Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model

Xiaoqin Zou,[†] Yaxiong Sun,[‡] and Irwin D. Kuntz*

Contribution from the Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, San Francisco, California 94143-0446

Received November 30, 1998. Revised Manuscript Received April 1, 1999

Abstract: Accounting for the effect of solvent on the strength of molecular interactions has been a long-standing problem for molecular calculations in general and for structure-based drug design in particular. Here, we explore the generalized-Born (GB/SA) model of solvation (Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−9) to calculate ligand−receptor binding energies. The GB/SA approach allows for the estimation of electrostatic, van der Waals, and hydrophobic contributions to the free energy of binding. The GB/SA formulation provides a good balance between computational speed and accuracy in these calculations. We have derived a formula to estimate the binding free energy. We have also developed a procedure to penalize any unoccupied embedded space that might form between the ligand and the receptor during the docking process. To improve the computational speed, the protein contribution to the electrostatic screening is precalculated and stored on a grid. Refinement of the ligand position is required to optimize the nonbonded interactions between ligand and receptor. Our version of the GB/SA algorithm takes approximately 10 s per orientation (with minimization) on a Silicon Graphics R10000 workstation. In two test systems, dihydrofolate reductase (dhfr) and trypsin, we obtain much better results than the current DOCK (Ewing, T. J. A.; Kuntz, I. D. *J. Comput. Chem.* 1997, 18, 1175−89) force field scoring method (Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 505−24). We also suggest a methodology to identify an appropriate parameter regime to balance the specificity and the generality of the equations.

I. Introduction

It is well-known that the desolvation effect during ligand−protein binding plays a critical role in determining the structure and free energy of the complex. Specifically, water molecules modulate the binding process in two ways: (1) They strongly screen the electrostatic interactions between charged atoms. (2) They contribute to hydrophobic interactions between nonpolar atom groups. The binding free energy is determined by a detailed and delicate balance between ligand−receptor interactions, ligand−water and receptor−water interactions, and water−water interactions in complicated, inhomogeneous environments. As a consequence, computing the solvation energy has been a challenge for structure-based drug design.

During the desolvation process in ligand binding, the change in electrostatic interactions can be divided into three components: partial desolvation of the ligand, partial desolvation of the protein, and screened electrostatic interactions between the bound ligand and protein. In an inhomogeneous medium the electric field depends upon the local environment which is altered with ligand binding. Therefore, determination of the three solvation components requires calculations before and after ligand binding and this computation can be very time consuming.

Numerous efforts have been made to deal with aqueous solutions (see refs 4-9 for reviews). The simplest model is to adjust the dielectric constant, ϵ, in some fashion, typically making ϵ distance dependent.[10−12] Because of its simplicity and its speed of computation, a distance-dependent dielectric term is extensively used in the drug design field, including the current force field scoring of DOCK.[3,13,14] However, such a parameterization has little theoretical justification and does not account for the dependence of charge−charge interactions on the local environment. It also tends to underestimate electrostatic interactions between charges in close proximity, such as those that occur in hydrogen-bonding interactions. Desolvation effects are totally ignored in this approach. Finally, the hydrophobic effect, a critical term in the binding of organic molecules, is not treated in standard molecular force fields.

The most obvious method to overcome these problems is to treat solvent molecules explicitly in molecular dynamics or

[†] Current address: Dalton Cardiovascular Research Center and Department of Biochemistry, University of Missouri, Columbia, MO 65211.

[‡] Current address: Computer-Assisted Drug Design, Bristol-Myers Squibb Company, 5 Research Parkway, Wallingford, CT 06492.

(1) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−9.

(2) Ewing, T. J. A.; Kuntz, I. D. *J. Comput. Chem.* **1997**, *18*, 1175−89.

(3) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 505−24.

(4) van Gunsteren, W. F.; Luque, F. J.; Timms, D.; Torda, A. E. *Annu. Rev. Biophys. Biomol. Struct.* **1994**, *23*, 847−63.

(5) Warshel, A.; Aqvist, J. *Annu. Rev. Biophys. Biophy. Chem.* **1991**, *20*, 267−98.

(6) Harvey, S. C. *Proteins* **1989**, *5*, 78−92.

(7) Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301−32.

(8) Gilson, M. K.; Given, J. A.; Head, M. S. *Chem. Biol.* **1997**, *4*, 87−92.

(9) Leach, A. R. *Molecular Modeling: Principles and Applications*; Longman: Singapore, 1996.

(10) Pickersgill, R. W. *Protein Eng.* **1988**, *2*, 247−8.

(11) Mehler, E. L.; Solmajer, T. *Protein Eng.* **1991**, *4*, 903−10.

(12) Solmajer, T.; Mehler, E. L. *Protein Eng.* **1991**, *4*, 911−7.

(13) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.

(14) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230.

10.1021/ja984102p CCC: $18.00
Published on Web 08/21/1999

Monte Carlo simulations of binding (see refs 15 and 16 for reviews). However, these approaches are currently impractical for screening large numbers of molecules.

Alternatively one can adopt an empirical approach. Binding energies can be calculated by using a set of parameters obtained from a "training set" of known interactions.[17−20] A simple empirical solvation model uses atom- or group-based solvent-exposed area terms.[21] Although this approach is extremely fast, its accuracy is limited by its simplicity. It also relies, critically, on the quality of the training database. The generality of an empirical method is often difficult to establish a priori.

An intermediate approach is to treat the solvent as a continuum dielectric medium.[22,23] Among a variety of such implicit models, the generalized-Born (GB) approximations[1,24,25] provides a good balance between speed and accuracy. According to the GB equation, the electrostatic interaction energy between two charges depends on both the intercharge distance and the effective solvation radii of the charges as a measure of their solvent exposure. This approach can also account for the hydrophobic effect in terms of the change in solvent-accessible surface area (SA) during binding. The GB/SA model accurately predicted solvation free energies (the free energy of transfer from the gas phase to solution) of a wide variety of small molecules and molecular ions.[24−27] Recently the GB/SA model has been successfully applied to study p*K*a shifts in small molecules,[28,29] HIV protease,[28] and the stability of nucleic acid helices.[30]

In this paper, we apply the general GB/SA model to compute ligand binding energies. This is, to our knowledge, the first application of the GB/SA model to study ligand binding. We also explore the valid parameter regime in our free energy scoring model. The parameters used in empirical models can be established by two different approaches. In the first approach, one uses a training set and regression techniques to produce the parameters that give the "best-fit" to the training data. Such models typically excel at interpolations, but it is often hard to predict their range of validity. The second approach selects parameters by examining the underlying physics principles. Such models often have a wide range of applications, but they do less well on fitting any specific training set. These two factors, specificity and generality, may even contain inherent contradictions due to the limitations of the model and data. In this paper, we search for an appropriate parameter regime in our model to find a balance between specificity and generality.

II. Method

1. Overview of the Generalized-Born (GB/SA) Model. Still and co-workers suggested a generalized-Born (GB/SA) model for the solvation of organic molecules.[1,24] In this model, the solvation free energy (G_{sol}) of a molecule consists of three terms: a solvent-solvent cavity term (G_{cav}), a solute−solvent van der Waals term (G_{vdw}), and an electrostatic polarization term (G_{pol}):

$$G_{sol} = G_{cav} + G_{vdw} + G_{pol} \tag{1}$$

The nonelectrostatic terms, G_{cav} and G_{vdw}, are approximated by a linear dependence on the solvent-accessible surface area (SA).[31] That is,

$$G_{cav} + G_{vdw} = \sum_i \sigma_i \mathrm{SA}_i \tag{2}$$

where SA_i is the solvent-accessible surface area of atom i; σ_i is an empirical atomic solvation parameter. σ_i was set to 7.2 cal/mol/Å^2 for all atoms in ref 1 and was nonzero only for nonpolar atoms in ref 24 (varying between 7 to 10 cal/mol/Å^2).

G_{pol}, defined as the change in electrostatic energy when a molecule is transferred from vacuum to solvent, is approximated by the following GB equation

$$G_{pol} = -\frac{1}{2}\left(1 - \frac{1}{\epsilon}\right)\sum_i \sum_j \frac{q_i q_j}{f_{ij}(r_{ij})} \tag{3}$$

where ϵ represents the dielectric constant of solvent ($\epsilon = 78.3$ for water); q_i and q_j represent the charges of atom i and j, respectively; r_{ij} represents the distance between atom i and j; and the function $f_{ij}(r_{ij})$ is defined as $f_{ij}(r_{ij},\alpha_i,\alpha_j) = \sqrt{r_{ij}^2 + \alpha_i\alpha_j e^{-r_{ij}^2/(4\alpha_i\alpha_j)}}$.

