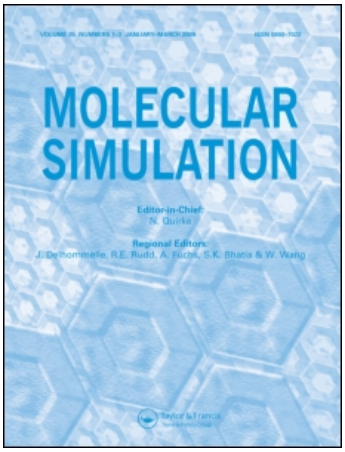

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

A combined steepest descent and genetic algorithm (SD/GA) approach for the optimization of solvation parameters

Ting Liu[a]; Lei Ye[a]; Huajun Chen[a]; Jingyuan Li[b]; Zhaohui Wu[a]; Ruhong Zhou[c]
[a] Department of Computer Science, Zhejiang University, Hangzhou, People's Republic of China [b] Department of Physics, Zhejiang University, Hangzhou, People's Republic of China [c] Department of Chemistry, Columbia University, New York, NY, USA

To cite this Article Liu, Ting , Ye, Lei , Chen, Huajun , Li, Jingyuan , Wu, Zhaohui and Zhou, Ruhong(2006) 'A combined steepest descent and genetic algorithm (SD/GA) approach for the optimization of solvation parameters', Molecular Simulation, 32: 6, 427 — 435
To link to this Article: DOI: 10.1080/08927020600812672
URL: http://dx.doi.org/10.1080/08927020600812672

Molecular Simulation, Vol. 32, No. 6, 15 May 2006, 427–435

Taylor & Francis
Taylor & Francis Group

A combined steepest descent and genetic algorithm (SD/GA) approach for the optimization of solvation parameters

TING LIU†, LEI YE†, HUAJUN CHEN†, JINGYUAN LI‡, ZHAOHUI WU† and RUHONG ZHOU¶§*

†Department of Computer Science, Zhejiang University, Hangzhou, People's Republic of China
‡Department of Physics, Zhejiang University, Hangzhou, People's Republic of China
¶IBM Thomas J. Watson Research Center, Yorktown Heights, New York, NY 10598, USA
§Department of Chemistry, Columbia University, New York, NY 10027, USA

(Received April 2006; in final form May 2006)

An accurate solvation model is essential for computer modeling of protein folding and other biomolecular self-assembly processes. Compared to explicit solvent models, implicit solvent models, such as the Poisson-Boltzmann (PB) with solvent accessible surface area model (PB/SA), offer a much faster speed—the most compelling reason for the popularity of these implicit solvent models. Since these implicit solvent models typically use empirical parameters, such as atomic radii and the surface tensions, an optimal fit of these parameters is crucial for the final accuracy of properties such as solvation free energy and folding free energy. In this paper, we proposed a combined approach, namely SD/GA, which takes the advantage of both local optimization with the steepest descent (SD), and global optimization with the genetic algorithm (GA), for parameters optimization in multi-dimensional space. The SD/GA method is then applied to the optimization of solvation parameters in the non-polar cavity term of the PB/SA model. The results show that the newly optimized parameters from SD/GA not only increase the accuracy in the solvation free energies for ~200 organic molecules, but also significantly improve the free energy landscape of a β-hairpin folding. The current SD/GA method can be readily applied to other multi-dimensional parameter space optimization as well.

Keywords: Parameter optimization; Genetic algorithm; Solvation parameters; Poisson-Boltzmann; Steepest descent

1. Introduction

Hydrophobic interactions play an important role in many areas, such as protein folding [1–5], self-assembly of amphiphiles [6], and capillary evaporation [7,8]. Many biomolecules are characterized by extended surfaces containing nonpolar regions, and the aggregation and subsequent removal of water molecules between these hydrophobic surfaces is believed to be central to their self-organization. A better understanding of solvation is a key step toward the understanding of protein folding and other self-assembly processes. Thus, an accurate solvation model is critical for computer modeling of these important processes [9–12].

Explicit solvent models, such as simple point charge (SPC) model [13], provide an explicit representation of water molecules with exclusion, polar and orientational characters, which offer a detailed description of many important phenomena, such as hydrogen bonds, bridge waters and dewetting (water drying) transitions, etc. However, they are often computationally expensive due to the large number of particles involved as well as the enormous conformational space allowed for these water molecules (which needs to be sampled for convergence). Alternatively, the implicit solvent models, such as the Poisson-Boltzmann (PB) with solvent accessible surface area (PB/SA) model [14] and the generalized born (GB) with solvent accessible surface area (GB/SA) model [15], consider the water solvent as a continuous medium surrounding the solute, and solve the PB equation or its approximations to incorporate the solvent effect implicitly. This drastically reduces the computational costs as compared to the explicit solvent models. The cost saving comes not only from the reduced degree of freedoms, but also from the fact that the expensive sampling (re-equilibration) of water molecules around the moving solute is no longer needed. The continuous medium assumes the water conformations being re-equilibrated instantaneously [14,15].

