# Novel Procedure for Developing Implicit Solvation Models for Peptides and Proteins

**Canan Baysal and Hagai Meirovitch***

*Supercomputer Computations Research Institute, Florida State University, Tallahassee, Florida 32306*

*Received: July 3, 1997*[⊗]

Solvation is an important factor in the structure stabilization of proteins. The free energy of solvation has been commonly approximated by summing the products of the atomic solvation parameters (ASPs) and the solvent accessible surface areas, where the ASPs were obtained from thermodynamic experiments of small molecules. This summation was usually added to the force field without calibration. We propose deriving an optimized set of ASPs *only* by requiring that the minimized total energy of the experimentally known structure is the *global* minimum. This method is applied to a cyclic hexapeptide in DMSO and can also be extended to loops in proteins in water.

The structure and function of biological macromolecules, such as peptides and proteins, are strongly affected by the solvent. Ideally, one would seek to study the protein in an explicit solvent; i.e. it is immersed in a large "box" of water molecules, and the protein−protein, water−water, and water−protein interactions are defined. The system can then be studied by molecular dynamics (MD) simulations. However, these simulations are very time consuming, and it is not feasible to calculate differences in the free energy between conformational states of a significant variance. Such stability calculations are important, for example, in the structure determination of loops in homologous proteins.

These problems can be alleviated by studying a less detailed partition function $Z(\mathbf{x})$, which is obtained by integrating the exponential $\exp[-E/k_BT]$ of the *system* energy, $E$, over the solvent coordinates only, for each set of the protein coordinates $\mathbf{x}$ ($\mathbf{x}$ is a $3N$ vector of the Cartesian coordinates of the $N$ protein atoms; $k_B$ and $T$ are the Boltzmann constant and the absolute temperature, respectively).[1,2] If $E$ is the *exact* potential energy, $Z(\mathbf{x})$ leads to the exact potential of mean force, $E_{exa}(\mathbf{x}) = -k_BT \ln Z(\mathbf{x})$. Note that although $E_{exa}$ is a free energy function, it will be referred to as an energy.

While $E_{exa}$ lacks the microscopic information about the solvent molecules, it correctly defines the stable regions of the protein. This energy surface is typically "decorated" by a tremendous number of potential energy wells (localized microstates) centered around local energy minima. The molecule is expected to visit a localized microstate for a very short time while staying for larger periods within a "wide microstate" $\Omega_j$, which is a larger energy well, consisting of a group of neighboring localized microstates. Therefore, one is interested in the most stable $\Omega_j$, i.e. those with the largest contribution, $Z_j$, to the total partition function of the molecule. $Z_j$ is obtained by integrating $\exp[-E_{exa}(\mathbf{x})/k_BT]$ over $\Omega_j$. $E_{exa}$ has been commonly approximated by $E_{tot}$

$$E_{tot} = E_{FF} + E_{sol} = E_{FF} + \sum_i \sigma_i A_i \qquad (1)$$

where $E_{FF}$ is the force field energy and the summation runs over all the atoms $i$, $A_i$ is the solvent accessible surface area (SASA), and $\sigma_i$ is the atomic solvation parameter (ASP) of atom $i$. Various sets of ASPs have been derived on the basis of the free energy of transfer of small molecules from the gas phase to water.[3−15] This approach is feasible due to the recent development of efficient methods for calculating the SASA and its derivatives.[6,10,16−22]

Such solvation models were shown in most cases to outperform the usual force fields, leading to relatively strong correlations between the (minimized) energy of a conformation and its root mean square deviation from the X-ray structure.[6,9−14,20] However, some of the sets of ASPs were found to lead to unsatisfactory results,[11,12,23−25] probably because typically $E_{sol}$ was added to the energy $E_{FF}$ of various force fields (using different dielectric constants) without further calibration. Indeed, Schiffer et al.[11,12] and more recently Fraternali and van Gunsteren[24] have checked several sets of ASPs in MD simulations and found that they should be properly scaled and sometimes changed in order to recover the correct experimental structural data of proteins and to capture the results obtained by MD simulations with explicit water.

We propose to derive the ASPs solely on the basis of free energy considerations and have applied the method to a peptide in DMSO. Thus, the ASPs are not obtained from the experimental free energy of transfer of small molecules from the gas phase to DMSO but are defined as the set for which the free energy of the wide microstate of the experimental structure is *globally* minimal. However, as a first step, in this report we determine the best ASPs on the basis of energy considerations alone; entropic effects will be studied later.