Here α_i is defined as the effective Born radius of atom i. It is a measure of the solvent exposure of an atom, approximating the average distance from the atomic charge center to the boundary of the dielectric medium. When an atom is fully exposed to solvent, α_i is its atomic radius. For an atom at the center of a spherical molecule, α_i equals to the radius of the molecule. Generally, α_i depends upon the geometry of all other atoms in the solute molecule but is independent of the solvent dielectric constant and the charge distribution. Formulas to calculate α_i are given in Appendix 1.

The asymptotic behavior of the function f_{ij} is as follows. When $r_{ij} = 0$, eq 3 reduces to the Born equation for superimposed charges. When $r_{ij} \to \infty$, eq 3 becomes the classical Coulomb law. For intermediate r_{ij}, f_{ij} approximates the increase of dielectric constant as a function of r_{ij}. With the introduction of the function f_{ij}, the GB/SA formula attempts to provide a relatively realistic description of the environmental dependence of the dielectric constant. The GB/SA method reported to successfully predict solvation free energies for a wide variety of organic molecules.[24,25]

2. Dielectric Properties of Water in the Receptor Binding Site. In the case of receptor−ligand binding, the evaluation of Born radii is more complicated. Generally, a uniform dielectric medium with an ϵ taken as 78.3 surrounds a solute molecule; the dielectric constant inside the solute is set to 1.[1] However,

(15) Beveridge, D. L.; Mezei, M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431.
(16) Kollman, P. A. *Chem. Rev.* **1993**, *93*, 2395−417.
(17) Bohm, H.-J. *J. Comput.-Aided Mol. Design* **1994**, *8*, 243−56.
(18) Jain, A. N. *J. Comput.-Aided Mol. Design* **1996**, *10*, 427−40.
(19) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. *J. Am. Chem. Soc.* 1996, *118*, 3959−69.
(20) Krystek, S.; Stouch, T.; Novotny, J. *J. Mol. Biol.* **1993**, *234*, 661−79.
(21) Eisenberg, D.; McLachlan, A. *Nature* **1986**, *319*, 199.
(22) Gilson, M. K.; Sharp, K. A.; Honig, B. *J. Comp. Chem.* **1988**, *9*, 327.
(23) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144−9.
(24) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem.* **1997**, *101*, 3005−14.
(25) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122−9.
(26) Scarsi, M.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. B* **1998**, *102*, 3637−41.
(27) Scarsi, M.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. A* **1997**, *101*, 8098−106.
(28) Luo, R.; Head, M. S.; Moult, J.; Gilson, M. K. *J. Am. Chem. Soc.* **1998**, *120*, 6138−46.
(29) Jayaram, B.; Liu, Y.; Beveridge, D. L. *J. Chem. Phys.* **1998**, *109*, 1465−71.
(30) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401−9.
(31) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151−76.

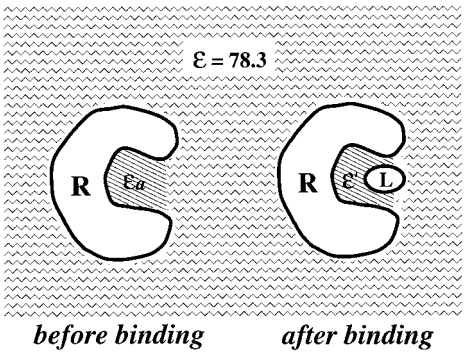

Figure 1. The dielectric properties around the binding site. A uniform dielectric medium with $\epsilon = 78.3$ surrounds the receptor molecule *R* and ligand molecule *L*. The dielectric constant inside the receptor and the ligand is set to 1. In the figure on the left, the dielectric constant of water in the core region of the active site (the shaded region), ϵ_a, is unknown; on the right, the dielectric constant in the unoccupied embedded space (the shaded region), ϵ', is 1. See the text for detail.

the dielectric properties of water in the receptor binding site are more complicated.

Before ligand binding, the dielectric constant of water in the core region of the active site, ϵ_a, is unknown (illustrated by the lightly shaded area in the molecule on the left of Figure 1). The value of ϵ_a varies from 1 for vacuum to 78.3 for bulk water, depending on the hydrophobicity of the site. Unfortunately there is no direct measure of the dielectric constant of water in the active site. The related literature includes simulations of average dielectric constants of the trypsin active site,[32] cytochrome *c*,[33] and many other globular proteins[34] and measurements of the screening of the electrostatic interactions with an α helix.[35,36] Specifically, molecular dynamic (MD) simulations lead to site-dependent average dielectric constants as large as 10 for protein and water groups in reference spheres of 15 Å radii centered in active sites, depends on the actual protein site, and can be as large as 10 in sites of catalytic importance. Other MD simulations showed that the overall dielectric constants of globular proteins range from 11 for myoglobin to 25 for cytochrome *c*, mostly due to the mobility of charged side chains located at the protein surface.[33,34] However, these quantities measure the average screening effect of water and solute molecules instead of the screening of water alone in the site of interest. The ϵ_a issue is discussed again in Section 5, but for simplicity ϵ_a is set to 78.3 for all calculations presented in this paper. The effect of ϵ_a on the Born radii has been accounted for in Appendix 2.

After ligand binding, poorly formed receptor−ligand complexes often contain regions of unoccupied space between the ligand and the receptor (see the lightly shaded area in the molecular complex in Figure 1). Two factors work to lower the local dielectric response in such regions. First, the occupancy by water molecules can be low, especially in hydrophobic pockets. Second, the mobility of water molecules that are present is often reduced due to hydrogen bonding. In either case, the "effective" dielectric constant would be very low. To account for this effect we must modify the original formulas for the Born radii, which assume that spaces that are not occupied by solute atoms are occupied by bulk solvent ($\epsilon = 78.3$ for water). The new formulas, again derived in Appendix 2, penalize the formation of unoccupied embedded volume between the ligand and the receptor.

3. Modifications of the GB/SA Formalism for Nonelectrostatic Contributions. Before we apply the GB/SA formalism to receptor−ligand binding, we need to modify eq 2 to account for the van der Waals interactions between the ligand and the solvent, and the receptor and the solvent. For the van der Waals interactions in the complex, we use the Lennard−Jones 6-12 potential. For the solute-solvent van der Waals interactions in the continuum solvent model, we assume a linear dependence on solvent-accessible surface areas (SA). Thus, for the ligand alone,

$$G_{\mathrm{vdw},L} = -\sigma_2 \sum_i^L \mathrm{SA}_i \tag{4}$$

where σ_2 is the linear coefficient parameter, and the minus sign reflects the attractive feature of the van der Waals interactions between the solute and the solvent. A similar equation is needed for the van der Waals interactions between the receptor and the solvent. For the ligand−receptor complex,

$$G_{\mathrm{vdw},LR} = \beta \cdot \mathrm{VDW} - \sigma_2 \sum_i^{L+R} \mathrm{SA}_i \tag{5}$$

where VDW stands for the Lennard−Jones (L−J) 6-12 potential.[3] The coefficient β is introduced to allow the scaling of the L−J potential to be different from that of the SA term.

For the G_{cav} term, we use a linear approximation proportional to the nonpolar SA[24,25]

$$G_{\mathrm{cav}} = \sigma_1 \sum_i \mathrm{SA}_{hp,i} \tag{6}$$

where (σ_1) is the solvation parameter for all nonpolar atoms, and $\mathrm{SA}_{hp,i}$ is the solvent-accessible surface area of the nonpolar atom *i*. The method we use to compute the SA is given in Appendix 3.

While eqs 5 and 6 both include solvent-accessible surface areas, eq 5 uses the total surface area and eq 6 refers only to the *nonpolar* surface area.

4. Applying the GB/SA Formalism to Ligand−Receptor Binding. We now generalize the GB/SA model of solvation free energy for a single molecule to the desolvation free energy of the ligand−receptor binding process. Notice that the free energy of a molecule X in solvent solvent (G_X^{solvent}) can be calculated as

$$G_X^{\mathrm{solvent}} = G_X^{\mathrm{vacuum}} + G_{\mathrm{sol}}^X \tag{7}$$

where *X* is substituted with (*L*) for the ligand alone, (*R*) for the receptor alone, or (LR) for the ligand−receptor complex. G_X^{vacuum} is the free energy of *X* in vacuum, and G_{so1}^X is the solvation free energy of *X*. In the GB/SA model, G_X^{vacuum} is simply the Coulombic interaction energy in vacuum, and G_{sol}^X is given in eq 1.

