DOI 10.5012/bkcs.2010.31.5.1247

Computational Prediction of Solvation Free Energies of Amino Acids with Genetic Algorithm

Jung-Hum Park, Jin-Won Lee,[†,*] and Hwangseo Park[*]

Department of Bioscience and Biotechnology, Sejong University, Seoul 143-747, Korea. [*]E-mail: hspark@sejong.ac.kr
[†]Department of Life Science, Hanyang University, Seoul 133-791, Korea. [*]E-mail: jwl@hanyang.ac.kr
Received February 17, 2010, Accepted March 5, 2010

We propose an improved solvent contact model to estimate the solvation free energies of amino acids from individual atomic contributions. The modification of the solvation model involves the optimization of three kinds of parameters in the solvation free energy function: atomic fragmental volume, maximum atomic occupancy, and atomic solvation parameters. All of these atomic parameters for 17 atom types are developed by the operation of a standard genetic algorithm in such a way to minimize the difference between experimental and calculated solvation free energies. The present solvation model is able to predict the experimental solvation free energies of amino acids with the squared correlation coefficients of 0.94 and 0.93 for the parameterization with Gaussian and screened Coulomb potential as the envelope functions, respectively. This result indicates that the improved solvent contact model with the newly developed atomic parameters would be a useful tool for the estimation of the molecular solvation free energy of a protein in aqueous solution.

Key Words: Solvation, Amino acids, Genetic algorithm, Atomic parameters, Envelope function

Introduction

Water plays an essential role in the stabilization of proteins and in optimizing their functionalities because about 30% of the residues including Asp, Glu, His, Lys, Tyr, and Arg are ionizable.[1] Ionization equilibria are indeed an important determinant for protein structure and function because they determine the charges of the ionizable groups and, consequently, the long-range electrostatic interactions that characterize intra- and intermolecular interactions associated with protein-solvent and protein-protein interactions. A realistic estimation of the stability and function of a protein in aqueous solution should therefore consider its solvation free energy as the interaction with the surrounding solvent. Despite the necessity for a deep understanding of solvation effects in biosystems, current experimental measurement techniques, such as osmotic stress[2] and far-infrared laser vibration-rotation tunneling spectroscopy,[3] have provided limited information only. Complementary to the experimental methods, computer modeling has drawn a particular interest as a tool for coping with the problem of protein solvation because it can describe the solvation effects directly from a molecular perspective.[4-6]

However, solvation free energy has been considered as one of the most calculation-difficult energy terms due to the complexity of solvent-solute interactions.[7] Although the explicit solvent models should be most accurate in calculating protein solvation energies, a high computational cost has made it difficult for them to be used in practical applications. Therefore, various implicit solvation models with a high efficiency have been developed as alternatives including solvent-accessible surface area model,[8-11] the appropriately defined first solvation shell model,[12] and the group contact model.[13] Poisson-Boltzmann (PB) equation approach and its analytical versions have also been successful in modeling the solvation effects in a realistic way. This method was developed to calculate the hydration free energy of spherical ions,[14] extended to treat arbitrary charge distributions in a spherical cavity,[15] and further improved to be solved analytically for simple boundary shapes[16] or numerically with finite-difference algorithms to treat the solute molecules with arbitrary shape.[17] However, a high computational cost for the finite-difference algorithm has limited the usefulness of the PB model. To reduce the computational time, the methods of potential of mean force have also been developed to estimate the solvation effects at the expense some accuracy.[18,19]

In the early 1990s, Stouten *et al*. suggested a solvation model for a protein molecule by extending the solvent contact model proposed by Colonna-Cesari and Sander.[20,21] The three key parameters in this model were the maximum atomic occupancy, the atomic fragmental volume, and the atomic solvation parameter representing the solvation free energy per unit of volume.[21] Under the assumption that the solvation free energy of an amino acid residue would be given by the sum over atomic contributions, they obtained the atomic parameters for six atom types (C, N, O, N^+, O, and S) using the standard linear least-squares procedures with the experimental solvation free energies of amino acids. This simple solvation model proved to be very successful in estimating the structural properties of a protein as well as in saving computation time in molecular dynamics simulations when compared to the explicit solvent model.[21] Due to such a small number of atom types, however, some proper modifications need to be made in order for the extended solvent contact model to be also useful in predicting solvation free energies of proteins and organic molecules.[22,23]

