBSA simulations is to generate the required complex, protein, and ligand snapshots from a single MD simulation of the complex. The value of this approach is questionable on two counts. First, the MD simulation itself is only a modest, small part of the total cost of the calculation. For the p38 series, one 5 ns MD simulation of the complex took roughly 50 h (or $16 \times 50 = 800$ CPU hours for the series). A separate simulation of the protein took another 50 h, and each simulation of a ligand took roughly 20 h. So the protocol where separate simulations of the three components were performed required an additional $50 + (20)(16) = 370$ CPU hours for the series, for a sum total of 1170 CPU hours. This net difference of 370 CPU hours is relative to a total MD + MM-PBSA CPU time with separate runs of ($1040 + 1170 = 2210$ CPU hours). Thus, the savings is roughly a sixth of the total CPU time required for the series, and while this is not inconsequential, it is only a moderate improvement in a fractional sense. More importantly, results using the separate run protocol are unquestionably better in every case. Given the intrinsic costs of these calculations, an attempt to shave a quarter off the total simulation time to attain appreciably poorer results seems foolish.

Not unexpectedly, with one exception we found rather poorer agreement with experiment in the cases where no restraints were placed on the system. Using the same force field, we found similar results previously using TI. This most likely reflects a combination of deficiencies in the force field and an inability to sample all the conformational substates available to the system when all degrees of freedom are free to move. The fact that in using TI we obtained such good agreement while using the restrained system implies that the removal of these substates is, by and large, an acceptable compromise.

The MM-PBSA method calculates absolute free energies for various states, and so with this method we should be able to calculate the absolute free energy of binding. This might suggest converting the experimental IC_{50} values into (approximate) binding free energies and plotting against these latter values. The reason we did not do this is to maintain consistency with our previous work on this data set. Whether to plot against the approximate experimental ΔG is really a matter of preference, and doing so would not change any of our conclusions or any of the analysis. In addition, it should be noted that the ΔG values we have calculated for this system using MM-PBSA are far from the experimental values on an absolute scale (Table 3). This is irrelevant when using the method as a scoring function, since the question is typically "which ligand binds better". That is the context in which all the scoring functions are being evaluated here. Nonphysical aspects of the protocol (nonmoving belly atoms, position restraints) probably contribute to the generally poor absolute agreement between the MM-PBSA and experimental results but do not (given the good results using TI and OWFEG with the exact same protocol) adversely affect the ability to make good relative predictions.

It is very important to note the limitations of our conclusions. We can say with certainty that the MM-PBSA method is poorly suited to our test system. Does this mean it is a poor choice for any congeneric set of ligands for any protein? No, surely not. As we have noted, some data sets can be reasonably well predicted by many/most reasonable scoring functions and we do not have evidence that MM-PBSA would not work in such a case. However, a scoring function must be predictive, reliable, and cost-effective, and our results raise doubts about whether MM-PBSA could be expected to fulfill those criteria for systems where other trusted methods are available.

This does not mean that the MM-PBSA method has no usefulness. Quite the contrary: Our results using 5 ns and separate runs (Figure 2, lower right) demonstrate that MM-PBSA is able to make passable predictions given enough care and enough computer resources. It fails in the present case where other, more precise methods can easily be applied. But MM-PBSA has a significant advantage over these other approaches in terms of the scope of applicability. That is because MM-PBSA calculates an absolute free energy for each species. TI and OWFEG calculate relative free energies, and the other, essentially empirical methods have significant troubles when comparing systems that are very different from one another. It has been shown that MM-PBSA can, in fact, be applied to different macrosystems that are well beyond the scope of these other approaches,[33] such as the helical preferences of DNA.[24] That is not to minimize the lessons of this paper: Significant amounts of sampling are surely required to get reliable, reproducible results, and one should avoid penny-wise, pound-foolish traps such as using one simulation to generate all the required snapshots.

