Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2007) 368–377
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

I-SOLV: A new surface-based empirical model for computing solvation free energies

Renxiao Wang *, Fu Lin, Yong Xu, Tiejun Cheng

State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 354 Fenglin Road, Shanghai 200032, PR China

Received 2 November 2006; received in revised form 3 January 2007; accepted 12 January 2007
Available online 17 January 2007

Abstract

We have developed a new empirical model, I-SOLV, for computing solvation free energies of neutral organic molecules. It computes the solvation free energy of a solute molecule by summing up the contributions from its component atoms. The contribution from a certain atom is determined by the solvent-accessible surface area as well as the surface tension of this atom. A total of 49 atom types are implemented in our model for classifying C, N, O, S, P, F, Cl, Br and I in common organic molecules. Their surface tensions are parameterized by using a data set of 532 neutral organic molecules with experimentally measured solvation free energies. A head-to-head comparison of our model with several other solvation models was performed on a test set of 82 molecules. Our model outperformed other solvation models, including widely used PB/SA and GB/SA models, with a mean unsigned error as low as 0.39 kcal/mol. Our study has demonstrated again that well-developed empirical solvation models are not necessarily less accurate than more sophisticated theoretical models. Empirical models may serve as appealing alternatives due to their simplicity and accuracy.

Keywords: Solvation free energy; Empirical model; Solvent-accessible surface; Atom type; I-SOLV

1. Introduction

Solvation refers to the partition between a gas phase and an aqueous solvent. The free energy change in this process is an essential thermodynamic quantity for characterizing solute molecules, which has significant implications on many subjects, such as protein folding and protein–ligand binding. A wide spectrum of theoretical models have been proposed in the past for computing this quantity. Historically, free energy perturbation (FEP) and thermodynamic integration (TI) were applied to compute the relative difference in the solvation free energies of structurally resembled molecules ($\Delta\Delta G_{solv}$) [1–3]. Both methods rely on extensive molecular dynamics or Monte Carlo sampling of solute molecules in explicit solvent. Such a simulation is unfortunately complicated to set up, and its accuracy is also limited by the force field and the sampling method employed in simulation. More importantly, for high-throughput projects such methods are computationally too expensive even for today's computers. Therefore, although methods like FEP and TI have theoretical significance, more practical solutions are much desired.

To simplify the description of aqueous environment and to reduce the degrees of freedom in simulation, a group of methods have been developed which treat the bulk solvent as a continuum media [4,5]. Among them, the Possion–Boltzmann (PB) method [6,7] and the Generalized Born (GB) method [8] are perhaps the two most popular options. The Possion equation describes the electrostatic potential $\varphi(\vec{r})$ generated by a charge distribution $\rho(\vec{r})$ in a continuum model of a polarizable solvent with dielectric constant $\varepsilon(\vec{r})$:

$$\nabla[\varepsilon(\vec{r})\nabla\varphi(\vec{r})] = -4\pi\rho(\vec{r})$$

Electrostatic energies are computed by integration of the solutions to this equation over the domain of interest. GB methods use analytical expressions based on the Born ion model to approximate electrostatic potentials of small molecules in solvent:

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

* Corresponding author. Tel.: +86 21 54925128.
E-mail address: wangrx@mail.sioc.ac.cn (R. Wang).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.01.006

When used with the dielectric screening algorithm introduced by Still et al. [9], GB models have been demonstrated to produce results close to those by PB models [4]. Note that both PB and GB methods only compute electrostatic energies of solute molecules, they are typically augmented by a term based on solvent-accessible surface (SAS) to account for the non-polar aspect in solvation free energy (for example, ref. [10]),

$$\Delta G_{\text{solv}} = \Delta G_{\text{elec}} + \Delta G_{\text{np}} = \Delta G_{\text{elec}} + \gamma \times \text{SAS}$$

Various extensions and adaptations of PB/SA and GB/SA models have been developed in recent years [11–23].

Models that rely completely on solvent-accessible surface areas have also been developed. One of the pioneering approaches is Eisenberg and McLachlan's model [24], which can be conceptually formulated as:

$$\Delta G_{\text{solv}} = \sum_i \sigma_i \text{SAS}_i$$

Here, SAS_i is the solvent-accessible surface area of atom i and σ_i is the so-called surface tension of atom i, which is in fact a weight factor. In such a model, the electrostatic and non-polar aspects in solvation free energies are taken into account implicitly by using atom types. A total of five atom types were used in Eisenberg and McLachlan's model, including carbon (C), neutral oxygen or nitrogen (O/N), charged oxygen (O^-), charged nitrogen (N^+), and sulfur (S). The surface tension of each atom type was derived from a regression analysis of the known solvation free energies of amino acid residues. Due to its simplicity, this model is still being widely used today in protein folding and binding studies.

This method has been extended by Wang et al. [25] and Hou et al. [26] to compute solvation free energies of small organic molecules. Basically, more extensive atom typing schemes were employed, and the surface tension of each atom type was derived from large data sets containing various organic molecules. Hawkins et al. also developed a pure surface-based model, i.e. SM 5.0R [27,28], in which atomic surface tensions were determined by atom types as well as local molecular geometries. Apparently, surface-based models have the advantage in speed. Their results are also fairly accurate on the data sets used for their parameterization. For example, Hou et al. recently conducted a comparison of their model with several standard PB/SA and GB/SA models on a common test set [29]. Their results demonstrated that surface-based models were even more accurate than computationally more expensive PB/SA or GB/SA models.

Inspired by the promising performance of surface-based models, we have developed a new empirical model for solvation free energy computation following the same approach. This model is a component of our in-house software package called Integrated Toolkits for Drug Design (ITDD), and will be referred to as I-SOLV throughout this manuscript. I-SOLV is based on solvent-accessible surfaces of united-atoms and utilizes a new set of atom types. It is parameterized with 532 neutral organic molecules with experimentally measured solvation free energies, the largest data set publicly reported so far for developing/validating solvation models. As validated on a test set, I-SOLV produced better results than other surface-based models as well as several PB/SA and GB/SA models.

