J Mol Model (2005) 11: 26–40
DOI 10.1007/s00894-004-0214-9

ORIGINAL PAPER

Tingjun Hou · Wei Zhang · Qin Huang · Xiaojie Xu

An extended aqueous solvation model based on atom-weighted solvent accessible surface areas: SAWSA v2.0 model

Received: 4 April 2004 / Accepted: 31 August 2004 / Published online: 24 November 2004

Abstract A new method is proposed for calculating aqueous solvation free energy based on atom-weighted solvent accessible surface areas. The method, SAWSA v2.0, gives the aqueous solvation free energy by summing the contributions of component atoms and a correction factor. We applied two different sets of atom typing rules and fitting processes for small organic molecules and proteins, respectively. For small organic molecules, the model classified the atoms in organic molecules into 65 basic types and additionally. For small organic molecules we proposed a correction factor of "hydrophobic carbon" to account for the aggregation of hydrocarbons and compounds with long hydrophobic aliphatic chains. The contributions for each atom type and correction factor were derived by multivariate regression analysis of 379 neutral molecules and 39 ions with known experimental aqueous solvation free energies. Based on the new atom typing rules, the correlation coefficient (*r*) for fitting the whole neutral organic molecules is 0.984, and the absolute mean error is 0.40 kcal mol^{-1}, which is much better than those of the model proposed by Wang et al. and the SAWSA model previously proposed by us. Furthermore, the SAWSA v2.0 model was compared with the simple atom-additive model based on the number of atom types (NA). The calculated results show that for small organic molecules, the predictions from the SAWSA v2.0 model are slightly better than those from the atom-additive model based on NA. However, for macromolecules such as proteins, due to the connection between their molecular conformation and their molecular surface area, the atom-additive model based on the number of atom types has little predictive power. In order to investigate the predictive power of our model, a systematic comparison was performed on seven solvation models including SAWSA v2.0, GB/SA_1, GB/SA_2, PB/SA_1, PB/SA_2, AM1/SM5.2R and SM5.0R. The results showed that for organic molecules the SAWSA v2.0 model is better than the other six solvation models. For proteins, the model classified the atoms into 20 basic types and the predicted aqueous free energies of solvation by PB/SA were used for fitting. The solvation model based on the new parameters was employed to predict the solvation free energies of 38 proteins. The predicted values from our model were in good agreement with those from the PB/SA model and were much better than those given by the other four models developed for proteins.

Keywords Solvation effect · SAWSA · PB/SA · GB/SA

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00894-004-0214-9

T. Hou · W. Zhang · Q. Huang · X. Xu (✉)
College of Chemistry and Molecular Engineering,
Peking University, Beijing, 100871, China
E-mail: xiaojxu@chem.pku.edu.cn

Introduction

The estimation of aqueous solvation free energy has been of long-standing practical interest in drug design, protein folding and stability analysis, and protein-ligand binding investigations [1–3]. Tremendous progress has been made in the development and validation of both implicit (continuum) [4–12] and explicit solvent models [13, 14]. Continuum solvation models employing molecular mechanics and semiempirical MO calculations have made it possible to assess desolvation faster than simulations employing explicit water. However, even these models are relatively computationally intensive and cannot be used routinely to assess rapidly the desolvation costs of large numbers of organic molecules such as combinatorial libraries. The invention of combinatorial synthesis and high-throughput screening has led to an increased need for quick and accurate determination or calculation of relevant physicochemical properties such as solvation free energy and lipophilicity of organic compounds. Moreover, the extensive applications of virtual screening based on molecular docking require universal methods for calculating solvation free energies for small molecules and macromolecules such as proteins, DNA or RNA, so that ligand binding can be estimated more quickly and precisely.

The fastest methods of calculating the solvation free energies are the charge-independent models based on atom or fragment addition [9–12]. The charge-independent models are of immense practical importance, since a fragment-additive or surface-based model can lower the required computer resources enormously. The earliest group contribution method proposed by Hine et al. [9] demonstrates that solvation free energies can indeed be predicted using additive models. They proposed a solvation model based on bond contribution and group contribution and discussed how these estimations can be used to assess the intrinsic hydrophilic character of organic compounds. But in order to construct a more predictive model, Hine et al. [9] eliminated some compounds with large predicted errors. Hine et al. [9] pointed out that the large deviations for these "outliers" were caused by long-range polar interactions. The model proposed by Hine et al. is quite simple and focuses on monofunctional organic molecules only. It is therefore not useful for estimating the desolvation cost of organic molecules in ligand-protein associations. In 1986, Eisenberg and McLachlan [10] developed a simple additive model based on the solvent accessible surface area to estimate the solvation free energies for protein. In 1986 and 1991, Ooi et al. [15] and Vila et al. [16] developed two models using a similar scheme to Eisenberg and McLachlan [10]. Since this time, additive-constitutive approaches have been used rarely. In 1999, Viswanadhan et al. [17] developed two group contribution methods employing atomic constants and molecular fingerprints based on the experimental database of aqueous solvation free energies. A database of 265 molecules with experimentally determined solvation free energies was used to derive the HLOGS and ALOGS models. In 1997, Hawkins et al. [18] proposed a solvation model for predicting aqueous free energies of solvation based entirely on geometry-dependent atomic surface tensions. In 1998, this model was extended to other solvents [19]. In 2001, Wang et al. [11] developed a model based on solvent accessible surface area, which can be used to predict the solvation free energies of both organic and biological molecules. Recently, Hou et al. [12, 20] proposed a method based on solvent accessible surface area (the SAWSA model), which can be used to predict the solvation free energies for both organic and biological molecules very quickly and precisely.

In the previous model, to achieve the best performance, we defined atom types for hydrogen. However, employing principles of physical chemistry, we should have given more elaborate definitions for the heavy atoms instead. Moreover, although the solvation free energies predicted by SAWSA model showed high linear correlation with those predicted by PB/SA, there were non-negligible differences if absolute values between the two methods. Thus, here we have re-evaluated the SAWSA parameters based on the new atom typing rules. In order to develop a more effective universal solvation model for proteins, we defined two different sets of atom typing rules: 65 atom types for organic molecules and 20 types for proteins. We expected that predictions of the solvation free energies would be improved based on the new atom typing rules. Additionally, we performed a systematic comparison of seven solvation models for small organic molecules (SAWSA v2.0, PB/SA_1, PB/SA_2, GA/SA_1, GB/SA_2, AM1/SM5.2R and SM5.0R) and five solvation models for proteins (SAWSA v2.0, Eisenberg, Wesson, Vila and Ooi). Hopefully this comparison will become the benchmark for evaluating such computational methods.

Method

Database preparation

In all, we used two data sets for the development and validation of solvation parameters for small organic molecules and proteins.

(1) The first data set includes 379 neutral organic molecules and 39 ions. The neutral organic molecules are divided into two sets: a training set with 293 molecules and a test set comprizing the remaining 86 molecules. The experimental free energies of solvation were collected from the literature (see Table S1 in the supporting materials) [8, 9, 10, 11, 12]. The molecular geometries of all compounds were modeled in the Cerius^2 molecular simulation package [21]. The initial structures were subjected to 1000 steps of conjugate-gradient molecular mechanics energy minimization using the MMFF force field [22]. Conformational analyses were performed for some molecules with flexible chains in order to find the global minimum geometries. For each molecule, only the global minimum conformation was used in the solvent accessible surface area (SASA) calculation and the subsequent parameterization. The MDL/SD files containing the 3D structural information and experimental solvation free energies were used to generate the molecular spreadsheet containing the SASA for each atom type. The MDL/SD database file including all small organic molecules cab be found in the supporting materials.

(2) The second data set includes 151 proteins randomly selected from the Brookhaven Protein Data Bank (PDB), from which 113 proteins were used as the training set to derive the solvation parameters, and 38 were used as the test set to evaluate the actual performance of the SAWSA v2.0 model for biopolymers. For these proteins, all crystallographic water molecules were eliminated from the structures. Some missing hydrogen atoms were added using the molecular design software InsightII, with a neutral sp^3 N terminus and a

carboxylic (COOH) C terminus assigned at neutral pH. Here it should be noted that all residues in the studied proteins are in neutral form [23]. Before commencing calculations, the structures were minimized using the AMBER force field and restraining the main chain to remove any steric overlap with [24]. The solvation free energies calculation is based on the solving of the Possion-Boltzmann (PB) equation and molecular surface area (SA) estimation was used as the standard value for parameterization. The electrostatic component of the solvation free energy was computed using the Delphi module in InsightII [6]. The program uses the finite difference method, which involves mapping the molecule onto a 3-D cubic grid within which the Poisson-Boltzmann (PB) equation must be satisfied at each point. As before, the electrostatic solvation energy is obtained as the difference between the results of two calculations. In the first, the region outside the solvent-accessible surface of the solute is a dielectric continuum of constant 1.0. In the second, the permittivity outside the molecule is 80. The difference in the electrostatic energy obtained from these two calculations provides the electrostatic component in solvation free energy calculation. The grid size was defined as 0.8×0.8×0.8 Å^3. The radius of the probe molecule was set to 1.4 Å. The partial charges used in the PB calculations were taken from the CFF91 force field [25]. The nonpolar contribution to solvation free energy is linear relative to SASA. The SASA values and the total solvation free energy were calculated using the solvation module in InsightII [23].

Atom typing rules

We defined two different kinds of atom classification systems, for small organic molecule and proteins, respectively:

1. The atom classification system for small organic molecule includes 65 atom types, of which 53 atom types used to classify carbon, nitrogen, oxygen, sulfur, phosphorus, and halogen atoms in neutral organic compounds and 12 types which are used to classify ions (Table 1).
2. The final atom classification system for proteins only includes 20 atom types (Table 2).