In the present study, we further improve the Stouten *et al*.'s solvent contact model by extending the parameter space to cope with as many atom types as commonly encountered in amino acids. All of the atomic parameters in the solvation free energy

function are optimized by the operation of a standard genetic algorithm (GA) using the experimental solvation free energy data. It will be shown that the improved solvent contact model with the newly developed atomic parameters can be an appropriate tool for predicting solvation free energies of amino acids in aqueous solution.

Theory and Computational Methods

Data set. In the optimization of the atomic parameters with genetic algorithm to calculate molecular solvation free energies, we worked with 20 amino acids for which experimental solvation energies have been reported.[24,25] All of the amino acids were subjected to the CORINA program to generate their 3-D coordinates in the Sybyl MOL2 format.[26] As implemented in CORINA, only a single conformation of each amino acid was generated based on the conformational parameters derived from the X-ray structures of small molecules. The 3-D structures obtained in this way have been shown to be similar to the molecular geometries optimized with the semiempirical AM1 calculations including solvation effects,[27] which indicates the reasonableness of the molecular structures derived with CORINA.

Definition of atom types. Different atom types should have different contributions to solvation free energy in the present solvation model under investigation. We used 17 basic atom types for the elements commonly found in amino acids. The atom type of a given atom in an amino acid was differentiated according to element, hybridization state, and chemical environment around the atom under consideration. Considering the portability and the simplicity of implementation of the classifications, all atom types were designated in the same fashion as in the Sybyl MOL2 format.

Optimization of atomic volume parameters with genetic algorithm. Three kinds of atomic parameters need to be optimized in order to calculate the solvation free energy of an amino acid based on the solvent contact model. Among them, the atomic volume parameter V_j represents the fragmental volume of atom j in a molecule. Because the V_j values exhibited a bad convergent behavior in the simultaneous optimization of the three kinds of parameters, they were optimized with the operation of an independent genetic algorithm as detailed below.

The total volume of an amino-acid molecule should be determined prior to the parametrization of V_j values. For this purpose, each amino acid was placed in a 3-D box whose length, width, and height correspond to the maximum distances along the three axes defining the coordinate system of its van der Waals volume. Monte Carlo simulations involving the random selections of a point in the predefined 3-D box were then carried out to calculate the total volume of the amino acid (V_{mol}) embedded in the box. In this simulation, V_{mol} could be obtained by the volume of the box (V_{box}) multiplied by the ratio of the number of trials to select a point in the molecular van der Waals volume (N_{hits}) to the total number of trials (N_{trials}). Thus, we have

$$V_{mol} = V_{box} \times \frac{N_{hits}}{N_{trials}} . \quad (1)$$

With the calculated V_{mol} values in hand, the atomic volume parameters were optimized with the standard genetic algorithm. A generation was defined with 100 vectors comprising the V_j parameters, followed by the removal of 50 with a bias toward preserving the most fit with the lowest error. The empty 50 vectors were then filled with point mutations to alter the value of one of the parameters with probability 0.01, and with cross breeds with probability 0.6 to select some parameters from one vector to replace the elements of another vector of the top 50. The 50 newly created vectors were then evaluated together with the top 50. This cycle was repeated as many times as desired. In the evaluation of the 100 vectors, we used a gradient-based minimization method on the error hypersurface (F_V). This hypersurface is defined by the sum of the absolute values of the differences between the calculated V_{mol} value of an amino acid and the sum of V_j values in the amino acid.

$$F_V = \sum_{k}^{molecules} \left| V_{mol}^{k} - \sum_{j}^{atoms} V_j \right| . \quad (2)$$

Calculation of solvation parameters with genetic algorithm. The solvent contact model to calculate the molecular solvation free energy is based on several fundamental assumptions. First, the solvation free energy of an amino acid k can be approximated by the sum of individual atomic contributions.

