Expedited Articles

Improved Scoring of Ligand−Protein Interactions Using OWFEG Free Energy Grids

David A. Pearlman* and Paul S. Charifson

Vertex Pharmaceuticals Incorporated, 130 Waverly Street, Cambridge, Massachusetts 02139-4242

Received August 30, 2000

A new approach to rapidly score protein−ligand interactions is tested on several protein−ligand systems. Results using this approach − the OWFEG free energy grid − are quite promising and are generally in better agreement with experiment (in some cases *much* better) than those obtained employing scoring techniques currently in wide use. The OWFEG free energy grid is generated from a one-window free energy perturbation MD simulation (Pearlman, D. A. *J. Med. Chem.* **1999**, *42*, 4313−4324). The OWFEG approach is applied to three protein systems: IMPDH, MAP kinase p38, and HIV-1 aspartyl protease. OWFEG scores are compared to experimental K_i and IC_{50} data in each case. Empirical scoring functions applied to the same systems for comparison include ChemScore, Piecewise Linear Potential (PLP), and Dock energy score.

Introduction

Molecular docking has been widely adopted as an integral part of the drug discovery paradigm.[1−3] A collection of potential ligands is docked into a protein binding site and then rapidly scored. Those compounds with the best scores are carried forward for further consideration. This procedure makes it practical to screen very large proprietary databases and is also essential when attempting to suggest candidates for synthesis from huge "virtual" databases of compounds that can be synthesized (either through combinatorial methods or through traditional bench chemistry).[4]

Screening in this fashion is only of value, of course, if one has confidence in the results obtained. Improvements in docking methods to include flexibility in the ligand molecule as well as to substantially speed the algorithm used have rendered this portion of the procedure reasonably reliable and generally applicable.[5−14] However, accurately scoring the docked configuration remains problematic. Although there are rapid scoring methods and techniques which can consistently indentify nanomolar and low-micromolar binders from large databases,[15] there are presently no structure-based methods for accurately ranking smaller sets of related analogues that are reliable across multiple protein systems and within a narrow range of affinities.

Significant shortcomings shared by scoring schemes currently in common use include: (1) they do not take into account flexibility in the protein binding site; (2) they treat entropic effects (which are frequently the primary driving force in ligand binding) in a crude, empirically based manner; and (3) they score on the basis of the bound complex alone, instead of as the difference between the bound (desolvated) and free (fully solvated) forms.[16] While scoring methods have been described that consider protein flexibility in some fashion[17−19] or that attempt to introduce a correction for the effects of solvation when binding,[20] none of these methods has simultaneously and completely considered both effects. Theoretical calculations exist which can address all these shortcomings at a detailed atomic level, such as free energy perturbation,[21] thermodynamic integration,[21] and (for entropic effects only) certain semiempirical quantum mechanical methods.[22] However, these shortcomings persist because the types of calculations necessary to quantitatively address them are simply too computationally intensive: we cannot possibly hope to apply methods such as free energy perturbation to the hundreds of thousands (or more) of compounds that might be run through a virtual screen.

Recently, we described a new method that attempts to bridge the gulf between high-accuracy low-throughput free energy simulations and empirical approaches.[23] This method, OWFEG (**o**ne-**w**indow **f**ree **e**nergy **g**rid), utilizes the same statistical mechanical equation that forms the basis of free energy perturbation. However OWFEG applies this equation in a way that makes it possible to generate from a single MD simulation a grid surrounding a molecule of interest that represents the free energy for insertion of a probe group at any point on that grid. The tradeoff is that the calculated free energy at each point is only qualitative. But taken as a whole, it has been demonstrated that the OWFEG grid is suggestive and reliable as a predictor of the effects of molecular changes.[23] Though free energy calculations are CPU intensive, the time penalty with OWFEG is paid only once, up-front, for the single MD simulation used to generate the grid. Once the OWFEG grid has been generated, the scoring is as rapid as with any empirical method.

* To whom correspondence should be addressed. Phone: 617-577-6522. Fax: 617-577-6400. E-mail: dap@vpharm.com.

10.1021/jm000375v CCC: $20.00
Published on Web 02/08/2001

In the previous study, OWFEG was applied to a ligand molecule to probe questions of how to change the ligand to improve its binding to a receptor protein. An OWFEG grid is generated about the ligand molecule, and additions to the ligand are made, based on the grid, in an attempt to optimize the net free $\Delta\Delta G$ of binding as desired. In this paper, OWFEG is used to generate a free energy grid in the (unbound) binding site of a protein. The free energy grid is then used to score a set of docked molecules. As will be shown, the results are quite promising and (depending on the system) either as good as or better than existing scoring methods.

We have applied OWFEG scoring to three protein systems for which appropriate sets of ligand data (K_i or IC_{50}) have been determined and for which crystallographic complexes have been reported: inosine monophosphate dehydrogenase (IMPDH), a 424-residue protein;[24,25] HIV-1 aspartyl protease, a 199-residue protein;[26−28] and p38 mAP kinase, a 330-residue protein.[29] In each case, the results of OWFEG scoring have been compared to those using several scoring schemes that are among the best in current use:[15] ChemScore,[30,31] Piecewise Linear Potential (PLP),[32] and Dock energy score.[33,34]

Background

From elementary equations of statistical mechanics, it can be shown that the free energy difference between two molecular states A and B is given by:[21]

$$\Delta G = G_B - G_A = -RT \ln \langle e^{-(V_B - V_A)/RT} \rangle_A \qquad (1)$$

This equation relates relates the free energy difference between states A and B to the ensemble average of a quantity that depends on the difference in the potential energies for those states. G_B and G_A are the respective free energies of states B and A. "$\langle\rangle_A$" refers to the average of the quantity within the brackets, evaluated from an ensemble representative of state A. V_B and V_A are the potential energies of the respective states, evaluated for the same configuration of the system. R is the gas constant, and T is the temperature. Free energy calculations carried out using this equation are typically termed FEP (**f**ree **e**nergy **p**erturbation) simulations.

If we run MD to generate an ensemble of states for the system of interest and use this ensemble to evaluate the average on the right-hand side of this equation, we can, in principle, derive the desired free energy difference. The difficulty in this approach arises primarily from difficulties in carrying out sufficient molecular sampling to accurately evaluate the ensemble average. These problems are exacerbated by the particular form of the ensemble that needs to be determined. Note that although the configurations from which the average is calculated are taken from a MD simulation carried out using the potential energy function representing state A, V_A, the quantity being averaged depends on both V_A and V_B. In other words, it is important, if we are going to obtain a good, quantitative, converged value for the ensemble, to sample configurations that are low in energy not only for state A but also for state B. However, if states A and B are disparate, then there may be configurations that correspond to low energies for state B but to high energies for state A. These will never get sampled. To circumvent this problem, a factor λ is introduced into the potential energy function such that $V(\lambda=0,\mathbf{r}) = V_A$ and $V(\lambda=1,\mathbf{r}) = V_B$. Then, instead of attempting to calculate the net free energy between disparate states A and B in a single simulation, this free energy is determined from a series of simulations between more similar nonphysical λ intermediates. As a state function, free energy is path-independent, and we can sum the free energies between these intermediates to give the total free energy:

$$\Delta G_{tot} = \sum_{i=1}^{NWINDOW} \Delta G_{\lambda(i-1)\lambda(i)} \qquad (2)$$

where

$$G_{\lambda(i-1)\lambda(i)} = -RT \ln \langle e^{-[V(\lambda(i),\mathbf{x}) - V(\lambda(i-1),\mathbf{x})]/RT} \rangle_{\lambda(i-1)} \qquad (3)$$

Each $\delta\lambda$ interval is termed a "window", and NWINDOW is the number of $\delta\lambda$ intervals used.