The definition of atom types is based on SMARTS. The atom types represented by SMARTS were determined by using the SMARTS system included in OELib, which is an open-source C++ library for small molecule chemical applications [26]. Based on the functionality in OElib, the programming was very simple. In the current work, two parameter files were used to store the SMARTS chains, one for atoms in small organic molecules and one for proteins. If we wanted to add some new typing rules or modify the typing rules, we would need to make some modifications to these parameter files.

Molecular surface areas were calculated using the MSMS program [27]. For proteins, the probe radius was set to 1.0 Å with a density of 3.0 vertex/Å^2. For small organic molecules, the probe radius was set to 0.5 Å with a density of 3.0 vertex/Å^2 as previously determined by us [12]. For each molecule, the atomic SASAs of the same atom types were added together.

The SAWSA model and parameters for the SAWSA model

In Viswanadhan's work, the authors used a simple atom-additive model based on the atom types used in the ALOGP method to calculate the solvation free energy [17, 28, 29]. The atomic solvation parameters were determined from the general equation, Eq. 1.

$$\Delta G_{\mathrm{wat}} = \sum_i a_i n_i \qquad (1)$$

where a_i is the contribution of atom type i, and n_i are the number of atoms (NA) with atom type i for a given molecule. Equation 1 has been widely used in most atom-additive approaches for the calculation of partition coefficient of small organic molecules.

In most previous work on the applications of the additive-constitutive models the calculations of solvation free energy have been based on the addition of the surface areas of atoms or fragments. For example, the models proposed by Eisenberg and McLachlan [10], Ooi et al. 15, Vila et al. 16, Wang et al. 11 and Hou et al. 12]. The solvation models based on the addition of surface areas construct direct connection between molecular conformation and solvation free energies. The solvation free energy of a molecule based on the addition of atom-weighted surface areas is described as

$$\Delta G_{\mathrm{wat}} = \sum_i b_i s_i \qquad (2)$$

where b_i is the contribution of atom type i, and s_i is the total *SAS* of type i.

Correction factors

For many systems, the model described by Eqs. 1 or 2 can give reasonably good results. However, we found that for many hydrocarbons or compounds with long hydrocarbon chains, the solvation free energies were often underestimated. The large deviation between experimental and predicted values may be accounted for the aggregation of these compounds through the inter- or intra-molecular group–group interactions in aqueous phase. In the current work, in order to consider the inter- and intra- molecular hydrophobic or van der Waals interactions, we introduced the correction factor of ‘‘hydrophobic carbon’’.

Table 1 Atom typing rules and their contributions to free energies of solvation for small organic molecules

Type	Description	Number of compounds	Frequency of use	Contribution1[a]	Contribution2[b]
Carbon in					
1	$\mathbf{C}\mathrm{H_4}$, $\mathbf{C}\mathrm{H_3R}$, $\mathbf{C}\mathrm{H_2R_2}$, $\mathbf{C}\mathrm{HR_3}$	251	731	0.215	0.477
2	$\mathrm{A_3}\text{–}\mathbf{C}\text{–}\mathrm{C=[N,O,S]}$, $\mathrm{A_3}\text{–}\mathbf{C}\text{–}\mathrm{C\equiv[C,N]}$	71	86	−2.626	−2.356
3	$\mathbf{C}\mathrm{A_3X}$, $\mathbf{C}\mathrm{A_2X_2}$, $\mathbf{C}\mathrm{H_3X}$, $\mathbf{C}\mathrm{H_2AX}$, $\mathbf{C}\mathrm{H_2X_2}$, $\mathbf{C}\mathrm{HA_2X}$, $\mathbf{C}\mathrm{HAX_2}$	217	327	−0.028	−0.019
4	$\mathbf{C}\mathrm{AX_3}$,$\mathbf{C}\mathrm{X_4}$, $\mathbf{C}\mathrm{HX_3}$	31	41	0.152	0.150
5	$\mathrm{R=}\mathbf{C}\mathrm{H_2}$	20	26	0.053	0.001
6	$\mathrm{R=}\mathbf{C}\mathrm{HA}$, $\mathrm{R=}\mathbf{C}\mathrm{A_2}$, $\mathrm{R=}\mathbf{C}\mathrm{H\text{–}c}$, $\mathrm{R=}\mathbf{C}\mathrm{H\text{–}(C=C)}$, $\mathrm{R=}\mathbf{C}\mathrm{(A)\text{–}c}$, $\mathrm{R=C(A)\text{–}C=*}$, $\mathrm{R=}\mathbf{C}\mathrm{(A)\text{–}C\equiv *}$, $\mathrm{R=}\mathbf{C}\mathrm{(A)\text{–}C=O}$	27	44	−0.141	−0.156
7	$\mathrm{R=}\mathbf{C}\mathrm{HX}$, $\mathrm{A=}\mathbf{C}\mathrm{AX}$, $\mathrm{A=}\mathbf{C}\mathrm{X_2}$	6	10	0.030	0.020
8	$\mathrm{R\equiv}\mathbf{C}\mathrm{H}$, $\mathrm{A\equiv}\mathbf{C}\mathrm{A}$	9	18	−0.041	−0.048
9	$\mathrm{[O,N]=}\mathbf{C}\mathrm{A_2}$, $\mathrm{[O,N]=}\mathbf{C}\mathrm{AX}$	16	16	−0.020	−0.117
10	$\mathrm{A\text{–}}\mathbf{C}\mathrm{OO}$	33	33	−0.083	−0.269
11	$\mathrm{O=}\mathbf{C}\mathrm{C_{sp2}}^{c}$, $\mathrm{O=}\mathbf{C}\mathrm{H\text{–}c}$, $\mathrm{O=}\mathbf{C}\mathrm{H\text{–}n}$, $\mathrm{O=}\mathbf{C}\mathrm{(A)\text{–}N}$	8	8	−0.041	−0.170
12	$\mathrm{O=}\mathbf{C}\mathrm{H_2}$, $\mathrm{O=}\mathbf{C}\mathrm{HA}$, $\mathrm{N=}\mathbf{C}\mathrm{HA}$, $\mathrm{S=}\mathbf{C}\mathrm{H_2}$, $\mathrm{S=}\mathbf{C}\mathrm{HA}$, $\mathrm{S=}\mathbf{C}\mathrm{A_2}$	16	16	−0.055	−0.190
13	$\mathrm{O=}\mathbf{C}\mathrm{HX}$, $\mathrm{[O,N]=}\mathbf{C}\mathrm{X_2}$	11	11	−0.032	−0.256
14	$\mathrm{c}\cdots\mathbf{c}\mathrm{H}\cdots\mathrm{c}$	93	411	−0.049	−0.047
15	$\mathrm{c}\cdots\mathbf{c}\mathrm{(c)}\cdots\mathrm{c}$, $\mathrm{c}\cdots\mathbf{c}\mathrm{(R)}\cdots\mathrm{c}$	54	74	−0.064	−0.102
16	$\mathrm{c}\cdots\mathbf{c}\mathrm{H}\cdots\mathrm{X_r}$, $\mathrm{A_r}\cdots\mathbf{c}\mathrm{H}\cdots\mathrm{X_r}$, , $\mathrm{A_r}\cdots\mathbf{c}\mathrm{(A)}\cdots\mathrm{X_r}$, $\mathrm{c}\cdots\mathbf{c}\mathrm{(R)}\cdots\mathrm{X_r}$	31	70	−0.667	−0.781
17	$\mathrm{c}\cdots\mathbf{c}\mathrm{(X)}\cdots\mathrm{c}$, $\mathrm{c}\cdots\mathbf{c}\mathrm{(X)}\cdots\mathrm{X_r}$ $\mathrm{X_r}\cdots\mathrm{c(X)}\cdots\mathrm{X_r}$	60	93	0.015	0.007
18	$\mathbf{C}\mathrm{H_4}$, $\mathbf{C}\mathrm{H_3R}$, $\mathbf{C}\mathrm{H_2R_2}$, $\mathbf{C}\mathrm{HR_3}$	21	81	−0.120	−0.124
$\mathrm{sp^3}$ oxygen in					
19	$\mathrm{R\text{–}}\mathbf{O}\mathrm{H}$	34	35	−0.521	−0.518
20	$\mathrm{c\text{–}}\mathbf{O}\mathrm{H}$, $\mathrm{N\text{–}}\mathbf{O}\mathrm{H}$, $\mathrm{R\text{–}}\mathbf{O}\mathrm{\text{–}X}$, $\mathrm{X\text{–}}\mathbf{O}\mathrm{\text{–}X}$	12	12	−0.505	−0.525
21	$\mathrm{R\text{–}}\mathbf{O}\mathrm{\text{–}R}$	27	33	−0.152	−0.135
22	$\mathrm{A\text{–}}\mathbf{O}\mathrm{\text{–}C=O}$	35	35	0.074	0.130
23	$\mathbf{O}\mathrm{\text{–}P}$	11	33	−0.057	−0.056
$\mathrm{sp^2}$ oxygen in					
24	$\mathbf{O}\mathrm{=C}$, $\mathbf{O}\mathrm{=c}$	80	81	−0.160	−0.137
25	$\mathbf{O}\mathrm{=[N, S, P]}$	4	4	−0.226	−0.208
Hydrogen in					
26	H	259	831	−0.009	−0.012
27	$\mathrm{H\ \text{–}}\mathbf{C}\mathrm{H_3}$, $\mathrm{H\text{–}}\mathbf{C}\mathrm{H_2R}(\pi=0)$	231	1195	0.003	0.002
28	$\mathrm{H\text{–}}\mathbf{C}\mathrm{H_1R_2}(\pi=0)$, $\mathrm{H\text{–}}\mathbf{C}\mathrm{H_1R_2}(\pi\neq 0)$, $\mathrm{H\text{–}}\mathbf{C}\mathrm{R_3}(\pi=0)$	166	832	−0.004	−0.004
29	$\mathbf{H}\mathrm{\text{–}CR_3}(\pi\neq 0)$	5	5	−0.026	−0.017
30	$\mathbf{H}\mathrm{\text{–}OH}$, $\mathbf{H}\mathrm{\text{–}SH}$	59	61	0.025	0.030
31	$\mathbf{H}\mathrm{\text{–}N}$	33	63	−0.112	−0.130
32	$\mathbf{H}\mathrm{\text{–}OC=O}$	6	6	−0.218	−0.232
33	$\mathbf{H}\mathrm{\text{–}C_{sp2}}$	151	582	0.001	0.006
Nitrogen in					
34	$\mathbf{N}\mathrm{_{sp2}}$	6	6	−0.254	−0.123
35	$\mathrm{R\text{–}}\mathbf{N}\mathrm{H_2}$, $\mathrm{X\text{–}}\mathbf{N}\mathrm{H_2}$	14	17	−0.211	−0.086
36	$\mathrm{O=C\text{–}}\mathbf{N}\mathrm{H_2}$	2	2	−0.716	−0.504
37	$\mathrm{R\text{–}}\mathbf{N}\mathrm{H\text{–}R}$	13	14	−0.412	−0.339
38	$\mathrm{O=C\text{–}}\mathbf{N}\mathrm{H}$	4	4	−1.201	−1.046
39	$\mathbf{N}\mathrm{R_3}$	7	8	−0.542	−0.476
40	$\mathrm{O=C\text{–}}\mathbf{N}\mathrm{A_2}$	2	2	−0.236	−0.120
41	$\mathbf{n}$	20	25	0.272	0.354
42	$\mathbf{n}\cdots\mathrm{c}\cdots\mathrm{n}$	7	12	−0.057	−0.017
Sulfur in					
43	$\mathrm{A\text{–}}\mathbf{S}\mathrm{H}$	4	4	−0.101	−0.109
44	$\mathbf{S}\mathrm{A_2}$	7	9	−0.042	−0.033
45	$\mathbf{S}\mathrm{=R}$	7	7	0.033	0.039
Phosphorus in					
46	P	11	11	−0.055	−0.044
Halogens in					
47	F	28	93	0.027	0.027