2. Methods and results

2.1. Data preparation

Experimentally determined solvation free energies of organic molecules were cited from refs. [21,23,25]. Removing the duplicates in the data sets used in those studies resulted in a compilation of 532 neutral organic molecules. All of these organic molecules, along with their experimentally determined solvation free energies, are tabulated in Supplementary material. Most of them are small to medium-sized: the average number of non-hydrogen atoms of these molecules is 7.2 ± 2.7; the average number of rotatable single bonds of these molecules is 1.4 ± 1.9 (see Supplementary material for more details). A structural model of each molecule was constructed by using the SYBYL software [30]. Since most of these molecules are not conformationally complicated, it is not necessary to attempt exhaustive conformational sampling to determine the lowest-energy conformation of each molecule. In our study, each molecule, if flexible, was simply constructed in its most extended conformation and then optimized in vacuum using the MMFF94 force field. The final optimized model was used in subsequent surface generation.

The Lee–Richards solvent-accessible surface [31] was adopted in our model. This type of surface is generated by rolling a spherical probe on the van der Waals surface of the given molecule and then tracing the center of the probe. In our study, the Lee–Richards solvent-accessible surface of each molecule was generated by an in-house computer program. Such a job could be roughly divided into two steps. First, the surface of each component atom was generated, which was represented by a set of dots evenly distributed on a sphere centered at the given atom. The classical atomic radii set proposed by Bondi [32] were adopted in our study, i.e. C(1.70 Å), N(1.55 Å), O(1.52 Å), P(1.80 Å), S(1.80 Å), F(1.47 Å), Cl(1.75 Å), Br(1.85 Å) and I(1.98 Å). The probe radius was set as 0.50 Å. A uniform density of four dots per $Å^2$ was applied to all atoms. Then, the overlapping areas on each atom were removed, and the remaining accessible areas were integrated for subsequent analyses.

2.2. Atom typing scheme

Atoms are grouped into different types according to their intrinsic properties and local chemical environment. Atoms of the same type are assumed to have a uniform contribution to solvation free energy. Our atom typing scheme uses a total of 49 atom types for C, N, O, P, S, F, Cl, Br and I (see Table 1). This scheme is essentially an extension to the basic atom typing scheme defined in the SYBYL software. A given atom is classified by considering the following aspects in order: (i) its element type, (ii) its hybridization state, (iii) its accessibility to

Table 1
The atom typing scheme implemented in I-SOLV

ID	Symbol	Model A[a]		Model B[b]		Description[c]
		Occurrence[d]	Surface tension (kcal/mol Å^2)	Occurrence	Surface tension (kcal/mol Å^2)	
1	C.3.3h.lipo	210	0.017	176	0.017	sp^3 lipophilic carbon C^*H_3R
2	C.3.2h.lipo	146	0.011	129	0.011	sp^3 lipophilic carbon $C^*H_2R_2$
3	C.3.h.lipo	31	0.006	27	0.004	sp^3 lipophilic carbon C^*HR_3 or C^*R_4
4	C.3.3h	147	0.002	126	0.003	sp^3 carbon C^*H_3R
5	C.3.2h	159	−0.020	138	−0.020	sp^3 carbon $C^*H_2R_2$
6	C.3.h	22	−0.075	17	−0.074	sp^3 carbon C^*HR_3 or C^*R_4
7	C.3.3h.X	66	−0.016	58	−0.018	sp^3 carbon C^*H_3X
8	C.3.2h.X	158	−0.030	129	−0.031	sp^3 carbon C^*H_2RX or $C^*H_2X_2$
9	C.3.h.X	67	−0.081	56	−0.094	sp^3 carbon C^*HR_2X, C^*HRX_2, C^*HX_3, C^*R_3X, $C^*R_2X_2$, C^*RX_3 or C^*X_4
10	C.2.2h	26	0.012	23	0.012	sp^2 aliphatic carbon H–C^*(=A)–H
11	C.2.h.(=C)	28	−0.013	24	−0.013	sp^2 aliphatic carbon H–C^*(=C)–R
12	C.2.h.(=C).X	7	−0.024	6	−0.027	sp^2 aliphatic carbon H–C^*(=C)–X
13	C.2.h.(=X)	19	0.008	16	0.008	sp^2 aliphatic carbon H–C^*(=X)–R
14	C.2.h.(=X).X	8	−0.052	8	−0.054	sp^2 aliphatic carbon H–C^*(=X)–X
15	C.2.(=A)	32	−0.090	26	−0.097	sp^2 aliphatic carbon R–C^*(=A)–R
16	C.2.(=A).X	44	−0.214	36	−0.224	sp^2 aliphatic carbon R–C^*(=A)–X
17	C.2.(=A).2X	7	0.022	6	0.018	sp^2 aliphatic carbon X–C^*(=A)–X
18	C.ar.h	179	−0.010	162	−0.008	Aromatic carbon A≈C^*(–H) ≈A in a five- or six-member aromatic ring
19	C.ar	168	−0.038	152	−0.052	Aromatic carbon A≈C^*(–A) ≈A in a five- or six-member aromatic ring
20	C.ar.ar	16	−0.105	14	−0.116	Aromatic carbon A≈C^*(≈A) ≈A in a five- or six-member aromatic ring
21	C.1.h	9	−0.001	9	0.001	sp carbon –C≡C^*–H
22	C.1	20	0.027	20	0.023	sp carbon –C≡C^*–A
23	N.3.2h.pi	22	−0.149	20	−0.149	sp^3 nitrogen A–N^*H_2 connected to a conjugated moiety
24	N.3.h.pi	7	−0.260	7	−0.260	sp^3 nitrogen A_2–N^*H connected to a conjugated moiety
25	N.3.pi	4	−0.649	4	−0.636	sp^3 nitrogen A_3N^* connected to a conjugated moiety
26	N.3.2h	12	−0.128	11	−0.128	sp^3 nitrogen A–N^*H_2
27	N.3.h	12	−0.174	10	−0.168	sp^3 nitrogen A_2–N^*H
28	N.3	7	−0.294	5	−0.244	sp^3 nitrogen A_3N^*
29	N.2	16	−0.265	16	−0.247	sp^2 aliphatic nitrogen –N^*=A
30	N.ar.5	6	−0.278	6	−0.280	Aromatic nitrogen on a five-member aromatic ring
31	N.ar.6	23	−0.231	18	−0.233	Aromatic nitrogen on a six-member aromatic ring
32	N.ar.6.2	5	−0.149	5	−0.147	Aromatic nitrogen on a six-member aromatic ring if there is another hetero-atom on the same ring, such as pyridazine, pyrimidine and pyrazine
33	N.1	12	−0.122	10	−0.116	sp nitrogen –C≡N^*
34	O.3.h.acid	7	−0.047	5	−0.056	sp^3 oxygen A–O^*–H in an acid group
35	O.3.h	76	−0.180	67	−0.178	sp^3 oxygen A–O^*–H
36	O.3.ether	40	−0.050	31	−0.046	sp^3 oxygen R–O^*–R in an ether group
37	O.3.ester	35	0.213	30	0.216	sp^3 oxygen R–O^*–R in an ester group
38	O.3.X	12	−0.355	11	−0.360	sp^3 oxygen R–O^*–X or X–O^*–X
39	O.2.(=C)	99	−0.143	84	−0.141	sp^2 oxygen –C=O^*
40	O.2.(=N)	16	−0.037	16	−0.038	sp^2 oxygen –N=O^*
41	O.2.(=P)	5	0.144	4	0.187	sp^2 oxygen –P=O^*
42	S.3.h	5	−0.017	4	−0.019	sp^3 sulfur A–S^*–H
43	S.3	13	−0.015	12	−0.018	sp^3 sulfur A–S^*–A
44	S.2	7	0.165	7	0.175	sp^2 sulfur –A=S^*
45	P.3	10	0.790	9	0.589	Phosphor A_3–P^*(=A)
46	F	31	0.024	24	0.024	Fluorine A–F^*
47	Cl	84	−0.001	67	0.000	Chlorine A–Cl^*
48	Br	33	−0.007	25	−0.005	Bromine A–Br^*
49	I	11	−0.007	9	−0.006	Iodine A–I^*

