Parameters for the Generalized Born Model Consistent with RESP Atomic Partial Charge Assignment Protocol

Wei Zhang, Tingjun Hou, Xuebin Qiao, and Xiaojie Xu*

College of Chemistry and Molecules Engineering, Peking University, Beijing 100871, China

Received: March 11, 2003; In Final Form: May 16, 2003

Here we report a set of new parameters for the generalized Born (GB) model consistent with the RESP atomic partial charge assignment protocol. Effective atomic radii and screen factors as parameters have been obtained through genetic algorithm optimization in the parameter space to minimize the differences between the calculated and experimental solvation free energies. Here, the calculated solvation free energies are based on a GB model using partial charges fitted from the electrostatic potentials based on the 6-31G* basis set with the nonelectrostatic contributions to the free energy of solvation modeled in terms of the solvent accessible surface area (SASA). The mean unsigned error in the solvation free energies calculated by the GB/surface area calculations using the final parameters of the 328 neutral molecules in the training set is 0.85 kcal/mol, and for the 30 charged molecules the value is 4.36 kcal/mol. The refined parameters were then applied to predict the solvation free energies of 44 neutral or charged organic molecules and 15 proteins, and reliable results were obtained for both organic molecules and proteins. For the 36 neutral organic molecules in the test set, our parameters incurred an unsigned mean error of 0.73 kcal/mol, and for the eight charged molecules in the test set, our parameters incurred an unsigned mean error of 3.65 kcal/mol. For the 44 organic molecules, the performance of the GB/SA model based on our new parameters was much better than Possion-Boltzmann (PB)/SA and GB/SA based on Jayaram's parameters. For the 15 proteins randomly selected from the Protein Data Bank, the calculated results from GB/SA based on our new parameters also gave consistent results with those from PB/SA and were much better than GB/SA based on Jayaram's parameters. This model might be widely applied in molecules dynamics, protein folding, molecular docking, free energy calculations, and conformation analysis. Moreover, we are now supplying a program to help AMBER users apply our new parameters to their MD simulations.

Introduction

Accurately and rapidly modeling solvation is crucial to quantitatively understanding the chemical and physical properties that underly many biochemical processes. To solve this problem, both molecules[1−3] and continuum[4−10] models of solvent have been developed. Explicit solvent models employ thousands of discrete solvent molecules and have been widely used for simulations in the liquid environment. Many properties of the solutions can be reproduced by calculations employing explicit solvent models, but such calculations converge very slowly because of the large number of particles and states involved. Because explicit solvent models are so computationally demanding, there is interest in developing more rapid continuum solvation models. Continuum solvation models treat the solvent as a continuous medium surrounding the solute beginning near its van der Waals surface. In principle, such models can predict solvation effects with relatively little computational resources, because the model includes no particles other than the atoms of solute.

A lot of continuum solvation models have been reported over the years. Many treatments of these were based on the surface area (SA) or solvent accessible surface area (SASA).[11−15] However, we are concerned that area-based representations provide poor approximations for the long-range electrostatic component of solvation. Another popular approach to continuum solvation treats the solvent as a high dielectric continuum, interacting with charges that are embedded in solute molecules of lower dielectric media.[16−18] Despite the severity of the approximation, this model gives a good account of the electrostatic component of solvation energy. However, such dielectric continuum models of the solvent do not include van der Waals solvent−solute interaction terms.

Because of the shortcomings of the previous models, the Possion-Boltzmann (PB)/SA model and the generalized Born (GB)/SA model had been developed, which provided solvation free energy based on the PB equation[19] or GB model[9] for the electrostatic component and SA for the nonpolar component of the solvation free energy.

In the PB/SA and GB/SA models, the solvation free energy is given as the sum of the solute−solvent van der Waals term and the solute−solvent electrostatic polarization term, as illustrated in eq 1.

$$G_{sol} = G_{vdw} + G_{pol} \quad (1)$$

Because hydrocarbons are nonpolar molecules ($G_{pol} \sim 0$), and their G_{sol} in water is approximately linearly related[13−15] to their SASA, we compute the solute−solvent van der Waals term by evaluating the SASA.

G_{pol} are usually presented by the PB equation, which is typically solved by finite-difference or boundary element numerical methods.[8,17−19] The equation solving procedure may become very expensive for proteins and nucleic acids, so there

* Corresponding author. E-mail: xiaojxu@chem.pku.edu.cn.

10.1021/jp034613k CCC: $25.00
Published on Web 07/23/2003

is a interest in finding an approximate solvation for the PB equation. One candidate is the GB approach.[9] In this model, the polar term of solvation free energy is represented by eqs 2 and 3.

$$\Delta G_{\mathrm{pol}} = -\frac{1}{2}\left(1 - \frac{1}{\epsilon_{\mathrm{w}}}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{f_{\mathrm{GB}}} \tag{2}$$

$$f_{\mathrm{GB}} = \sqrt{r_{ij}^{\ 2} + R_i R_j \exp(-r_{ij}^{\ 2}/4R_i R_j)} \tag{3}$$

Among the equations, q_i and q_j is the partial charge of atoms i and j, respectively; ϵ_{w} is the solvent dielectric constant of the media; r_{ij} is the distance between atoms i and j; and R_i and R_j is the effective Born radii of atoms i and j, respectively. In its original form, R_i was estimated by a numerical integration procedure, but recently a pairwise approximation calculation of effective Born radii has been reported, and was widely accepted in estimating the solvation free energies of proteins.[20] The continuum solvation model has been reviewed heavily in the recent past, to include GB models.[21,22] A nice review of both continuum and explicit models applied to biological problems is that by Orozco and Luque.[23]

The energy calculations of protein and nucleic acids are typically based on force fields such as AMBER, CHARMM, and GROMOS, and each force field has its own rule in determining the atomic partial charges. For instance, the AMBER 6.0 suite software package employs a so-called RESP atomic partial charge assignment protocol, so for GB models run on these force fields, it is necessary to determine a set of parameters to estimate effective Born radius accurately. In 1998, Jayaram et al. reported a set of parameters for GB models consistent with the RESP protocol.[24] In 2000, Cheng et al. reported a set of parameters for GB models consistent with the MMFF force field.[25] Recently, Liu et al. reported GB parameters consistent with the GROMOS96 force field.[26]

As we mentioned above, there was already a set of GB parameters consistent with the RESP protocol reported by Jayaram et al. in 1998, but their parameters were designed only for proteins, and not suitable for studying the protein−inhibitor complex system; so for many organic molecules, the performance of GB/SA based on the Jayaram's parameters is very poor. Moreover, Jayaram's group employed a relatively small training set (32 molecules) to derive their parameters, which is insufficient for dealing with small organic molecule systems, so there is interest in a set of GB parameters that could predict the solvation free energy of small organic molecules accurately.