*Corresponding author. Email: ruhongz@us.ibm.com

Molecular Simulation
ISSN 0892-7022 print/ISSN 1029-0435 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/08927020600812672

However, the implicit solvent models usually involve many empirical parameters to describe the solvent effect. For example, the so-called non-polar cavity term is often calculated based on the solvent accessible surface area. It is simply the sum of each atom's solvent accessible surface area weighted by an adjustable parameter referred to as the surface tension [9,11,12,14,15]. These adjustable parameters are essential to the accuracy of the implicit solvent models. In our previous studies of the folding free energy landscape of a well-studied peptide, the C-terminus β-hairpin of protein G [16–19], we found that some non-native states are heavily overweighted in implicit solvent models (both PB and GB), and the lowest free energy states are not the native state. An overly strong salt-bridge between charged residues was found to be responsible for this behavior, particularly in the GB model [17,19–21]. On the other hand, the explicit solvent model reproduces the experimental results quite well near the biological temperatures for the same β-hairpin. It was found that the balance between the polar electrostatic and non-polar cavity interactions is shifted towards favoring the electrostatic interactions in some residues, which causes overly strong salt-bridges between some charged groups and results in an over-weighting of non-native structures [17,19]. Thus, an optimized set of parameters for the non-polar cavity term is of great current interest.

The present work will focus on the optimization of the surface tension parameters used in the PB/SA model. The PB model is selected over the GB model for its more accurate description of the polar electrostatic solvation free energy. The parameterization is done within the context of the optimized potential for liquid simulation—all-atom (OPLSAA) force field [22]. It should be noted that the current PB/SA model based on Honig *et al.* Delphi implementation [14] adopts all atomic radii from the OPLSAA van der Waals radii, so there are no extra "free" parameters to be optimized. Of course, as pointed out by Nina *et al.* and Swanson *et al.* [23,24], the electrostatic solvation free energies could not be reproduced accurately by PB if radius parameters were not optimized. In this study, however, we will focus on the non-polar cavity term optimization as an initial test for the SD/GA approach. Nonetheless, it is still a multi-dimensional optimization problem due to the many different surface tensions associated with various atom types. Following an earlier work by Levy *et al.* [9] for the surface GB (the SGB model [25]) parameter optimization with the same OPLSAA force field, we used a dataset of 199 small organic molecules for the hydration free energy fitting. The dataset of small molecules is composed of 199 entries comprising alkanes, alkenes, alkynes, aromatics, alcohols, phenols, ethers, esters, acetals, amines, amides, ketones, aldehydes, carboxylic acids, nitriles, nitro compounds, aromatic heterocyclic, thiols and sulfides (for a complete list of the entries, consult Ref. [9]). Similarly, we also assign all atoms into 19 different types (see below for details) according to their similarity in chemical environment [9]. These 19 surface tension parameters are thus the object of optimization in this study. The optimization process can be described as minimizing an object function—defined as the root mean square (RMS) error between the experimental and calculated solvation free energies:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i^N \left(\Delta G^i_{\mathrm{cal}}(\{\sigma_j\}) - \Delta G^i_{\mathrm{expt}}\right)^2}{N}} \qquad (1)$$

where N is the total number ($N = 199$) of small organic molecules [9], and $\{\sigma_j\}$ is the set of surface tension parameters to be optimized. $\Delta G_{\mathrm{cal}}(\{\sigma_j\})$ represents the solvation free energy from the PB/SA calculation, and ΔG_{expt} the solvation free energy from experiment. Even though a single point solvation free energy calculation with the PB/SA model does not cost much CPU time (less than 1 s in a modern PC), the 19 dimensional (19D) optimization space can be a great challenge, since each point in this 19D space represents a PB/SA calculation (since conformations might change slightly with different parameters, a minimization is often needed). Therefore, an efficient optimization method is highly demanded. In general, non-global optimization methods, such as the steepest descent (SD) method, are quick in local searches but often trapped in local minima; whereas the global optimization methods, such as genetic algorithm (GA), are efficient in escaping the local minima but slow in locating the nearby minimum. Thus, in this study, we propose a combined approach for the solvation parameter optimization, namely SD/GA, which takes advantages from both the local and global optimization methods. We believe the current SD/GA approach is more systematic than the previous heuristic approaches for large parameter space optimization [9]. In addition, we have also provided a parallel implementation of the SD/GA algorithm for high performance. It should be noted that the idea of combining the local and global minimization methods is not new and others have proposed similar approaches in other areas such as operational research [26]; however, there are no existing systematic tools, to our knowledge, for the optimization of solvation parameters or other force field parameters, particularly in a parallel fashion for high performance.

The optimized parameters are then applied to the folding free energy landscape of the C-terminus β-hairpin of protein G, which showed a distorted free energy landscape in the previous study with a non-optimized PB/SA model [19]. The newly optimized PB/SA model significantly improves the folding free energy landscape of the β-hairpin as compared to experiment and results from the explicit solvent model.

2. Methods and algorithms

Many optimization algorithms require that the object function be differentiable with the parameter variables. Unfortunately, here we do not have a direct analytical function for the RMS error in equation (1) versus the surface tension parameters due to the complex nature of

the solvent accessible surface area (Connolly surface [27] in this case). Thus, we use a numerical mesh instead, for all the parameters to be optimized. Both SD and GA are then operated on this parameter mesh space, with the solvation free energies and the object function being calculated numerically on the fly. In the following, we will briefly discuss the basic ideas of SD and GA, and then describe the details of our combined SD/GA algorithm, as well as its parallel implementation.

2.1 Steepest descent (SD) method

The basic idea of the SD algorithm can be described as a two-step process: probing and accelerating. The probing step is that for a given starting point in the n-dimensional space, such as $P_0(X_0, X_1, \ldots, X_n)$, SD probes the two opposite directions of each coordinate axes X_i and chooses the direction which decreases the overall object function most—the steepest decent direction. Then, SD tentatively moves to a new point P_1 along that SD direction. The second step is the accelerating step, which means when SD reaches a new point P_1 in the above probing step, it makes a larger step along the direction $\overrightarrow{P_0P_1}$ to reach a newer point P_2. If P_2 offers a lower object function than P_1, SD is said to have succeeded in accelerating and P_2 is then made as the new final point; otherwise SD stays with the previous point P_1. This process iterates until SD locates the minimal object function or converges to a predefined criterion.