[a] Results derived from all of the 532 molecules.
[b] Results derived from the 450 molecules after excluding the 82 molecules in the test set.
[c] The following symbols are used: (–) single bond, (=) double bond, (≡) triple bond, (≈) aromatic bond, (R) group linked through a carbon atom, (X) a group linked through a hetero-atom, (A) group linked through any atom. The asterisk indicates the relevant atom.
[d] Total number of molecules containing this atom type.

the solvent, which is measured by the number of attached hydrogen atoms and (iv) its chemical environment, such as the existence of neighboring hetero-atoms. Note that our atom typing scheme is based on the concept of united-atoms, i.e. hydrogen atoms are included implicitly in their root atoms. No atom type is thus necessary for classifying hydrogen atoms.

Our atom typing scheme is generally straightforward to understand. However, the definition of ‘‘lipophilic carbon

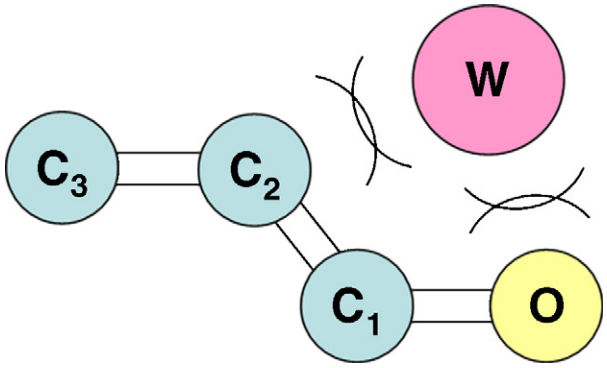

Fig. 1. Definition of "lipophilic carbon atoms". C_2 is not defined as "lipophilic carbon atom" since it can still be affected by a solvent molecule (W) anchored by a nearby hetero-atom (an oxygen atom in this case). C_3 is defined as a "lipophilic carbon atom".

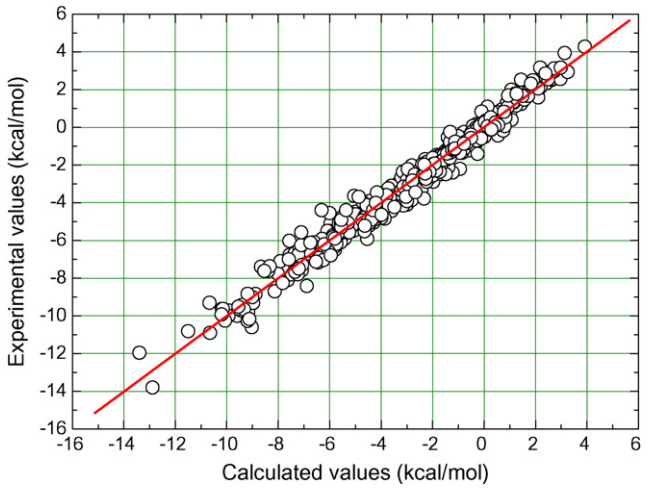

Fig. 2. Correlation between experimentally measured solvation free energies and fitted values by I-SOLV for the entire data set ($N = 532$, $R = 0.989$, $SD = 0.47$ kcal/mol, $MUE = 0.35$ kcal/mol).

atoms" needs to be further explained. In our scheme, sp^3 hybridized carbon atoms that are separated from any hetero-atoms by at least three chemical bonds are defined as "lipophilic carbon atoms" (atom types 1–3 in Table 1); while other sp^3 hybridized carbon atoms that are not covalently connected to any hetero-atoms are classified differently (atom types 4–6 in Table 1). A carbon atom in types 1–3 is apparently more isolated from any hetero-atoms as compared to its counterpart in types 4–6 (see Fig. 1). It is thus expected to be more lipophilic. A similar concept of "lipophilic carbon atoms" was also applied successfully in our empirical models for computing octanol–water partition coefficients ($\log P$) [33,34].

2.3. Model development

We assume that: (1) the solvation free energy of a solute molecule can be computed by summing up the contributions from its component atoms; (2) the contribution from a certain atom is characterized by its "surface tension", which is determined by the atom type applicable to this atom; (3) if treating solvent as a continuum media, the magnitude of the contribution from a certain atom is proportional to the solvent-accessible surface area of this atom. Based on these assumptions, the solvation free energy of a given solute molecule is computed as:

$$\Delta G_{\text{solv}} = \sum_{i=1}^{N} \sigma_i \text{SAS}_i \quad (1)$$

Here, SAS_i is the solvent-accessible surface area of atom i and σ_i is the surface tension, i.e. weight factor, of atom i. Note that only non-hydrogen atoms are considered in the above equation.

In order to derive the surface tension of each atom type, we performed a least-square multivariate regression analysis on all of the 532 molecules in our data set by applying Eq. (1). It produced a correlation coefficient (R) of 0.989, a standard deviation (SD) of 0.47 kcal/mol, a mean unsigned error (MUE) of 0.35 kcal/mol, and a Fisher-value (F) of 423 (Table 2). The correlation between experimentally measured solvation free energies and fitted values of these molecules is plotted in Fig. 2. Regression coefficients of all 49 atom type from this analysis are summarized in Table 1. This regression model is referred to as Model A in this paper, which is the final form of I-SOLV.