The structure of the cyclic peptide *cyclo*-(D-Pro[1]-Phe[2]-Ala[3]-Ser[4]-Phe[5]-Phe[6]) in DMSO was investigated using 2-D NMR and X-ray crystallography by Kessler et al.[26] In that study, a *single* conformation that completely satisfies the nuclear Overhauser enhancement (NOE) distance constraints was not found. However, the data could satisfactorily be explained by two conformations that have the same $\beta$II′-turn around D-Pro[1] and Phe[2] but different turns ($\beta$I and $\beta$II) around Ser[4] and Phe[5]. This analysis included MD simulations using the GROMOS force field, $E_{GRO}$. These two structures are defined with respect to the backbone only; therefore, they should be viewed as wide microstates with backbone and side chain conformational flexibility; i.e. many energy minimized structures belong to each of these wide microstates. We shall refer to them as the $\beta$I and $\beta$II motifs. We have first reconfirmed the findings of Kessler et al.[26] by a best-fit analysis based on a large set of energy-minimized structures generated by our "local torsional deformations" (LTD) method,[27,28] using $E_{GRO}$ alone and $E_{GRO}$ with an

* To whom correspondence should be addressed. E-mail: hagai@scri.fsu.edu.

⊗ Abstract published in *Advance ACS Abstracts,* September 1, 1997.

S1089-5647(97)02175-5 CCC: $14.00 

artificial potential that forces the structures to satisfy the NOE distance restraints.

The results demonstrate that $E_{GRO}$ alone is inappropriate to describe this system because the lowest minimized values found for the $\beta$I and $\beta$II motifs are, respectively, ~15 and 5 kcal/mol above the lowest minimized energy obtained by LTD. One would expect that for $E_{exa}$ (which is unknown), the lowest energy minimized structures including the global energy minimum (GEM) would pertain to the $\beta$I and $\beta$II motifs. This is because these wide microstates have the lowest global free energy $-k_BT$ ln $Z_j$ and energetic effects are still dominant at $T = 300$ K.

These considerations dictate our "working criterion" for calculating the optimal ASP set: The energy of the lowest energy-minimized structures that represent the above two microstates should be as close as possible to the GEM($E_{tot}$) and within the 2 kcal/mol range above it (see discussion in refs 29 and 30). This is achieved by applying an iterative procedure that relies on an extensive conformational search with LTD. Entropic effects will be taken into account in a following study, where the optimal ASPs will be used in MD simulations of the wide microstates of the $\beta$I, $\beta$II, and other structures. Their free energies, and hence relative populations, will be obtained with the local states method,[30−32] which might lead to a further refinement of the ASPs derived here.

The optimization process consists of several stages: We started from a set of 55 test structures selected from the large sample generated with LTD based on $E_{GRO}$. These structures, which include representatives of the $\beta$I and $\beta$II motifs, are of diverse energies, backbone motifs, and $\chi_1$ values. The radius of the spherical probe that represents a DMSO molecule is taken as 3 Å, and the SASA is calculated by the program MSEED.[17] At the first stage only a single ASP is considered; i.e. $\sigma_i = \sigma$ for all the atoms in eq 1. After a value of $\sigma$ is selected, all the 55 structures are reminimized with respect to $E_{tot}(\sigma)$, and it is verified that the resulting changes in the dihedral angles do not exceed several degrees; i.e. the structural motifs are preserved. Then, several structures with the lowest energies, as well as those of the $\beta$I and $\beta$II motifs, are given further consideration. Thus, for each of the selected backbone motifs, the 27 combinations of side chain conformations, based on the three $\chi_1$ values (60, −60, and 180°) of the three Phe residues, are generated and minimized with respect to $E_{tot}(\sigma)$. Finally, the energy differences between the lowest energy found, $E_L(\sigma)$, and the best representatives of the $\beta$I and $\beta$II motifs are calculated. This procedure is repeated for many different values of $\sigma$, and its optimal value $\sigma^*$ is determined according to the criterion discussed above. Then, an extensive conformational search is carried out with LTD, based on 5000 minimizations of $E_{tot}(\sigma^*)$; if new structures with energy lower than $E_L(\sigma^*)$ are not obtained, the optimization of $\sigma$ is stopped. Otherwise, the new low-energy structures are added to the set, and a new round of optimization for $\sigma$ is carried out.

The final set of structures obtained in the search for the best single parameter is used as the starting set for the two-parameter optimization, where the O atoms are allowed to take on a different ASP value than the rest of the atoms. The optimization procedure is continued as before and is finally extended to three $\sigma_i$, where an additional parameter is assigned to the H atoms. As for the single-parameter case, extensive LTD runs of 5000 minimizations are performed. The set of test structures increased from 55 to 75 during the overall process.