2.2 Genetic algorithm (GA)

The GA is an intelligent global searching method which simulates the evolution of biology and is based on Darwin's concept of natural selection. It is robust, flexible, and most of all, simple. It can overcome the local minima trap problem that local optimization methods, such as SD, often encounter in global minimum searching.

In GA, a single solution to a specific problem is represented by a "genome", which can be a sequence of nucleotides (true genome), or a sequence of binary bits, an array of floats or anything else. Two of the basic operators of GA are "crossover" and "mutation". The crossover operator crossovers two selected parent genome based on probability p_c to generate a new genome. The mutation operator mutates each genome based on probability p_m to get a new genome (both p_c and p_m are parameters in GA).

In the current optimization problem, the solution "genome" is simply a 19D solvation parameter vector. The crossover and mutation operators can be easily defined as following. For the crossover operator, the total 19 parameters are divided into three subgroups randomly (two 6D and one 7D), which can be optimized separately for a certain number of steps. Then, every so often (probability p_c), the three partially optimized subgroups are re-combined (crossovered) into one single group with 19D. The mutation operator is to randomly choose a parameter (probability p_m) and set its value to a random number (within a predefined boundary). Both the crossover and mutation operations help the system to escape the local minimum traps.

2.3 Combination of SD and GA

As discussed above, the SD method, though fast in local search, might get trapped in some local minima, while GA can help escape these local traps by applying the crossover and mutation operators from time to time. Thus we incorporate both the crossover and mutation operators of GA into the SD process. The combined SD/GA algorithm is summarized in the following:

(1) When looping over the 19D SD process, randomly divide the 19 parameters into three subgroups, each with about 6D, then minimize these subgroups separately using the SD method. The partially optimized three subgroups are then re-combined into one single group—this is equivalent to "crossover" operator in the GA algorithm. Finally, we run the 19D SD for some number of steps.
(2) Check if the results satisfy the predefined criterion. If not, then randomly pick some parameters and set random values for them (within their boundaries). This is equivalent to the "mutation" operator in the GA algorithm. Run the 19D SD for some number of steps.
(3) Check if the results satisfy the predefined criterion. If still not, run a certain number of extra steps of the 19D SD to see if it satisfies the overall criterion. Repeat step (1)–(3) until either (a) the predefined accuracy criterion is satisfied, or (b) all the variables hit their predefined boundaries (which means boundaries need to be adjusted), or (c) no meaningful changes in last two cycles (satisfying the convergence criterion).

A pseudocode is also provided for a better understanding of the flow of the above SD/GA algorithm.

```
Step One:
    Divide the 19 parameters into three subgroups randomly;
    Run SD process on the first subgroup;
    Run SD process on the second subgroup;
    Run SD process on the third subgroup;
    Combine subgroup results to form a 19-dimension solution;
    Run some number of steps of SD on 19 parameters;
Step Two:
    If (δaccuracy >= THERESHOLD){
        Randomly select some parameters;
        Set random values for these parameters;
        Run some number of steps of SD on 19 parameters;}
    Else
        Exit;
```

Step Three:
If ($\delta accuracy >= THERESHOLD$)
Run some number of steps of SD on 19 parameters;
Else
Exit;
Step Four:
If ($\delta accuracy >= THERESHOLD \&\& !No\ Changes \&\& !HitBoundary$)
Go to Step One;
Else
Exit;

2.4 Parallel implementation

As mentioned above, even though each solvation free energy calculation with the PB/SA model does not cost that much CPU time (less than 1 s on a modern PC), the high dimensionality of the parameter space quickly surmounts to an enormous CPU request. For example, in a 19D space, suppose each parameter has a mesh size of 10 (in practice, we have used up to 100 for some parameters, see below), then there are total 10^{19} possible points in the 19-dimensional space. Since most of the CPU time is spent on the PB/SA calculations in the SD process, we can parallel the SD process to achieve a higher performance. Fortunately, each PB/SA calculation is independent given the 19D parameter set, thus this is a so-called embarrassingly parallel problem. We can perform all necessary PB/SA calculations at a certain point in parallel, and then collect RMS errors to determine which direction is the SD direction. In general, for n parameters to be optimized in one SD process, we launch 2^n independent jobs to 2^n processors. In the current implementation, we used 64–128 processors ($n = 6, 7$) in parallel.

3. Results and discussion

The combined SD/GA algorithm is used to optimize the solvation parameters for the implicit solvent PB/SA model in the context of the OPLSAA force field [22]. The atomic radii used to define the dielectric boundary surface in the PB model are taken directly from the van der Walls radii in the OPLSAA force field, and the fixed charge distribution of the solute is constructed from the OPLSAA partial atomic charges. The non-polar cavity term is measured by the usual solvent accessible surface areas weighted by the surface tensions as mentioned above. A dielectric constant of 1.0 is used for the small organic molecules, while the aqueous solvent region is assigned a dielectric constant of 78.5. Following a similar previous work on the optimization of SGB parameters, [9] a total of 19 different solvation atom types are assigned according to their similarity in chemical environment. These atom types as well as a brief description for each type are summarized in table 1.

Table 1. Atom types assigned for the PB/SA model in the context of OPLSAA force field following the pioneer work of Levy and coworkers.