Applying eqs 3, 4, 6, and 7, we have

(32) King, G.; Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1991**, *95*, 4366−77.
(33) Simonson, T.; Perahia, D. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *92*, 1082−6.
(34) Simonson, T.; Brooks, C. L., III *J. Am. Chem. Soc.* **1996**, *118*, 8452−8.
(35) Lockhart, D. J.; Kim, P. S. *Science* **1992**, *257*, 947−51.
(36) Lockhart, D. J.; Kim, P. S. *Science* **1993**, *260*, 198−202.

$$G_L^{\text{solvent}} = \frac{1}{2}\sum_i^L\sum_{j\neq i}^L \frac{q_i q_j}{r_{ij}} + \left(\sigma_1\sum_i^L \text{SA}_{hp,i} - \sigma_2\sum_i^L \text{SA}_i - \frac{1}{2}\left(1-\frac{1}{\epsilon}\right)\sum_i^L\sum_j^L \frac{q_i q_j}{f_{ij}^L(r_{ij})}\right) \quad (8)$$

for the ligand and a similar equation for the receptor. For the ligand−receptor complex, we use eqs 3, 5, 6, and 7:

$$G_{LR}^{\text{solvent}} = \frac{1}{2}\sum_i^{L+RL+R}\sum_{j\neq i} \frac{q_i q_j}{r_{ij}} + \left(\sigma_1\sum_i^{L+R} \text{SA}_{hp,i} + \beta\cdot\text{VDW} - \sigma_2\sum_i^{L+R} \text{SA}_i - \frac{1}{2}\left(1-\frac{1}{\epsilon}\right)\sum_i^{L+RL+R}\sum_j \frac{q_i q_j}{f_{ij}^{LR}(r_{ij})}\right) \quad (9)$$

We now compute the binding free energy G_{binding}. By definition,

$$G_{\text{binding}} = G_{LR}^{\text{solvent}} - G_L^{\text{solvent}} - \text{G}_R^{\text{solvent}} \quad (10)$$

Substituting eq 10 with eqs 8 and 9 and expanding the summations over atoms, we have

$$G_{\text{binding}} = \sigma_1\Delta(\text{SA}_{hp}) + \beta\cdot\text{VDW} - \sigma_2\Delta(\text{SA}) + G_{\text{POL}} \quad (11)$$

where

$$G_{\text{POL}} = G_{\text{screened es}} + G_{L\ \text{desolve}} + G_{R\ \text{desolve}}$$

Here $\Delta(\text{SA}_{hp})$ and $\Delta(\text{SA})$ denote the change in the hydrophobic and total solvent-accessible surface area due to ligand binding; while $G_{\text{screened es}}$ is the screened ligand−receptor electrostatic energy, $G_{L\ \text{desolve}}$ is the electrostatic polarization energy due to desolvating part of the ligand (the partial desolvation energy of the ligand), and $\text{G}_{R\ \text{desolve}}$ represents the partial desolvation energy of the receptor. The last three components are calculated by

$$G_{\text{screened es}} = \sum_i^L\sum_j^R \frac{q_i q_j}{r_{ij}} - \left(1-\frac{1}{\epsilon}\right)\sum_i^L\sum_j^R \frac{q_i q_j}{f_{ij}^{LR}(r_{ij})} \quad (12)$$

$$G_{L\ \text{desolve}} = \frac{1}{2}\left(1-\frac{1}{\epsilon}\right)\sum_i^L\sum_j^L\left(\frac{q_i q_j}{f_{ij}^{L}(r_{ij})} - \frac{q_i q_j}{f_{ij}^{LR}(r_{ij})}\right) \quad (13)$$

$$G_{R\ \text{desolve}} = \frac{1}{2}\left(1-\frac{1}{\epsilon}\right)\sum_i^R\sum_j^R\left(\frac{q_i q_j}{f_{ij}^{R}(r_{ij})} - \frac{q_i q_j}{f_{ij}^{LR}(r_{ij})}\right) \quad (14)$$

At large distances (r_{ij}), terms within each component of G_{POL} cancel. By clustering these components, it is evident that a long-range interaction problem reduces to a short-range problem. As a result, we can use relatively short cutoffs in the energy calculations.

5. Implementation of the GB/SA Model in DOCK. We have implemented the above GB/SA scheme in the DOCK software package,[2,37] providing an additional independent scoring function—the "free energy score". To improve the computational speed, we have adopted the following three strategies:

(1) The receptor contributions to electrostatic screening are precalculated and stored on a grid. These contributions include the solvation energy term for the receptor alone (i.e., the first summation term in eq 14) and the effect of the receptor atoms on the Born radii α_is of the atoms in the receptor and the ligand (in f^{LR} of eqs 12−14). The precalculation is limited because eqs 12−14 are not pairwise. Therefore, f^L and contributions from ligand atoms to f^{LR} are calculated during ligand docking.

(2) The effective Born radii α_is are determined by the geometrical relationship of ligand and receptor atoms. However, receptor atoms away from the binding site have little effect and can be ignored. To accomplish this, two boxes are used for the precalculation (Figure 2): a smaller grid box containing receptor atoms that are involved in the desolvation process of a ligand, and a larger auxiliary box to account for the receptor atoms that contribute only to α_is of the receptor atoms that lie in the grid box near the boundary regions. Receptor atoms outside the auxiliary box have little effect on the ligand binding and are ignored. All terms in the solvation energy that are independent of ligand features are precalculated and saved on grid points in the grid box for docking use. The receptor atoms between the auxiliary box and the grid box are used only once to establish the α_is of the receptor atoms inside the grid box; these calculations are not repeated during ligand docking.

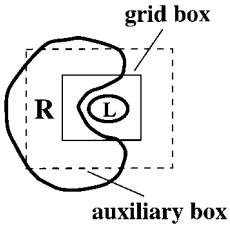

Figure 2. Schematic diagram of docking a ligand (*L*) to a receptor (*R*), illustrating how grids are defined for evaluating the free energy scoring function. Two boxes are used for the calculation: a grid box which contains protein atoms involved in the desolvation process with ligands, and an auxiliary box. The auxiliary box defines a rectangular shell containing receptor atoms that contribute to the Born radii of the receptor atoms just inside the grid box. It is larger than the grid box in each dimension by an electrostatic cutoff distance. All terms in the solvation energy that are independent of ligand features are precalculated and saved on grid points in the grid box; but the receptor atoms between the grid box and the auxiliary box are used only once to calculate the Born radii of receptor atoms.

(3) A short distance cutoff in energy calculations is used, taking advantage of the cancellation effect discussed at the end of Section 4.

As with the current DOCK force field scoring,[38] we find it useful to optimize the ligand position during free energy scoring. However, complete optimization of ligand−receptor interactions is very time consuming. As a compromise, σ_is and the partial desolvation energies of the ligand and of the receptor ($G_{L\ \text{desolve}}$ and $G_{R\ \text{desolve}}$ in eqs 13 and 14) are not updated during minimization steps. This has been shown to be a good approximation.

To test the robustness of our free energy scoring scheme, we varied parameters, such as grid spacing, distance cutoff, and size of the grid box. In Appendix 4, we show how these parameters affect the calculated binding free energies of (1) the dihydrofolate reductase (dhfr)−methotrexate (MTX) complex (4dfr) and (2) the benzamidine−trypsin (bovine pancreas trypsin) complex.

From Table 5 in Appendix 4, we conclude that a distance cutoff of 5−8 Å and a grid spacing of 0.4 Å provides a good balance between speed and accuracy. Generally, we set the grid

(37) Ewing, T. J. A. Thesis, 1997.

(38) Gschwend, D. A.; Kuntz, I. D. *J. Comput.-Aided Mol. Design* **1996**, *10*, 123−32.

spacing to be 0.4 Å and the cutoff to be 8 Å. The size of the grid box depends on the shape of the active site. The grid box should enclose all spheres generated with SPHGEN[3] and receptor atoms within the cutoff distance from these spheres. It is possible that some long ligand molecules contained within a database will partly fall out of the grid box. The portions of ligands that are outside the box are assumed to be screened by water and their contributions to the binding free energies are ignored.