Finally, it is important to discuss the results we have obtained in light of other, apparently better, results using MM-PBSA to predict binding for a series of ligands that have appeared in the literature.[27,30,32,34] When interpreting the results from a study that attempts to predict series data, one must consider several factors. First, how many compounds are predicted? The greater the number of compounds in the series, the better will be the reliability of the fit (or lack thereof). Second, what range in K_i (IC_{50}) do the compounds span? Preferably, they should evenly span at least 2−3 orders

Table 3. Components of Free Energies[a]

no.	ΔV_{mm}	$\Delta\Delta G_{nonpolar}$	$\Delta\Delta G_{elec}$(PB)	$-T\Delta S$	ΔG_{total}	ΔG_{expt}
			1 ns, Single Run, Restraints			
1	−58.23(3.34)	−3.90(0.06)	37.88(2.72)	21.00(4.05)	−3.26(4.92)	−9.06
2	−59.62(3.32)	−3.95(0.06)	38.42(2.76)	20.03(4.99)	−5.12(5.78)	−9.61
3	−60.78(3.29)	−4.02(0.09)	38.69(2.58)	16.61(4.26)	−9.51(5.15)	−8.03
4	−59.99(3.23)	−4.19(0.09)	39.16(2.96)	15.93(4.49)	−9.09(5.23)	−8.37
5	−59.48(3.15)	−4.13(0.14)	40.56(4.46)	25.04(5.48)	2.00(6.38)	−8.03
6	−59.58(3.01)	−4.16(0.10)	39.29(3.00)	21.37(4.84)	−3.08(5.58)	−7.85
7	−45.25(8.72)	−1.77(0.17)	46.86(9.99)	19.48(9.68)	19.33(10.22)	−8.71
8	−57.37(3.10)	−4.05(0.10)	38.01(3.43)	18.95(5.30)	−4.47(6.08)	−9.20
9	−57.85(3.17)	−4.23(0.12)	38.63(2.85)	20.25(5.62)	−3.20(6.49)	−8.65
10	−60.67(3.60)	−4.13(0.10)	40.91(3.24)	19.91(6.98)	−3.98(7.70)	−9.00
11	−61.01(3.09)	−4.15(0.06)	38.38(2.82)	24.47(3.67)	−2.30(4.47)	−9.26
12	−62.73(3.70)	−4.24(0.09)	45.15(3.25)	20.08(4.42)	−1.74(5.34)	−9.06
13	−61.83(3.58)	−4.19(0.10)	39.22(2.81)	20.73(4.25)	−6.08(5.17)	−9.03
14	−60.37(4.45)	−4.01(0.08)	44.12(4.38)	20.94(4.64)	0.68(5.43)	−8.85
15	−62.51(3.68)	−4.14(0.08)	45.39(3.84)	18.19(7.30)	−3.07(7.97)	−9.14
16	−64.31(3.54)	−4.37(0.14)	43.39(3.29)	19.46(5.06)	−5.84(5.91)	−10.22
			5 ns, Single Run, Restraints			
1	−58.84(3.44)	−3.91(0.07)	38.57(2.63)	20.36(4.66)	−3.82(5.55)	−9.06
2	−59.63(3.13)	−3.97(0.07)	38.25(2.72)	18.88(5.15)	−6.46(5.83)	−9.61
3	−60.68(3.27)	−4.00(0.08)	38.68(2.44)	18.71(4.42)	−7.29(5.21)	−8.03
4	−60.39(3.45)	−4.17(0.09)	39.33(2.97)	20.82(4.96)	−4.41(5.78)	−8.37
5	−59.47(3.77)	−4.09(0.14)	40.47(4.31)	20.97(4.88)	−2.12(5.88)	−8.03
6	−59.61(3.25)	−4.16(0.10)	39.17(2.77)	21.48(4.40)	−3.11(5.34)	−7.85
7	−47.25(1.93)	−1.81(0.09)	49.69(4.02)	16.19(7.34)	16.81(8.15)	−8.71
8	−57.57(3.29)	−4.05(0.10)	37.84(3.12)	18.56(6.02)	−5.22(6.76)	−9.20
9	−58.22(3.50)	−4.21(0.12)	38.10(2.71)	19.79(4.46)	−4.53(5.38)	−8.65
10	−60.35(3.59)	−4.13(0.09)	40.91(3.36)	19.73(3.91)	−3.84(5.08)	−9.00
11	−60.56(3.15)	−4.15(0.07)	38.33(2.79)	22.70(6.22)	−3.67(6.76)	−9.26