Type	*OPLS symbol(s)*	*Description*
1	C	sp2 Carbon in carbonyl and carboxylic group in ketones, aldehydes, amides and carboxylic acids
2	CM,C=	sp2 Carbon in alkenes and dienes
3	CA	sp2 Carbon in aromatic rings
4	CT, CY, CO	sp3 Carbon
5	CZ	sp Carbon in nitriles and alkynes[†]
6	H, HO, HS	Hydrogen on heteroatoms
7	HA	Hydrogen on aromatic rings
8	HC	Hydrogen on sp3 carbons and on aldehyde carbonyl group
9	N	sp2 Nitrogen in amides
10	NC	sp2 Nitrogen in aromatic heterocyclic compounds
11	NZ	sp Nitrogen in nitriles
12	NO, ON	Nitrogen or oxygen in nitro($-NO_2$)group[‡]
13	NT	sp3 Nitrogen in ammonia and amines
14	O	Carbonyl and carboxyl oxygen in ketones, aldehydes, amides, and carboxylic acids
15	OH,OH2	Oxygen on carbon carrying one hydrogen as in alcohols, diols, phenols, and carbonyl and carboxylic acids
16	OS	Oxygen in ethers and noncarbonyl oxygen in esters
17	S,SH	Sulfur in thiols and sulfides
18	O2	Oxygen in carboxylate group
19	N3,N3A,NZ1	Nitrogen in ammonium group

[†] An all-atom OPLS type for sp aliphatic carbons is not available.
[‡] As defined type 12 corresponds to the entire $-NO_2$ group.

In order to test the current SD/GA approach, we start with a relatively small parameter space by setting a narrow range for each parameter. In principle, the SD/GA should converge to the same final position no matter where it starts. If SD/GA can locate the global minimum successfully in a smaller space repeatedly, then we will enlarge the parameter space to perform the full-scale optimization. Fortunately, these solvation parameters, i.e. surface tensions, under optimization have physical meanings, so they generally possess values within some reasonable boundaries, which greatly reduce the optimization search space. For example, many molecular dynamics programs set the default values for these surface tensions to 5 cal/mol/Å^2, based on experimental observations of the solvation free energies of simple alkane solutes [11,14,25]. We use this default value of 5 cal/mol/Å^2 as our starting point for all parameters in the final full scale optimization (see below). As a first trial, however, four different optimization runs are performed starting from four different initial points, all within a small

Table 2. The convergence of the SD/GA optimization starting from four different initial positions within a small range of parameter space.

OPLSAA types	*Range*	*Four start points*				*Four final points*			
C	(1 9)	4	7	1	3	9	9	9	9
CM, C=	(1 9)	6	8	3	7	9	9	9	9
CA	(4 10)	6	6	8	5	4	4	4	4
CT, CY, CO	(1 9)	8	3	5	5	5.5	5.5	5.5	5.5
CZ	(0 10)	5	8	2	7	0	1	0	1
H, HO, HS	(2 8)	5	4	7	7	2	2	2	2
HA	(2 8)	4	7	4	6	2	2	2	2
HC	(8 14)	12	10	9	10	8	8	8	8
N	(0 10)	3	7	5	4	0	0	0	0
NC	(3 11)	8	5	10	4	3	3	3	3
NZ	(−5 5)	4	3	4	−4	5	5	5	5
NO, ON	(−5 5)	0	−3	−1	3	1.5	1.5	2	2
NT	(4 10)	8	5	3	8	4	4	4	4
O	(0 10)	5	7	2	7	0	0	0	0
OH, OH2	(3 11)	7	5	10	5	3	3	3	3
OS	(3 11)	7	6	5	8	3	3	3	3
S, SH	(0 10)	4	2	9	3	8	8	8	8
O2	(0 10)	2	6	8	6	10	10	10	10
N3, N3A, NZ1	(0 10)	7	5	4	9	10	10	10	10

range with sizes (max−min) from 6 to 10 cal/mol/Å^2. Table 2 lists these small ranges for all parameters (first column). The four different starting points (19D) are also listed in table 2, second column. The SD step size in each dimension is set to be 0.5 cal/mol/Å^2. The final results are summarized in the third column of table 2. The four solutions converge quickly into the same final point in space, with only small differences in the 5th (type CZ) and 12th (type NO and ON) dimensions. Even with the small differences in these 2Ds, the four solutions still show the same general direction of the "descending". Some of the parameters hit the boundaries, since as we stated earlier, the optimization process can not go beyond the parameter boundaries.

Then, as a second trial, the solution space is enlarged by setting larger ranges with sizes (max−min) from 10 to 14 cal/mol/Å^2 for various parameters. Again, four different starting points are randomly selected. The various ranges, starting points and final results are summarized in table 3. The four final points in this case differ slightly more than those in the previous trial, but they still display a very good convergence. From both tables 2 and 3, and many other experiments we did (data not shown), we found that no matter where we start, the final solutions always converge to roughly the same point, even with a much larger parameter space. Thus, we conclude that the current SD/GA approach does offer a systematic way to optimize the multi-dimensional solvation parameters.

Finally, a much longer full scale optimization is run with the parallel SD/GA implementation on 64–128 nodes Linux cluster, starting from a single initial point with all parameters setting at default values of 5 cal/mol/Å^2. During this long run, the SD/GA method automatically enlarges the parameter ranges once more than 1/4 of parameters hit their boundaries. The process

Table 3. The convergence of the SD/GA optimization starting from four different initial positions within a large range of parameter space.