We also studied the dependence of G_{binding} on the dielectric constant of water in the active site before binding (ϵ_a). In fact, only the first term in $G_{R\ \text{desolve}}$ (see eq 14), is affected by ϵ_a, and it is calculated only once before a database search for a given receptor. Furthermore, G_{binding} is insensitive to ϵ_a for $\epsilon_a > 5$ (which is very likely[32−34]). For simplicity, we set $\epsilon_a = 78.3$ for the rest of this paper.

6. Optimization for the Parameter Set (β,σ_1,σ_2). In structure-based drug design, an important goal of any binding energy calculations is the ability to rank inhibitors correctly. The predicted binding energies should have good correlation with the experimental measurements, and the percentage of the predicted hits should be high. It is also desirable, if the model purports to describe binding free energies, to predict binding energies in the physically reasonable range. For example, a good inhibitor would be expected to have an energy in the range of −6 to −20 kcal/mol. For the remaining molecules in databases, the binding energies should be more positive. We will use these two criteria to determine the appropriate parameter regime (β,σ_1,σ_2) for our free energy scoring function.

Specifically, we define two types of error functions. For inhibitors with known binding affinities, we define the error function as

$$\text{Err}^1 = \frac{1}{N}\sum_{i=1}^{N}(G_{\text{bind},i}^{\text{pred}} - G_{\text{bind},i}^{\text{expt}})^2 \qquad (15)$$

where $G_{\text{bind},i}^{\text{pred}}$ and $G_{\text{bind},i}^{\text{exp}}$ represent the calculated and measured binding energies of each inhibitor and N represents the total number of known inhibitors in the training set.

For molecules in a random database which are unlikely to be good inhibitors, we define the error function for a particular database search as

$$\text{Err}^2 = \frac{1}{M}\sum_{i=1}^{M}(G_{\text{bind},i}^{\text{pred}} - G_{\text{bind}}^{\text{thres}})^2 \qquad (16)$$

where $G_{\text{bind}}^{\text{thres}}$ represents the most negative binding energies considered to be reasonable for these molecules, and M represents the total number of molecules in the database that have lower binding energies than $G_{\text{bind}}^{\text{thres}}$. The appropriate parameter regime should optimize all these error functions simultaneously. In this paper, we arbitrarily take $G_{\text{bind}}^{\text{thres}}$ to be −5.5 kcal/mol, e.g., ca. 100 μM inhibition constants. The issue of parameter optimization is discussed further in Section III.2.

III. Results

1. Test for the Effect of the Location of Polar/Charged Groups on G_{pol}. We have implemented the GB/SA model as an alternative scoring method in the DOCK program.[2,37] The first test of this scoring method examines a ligand containing a solvent exposed polar or charged functional group. Such a group should contribute little to the electrostatic interaction with a receptor as compared to the same functional group when embedded in a solvent excluded binding site. This distinction cannot be made by a simple Coulombic force field[3] with a distance dependent dielectric constant. The crystal structure of the dhfr−MTX complex (4dfr) was used (see Appendix 4) for this test. The MTX was assumed to be protonated in the bound state.[39] The grid spacing and the cutoff were set to 0.4 and 5 Å, respectively (Section II.5). Each of the following artificial modifications were made to methotrexate in the minimized dhfr−MTX crystal structure: The γ- and α-carboxylate groups were removed, MTX was deprotonated, and both amino groups were removed (Figure 3). To allow a direct comparison, we did not minimize the ligand structures for each modification. We give only the polarization energy results in Table 1 to focus on the electrostatic interactions. For comparison, we also scored these ligand molecules with the DOCK force field[3] (grid spacing of 0.3 Å and a distance cutoff of 10 Å). Because of the sensitivity of the force field score to ligand orientations, orientational minimization was performed for this case. These results are also given in Table 1.

Intuitively, we expect that removing the γ-carboxylate group would resulted in an insignificant change of the electrostatic interaction with the complex as this charged group is quite exposed to water. This is consistent with the calculated results. The slight decrease of G_{POL} may result from the absence of several structural water molecules that form hydrogen bonds with the γ-carboxylate group. We also expect that removing an embedded charged group (e.g., deprotonating MTX or removing α-carboxylate group) or polar groups (e.g., removing both amino groups) will decrease $G_{\text{screened es}}$ and therefore raise G_{POL}. Both modifications are unfavorable for binding, in agreement with the calculated results in Table 1. In contrast, the force field scoring results do not give the expected results.

In conclusion, the free energy scoring function correctly differentiated the desolvation effects of charged/polar groups at different positions in a ligand−receptor complex.

2. Rank Ordering of Binding Affinities. We next ask whether the solvation calculation can produce a sensible rank ordering of known enzyme inhibitors seeded into a database for two enzymes: dihydrofolate reductase (dhfr), and trypsin.

We set the grid spacing to be 0.4 Å and the distance cutoff to be 8 Å. Our implemented GB/SA algorithm takes about 10 s per orientation (with minimization) on a Silicon Graphics Octane workstation. Because this is not fast enough for screening the complete Available Chemicals Directory (ACD, distributed by Molecular Design Ltd., San Leandro, CA), we use the free energy scoring as a post-DOCK screening. Thus, we first use DOCK to identify the 10 000 top force field scoring molecules from the ACD and then carry out the GB calculations to rank these candidates.

We also scored a set of known inhibitors for comparison. We identified the physical parameter regime by optimizing the average of the three error functions for dihydrofolate reductase (dhfr), trypsin, and the set of known inhibitors. We also tested the capability of our free energy scoring function to select the right conformations of a binding ligand out of a variety of possible conformations.

2.1. Dihydrofolate Reductase and Trypsin. The best-known inhibitors of dihydrofolate reductase are protonated methotrexate (MTX)[40] and protonated trimethoprim (TMP)[39] (Figure 4). Starting with the crystal using the dhfr−MTX (4dfr) and dhfr−

(39) Matthews, D. A.; Bolin, J. T.; Burridge, J. M.; Filman, D. J.; Volz, K. W.; Kaufman, B. T.; Beddell, C. R.; Champness, J. N.; Stammers, D. K.; Kraut, J. *J. Biol. Chem.* **1985**, *260*, 381−91.

(40) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. *J. Biol. Chem.* **1982**, *257*, 13650−62.

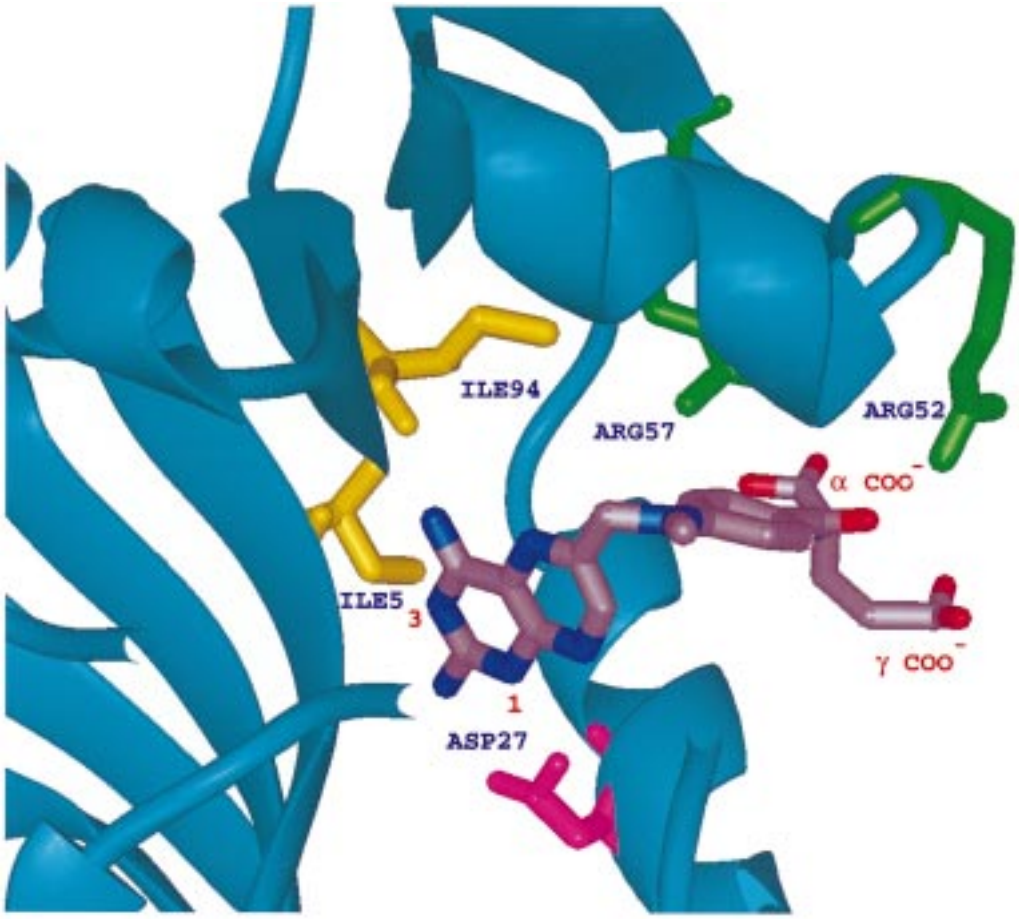

Figure 3. Bound dhfr−MTX crystal structure.[39] The important functional groups of MTX and residues making hydrogen bonds with the ligand are labeled.