The results are summarized in Table 1, which reveals that as the number of $\sigma_i$ is increased from 0 to 3, the energy difference $E_{\beta I} - E_L$ decreases monotonically from 15.2 to 0.8 kcal/mol. The corresponding decrease for the $\beta$II structure is from 5.7 to 0 kcal/mol. It should be noted that in the three-parameter case, two additional energy-minimized structures of different backbone motifs were also found within the 0.8 kcal/mol energy range; the relative stability of these four wide microstates will be determined in our next study where the entropic contributions will be calculated. Nevertheless, for three ASPs our optimization criterion can be considered as fully satisfied, since the $\beta$II motif represents the GEM structure. Consequently, optimization with respect to a larger number of ASPs is not attempted. An LTD run of the current cyclic hexapeptide with 5000 LTD minimizations is expected to find the GEM, based on previous conformational search results using molecules of comparable size. For cycloheptadecane the GEM is found within several hundred LTD minimizations, and for the linear pentapeptide Leu-enkephalin within 3000 minimizations using a related method.[28] However, one cannot rule out the possibility that the GEM structure may have been missed.

**TABLE 1: Optimized ASP Sets[a]**

| $\sigma_C$ | $\sigma_H$ | $\sigma_O$ | $E_{\beta II}-E_L$ | $E_{\beta I}-E_L$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 5.7 | 15.2 |
| −90 | −90 | −90 | 3.2 | 8.6 |
| −55 | −55 | −175 | 2.1 | 3.7 |
| −45 | −150 | −205 | 0.0 | 0.8 |

[a] The ASPs ($\sigma_i$) are in cal/mol/Å$^2$, the energy $E$ is in kcal/mol. $E_L$ is the lowest minimized value of $E_{tot}$ (eq 1) for the given ASP set; $E_{\beta I}$ and $E_{\beta II}$ are the lowest minimized energies obtained for structures that belong to the experimental $\beta_I$ and $\beta_{II}$ motifs.

All the ASPs in Table 1 are negative, which means that $E_{tot}$ prefers structures with larger SASA than those preferred by $E_{GRO}$ alone. $\sigma_O = -205$ cal/mol/Å$^2$ is the most effective ASP. This can be understood in terms of electrostatic interactions between the molecule and the explicit model of DMSO defined within the GROMOS package (note, however, that we have not carried out simulations with explicit DMSO). Due to the relatively large negative partial charge (−0.38 eu) of the carbonyl oxygen, these atoms prefer to be exposed to the solvent and interact with the S atom and the two $CH_3$ groups of DMSO with partial charges of +0.139 and +0.160, respectively. The partial charge, +0.28, of H(N) is smaller in absolute value than that of O, which expresses the somewhat weaker preference of H to be exposed ($\sigma_H = -150$ cal/mol/Å$^2$). Most of the C atoms are uncharged, and their preference for exposure to the solvent ($\sigma_C = -45$ cal/mol/Å$^2$) is therefore relatively small, mainly originating from Lennard-Jones interactions with the DMSO molecules and entropic effects. Notice that due to geometrical constraints of the cyclic molecule, the SASA of the N atoms is zero; therefore, assigning them the same ASP as the C atoms does not affect the results. Obviously these results should be verified for other peptides in DMSO; it is also of interest to see the effect of the force field on the ASPs.

The dependence of the relevant energy differences on $\sigma_i$ is presented in Figure 1. The figure shows that the energy of the $\beta$I structure is more sensitive to the change of the ASPs than that of the $\beta$II structure. However, one should bear in mind that the graphs are mainly based on the set of 75 structures that were accumulated during the optimization process; further conformational search for the lowest energy structure was carried only for part of the parameter sets displayed. This might be the reason why the $\beta$II structure remains the $E_L$ for a relatively large range of parameter values.

In summary, we have described a new way for deriving atomic solvation parameters. This is the first time ASPs have been obtained for a peptide in DMSO, and they can readily be used to analyze NMR data using GROMOS. The present method is applicable to a wide range of solvents, and in