OPLSAA types	*Range*	*Four start points*				*Four final points*			
C	(−1 12)	5	8	3	0	11	11	11	11
CM, C=	(−1 11)	10	0	2	7	11	11	11	11
CA	(2 12)	3	6	7	5	2	5	2	2
CT, CY, CO	(−1 11)	6	7	9	3	5	5	6	5
CZ	(−2 12)	0	4	10	7	−2	−2	−2	−2
H, HO, HS	(0 10)	7	8	3	4	1	4	1	4
HA	(0 10)	4	9	1	6	4.5	2	5	5
HC	(6 16)	7	8	14	10	6	6	6	6
N	(−2 12)	7	10	4	0	−2	−2	−2	−1
NC	(1 13)	10	6	7	4	1	1	1	1
NZ	(−7 7)	5	3	0	−4	7	7	7	7
NO, ON	(−7 7)	3	1	−5	−2	−6	−5	−7	−6
NT	(2 12)	6	7	3	4	2	2	2	2
O	(−2 12)	4	−1	7	0	−2	−2	−4	−2
OH, OH2	(1 13)	9	5	2	4	9	7	10	8
OS	(1 13)	3	10	4	9	1	1	1	1
S, SH	(−2 12)	6	5	3	−1	12	10	12	10
O2	(−2 12)	5	1	3	8	12	12	12	12
N3, N3A, NZ1	(−2 12)	1	−1	10	3	12	12	12	12

Table 4. The optimization path (only partial path shown) of the SD/GA method with an auto adjustment of the parameter ranges.

5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
5	6	4	5	5	5.5	5	3	6	2	5	6	5	7	5	4	5	7	7
5	7	4	5	5	7	5	3	8	2	5	7	5	7	5	5	5	7	7
7	7	4	5	2	7	5	3	8	2	7	7	2	7	8	2	8	7.5	7
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
11	10	6	5	1	9	7	5	12	1	7	7	2	12	13	1.5	11	10	10
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
14	13	4	5.5	−2	6	5	3	14	1	7	9	2	14	15	−1	13	14	14
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
15	13	2	6	−2	7	7	3	−5	−4.5	15	−6	0	−3	24	−3	14	18	20
15	13	2	6.5	−2	7	7	5.5	−8	−5	19	−7.5	0	−6.0	25	−3	14	20	20

keeps running until the final accuracy/convergence criterion is met, with either an RMS error (accuracy) of less than 0.5 kcal/mol or a solution convergence of less than 0.001 kcal/mol. Table 4 displays the final results, as well as the starting and a few middle points (parameter trajectory) during the optimization process.

During this long optimization process, the RMS error from the experimental solvation free energies decrease from 1.06 kcal/mol (when all use default values) to 0.82 kcal/mol. The final results as well as the corresponding atom types and starting values are summarized in table 5. Figure 1 also shows the comparison of the solvation free energies before and after the optimization for the ~200 small organic molecules. Before the optimization, the calculated solvation free energies show large errors for those very polar molecules with solvation free energies near −10 cal/mol. After the optimization, the overall agreement with experiment is much better, even though there can be still some significant deviations for these very polar molecules. A closer look reveals that some of these molecules, such as ethandiol (expt. −9.6, calc. −12.82 kcal/mol) and hydroxybenzenitrile (expt. −10.17, calc. −11.83 kcal/mol), have two or more polar functional groups in the molecule. This larger error in molecules with multi-functional groups is related to the fact that a fixed partial charge is used for atoms in a functional group, no matter whether another such group is presented nearby or not in the standard OPLSAA force field. A more rigorous fix for this problem might be to use a polarizable force field so that the atomic partial charges can adjust with the changing environment accordingly. These deviations also indicate that a simple fix for the cavity term might not be enough for the high accuracy of implicit solvent models, and more complicated cavity terms [9] and a refit of the polar term (with polarizable electrostatic interactions) might be needed. Nevertheless, the current simple model with the SD/GA optimization does improve the accuracy significantly. Also, the combined SD/GA method offers a systematic way for parameter optimizations, which improves the PB/SA model without much of human intervention as in previous approaches.

Table 5. Optimized PB/SA solvation parameters for the 19 atom types assigned within the context of the OPLSAA force field.

Type	*OPLS symbol(s)*	*Starting parameters*	*Optimized*
1	C	5.0	15.0
2	CM,C=	5.0	13.0
3	CA	5.0	2.0
4	CT, CY, CO	5.0	6.5
5	CZ	5.0	−2.0
6	H, HO, HS	5.0	7.0
7	HA	5.0	5.5
8	HC	5.0	5.5
9	N	5.0	−8.0
10	NC	5.0	−5.0
11	NZ	5.0	19.0
12	NO, ON	5.0	−7.5
13	NT	5.0	0.0
14	O	5.0	−6.0
15	OH, OH2	5.0	25.0
16	OS	5.0	−3.0
17	S, SH	5.0	14.0
18	O2	5.0	20.0
19	N3, N3A, NZ1	5.0	20.0

The newly optimized solvation parameters are then applied to the folding free energy landscape of a well-studied protein peptide—the C-terminus β-hairpin of protein G. This β-hairpin (Ace-GEWTYDDATKTF-TVTE-Nme, where Ace and Nme are normal capping groups) has become a system of choice for β-sheets studies due to its stability in isolated solution and its fast folding speed (folds in 6 μs), [28–30] and it is also an ideal small protein system for new parameter tests since it is generally more challenging to model β-sheets than other small peptides such as α-helices [31]. In our previous study [19], we found that the β-hairpin's folding free energy landscape from the PB/SA model is less accurate than that of an explicit solvent model (SPC model), with the same protein OPLSAA force field, even though it is significantly better than that of the GB model [17]. Thus, it is of current interest to see if the newly optimized parameters from the above SD/GA approach will improve the folding free energy landscape of this β-hairpin.