The problem with FEP simulations carried out as described above is that they are very CPU intensive and a different λ path is defined for each pair (A, B) of endpoints, so each free energy change requires a separate full set of costly simulations, both for the solvated protein complexed with the ligand and separately for the solvated ligand alone. While this is acceptable when one needs only a handful of free energy differences to answer a question, it is not appropriate to answer general questions such as 'where should we change this molecule?' or 'which of these many molecules will best bind to this active site?' These types of questions are much better addressed by a construct such as an energy grid in the region of interest.

Typically, a potential energy grid is created to address these questions. Many of scoring functions in wide use for screening libraries are of this type.[32,33,35] Potential energy grids are fine for representing simple (dispersion/electrostatic) attractive and repulsive forces. However, a potential energy grid will generally reflect neither conformational flexibility of the binding site nor the entropic solvent effects of binding. These effects can be of tremendous importance, and many binding phenomena are *dominated* by the solvation effects. These are phenomena that FEP simulations are well-equipped to describe.

It is important to keep in mind that when scoring compounds for screening, one is not typically concerned with quantitatively precise scores. What is significantly more important is that the scoring function be able to qualitatively separate "good" binders from "bad" ones. In light of this reduced need for accuracy, the FEP method merits another look. What if we can determine a *qualitatively* predictive free energy difference directly from eq 1, without introducing the λ dependence of eqs 2 and 3? If we can, then it becomes straightforward to generate a free energy grid from a single MD simulation. This is because each free energy simulation shares the exact same reference state A, and the free energy is calculated from a single simulation, carried out using V_A. At each grid point we carry out a "one-window" simulation where the reference state A corresponds to "nothing" at the grid point and the final state B corresponds to a probe atom in the same location. This

is the OWFEG method.[23] In essence, we are calculating at each grid point:

$$\Delta G_{\text{grid-point}} = -RT \ln \langle e^{-V_{\text{probe}}/RT} \rangle_{\text{no-probe}} \qquad (4)$$

which derives from eq 1 substituting "probe" for state B, "no probe" for state A, and 0 for $V_{\text{no-probe}}$. The probe atom can be neutral or charged, and typically three grids are generated corresponding to a neutral probe, a positively charged probe, and a negatively charged probe. From the resulting three grids, linear interpolation can be used to evaluate the grid score for an atom of any charge. Since the regions about which the grid is generated will typically be flexible, it is necessary to define a reference frame for each grid point. OWFEG utilizes a FIRF (**f**loating **i**ndependent **r**eference **f**rame). With the FIRF, the reference system for each grid point is defined at the onset of the simulation by the closest atom (and its 1−2 and 1−3 neighbors) of the molecule about which the grid is being generated. This is critical[23] and allows an appropriate grid to be generated for any molecule, regardless of its innate flexibility. While the reference frame is flexible (i.e. the location of the grid point moves as the MD simulation progresses), at the end of the simulation, the ΔG at each grid point is that which is appropriate for the initial position of that grid point. This allows us to reference back all the ΔG values to the initial grid, which is necessary if we are to use them for scoring.

OWFEG simulations are carried out allowing weakly restrained motion of the region of the protein near the active site and free movement of all solvation waters. As a result, the free energies determined include enthalpic and entropic contributions that are averaged over conformational states available to the protein and the configurational states available to the solvent. This is a critical advantage over methods that fail to explicitly consider protein flexibility or solvent effects and over methods that fail to include entropic contributions in some form (either implicitly or explicitly).

Methods

All simulations were carried out using a version of the Amber/Sander 5.0 MD program[36] that has been modified to perform OWFEG calculations.[23] The statistics required to determine the grid point free energies according to eq 4 were determined from 1 ns of MD sampling. All MD was run using a 2-fs time step, with bonds constrained using SHAKE.[37] A nonbonded cutoff of 8 Å was used, with a pairlist update every 20 steps. All-atom representations of the protein and solvent were used. For each run, a "belly" of moving protein residues was defined as those within 15 Å of the binding site, along with all water molecules. Protein residues outside this belly remained fixed during the simulation. Explicit water molecules were added to solvate the binding site. Waters were added to fill a spherical volume with radius 25 Å centered on the binding site. All water molecules were free to move in all simulations. In each simulation, a positional restraint with a modest force constant of $K_{\text{rest}} = 0.5$ kcal/mol was applied to the protein residues within the moving belly, to keep the protein conformation from deforming too severely from the crystal conformation.

All models were generated starting with X-ray crystallographic coordinates. Prior to running MD/OWFEG, minimization was carried out to remove any steric overlaps and to fully regularize the structure. The resulting minimized coordinates were used as the reference coordinates for the positional restraints. Then, between 50 ps and 1 ns of MD equilibration (see below), using the same simulation parameters as described above, was performed prior to collecting data for the OWFEG map. The standard form of the Amber force field was used to evaluate energies, using parameters from Weiner et al.[38] This force field contains terms to reflect the contributions from bond, valence angle, torsion, Leonard−Jones, electrostatic (point charge), and H-bond interatctions.

OWFEG scoring was performed for three protein−ligand sets. The proteins examined were IMPDH,[24,25] HIV-1 aspartyl protease,[26−28] and p38.[29] For IMPDH and p38, the set of ligands to be scored was derived from the Vertex Pharmaceuticals[39] project database for the relevant protein. For HIV-1 PR, the structures and binding data were previously available.[40] The ligands for each protein were selected to span a wide range of binding efficacies and to include several scaffold classes. More details on the design of the test sets used and the docking methods employed are presented elsewhere.[15] All structures used in this study were evaluated graphically to ensure that they were properly docked (by visual inspection and comparison to known crystallographic ligands). In each case, IC_{50} (or K_i) data were measured using the same assays and so should be self-consistent. A total of 37 ligands were scored against IMPDH, 33 against HIV-1 PR, and 44 against p38. To put the OWFEG results into context, the same sets of ligands were scored against several scoring functions in current use, including ChemScore,[30,31] PLP,[32] and Dock energy score.[33,34] In a recent study, it was demonstrated that these functions are among the best currently available for correctly selecting out known active compounds from large databases of inactives.[15]