Here we report a set of parameters for the GB model consistent with the RESP atomic partial charge assignment protocol. This set of parameters was derived by fitting to the experimental solvation free energies of 358 small organic molecules. To our knowledge, the training set used here is significantly larger than those in previous work.[24−26] The adoption of a large training set gives us more opportunity to define more elaborate definitions of atom typing rules. Furthermore, we used the genetic algorithm (GA) to optimize the difference between the calculated solvation free energies and the experimental solvation free energies. As a very efficient stochastic optimization method, it has been widely used to solve the minimization problems such as conformational search,[27] molecular docking,[28−30] and QSAR.[31−33] We expect that by applying GA as the optimization method, the parameters for each atom type should achieve the most optimal values. The Minnesota group has used GA to develop all of their quantum mechanical GB models (the SMx models) and have gotten brilliant results.[34]

Methodology

Data Set. We selected 402 organic molecules to perform the parametrization. Their names and experimental solvation free energies are included in Supporting Information. The experimental solvation free energies were determined at 298 K, 1 atm. The molecular geometries of all compounds were modeled using the Cerius^2 molecular simulation package.[35] The initial structures were fully minimized using molecular mechanics with the MMFF force field.[36] Conformational analyses were performed for some molecules with flexible chains to find the global minimum geometries. For each molecule, only the global minimum conformation was used in the subsequent parametrization. The whole data set was divided into a training set with 358 molecules and a test set with 44 molecules. For all molecules in the training set, 328 molecules were neutral, while the other 30 molecules were charged. The solvation parameters were determined based on the training set, and the actual prediction ability was validated by the test set. The parametrization procedure is basically searching for a set of parameters that could reproduce the solvation free energies of the parametrization set molecules. The electrostatic contribution to solvation free energy can be estimated according to eqs 2 and 3, which requires an input of atomic partial charge, atomic Cartesian coordinates, and atomic initial Born radii. The nonpolar contribution to the solvation free energy is estimated by a molecule's SASA. The derived model was then applied to predict the solvation free energy of two test sets.

Among the first test set made up of 44 simple organic molecules, 36 molecules were neutral and 8 molecules were charged. As the second test set, 15 proteins were chosen randomly from the Brookhaven Protein Data Bank (PDB). For these proteins, all crystallographic water molecules were eliminated from the structures. Some missing hydrogen atoms were added using the InsightII molecular simulation package,[37] with a neutral sp3 N-terminus and a carboxylic (COOH) C-terminus assigned at neutral pH. These structures were minimized using the AMBER force field to remove any steric overlap with a restraint of the main chain.

Atom Typing Rule. We classified the atoms in molecules according to its element and hybridization. A total of 21 atom types were introduced, and four of them were designed for representing charged molecules. The definition of these atom types was based on the SMARTS description (see Table 1).[38] SMARTS is a language that allows you to specify substructures using rules that are straightforward extensions of SMILES. In fact, almost all SMILES specifications are valid SMARTS targets. As SMILES, in SMARTS one can use atomic and bond symbols to specify a graph. However, in SMARTS the labels for the graph's nodes and edges (its "atoms" and "bonds") are extended to include "logical operators" and special atomic and bond symbols; these allow SMARTS atoms and bonds to be more general. Using SMARTS, flexible and efficient substructure search specifications can be made in terms that are meaningful to chemists. In the current work, a parameter file was used to store the SMARTS chains defined for all atom types. If we want to add some new typing rules or modify the typing rules, we only need to make some modifications to this parameter file.

Derivation of Atomic Coordinate and Atomic Partial Charges. Partial charges were derived to be consistent with the AMBER charge derivation protocols. All the studied organic

TABLE 1: Typing Protocol of Atom Types and Derived Parameters for Generalized Born Solvation Model

no.	name	description	occurrence in parametrization set	radii	screen parameter
1	HC	hydrogen atom connected to alkane carbon	2707	1.6	0.5
2	H1	hydrogen atom connected to polar atom	100	0.9	1.2
3	HA	hydrogen atom connected to aromatic carbon	376	0.9	0.8
4	C1	sp^1 carbon atom	25	2.5	0.9
5	C2	sp^2 carbon atom	155	2.4	0.5
6	C3	sp^3 carbon atom	1151	2.1	0.6
7	CA	aromatic carbon	538	1.7	0.8
8	N1	sp^1 nitrogen atom	7	1.4	1.5
9	N2	sp^2 nitrogen atom	51	1.3	0.7
10	N3	sp^2 nitrogen atom	28	1.2	0.9
11	O1	sp^2 oxygen	95	1.85	1.0
12	O2	sp^3 oxygen	132	1.5	0.8
13	F	fluorine	89	2.0	1.0
14	P	phosphorus	9	2.5	0.5
15	S	sulfur atom	25	2.0	1.1
16	Cl	chlorine	128	2.2	0.8
17	Br	bromine	37	2.5	0.7
18	N2C	charged sp^2 nitrogen atom	24	2.8	0.6
19	N3C	charged sp^3 nitrogen atom	14	1.4	1.0
20	OC	charged oxygen atom	8	1.3	1.9
21	SC	charged sulfur atom	4	1.8	1.1

molecules after minimization of molecular mechanics were further optimized using quantum mechanics with HF/6-31G basis set, and then the HF/6-31G* electrostatic potential (ESP) charges of the small organic molecules were obtained using Gaussian-98.[39] For proteins, their ESP charges were assigned using *xleap*, the graphics interface of AMBER 6.0 software package,[40] which assigned the predetermined RESP charge to each atom in protein. The charge we used here was the gas-phase HF/6-31G* charge, which is larger than the condensed-phase charge, but this error mimics the increased partial atomic charges that would be expected after solute polarization were the solute to be treated by a quantum mechanical self-consistent reaction field procedure. Carlson et al. were the first to point this out explicitly.[40]