Table 1 (Contd.)

Type	Description	Number of compounds	Frequency of use	Contribution1[a]	Contribution2[b]
48	Cl	59	124	−0.014	−0.014
49	Br	29	37	−0.029	−0.028
50	I	8	9	−0.028	−0.026
United atom types					
51	A–**NO_2**	4	12	−0.082	−0.078
52	c–**NO_2**	6	18	−0.078	−0.080
53	–**CN**	7	14	−0.084	−0.088
Ions in					
54	**NH4+**	1	5		−1.265
55	**NH3+**	6	24		−1.498
56	**NH2+**	4	12		−2.181
57	**NH+**	4	8		−4.544
58	**NC+**	4	12		−1.540
59	**ND+**	4	8		−1.929
60	**NE+**	4	8		−2.545
61	**O-**	3	3		−4.846
62	**O**A-	1	1		−3.761
63	**COO-**	4	4		−3.786
64	**S-**	3	3		−3.530
65	**S**A-	1	1		−3.097
66	Correct factor			0.348	0.349

The atom described is shown in bold

R Any group linked through carbon, * any atom, *A* any atom except hydrogen, *X* any heteroatom (O, N, S, P and halogens), *c* aromatic carbon, *n* aromatic nitrogen, X_r aromatic atom except aromatic carbon, *o* aromatic oxygen, – single bond, = double bond, ≡ triple bond, ··· aromatic bond, $\pi=0$ the atom has π electrons, $\pi\neq 0$ the atom does not have π electrons, *sp*^{2} the hybridized state

[a] The solvation parameters using neutral molecules

[b] The solvation parameters using all molecules including ions

Here, we defined sp^3- or sp^2-hybridized carbon without any attached heteroatoms with the one to four relationship, as "hydrophobic carbons" (see Fig. 1). It should be noted that sp^2-hybridized aromatic carbons were not considered as hydrophobic carbons. Moreover, the sp^2-hybridized carbon in a ring was also not considered as a hydrophobic carbon because the sp^2-hybridized carbon in a ring is relatively rigid and does not have an easily adjustable conformation that would facillate aggregation.

After including the correction factor, the solvation free energy is given by

$$\Delta G_{\mathrm{wat}} = \sum_i b_i s_i + \sum_j c_j B_j \qquad (3)$$

where b_i and c_i are regression coefficients s_i the total *SASA* of atom type *i* and B_j is the number of the correction factor of atom type *j*.

Fitting procedure

In the current work, least-squares fitting was applied to derive the solvation parameters. It should be noted that all models developed in this paper do not consider the dielectric constants, so the parameters developed for water cannot be transferred to other solvents. For small organic molecules, we constructed four solvation models

Table 2 Atom typing rules and their contributions to free energies of solvation for proteins

Type	Description	Contribution
1	**C**H_4, **C**H_3R, **C**H_2R_2, **C**HR_3	0.425
2	A_3–**C**–C=A	−1.932
3	**C**A_3X, **C**A_2X_2, **C**H_3X, **C**H_2AX, **C**H_2X_2, **C**HA_2X, **C**HAX_2	−1.486
4	R=**C**HX, A=**C**AX, A=**C**X_2	0.608
5	**C**=O	0.340
6	**c**	0.193
7	**C**$_{\mathrm{sp2,\ 5R}}$[a]	0.147
8	R–**O**H	−0.371
9	A–**O**–C=O	−0.162
10	**O**=C	−0.238
11	**H**	−0.027
12	**H**–CH_1R_2($\pi=0$), **H**–CH_1R_2($\pi\neq 0$), **H**–CR_3($\pi=0$)	0.005
13	**H**–OH, **H**–SH	−0.476
14	**H**–N	−0.382
15	**N**$_{\mathrm{sp2}}$	−2.257
16	R–**N**H_2, R_2–**N**H, R_3–**N**, X–**N**H_2	0.967
17	**N**–O=C	0.233
18	**n**	−0.745
19	A–**S**H	−0.769
20	**S**A_2	−0.349

The atom described is shown in bold

R Any group linked through carbon, *A* any atom except hydrogen, *X* any heteroatom, *c* aromatic carbon, *n* aromatic nitrogen, – single bond, = double bond, $\pi=0$ the atom has π electrons, $\pi\neq 0$ the atom does not have π electrons, *sp*^{2} the hybridized state

[a] Carbon atom with *sp*^{2} hybridized state in five-membered ring

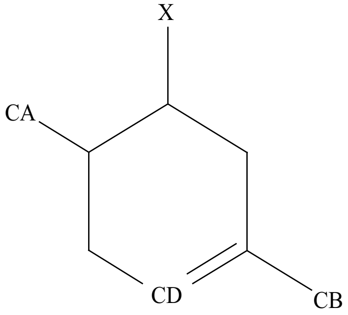

Fig. 1 The definition of hydrophobic carbons. Here CA, CB and CD are three carbon atoms; X represents a heteroatom. According to our definition, CB is a hydrophobic carbon, CA is not a hydrophobic carbon because a heteroatom is within four atoms and CD is not a hydrophobic carbon because CD is sp^2-hydridized and in a six-member ring

(see model I, model II, model III and model IV in Table 3). In model I and model II, the correction factor was used in the fitting process. In model I, the 293 neutral molecules in the training set were used to obtain the solvation parameters, and the test set was used to make the actual prediction. In model II, the parameters were derived using the data set with all molecules in Table S1 including the 39 charged ions. In model III, the correction factor for the whole data set was not considered in fitting. Models I, II and III were based on the addition of atom-weighted solvent-accessible surface areas (Eqs. 2 and 3). Moreover, we developed a model (model IV in Table 3) based on the simple addition of the number of atoms with the same atom type (Eq. 1).

For proteins, linear correlation was applied to minimize the differences between the predicted solvation free energies of SAWSA v2.0 and the predicted values of PB/SA. The solvation model used in fitting was based on Eq. 2.

Assessment of solvation models for small organic compounds

The molecular models of the 69 compounds in the test set were stored in MACCS/SD format and then used for the calculation. In Tables 4 and 5, the six solvation models used for comparison with SAWSA v2.0 can be divided into four categories: PB/SA solvation models, GB/SA solvation models, SM5.0R solvation model and AM1/SM5.2R solvation models.