A leave-one-out cross-validation was also performed on the entire data set in order to test the predictive power of our model. This cross-validation produced a correlation coefficient

Table 2
Comparison of the regression results of several surface-based models

Model	Number of adjustable parameters[a]	Number of molecules in regression[b]	Correlation coefficient (R)	Standard deviation (kcal/mol)	Mean unsigned error (kcal/mol)	Reference
I-SOLV	49	532	0.989	0.47	0.35	–
SAWSA 2.0	54	379	0.984	–[c]	0.40	29
SAWSA 1.0	46	377	0.97	0.70	0.52	26
WSAS	47	387	–	0.79	0.54	25
SM 5.0R	24 + 17[d]	248	–	–	0.53	27, 28

[a] For neutral atom types only. SAWSA and WSAS use a number of charged atom types as well.
[b] For neutral solute molecules only.
[c] Not given in the corresponding reference.
[d] SM 5.0R uses 24 parameters for atomic surface tensions and 17 additional parameters to measure the impacts of local molecular geometries on atomic surface tensions.

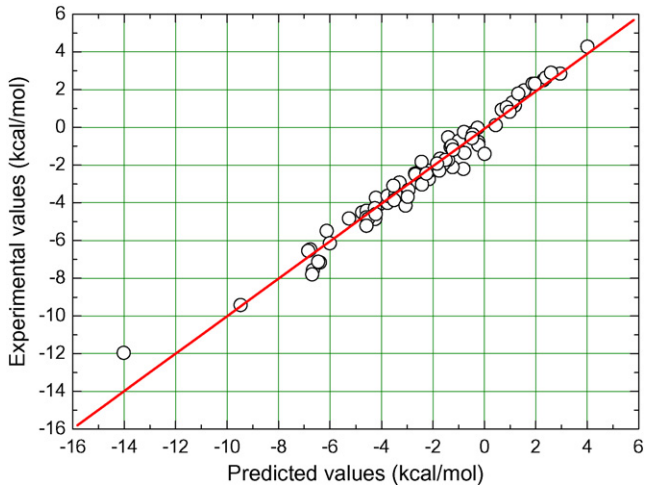

Fig. 3. Correlation between experimentally measured solvation free energies and predicted values by I-SOLV for the test set ($N = 82$, $R = 0.985$, $SD = 0.52$ kcal/mol, $MUE = 0.39$ kcal/mol, $Y = 0.99 \times X - 0.08$).

between experimental and predicted values of 0.985, a standard deviation of 0.54 kcal/mol, and a mean unsigned error of 0.40 kcal/mol.

2.4. Validation

In order to further test the performance of our model, a total of 82 molecules were selected from our data set as a test set (see Supplementary material). Note that this test set is identical, in terms of contents, to the one used by Hou et al. in their recent comparative evaluation of some solvation models [29] so that we can make direct comparison of our results with theirs.

In order to derive a model which is independent on this test set, another least-square multivariate regression analysis was performed on the remaining 450 molecules in our data set. It produced a correlation coefficient (R) of 0.989, a standard deviation (SD) of 0.47 kcal/mol, and a mean unsigned error (MUE) of 0.35 kcal/mol. Regression coefficients of all 49 atom type from this analysis are also summarized in Table 1, which is referred to as Model B in this paper. Model B was then applied to predict the solvation free energies of the 82 molecules in the test set. The correlation between experimentally measured and predicted solvation free energies of the test set is plotted in Fig. 3. It produced a correlation coefficient (R) of 0.985, a standard deviation (SD) of 0.52 kcal/mol, a mean unsigned error (MUE) of 0.39 kcal/mol. These results are also summarized in Table 3 together with the statistical results given by several other solvation models on the same test set.

2.5. PB/SA and GB/SA computations

We also applied standard PB/SA and GB/SA models to the test set used in our study. Each molecular model in our test set was further optimized at the HF/6-31G* level by using the Gaussian 98 program [35]. The LanL2DZ basis set was also adopted for processing molecules containing bromine and iodine atoms. Electrostatic potentials of all molecules were computed at the same level as structural optimization. Atomic point charges were consequently derived by using the restrained electrostatic potential (RESP) module in the AMBER program [36].

In all PB/SA computations, the electrostatic component (ΔG_{elec}) was calculated using the DELPHI program (Version 4) [37,38] with the PARSE atom radii set. Dielectric constant was set to 1.0 for the solute molecule and 80.0 for solvent, respectively. The non-polar component (ΔG_{np}) was computed by the algorithm proposed by Sitkoff et al. [10], i.e. $\Delta G_{np} = 0.00542 \times SAS + 0.092$. Solvent-accessible surface (SAS) areas were computed by using the MSMS program [39] with a probe radius of 1.4 Å and a density of 3.0 vertex/Å^2. In all GB/SA computations, the electrostatic component (ΔG_{elec}) was computed using the MM-GB/SA module in the AMBER program. The non-polar component (ΔG_{np}) was computed as $\Delta G_{np} = 0.0072 \times SAS$ according to the default setting in AMBER. Correlations between experimentally measured solvation free energies and predicted values by our PB/SA and GB/SA computations are plotted in Figs. 4 and 5, respectively. The statistical results of these computations are summarized in Table 3.

Table 3
Performance of several solvation models on the test set ($N = 82$)

Model	Correlation coefficient (R)	Standard deviation (kcal/mol)	Mean unsigned error (kcal/mol)	Acceptable (%)[a]	Disputable (%)[b]	Unacceptable (%)[c]	Uncalculated (%)
I-SOLV	0.985	0.52	0.39	89.0	9.8	1.2	0.0
SAWSA 2.0[d]	0.98	0.57	0.43	84.2	13.4	2.4	0.0
PB/SA	0.891	1.37	2.47	9.8	19.5	70.7	0.0
PB/SA (Hou)[d]	0.80	1.85	1.44	31.7	28.1	37.8	2.4
GB/SA	0.641	2.33	1.92	29.3	18.3	52.4	0.0
GB/SA (Hou)[d]	0.65	2.36	1.72	32.5	26.8	39.0	2.4
SM 5.2R[d]	0.94	1.71	1.17	50.0	20.7	29.3	0.0
SM 5.0R[d]	0.97	0.70	0.54	75.6	18.3	4.9	1.2

[a] Acceptable predictions (absolute errors $\leq$ 0.75 kcal/mol).
[b] Disputable predictions (0.75 kcal/mol $<$ absolute errors $\leq$ 1.50 kcal/mol).
[c] Unacceptable predictions (absolute errors $>$ 1.50 kcal/mol).
[d] Results cited from ref. [29].