As mentioned earlier, in the previous free energy landscapes from replica exchange molecular dynamics (REMD) method [16,17,32,33] for both the PB [19] and GB models [17], some non-native states (H-state, hydrophobic core formed but no native beta-strand H-bonds) are heavily overweighted and the lowest free energy states are not the native state [17–19]. Some overly strong salt-bridges between charged residues (ASP46, ASP47, GLU56 and LYS50) were found to be responsible for this behavior in these implicit solvent models,

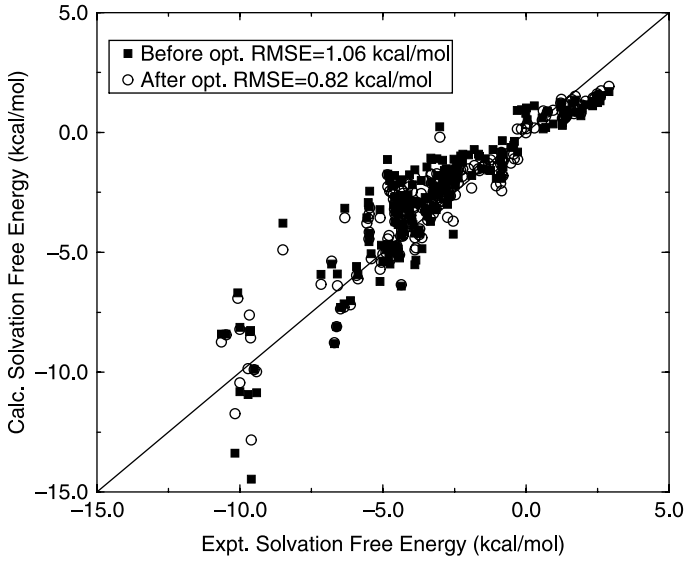

Figure 1. Comparison of solvation free energies of 199 small organic molecules before and after the cavity term optimization with SD/GA. The RMS error before the optimization is 1.06 kcal/mol, while it is 0.82 kcal/mol after the optimization.

particularly in the GB model [17–19]. The balance between polar electrostatic and non-polar hydrophobic (cavity) interactions seems to be shifted towards favoring electrostatic over hydrophobic interactions in some residues, which causes overly strong electrostatic interactions between some charged groups and results in an over-weighting of non-native structures. In the present work, we combined the newly optimized PB/SA model (Delphi implementation [14] from Honig's group) with the REMD method (the use of Delphi in molecular dynamics has been described previously [19]) to calculate the free energy landscape of the β-hairpin. A total of 18 replicas are simulated with the same temperature range (from 270 to 690 K) as in the previous studies [17,19] and an acceptance ratio of ~30% between neighboring replicas. Each replica is run for 3.0 ns for data collection as before. The replica exchanges are attempted every 2.0 ps, and protein configurations are saved every 80 fs, giving a total of 0.675 million configurations. Here, a dielectric constant of 2.0 is used for the protein region, while the dielectric constant of the aqueous solvent region is still 78.5. It should be noted that assigning a dielectric constant for proteins can be very tricky, and values from 1.0 to 4.0 have been suggested before [14,34]. Some others have also suggested that the choice of a single dielectric constant, assuming that the entire protein interior can be treated as a single dielectric continuum, may be problematic altogether [35,36]. Anyway, in the previous study, [19] we had tried dielectric constants of 1.0, 2.0 and 4.0 for the β-hairpin in the PB/SA model, and the free energy landscapes do not change much and the general conclusion stays roughly the same.

The free energy landscape is then obtained through a histogramming analysis on the conformation probability distributions along some reaction coordinates at 310 K from the REMD simulations [17]. Figure 2 shows the comparison of the free energy contour maps from the explicit solvent model (figure 2a), the PB/SA model with old parameters (figure 2b), and the PB/SA model with new parameters (figure 2c). The free energy landscape is plotted against the two reaction coordinates used previously [17,18], i.e. the number of beta-strand hydrogen bonds (N_{HB}^{β}) and the radius of gyration of the hydrophobic core ($R_{\mathrm{g}}^{\mathrm{core}}$). N_{HB}^{β} is defined as the number of native beta-strand backbone-backbone hydrogen bonds excluding the two at the turn of the hairpin (total 5) [17]. $R_{\mathrm{g}}^{\mathrm{core}}$ is the radius of gyration of the side chain atoms on the four hydrophobic residues, W43, Y45, F52 and V54. The free energy contour map from the PB/SA model with newly optimized parameters is found to be significantly better than the PB/SA model with old parameters, when compared to the explicit solvent model and experiment. The improvements include: (1) most importantly, the lowest free energy state is now the folded state (F-state) in the new PB/SA model, whereas the lowest free energy state was the so-called H-state (hydrophobic core formed but no beta-strand hydrogen bonds) in the old PB/SA model. In the new PB/SA model, the F-state is approximately 0.43 kT lower than the H-state, while it was 0.70 kT higher than the H-state in the old PB/SA model (as for comparison, the F-state is about 2.3 kT lower than the H-state in the explicit solvent model). (2) The F-state in the new PB/SA model has a larger number of native beta-strand hydrogen bonds ($N_{HB} \approx 3.4$) than the old PB/SA model ($N_{\mathrm{HB}} \approx 2.2$), which indicates that the F-state in the new PB/SA model resembles better the truly native state (again, as for comparison, $N_{\mathrm{HB}} \approx 4.5$ in the explicit solvent model). The representative structures from the lowest free energy state also confirms that these structures indeed mimic the native structure, with the overly strong salt-bridges largely fixed. In summary, the new PB/SA model, though not perfect, has been significantly improved with these newly optimized