The parameters of the probe group at each grid position were $r^* = 2.0$ Å and $\epsilon = 0.15$ kcal/mol. These parameters are representative of a united atom methyl group.[38] Three grids were generated during each OWFEG run, corresponding to probe groups with +0.3e, 0.0e, and −0.3e net charges. The ±0.3 charges were taken as crudely representative of nitrogen- and oxygen-containing groups, respectively. The appropriate free energy for any charged atom at a grid point can then be determined by linear interpolation among the three grid point values. A rectilinear OWFEG grid was formed in and around the binding site of the crystallographic structure of the protein. The dimensions of the grid initially extended 3 Å beyond any atom in the ligand test set in each (x,y,z) dimension. Grid points were then pruned away so that only those points within 4.0 Å of at least one protein atom remained. Grid points were also pruned away that were closer than 0.75 Å from any protein atom. Pruning the grid point list of unnecessary points is valuable due to the considerable expense of running simulations with very large numbers of grid points. Through pruning (and relatively wide grid spacing), appreciable decreases in total simulation times can be achieved with little or no affect on the quality of the results. A total of 8874 grid points were calculated for IMPDH, 14038 points for HIV-1 PR, and 8120 points for p38.

The free energy at each grid point was calculated from eq 4 with V_{probe} given as the nonbonded part of the force field, i.e.:

$$V_{\text{probe}} = \sum_{i \neq j} \left\{ \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right] + q_i q_j / \epsilon R_{ij} \right\} \qquad (5)$$

The nonbonded pairlist used in the summations was that of the closest protein atom to the grid point.

The ligands were docked into the active site using the Dock 4.0.1 method,[41] as previously described.[15] The docked conformations were then scored by the OWFEG grid. Any hydrogen atoms were ignored and their charges were summed into the attached heavy atom. For each atom, two of the three grids were chosen, either the neutral and positively charged probe grids or the neutral and negatively charged probe grids, depending on whether the atom had a partial positive or negative charge. The cell in the grids corresponding to the atom was located in both grids, and the OWFEG energy trilinearly interpolated on that cell was calculated in each grid. The energy for an atom of charge X was then given by

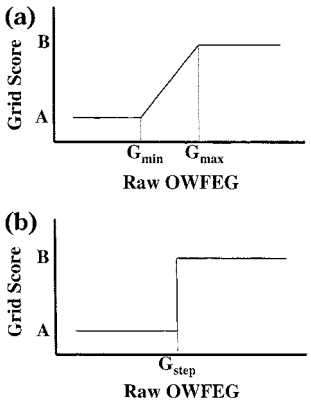

Figure 1. Forms of the simple conversion function that scales and limits the raw OWFEG grid data for use in scoring. The abcissa represents the value in the OWFEG grid. The ordinate represents the modified value. (a) Functional form originally used to scale the OWFEG data. Values less that G_{min} in the original OWFEG grid are set equal to A. Values greater than G_{max} are set equal to B. Values between G_{min} and G_{max} are scaled linearly. Note that the maximum limit value, B, can be set arbitrarily. The important value is the difference between this value and A. In our work, we have set $B = 1.0$ kcal/mol. Thus, this function has three fitted coefficients. (b) Best results using functional form (a) were obtained with $G_{min} = G_{max}$, indicating that the best conversion function between raw OWFEG and a scoring grid is actually given by this simple step function. This functional form has only two fitted parameters: G_{step} and B (A is again arbitrarily set to 1.0 kcal/mol in our work).

interpolation between the two grid scores:

$$G_{score} = X\frac{(G_{grid-2} - G_{grid-1})}{Q_{probe}} + G_{grid-1} \qquad (6)$$

where G_{grid-1} and G_{grid-2} are the scores on the neutral and charged probe grids, and Q_{probe} is the charge of the probe group used to generate the charged OWFEG grid (here, ±0.3e). The net score for each ligand is given by the sum of G_{score} over all atoms of the ligand.

The grids generated by OWFEG ultimately fall somewhere between qualitative and quantitative in nature. As has been shown, the predictions they allow are excellent in terms of differentiating "good" and "bad" locations for probe group addition. However, the raw free energy values in an OWFEG map are not directly numerically predictive of experimental values. In particular, the range of values (maximum to minimum) calculated directly by OWFEG for grid points that could "reasonably" be occupied will be much larger than is realistic. For this reason, it is necessary to rescale the OWFEG values before using them in a scoring function.

To convert the raw OWFEG data to a scoring function, a very simple capped linear conversion approach was used (Figure 1a). Note that this conversion requires the specification of three parameters: A, G_{min}, and G_{max}. Because the energies predicted by the resulting modified OWFEG grid are to be used to calculate *relative*, rather than absolute, binding energies, B in Figure 1a can be specified arbitrarily. In our work, we chose it to be 1.0. While one could obviously devise more complicated conversion routines, it was our aim to keep the function intuitive and at the same time to keep the number of variables one must specify down to a minimum.

For the OWFEG scoring grids to be truly useful in the general case, it is necessary that they can be used *before* we have substantial binding data for a particular system. Therefore, it is important that we be able to derive a set of conversion constants (A, G_{max}, and G_{min}) that can be applied to all cases. For this reason, these constants were chosen so that a single set of three constants optimized the scoring performance averaged over all three systems examined here. The same constants were applied to each of the three grids in each case (neutral, positively charged and negatively charged probe groups). Optimization of A, G_{min}, and G_{max} was performed by systematically varying these three constants to form a set of starting guesses. The set of starting guesses consisted of all combinations of $G_{max} = 15$ to -15 kcal/mol in increments of 5 kcal/mol, $G_{min} = G_{max}$ to -15 kcal/mol in increments of 5 kcal/mol, and $A = 0.5$ to -2 kcal/mol in increments of 0.5 kcal/mol, with $B = 1.0$ kcal/mol. Simplex optimization of these constants was performed for each starting set, optimizing on $<R^2>$, the value of the linear correlation coefficient between OWFEG energy and experimentally measured pK_i (or pIC_{50}) determined separately for all three systems (HIV-1 PR, IMPDH, p38), and then averaged. A program, OPTIMIZE_OWFEG, was written to perform the simplex optimizations. In the end, the minimized set of variables that resulted in the largest averaged $<R^2>$ was used in all scoring.

Before scoring on the OWFEG grid, each molecule was independently docked into the appropriate active site using the Dock 4.0.1 program and evaluated with the Dock energy scoring function. Atomic point charges were assigned to each atom using the Gasteiger Marselli method.[42] Any hydrogens were removed and their associated charges were added into the associated heavy atom before scoring.

Results

Optimization of the three constants required to convert an OWFEG grid to a scoring grid was performed using the OPTIMIZE_OWFEG program, as described in Methods. We obtained optimized values of $A = -14.804$, $G_{min} = 5.988$, and $G_{max} = 5.998$, yielding an averaged value of $<R^2> = 0.702$. Note that we imposed no restraint on the relative values of G_{min} and G_{max}, so it is surprising that the refinement leads to these two values being equal. The effective scoring function is shown in Figure 1b. This is a single-step function, with only two fitted parameters: the location of the step G_{step} and the difference between the energies on the two sides of the step. In a sense, optimization leads to a result reflective of the qualitative nature of these maps: grid points are either "favorable" (lower than G_{step}) or "unfavorable" (greater than G_{step}). The OWFEG score for each atom represents a linear interpolation between two step function grids: one for neutral probes and one for (either positively or negatively) charged probes.