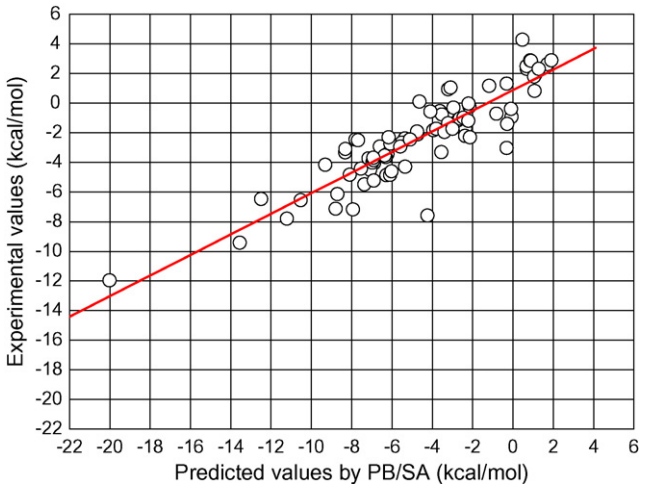

Fig. 4. Correlation between experimentally measured solvation free energies and predicted values by PB/SA for the test set ($N = 82$; $R = 0.891$, $SD = 1.37$ kcal/mol, $MUE = 2.47$ kcal/mol, $Y = 0.69 \times X + 0.87$).

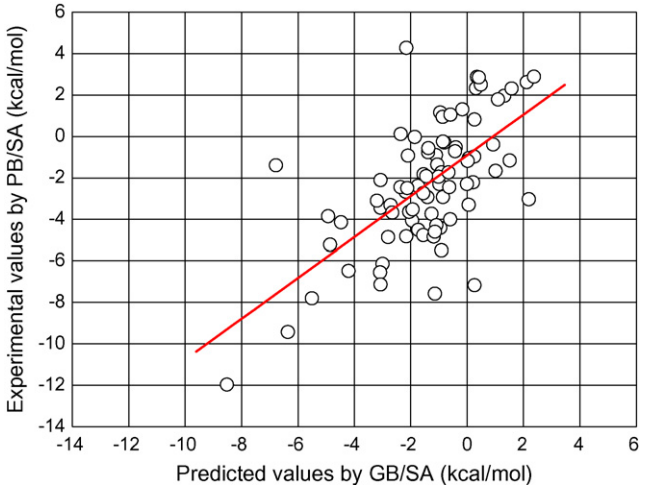

Fig. 5. Correlation between experimentally measured solvation free energies and predicted values by GB/SA for the test set ($N = 82$, $R = 0.641$, $SD = 2.33$ kcal/mol, $MUE = 1.92$ kcal/mol, $Y = 0.98 \times X - 0.91$).

3. Discussion

3.1. Performance evaluation

It is appropriate to compare our model with Wang's model (WSAS) [25], Hou's model (SAWSA) [26,29] and Hawkins' model (SM 5.0R) [27,28] since they are all empirical solvation models that use only solvent-accessible surfaces as descriptors. Table 2 summarizes the statistical results of these models which are cited from original references. One can see that I-SOLV produced better statistical results than the other three models in fitting experimental data. This performance is remarkable considering that our data set is at least 40% larger than the data sets used by the other three models. A larger data set normally represents a greater complexity and is thus more challenging to reconcile for an empirical model. Note that the complexity of our model, i.e. total number of adjustable parameters, is approximately at the same level as WSAS and SAWSA. Therefore, the better performance of our model is not due to more extensive parameterization. SM 5.0R uses a somewhat smaller set of adjustable parameters. However, it is unclear if SM 5.0R would need more parameters to maintain its accuracy if it was calibrated on a larger data set.

When validated on the test set, I-SOLV made excellent predictions with a mean unsigned error of only 0.39 kcal/mol. The predicted values of 89% samples were identified as "acceptable", i.e. absolute error < 0.75 kcal/mol. I-SOLV again outperformed SAWSA 2.0 and SM 5.0R in this test (see Table 3). WSAS has not been tested on this test set, and thus a direct comparison cannot be made. However, judged by the results from whole set fitting (Table 2), it is reasonable to expect that I-SOLV outperforms WSAS as well.

The results discussed above are based on a particular test set, which was cited from Hou et al.'s study [29] for the sake of convenience. Hou et al. did not make it clear how this test set was selected. It seems to us that it was carefully constructed so that a number of representative molecules from each category, such as hydrocarbons, alcohols, ethers, acids, esters, amines and so on, were included. We have noticed that the mean unsigned error produced by I-SOLV on this test set (0.39 kcal/mol) is virtually the same as the one from the leave-one-out cross-validation of I-SOLV on the entire data set (0.40 kcal/mol). We thus expect that our model will demonstrate approximately the same level of accuracy on other subsets selected from the data set used in our study as well.

3.2. On atom tying scheme

We presume that the superior performance of I-SOLV is attributed primarily to our atom typing scheme. The most notable feature of our atom typing scheme is perhaps the concept of united atoms, i.e. hydrogen atoms are considered implicitly on their root atoms. A clear advantage of this method is that it saves atom types for classifying hydrogen atoms. Consequently, saved atom types can be used to characterize other atoms which may be more important for improving accuracy. In contrast, both SAWSA and WSAS have to assign a number of atom types to hydrogen atoms. Also, in SAWSA and WSAS hydrogen atoms are actually classified according to the nature of their root atoms. Classification of a given hydrogen atom is thus linked with the classification of its root atom. This may cause inner-correlations between the variables in regression analysis, which will eventually impair the predictive power of the final model. This potential risk will be removed if hydrogen atoms are treated implicitly like in our model.

Our atom typing scheme is designed in a logical and systematic manner, which can be easily reproduced by other researchers. Unlike SAWSA 2.0, our model does not need "super" atom types for certain chemical groups, such as $–NO_2$, –NO and –CN, or any special correction factors. Our atom typing scheme is also fairly comprehensive for handling common organic molecules. Some chemical structures, such as sulfone and sulfoxide, are missing in our data set and thus are not covered by our current atom typing scheme. However, this

problem can be easily fixed by supplying additional atom types if experimentally measured solvation free energies of such molecules become available.