Calibration of Atomic Effective Born Radii. The effective Born radii were calculated following a procedure recommended by Hawkins, Cramer, and Truhlar, in which the effective Born radii are estimated from a sum over atom pairs as described in eq 4.[18]

$$R_i^{-1} = r_i^{-1} - \left(\frac{1}{2}\right)\sum_j \frac{1}{L_{ij}} - \frac{1}{U_{ij}} + \frac{r_{ij}}{4}\left(\frac{1}{U_{ij}^2} - \frac{1}{L_{ij}^2}\right) + \frac{1}{2r_{ij}}\ln\frac{L_{ij}}{U_{ij}} + \frac{\rho_j^2}{2r_{ij}}\left(\frac{1}{U_{ij}^2} - \frac{1}{L_{ij}^2}\right) \quad (4)$$

where

$$L_{ij} = 1, \quad \text{if } r_i > r_{ij} + \rho_j$$

$$L_{ij} = r_i, \quad \text{if } r_{ij} - \rho_j < r_i < r_{ij} + \rho_j$$

$$L_{ij} = r_{ij} - \rho_j, \quad \text{if } r_i < r_{ij} - \rho_j$$

$$U_{ij} = 1, \quad \text{if } r_i > r_{ij} + \rho_j$$

$$U_{ij} = r_{ij} + \rho_j, \quad \text{if } r_i < r_{ij} + \rho_j$$

In the above equations, r_i and r_j are the initial Born radii of atom i and atom j, while and ρ_i is the product of atom i's initial Born radii and screen parameters. The screen parameters were introduced by Hawkins et al. to correct for systematic errors introduced by the pairwise screening approximation.

Nonpolar Contribution of the Solvation Free Energy. As we had mentioned above, the nonpolar contribution solvation free energy is linear relative to the SASA. We calculated the surface using MSMS with a probe radius of 1.4 Å.[41] The nonpolar contribution of solvation free energy is then estimated with a coefficient of 0.005 kcal/(mol Å^2), which is the default value of the SANDER module of the AMBER 6.0 software package.

Parameters Fitting Procedure. Now we have all the inputs required by a GB/SA solvation model. Given a set of parameters, we can predict the solvation free energy of each molecule in the training set, and given the experimental solvation free energy, we can estimate its mean unsigned error.

Here, the fitting procedure was based on GA, which was under development in our laboratory.[30−32] GA can effectively deal with the multiple-dimension problem, no matter whether those variables are highly coupled or not, which makes it an ideal optimization method for the problem of parametrization.[33,42,43] The brief fitting process based on GA is in four steps: creation of the initial population, selection operation, crossover operation, and mutation operation. According to the GA, an individual should be represented as a linear string, which plays the role of the DNA for the individual, so the parameters for all atom types were treated as a string. The initial population was generated by randomly generating the initial parameters. Then these individuals were scored according to their fitness. In the parametrization of GA, the sum of mean unsigned errors of the training set was used as the score function. After some cycles of the selection, crossover, and mutation operations, the model with the highest fitness score was obtained. The GA optimizations were terminated if the total mean unsigned error did not change after a certain number of iterative cycles; for example, after 100, the optimization would end. More detailed description of the fitting process based on GA can be found in our previous work.[33,42,43]

The fitting procedure was taken in two stages. In the first stage, the parameters of 17 normal atom types were determined by fitting the experimental solvation free energy with the calculated solvation free energies of the 328 neutral using a genetic algorithm. In the second stage, the derived parameter in the first stage was fixed and the parameters of four ion atom types were determined by fitting the experimental solvation free

energies with the calculated solvation free energies of the 30 charged molecules using a systemic search method.

PB/SA Solvation Model. As we had mentioned above, PB/SA solvation model is a widely accepted solvation model, so we used the model to predict the solvation free energy of test set molecules, and 15 proteins for a reference.

In PB/SA model, the electrostatic contribution to the solvation free energy was calculated by taking the difference between the total energy of the system obtained with ϵint = 2; ϵext = 1 and ϵint = 2; ϵext = 78.5. It requires an input of atomic coordinate, partial charge, and van der Waals radii. In our calculations, we used the same partial charge and coordinate as in the GB calculation. For the van der Waals radii, we used the default value supplied by the DelphiII software package.[19] The radii of atoms were taken from the PARSE parameter set.[44] It should be noted that the PARSE parameter set does not provide the van der Waals radii of halogen, so we used the van der Waals radii in the AMBER force field instead, which is 1.75 Å for the fluorine atom, 1.95 Å for the chlorine atom, and 2.22 Å for the bromine atom.

The calculations based on PB equations were performed using the DelphiII software package. In the prediction of solvation free energies of test set molecules, the resolution employed was 4 grids/Å. For proteins, the resolution was 1.8 grids/Å. The calculations of SASA were performed using the MSMS program.[41]

Jayaram's GB/SA Model. Jayaram's generalized Born model was employed to predict the solvation free energy of all the reference systems.[21] The setting is the same as that in our GB calculations.

It should be noted that many molecules in the training set and the test set contain halogen atoms, but they do not have corresponding GB parameters in Jayaram's parameter set, so for these atoms, the initial Born radii were set to van der Waals radii in the AMBER force field and the screen parameter was set to 1.0. The calculations based on Jayaram's solvation model were performed using SANDER in AMBER6.0 with the control parameter GBPARM set to 1.