The results, with experimental pK_i or pIC_{50} plotted against OWFEG grid score, are plotted in Figure 2. The individual linear correlation coefficients are -0.767 for IMPDH, -0.630 for HIV-1 PR, and -0.702 for p38. Cursory examination of these plots indicates that the OWFEG scoring has been successful in the ultimate goal of screening: providing the ability to choose an energy cutoff that yields an appreciable enrichment of strong binders in the compounds that survive the cutoff. In each case, it is clear that we can set an energy cutoff that will result in a dramatically reduced set of compounds that still includes all, or most, of the compounds with the best binding constants (highest values of pK_i or pIC_{50}).

The significance of a new scoring function is, of course, dependent on how that function improves on those already available. A recent paper presented detailed comparisons of the scoring functions currently in general use.[15] For comparison, we applied the best scoring functions, as determined in that paper, to the same data sets to which OWFEG has been applied. In Figures 3−5, we present the scoring data for these same data sets when scored using the ChemScore, PLP, and Dock

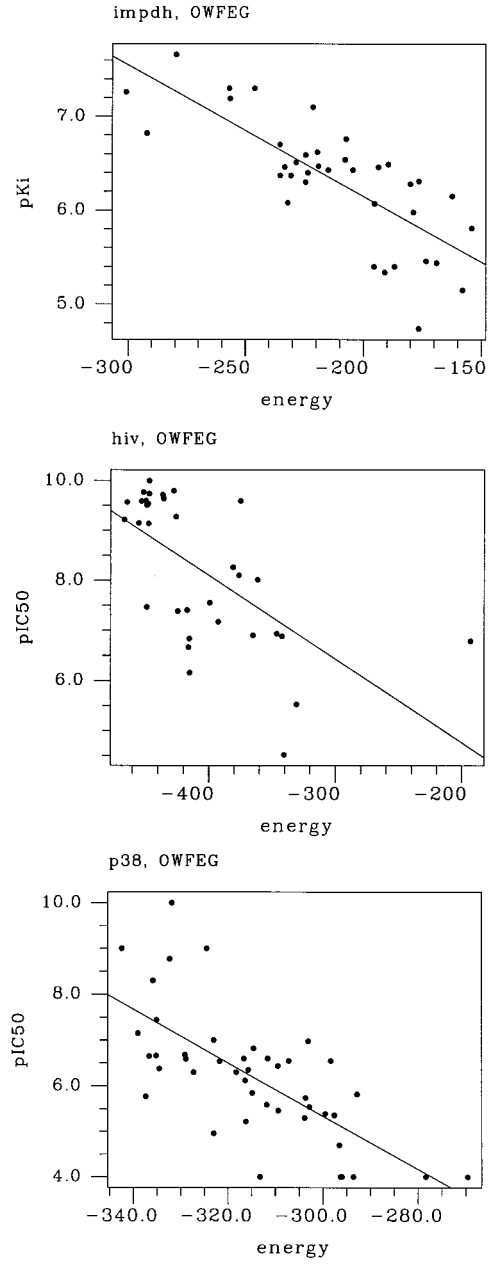

Figure 2. pK_i (or pIC_{50}) plotted against the OWFEG grid score for the IMPDH, HIV-1 PR, and p38 model systems. The scoring grid for each system was obtained from the raw OWFEG map using the scaling function in Figure 1b, with $G_{step} = 5.993$ kcal/mol, $A = -14.804$, and $B = 1.0$ kcal/mol. The line plotted for each set represents the least-squares best-fit line through the data. The correlation coefficients for the least-squares fits are −0.767, −0.630, and −0.702, respectively, for IMPDH, HIV-1 PR, and p38. Note that the absolute values of the OWFEG scores are arbitrary.

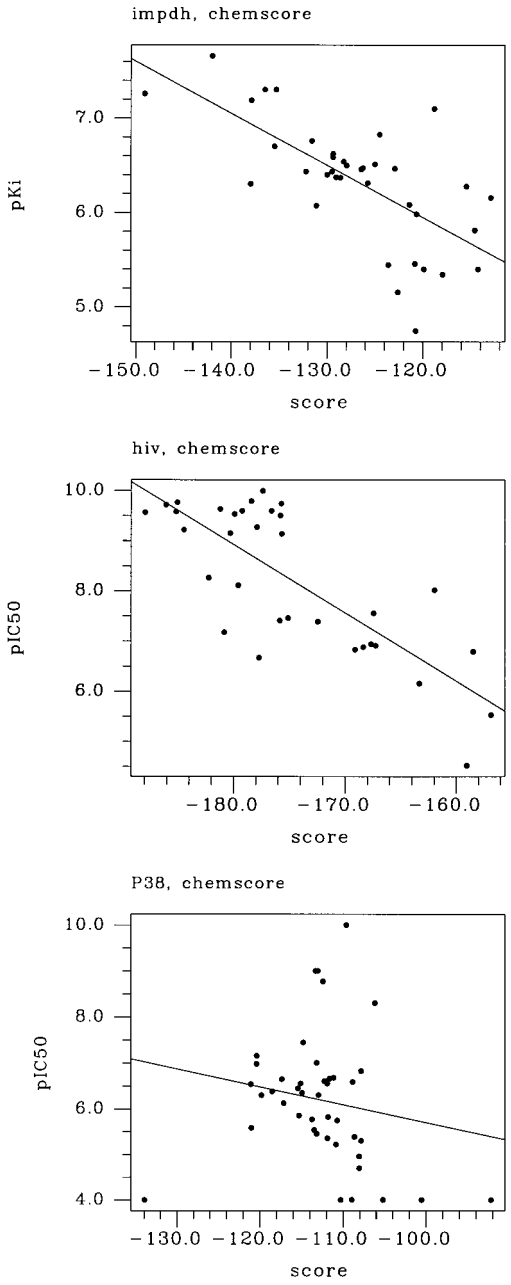

Figure 3. pK_i (or pIC_{50}) plotted against the ChemScore grid score for the IMPDH, HIV-1 PR, and p38 model systems. The line plotted for each set represents the least-squares best-fit line through the data. The correlation coefficients for the least-squares fits are −0.679, −0.776, and-0.175, respectively, for IMPDH, HIV-1 PR, and p38.

energy scoring methods. From cursory visual inspection as well as on the basis of correlation coefficient, it can be seen that OWFEG does appreciably better than any of these scoring functions in the p38 case, somewhat better than any of these scoring functions in the case of IMPDH, and roughly about the same in the case of HIV-1 PR. The only scoring function that results in a better correlation coefficient than OWFEG, and then only for the HIV-1 PR case, is ChemScore (not surprisingly, since the ChemScore function was originally parameterized against this set of HIV-1 PR data). On the basis of the correlation coefficient, the superiority of OWFEG for scoring p38 is actually quite striking. Compared to the correlation coefficient of −0.702 achieved by OWFEG, the best correlation coefficient achievable by any of the three other scoring functions we applied was −0.542 using the Dock scoring function.