Both PB/SA_1 and PB/SA_2 are based on solving the PB equation. A more detailed description of PB can be found else where [6]. The two PB/SA models were constructed using the AMBER force field. The only difference between those two PB/SA models is the usage of different sets of van der Waals parameters. The van der Waals parameters used in PB/SA_1 were developed by D. Sitkoff et al. [30], while those used in PB/SA_2 were the same as those used in the AMBER force field [24]. Partial charges for PB calculations were derived in order to maintain consistency with the AMBER charge derivation protocols. All the studied organic molecules after minimization of molecular mechanics were further optimized using quantum mechanics with the HF/6-31G basis set, and then, using

Table 3 Performance of the SAWSA v2.0 models by solute function class

Solute class	Model I		Model II		Model III		Model IV	
	Number	Error[a]	Number	Error	Number	Error	Number	Error
Alkanes	21	0.50	21	0.51	21	1.03	21	0.50
Alkennes	21	0.39	21	0.32	21	0.37	21	0.30
Alkynes	8	0.11	8	0.13	8	0.13	8	0.13
Aromatics hydrocarbon	18	0.44	18	0.62	18	0.60	18	0.57
Fluorides	21	0.58	21	0.61	21	0.62	21	0.57
Chlorides	39	0.31	39	0.33	39	0.35	39	0.29
Bromides	20	0.28	20	0.36	20	0.26	20	0.36
Iodinates	8	0.29	8	0.39	8	0.26	8	0.41
Alcohols	43	0.42	43	0.46	43	0.55	43	0.45
Ethers	20	0.51	20	0.52	20	0.59	20	0.62
Aldehydes	16	0.24	16	0.23	16	0.26	16	0.30
Ketones	17	0.19	17	0.20	17	0.20	17	0.18
Acids	6	0.17	6	0.16	6	0.18	6	0.31
Esters	29	0.20	29	0.22	29	0.21	29	0.22
Amines	28	0.36	28	0.34	28	0.52	28	0.50
Amides	6	0.09	6	0.10	6	0.20	6	0.30
Nitriles	6	1.02	6	0.79	6	0.94	6	0.73
Nitro compounds	7	0.30	7	0.40	7	0.34	7	0.30
Compounds with N in heterorings	23	0.63	23	0.73	23	0.71	23	0.72
Compounds with S	6	0.29	6	0.25	6	0.59	6	0.41
Compounds with P	11	1.12	11	1.10	11	1.10	11	0.73
Ions			39	1.65	39	1.62	39	1.82
Total	379	0.40	418	0.54	418	0.59	418	0.56

[a] Mean unsigned error (kcal mol^{-1})

Table 4 Predictions of seven aqueous solvation models using the compounds in the test set (in kcal mol^{-1})

Number	Name	Expt	SAWSA v2.0	PB/SA_1	PB/SA_2	GB/SA_1	GB/SA_2	SM5.0R	SM5.2R
3	Propane	1.96	1.16	1.04	1.03	0.80	0.90	1.49	1.23
5	2-methyl propane	2.32	1.64	1.00	0.97	0.25	0.75	1.89	1.53
6	2,2-dimethyl propane	2.50	2.19	0.98	0.94	−0.96	0.66	2.44	1.89
13	2,4-dimethyl pentane	2.59	2.48	1.11	1.08	−0.66	0.77	2.60	2.15
14	2,2,4-trimethyl pentane	2.88	2.65	1.27	1.24	0.14	0.99	2.60	2.42
20	n-heptane	2.62	2.13	1.68	1.66	1.38	1.46	2.04	2.03
21	n-octane	2.89	2.38	1.86	1.84	1.62	1.70	2.18	2.23
24	2-methyl propene	1.16	1.47	−0.29	−0.09	−1.76	−0.61	1.54	0.98
26	2-methyl-2-butene	1.31	1.06	0.44	0.50	−0.44	0.12	1.43	0.77
40	1,4-pentadiene	0.94	0.96	−1.61	−1.19	−2.23	−1.56	1.22	1.34
49	1-nonyne	1.05	1.03	−2.53	−1.42	−3.09	−2.13	0.74	0.13
56	o-xylene	−0.90	−1.46	−1.42	−1.09	−1.16	−1.31	−0.15	−2.11
59	1-propylbenzene	−0.30	−0.56	−1.14	−0.80	−0.74	−0.97	0.19	−0.93
64	Naphthalene	−2.41	−2.65	−2.95	−2.36	−2.07	−2.55	−2.00	−5.09
67	Phenanthrene	−3.40	−2.88	−3.28	−2.56	−1.58	−2.45	−2.67	−5.27
68	p-chlorotoluene	−1.92	−1.26	−1.33	−1.18	−0.95	−1.10	−0.52	−2.08
74	Octafluoropropane	4.28	3.77	1.00	0.23	3.90	−1.90	4.00	5.54
77	Chlorofluoromethane	−0.77	−0.13	−2.65	−2.58	−2.04	−2.27	0.46	−2.12
83	1,1,2,2-tetrachlorodifluoroethane	0.82	0.41	1.09	1.05	1.05	0.99	0.37	0.60
86	1,2-dichlorotetrafluoroethane	2.31	2.04	1.17	1.06	1.39	1.09	1.81	2.91
88	Bromotrifluoromethane	1.79	0.98	0.97	0.85	0.98	0.62	0.22	0.71
91	Dichloromethane	−1.36	−1.06	−2.11	−1.66	−1.55	−1.16	−1.36	−2.00
96	E-1,2-dichloroethane	−1.73	−1.06	−2.34	−2.19	−1.67	−1.52	−1.63	−1.43
102	Hexachloroethane	−1.40	−0.24	0.87	0.87	3.83	2.98	−1.31	−0.47
104	2-chloropropane	−0.24	−0.05	−1.34	−1.38	−1.33	−1.29	0.41	−0.57
116	Trichloroethylene	−0.44	−0.44	0.15	0.20	0.42	0.40	0.70	−0.31
118	3-chloropropane	−0.57	−1.02	−1.93	−1.74	−2.08	−1.70	−0.28	−0.34
122	1,3-dichlorobenzene	−0.98	−1.02	−1.03	−0.96	−0.37	−0.71	−0.95	−2.18
125	2,3-dichlorobiphenyl	−2.45	−2.70	−2.39	−2.01	−1.31	−1.88	−2.20	−4.27
127	Bromotrichloromethane	−0.93	−0.50	0.92	0.92	0.95	0.98	−1.03	−1.13
128	1-chloro-2-bromoethane	−1.95	−1.53	−1.58	−1.59	−0.80	−0.78	−2.01	−1.62
130	Dibromomethane	−2.11	−2.10	−1.03	−1.20	−1.44	−1.00	−2.25	−1.72
138	1-bromo-2-methylpropane	−0.03	−0.77	−0.53	−0.62	−0.72	−0.46	0.07	0.09
145	1,4-dibromobenzene	−2.30	−1.68	−0.50	−0.42	0.43	0.05	−1.21	−3.47
147	1-bromo-2-ethylbenzene	−1.19	−1.21	−0.93	−0.68	0.06	−0.46	−0.50	−2.89
151	Iodoethane	−0.72	−0.78	–	–	–	–	−0.52	−0.58
156	Iodobenzene	−1.73	−1.20	–	–	–	–	−2.51	−4.02
161	1-propanol	−4.85	−5.24	−4.51	−5.77	−4.14	−4.96	−4.64	−5.52
165	2,2,3,3,3-pentafluroproanol	−4.15	−3.21	−5.94	−8.37	−3.60	−8.68	−2.99	−5.79
167	2-methyl-1-propanol	−4.51	−4.63	−3.91	−5.34	−3.73	−4.46	−4.13	−5.03
171	2-methyl-1-butanol	−4.42	−4.45	−4.31	−5.74	−3.74	−4.40	−4.08	−4.88
181	4-methyl-2-pentanol	−3.74	−3.10	−2.93	−4.22	−3.73	−3.28	−2.90	−3.58
182	Cyclopentanol	−5.49	−5.64	−4.08	−5.50	−2.82	−3.97	−5.32	−5.72
186	4-heptanol	−4.01	−3.98	−2.11	−2.99	−1.15	−2.12	−3.58	−3.82
192	4-bromophenol	−7.00	−6.69	−6.49	−7.62	−4.34	−7.02	−6.36	−8.30
195	4-cresol	−6.12	−6.61	−6.53	−7.65	−5.12	−7.27	−5.81	−7.35
200	Dimethoxymethane	−2.93	−4.23	−3.62	−3.08	−2.27	−2.24	−3.04	−4.26
201	Methyl propyl ether	−1.66	−2.17	−1.31	−1.09	−0.43	−0.56	−1.52	−2.13
203	Methyl tert-butyl ether	−2.21	−1.43	−1.33	−1.18	−2.11	−0.74	−0.71	−1.64
206	Dipropyl ether	−1.16	−1.89	−0.81	−0.58	0.23	−0.06	−1.40	−1.61
207	Diisopropyl ether	−0.53	−1.05	−0.68	−0.63	−0.17	0.16	−0.77	−1.31
210	2-methyltetrahydrofuran	−3.30	−2.23	−1.95	−1.73	−0.70	−0.95	−2.76	−3.83
219	1-chloro-2,2,2-trifluroroethyl difluromethyl ether	0.11	0.12	−2.79	−2.99	−1.14	−3.30	−1.24	−3.25
221	Propanal	−3.44	−3.66	−5.22	−4.59	−5.35	−5.02	−3.61	−5.41
225	Heptanal	−2.67	−2.69	−4.51	−3.88	−4.57	−4.18	−3.05	−4.57
229	Trans-2-hexenal	−3.68	−3.50	−5.46	−4.75	−5.52	−5.33	−3.66	−5.96
236	2-butanone	−3.71	−3.84	−4.51	−4.00	−3.86	−4.06	−3.39	−6.55
239	2-pentanone	−3.52	−3.58	−4.38	−3.92	−3.64	−3.86	−3.23	−6.06
242	2,4-dimethyl-3-pentanone	−2.74	−3.07	−2.88	−2.65	−2.41	−2.80	−2.28	−4.51
246	4-heptanone	−2.93	−3.16	−3.68	−3.20	−2.58	−2.98	−2.77	−5.10
253	Propionic acid	−6.46	−6.74	−9.60	−10.76	−7.72	−10.40	−5.90	−8.16
257	4-amino-3,5,6-trichloropyridine-2-carboxylic acid	−11.96	−12.22	−11.22	−11.45	−4.82	−10.06	−14.21	−12.44
261	Methyl acetate	−3.31	−3.30	−6.19	−5.39	−4.73	−4.81	−3.60	−5.74
265	Ethyl acetate	−3.08	−3.03	−5.95	−5.18	−4.35	−4.51	−3.61	−5.43
270	Amyl acetate	−2.45	−2.34	−5.42	−4.69	−3.74	−3.95	−3.26	−4.53

Table 4 (Contd.)