Table 1. Effects of the Locations of Ligand Polar Groups on Polarization Free Energy

	force field score	G_{POL} (kcal/mol)	$G_{screened\ es}$ (kcal/mol)	$G_{L\ desolve}$ (kcal/mol)	$G_{R\ desolve}$ (kcal/mol)
(a) MTX	−71.6	−5.1	−79.6	46.2	28.3
(b) removing γ-COO^-	−64.2	−7.6	−78.1	42.5	27.9
(c) removing α-COO^-	−52.5	−1.3	−50.5	27.4	21.8
(d) removing both NH_2	−67.1	1.6	−74.1	47.4	28.3
(e) deprotonating MTX	−62.8	17.6	−36.1	25.4	28.3

TMP (0dfr)[41] complexes, the four terms in $G_{binding}$ (see eq 11), G_{POL}, $\Delta(SA_{hP})$, VDW, and $\Delta(SA)$, were computed. This calculation was repeated with unprotonated MTX. A more difficult test is to properly rank a small ligand molecule because additive force fields favor large molecules.[42] We therefore tried to dock the most important functional group in MTX, the pteridine ring, alone (Figure 4). Both the unprotonated and protonated pteridine ring were tested and the results are listed in Table 2.

We also tested a series of known inhibitors for trypsin: benzamidine, APPA, TAPAP, and NAPAP. These inhibitors were scored using the bovine trypsin crystal structures: benzamidine−trypsin (3ptb),[43] APPA−trypsin (ltpp),[43,44] TAPAP−trypsin (lpph),[45] and NAPAP−trypsin (lppc).[45] The embedded

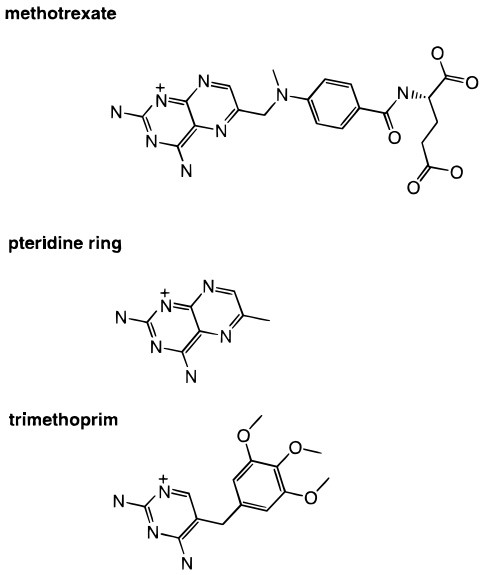

Figure 4. Chemical structures of methotrexate, pteridine ring, and trimethoprim.

(41) The X-ray structure of 0dfr−TMP was given by D. A. Matthews via personal communication.

(42) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. *Proteins* **l999**, *34*, 4−16.
(43) Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. *Acta Crystallogr.* **1983**, *39*, 480−90.
(44) Walter, J.; Bode, W. *Hoppe-Seyler's Z. Physiol. Chem.* **1983**, *364*, 949−59.
(45) Turk, D.; Sturzebecher, J.; Bode, W. *FEBS Lett.* **1991**, *287*, 133−8.

Table 2. Free Energy Scores and Rankings of MTX and TMP Docked to dhfr

	ΔG_{exp} (kcal/M)	force field score	rank	G_{POL} (kcal/M)	VDW (kcal/M)	Δ(SA) (Å^2)	Δ(SA$_{hp}$) (Å^2)	parameter set 1[a]		parameter set 2[a]	
								$G_{binding}$ (kcal/M)	rank	$G_{binding}$ (kcal/M)	rank
MTX	−11.7	−70.0	7	−6.7	−32.9	−915	−500	−11.9	2	−20.7	1
TMP	−12.1	−28.9	2402	−2.8	−18.6	−714	−486	−12.6	1	−11.8	2
MTX (unprot.)		−62.1	16	9.3	−27.1	−913	−496	5.2	451	−1.1	9
pteridine		−29.7	1943	−3.8	−16.5	−461	−225	−4.4	16	−10.1	3
pteridine (unprot.)		−23.7	6440	20.6	−20.4	−460	−223	19.5	6299	12.0	1940

[a] Parameters used: set 1 (β, σ_1, σ_2) = (0.151, 0.0719, 0.0391); set 2 (β, σ_1, σ_2) = (0.6, 0.025, 0.02). See text and eq 11 for detail.

Table 3. Free Energy Scores and Rankings of Benzamidine, APPA, TAPAP, and NAPAP Docked to Trypsin

	ΔG_{ex} (kcal/M)	force field score	rank	G_{POL} (kcal/M)	VDW (kcal/M)	Δ(SA) (Å^2)	Δ(SA$_{hp}$) (Å^2)	parameter set 1[a]		parameter set 2[a]	
								$G_{binding}$ (kcal/M)	rank	$G_{binding}$ (kcal/M)	rank
benzamidine	−6.4	−28.9	2988	−4.0	−21.4	−406	−246	−9.0	10	−14.9	3
APPA	−7.9	−33.0	711	−6.6	−18.5	−504	−244	−7.3	19	−13.7	4
TAPAP	−8.0	−43.2	36	4.6	−34.6	−743	−481	−6.1	23	−13.3	5
NAPAP	−8.4	−50.5	3	2.5	−36.5	−855	−540	−8.5	13	−15.9	1

[a] Parameters used: set 1 (β, σ_1, σ_2) = (0.151, 0.0719, 0.0391); set 2 (β, σ_1, σ_2) = (0.6, 0.025, 0.02). See text and eq 11 for detail.