Results and Discussion

Atom Typing Rules. As we mentioned above, for each atom two parameters should be determined: the initial Born radii and the screen parameter, so the determination of the parameter of every atom is impossible. One strategy is to classify atoms into different categories according to its chemical environment. Atoms in the same category have the same parameters. Each category is called an atom type, and their definitions are called the atom typing rule. The atom typing rule is the key of the parametrization work, and it should meet some demands. First, it should be fully contained, which means that every atom in the concerned system can be classified into one atom type; second, each atom type should be exclusive, which means there should not be an atom that can be classified into two or more categories; third, the number of atom types should be as small as possible to avoid the overfitting problem.

In the work of Jayaram et al.,[24] the authors employed relative simple atom typing rules, in which atoms are classified into six types according to their element. They chose this typing rule partly because of their relative small training set (32 molecules). Their parameters would be efficient in predicting the solvation free energy of proteins and nuclear acids, because atom types in these systems are very limited, but it may not behave well in predicting the solvation free energies of organic molecules because the chemical environments in organic molecules are much more complicated than those in proteins. Therefore, Jayaram's parameters may not be suitable for studying the solvation contribution to the interaction free energy between enzyme and inhibitor, which is of great importance in drug design.

TABLE 2: Neutral Organic Molecules in the Training Set with Deviations Larger than 2.0 kcal/mol

no.	molecular name	ΔG_{exp}	ΔG_{calc}	residue
65	tetrafluoromethane	3.16	0.33	−2.83
66	hexafluoroethane	3.94	1.07	−2.87
67	octafluoropropane	4.28	2.01	−2.27
68	fluorobenzene	0.78	−1.71	−2.49
70	chlorofluoromethane	0.77	−1.36	−2.13
90	1,1,2,2-tetrachloroethane	−2.36	−0.08	2.28
225	methyl formate	−2.78	−5.91	−3.13
226	ethyl formate	−2.65	−5.29	−2.64
227	propyl formate	−2.48	−5.21	−2.73
251	ethyl heptanoate	−4.60	−7.87	−3.27
252	methyl octanoate	−4.61	−6.72	−2.11
253	methyl benzoate	−4.50	−7.55	−3.05
254	butylamine	−4.38	−6.65	−2.27
255	pentylamine	−4.09	−6.67	−2.58
256	hexylamine	−4.04	−6.82	−2.78
264	*N*,*N*-dimethylpiperazine	−7.58	−2.04	5.55
265	*N*-methylpiperazine	−7.77	−5.44	2.33
266	1,1-dimethyl-3-phenyl urea	−11.87	−9.00	2.87
268	ethylenediamine	−9.75	−18.23	−8.48
271	*N*-methylmorpholine	−6.34	−3.59	2.75
272	*N*-methylpyrrolidine	−3.97	−0.80	3.17
273	*N*-methylpiperidine	−3.89	−0.78	3.11
291	9-methyladenine	−13.60	−20.19	−6.59
307	*N*-methylformamide	−10.00	−7.94	2.06
309	*E*-*N*-methylacetamide	−10.00	−7.75	2.25
310	*Z*-*N*-methylacetamide	−10.00	−7.14	2.86
328	dimethyl 3-methyl-4-thio-methoxyphenyl thiophosphate	−6.92	−3.85	3.07
329	ethyl 4-cyanophenyl phenylthiophosphonate	−5.10	−7.91	−2.81

To predict the solvation free energy of organic molecules, more complicated atom typing rules must be employed. Here we employed an atom typing rule containing 21 atom types, and four of them are specially designed for representing ions. Their definitions and number of appearance in the training set are listed in Table 1. To avoid the overfitting problem, we employed a much bigger training set made up of 358 molecules, in which 328 are neutral and 30 are charged.

We took this typing rule from the one SANDER used in the calculation of molecular SASA. As we know SANDER employed a method named LCPO to calculate molecular SASA. In LCPO, atoms were classified into 21 types according to their element, hybridization, and number of hydrogen atoms linked to it; our typing rule is much like theirs except that we neglect the number of hydrogen atoms which will enlarge our type set and consequently cause the overfitting problem.

Solvation Free Energy of Training Set and Test Set. The derived parameters are listed in Table 1. If we do not consider the four charged atom types, the GB/SA model based on the new parameters yielded fairly satisfactory results, $n = 328$, $r = 0.911$, $s = 1.241$, $F = 1590.841$. From the predictions to the molecules in the training set, there are two compounds (compounds 268 and 291 in Table A, Supporting Information) with deviations larger than 6.0 kcal/mol. If we eliminate these two compounds as outliers, the correlation between the experimental solvation free energies and the calculated values was improved obviously ($n = 326$, $r = 0.924$, $s = 1.115$, $F = 1911.039$). For the 328 neutral molecules in the training set, this new set of parameters gives a mean unsigned error of 0.85 kcal/mol in predicting the solvation free energy. For the charged