Number	Name	Expt	SAWSA v2.0	PB/SA_1	PB/SA_2	GB/SA_1	GB/SA_2	SM5.0R	SM5.2R
278	Ethyl butyrate	−2.84	−2.66	−5.42	−4.60	−3.46	−3.77	−3.28	−4.48
284	Ethyl heptanoate	−2.30	−1.65	−4.53	−3.85	−2.86	−3.20	−2.87	−3.60
293	Dimethylamine	−4.28	−4.39	−0.98	−1.31	−0.81	−0.95	−3.57	−3.76
298	Triethylamine	−3.03	−2.65	0.92	0.62	1.19	1.45	−1.37	−1.46
301	N,N-dimethylpiperazine	−7.58	−7.31	−0.46	−0.87	0.63	−0.16	−7.38	−8.00
308	2-methoxyethanamine	−6.55	−6.61	−2.98	−3.33	−1.25	−2.70	−7.23	−7.65
309	Morpholine	−7.17	−6.16	−3.68	−3.66	−0.96	−1.92	−8.37	−8.68
317	3-methylpyridine	−4.77	−4.20	−2.56	−3.55	−2.02	−3.76	−3.85	−5.30
320	3-ethylpyridine	−4.60	−3.49	−2.47	−3.53	−1.29	−3.34	−3.67	−4.68
323	2,4-dimethylpyridine	−4.85	−4.32	−2.52	−3.44	−2.54	−3.31	−3.66	−5.69
327	3,5-dimethylpyridine	−4.84	−4.61	−2.31	−3.51	−2.20	−3.65	−3.48	−5.09
339	Propionitrile	−3.85	−2.67	−3.74	−5.23	−4.10	−6.13	−4.20	−4.72
341	2,6-dichlorobenzonitrile	−5.22	−3.49	−3.75	−4.51	−3.01	−4.66	−5.62	−8.21
355	Propionamide	−9.42	−9.47	−9.32	−8.26	−7.29	−8.69	−9.98	−14.48
358	1-propandethiol	−1.05	−1.40	−1.97	−2.52	−2.11	−2.25	−0.91	−0.57
365	Dimethyl disulfide	−1.83	−2.61	−1.69	−1.58	−1.66	−1.73	−1.61	−2.29
368	Triethyl phosphate	−7.80	−6.19	−9.99	−8.85	−7.78	−7.30	–	−16.66

the HF/6-31G* electrostatic potential (ESP), charges of the small organic molecules were obtained using Gaussian-98 [31]. It should be noted that in Gaussian the ESP charges were derived based on the scheme proposed by Merz and Kollman [32]. The basic principle of this method is to fit the electrostatic potential onto the molecular van der Waals surface. However, they do not supply the radii for iodine, so we were unable to derive the ESP charges for the molecules containing iodine.

The two GB/SA solvation models were also constructed in conjunction with the AMBER force field. Generalized Born (GB) calculations can be considered as a means of approximating finite difference PB free energies and related quantities [4, 7]. In this model the polar term of solvation free energy is represented by Eqs. 4 and 5.

$$\Delta G_{\mathrm{pol}} = -\frac{1}{2}\left(1-\frac{1}{\varepsilon_w}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{f_{\mathrm{GB}}} \tag{4}$$

$$f_{\mathrm{GB}} = \sqrt{r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j)} \tag{5}$$

In these equations, q_i and q_j are the partial charges of atoms i and j; ϵ_w is the solvent dielectric constant of the media; r_{ij} is the distance between atom i and atom j; R_i and R_j are the effective Born radii of atoms i and j. In its original form, R_i was estimated by a numerical integration procedure, but recently a pair-wise approximation calculation of effective Born radii has been reported, and is widely accepted for the estimation of protein solvation free energy [33].

The difference between these two GB/SA models is in the applications of two different sets of van der Waals parameters. The effective Born radii and the screening parameters used in GB/SA_1 were developed by Jayaram et al. [34] The effective Born radii used in GB/SA_2 were developed by Bondi, [35] while the screening parameters in GB/SA_2 were obtained from the TINKER molecular modeling package, version 3.6 [36].

Table 5 Comparison of aqueous solvation models for the prediction of the test set

Method	Acceptable[a]	Disputable[b]	Unacceptable[c]	Uncalculated[d]	UME[e]	MSD[f]	r[g]	SD[h]	F[i]
Charge-dependent									
PB/SA_1	31.71	28.05	37.80	2.44	1.44	1.85	0.80	0.63	133.78
PB/SA_2	32.93	34.15	30.49	2.44	1.34	1.77	0.82	0.68	165.02
GB/SA_1	32.50	26.83	39.02	2.44	1.72	2.36	0.65	2.32	57.40
GB/SA_2	30.49	29.27	37.80	2.44	1.52	2.08	0.74	2.05	96.09
SM5.2R	50.00	20.73	29.27	0.00	1.17	1.71	0.94	1.05	583.77
Charge-indepdendent									
SM5.0R	75.61	18.29	4.88	1.22	0.54	0.70	0.97	0.69	1391.41
SAWSA v2.0	84.15	13.41	2.44	0.00	0.43	0.57	0.98	0.56	2243.41

SD Standard deviations

[a] Percentage of acceptable results (estimation error < ±0.75)

[b] Percentage of disputable results (estimation error > ±0.75 and < ±1.50)

[c] Percentage of unacceptable results (estimation error > ±1.50)

[d] Percentage of uncalculated results

[e] Mean unsigned error

[f] Mean squared deviations

[g] Correlation coefficient between the experimental and calculated solvation free energies

[h] Fisher values

Moreover, the SM5.0R [18] and AM1/SM5.2R solvation models [37] within the AMSOL program [38] were also applied to estimate the free energies of solvation. The SM5.0R model is charge-independent, and predicts aqueous or organic solvation free energies based entirely on geometry-dependent atomic surface tension. The AM1/SM5.2R model predicts aqueous or organic solvation energies based on geometry-dependent atomic surface tensions and electrostatic polarization energies calculated with class II zero overlap Mulliken charges obtained from the wavefunction produced by either the MNDO, AM1, or PM3 semiempirical Hamiltonians. Here, we used the AM1 semiempirical Hamiltonian [39].

Assessment of solvation models for proteins

As a comparison, the solvation free energies for those proteins in the test set were calculated using four solvation models based on SASA, including the Ooi et al. [15] model, the Vila et al. [16] model, the Eisenberg and McLachlan [10] model and the Wesson and Eisenberg 40] model. The calculation of the solvation free energies using these four models was supported by the Solvation module in InsightII.

Results and discussion

Performance of the SAWSA v2.0 model for organic molecules

The initial model (model I in Table 3) based on the training set yields satisfactory results, $n = 293$, $r = 0.99$, $SD = 0.51$, $F = 9666.98$. Figure 2 shows a plot of observed versus calculated solvation free energies for the compounds in the training set. Moreover, this model expresses good predictive ability for the external test set ($r_{pred} = 0.98$, $SD = 0.56$) with an average absolute error of 0.43 units across a range of 16.24 units. The predictions are even better than those of the calibration set. Figure 3 shows the correlation between the observed and the calculated solvation free energies of the tested compounds. In total, the average error for the whole data set is 0.40 kcal mol^{-1}, which is much better than that obtained by us previously (0.52 kcal mol^{-1}) [12] and that obtained by Wang et al. (0.54 kcal mol^{-1}) [11]. Theis accurate prediction for the test set implys that the model is reliable and not overfitting. Figure 4 show a histogram of the deviation of the calculated values from the experimental results, with a near-Gaussian error distribution curve centered on zero.

In model II, the whole data set including the 39 ions was used in fitting. We achieved an average error of 0.54 kcal mol^{-1}, which is a greater error than that found using the neutral molecules only. Compared with model I, the increase of the mean unsigned error of model II is mainly due to the poor predictions for the ions. It should

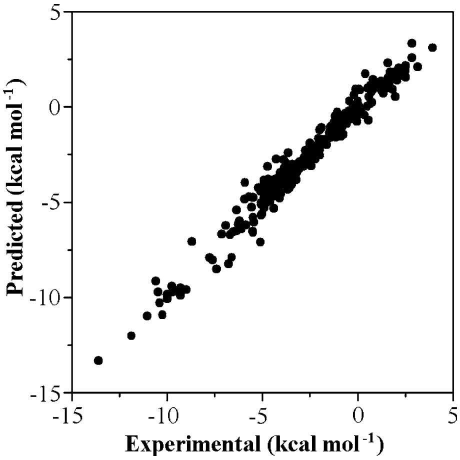

Fig. 2 Predicted and experimental aqueous solvation free energies for the 273 neutral molecules in the training set

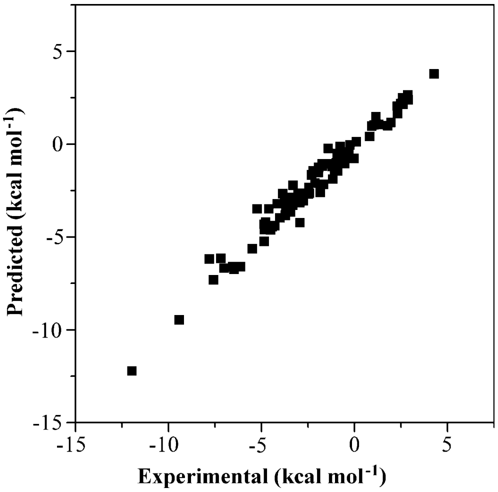

Fig. 3 Predicted and experimental aqueous solvation free energies for the 84 neutral molecules in the test set

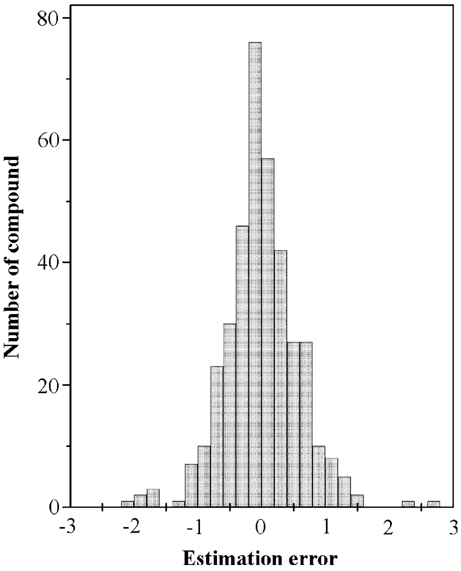

Fig. 4 Distribution histogram of the estimation errors

be noted that the ionic solutes have also be used with SM5.0R as reported by Hawkins et al. [37]. A mean unsigned error of 4.41 kcal mol^{-1} was obtained when using the SM5.0R/AM1 model [37]. However, in the current work, the mean unsigned error for the ionic solutes is only 1.65 kcal mol^{-1}, which is much better than using the SM5.0R/AM1 model and also a little better than that obtained in our previous work [12].