It needs to be mentioned that our model, in its current form, is applicable to neutral organic molecules only. Experimentally measured solvation free energies of charged molecules are relatively rare. Sometimes the quality of such data is also in question. Both SAWSA and WSAS have attempted to handle charged molecules by introducing a couple of charged atom types. Although these models showed acceptable results on some charged molecules, we believe that it is still premature for an empirical solvation model to do so.

3.3. On solvent-accessible surface

Another notable feature of our model is that it uses the Lee–Richards solvent-accessible surfaces [31] as descriptors. There is some confusion in literature when the term of ''solvent-accessible surface'' is used because some people also refer to the molecular surface as solvent-accessible surface. Both types of solvent-accessible surfaces are generated by rolling a spherical probe on the van der Waals surface. The molecular surface is traced by the inward-facing part of the probe; while the Lee–Richards surface is traced by the center of the probe (Fig. 6) [40]. The molecular surface is also referred to as the Connolly surface because it was Connolly [41] who popularized this concept. The Connolly surfaces are used as descriptors in SAWSA and WSAS.

Both types of solvent-accessible surfaces have been applied in a variety of molecular modeling studies. Our results have demonstrated that the Lee–Richards surface is at least as good as the Connolly surface for an empirical solvation model. From a technical point of view, however, generation of the Lee–Richards surface is more straightforward since it is actually a simple expansion to the van der Waals surface. Anybody with necessary programming skills is able to implement such a routine. Generation of the Connolly surface, however, needs sophisticated analytical algorithms to determine ''contact'' and ''re-entrant'' regions (Fig. 6), which is a subject of study even today [42]. Both WSAS and SAWSA rely on an external program, i.e. MSMS [39], to generate the Connolly surfaces. In contrast, I-SOLV is an integrated program, which generates the solvent-accessible surfaces matching its own requirements. An integrated program is certainly more efficient since it saves the users from the troublesome pipelining of multiple programs.

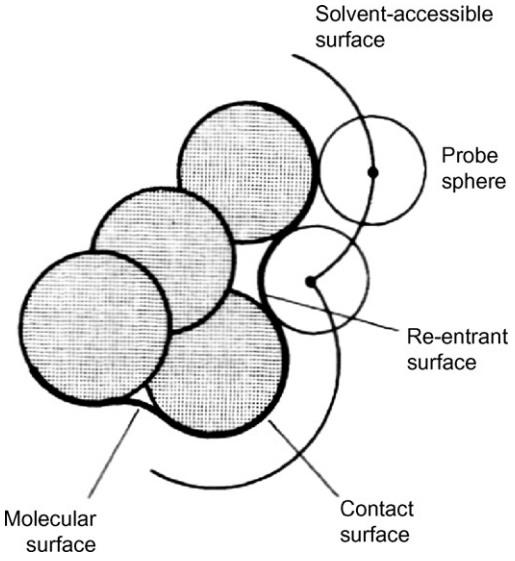

Fig. 6. Illustration of the solvent-accessible surface and the molecular surface (figure cited from ref. [40]).

As described in Section 2, the standard atomic radii proposed by Bondi [32] were adopted by us in the generation of the Lee–Richards solvent-accessible surfaces. Another parameter in such a surface generation is the radius of the solvent probe, which was set to 0.50 Å in our model. We actually tested a series of probe radii, ranging from 0 to 1.50 Å, with our model. The results are summarized in Table 4. As one can see there, a probe radius around 0.50 Å gave the best results. We noticed in our study that when the probe radius was increased to a certain level, e.g. 1.50 Å, some atom types would have no contribution to the final accessible surfaces of the given molecule. One example is the phosphor atoms: they are typically surrounded by other atoms in a given molecule and thus not accessible to a large solvent probe. This may occur to the Lee–Richards surface as well as the Connolly surface. When such a circumstance occurs, some parts of the given molecule actually become invisible to Eq. (1), which consequently leads to a reduced accuracy of our model. Interestingly, both WSAS and SAWSA also use a probe radius around 0.50 Å in their surface generation; while SM 5.0R is directly based on the van der Waals surfaces, i.e. probe radius equals to zero. It seems that the use of a small solvent probe in surface generation is a common feature for surface-based solvation models.

Table 4
Impact of the solvent probe radius on the regression results of I-SOLV

Probe radius (Å)	Correlation coefficient (R)	Standard deviation (kcal/mol)	Mean unsigned error (kcal/mol)	F-value
0.00	0.986	0.52	0.38	353
0.25	0.988	0.49	0.36	389
0.50	0.989	0.47	0.35	423
0.75	0.987	0.50	0.37	380
1.00	0.987	0.51	0.39	363
1.25	0.986	0.52	0.40	347
1.50	0.985	0.55	0.41	314

Table 5
Statistical results of some continuum solvation models cited from literature

Model	Category	Neutral molecules in data set	Correlation coefficient (R)	Standard deviation (kcal/mol)	Mean unsigned error (kcal/mol)	Reference
Gallicchio et al. (SGB/NP)	GB/SA	199	–[a]	–	0.50	15
Bordner et al.	PB/SA	410	–	0.72	–	17
Jorgensen et al.	GB/SA	75	0.964	–	0.61	18
Thompson et al. (SM 5.43R)	GB/SA	257	–	–	0.49[b]	20
Kelly et al. (SM6)	GB/SA	273	–	–	0.47[b]	21
Basilevsky et al.	PCM	278	–	0.60	–	22
Rizzo et al.	PB/SA and GB/SA	460	0.90[b]	–	0.99[b]	23

[a] Not given in the corresponding reference.
[b] Given by the best model among a number of options.

3.4. Comparison to PB/SA and GB/SA models

Hou et al. performed PB/SA and GB/SA computations in their study [29] using DELPHI, AMBER and MSMS. Since these programs are available to us, we have repeated these computations using the same parameters reported by Hou et al. (see Section 2). The purpose is to examine whether the results given by these standard PB/SA and GB/SA procedures are reproducible. Interestingly, we can only observe modest correlations between our results and Hou's results: the correlation coefficients (R) are 0.937 and 0.802 for PB/SA and GB/SA results, respectively. This is probably because the geometries of some molecules in our test set are not identical to the ones used by Hou et al. since we build the molecular models in our test set independently. Our finding prompts that it may not be straightforward to reproduce the results of charge-dependent models like PB or GB because their results are sensitive to the geometry of solute molecules.