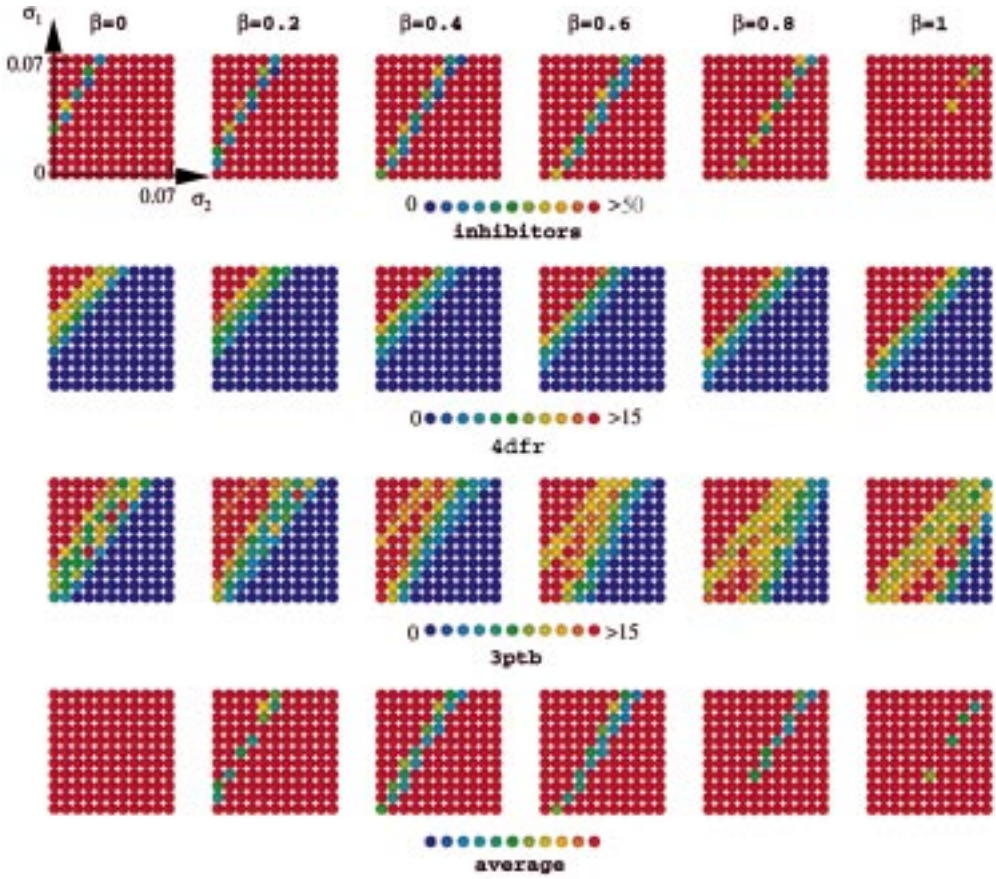

Figure 5. Physical parameter search. For each panel, the vertical and horizontal axes represent σ_1 and σ_2 in the unit of kcal/mol/Å^2. The panels show results for different β. The colors represent the magnitude of the error function, with blue for low and red for high. The first row shows the fitting of the inhibitor set (eq 15). The second and third rows display the error from the database search for dhfr and trypsin, respectively (eq 16). The last row shows the optimized error function, which is defined as the average value of the rescaled error functions in the first three rows with the rescaling factors of 50, 15, and 15, respectively.

volume (Figure 1) is defined as the space occupied by the benzamidine. Results are given in Table 3.

Next, we calculated $G_{binding}$ for the best scoring 10 000 molecules from the ACD for dhfr and trypsin as determined by the standard DOCK force field score. Using these database results with the training set of the six known inhibitors, we tried to identify the appropriate parameter regime for β, σ_1,σ_2. Only the physically reasonable ranges of 0−1 for β and 0−0.07 kcal/mol/Å^2 for σ_1 and σ_2 were explored. We varied the parameters and calculated the resulting three error functions using eqs 15 and 16. The predicted binding energies for each molecule in the training set were not allowed to be more positive than G_{bind}^{pred} even if Err[1] was low. The results are plotted in Figure 5. The errors are represented in colors, with blue for low error and red for high error values. The top three rows are for the training set, dhfr database search, and trypsin database search, respec-

tively. The appropriate parameter regime should optimize all three error functions, as shown in the blue and green regions in the bottom row.

We observe from the bottom row in Figure 5 that parameter β is weakly constrained between 0.2 and 0.8 while σ_1 and σ_2 show strong covariance over the allowable parameter range. Though σ_1 and σ_2 vary linearly, it is not possible to remove either parameter because of the shift in the linear relationship depending on the value of β.

To illustrate these results, we considered two sets of (β,σ_1, σ_2) in the ideal region. The first set of parameters, (0.151, 0.0719, 0.0391), yields the best fit for the binding energies for the six inhibitors. The predicted binding energies, experimental data, and the corresponding ranking in the dhfr database (for MTX and TMP) or trypsin database (for the trypsin inhibitors) are listed in Tables 2 and 3. TMP and MTX rank no. 1 and no. 2 among top scoring 10 000 ACD molecules for dhfr. As a comparison, the force field function also gives a high score to MTX (7), but poorly ranks TMP (2402). Furthermore, Table 2 shows how important protonation is for binding energies. The pteridine ring, a small ligand, moves from 6299 (deprotonated) to 16 (protonated), consistent with the experimental findings.[39] Force field scoring ranks both poorly (6440 and 1943, respectively) because it favors large molecules.[42] Finally, the same set of parameters gives a good fit to the measured binding energies of the four trypsin inhibitors, although it did a less satisfactory job in ranking these inhibitors among the top 10 000 ACD molecules for trypsin with the free energy scoring. The rankings of benzamidine, APPA, TAPAP, and NAPAP are 10, 19, 23, and 13, respectively, as shown in Table 3. In comparison, the force field scoring ranks NAPAP well (3) but not the others.

Importantly, the above set of parameters is clearly not a unique choice. For example, we could take (β,σ_1,σ_2) to be (0.6, 0.025, 0.02), a set of parameters chosen from the blue and green regions in the bottom row of panels in Figure 5. This set raises the contributions from the VDW term and lowers the impact of the surface terms. Though this set gives less satisfactory binding energy predictions (shown in Tables 2 and 3), it does a better job on database searching, ranking the protonated MTX and TMP the first and second position for dhfr; and ranking NAPAP, benzamidine, APPA, and TAPAP 1, 3, 4, and 5 for trypsin. In summary, there is no single set of "best" parameters for the data at hand. Rather, a strong covariance of all these parameters allows a range of "reasonable" choices, implying that the parameter choices are undetermined.

2.2. Orientational Test. Another criterion for a successful scoring function is the ability to act as a sieve for good ligand orientations (e.g., orientations near the X-ray position) out of a sampling of orientations. As a simple test, we generated the top 10 orientations with DOCK4.0 by performing flexible docking with the dhfr−MTX complex. Specifically, we started with CONCORD-generated coordinates of MTX[46] and docked it to the X-ray position of dhfr. We chose 1000 anchor orientations and 50 pruned configurations for orientational sampling.[37] Each orientation was scored with the force field function. The top 10 best scoring results are presented in Table 4. Among these orientations, only no. 3 is close to the crystal structure of MTX.

These 10 orientations were then re-scored with our free energy scoring function allowing optimization. We used both sets of parameters in Section 2.1. The results are also given in Table 4. Both parameter sets successfully identified the correct ligand orientation. The RMSD (root mean squared deviation) between this orientation and the crystal structure of the bound MTX is 1.02 Å, while the other orientations have RMSDs ranging from 2.73 to 3.17 Å.

Table 4. Free Energy Scores for the Conformations of MTX from Flexible Docking

orientations	force field score	$G_{binding}$ (set 1)[b] (kcal/mol)	$G_{binding}$ (set 2)[b] (kcal/mol)
1	−60.1	32.4	18.4
2	−59.4	25.7	13.5
3[a]	−59.3	−0.3	−10.8
4	−57.8	23.5	11.8
5	−57.7	26.1	13.1
6	−57.5	27.0	13.4
7	−57.2	22.5	9.6
8	−55.4	21.3	10.2
9	−55.3	22.7	9.9
10	−55.2	31.3	19.6

[a] The conformation (no. 3) is closest to the crystal structure (RMSD= 1.2 Å). [b] Parameters used: set 1 (β, σ, σ_2) = (0.151, 0.0719, 0.0391); set 2 (β, σ_1, σ_2) = (0.6, 0.025, 0.02). See text and eq 11 for detail.

IV. Discussion

We have shown that it is feasible to use a continuum model to develop a "solvation correction" for electrostatic interactions that is both accurate enough and rapid enough to rank order 10 000 complexes per day of computation on a workstation. Other approaches, using either a simple Coulombic force field[3,19,20,47] or ad hoc assumptions,[17,18,21] tend to underestimate electrostatic interactions or to have difficulty in accurately estimating the electrostatic contribution for complicated geometries. Furthermore, desolvation of charged centers was not considered in these approaches. A simple and fast model was suggested recently to crudely estimate the "solvation correction".[42] The extensions to the generalized-Born model proposed here allow physically reasonable first-order corrections to be applied with a modest computational effort.

We also show that other contributions to binding free energies can be added to a force field leading to estimates of receptor−ligand binding energies that are in reasonable agreement with experiments. Specifically, we have added two nonelectrostatic terms: the change of the van der Waals interactions of the system and a surface area term accounting for the hydrophobic effect. The conformational entropy contributions, which are usually approximated by a constant term associated with the loss of translational and rotational freedom of the ligand, and two terms associated with loss of conformational entropies, respectively proportional to the numbers of rotatable bonds and of the heavy atoms,[17−20,47] have not been considered in this paper, and will be briefly discussed below.