Correlation coefficient only tells part of the story, however. To fully appreciate this, it is necessary to place the scoring results in context. Scoring is performed, in practice, to reduce the size of the database that must be probed experimentally. Sceening has value only if the ability of the scoring function to select out active compounds is greater than random. In other words, after the screen, the remaining portion of the database must be richer in "hits" than was the initial

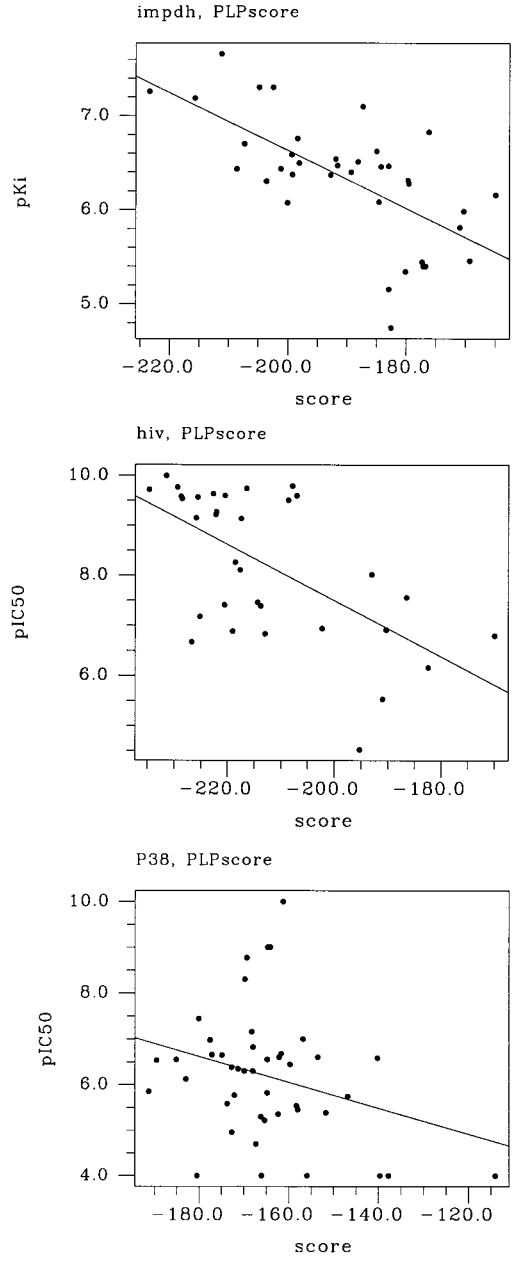

Figure 4. pK_i (or pIC_{50}) plotted against the PLP grid score for the IMPDH, HIV-1 PR, and p38 model systems. The line plotted for each set represents the least-squares best-fit line through the data. The correlation coefficients for the least-squares fits are −0.658, −0607, and −0.289, respectively, for IMPDH, HIV-1 PR, and p38.

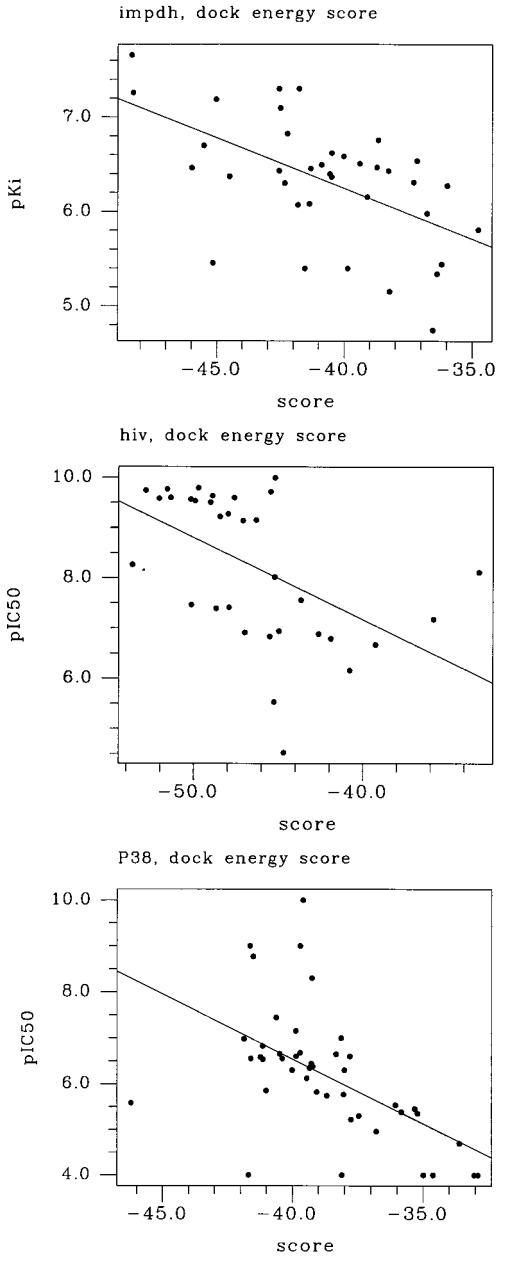

Figure 5. pK_i (or pIC_{50}) plotted against the Dock grid score for the IMPDH, HIV-1 PR, and p38 model systems. The line plotted for each set represents the least-squares best-fit line through the data. The correlation coefficients for the least-squares fits are −0.556, −0.511, and −0.542, respectively, for IMPDH, HIV-1 PR, and p38.

database. Note that this sets a necessary condition for a screening function in terms of its ability to isolate compound "hits" as better scoring compounds but does not dictate *anything* about the ability of that function to properly score poorly binding compounds. Consider the extreme case of a function that assigns the exact same low energy score to all "hits" and that assigns a completely random high energy score to all nonhit compounds. This function would have a very poor correlation coefficient. However, this function would also do an excellent job of separating out "hits" during a screen. Now, in practice, scoring functions do not typically exhibit this type of extreme dichotomy between the quality of scoring for hits and nonhits, so the correlation coefficient *is* at least a decent qualitative indicator of general scoring ability. But one needs to go further.

In fact, what one needs to consider are "hit enrichment" plots. In such a plot, the enrichment factor (EF) is plotted versus the fraction of the database that is screened. To calculate the EF, we first identify the compounds in the database that will be considered "hits". As screening is primarily a qualitative exercise − we do not expect the screen to differentiate 50 and 250 nM compounds but we do desire that the screen be able to differentiate these from millimolar hits − all the best compounds can be considered equivalent in terms of evaluating the quality of the function, and these will all be deemed "hits". For HIV-1 PR, the best 16 binders

form a distinct cluster spanning 3 orders of magnitude in IC_{50}, and this was taken as the "hit" set. Similarly, for p38, the best five binders form a distinct cluster that can be taken as the "hit" class. For IMPDH, the top 7 compounds form a sparsely populated tail in the K_i distribution and span 3 orders of magnitude in K_i, and these were taken as the "hit" set (see any of Figures 2−4).