TABLE 3: Solvation Free Energy of Molecules in the Test Set

no.	name	ΔG_{exp}	our GB/SA model	PB/SA model	Jayaram's GB/SA model
A1	*n*-pentane	2.33	1.18	1.13	1.67
A2	*n*-heptane	2.62	1.32	1.29	1.96
A3	4-methyl-1-pentene	1.91	1.15	−1.01	−0.19
A4	1,4-pentadiene	0.94	0.98	−2.46	−1.29
A5	butenyne	0.04	−0.07	−4.01	−3.13
A6	butylbenzene	−0.40	−1.30	−2.16	0.67
A7	1,1-difluoroethane	−0.11	−1.34	−2.92	−0.59
A8	dichlorodifluoromethane	1.69	1.15	1.14	1.68
A9	1-bromo-1,2,2,2-tetrafluoroethane	0.52	−1.30	−0.86	2.88
A10	1,1,1,2-tetrachloroethane	−1.15	−0.36	−1.53	1.39
A11	1,1-dichlorobutane	−0.70	−0.51	−1.47	0.84
A12	chlorobenzene	−1.01	−1.60	−2.17	0.36
A13	1-chloro-2-bromoethane	−1.95	−0.82	−2.81	0.28
A14	1-bromo-2-methylpropane	−0.03	−0.48	−1.61	0.17
A15	*o*-bromocumene	−0.85	−1.31	−2.03	0.51
A16	1-butanol	−4.72	−5.24	−4.12	−1.97
A17	2-methyl-1-pentanol	−3.93	−4.62	−3.79	−1.51
A18	1-heptanol	−4.25	−4.94	−3.79	−1.42
A19	3-cresol	−5.49	−5.86	−7.02	−3.04
A20	ethyl propyl ether	−1.81	−1.23	−0.89	1.25
A21	1,2-diethoxyethane	−3.53	−2.43	−2.16	1.46
A22	pentanal	−3.03	−2.89	−4.26	−3.16
A23	*m*-hydroxybenzaldehyde	−9.51	−8.80	−10.22	−5.51
A24	cyclopentanone	−4.68	−3.32	−4.76	−2.98
A25	propionic acid	−6.46	−6.36	−8.11	−4.80
A26	isobutyl formate	−2.22	−3.13	−5.24	−1.68
A27	methyl propionate	−2.97	−3.01	−4.93	−1.54
A28	methyl hexanonate	−2.48	−2.57	−4.58	−0.84
A29	dimethylamine	−4.28	−4.99	−1.98	−0.36
A30	aniline	−5.49	−7.42	−7.69	−2.31
A31	pyrrolidine	−5.47	−2.88	−1.24	0.84
A32	2,4-dimethylpyridine	−4.85	−4.39	−4.33	−1.44
A33	2-ethyl-3-methoxypyrazine	−4.39	−4.20	−3.48	−0.10
A34	*N*′,*N*-dimethyl formamide	−4.90	−5.38	−6.81	−4.45
A35	thioanisole	−2.73	−2.09	−3.20	−0.51
A36	tripropyl phosphate	−6.10	−6.47	−8.58	−3.14
	mean unsigned error	0.00	0.73	0.73	1.56
A37	(*cyclo*-C_6H_{11})NH_3^+	−62.00	−56.91	−66.08	−42.55
A38	(CH_3)(C_6H_5)NH_2^+	−56.50	−56.16	−62.38	−40.48
A39	$C_5H_5NH^+$[c]	−53.50	−59.05	−64.08	−46.55
A40	(CH_3)$_2$(*n*-C_3H_7)NH^+	−59.00	−51.63	−58.67	−35.84
A41	(*n*-C_3H_7) $H_5N_3C^+$	−65.50	−61.15	−63.21	−34.23
A42	(*i*-C_3H_7) $H_5N_3C^+$	−63.00	−62.00	−63.64	−30.95
A43	$C_6H_5CO_2^-$	−73.50	−68.05	−68.42	−62.06
A44	(*n*-C_3H_7)S^-	−79.00	−78.88	−78.81	−71.71
	mean unsigned error	0.00	3.65	3.65	3.63

molecules in the training set, the GB/SA calculations give much worse prediction than the neutral molecules. For the 30 charged molecules in the training set, the new parameters produced a mean unsigned error of 4.71 kcal/mol. The experimental and calculated solvation free energies using the new parameters are summarized in Tables A and B in Supporting Information. The GB/SA model based on our new parameters predicts well for most of the 328 neutral compounds in training set, but as listed in Table A and B, 28 compounds showed deviations greater than 2.0 kcal/mol. We think that these deviations can be explained by two reasons. The first may be suggested by the principle difference between the experimental solvation free energies and the calculated values of GB/SA. In our fitting process, the experimental solvation free energies were treated as the standard values. But it should be noted that the experimental value of a solute is not induced by a single molecule, but by a group of solute molecules in solvent. The solute molecules may produce intermolecular or intramolecular group−group interactions, for example, the intermolecular hydrophobic interactions and intra- or intermolecular hydrogen bonds. Jayaram et al. used the solvation free energy predicted by the PB/SA model as the standard value; it can be accurate in dealing with protein systems, but it is unacceptable in dealing with small organic molecule systems, as we illustrated in Table 3. The PB/SA model cannot give accurate solvation free energies of small organic molecules, so we think that using the experimental solvation free energies in parametrization is more reasonable than the calculated solvation free energies from PB/SA calculations. Second, in our work, we may not define the best atom typing rules. In principle, if we define enough atom types and can obtain the parameters for them, the electrostatic contribution may be well estimated. But unfortunately, the chemical environments in organic molecules are so complicated that it is very difficult for us to define unlimited atom types to differentiate all chemical environments. It could explain the stunningly large radius of type N2C (2.8 Å), but we have very little occurrence of this type of atom (24 times), so the radius of this type is not reliable. It could be overcome by adding a new experimental value to the training set. Moreover, the data set with experimental solvation free energies is limited. The data do not allow us to define so many atom types, otherwise overfitting cannot be avoided.

The derived parameters were then applied to predict the solvation free energy of the test set molecules for validation.