Lastly, we show that even for widely used models of free energy there are unlikely to be unique sets of parameters if the model involves any non ab initio parameters. We find that mapping reasonable parameter space yields several parameter regimes that lead to comparable answers. Optimizing parameters only for a training set of known inhibitors is not sufficiently rigorous. One strategy to seek a more robust parameter set is to optimize simultaneously the parameters which discriminate against molecules in databases that are unlikely to contain good inhibitors. This procedure should produce parameters which minimize "false positives".

There are differences between the GB/SA scheme for solvation of a single molecule and our generalized GB/SA model

(46) Rusinko, A.; Sheridan, R. P.; Nilakantan, R.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. *J. Chem. Inform. Comp. Sci.* **1989**, *29*, 251−5.

(47) Bardi, J. S.; Luque, I.; Freire, E. *Biochemistry* **1997**, *36*, 6588−96.

Table 5. Effects of Grid Spacing and Cutoff on the Polarization Free Energy Term

cutoff (Å)	grid spacing (Å)	$G_{\text{screened es}}$ (kcal/mol)	$G_{L\ \text{desolve}}$ (kcal/mol)	$G_{R\ \text{desolve}}$ (kcal/mol)	G_{POL} (kcal/mol)	CPU[a] (s)
(A) Bound Crystal Structure of dhfr−MTX						
5	0.5	−32.6	25.5	28.6	21.5	1.0
	0.4	−36.1	25.4	28.1	17.4	1.6
	0.3	−35.3	25.0	28.2	17.8	3.7
8	0.5	−37.8	29.1	27.1	18.4	4.9
	0.4	−40.7	29.2	27.4	15.9	9.4
	0.3	−41.0	29.4	27.4	15.8	30.9
10	0.5	−40.4	30.3	27.2	17.2	11.3
	0.4	−43.4	30.5	27.6	14.7	26.3
	0.3	−42.3	30.9	27.6	16.2	98.6
(B) Trypsin−Benzamidine						
8	0.5	−40.9	21.5	17.5	−2.0	2.7
	0.4	−39.8	20.5	17.2	−2.2	5.2
	0.3	−39.3	20.9	16.4	−2.0	13.1
10	0.5	−41.2	22.3	16.1	−2.7	7.3
	0.4	−39.2	21.3	15.7	−2.9	14.8
	0.3	−37.6	20.8	15.1	−1.7	59.3

[a]Silicon Graphics Octane.

for estimation of ligand binding energies. First, obviously, the electrostatic contribution to binding free energies consists of three terms; the screened ligand−receptor electrostatic energy, the partial desolvation energy of the ligand, and the partial desolvation energy of the receptor. In the single molecule case, there is only one term for the solvation energy. Second, there is the possibility of an unoccupied space between the ligand and the receptor. Such a cavity may be unfavorable for occupation by water, leading to a low dielectric unoccupied volume; a phenomenon absent in the case of solvation of a single molecule. Third, the changes of the van der Waals interactions of the system before and after binding are more subtle than that of a single molecule before and after solvation. We used the Lennard−Jones 6-12 potential to characterize the solute−solute van der Waals interactions and assumed the solute−solvent van der Waals interactions to be linearly related to the solvent-accessible surface area (SA). In addition, uncertainties associated with the original GB/SA model also remain. These include the validity of continuum models for macromolecules and the required reparameterization for different force field energy functions.[29] Recent applications on HIV protease[28] and in particular nucleic acid helices[30] show promising results. Ideally, one may use the charge sets derived from electrostatic potential fitting to ab initio wave functions[29,48] (e.g., RESP[49] or CM1[50,51]). Unfortunately these sets are usually sensitive to the ligand structures. The ultimate resolution of these matters are for future studies.

Several lines of investigation remain open for further exploration. The computational speed of the solvation algorithm needs to be improved if we are to screen, directly, a large database. This requires more efficient programing of the grid calculations and a way to convert the non-pairwise feature of effective Born radii calculations to some approximate pairwise calculations as in refs 24, 25, 29, and 52. Currently, nearly 99% of the total CPU time for binding energy calculations is spent on the polarization term calculations. Pairwise calculations of effective Born radii will increase the computational speed dramatically. Also, we need to improve the robustness of the calculated binding energies to ligand structure variations. One may attempt to solve this problem (1) by taking into account the dielectric constant of proteins[23] to partly screen the electrostatic interactions between ligand and receptor and (2) by applying the soft-core approximation[53] to the GB force field to avoid singularities at close atom−atom distances.

Another difficult but important problem that remains to be addressed is the generalization of the enthalpic GB energy to free energy. Calculation of the entropic contribution to free energy is a long-standing bottleneck for accurately estimating the binding energy. The entropy loss upon binding involves translational, rotational, and conformational entropy changes which all depend on the tightness of binding and the flexibility of the ligand.[54] The conformational sampling issue is critical. A smart sampling algorithm is needed to save computational time, e.g., predominant states[55] or configurational bias Monte Carlo[56,57] methods.

In summary, our free energy scoring function can be applied in two ways. First, it can be used as a post-DOCK screening of a large database. In this scenario, the top 10000−100000 compounds/orientations are selected by rigid/flexible/combinatorial docking using the force field scoring function. This subset of the database is then to re-ranked with GB free energy scoring. Second, the free energy scoring function can be applied directly to rigid/flexible docking of a small database.

The authors are grateful to helpful discussions with Todd Ewing, Geoffrey Skillman, Ken Brameld, Connie Oshiro, Makino Shingo, Kaiqi Chen, Michelle Lamb, and David

(48) Reddy, M. R.; Erion, M. D.; Agarwal, A.; Viswanadhan, N.; McDonald, D. Q.; Still, W. C. *J. Comput. Chem.* **1998**, *19*, 769−80.
(49) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620.
(50) Hawkins, G. D.; Lynch, G. C.; Giesen, D. J.; Rossi, I.; Storer, J. W.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. AMSOL-version 5.4, QCPE program 606. *QCPE Bull.* **1996**, *16*, 11.
(51) Hawkins, G. D.; Giesen, D. J.; Lynch, G. C.; Chambers, C. C.; Rossi, I.; Storer, J. W.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. AMSOL-version 6.1.1. Oxford Molecular Group: London 1997.
(52) Dominy, B. N.; Brooks, C. L., III *J. Phys. Chem. B* **1999**, *103*, 3765−73.
(53) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529−39.
(54) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *101*, 1047−69.
(55) Head, M. S.; Given, J. A.; Gilson, M. K. *J. Phys. Chem.* **1997**, *101*, 1609−18.
(56) Frenkel, D.; Mooij, G.; Smit, B. *J. Phys.-Condens. Matter* **1992**, *4*, 3053−76.
(57) Dodd, L.; Boone, T.; Theodorou, D. *Mol. Phys.* **1995**, *78*, 961−96.
(58) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535.
(59) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. Crystallographic databases: information content, software systems, scientific applications. In *Data Commission of the International Union of Crystallography*; Allen, F. H., Bergerhoff, G., Seivers, R. Eds.; Bonn, 1987; pp. 107−32.

Sullivan. We also gratefully acknowledge Dr. D. A. Matthews for providing us with the crystal structure of the bound dhfr−TMP (Odfr) complex. We thank the UCSF Computer Graphics Laboratory for use of the Midas molecular graphics software. This work is supported by NIH grants GM31497 and GM56531 (P. O. Montellano, Principal Investigator) and by the Daiichi Corporation.

Appendix 1: Method for Numerical Calculation of α_i (the Effective Born Radius of Atom *i*)

We used a slightly different method from Still et al.'s shell-based algorithm[1] to calculate α_i, instead following Scarsi et al. in using a grid-based approach.[27] The grid-based method is more efficient for ligand binding systems in which a macromolecule is involved and the atomic Born radii vary before and after binding. This efficiency is associated with precalculating contributions of the receptor atoms to α_is of atoms in the receptor and the ligand.

Still et al. define α_i, the effective Born radius of atom *i*, as the atomic radius which would give the actual electrostatic energy of the molecule−dielectric system by the Born equation if all other atoms in the molecule were uncharged (i.e., these atoms simply serve to displace the solvent dielectric medium).[1] For an isolated atom *i*, α_i is just the atomic radius a_i. The atomic radius is calculated from the van der Waals radius,[1] $a_i = R_{VDW,i} - 0.09$). For a polyatomic solute, α_i is defined by

$$G_{pol,i} = G_i^{solvent} - G_i^{vacuum} = -\frac{1}{2}\left(1-\frac{1}{\epsilon}\right)\frac{q_i^2}{\alpha_i} \qquad (17)$$

where ϵ represents the dielectric constant of solvent ($\epsilon = 78.3$ for water) and q_i represents the atomic charge.