Having chosen the target hits, the EF can now be calculated. The EF is given for any fraction of the database (the "sampled set") as:

$$EF = \frac{HITS_{\text{sampled-set}}/N_{\text{sampled-set}}}{HITS_{\text{total-database}}/N_{\text{total-database}}} \qquad (7)$$

where *HITS* is the number of experimentally best binding hits and *N* is the number of compounds. By this definition, complete random selection will yield EF = 1.0. An EF = 6 means the selected fraction of the database is 6 times as rich in hits as was the initial database. An EF less than 1 means the scoring function is performing worse than random. Note that the sampled set is always chosen as the $N_{\text{sampled-set}}$ compounds with the lowest energy scores.

The EF is plotted versus the fraction of the screened database for all four scoring functions and for all three protein systems in Figure 6. From these plots, the suggestions of the raw data plots are made explicit: OWFEG scoring performs appreciably better than any of the other scoring functions in the IMPDH and p38 cases. For HIV-1 PR, OWFEG performs somewhat better than any of the other scoring routines when >20% of the database is selected and performs about the same as ChemScore and PLP (and better than Dock) for tighter screens. Note that it is the behavior in these plots for low fractions of the database selected that is of the most significance. This is the region of the curve that represents the behavior we will be exploiting when actually performing a screen. Note also that the first (fraction selected = 0.1) point in these plots represents a sample size of only 4 points, and so some noisiness at this end of the plots is expected.

It is of particular note that in the case of p38, ChemScore and PLP actually do appreciably *worse than random* when >40% of the database is selected. For p38, the Dock energy score performs much better than ChemScore or PLP (though much worse than OWFEG). Surprisingly, outside of the superiority of OWFEG in each case, there is no consistency among the other scoring functions. Of the non-OWFEG scores, the Dock energy score performs worst for small sample sets for IMPDH and HIV-1 PR but performs best for p38. After OWFEG, ChemScore performs best for IMPDH but worst for p38.

To determine how dependent the OWFEG results are on the electrostatic contribution, we re-ran the OWFEG scoring using only a single grid corresponding to the neutral probe atom. The results from this scoring are presented as enrichment factors (EF) in Figure 7. As can be seen, the results for IMPDH and p38 scoring are clearly worse (though still better than obtained using other scoring methods). In comparison, the HIV-1 PR results are only slightly changed. The implication is that binding to the HIV-1 PR binding site is primarily hydrophobic in nature, while binding to both IMPDH and p38 sites reflects a greater hydrophilic component. Note that in all three cases, the ligands considered are all net neutral. We would expect the electrostatic component to play a significantly larger role for solutes with a net charge.

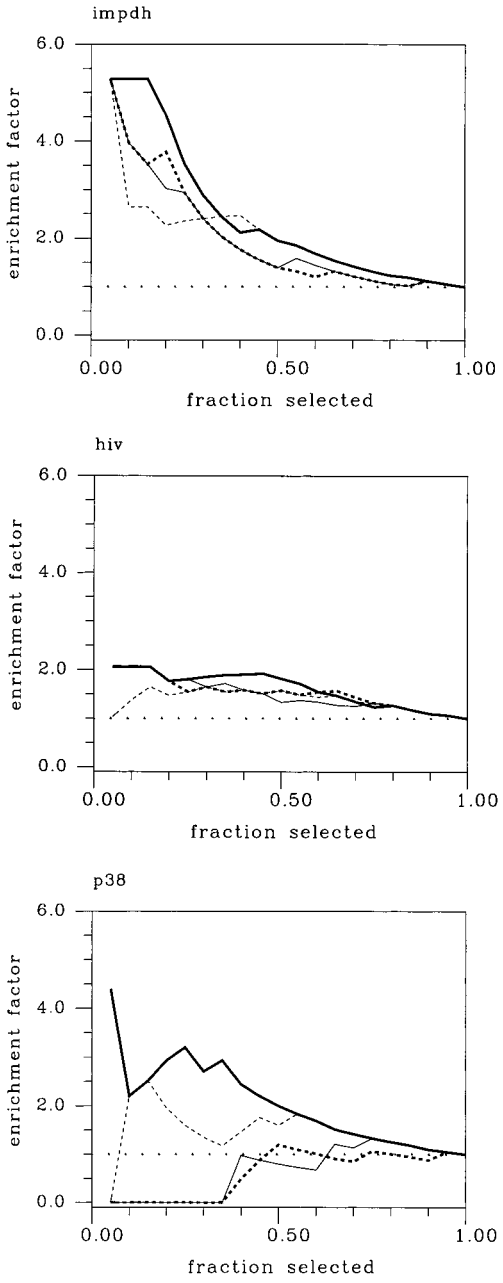

Figure 6. Enrichment factor (EF) as a function of how much of the database is selected. The EF is the ratio of hits to chosen compounds in the selected set, divided by this same ratio in the complete database (see eq 7): e.g. EF = 6 means each compound in the selected set is 6 times as likely to be a hit as a compound randomly chosen from the database as a whole; EF of less than 1 means the scoring function is worse than random. Legend: thick solid lines, OWFEG; thick dashed lines, ChemScore; thin solid lines, PLP scoring; thin dashed lines, Dock energy score; horizontal dotted line, EF = 1.0, the value random selection would yield.

Discussion

We have applied the OWFEG free energy grid approach to scoring sets of ligands docked to the IMPDH,