12	−61.57(3.84)	−4.23(0.12)	44.41(3.57)	20.24(4.75)	−1.15(5.71)	−9.06
13	−62.47(3.34)	−4.20(0.10)	39.97(2.60)	20.02(5.79)	−6.69(6.47)	−9.03
14	−60.03(3.91)	−4.00(0.08)	43.62(3.63)	20.08(5.26)	−0.33(6.15)	−8.85
15	−63.00(4.02)	−4.15(0.09)	45.85(4.26)	19.33(6.57)	−1.97(7.25)	−9.14
16	−64.25(3.54)	−4.39(0.11)	43.50(3.39)	22.94(4.98)	−2.20(5.94)	−10.22
			1 ns, Separate Runs, Restraints			
1	−44.41(36.35)	−4.01(0.32)	46.24(22.90)	19.24(4.63)	17.06(34.35)	−9.06
2	−43.34(33.94)	−3.99(0.40)	41.93(21.61)	18.70(5.14)	13.31(30.97)	−9.61
3	−52.50(36.71)	−4.04(0.37)	53.72(22.94)	16.71(5.23)	13.89(32.77)	−8.03
4	−29.51(35.51)	−4.18(0.36)	39.93(20.89)	17.67(5.07)	23.91(32.19)	−8.37
5	−45.06(33.33)	−4.22(0.32)	47.61(22.23)	21.51(3.71)	19.84(32.81)	−8.03
6	−42.51(32.26)	−4.18(0.34)	46.34(22.60)	18.92(5.54)	18.58(28.72)	−7.85
7	−39.23(27.47)	−1.80(0.35)	57.09(24.42)	18.94(8.08)	35.00(28.37)	−8.71
8	−38.42(38.34)	−4.05(0.37)	41.70(22.07)	18.40(4.23)	17.64(34.69)	−9.20
9	−40.73(36.47)	−4.30(0.36)	39.85(22.07)	18.45(5.20)	13.28(32.43)	−8.65
10	−41.05(31.47)	−4.19(0.34)	47.36(20.72)	18.79(4.57)	20.91(32.32)	−9.00
11	−43.89(33.97)	−4.22(0.38)	44.31(21.39)	22.57(2.86)	18.77(30.48)	−9.26
12	−32.94(38.60)	−4.18(0.35)	45.65(20.58)	19.08(3.17)	27.61(32.34)	−9.06
13	−45.45(37.73)	−4.23(0.36)	41.33(23.69)	19.31(3.77)	10.95(32.14)	−9.03
14	−47.65(34.20)	−4.07(0.30)	55.47(23.37)	18.67(4.06)	22.42(28.01)	−8.85
15	−37.35(33.77)	−4.20(0.39)	47.35(21.69)	17.42(4.01)	23.22(30.02)	−9.14
16	−44.59(37.04)	−4.32(0.38)	45.85(20.86)	18.81(6.02)	15.76(34.24)	−10.22
			5 ns, Separate Runs, Restraints			
1	−49.75(36.06)	−3.95(0.35)	45.37(20.57)	18.66(4.98)	10.33(33.19)	−9.06
2	−44.41(33.63)	−4.00(0.34)	40.94(20.28)	18.03(5.61)	10.56(30.71)	−9.61
3	−58.79(36.32)	−3.98(0.35)	52.44(20.83)	17.93(4.94)	7.60(32.90)	−8.03
4	−36.41(34.01)	−4.12(0.35)	37.48(20.87)	18.44(5.80)	15.39(31.14)	−8.37
5	−42.15(35.83)	−4.07(0.34)	42.10(20.75)	19.90(5.89)	15.78(33.88)	−8.03
6	−38.67(37.06)	−4.10(0.37)	39.05(21.92)	20.42(5.46)	16.70(33.89)	−7.85
7	−44.10(35.04)	−1.74(0.35)	54.40(20.39)	14.20(7.90)	22.77(33.13)	−8.71
8	−45.24(36.48)	−4.03(0.36)	38.70(21.74)	18.34(6.28)	7.77(32.58)	−9.20
9	−33.84(33.88)	−4.19(0.36)	31.72(20.71)	18.73(5.90)	12.43(31.36)	−8.65
10	−42.57(36.45)	−4.12(0.36)	41.63(21.00)	18.68(5.42)	13.62(34.23)	−9.00
11	−46.17(37.02)	−4.12(0.35)	41.13(21.69)	21.08(5.39)	11.92(32.74)	−9.26
12	−36.61(34.46)	−4.10(0.38)	42.67(20.64)	18.85(5.18)	20.81(33.07)	−9.06
13	−48.60(35.01)	−4.21(0.35)	41.04(21.22)	18.92(5.22)	7.15(31.48)	−9.03
14	−43.51(34.68)	−3.97(0.34)	47.57(20.93)	19.57(5.90)	19.66(31.70)	−8.85
15	−41.02(36.77)	−4.12(0.36)	43.53(20.96)	17.31(6.74)	15.69(33.83)	−9.14
16	−40.31(35.14)	−4.28(0.37)	41.87(21.35)	20.35(5.34)	17.62(32.39)	−10.22