In order to investigate the influence of the correction factor, we constructed model III using the contributions from the atom-weighted surface areas only. Compared with model II, the predictive power of model III is obviously weaker and the mean unsigned error was increased from 0.54 to 0.59 kcal mol^{-1}, which means that the correction factor of ''hydrophobic carbon'' is very important in our model. In the first instance, we only introduced this correction factor for aliphatic and aromatic hydrocarbons, to improve the correlation of the model. However, we found that the results obtained with some heteroatom-containing compounds, especially those bearing long aliphatic chains, were greatly underestimated. We think these long aliphatic chains can also introduce inter-molecular aggregation. Therefore, we extended the ''hydrophobic carbon'' to all kinds of organic compounds: if there is no heteroatom at a certain range, a carbon atom is a ''hydrophobic carbon''. Adopting the new concept, the correlation of the model was further improved. Table 3 lists the unsigned average errors for models I, II and III by compound class. The calculated solvation free energies using model I, II and III are listed in Table S1 in the supporting material.

Model I, II and III are based on SASA according to Eqs. 2 and 3. However, in some atom-based additive methods, for example, the solvation model developed by Viswanadhan et al. [17] the authors only correlated the NA with the experimental solvation free energies according to Eq. 1. In fact, most prediction models for log P are based on the simple addition of NA. Here, we also propose a model based on Eq. 1 (see model IV in Table 3). The model based on NA is a little worse than the model based on SASA, but the difference is not very significant. It seems that both models based on NA and that based on SASA can generate good results. Indeed, for small organic molecules, these two models do not differ to any great extent because nearly all atoms in small compounds are exposed to solvent. Here it should be noted that although for small organic molecules the performances of the method based on NA and that based on SASA are similar, we think the method based on SASA should be a more universal model especially for large molecules such as proteins. For the method based on NA, all atoms of the same atom type are considered equivalent, which means that all atoms should be exposed to solvent. However, we know if some atoms in a molecule are surrounded by other atoms and located in the interior of a molecule, for example in peptides or proteins, then these atoms contribute little or even nothing to solvation free energies. If we use the method based on NA, these interior atoms are considered to be equivalent to the other atoms with the same atom types exposed to solvent. But if we use the method based on SASA, the contribution of the exposed atoms and the interior atoms can be separated effectively. Thus, the method based on NA is only applicable for small organic molecules, but for proteins, this method is completely meaningless.

Atom typing rules in the SAWSA v2.0 model

Any additive method, either by NA or SAS of atom types, needs a proper scheme for fragment/atom classification. The quality of such a classification scheme can be evaluated by how well the calculated solvation free energies agree with their experimental counterparts. To some extent, an additive method is the art of fragment/atom classification.

Here, the classification scheme differentiates atoms according to (1) element, (2) hybridization state, (3) nature of the neighboring atoms, and (4) adjacency to π systems. Thus, atoms belonging to the same atom type generally have similar charge densities. This establishes support for the assumption that a certain type of atom with similar surface area has a specific contribution to solvation. Compared with the atom classification scheme that we used previously, [12] the new scheme is more systematic and more easily understood. Our new scheme pays more attention to the definition of heavy atoms than to hydrogen atoms. In our previous work, in order to achieve the best correlation, we defined 16 atom types for hydrogen. Generally, the solvation free energies in neutral organic molecules are mainly derived from heavy atoms. According to the general principle of physical chemistry, the excessive definition of hydrogen atoms seems of little benefit. In our new scheme, we only defined 8 atom types for hydrogen atoms in neutral molecules. Moreover, we gave more elaborate definitions to atoms adjacent to any π-system, which proved to be important for affecting the charge densities. In the atom typing rules for small organic molecules, we used 65 atom types rather than the 58 atom types that we used previously. However, this number is still smaller than Viswanadhan's set of 67. Moreover, Viswanadhan et al. [17] only used a database of 265 molecules to derive the ALOGS model. Their database is much smaller than that used in our work. Additionally, the database used by Viswanadhan et al. [17] does not include any ions.

Comparison of the performance of seven solvation models

Our solvation model can only be evaluated from the correlation between the experimental solvation free energies and the calculated values. It is well known that the actual predictive power may only be determined based on a list of compounds in a test set. Moreover, we want to know if our model can give comparative prediction with other calculation procedures.

The calculated solvation free energies using SAWSA v2.0 and the other six solvation models are listed in Table 4. In assessing the calculated solvation free energies from the seven models, we used the following the criteria: (1) The individual estimation errors are grouped: errors less than ±0.75 kcal mol^{-1} are considered as acceptable, errors greater than ±0.75 kcal mol^{-1} and less than ±1.50 kcal mol^{-1} are considered as disputable and errors exceeding ±1.50 kcal mol^{-1} are considered as unacceptable. The missing calculations are also counted. All these results are given as a percentage of the entire test set. (2) The experimental and the calculated values are correlated using linear regression analysis. The statistical results (i.e. r, SD, and F-value) are recorded. The mean unsigned error (UME) and the mean squared deviations (MSD) are also calculated. All the results are summarized in Table 5.

We roughly divided the solvation models used into two categories: charge-dependent and charge-independent. From the correlation coefficients and the mean unsigned errors, the two charge-independent methods give better results than the other five charge-dependent ones. In those seven methods, SAWSA v2.0 performs best, as indicated by the high linear regression coefficient ($r = 0.98$) and low mean unsigned error (0.43 kcal mol^{-1}). Moreover, the other statistical tests produce better results for SAWSA v2.0 than those for the other solvation models.

The performance of SM5.0R is a little worse than that of SAWSA v2.0., but this method cannot give effective results for compound **368**, because SM5.0R does not include the solvation parameters for phosphorus. In the charge-dependent groups, the best performance is from AM1/SM5.2R. In fact, for most compounds in the test set, the prediction by AM1/SM5.2R is acceptable, but for several compounds, the predictions by AM1/SM5.2R are very poor, such as compound **355** (propionamide) and **368** (triethyl phosphate). We think that the large deviation for these molecules may be due to the deficient parameterization of AM1/SM5.2R. Such compounds similar to compounds **355** and **368** are not included in the training set, and some atom types in these compounds are not fully considered in parameterization. Moreover, SM5.2R uses uncorrected AM1 Mulliken charges for electrostatics, which are known to be extremely poor for certain functionalities.

The performance of the two GB/SA models is the worst (UME = 1.72 and 1.52 kcal mol^{-1}). As far as we know, the predictive power of the GB/SA model is significantly affected by two parameters, the initial Born radii and the screening parameters. GB/SA_1 adopted the parameters developed by Jayaram et al. [34]. In fact, the atom typing rules defined by Jayaram are very limited, and only include several atom types for H, C, O, N, S, P and Na^+. Thus, the GB/SA model based on Jayaram's parameters cannot perform well for compounds which contain any halogen atoms. In order to expand the predictive scope of GB/SA_1, we directly adopted the van der Waals radii of F, Cl and Br used in AMBER6.0 into the Jayaram's parameter set. The scale parameters and screening factors for F, Cl and Br are all set to 1.0. Interestingly, the predictions by GB/SA_1 of compounds containing F, Cl and Br are acceptable. Additionally, because the ESP charges cannot be derived for molecules with I, GB/SA also cannot give effective prediction for compounds **151** and **156**. Moreover, we also found that Jayaram's GB/SA model could only perform well for relatively simple organic molecules, and could not give good predictions for compounds with relatively complicated functional groups, for example, compounds containing pyrrole or pyridine functional groups. Overall, the current atom typing rules, the corresponding initial Born radii and screening parameters in the Jayaram's parameter set need improvement. The improvement of Jayaram's parameter set may be accomplished according to the following two aspects. Firstly, we should define more complete atom typing rules, for example, the sp^2 nitrogen and sp^3 nitrogen may be defined as two different atom types. Secondly, in parameterization, we should use a large training set to obtain reliable parameters for different atom types. In Jayaram's work, the authors adopt a training set of only 32 molecules for parameterization. Compared with GB/SA_1, the predictions by GB/SA_2 seem a little better. However, the parameters from Bondi and Tinker are also not very satisfactory [35, 36].