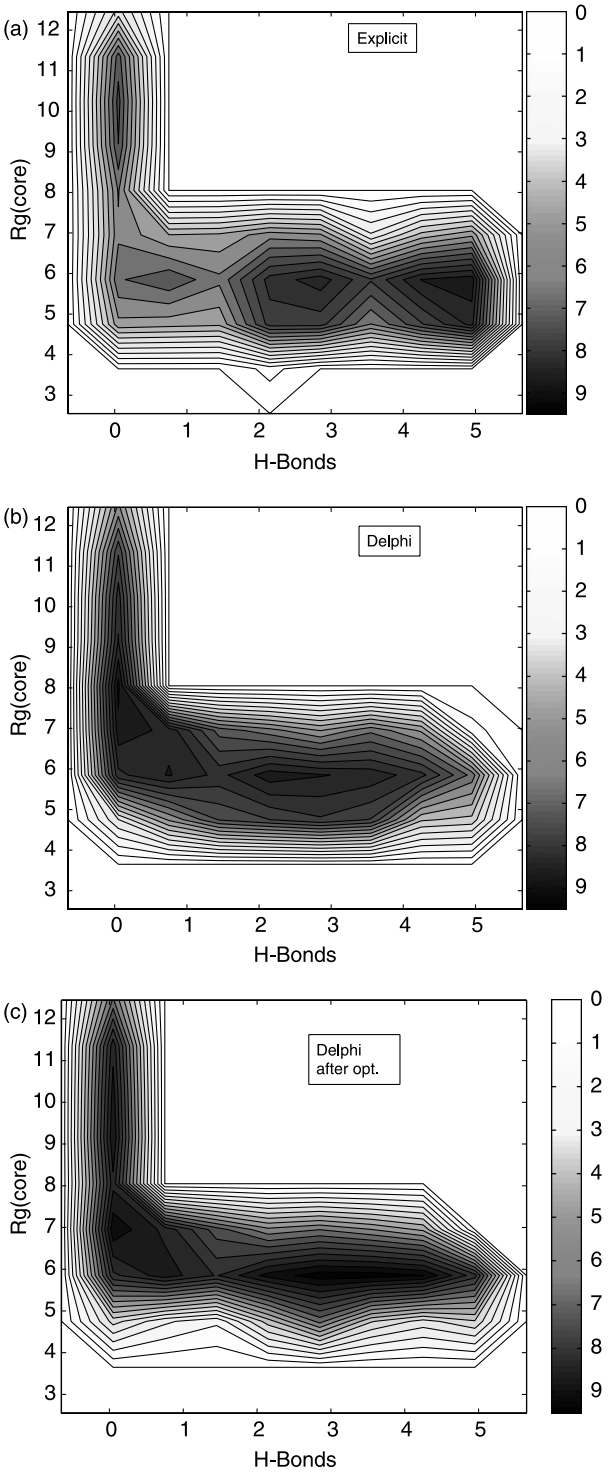

Figure 2. Comparison of the free energy versus the number of β-sheet H-bonds N_{HB}^{β} and the hydrophobic core radius gyration $R_{\mathrm{g}}^{\mathrm{core}}$ at 310 K: (a) from the explicit solvent model, (b) from the implicit solvent PB/SA model with old cavity parameters, and (c) from the PB/SA model with new optimized cavity parameters. A hydrogen bond is counted if the distance between two heavy atoms (N and O in this case) is less than 3.5 Å and the angle N–H. . .O is larger than 120.0°. The free energy is in units of kT, and contours are spaced at intervals of 0.5 kT.

parameters, which correctly predicts the F-state to be lowest free energy state and the representative structures in the F-state to resemble the true native structure.

4. Conclusion

In the present paper, we introduced a combined SD and genetic algorithm (SD/GA) method for the global minimum search, which utilizes the advantages of both local and global optimization methods. The SD/GA method is then used to optimize the solvation parameters for the implicit solvent Poisson-Boltzmann (PB/SA) model using ∼200 small organic molecules. The results from various convergence tests indicate that the SD/GA method can effectively escape the local minima traps and locate the global minimum in a high dimensional parameter space. The newly optimized solvation parameters are then applied to the folding free energy landscape of a β-hairpin. Compared to the previous results from the same PB/SA model with old parameters, the new PB/SA model significantly improves the folding free energy landscape of the β-hairpin. It correctly predicts the F-state to be the lowest free energy state, whereas the old model predicted the so-called H-state (hydrophobic core formed but no native beta-strand hydrogen bonds) to be the lowest free energy state. In addition, the previous erroneous salt-bridges between charged residues are largely fixed in the new PB/SA model.

The future work will be focused on further improvements in the cavity term by exploring other more complicated models such as the one proposed by Levy *et al.* [9] (rather than the simple solvent accessible surface area model), as well as new improvements in the polar electrostatic term by optimizing the atomic radii or incorporating some polarizable measures.

Acknowledgements

We would like to thank Barry Honig for sending us the Delphi code for the Poisson-Boltzmann continuum solvent model in a previous study, and Ray Luo, Emilio Gallicchio and Ron Levy for many helpful discussions.