As shown in Table 3, correlation between the experimental values and the predicted values by our PB/SA computation of the entire test set produced a correlation coefficient of 0.891 and a standard deviation of 1.37 kcal/mol. The overall correlation is acceptable. However, an obvious problem here is that PB/SA did not predict the absolute values of solvation free energies very well. PB/SA produced a large mean unsigned error of 2.47 kcal/mol, and only about 10% predictions were classified as acceptable. As shown in Fig. 4, solvation free energies of many molecules in this test set were heavily underestimated by PB/SA. It seems that the electrostatic component in PB/SA is unrealistically exaggerated since the absolute errors tend to amplify when solute molecules are more hydrophilic. Re-calibration of the standard PB/SA model may be a solution to this problem. In our recent analysis of a series of FKBP12 inhibitors [43], re-calibration of the PB/SA term in the MM-PB/SA method indeed led to improved predictions of protein–ligand binding affinities.

Correlation between the experimental values and the predicted values by our GB/SA computation of the entire test set produced a correlation coefficient of 0.641 and a standard deviation of 2.33 kcal/mol. This correlation is worse than the one given by PB/SA. However, the regression line in Fig. 5 shows a reasonable slope and intercept. Consequently, GB/SA tended to make better predictions of the absolute values of solvation free energies than PB/SA: about 30% predictions were classified as acceptable.

In summary, judged either by our results or Hou's results (Table 3), I-SOLV performs significantly better than standard PB/SA and GB/SA models on this test set. Performance of the other two surface-based models, i.e. SAWSA 2.0 and SM 5.0R, is also impressive. It needs to be mentioned though that only a few continuum solvation models are included in this comparison, which does not reflect the latest progresses in this area. In Table 5, we have summarized some recently developed continuum solvation models that came to our attention. We have not tested these models in our study because most of them are currently not available to the public. According to the data reported in literature (Table 5), these models are generally able to reduce the average errors in solvation free energy computation below 1.00 kcal/mol, apparently better than conventional PB/SA and GB/SA models. Even so, considering that the average error of I-SOLV in leave-one-out cross-validation is as low as 0.40 kcal/mol, I-SOLV is at least comparable to these continuum solvation models in terms of overall accuracy.

A common feature of these recently developed continuum solvation models is the acceptance of a higher level of parameterization. A continuum solvation models typically needs to combine a point-charge model, a set of atomic radii (or Born radii), and a set of surface tension parameters. The recent study by Rizzo et al. [23] has demonstrated the importance of making an optimized combination of these factors for a PB/SA or GB/SA model. The most significant improvement observed by them was to replace the uniformed surface tension in the non-polar term with a set of 14 atom-based surface tension parameters. Another example is the latest release from the famous series of SMx models, i.e. SM6 [21], which uses a set of 25 surface tension parameters. An earlier study by Gallicchio et al. [15] even used a set of 38 surface tension parameters. The surface tension parameters used in these models were all derived by fitting to certain data sets. These solvation models in fact have considerable empirical elements in themselves.

4. Conclusions

I-SOLV is an empirical solvation model completely based on solvent-accessible surface areas. It produced an excellent

correlation when fitting to a diverse data set of 532 neutral organic molecules. When validated on an independent test set, I-SOLV predicted absolute solvation free energies with a mean unsigned error as low as 0.39 kcal/mol. Our results have demonstrated that surface-based empirical models are not necessarily less accurate than more sophisticated models like PB/SA and GB/SA. This is somewhat contradicting to conventional thoughts in this area. Nevertheless, empirical models are dominant in some other areas, such as the computation of octanol/water partition coefficients (log *P*). We do not see any reason why empirical models cannot be successful in this area if more high-quality data of solvation free energies are available. Currently, empirical models like I-SOLV, SAWSA, WSAS and SM 5.0R can serve as useful alternatives for computing solvation free energies. Because of their simplicity and other technical advantages, they may be even more appealing to high-throughput studies.

Acknowledgements

The authors are grateful to the financial supports from the Chinese National Natural Science Foundation (Grant No. 20502031) and the Chinese Ministry of Science and Technology (the 863 project, Grant No. 2006AA02Z337). The technical aids provided by Chunni Lu and Weiqi Zhang from the computational core facilities of the Shanghai Institute of Organic Chemistry are also appreciated.

Appendix A. Supplementary data

Tables of the entire data set used in this study, summarizing name, molecular formula and the experimentally determined solvation free energy of each molecule can be found in the online version of this article at doi:10.1016/j.jmgm.2007.01.006.