$$\Delta G_{calc}^{k} = \sum_{i}^{atoms} \Delta G_{sol}^{i} \quad (3)$$

Second, the individual solvation energy of an atom i can be given by the product of the atomic solvation parameter (S_i) and the degree of its exposure to bulk solvent (F_i).

$$\Delta G_{sol}^{i} = S_i F_i \quad (4)$$

Third, the atomic degree of exposure is approximated as the percentage of the unoccupied volume around the atom in the amino acid. The occupied volume around the atom i (O_i) can then be determined by summing the atomic volume parameters representing the fragmental volumes of all other atoms in the amino acid multiplied by a suitable envelope function, $E(r_{ij})$, with respect to the distance between the centers of atoms i and j.

$$O_i = \sum_{j \neq i}^{atoms} V_j E(r_{ij}) \quad (5)$$

Here, the two kinds of envelope function are taken into account: Gaussian and screened Coulomb potential (SCP) types. Because F_i is the difference between the maximum occupancy of atom i (O_i^{max}) in an amino acid[21] and O_i, the solvation free energy of an amino acid k can be expressed in the following two forms:

$$\Delta G_{calc}^{k} = \sum_{i}^{atoms} S_i \left(O_i^{max} - \sum_{j}^{i \neq j} V_j e^{-\frac{r_{ij}^2}{2\sigma^2}} \right), \quad (6)$$

$$\Delta G_{calc}^{k} = \sum_{i}^{atoms} S_i \left(O_i^{max} - \sum_{j}^{i \neq j} V_j \frac{e^{-\frac{r_{ij}}{\sigma}}}{r_{ij}} \right) \quad (7)$$

Therefore, the two atomic parameters (S_i and O_i^{max}) need also to be optimized in addition to V_j to estimate the solvation free energy of amino acids. These parameterizations were carried out by operating the genetic algorithm with the same procedure as in the optimization of atomic volume parameters. We used a gradient-based minimization method on the error hypersurface defined by the sum of the absolute values of the differences between the calculated and experimental solvation free energies. Formally this fitness function is defined as

$$F_s = \sum_{i=1}^{molecules} \left| \Delta G_{exp}^{i} - \Delta G_{calc}^{i} \right|. \quad (8)$$

Results and Discussion

Listed in Table 1 are the optimized atomic volume (V_j), maximum atomic occupancy (O_i^{max}), and atomic solvation parameters (S_i) for the 17 atom types that are necessary to depict 20 amino acid molecules. Three kinds of atomic solvation energy parameters are thus extended from the earlier ones obtained by Stouten *et al*. to those optimized in this study to cope with a variety of atom types in amino acids. The V_j values calculated in this study are very different from the atomic volumes of isolated atoms. The reason lies in that each V_j value represents the average of the contributions of the atom with type *j* to the van der Waals volumes of various sizes and shapes the amino acids can have. This indicates that the V_j values may exhibit a strong dependence on the amino acids whereas the atomic volume of an isolated atom has to be a constant value.

The optimized S_i parameters reveal a trend consistent with general atomic properties. We note, for example, that the S_i values get more negative in going from sp^3 to sp^2 and carbocation in the case of carbon atoms. This indicates that atomic solvation would be more favorable with the increase of the *s*-character in the hybridization of atomic orbitals of a carbon atom. This is not surprising because the increase in *s*-character raises the electronegativity of carbon atom, which has an effect of increasing the stability in aqueous solution by facilitating the hydrophilic interactions with solvent molecules. In the case of nitrogen atom, on the contrary, the S_i values are shown to be more negative in the order of $sp^3 > sp^2$, which exhibits the same trend as in the order of basicity. The greater basicity of a nitrogen atom with the lower *s*-character is attributed, in general, to the reduced electronegativity that is responsible for the increase in the tendency of the lone electron pair to react with proton in aqueous solution. It is thus apparent that the nitrogen atom with higher basicity has a greater tendency to be stabilized by the establishment of hydrogen bond interactions with solvent molecules, which can be invoked to explain its more negative atomic solvation parameter than the less basic nitrogen atom. An oxygen atom appears to have smaller absolute S_i values than the nitrogen atom with the same hybridized atomic orbitals due most probably to the decrease in basicity that has an effect of reducing the interactions with solvent molecules.