Table 3. (Continued)

no.	ΔV_{mm}	$\Delta\Delta G_{nonpolar}$	$\Delta\Delta G_{elec}$(PB)	$-T\Delta S$	ΔG_{total}	ΔG_{expt}
			1 ns, Single Run, No Restraints			
1	−59.30(3.37)	−4.18(0.18)	40.42(4.00)	18.55(5.52)	−4.51(6.41)	−9.06
2	−55.48(4.15)	−4.05(0.15)	35.20(3.66)	17.81(4.18)	−6.51(5.54)	−9.61
3	−60.32(4.54)	−4.12(0.18)	42.75(4.74)	15.21(6.82)	−6.48(8.23)	−8.03
4	−58.30(5.83)	−4.22(0.23)	40.09(5.96)	14.79(3.58)	−7.64(5.32)	−8.37
5	−60.35(4.60)	−4.17(0.13)	39.26(4.34)	20.04(6.86)	−5.23(7.93)	−8.03
6	−61.89(4.06)	−4.23(0.15)	44.12(4.19)	18.52(6.27)	−3.49(7.25)	−7.85
7	−60.33(4.75)	−4.07(0.14)	43.12(5.01)	17.49(5.79)	−3.79(7.10)	−8.71
8	−52.26(5.45)	−4.00(0.16)	36.11(4.11)	19.02(3.58)	−1.13(5.32)	−9.20
9	−54.49(5.20)	−4.23(0.16)	36.49(4.25)	13.08(5.25)	−9.15(6.36)	−8.65
10	−57.11(3.87)	−4.24(0.14)	37.41(4.28)	17.79(6.66)	−6.15(7.75)	−9.00
11	−57.79(3.98)	−4.12(0.16)	35.59(3.88)	16.00(4.72)	−10.32(5.84)	−9.26
12	−59.55(4.78)	−4.23(0.15)	42.26(4.25)	15.70(3.17)	−5.81(4.87)	−9.06
13	−59.73(3.51)	−4.25(0.14)	40.79(3.54)	16.04(7.15)	−7.15(7.89)	−9.03
14	−57.05(5.87)	−4.03(0.15)	41.03(5.71)	21.20(3.99)	1.15(5.75)	−8.85
15	−61.15(6.59)	−4.09(0.20)	45.92(5.95)	21.97(7.48)	2.65(8.34)	−9.14
16	−56.25(6.21)	−4.17(0.16)	38.38(5.39)	16.41(8.26)	−5.63(9.17)	−10.22
			5 ns, Single Run, No Restraints			
1	−57.97(3.83)	−4.11(0.19)	39.50(4.27)	18.29(5.28)	−4.28(6.49)	−9.06
2	−56.89(3.84)	−3.95(0.15)	36.67(3.67)	18.16(5.37)	−6.01(6.40)	−9.61
3	−58.98(5.34)	−3.88(0.23)	43.24(5.04)	18.10(6.08)	−1.52(7.42)	−8.03
4	−58.99(4.84)	−4.16(0.16)	39.49(4.59)	15.27(6.09)	−8.39(7.09)	−8.37
5	−57.71(3.87)	−4.08(0.21)	36.07(4.17)	19.34(5.91)	−6.39(6.99)	−8.03
6	−61.02(5.08)	−4.12(0.24)	42.08(5.04)	19.13(6.37)	−3.94(7.32)	−7.85
7	−61.32(4.06)	−4.13(0.15)	43.18(4.58)	17.51(5.78)	−4.76(7.06)	−8.71
8	−45.85(5.29)	−4.03(0.17)	33.60(4.81)	17.29(5.84)	1.01(7.36)	−9.20
9	−56.25(4.23)	−4.21(0.17)	36.93(3.96)	14.98(5.71)	−8.55(6.83)	−8.65
10	−46.55(7.33)	−4.13(0.16)	31.76(5.75)	15.88(5.90)	−3.04(7.99)	−9.00
11	−60.06(4.23)	−4.23(0.21)	39.88(4.84)	17.44(6.33)	−6.97(7.43)	−9.26
12	−59.35(4.04)	−4.19(0.16)	41.82(4.18)	18.89(6.88)	−2.83(7.90)	−9.06
13	−60.50(4.24)	−4.27(0.15)	40.89(4.01)	16.91(6.36)	−6.97(7.53)	−9.03
14	−68.24(10.90)	−4.11(0.17)	51.37(10.19)	18.13(6.26)	−2.85(7.57)	−8.85
15	−65.32(8.84)	−4.10(0.21)	50.55(8.08)	19.88(6.26)	1.01(7.47)	−9.14
16	−59.60(4.67)	−4.12(0.21)	38.76(4.35)	18.21(6.48)	−6.75(7.50)	−10.22