References

[1] A.R. Fersht. *Structure and Mechanism in Protein Science*, W.H. Freeman and Company, New York (1999).
[2] C.L. Brooks, J.N. Onuchic, D.J. Wales. Taking a walk on a landscape. *Science*, **293**, 612 (2001).
[3] C.M. Dobson, A. Sali, M. Karplus. Protein folding: a perspective from theory and experiment. *Angew Chem. Int. Edit. Engl.*, **37**, 868 (1998).
[4] R. Zhou, X. Huang, C.J. Margulius, B.J. Berne. Hydrophobic collapse in multidomain protein folding. *Science*, **305**, 1605 (2004).
[5] P. Liu, X. Huang, R. Zhou, B.J. Berne. Observation of a dewetting transition in the melittin tetramer collapse. *Nature*, **437**, 159 (2005).
[6] G. Hummer, S. Garde, A.E. Garcia, L.R. Pratt. New perspectives on hydrophobic effects. *Chem. Phys.*, **258**, 349 (2000).
[7] A. Luzar, K. Leung. Dynamics of capillary evaporation. i. Effect of morphology of hydrophobic surfaces. *J. Chem. Phys.*, **113**, 5836 (2000).
[8] K. Leung, A. Luzar, D. Bratko. Dynamics of capillary drying in water. *Phys. Rev. Lett.*, **90**(1–4), 65502 (2003).

[9] E. Gallicchio, L.Y. Zhang, R.M. Levy. The sgb/np hydration free energy model baed on the surface generalized born solvent reaction field and novel non-polar hydration free energy estimators. *J. Comput. Chem.*, **23**, 517 (2002).
[10] E. Gallicchio, L.Y. Zhang, R.M. Levy. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the agbnp implicit solvent model. *Proteins*, **56**, 310 (2004).
[11] D. Bashford, D. Case. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, **51**, 129 (2000).
[12] A. Onufriev, D. Bashford, D. Case. A modification of the generalized born model suitable for macromolecules. *J. Phys. Chem. B*, **104**, 3712 (2000).
[13] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, J. Hermans. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*, B. Pullman (Ed.), pp. 331–342, Dordrecht, Reidel (1981).
[14] B. Honig, A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, **268**, 1144 (1995).
[15] W.C. Still, A. Tempczyk, R.C. Hawley, T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, **112**, 6127 (1990).
[16] R. Zhou, B.J. Berne, R. Germain. The free energy landscape for β-hairpin folding in explicit water. *Proc. Natl. Acad. Sci. USA*, **98**, 14931 (2001).
[17] R. Zhou, B.J. Berne. Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water? *Proc. Natl. Acad. Sci. USA*, **99**, 12777 (2002).
[18] R. Zhou. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins*, **53**, 148 (2003).
[19] R. Zhou, G. Krilov, B.J. Berne. Comment on "can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water?". *J. Phys. Chem. B.*, **108**, 7528 (2004).
[20] T.Z. Lwin, R. Zhou, R. Luo. Is poisson-boltzmann theory insufficient for protein folding simulations? *J. Chem. Phys.*, **124**, 34902 (2006).
[21] R. Geney, M. Layten, R. Gomperts, V. Hornak, C. Simmerling. Investigation of salt bridge stability in a generalized born solvent model. *J. Chem. Theory Comput.*, **2**, 115 (2006).
[22] W.L. Jorgensen, D. Maxwell, J. Tirado-Rives. Development and testing of the opls all-atom froce field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, **118**, 11225 (1996).
[23] M. Nina, D. Belglov, B. Boux. Atomic radii for continuum electrostatics calculations based on molecular dynamics free energy simulations. *J. Phys. Chem. B*, **101**, 5239 (1997).
[24] J.M. Swanson, S.A. Adcock, J.A. McCammon. Optimized radii for Poisson-Boltzmann calculations with the amber force field. *J. Chem. Theory Comput.*, **1**, 484 (2005).
[25] A. Ghosh, C.S. Rapp, R.A. Friesner. Generalized born model based on a surface integral formulation. *J. Phys. Chem. B*, **102**, 10983 (1998).
[26] H.A. Thi, D.T. Pham, N.V. Thoai. Combination between global and local methods for solving an optimization problem over the efficient set. *Eur. J. Oper. Res.*, **142**, 258 (2002).
[27] M.L. Connolly. Analytical molecular surface calculation. *J. Appl. Cryst.*, **16**, 548 (1983).
[28] V. Munoz, P.A. Thompson, J. Hofrichter, W.A. Eaton. Folding dynamics and mechanism of β-hairpin formation. *Nature*, **390**, 196 (1997).
[29] V. Munoz, E.R. Henry, J. Hofrichter, W.A. Eaton. A statistical mechanical model for β-hairpin kinetics. *Proc. Natl. Acad. Sci. USA*, **95**, 5872 (1998).
[30] F.J. Blanco, L. Serrano. Folding of protein g b1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *Eur. J. Biochem.*, **230**, 634 (1995).
[31] B. Zagrovic, E.J. Sorin, V.S. Pande. β-Hairpin folding simulation in atomistic detail. *J. Mol. Biol.*, **313**, 151 (2001).
[32] Y. Sugita, Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, **314**, 141 (1999).
[33] R. Zhou. Trp-cage: folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci. USA*, **100**, 13280 (2003).
[34] T. Simonson. Macromolecular electrostatics: continuum models and their growing pains. *Curr. Opin. Struct. Biol.*, **11**, 243 (2001).
[35] C.N. Schutz, A. Warshel. What are the dielectric 'constants' of proteins and how to validate electrostatic models. *Proteins*, **44**, 400 (2001).
[36] E. Garcia-Moreno, J. Dwyer, A. Gittis, E. Lattman, D. Spenser, W. Stites. Experimental measurement of the effective dielectric in the hydrophobic core of a protein. *Biophys. Chem.*, **64**, 211 (1997).