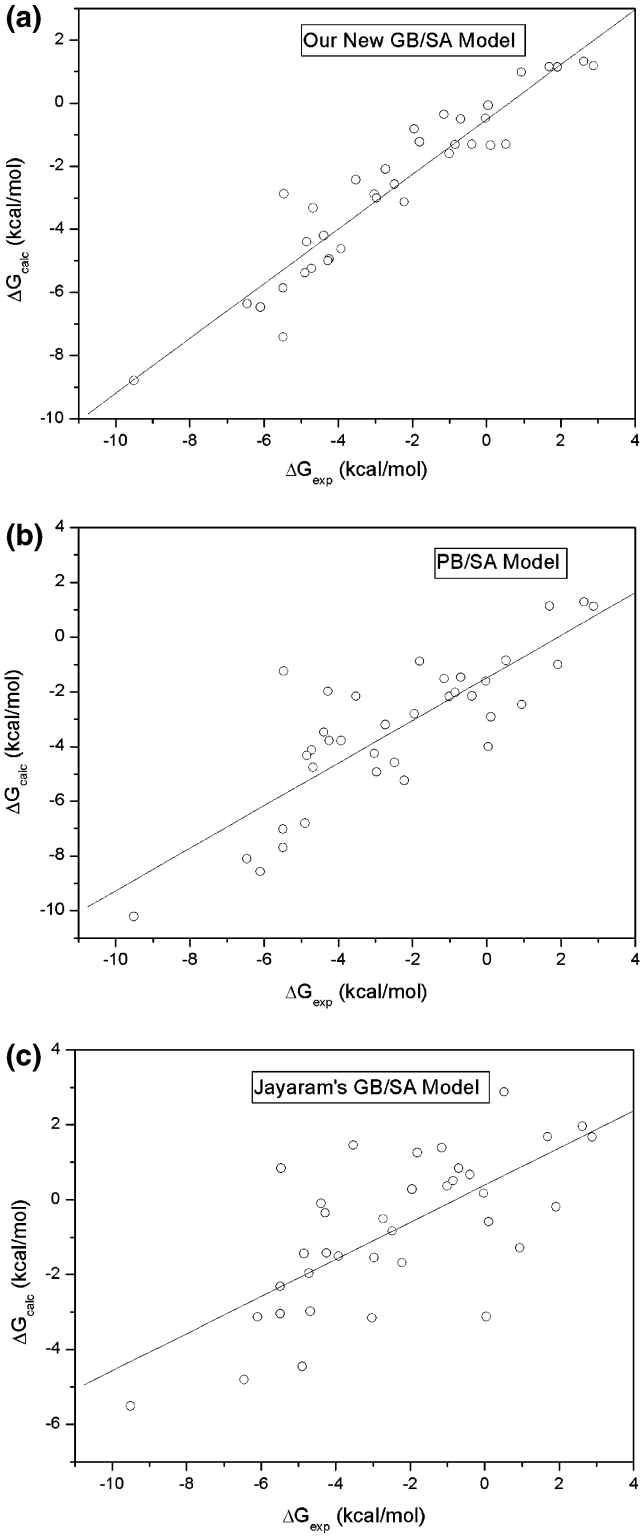

Figure 1. (a) Experiment solvation free energy vs predicted value using our new parameters. (b) Experiment solvation free energy vs predicted value using our new parameters. (c) Experiment solvation free energy vs predicted value using Jayram's GB/SA model.

The experimental solvation free energy and the predicted value of the test set are listed in Table 3. For the 36 neutral molecules in the test set, the obtained mean unsigned error is 0.73 kcal/mol, and for the left 8 charged molecules this value is 3.65 kcal/mol. Figure 1a shows the linear correlation between the experimental values and the calculated values using GB/SA based on our new parameters for the neutral molecules in the test set, which has a correlation coefficient of 0.95 and a standard deviation of 0.86. The good prediction for the test set indicates that the obtained parameters are reliable.

TABLE 4: Solvation Free Energy of Proteins

PDB entry	PB/SA	Jayaram's GB/SA	our GB/SA
1ahh	−4013	−1170	−4047
1bbh	−2221	−2966	−3387
1bbs	−3769	−3200	−3745
1c7c	−5260	−1615	−4057
1ctf	−1087	−1018	−1029
1dyv	−1586	−1648	−2100
1eol	−3100	−2269	−3320
1fkh	−1074	−809	−983
1g54	−2283	−1731	−2279
1gky	−2268	−2047	−2286
1htr	−4517	−3666	−4084
1mpp	−6090	−6001	−6056
1prn	−7788	−7930	−7503
1ypa	−681	−625	−697
2alp	−1468	−742	−1537

In a comparative fashion, the PB/SA model and Jayaram's GB/SA model had also been employed to predict the solvation free energy of the test set. For the charged molecules in the test set, the performance of our GB/SA model, the PB/SA model, and Jayaram's GB/SA is similar. But for the neutral molecules in the test set, the performance of these three solvation models shows obvious differences. Figure 1b shows the linear correlation between the experimental values and the calculated values using PB/SA for the neutral molecules in the test set, which has a correlation coefficient of 0.82 and a standard deviation of 1.57. For the neutral molecules in Table 3, the mean unsigned error between the experimental values and the calculated values using PB/SA is 1.55 kcal/mol. It is obvious that the predictive ability of our GB/SA model is obviously better than that of the PB/SA model. Figure 1c shows the linear correlation between the experimental values and the calculated values using Jayram's GB/SA model for the neutral molecules in the test set, which has a correlation coefficient of 0.70 and a standard deviation of 1.48. Meanwhile, the unsigned mean error is 2.18 kcal/mol, which means that Jayaram's GB/SA model does not have good predictive power for small organic molecules. From the calculated results in Table 3, we also find that the PB/SA model and Jayaram's GB/SA model failed in predicting the solvation free energy of molecules containing halogen. That is because the two models do not contain parameters for halogen atoms. We also see that Jayaram's GB/SA model fails to predict the solvation free energy of molecules containing pyrrole or pyridine functional groups. This may be because they did not include this kind of molecule in their training set. The fact that the PB/SA model behaved much better than Jayaram's model implies that the PB/SA model is a more reliable model than the GB/SA model. Though our new model behaved better than the PB/SA model, our model employed 34 parameters, while the PB/SA employed only nine parameters. The only superiority of the GB/SA model to the PB/SA model is that the GB/SA model is much faster than the PB/SA model. To perform PB calculations on the 36 molecules in test set, it took a PIV1.4GHz processor 33 s. For the GB model, the cost time is only 3 s.