			1 ns, Separate Runs, No Restraints			
1	−81.10(58.50)	−4.75(0.58)	42.58(64.24)	17.75(7.25)	−25.51(44.10)	−9.06
2	−102.05(60.33)	−5.28(0.79)	38.97(48.58)	17.35(6.84)	−51.01(39.53)	−9.61
3	−55.00(57.99)	−4.51(0.78)	19.92(47.31)	17.42(6.35)	−22.17(39.92)	−8.03
4	−45.52(74.38)	−6.41(7.82)	28.08(64.83)	18.60(7.39)	−5.25(43.50)	−8.37
5	−41.08(55.03)	−4.43(0.80)	16.87(49.05)	22.05(7.43)	−6.59(38.56)	−8.03
6	−96.67(71.31)	−5.43(0.83)	64.69(56.97)	20.43(7.32)	−16.98(39.36)	−7.85
7	−108.81(77.66)	−5.43(0.69)	52.24(59.68)	18.15(11.09)	−43.86(47.03)	−8.71
8	−43.90(68.39)	−4.92(0.77)	11.22(59.83)	19.50(8.75)	−18.10(41.17)	−9.20
9	−100.24(64.83)	−5.77(0.86)	30.66(46.71)	16.79(5.03)	−58.56(42.08)	−8.65
10	−47.40(57.62)	−5.86(0.68)	1.10(45.47)	21.67(9.05)	−30.49(44.85)	−9.00
11	−84.64(67.03)	−5.18(0.67)	55.38(52.72)	18.68(7.04)	−15.77(44.12)	−9.26
12	−66.00(75.39)	−5.70(0.76)	24.75(60.46)	17.81(7.07)	−29.14(42.74)	−9.06
13	−82.80(61.57)	−5.73(0.68)	50.26(51.45)	17.87(7.45)	−20.39(41.18)	−9.03
14	−100.51(52.41)	−4.89(0.71)	62.79(44.68)	20.10(7.14)	−22.52(38.61)	−8.85
15	−62.47(55.26)	−5.09(0.72)	32.30(40.79)	20.33(7.72)	−14.93(41.15)	−9.14
16	−92.77(61.74)	−5.14(0.86)	46.85(38.46)	19.64(7.94)	−31.42(50.94)	−10.22
			5 ns, Separate Runs, Restraints			
1	−139.66(69.29)	−5.80(1.02)	90.18(61.91)	17.77(6.52)	−37.50(39.45)	−9.06
2	−157.17(67.13)	−6.62(1.03)	98.35(57.08)	17.64(6.19)	−47.80(38.24)	−9.61
3	−57.36(59.21)	−6.05(1.16)	35.79(49.69)	20.83(6.02)	−6.80(38.91)	−8.03
4	−76.54(65.36)	−6.39(0.95)	57.08(54.84)	18.36(6.00)	−7.49(38.10)	−8.37
5	−108.93(66.86)	−5.80(1.11)	38.04(57.51)	18.83(6.13)	−57.86(39.71)	−8.03
6	−139.12(65.79)	−6.81(1.04)	103.34(54.51)	20.76(6.02)	−21.83(37.57)	−7.85
7	−163.73(70.20)	−6.86(1.08)	108.14(60.19)	18.16(6.97)	−44.29(37.52)	−8.71
8	−65.69(70.30)	−6.32(1.05)	29.00(62.51)	19.28(5.88)	−23.73(39.74)	−9.20
9	−123.75(65.06)	−6.87(0.92)	60.98(51.77)	18.26(6.10)	−51.39(40.15)	−8.65
10	−60.41(63.67)	−6.44(0.78)	10.17(52.79)	19.97(6.36)	−36.72(37.79)	−9.00
11	−128.67(68.17)	−6.14(0.90)	103.98(59.27)	19.72(6.62)	−11.11(39.26)	−9.26
12	−117.48(67.22)	−6.97(0.97)	53.94(53.72)	18.33(5.84)	−52.18(38.25)	−9.06
13	−112.08(70.90)	−6.83(0.90)	61.19(58.19)	18.73(7.34)	−38.98(39.94)	−9.03
14	−146.82(73.31)	−4.95(0.85)	134.95(71.37)	17.54(5.85)	0.72(39.42)	−8.85
15	−110.36(72.02)	−6.12(0.99)	79.61(61.84)	20.41(6.69)	−16.46(37.56)	−9.14
16	−123.71(60.19)	−6.84(1.13)	74.13(49.44)	18.32(5.69)	−38.11(39.45)	−10.22