Table 2 lists the calculated solvation free energies for the amino acids with the optimized atomic parameters shown in Table 1, in comparison with the experimental ones. It is seen that the calculated solvation free energies compare reasonably well with the experimental results irrespective of the envelope functions used in the optimization. 14 out of 20 amino acids

Table 1. The Atomic fragmental volume (V_j), maximum atomic occupancy (Occ_i^{max}), and atomic solvation parameters (S_i) optimized with genetic algorithm

atom type	description	V_j (Å^3)	Occ_i^{max} (Å^3)	S_i (kcal/mol Å^3)
C.3	sp^3 carbon	19.682	485.714	1.190
C.2	sp^2 carbon	17.698	466.257	0.905
C.ar	aromatic carbon	19.286	457.143	0.809
C.cat	carbocation	18.889	461.905	−4.523
N.3	sp^3 nitrogen	14.508	380.952	−16.508
N.2	sp^2 nitrogen	12.539	338.095	−11.766
N.am	amidic nitrogen	12.935	352.381	−12.381
N.4	positively charged sp^3 nitrogen	12.143	385.714	−23.353
N.pl3	trigonal planar nitrogen	12.381	328.571	−11.746
O.3	sp^3 oxygen in hydroxyl group	13.432	309.524	−4.416
O.2	sp^2 oxygen	12.952	290.476	−5.714
O.co2	carboxylate oxygen	16.571	276.190	−13.431
S.3	sp^3 sulfur	25.714	585.714	−5.397
H	hydrogen bonded to carbon	5.429	247.619	2.619
H.2	hydrogen bonded to nitrogen	3.857	238.095	−2.857
H.3	hydrogen bonded to oxygen	3.396	214.286	−3.809
H.4	hydrogen bonded to sulfur	2.142	258.143	0.238

Table 2. Experimental and calculated solvation free energies (in kcal/mol) of 20 amino acids

amino acid	experiment	calculation (gaussian envelope function)	calculation (SCP envelope function)
alanine	0.79	−2.18	−1.86
arginine	−21.07	−19.13	−21.01
asparagine	−10.83	−9.11	−8.25
aspartate	−12.1	−9.43	−12.76
cystein	−2.39	−2.76	0.49
glutamine	−10.53	−8.68	−8.01
glutamate	−11.35	−8.02	−11.54
glycine	−1.15	−3.86	−3.12
histidine	−11.42	−9.95	−8.33
isoleucine	1.00	2.47	1.86
leucine	1.13	2.39	1.85
lysine	−10.67	−8.31	−8.67
methionine	−2.63	−1.17	−0.99
phenylalanine	−1.91	−1.34	−2.13
proline	0.20	1.37	1.14
serine	−6.21	−5.98	−4.03
threonine	−6.03	−5.30	−3.78
tryptophan	−7.03	−6.07	−6.37
tyrosine	−7.26	−6.17	−5.30
valine	0.84	0.96	0.61