The performances of the two PB/SA models are better than those of the two GB/SA models, but their predictions are also not entirely satisfactory. For the molecules in the test set in Table 5, the mean unsigned errors for PB/SA_1 and PB/SA_2 are 1.44 and 1.34 kcal mol^{-1}, respectively. Because of the absence of parameters for iodine these two PB/SA models failed to give predictions for compounds **151** and **156**. For a PB model to be successful on a system, it is necessary that the van der Waals radii for each type should be parameterized to its optimal value. At present, the available van der Waals radii supported by Delphi are not good enough to produce the best prediction.

It should be noted that the predictions for these 69 compounds may not give a decisive rank of all these solvation models because the number of compounds in the test set is rather limited. However, the comparison does at least demonstrates that SAWSA v2.0 gives the best results among all these methods and yields acceptable estimations for the tested compounds.

SAWSA v2.0 model for proteins

Because of the availability of the solvation free energies of small organic molecules, researchers attempt to transfer the solvation parameters for small organic molecules to the prediction of protein solvation. Our previous work shows that the solvation parameters for small organic molecules can be used to rank the solvation abilities of proteins, but the absolute values calculated by SAWSA and PB/SA still exhibit large differences. The large differences between SAWSA and PB/SA may be simply due to the intrinsic differences between the chemical environment of the same atom types in small organic molecules and in proteins. So the simple transfer of the solvation parameters for organic molecules to the predictions of solvation of proteins may be not very suitable. In order to obtain more reasonable parameters we used a novel strategy for the parameterization of proteins. The solvation free energies were calculated using PB/SA and used as the standard values in the parameterization for proteins. Moreover, we defined new and simple atom typing rules for proteins. The new atom classification system includes 20 atom types. The obtained solvation parameters are shown in Table 2.

In our previous work, we found that the predictions of SAWSA for small organic molecules, were influenced by the probe radius in the SASA calculation. Here, the influence of the probe radius on the calculated results was also investigated, and different probe radii from 0.5 to 1.4 Å were used. From the calculated results, we found that the effect was not very significant. For the mean unsigned errors and the standard deviations, a probe radius of 1.0 Å was found to be the best, so a probe radius of 1.0 Å was applied for the SASA calculations. For the proteins in the training set, the calculated free energies of solvation and the experimental values show very good linear correlation with $r = 0.99$. Figure 4 shows the correlation between the experimental and calculated solvation free energies.

The solvation model for proteins was based on Eq. 2. For the training set, the SAWSA v2.0 model gives a mean unsigned error of 53.71 kcal mol^{-1}. The result of this model is further summarized by the scatter plot in Fig. 5. The experimental and calculated solvation free energies for the training set are listed in Table S2 in the supporting materials. The correlation in this figure is very good, $r = 0.999$, $SD = 71.31$, $F = 72367.36$. The predictive power of the new model was validated by the test set. Table S3 in the supporting materials reports the predicted solvation free energies for the test set using the new solvation parameters. Figure 6 shows the correlation between the calculated solvation free energies by PB/SA and the calculated values by SAWSA 2.0 for the proteins in the test set. The correlation in Fig. 6 is also very good, $r = 0.986$, $SD = 153.28$, $F = 1240.67$. The mean unsigned error for the test set is 116.33 kcal mol^{-1}. Although the correlation and the prediction for the test set is a little worse than those for the training set, the predictive power of the SAWSA v2.0 model using the new parameters is very good.

Here, the predictive abilities of four other models proposed by Eisenberg et al. Wesson et al. Ooi et al. and Vila et al. were also investigated [10, 15, 40, 16]. These four models have been widely used to empirically calculate the solvation effect for much large molecules such as protein. The predictions using these four models are listed in Table S3 in the supporting information.

The first model was proposed by Eisenberg and McLachlan [10]. In this model, the author defined five kinds of atom types for proteins, and the experimental free energies of transfer were used to derive the solvation parameters. The Eisenberg model is still widely used today owning to its simplicity. The correlation between the predictions of the Eisenberg model and those of the PB/SA model is shown in Fig. 7. The predictions by the Eisenberg model and PB/SA show a high linear correlation ($r = 0.967$), which is a little worse than that shown in Fig. 6. However, the absolute solvation free energies predicted using PB/SA and Eisenberg show large differences. If we do not consider the absolute discrepancy between those two models, the solvation abilities of the proteins in the test set can be effectively ranked by the

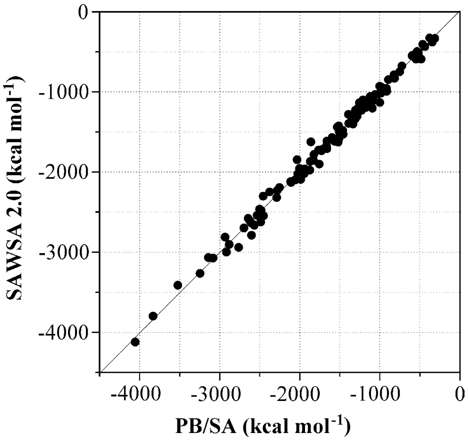

Fig. 5 Comparison of the predictions using the SAWSA v2.0 model and the PBSA model for the proteins in the training set

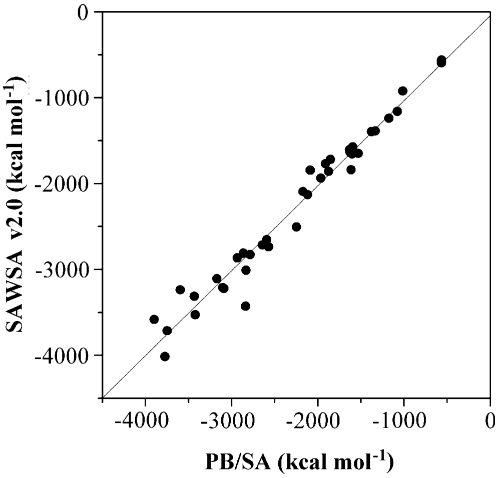

Fig. 6 Comparison of the predictions using the SAWSA v2.0 model and the PBSA model for the tested proteins

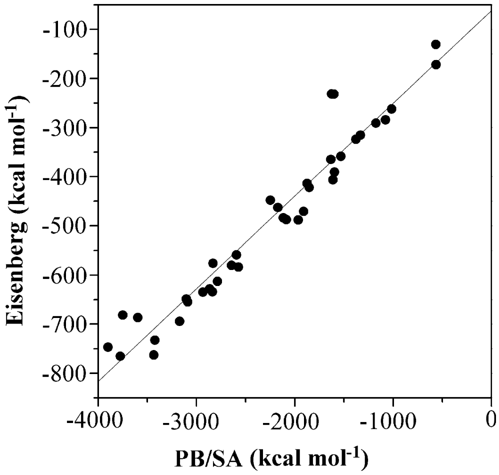

Fig. 7 Comparison of the predictions using the Eisenberg model and the PBSA model for the tested proteins

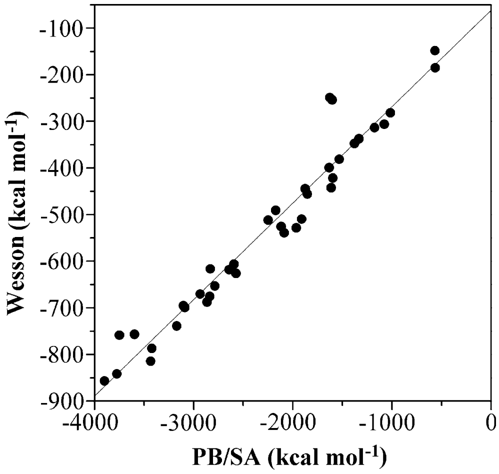

Fig. 8 Comparison of the predictions using the Wesson model and the PBSA model for the tested proteins

Eisenberg model. The Eisenberg model is widely employed in molecular simulation, but its applications in drug design is very limited, because this model is only developed for proteins. Therefore, the Eisenberg model cannot be applied in modeling docking and free energy calculations. The Wesson model is quite similar to the Eisenberg model, [40] and the atom typing rules and the solvation parameters used in these two models are nearly identical. As shown in Table S3 and Fig. 8, the predictions using the Wesson model are similar to those using the Eisenberg model.

The third model is produced by Ooi et al. [15]. They used, in part vapor-to-water transfer free energies of small solute molecules given by Cabani et al. [41] to find their atomic parameters. The model proposed by Ooi et al. [15] is based on united atom and contains seven types of atoms or groups that allow the method to be applied to a neutral training set. The correlation between the predictions with Ooi et al. [15] and those with *PB*/*SA* is shown in Fig. 9. The predicted values using these two models show an obvious linear correlation ($r = 0.901$), but the linear correlation is worse than that shown in Figs. 6, 7 and 8. That is to say, the predictive ability of the SAWSA v2.0 model for proteins may be significant better than that of the Ooi model. As for the Eisenberg and Wesson models, the absolute solvation free energies predicted using the Ooi model are also significantly different from those given by the PB/SA model.

The fourth solvation model was developed by Vila et al. [16] which is a revised version of the model proposed by Ooi et al. [16]. But compared with the Ooi model, two pairwise-distance-dependent modifications were applied: the atomic interaction using pairwise distances (AIPD) modification in which interactions between all atoms are taken into account, and the unified interacting side chains (UISC) modification in which the side chains are represented as unified interacting groups. Figure 6 shows the correlation between the predictions with Ooi and those with *PB*/*SA*. As shown in Fig. 10, the data show some linear correlation, but the correlation is very poor ($r = 0.651$). It seems that the Vila model cannot rank the solvation abilities of the proteins effectively.

In our work, we only developed solvation parameters for neutral proteins. But in most environments, some residues in proteins are charged, so the development of a solvation model for charged residues is very important. Recently, we have begun to develop a more universal model for charged proteins. This new solvation model will be reported in the near future. Moreover, we are trying to create a more effective atom classification system in order to improve the prediction.