Solvation Free Energy of Proteins. The parameters for GB/SA are derived based on a set of small molecules. Certainly, the functional groups of protein can also be found in these small organic molecules, so we believe that the parameters can be extended to proteins. The calculated results using our GB/SA model and the PB/SA model for the 15 proteins in the test set are shown in Table 4. Figure 2a shows that the plot of

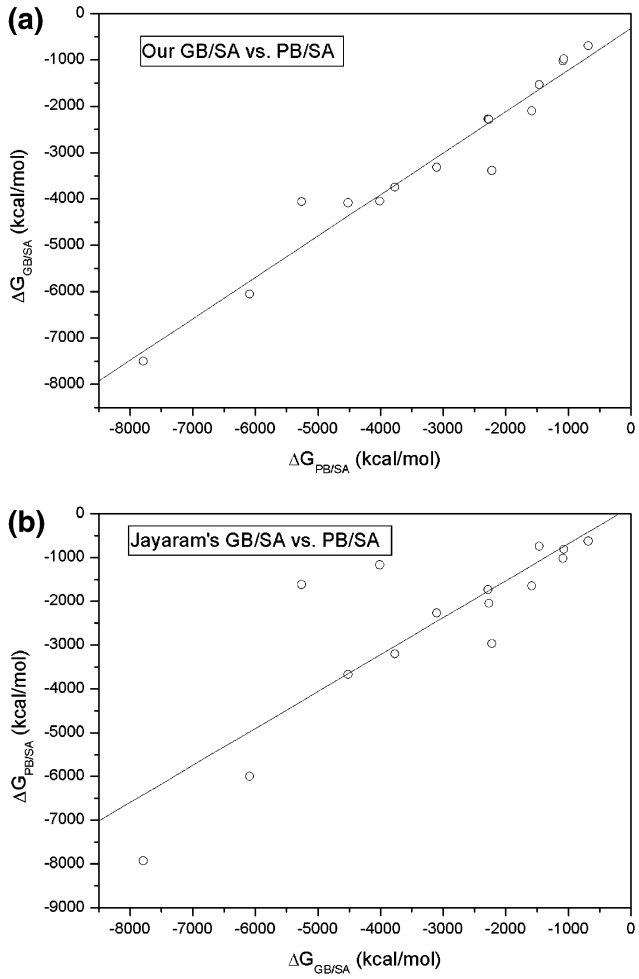

Figure 2. (a) The predicted solvation free energy of proteins using our GB/SA model vs PB/SA model. (b) The predicted solvation free energy of proteins using Jayaram's GB/SA vs PB/SA model.

predictions using PB/SA versus predictions using our GB/SA model. The good linear correlation ($r = 0.97$) indicates that the solvation abilities of these 15 proteins can be well ranked by the predictions using our GB/SA model. From the absolute values, the predictions using PB/SA are in good agreement with those using our GB/SA model besides 1bbh and 1bbs.

Here, the predictive ability of Jayaram's GB/SA model was also investigated. The predicted values are shown in Table 4. The correlation between the predictions using Jayaram's GB/SA model and those using PB/SA is shown in Figure 2b. The predicted values using these two models show obvious linear correlation ($r = 0.85$). But the linear correlation is obviously worse than that shown in Figure 2b. Moreover, the data in Table 4 indicate that the unsigned mean error between the predicted values using PB/SA and those using Jayaram's GB/SA model is 778 kcal/mol, which is much larger than the unsigned mean error (278 kcal/mol) between the predicted values using PB/SA and those using our GB/SA model. This fact implies that for proteins the predictive power of our GB/SA model is much better than that of Jayaram's GB/SA model.

Further Development and Applications of the GB/SA Model. Due to the simplicity and efficiency of the GB/SA model, it may be widely used in many fields. But further application of this method is also significantly restricted by its predictive power. If we want to improve the predictive ability of the GB/SA model, we should provide more elaborate atom typing rules and corresponding parameters. Although the predictions of the GB/SA model using our new parameters have been improved a lot, we also believe that the atom typing rules used here should not be the optimum. In our further work, we will attempt to give more rational definitions for these atoms in a complicated chemical environment. Certainly, the number of atom types is strongly limited by the available experimental data. The predictive ability of the charge-independent model may be improved by choosing a balanced training set that represents as many chemical functionalities as possible.

The further potential applications of GB/SA should be promising. First, the SASA model has potential applications in molecular dynamics, conformational analysis, and protein folding. For example, the special program named *pdb_typing* was developed to help current AMBER users apply our new parameters to their MD simulations. More detailed descriptions of the program *pdb_typing* can be found in Supporting Information. In our previous work, GA was used to sample the conformational spaces and thoroughly search the global conformations of peptides.[41] But in our program, only the potentials of the peptides were considered. In future work, we will apply this model to calculate the solvation free energy in protein folding or the installation of side chains. We expect that the consideration of the solvation free energy will improve the performance of our method. Second, we will apply this model to calculate the relative binding free energy for a set of protein/ligand complexes and incorporate this model into our docking program. In our group, we have developed different score functions for the following two stages of conformation searching. In the first stage, surface complementarity is considered, while in the second stage only energetic complementarity is considered. In the current release of our SFDOCK program, only the van der Waals and electrostatic interactions were used to estimate the energetic complementarity. Soon, the GB/SA model will be incorporated into our program.

Conclusion

We derived a set of parameters for the GB/SA model consistent with the AMBER force field. We employed a much larger training set (358 molecules) in the parameter's derivation procedure than Jayaram et al. did in deriving their parameters. In the current work, we employed atom typing rules containing 21 atom types, of which four are specially designed for ions. The definition of atom types was based on the SMARTS string. Predictions using the solvation model based on the 358-molecule set give an average unsigned error of 0.85 kcal/mol for the neutral molecules and 4.71 kcal/mol for the ions.

We applied the parameters developed in this paper to calculate the solvation free energies for 44 small organic molecules. The calculated results using our new parameters are consistent with those from experiments. Comparison of the results from our GB/SA model, PB/SA model, and Jayaram's GB/SA model shows that the calculations with our GB/SA model are obviously better than those with the other two solvation models. We have also applied our model to predict the solvation free energies for 15 proteins. For the 15 proteins randomly selected from the Brookhaven PDB database, the solvation free energies predicted by the SASA model bear high linear correlations ($r = 0.97$) with those predicted by the PB/SA model, which were much better than those given by Jayaram's GB/SA model.