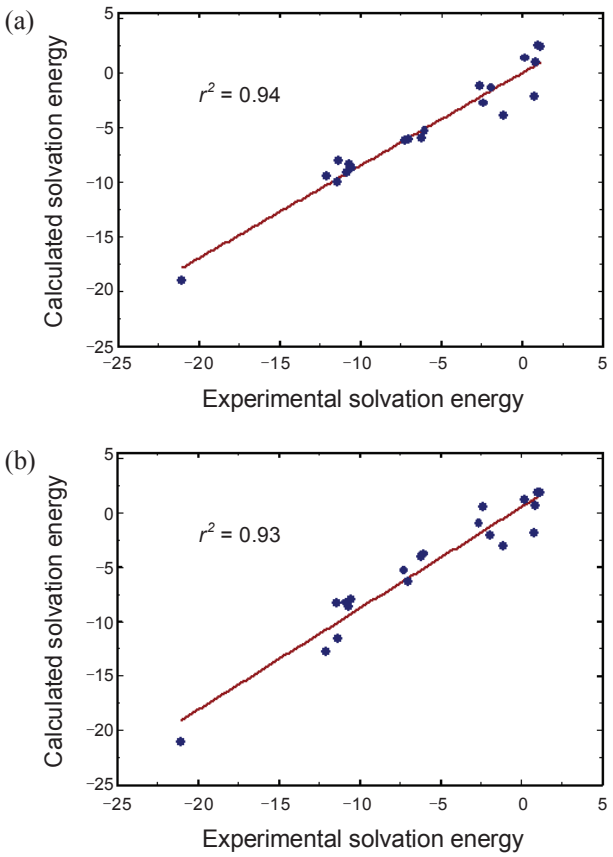

Figure 1. Correlation between experimental versus calculated solvation free energies of amino acids with the parametrization with (a) Gaussian and (b) SCP envelope functions. All energy values are given in kcal/mol.

have deviations of < 2 kcal/mol from the corresponding experimental values if the Gaussian envelope function is used, while 13 amino acids have such deviations in the optimization with SCP envelope function. It is noteworthy that SCP envelope function should be more suitable than the Gaussian function in estimating solvation free energies of the charged amino acids: the deviation from the experimental solvation free energies is no more than 0.73 kcal/mol on average for the SCP function, as compared to 2.5 kcal/mol for the Gaussian function. This is because SCP is more long-range than Gaussian, and therefore should be more accurate in describing the electrostatic interactions with bulk solvent. On the other hand, Gaussian envelope function seems to be superior to the SCP for the neutral residues, which is not surprising for the relative insignificance of the electrostatic interactions.

The correlation between the experimental and the calculated solvation free energies are illustrated in Figure 1. In case of the Gaussian envelope function, we obtain the squared correlation coefficient (r^2) of 0.94, which is a little better than the fitting with the SCP envelope function. The accuracy of the present method in predicting molecular solvation free energy is better than that of the QSPR model trained with 775 compounds in which some drug-like properties of organic compounds computed from their 2-D structures were used as molecular descriptors.[28] The quality of the present solvation model is also superior to that of the artificial neural network (ANN) model reported by Liu and So which was trained with 1033 compounds using 19 adjustable variables.[29] The comparisons thus indicate that our GA-based parameterization method would be more efficient in estimating molecular solvation free energies. Most probably, such an enhanced efficiency is due to the direct use of 3-D structures in the parameterizations rather than 1-D or 2-D molecular descriptors as in the other methods.

The present GA-based solvation model involving the atomic parameterizations has a few advantages over the traditional statistical models. First, 3-D molecular coordinates have only to be provided prior to the optimizations of the individual atomic parameters. The computational cost for molecular solvation free energy can therefore be reduced to a substantial extent as compared to the other methods in which molecular descriptors should be calculated to construct a statistical model for solvation. Such a computational acceleration enables the present solvation model to be an appropriate tool to cope with large chemical libraries as well as with amino acids. Second, the solvation free energy function given in Eqs. (6) and (7) and the newly developed parameters can be incorporated into the potential energy function of a protein in aqueous solution as an implicit solvation model. This effective solvation term is likely to be efficient in terms of both saving computational cost for atomistic simulations and exploring structural properties of proteins just as Stouten *et al.*'s previous solvent contact model revealed such an efficiency and accuracy in molecular dynamics simulation of proteins in aqueous solution.[21] Finally, the accuracy of the present GA-based solvation model can be enhanced in a straightforward way by subdividing the atom types according to chemical environment around the atom of interest. The

atomic parameters of some atom types including C.cat, N.4, and H.4 could not be determined with accuracy due to the small number of amino acids. We will improve the atomic solvation parameters using the experimental solvation energy data for polypeptides as many as possible.

At the present, it is difficult to optimize the atomic parameters so as to be able to predict the solvation free energy of a protein due to the lack of experimental data. We will, in this regard, also try to obtain a proper solvation model for proteins in collaboration with experimental groups.