Rather than comparing the predicted results of absolute free energies given by different salvation models, it would be more interesting to compare the structure and dynamics of protein using the SAWSA model

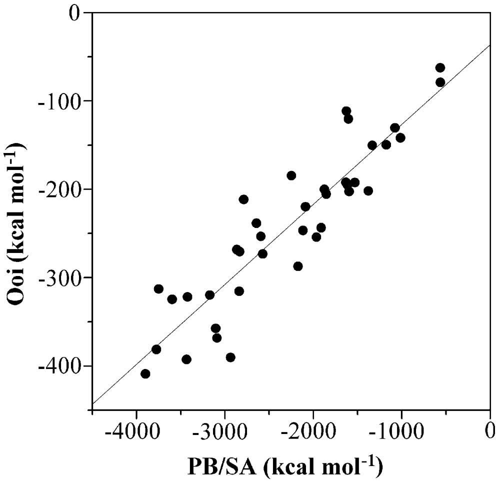

Fig. 9 Comparison of the predictions using Ooi and PBSA for the tested proteins

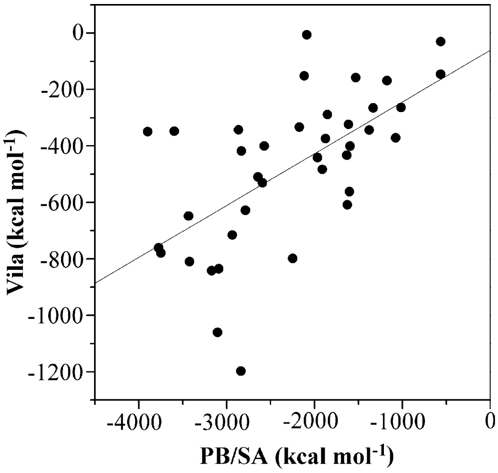

Fig. 10 Comparison of the predictions using Vila and PBSA for the tested proteins

using other solvation models, such as (1) explicit water, (2) PB/SA, or (3) GB/SA. In our future work, we will perform these comparisons.

Conclusions

In the current work, we have developed several solvation models based on atom-weighted solvent accessible surface areas for small organic molecules and proteins. For small organic molecules, an atom classification system with 65 atom types was used; and the experimental aqueous free energies of solvation were used in fitting. Moreover, we proposed a correction factor of "hydrophobic carbon" to account for the aggregation of hydrocarbons and compounds with long hydrophobic aliphatic chains. For small organic molecules, the prediction using the best solvation model based on all 379 neutral molecule set gives a mean unsigned error of 0.40 kcal mol^{-1}, which was better than the model proposed by us previously and the model proposed by Wang et al. [11]. The systematic comparison was performed on SAWSA v2.0, PB/SA_1, PB/SA_2, GB/SA_1, GB/SA_2, AM1/SM5.2R and SM5.0R. The calculated results showed that for organic molecules the SAWSA v2.0 model gave better results than the other six solvation models.

For proteins, an atom classification system with 20 basic atom types was used, and the predicted aqueous free energies of solvation for the PB/SA model were used in fitting. For the proteins in the training set, the solvation free energies from PB/SA and those from SAWSA v2.0 show a high linear correlation of $r = 0.999$ and a very low mean unsigned error of 53.71 kcal mol^{-1}. The solvation model based on the new parameters was used to predict the solvation free energies of 38 proteins. Although the test set has a lower mean unsigned error (116.33 kcal mol^{-1}) compared with that of the training set, the overall predicted values from our model were in good agreement with those from the PB/SA model, and were much better than those given by the other four models reported for proteins. Due to the simplicity and efficiency of the SAWSA v2.0 model, it may be widely used in many fields, including molecular docking, [42, 43] conformational analysis [44], protein folding and free energy calculations [45, 46, 47].

Supplementary material

The methods proposed here and all the parameters for calculations on solvation free energies have been incorporated into a computer program called SolAWSA. The SolAWSA computer code can be obtained by contacting the authors. In SolAWSA, two sets of solvation parameters are afforded: parm1.prm and parm2.prm. parm1.prm is used for the solvation of small organic molecules in water, and parm2.prm for the solvation of proteins in water. The SolAWSA program has been tested on IRIX and Linux operation systems. The experimental and calculated solvation free energies for small organic molecules are listed in Table S1. The structures of the training databases and the test set for small organic molecules are saved in MACCS/SD database files named data_set.sd (the SD database files include the experimental solvation free energies), which can be downloaded from internet freely. The experimental solvation free energies of the proteins in the training set and test set are listed in Table S2 and S3.

Acknowledgements Project (No. 29992590-2) supported by the National Natural Science Foundation of China.

References

1. Nemethy G, Peer WJ, Scheraga H (1981) Annu Rev Biophys Bioeng 10:459–497
2. Kellis JT, Nyberg Jr K, Fersht AR (1998) Nature 333:784–786
3. Alkorta I, Villar HO, Perez JJ (1993) J Comput Chem 14:620–626
4. Cramer CJ, Truhlar DG (1992) Science 256:213–217
5. Mohan V, Davis ME, MaCammon JA, Pettitt BM (1992) J Phys Chem 96:6428–6431
6. Honig B, Nicholls A (1995) Science 268:1144–1149
7. Still WC, Tempczyk A, Hawley RC, Hendrickson TA (1990) J Am Chem Soc 112:6127–6129
8. Cramer CJ, Truhlar DG (1991) J Am Chem Soc 113:8305–8311
9. Hine J, Mookerjee PK (1975) J Org Chem 40:292–298
10. Eisenberg D, McLachlan AD (1986) Nature 319:199–203
11. Wang JM, Wang W, Huo SH, Lee M, Kollman PA (2001) J Phys Chem B 105:5055–5067
12. Hou TJ, Qiao XB, Zhang W, Xu XJ (2002) J Phys Chem B 106:11295–11304
13. Allen MP, Tildesley DJ (1987) Computer simulations of liquids. Oxford University Press, London
14. MaCammon JA, Harvey SC (1987) Dynamics of proteins and nucleic acids. Cambridge University Press, Cambridge
15. Ooi T, Oobatake M, Nemethy G, Scherage HA (1987) Proc Natl Acad Sci U S A 84:3086–3090
16. Vila J, Vasquez M, Scherage HA (1991) Proteins 10:199–218
17. Viswanadhan VN, Ghose AK, Singh UC, Wendoloski JJ (199) J Chem Comp Sci 39:405–412

18. Hawkins GD, Cramer CJ, Truhlar DG (1997) J Phys Chem B 101:7147–7157
19. Hawkins GD, Liotard D A, Cramer CJ, Truhlar DG (1998) J Org Chem 63:4305–4313
20. Hou TJ, Xu XJ (2002) Acta Phys Chim Sin 18:1052–1055
21. Cerius2 User Guide (1998) MSI, San Deigo
22. Halgren TA (1996) J Comput Chem 17:490–519
23. InsightII User Guide (1998) MSI, San Deigo
24. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) J Am Chem Soc 117:5179–5197
25. Schmidt AB, Fine RM (1994) Mol Simulat 13:347–365
26. James CA, Weininger D, Delany J (2001) Daylight theory manual daylight 4.62. Daylight Chemical Information Systems Inc., Los Altos
27. Sanner MF, Olson AJ, Spehner J (1996) Biopolymers 38:305–320
28. Ghose AK, Crippen GM (1986) J Comput Chem 7:565–577
29. Ghose AK, Viswanadhan VN, Wendoloski J (1998) J Phys Chem B 102:3762–3772
30. Sitkoff D, Sharp KA, Honig B (1994) J Phys Chem 98:1978–1988
31. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratman RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu C, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson BG, Chen W, Wong MW, Andres JL, Gonzales C, Head-Gordon M, Replogle ES, Pople JA (1998) Gaussian 98. Gaussian Inc., Pittsburgh
32. Besler BH, Merz Jr KM, Kollman PA (1990) J Comp Chem 11:431–439
33. Hawkins GD, Cramer CJ, Truhlar DG (1996) J Phys Chem 100:19824–19839
34. Jayaram B, Sprous D, Beveridge DL (1998) J Phys Chem B 102:9571–9576
35. Bondi A (1964) J Phys Chem 68:441–451
36. Ponder J (1999) Tinker Molecular Simulation Package. http://dasher wustl.edu/tinker
37. Hawkins GD, Cramer CJ, Truhlar DG (1998) J Phys Chem B 102:3257–3271
38. Hawkins GD, Giesen DJ, Lynch GC, Chambers CC, Rossi I, Storer JW, Li JB, Zhu TH, Winget P, Rinaldi D, Liotard DA, Cramer CJ, Truhlar DJ (2002) AMSOL 6.7.2 University of Minnesota, Minneapolis, 55455–0431
39. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J AM Chem Soc 107:3902–3909
40. Wesson L, Eisenberg D (1992) Protein Sci 1:227–235
41. Cabani S, Mollica GV, Lepori L (1981) J Solution Chem 10:563–595
42. Wang JM, Hou TJ, Chen LR, Xu XJ (1999) Chemometr Intell Lab 45:281–286
43. Hou TJ, Wang JM, Chen LR, Xu XJ (1999) Protein Eng 12:639–647
44. Wang JM, Hou TJ, Chen LR, Xu XJ (1999) Chemometr Intell Lab 45:347–351
45. Hou TJ, Zhang W, Xu XJ (2001) J Phys Chem B 105:5304–5315
46. Hou TJ, Guo SL, Xu XJ (2002) J Phys Chem B 106:5527–5535
47. Hou TJ, Zhu LL, Chen LR, Xu XJ (2002) J Chem Inf Comp Sci 43:273–287