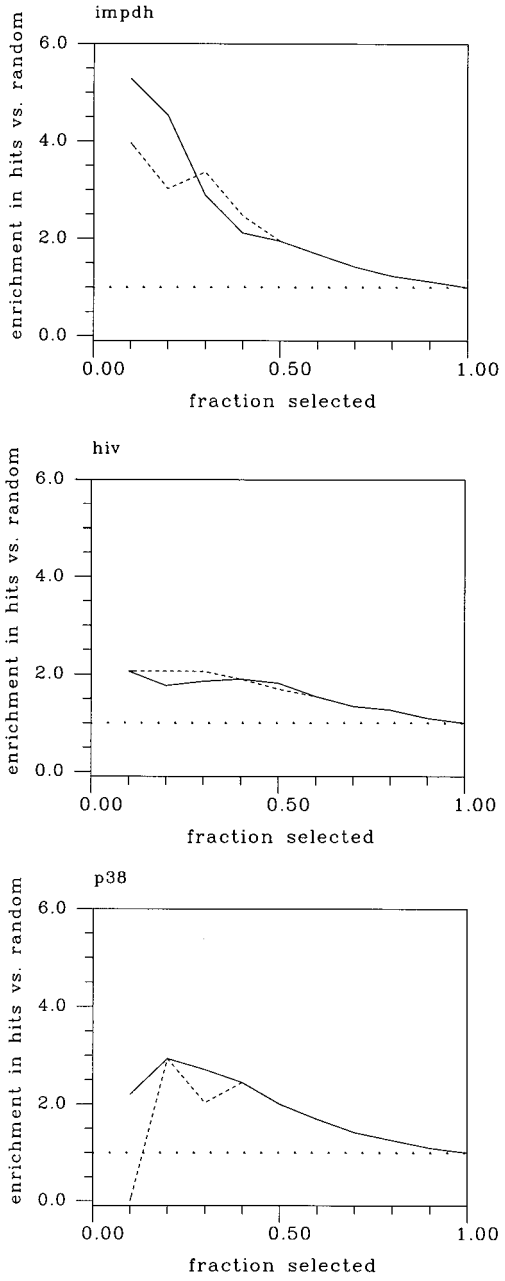

Figure 7. Enrichment factor (EF) as a function of how much of the database is selected, plotted for OWFEG scoring with (solid lines) and without (dashed lines) use of the charged probe grids along with the neutral probe grid which is used in both cases. The horizontal dotted line represents EF = 1.0, the value random selection would yield.

HIV-1 PR, and p38 proteins. The results from this approach are quite impressive, with linear correlation coefficients between OWFEG score and experimental pK_i of −0.767, −0.630, and −0.702 for IMPDH, HIV-1 PR, and p38, respectively. These results for p38 are even more impressive when compared to the considerably poorer results obtained using the best scoring functions currently in use (ChemScore, PLP and Dock). On the basis of correlation coefficient alone, OWFEG also performs better than the other scoring functions in the case of IMPDH and better than all but ChemScore for HIV-1 PR.

Of greater practical importance than the correlation coefficient is the ability of a scoring function to actually separate "good" and "poor" binders, since such a function will ultimately be used to prune a database. For example, the energy cutoff may be set to arbitrarily remove 90% of the compounds in the database. In such a case, the worth of a scoring function is measured by its ability to leave a significantly greater percentage of hits in the set of remaining compounds than were in the original database − the EF. By reference to the thick solid lines in Figure 6, it is clear that OWFEG does an excellent job of enhancing the enrichment ratio. More significantly, OWFEG does an appreciably better job at enhancing the enrichment ratio than any of the other scoring functions tested for both the IMPDH and p38 cases and slightly better in the case of HIV-1 PR. It is also worth noting that two of the scoring functions (ChemScore and PLP) actually do worse than *random* for p38 when judged by the enrichmnent factor criterion.

A particularly satisfying aspect of the OWFEG method is that the scoring function depends on only two fitted coefficients, which determine the step position and height in the function that translates the raw OWFEG data to a scoring function. The same two parameters are used for all three systems to be transferable to other systems. Apart from this, the methodology is completely general: the same simulation protocol and parameter set was used to set up each system. Similarly, solvent water was represented identically in each case; no special provision was made for example the catalytic water in HIV-1 PR. We note that fitted coefficients can present a problem if a general set of coefficients that spans multiple systems cannot be derived. Our ability to make good predictions across three systems using the same minimal set of two fitted coefficients suggests that OWFEG may not suffer this difficulty. However, until the method has been applied to additional systems, we cannot say with this with certainty. In support of the transferability of the parameters, we have recently demonstrated[43] that coefficients parameterized to only two of the data sets (IMPDH and HIV-1 PR) are similar to those parameterized to all three *and* that when the coefficients parameterized on IMPDH and HIV-1 PR are used to make predictions for p38 compounds not included in the work here, the predictions are of comparable quality to what has been seen here.

The OWFEG grids lie somewhere between qualitative and quantitative in nature, and other choices with respect to these calculations have been made in light of this. For example, in the interest of making the OWFEG simulations proceed more quickly, a "belly" was used to restrict the moving atoms of the protein to a 15 Å region about the active site. Similarly, positional restraints have been imposed on the atoms of the protein, to reduce the conformational sampling requirements (test runs without positional restraints on the protein yielded poorer results). In addition, we have neglected ligand solubility in the estimation of relative binding efficacy. That is, the measured binding data reflects the process:

$$\mathrm{P(aq) + I(aq) \rightarrow P{:}I(aq)} \qquad (8)$$

which can be represented as the summation of the simulatable (nonphysical) processes:

$$\mathrm{P(aq) \rightarrow P{:}I(aq)} \qquad (9)$$

and

$$\mathrm{I(aq)} \rightarrow \mathrm{nothing(aq)} \qquad (10)$$

where "P" is the protein and "I" is the inhibitor. The OWFEG grid represents eq 9. In theory, the solvation free energy of the inhibitor in water must also be determined. Here, we have made the assumption that this free energy can be neglected. This assumption was made more out of necessity than anything else: there is no simple, meaningful way to extend the OWFEG scoring approach to a pure solvent box.

We have attempted to include a very simple approximation to the solvation free energy for each binder. This was done by first calculating the free energies for insertion of neutral, positively charged, and negatively charged probe groups in a pure water box, using OWFEG simulations (three separate simulations). Each atom of a binder was assigned a solvation free energy value by linear interpolation based on the partial atomic charge of that atom and the probe group insertion energies. Then the solvation free energies of all the atoms of a binder were summed to give a total solvation free energy correction for the binder molecule. Although inclusion of this term changes the fitted coefficients for the function that converts OWFEG grid points to the scoring function, we find the net results (correlation between OWFEG score and experiment, as well as the qualitative look of the corresponding plots) to be effectively unchanged. Since the correction introduces more variables to the scoring functions without having any significant affect on the quality of the predictions, we did not pursue this further.

One further limitation on the OWFEG is that the OWFEG scoring grid does not provide any contribution arising from molecular internal coordinates (entropy of bond rotation, etc.). And, of course, all of the approximations and limitations inherent in free energy calculations relating to such facets as charge selection and molecular mechanical parameters will also be reflected in the OWFEG results. In particular, we believe that the charge model used in this study (Gasteiger Marselli[42]) could be improved upon.

Despite these admitted approximations and assumptions, the OWFEG scores are quite predictive. It is worthwhile to consider why OWFEG appears to work better than other scoring functions. The answer may well lie in the fact that the OWFEG energies are derived from an MD simulation where the solvent is represented explicitly and allowed to move and where limited flexibility has been allowed to the protein. This differentiates the OWFEG approach from other scoring functions in common use, which either ignore the flexibility of the system or else attempt to correct for it with softer repulsive terms, and where solvent effects, if addressed at all, are included implicitly. As is well-known, residues of the binding site (as well of the protein in general) can undergo significant conformational motion, which can substantively affect both the entropic contributions to the free energy of binding and the effective van der Waals envelope the binder experiences. The beauty of the free energy approach is that all of these factors − flexibility of the active site, enthalpic, *and* entropic effects − are implicitly included in the grid score (albeit at a reduced level in the case of the protein, since positional restraints have been imposed). It may well be the case that the degree to which OWFEG improves on existing methods reflects the inherent flexibility of the active site and surrounding water. If this is the case, then one would expect the IMPDH and p38 sites to exhibit considerably more relevant flexibility than HIV-1 PR. In fact, analysis of both temperature factors and conformations from multiple crystallographic complexes is consistent with this (unpublished data).