Conclusions

We have shown the outperformance of the modified solvent contact model involving the GA-based atomic parameterizations in predicting molecular solvation free energies of amino acids in aqueous solution. The present solvation model is based only on 3-D molecular coordinates with no additional molecular descriptors being required to calculate solvation free energy. Using the newly developed atomic parameters for 17 atom types with genetic algorithm, the solvation model was able to predict the experimental solvation free energies of amino acids with the r^2 values of 0.94 and 0.93 for the parameterization with Gaussian and SCP envelope functions, respectively. Considering the efficiency in energy calculation and the simplicity in model refinement by subdividing the atom types, we expect that the present solvation model will be a new useful tool for rapid calculation of the molecular solvation free energy of proteins.

Acknowledgments. This work was supported by the grant from Korea Institute of Oriental Medicine (Grant No. K10060).

References and Notes

1. Rost, B.; Sander, C. *J. Mol. Biol.* **1993**, *232*, 584.
2. Parsegian, V. A.; Rand, R. P.; Fuller, N. L.; Rau, D. C. *Methods Enzymol.* **1986**, *127*, 400.
3. Liu, K.; Cruzan, J.; Saykally, R. *Science* **1996**, *271*, 929.
4. Makarov, V.; Pettitt, B. M.; Feig, M. *Acc. Chem. Res.* **2002**, *35*, 376.
5. Gu, C.; Lustig, S.; Trout, B. L. *J. Phys. Chem. B* **2006**, *110*, 1476.
6. Vorobjev, Y. N.; Vila, J. A.; Scheraga, H. A. *J. Phys. Chem. B* **2008**, *112*, 11122.
7. Jorgensen, W. L.; Duffy, E. M. *Adv. Drug Delivery Rev.* **2002**, *54*, 355.
8. Wesson, L.; Eisenberg, D. *Protein Sci.* **1992**, *1*, 227.
9. Eisenberg, D.; Mclachlan, A. D. *Nature* **1986**, *319*, 199.
10. Ooi, T.; Oobatake, M. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 2859.
11. Schiffer, C. A.; Caldwell, J. W.; Stroud, R. M.; Kollman, P. A. *Protein Sci.* **1992**, *1*, 396.
12. Kang, Y. K.; Nemethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1987**, *91*, 4105.
13. Colonnacesari, F.; Sander, C. *Biophys. J.* **1990**, *57*, 1103.
14. Born, M. Z. *Physics* **1920**, *1*, 45.
15. Kirkwood, J. G. *J. Chem. Phys.* **1934**, *2*, 351.
16. Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333.
17. Klapper, I.; Hagstrom, R.; Fine, R.; Sharp, K. *Proteins* **1986**, *1*, 47.
18. Garde, S.; Hummer, G.; Garcia, A.; Pratt, L.; Paulaitis, M. *Phys. Rev. E* **1996**, *53*, R4310.
19. Hummer, G.; Soumpasis, D. *Phys. Rev. E* **1994**, *50*, 5085.
20. Colonna-Cesari, F.; Sander, C. *Biophys. J.* **1990**, *57*, 1103.
21. Stouten, P. F. W.; Frömmel, C.; Nakamura, H.; Sander, C. *Mol. Simul.* **1993**, *10*, 97.
22. Park, J.-H.; Ko, S.; Park, H. *Bull. Korean Chem. Soc.* **2008**, *29*, 921.
23. Park, H.; Jung, S.-K.; Bahn, Y. J.; Jeong, D. G.; Ryu, S. E.; Kim, S. J. *Bull. Korean Chem. Soc.* **2009**, *30*, 1313.
24. Dixit, S. B.; Bhasin, R.; Rajasekaran, E.; Jayaram, B. *J. Chem. Soc. Faraday Trans.* **1997**, *93*, 1105.
25. Chang, J.; Lenhoff, A. M.; Sandler, S. I. *J. Phys. Chem. B* **2007**, *111*, 2098.
26. Sadowski, J.; Gasteiger, J.; Klebe, G. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000.
27. Goller, A. H.; Hennemann, M.; Keldenich, J.; Clark, T. *J. Chem. Inf. Model* **2006**, *46*, 648.
28. Cheng, A.; Merz, K. M., Jr. *J. Med. Chem.* **2003**, *46*, 3572.
29. Liu, R.; So, S.-S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633.