[a] All energies in kcal/mol. Energy values in parentheses are the associated root-mean-squared deviations.

of magnitude. If they do not (if, for example, you have a cluster of points around one value and another point far from that point), it is possible to obtain an apparently good measure of correlation even with an unpredictive function. Third, are the experimental data (K_i, IC_{50}) reliable? Have they been measured in a consistent fashion? If not (if, for example, they are extracted from a variety of publications for several laboratories), one cannot be sure that the variance in the experimental data is so large that it will mask any true predictivity (or lack thereof) of the scoring functions. Fourth, and most critically, does reproducing the ligand data present a challenge? For example, typical ligand binding data sets for some proteins, such as HIV-1 protease, are reasonably well predicted by any acceptable function.[45] There are some protein/binder data sets that are acceptably well predicted by almost any reasonable function, while other data sets are extremely difficult to predict. Unfortunately, it is difficult to ascertain the answers to most of these questions by simply reading a publication. Only the number of compounds and the range they span can typically be evaluated. If the results of applying other methods to a data set are not presented, it is difficult to ascertain if the data set is truly selective for good scoring functions.

Looking at the previous publications that have applied MM-PBSA to series data with more success, it is not entirely clear why the method is so much less successful in the present case. The following observations can be made, though. None of these prior studies evaluated as many ligands as ours; the number in previous studies ranges from 7 to 12. In one study, good correlation is only obtained for two subgroups generated by dividing the 10 studied ligands of CDK2 into two groups of 4 and 6, respectively (the division of subgroups based on chemical structure).[27] In another study of 12 ligands of HIV-1 reverse transcriptase, a surprisingly good rms deviation of only 1.12 kcal/mol is obtained between MM-PBSA prediction and experiment.[30] However, a plot of calculation versus experiment, while better than the results we obtained for p38, is not entirely convincing that the method can reliably predict small differences in binding energy among the ligands. In a study of inhibitors of cathepsin D, amazingly good agreement between MM-PBSA prediction and experiment, both in absolute ΔG and in correlation coefficient (0.98), is found. However, only seven compounds were studied. Finally, in a study of ligands to Avidin, correlation between prediction and experiment was an impressive 0.92 for nine studied compounds.[34] One thing missing from all these studies is a comparison with other methods on the same series of data. Without this comparison, it is impossible to determine if, perhaps, these data are more easily predicted in general by all scoring functions.

It may simply be that the MM-PBSA method is better suited to some of these other systems. But if that is the case (if MM-PBSA only reliably works for some ligand−protein systems), then a rapid means to figure out whether a particular system is amenable to this technique is required. Otherwise, given the enormous expense of MM-PBSA, it does not make sense to apply it to these kinds of problems when other, cheaper methods (which are at least as good) exist.