It should be noted that while OWFEG does a better job of accounting for protein flexibility, the OWFEG grids can only account for, at best, motions that occur on the nanosecond time scale of the simulations used to generate them. More substantial, long time scale changes, such as substantial rearrangements in the binding site, are not likely to be reflected in the OWFEG grids unless much longer (computationally infeasible) simulations are performed. From the quality of the results obtained, such long-scale motions are apparently not critical in understanding the data sets considered here.

Overall, the OWFEG grids are straightforward to generate, can be implemented using a generally applicable protocol, and have allowed good to exceptionally good results for the test systems studied here. From this, we expect that OWFEG will be a great addition to the suitcase of scoring functions that one applies when performing database screening.

Acknowledgment. We thank Vicki Sato for critical reading of the manuscript and thoughtful contributions.

References

(1) Gschwend, D. A.; Good, A. C.; Kuntz, I. D. Molecular docking towards drug discovery. *J. Mol. Recognit.* **1996**, *9*, 175−186.
(2) Jones, G.; Willett, P. Docking Small-Molecule Ligands Into Active Sites. *Curr. Opin. Biotechnol.* **1995**, *6*, 652−656.
(3) Knegtel, R. M. A.; Wagener, M. Efficacy and selectivity in flexible database docking. *Proteins: Struct. Funct. Genet.* **1999**, *37*, 334−345.
(4) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − An Overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(5) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293−304.
(6) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.* **1996**, *9*, 1−5.
(7) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(8) Kramer, B.; Rarey, M.; Lengauer, T. CASP2 experiences with docking flexible ligands using FlexX. *Proteins* **1997**, Suppl., 221−225.
(9) Rarey, M.; Kramer, B.; Lengauer, T. Multiple automatic base selection: protein−ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 369−384.
(10) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(11) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor-Sites Using a Genetic Algorithm With a Description of Desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(12) *Flexdock*, SYBYL version 6.1; Tripos & Associates: St. Louis, MO, 1998.
(13) Charifson, P. S.; Leach, A. R.; Rusinko, A., III. The generation and use of large 3D databases in drug discovery. *Network Sci.* [electronic publication] **1995**, 1. URL: http://www.awod.com/netsci/Issues/Sept95/feature3.html.
(14) Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: a way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 565−582.

(15) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A method of obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(16) Ajay; Murcko, M. A. Computational Methods to Predict Binding Free Energy in Ligand−Receptor Complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.
(17) Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424−440.
(18) Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235*, 345−356.
(19) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(20) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins: Struct. Funct. Genet.* **1999**, *34*, 4−16.
(21) Pearlman, D. A.; Rao, B. G. Free energy calculations: methods and applications. In *Encyclopedia of Computational Chemistry*; Schleyer, P. V. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, R. P., Eds.; John Wiley & Sons: New York, 1998; pp 1036−1061.
(22) *Modem Theoretical Chemistry;* Segal, G., Ed.; Plenum: New York, 1977; Vols. 7−8: Semiempirical Methods of Electronic Structure Calculation.
(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) Sintchak, M. D.; Fleming, M. A.; Futer, O.; Raybuck, S. A.; Chambers, S. P.; Caron, P. R.; Murcko, M. A.; Wilson, K. P. Structure and mechanism of inosine monophosphate dehydrogenase in complex with the immunosuppressant mycophenolic acid. *Cell* **1996**, *85*, 921−930.
(25) Fleming, M. A.; Chambers, S. P.; Connelly, P. R.; Nimmesgern, E.; Fox, T.; Bruzzese, F. J.; Hoe, S. T.; Fulghum, J. R.; Livingson, D. J.; Stuver, C. M.; Sintchak, M. D.; Wilson, K. P.; Thomson, J. A. Inhibition of IMPDH by mycophenolic acid: dissection of forward and reverse pathways using capillary electrophoresis. *Biochemistry* **1996**, *35*, 6990−6997.
(26) Salituro, F. G.; Baker, C. T.; Court, J. J.; Deininger, D. D.; Li, B.; Novak, P. M.; Rao, B. G.; Pazhanisamy, S.; Porter, M. D.; Schairer, W. C.; Tung, R. D. Design and synthesis of novel conformationally restricted HIV protease inhibitors. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 3637−3642.
(27) Baker, C. T.; Salituro, F. G.; Court, J. J.; Deininger, D. D.; Kim, E. E.; Li, B.; Novak, P. M.; Rao, B. G.; Pazhanisamy, S.; Schairer, W. C.; Tung, R. Design, synthesis, and conformational analysis of a novel series of HIV protease inhibitors. *Bioorg. Med. Chem.* **1998**, *8*, 3631−3636.
(28) Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J. Am. Chem. Soc.* **1995**, *117*, 1181−1182.
(29) Wilson, K. P.; McCaffrey, P. G.; Hsiao, K.; Pazhanisamy, S.; Galullo, V.; Bemis, G. W.; Fitzgibbon, M. J.; Caron, P. R.; Murcko, M. A.; Su, M. S. The structural basis for the specificity of pyridinylimidazole inhibitors of p38 MAP kinase. *Chem. Biol.* **1997**, *4*, 423−431.
(30) 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.
(31) 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.
(32) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Freer, S. T. Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease − Conformationally Flexible Docking by Evolutionary Programming. *Chem. Biol.* **1995**, *2*, 317−324.
(33) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking With Grid-BasedEnergy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.
(34) 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.
(35) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG − A system to select quasi-flexible ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153−174.
(36) Case, D. A.; Pearlman, D. A.; Caldwell, J. C.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C.; Darden, T.; Merz, K. M.; Stanton, R. V.; Cheng, A.; Vincent, J. J.; Crowley, M.; Ferguson, D. M.; Radmer, R.; Seibel, G. L.; Singh, U. C.; Weiner, P.; Kollman, P. A. *AMBER 5.0*; University of California: San Francisco, 1997.
(37) 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.
(38) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comput. Chem.* **1986**, *7*, 230−252.
(39) Vertex Pharmaceuticals Inc., Cambridge, MA.
(40) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M., D.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; Desolms, S. J.; Gaffin, N.; Ghosh, A. K.;Giuliani, E. A.; Graham, S. L.; Guare, J. P.; Hungate, R. W.; Lyle, T. A.; Sanders, W. M.; Tucker, T. J.; Wiggins, M.; Wiscount, C. M.; Woltersdorf, O. W.; Young, S. D.; Darke, P. L.; Zugay, J. A. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J. Med. Chem.* **1995**, *38*, 305−317.
(41) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Prot. Eng.* **1993**, *6*, 723−732.
(42) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3222.
(43) Pearlman, D. A.; Charifson, P. S. Manuscript in preparation.

JM000375V