Notice that

$$G_i^{vacuum} = \frac{q_i^2}{2a_i}$$

and

$$G_i^{solvent} = \frac{q_i^2}{2\epsilon a_i} + \frac{q_i^2}{2}\left(1-\frac{1}{\epsilon}\right)\frac{dV}{4\pi}\sum_k^V \frac{1}{r_{ik}^4} \qquad (18)$$

where *k* represents a cell in the grid space that is occupied by the solute molecule, d*V* is the volume of the cell, and r_{ik} is the distance to the center of atom *i*. The second term in $G_i^{solvent}$ accounts for the increase in electrostatic energy of atom *i* when elements of the solvent are displaced by the remaining atoms of the solute molecule which have a dielectric constant of unity[1,22] and "pseudo-neutral" charges. Applying these equations to eq 17 we have

$$\frac{1}{\alpha_i} = \frac{1}{a_i} - \frac{dV}{4\pi}\sum_k^V \frac{1}{r_{ik}^4} \qquad (19)$$

Equation 19 shows that the effective Born radii depend only on the geometry of the solute molecule; a direct consequence of the basic assumption in the GB model that effective Born radii do not depend on the charge distribution in the system. The same equation was derived by Scarsi et al as eq 14 in their paper.[27]

Calculation of α_i by eq 19 is very time consuming. Recently several different approximate approaches have been suggested to calculate Born radii in a pairwise manner.[24,25,29] However, because of the different atomic densities in macromolecules and in organic molecules and because of the frequent existence of unoccupied embedded volume between the ligand and the receptor, we use a modified form of eq 19.

Appendix 2: Calculation of the Effective Born Radii with Consideration of Dielectric Properties of Water in the Receptor Binding Site

Though the dielectric property of water in the active site before binding and the existence of an unoccupied volume after binding will not affect the formalism for solvation free energy calculations in the GB model (eqs 1−3), it will change the effective Born radii of solute atoms. Equations 18 and 19 in Appendix 1 must be modified as follows.

Before a ligand binds, $G_i^{solvent}$ in eq 18 is now given by

$$G_i^{solvent} = \frac{q_i^2}{2\epsilon a_i} + \frac{q_i^2}{2}\left(1-\frac{1}{\epsilon}\right)\frac{dV}{4\pi}\sum_k^V \frac{1}{r_{ik}^4} + \frac{q_i^2}{2}\left(\frac{1}{\epsilon_a}-\frac{1}{\epsilon}\right)\frac{dV}{4\pi}\sum_k^{V_a}\frac{1}{r_{ik}^4}$$

where ϵ_a represents the dielectric constant of water in the core region of the active site, and the summation is done for all grid elements in this region (denoted by V_a, see the lightly shaded region in the molecule on the left of Figure 1). Substitution of this expression in eq 17 yields

$$\frac{1}{\alpha_i^{(1)}} = \frac{1}{a_i} - \frac{dV}{4\pi}\sum_k^V \frac{1}{r_{ik}^4} - \frac{\frac{1}{\epsilon_a}-\frac{1}{\epsilon}}{1-\frac{1}{\epsilon}}\frac{dV}{4\pi}\sum_k^{V_a}\frac{1}{r_{ik}^4} \qquad (20)$$

Equation 20 is used in this paper to compute α_i numerically before ligand binding.

Similarly, after a ligand binds, α_i is defined by

$$\frac{1}{\alpha_i^{(2)}} = \frac{1}{a_i} - \frac{dV}{4\pi}\sum_k^V \frac{1}{r_{ik}^4} - \frac{\frac{1}{\epsilon'}-\frac{1}{\epsilon}}{1-\frac{1}{\epsilon}}\frac{dV}{4\pi}\sum_k^{V'}\frac{1}{r_{ik}^4} \qquad (21)$$

where ϵ' is the dielectric constant of the unoccupied embedded volume, and the summation is done for all grid elements in this region (denoted by V', see the shaded region for the molecule on the right of Figure 1). Because ϵ' is 1, eq 21 reduces to

$$\frac{1}{\alpha_i^{(2)}} = \frac{1}{a_i} - \frac{dV}{4\pi}\sum_k^{V+V'}\frac{1}{r_{ik}^4} \qquad (22)$$

That is, *k* represents a unit volume in the grid space that is occupied by a solute atom or overlapped with an unoccupied embedded volume. Equation 22 is used to compute α_i numerically after ligand binding.

ϵ_a is unknown and is determined by the hydrophobicity of the active site. The upper and lower bound for ϵ_a are the dielectric constant of the solvent (ϵ) and the dielectric constant of vacuum. When ϵ_a equals ϵ, eq 20 reduces to eq 19. When ϵ_a equals 1, eq 20 is the same as eq 19 except that the volume for summation expands from the solute molecule only to inclusion of the active site.

For a quick estimation of the size of the core region of the active site, we take the region occupied by the core of a known inhibitor of a receptor as the active site of the receptor. That is, replace the core of such an inhibitor by small spheres, each with a radius of 2Å. The resulting volume, excluding the overlap with the receptor, is taken to be the core active site. The embedded volume is approximated by the space complementary to the ligand molecule in the core region of the active site.

Appendix 3: Calculation of solvent-accessible surface area (SA)

To compute SA, uniform atom-based spherical grids are matched with a regular cubic grid. A set of evenly spaced elements are defined on each atom surface. The coordinates of the centers of these surface elements in the cubic grid box holding a solute molecule or an active site are then computed. Those element centers that do not overlap with other atoms are identified, and the corresponding accumulated surface area is the solvent-accessible surface area of interest.

The evenly spaced surface grid on an atom is defined by varying the azimuth angle (θ) from 0 to π and the polar angle (φ) from 0 to 2π with specified angular intervals. Taking R as the radius of the atom (=van der Waals radius + probe radius, where the probe radius is usually set to 1.4 Å[23]), r as $r = R \sin \theta$, and d as the spacing distance, then the angular intervals are $\Delta\theta = d/R$ and $\Delta\varphi = d/r$. *If* (x_0,y_0,z_0) are the coordinates of the atomic center, the coordinates of the element centers can be calculated by $x = x_0 + r \cos \varphi$, $y = y_0 + r \sin \varphi$, and $z = z_0 + R \cos \theta$. To save computational time, the surface grids are preset for each van der Waals atom type and saved for future SA calculations. The relative error of the numerical method here for SA calculations is approximately 2% in symmetrical cases for which analytical formulas for SA are available.

Appendix 4:- Parameter optimization for the implemented GB/SA model

The bound dhfr−MTX crystal structure[40] was taken from the Brookhaven Protein Data Bank (PDB).[58,59] We used unprotonated MTX in this test case. The core region of the active site is defined as the space occupied by the bound-MTX in the crystal structure excluding the dicarboxylate functionality up to the amide. These heavy atoms are replaced by dummy atoms with radii of 2 Å each. The embedded volume is the region complementary to the ligand molecule in the core active site. (*G*$_{binding}$ is not sensitive to the size of the active site. A significant reduction of the volume to that of the pteridine ring causes a change of only 3 kcal/mol in *G*$_{binding}$.) The same volume definitions were used for all dhfr−MTX-related calculations in this paper. ϵ_a was set to 78.3. Three sets of distance cutoff were used for energy and Born radii calculations: 5, 8, or 10 Å. Each cutoff was associated with different grid boxes. Specifically, the grid boxes enclose the spheres generated with the program SPHGEN,[3] with the box surfaces placed at 5 Å (for cutoff = 5 Å), 8 Å (for cutoff = 8 Å), or 10 Å (for cutoff = 10 Å) from the nearest sphere. For each cutoff, we tried three different grid spacings: 0.3, 0.4, and 0.5 Å. Table 5A shows how these parameters affect the GB term, *G*$_{POL}$. All the calculations were performed on a Silicon Graphics Octane workstation, equipped with a 195 MHz R10000 processor. The same procedures were also applied to the bound benzamidine−trypsin (bovine pancreas trypsin) crystal structure.[39] We used 8 and 10 Å for the cutoff distances because 5 Å appeared to be insufficient. The results are listed in Table 5B.

JA984102P