References

[1] W.L. Jorgensen, C. Ravimohan, Monte Carlo simulation of differences in free energies of hydration, J. Chem. Phys. 83 (1985) 3050–3054.
[2] W.L. Jorgensen, Free energy calculations: a breakthrough for modeling organic chemistry in solution, Acc. Chem. Res. 22 (1989) 184–189.
[3] P. Kollman, Free energy calculations: applications to chemical and biochemical phenomena, Chem. Rev. 93 (1993) 2395–2417.
[4] C.J. Cramer, D.G. Truhlar, Implicit solvation models: equilibria, structure, spectra, and dynamics, Chem. Rev. 99 (1999) 2161–2200.
[5] J. Tomasi, B. Mennucci, R. Cammi, Quantum mechanical continuum solvation models, Chem. Rev. 105 (2005) 2999–3093.
[6] G. Lamm, in: K.B. Lipkowitz, R. Larter, T.R. Cundari (Eds.), Reviews in Computational Chemistry, vol. 19, John Wiley & Sons Inc., New Jersey, 2003, pp. 147–365.
[7] N.A. Baker, in: K.B. Lipkowitz, R. Larter, T.R. Cundari (Eds.), Reviews in Computational Chemistry, vol. 21, John Wiley & Sons Inc., New Jersey, 2005, pp. 349–379.
[8] D. Bashford, D.A. Case, Generalized Born models of macromolecular solvation effects, Annu. Rev. Phys. Chem. 51 (2000) 129–152.
[9] W.C. Still, A. Tempczyk, R.C. Hawley, T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics, J. Am. Chem. Soc. 112 (1990) 6127–6129.
[10] D. Sitkoff, K.A. Sharp, B. Honig, Accurate calculation of hydration free-energies using macroscopic solvent models, J. Phys. Chem. 98 (1994) 1978–1988.
[11] D. Qiu, P.S. Shenkin, F.P. Hollinger, W.C. Still, The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii, J. Phys. Chem. A 101 (1997) 3005–3014.
[12] B.N. Dominy, C.L. Brooks III, Development of a Generalized Born model parameterization for proteins and nucleic acids, J. Phys. Chem. B 103 (1999) 3765–3773.
[13] X. Zou, Y. Sun, I.D. Kuntz, Inclusion of solvation in ligand binding free energy calculations using the Generalized Born model, J. Am. Chem. Soc. 121 (1999) 8033–8043.
[14] A. Onufriev, D.A. Case, D. Bashford, Effective Born radii in the Generalized Born approximation: the importance of being perfect, J. Comput. Chem. 23 (2002) 1297–1304.
[15] E. Gallicchio, L.Y. Zhang, R.M. Levy, The SGB/NP hydration free energy model based on the surface Generalized Born solvent reaction field and novel nonpolar hydration free energy estimators, J. Comput. Chem. 23 (2002) 517–529.
[16] E. Gallicchio, R.M. Levy, AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling, J. Comput. Chem. 25 (2004) 479–499.
[17] A.J. Bordner, C.N. Cavasotto, R.A. Abagyan, Accurate transferable model for water, *n*-octanol, and *n*-hexadecane solvation free energies, J. Phys. Chem. B 106 (2002) 11009–11015.
[18] W.L. Jorgensen, J.P. Ulmschneider, J. Tirado-Rives, Free energies of hydration from a Generalized Born model and an all-atom force field, J. Phys. Chem. B 108 (2004) 16264–16270.
[19] M. Feig, A. Onufriev, M.S. Lee, W. Im, D.A. Case, C.L. Brooks III, Performance comparison of Generalized Born and Poisson methods in the calculation of electrostatic solvation energies for protein structures, J. Comput. Chem. 25 (2004) 265–284.
[20] J.D. Thompson, C.J. Cramer, D.G. Truhar, New universal solvation model and comparison of the accuracy of the SM5.42R, SM5.43R, C-PCM, D-PCM, and IEF-PCM continuum solvation models for aqueous and organic solvation free energies and for vapor pressures, J. Phys. Chem. A 108 (2004) 6532–6542.
[21] C.P. Kelly, C.J. Cramer, D.G. Truhlar, SM6: a density functional theory continuum solvation model for calculating aqueous solvation free energies of neutrals, ions, and solute–water clusters, J. Chem. Theory Comput. 1 (2005) 1133–1152.
[22] M.V. Basilevsky, I.V. Leontyev, S.V. Luschekina, O.A. Kondakova, V.B. Sulimov, Computation of hydration free energies of organic solutes with an implicit water model, J. Comput. Chem. 27 (2006) 552–570.
[23] R.C. Rizzo, T. Aynechi, D.A. Case, I.D. Kuntz, Estimation of absolute free energies of hydration using continuum methods: accuracy of partial charge models and optimization of nonpolar contributions, J. Chem. Theory Comput. 2 (2006) 128–139.
[24] D. Eisenberg, A.D. McLachlan, Solvation energy in protein folding and binding, Nature 319 (1986) 199–203.
[25] J. Wang, W. Wang, S. Huo, M. Lee, P.A. Kollman, Solvation model based on weighted solvent accessible surface area, J. Phys. Chem. B 105 (2001) 5055–5067.
[26] T. Hou, X. Qiao, W. Zhang, X. Xu, Empirical aqueous solvation models based on accessible surface areas with implicit electrostatics, J. Phys. Chem. B 106 (2002) 11295–11304.
[27] G.D. Hawkins, D.A. Liotard, C.J. Cramer, D.G. Truhlar, OMNISOL: fast prediction of free energies of solvation and partition coefficients, J. Org. Chem. 63 (1998) 4305–4313.
[28] G.D. Hawkins, C.J. Cramer, D.G. Truhlar, Parameterized model for aqueous free energies of solvation using geometry-dependent atomic surface tensions with implicit electrostatics, J. Phys. Chem. B 101 (1997) 7147–7157.
[29] T. Hou, W. Zhang, Q. Huang, X. Xu, An extended aqueous solvation model based on atom-weighted solvent accessible surface areas: SAWSA v2.0 model, J. Mol. Model. 11 (2005) 26–40.
[30] The SYBYL Software, Version 7.2, Tripos Inc., http://www.tripos.com/.
[31] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. 55 (1971) 379–400.
[32] A. Bondi, van der Waals, Volumes and radii, J. Phys. Chem. 68 (1964) 441–452.

[33] R. Wang, Y. Fu, L. Lai, A new atom-additive method for calculating partition coefficients, J. Chem. Inf. Comput. Sci. 37 (1997) 615–621.
[34] R. Wang, Y. Gao, L. Lai, Calculating partition coefficient by atom-additive method, Persp. Drug Discov. Des. 19 (2000) 47–66.
[35] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, R.E. Stratmann Jr., J.C. Burant, S. Dapprich, J.M. Millam, A.C. Daniesl, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Rotiz, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskora, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, J.L. Andres, C. Gonzalez, M. Head-Gordon, E.S. Replogle, J.A. Pople, Gaussian Inc., Pittsburgh, PA, 1998.
[36] D.A. Case, D.A. Pearlman, J.W. Caldwell, T.E.I. Cheatham, J. Wang, W.S. Ross, C. Simmerling, T. Darden, K.M. Merz, R.V. Stanton, A. Cheng, J.J. Vincent, M. Crowley, V. Tsui, H.R.R. Gohlke, Y. Duan, J. Pitera, I. Massova, G.L. Seibel, U.C. Singh, P. Weiner, P.A. Kollman, AMBER Version 8, University of California, San Francisco, 2002.
[37] W. Rocchia, E. Alexov, B. Honig, Extending the applicability of the nonlinear Poisson–Boltzmann equation: multiple dielectric constants and multivalent ions, J. Phys. Chem. B 105 (2001) 6507–6514.
[38] W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, B. Honig, Rapid grid-based construction of the molecular surface for both molecules and geometric objects: applications to the finite difference Poisson–Boltzmann Method, J. Comput. Chem. 23 (2002) 128–137.
[39] M.F. Sanner, A.J. Olson, J.C. Spehner, Reduced surface: an efficient way to compute molecular surfaces, Biopolymers 38 (1996) 305–320.
[40] A.R. Leach, Molecular Modeling: Principles and Applications, second ed., Pearson Education Ltd., Harlow, England, 2001, pp. 6–8.
[41] M. Connolly, Solvent-accessible surfaces of proteins and nucleic acids, Science 221 (1983) 709–713.
[42] S. Bhat, E.O. Purisima1, Molecular surface generation using a variable-radius solvent probe, Proteins 62 (2006) 244–261.
[43] Y. Xu, R. Wang, A computational analysis of the binding affinities of FKBP12 inhibitors using the MM-PB/SA method, Proteins 64 (2006) 1058–1068.