Supporting Information Available: Experimental and calculated solvation free energy values for molecules of the training set (Table A) and experimental and calculated free

energy values for molecules in the test set (Table B). This material is available free of charge via the Internet at http://pubs.acs.org. The special program *pdb_typing* was developed to help AMBER 6.0 users apply our parameters to their MD simulations. *pdb_typing* reads the coordinates with PDB format and writes a radii file that contains the radii and screen parameter of each atom. This program needs two input files: gbparm.dat and atomtyp.txt. File gbparm.dat contains the derived GB/SA parameters of our work, and atomtyp.txt contains the atom typing definitions represented in SMARTS language. AMBER users should do a minor revision to SANDER's source code, which is under the directory $AMBERHOME/src/sander/. The revision is adding one line "read (18,*) (x(L96 − 1 + i), i = 1, natom)" to the file mdread.f after line 637, and then recompiling it. (You can also download the revised file from our web site.) The PDB format file can be generated using the SAVEPDB command available in XLEAP module of AMBER6.0. The generated radii file can be used by SANDER in MD simulations and optimizations by specifying -radii <filename> in the command line and setting the parameter READRAD to 1 in SANDER's control parameter file. The program *pdb_typing* and the corresponding parameter files can be obtained from us upon request.

References and Notes

(1) Rossky, P. J.; Karplus, M. *J. Am. Chem. Soc.* **1979**, *101*, 1913−1937.

(2) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(3) Jorgensen, W. L.; Ravimohan, C. J. *J. Chem. Phys.* **1985**, *83*, 3050−3054.

(4) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199−203.

(5) Ooi. T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086−3090.

(6) Kang, Y. K.; Nemethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1987**, *91*, 4105−4109/4109−4118.

(7) Kang, Y. K.; Nemethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1987**, *91*, 4118−4122.

(8) Gilson, M.; Honig, B. *Proteins* **1988**, *4*, 7−18.

(9) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. J. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6133.

(10) Cramer, C. J.; Truhlar, D. G.*J. Am. Chem. Soc.* **1991**, *113*, 8305−8311.

(11) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754−2759.

(12) Amidon, G. L.; Yalkowsky, S. H.; Anik, S. T.; Valvani. S. C. *J. Phys. Chem.* **1975**, *79*, 2239−2246.

(13) Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616−627.

(14) Hou, T. J.; Qiao, X. B.; Zhang, W.; Xu, X. J. *J. Phys. Chem. B* **2002**, *106*, 11295−11304.

(15) Hou, T. J.; Xu, X. J. *Acta Phys. Chim. Sin.* **2002**, *18*, 1052−1056.

(16) Jayaram, B. *J. Phys. Chem.* **1994**, *98*, 5773−5777.

(17) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435−445.

(18) Gilson, M.; Sharp, K. A.; Honig, B. *J. Comput. Chem.* **1988**, *9*, 327−335.

(19) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144−1149.

(20) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122−129.

(21) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161−2200.

(22) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−152.

(23) Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187−4225.

(24) Jayaram, B.; Sprous, D.; Beveridge, D. L. *J. Phys. Chem. B* **1998**, *102*, 9571−9576.

(25) Cheng, A.; Best, S. A.; Merz, J. K. M.; Reynolds, C. H. *J. Mol. Graph. Mod.* **2000**, *18*, 273−282.

(26) Zhu, J.; Shi Y. Y.; Liu, H. Y. *J. Phys. Chem. B* **2002**, *106*, 4844−4853.

(27) Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. *J. Comput. Chem.* **1993**, *14*, 1407−1414.

(28) Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. *J. Comput. Aid. Mol. Des.* **1995**, *9*, 113−130.

(29) Jones, G.; Willett, P.; Glen, R. C. *J. Mol. Biol.* **1995**, *245*, 43−53.

(30) Hou, T. J.; Wang, J. M.; Chen, L. R.; Xu, X. J. *Protein Eng.* **1999**, *12*, 639−647.

(31) Kubinyi, H. *Quantum Struct.-Act. Relat.* **1994**, *13*, 285−294.

(32) Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(33) Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 775−781.

(34) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100* (51), 19824−19839.

(35) *Cerius2 User Guide*; MSI: San Diego, USA, 1998.

(36) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490−519.

(37) *InsightII User Guide*; MSI: San Deigo, USA, 1998.

(38) James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual Daylight 4.62*; Daylight Chemical Information Systems Inc.: Los Altos, 2001.

(39) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Bobb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chem, W.; Wong, M. W.; Andres, J. L. Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(40) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 6*; University of California: San Francisco, 1999.

(41) Sanner, M. F.; Olson, A. J.; Spehner, J. *Biopolymers* **1996**, *38*, 305−320.

(42) Hou, T. J.; Wang, J. M.; Li, Y. Y.; Xu, X. J. *Chin. Chem. Lett.* **1998**, *9*, 651−654.

(43) Hou, T. J.; Wang, J. M.; Xu, X. J. *Chemometr. Intell. Lab.* **1999**, *45*, 303−310.

(44) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* **1994**, *98*, 1978−1988.

(45) Wang, J. M.; Hou, T. J.; Chen, L. R.; Xu, X. J. *Chemometr. Intell. Lab.* **1999**, *45*, 347−351.

(46) Hou, T. J.; Wang, J. M.; Chen, L. R.; Xu, X. J. *Protein Eng.* **1999**, *12*, 639−647.

(47) Hou, T. J.; Wang, J. M.; Xu, X. J. *Chin. Chem. Lett.* **1999**, *10*, 615−618.

(48) Hou, T. J.; Xu, X. J. *J. Mol. Graph. Model.* **2001**, *19*, 455−465.

(49) Weiser, J.; Peter S. S.; Still, W. C. *J. Comput. Chem.* **1999**, *20*, 217−230.