In the end, it is not really surprising that the MM-PBSA method did so poorly relative to the other approaches on this system. After all, this is an approximate technique that depends on a large number of assumptions and the total of energies derived from vastly different calculations. The chances that the errors from all those assumptions and different calculations will cancel out in a way that provides a net result accurate enough to score the binding of a set of ligands that only differ by a few kcal/mol in their experimental binding is probably unrealistic. Outside of TI, the other methods used to score the p38 series are also based on assumptions and approximations, but not as many, and the scoring is done using one or more related functions. MM-PBSA offers the chance to approximate the net free energy for a system, but even in the best circumstances, the resultant will be associated with a large error bar. Thus, in addition to the lessons enumerated above, we should be sure to pick carefully the systems to which it is applied, avoiding those where the difference of interest is small.

Acknowledgment. Thanks are extended to Dr. W. Patrick Walters for helpful discussions.

References

(1) Postma, J. P. M.; Berendsen, H. J. C.; Haak, J. R. Thermodynamics of cavity formation in water. *Faraday Symp. Chem. Soc.* **1982**, *17*, 55−67.
(2) Tembe, B. L.; McCammon, J. A. Ligand−receptor interactions. *Comput. Chem.* **1984**, *8*, 281−283.
(3) Jorgensen, W. L.; Ravimohan, C. Monte Carlo simulation of differences in free energies of hydration. *J. Chem. Phys.* **1985**, *83*, 3050−3054.
(4) Zwanzig, R. W. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.
(5) Straatsma, T. P.; Berendsen, H. J. C.; Postma, J. P. M. Free energy of hydrophobic hydration: A molecular dynamics study of noble gases in water. *J. Chem. Phys.* **1986**, *85*, 6720−6727.
(6) Lybrand, T.; McCammon, J. A.; Wipff, G. Theoretical calculation of relative binding affinity in host−guest systems. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 833−835.
(7) Hwang, J. K.; Warshel, A. Semiquantitative calculations of catalytic free energies in genetically modified enzymes. *Biochemistry* **1987**, *26*, 2669−2673.
(8) Mitchell, M. J.; McCammon, J. A. Free energy difference calculations by thermodynamic integration. Difficulties in obtaining a precise value. *J. Comput. Chem.* **1991**, *12*, 271−275.
(9) Pearlman, D. A.; Kollman, P. A. The overlooked bond-stretching contribution in free energy perturbation calculations. *J. Chem. Phys.* **1991**, *94*, 4532−4545.
(10) Pearlman, D. A. Determining the contributions of constraints in free energy calculations: development, characterization, and recommendations. *J. Chem. Phys.* **1993**, *98*, 8946−8957.
(11) Straatsma, T. P.; McCammon, J. A. Multiconfiguration thermodynamic integration. *J. Chem. Phys.* **1991**, *95*, 1175−1188.
(12) Pearlman, D. A. Free energy derivatives: a new method for probing the convergence problem in free energy calculations. *J. Comput. Chem.* **1994**, *15*, 105−123.
(13) Mazor, M.; Pettitt, B. M. Convergence of the chemical potential in aqueous simulations. *Mol. Simul.* **1991**, *6*, 1−4.
(14) Helms, V.; Wade, R. C. Free energies of hydration from thermodynamic integration: Comparison of molecular mechanics force fields and evaluation of calculation accuracy. *J. Comput. Chem.* **1997**, *18*, 449−462.
(15) Hodel, A.; Simonson, T.; Fox, R. O.; Brunger, A. T. Conformational substates and uncertainty in macromolecular free energy calculations. *J. Phys. Chem.* **1993**, *97*, 3409−3417.
(16) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P.; Gallop, M. A. Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.
(17) Frank, R. Simultaneous and combinatorial chemical synthesis techniques for the generation and screening of molecular diversity. *J. Biotechnol.* **1995**, *41*, 259−272.
(18) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening. An overview. *Drug Discovery Today* **1998**, *3*, 160−178.

(19) 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.
(20) Wang, W.; Wang, J.; Kollman, P. A. What determines the van der Waals coefficient beta in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations? *Proteins* **1999**, *34*, 395−402.
(21) Carlson, H. A.; Jorgensen, W. L. An extended linear response method for determining free energies of hydration. *J. Phys. Chem.* **1995**, *99*, 10667−10673.
(22) Radmer, R. J.; Kollman, P. A. The application of three approximate free energy calculations methods to structure based ligand design: trypsin and its complex with inhibitors. *J. Comput.-Aided Mol. Des*. **1998**, *12*, 215−227.
(23) Pearlman, D. A. Free energy grids: a practical qualitative application of free energy perturbation to ligand design using the OWFEG method. *J. Med. Chem.* **1999**, *42*, 4313−4324.
(24) Srinivasan, J.; Miller, J.; Kollman, P. A.; Case, D. A. Continuum solvent studies of the stability of RNA hairpin loops and helices. *J. Biomol. Struct. Dyn.* **1998**, *16*, 671−682.
(25) Vorobjev, Y. N.; Almagro, J. C.; Hermans, J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. *Proteins* **1998**, *32*, 399−413.
(26) Jayaram, B.; Sprous, D.; Yong, M. A.; Beveridge, D. L. Free energy analysis of the conformational preferences of A and B forms of DNA in solution. *J. Am. Chem. Soc.* **1998**, *120*, 10629−10633.
(27) 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.
(28) Honig, B.; Sharp, K. A.; Yang, A.-S. Macroscopic models of aqueous solutions. Biological and chemical applications. *J. Phys. Chem.* **1993**, *97*, 1101−1109.
(29) Pearlman, D. A.; Rao, B. G. Free energy calculations: methods and applications. *Encycl. Comput. Chem.* **1998**, 1036−1061.
(30) Wang, J.; 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.
(31) Gohlke, H.; Kiel, C.; Case, D. A. Insights into protein−protein binding by binding free energy calculation and free energy decomposition for the Ras−Raf and Ras−RalGDS complexes. *J. Mol. Biol.* **2003**, *330*, 891−913.
(32) Huo, S.; Wang, J.; Cieplak, P.; Kollman, P. A.; Kuntz, I. D. Molecular dynamics and free energy analyses of cathepsin D−inhibitor interactions: insight into structure-based ligand design. *J. Med. Chem.* **2002**, *45*, 1412−1419.
(33) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889−897.
(34) Kuhn, B.; Kollman, P. A. Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J. Med. Chem.* **2000**, *43*, 3786−3791.
(35) Gouda, H.; Kuntz, I. D.; Case, D. A.; Kollman, P. A. Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. *Biopolymers* **2003**, *68*, 16−34.
(36) Chong, L. T.; Duan, Y.; Wang, L.; Massova, I.; Kollman, P. A. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14330−14335.
(37) Pearlman, D. A.; Charifson, P. S. Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *J. Med. Chem.* **2001**, *44*, 3417−3423.
(38) 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.
(39) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand−receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503−519.
(40) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, 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.
(41) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; et al. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317−324.
(42) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.
(43) 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.
(44) Wilson, K. P.; McCaffrey, P. G.; Hsiao, K.; Pazhanisamy, S.; Galullo, V.; et al. The structural basis for the specificity of pyridinylimidazole inhibitors of p38 MAP kinase. *Chem. Biol.* **1997**, *4*, 423−431.
(45) Pearlman, D. A.; Charifson, P. S. Improved scoring of ligand−protein interactions using OWFEG free energy grids. *J. Med. Chem.* **2001**, *44*, 502−511.
(46) Pearlman, D. A.; Walters, W. P. Manuscript in preparation.
(47) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(48) Pearlman, D. A.; Case, D. A.; Caldwell, J. C.; Ross, W. S.; Cheatham, T. E., III; et al. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Comm.* **1995**, *91*, 1−41.
(49) Warwicker, J.; Watson, H. C. Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. *J. Mol. Biol.* **1982**, *157*, 671−679.
(50) Sharp, K. A.; Honig, B. Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301−332.
(51) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144−1149.
(52) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978−1988.
(53) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, revision A.3; Gaussian, Inc.: Pittsburgh, PA, 1998.
(54) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305−320.
(55) Brooks, B. R.; Janezic, D.; Karplus, M. Harmonic analysis of large systems. *J. Comput. Chem.* **1995**, *16*, 1522−1553.
(56) MacKerell, J. A. D. D.; Bashford, D.; Bellot, M.; Dunbrack, R. L.; Evanseck, J. D.; et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(57) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating water. *J. Chem. Phys.* **1983**, *79*, 926−935.

JM050306M