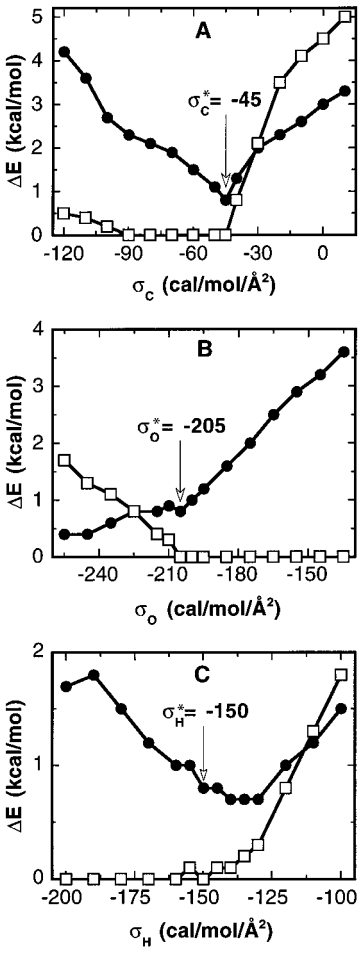


**Figure 1.** Energy differences $E_{\beta II}-E_L$ (●) and $E_{\beta II}-E_L$ (□) as a function of the ASPs. One parameter is varied while the other two are held fixed in their optimal values of Table 1. For each parameter the energy $E_{tot}$ is minimized for all the 27 different combinations of the side chain dihedral angles $\chi_1$ of the Phe residues, and the lowest energy is used for calculating the energy difference. For almost the same range of parameters, the range of the energy differences is the largest for $\sigma_C$ (A), decreasing for $\sigma_O$ (B) and $\sigma_H$ (C).

particular to surface loops of a protein in water. Implicit models based on ASP's are useful in various areas, such as the analysis of NMR and X-ray data, docking of a ligand to an active site of an enzyme, structure determination based on homology, and the inverse protein-folding problem.

**Acknowledgment.** We thank Dr. M. Vásquez for helpful discussions and acknowledge support from the Florida State University Supercomputer Computations Research Institute, which is partially funded by the U.S. Department of Energy (DOE) under Contract DE-FC05-85ER250000. This work was also supported by DOE Grant DE-FG05-95ER62070.

## References and Notes

(1) Lifson, S.; Oppenheim, I. *J. Chem. Phys.* **1960**, *33*, 109.
(2) Gō, N.; Scheraga, H. A. *J. Chem. Phys.* **1969**, *51*, 4751.
(3) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199.
(4) Wesson, L.; Eisenberg, D. *Protein Sci.* **1992**, *1*, 227.
(5) Ooi, T.; Oobatake, M.; Némethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086.
(6) Stouten, P. F. W.; Frömmel, C.; Nakamura, H.; Sander, C. *Mol. Simul.* **1993**, *10*, 97.
(7) Kang, Y. K.; Gibson, K. D.; Némethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1988**, *92*, 4739.
(8) Smith, K. C.; Honig, B. *Proteins* **1994**, *18*, 119.
(9) Vila, J.; Williams, R. L.; Vásquez, M.; Scheraga, H. A. *Proteins* **1991**, *10*, 199.
(10) Luty, B. A.; Wasserman, Z. R.; Stouten, P. F. W.; Hodge, C. N.; Zacharias, M.; McCammon, J. A. *J. Comput. Chem.* **1995**, *16*, 454.
(11) Schiffer, C. A.; Caldwell, J. W.; Stroud, R. M.; Kollman, P. A. *Protein Sci.* **1992**, *1*, 396.
(12) Schiffer, C. A.; Caldwell, J. W.; Kollman, P. A.; Stroud, R. M. *Mol. Simul.* **1993**, *10*, 121.
(13) von Freyberg, B.; Richmond, T. J.; Braun, W. *J. Mol. Biol.* **1993**, *233*, 275.
(14) Williams, R. L.; Vila, J.; Perrot, G.; Scheraga, H. A. *Proteins* **1992**, *14*, 110.
(15) Eisenhaber, F. *Protein Sci.* **1996**, *5*, 1676.
(16) Richmond, T. J. *J. Mol. Biol.* **1984**, *178*, 63.
(17) Perrot, G.; Cheng, B.; Gibson, K. D.; Vila, J.; Palmer, K. A.; Nayeem, A.; Maigret, B.; Scheraga, H. A. *J. Comput. Chem.* **1992**, *13*, 1.
(18) Eisenhaber, F.; Argos, P. *J. Comput. Chem.* **1993**, *14*, 1272.
(19) Sridharan, S.; Nichols, A.; Sharp, K. A. *J. Comput. Chem.* **1996**, *15*, 1038.
(20) Mumenthaler, C.; Braun, W. *J. Mol. Model.* **1995**, *1*, 1.
(21) Hasel, W.; Hendrickson, T. F.; Still, W. C. *Tetrahedron Comput. Methodol.* **1988**, *1*, 103.
(22) Augspurger, J. D.; Scheraga, H. A. *J. Comput. Chem.* **1996**, *17*, 1549.
(23) van Aalten, D. M. F.; Amadei, A.; Bywater, R.; Findlay, J. B. C.; Berendsen, H. J. C.; Sander, C.; Stouten, P. F. W. *Biophys. J.* **1996**, *70*, 684.
(24) Fraternali, F.; van Gunsteren, W. F. *J. Mol. Biol.* **1996**, *256*, 939.
(25) Juffer, A. H.; Eisenhaber, F.; Hubbard, S. J.; Walther, D.; Argos, P. *Protein Sci.* **1995**, *4*, 2499.
(26) Kessler, H.; Matter, H.; Gemmecker, G.; Kottenhahn, M.; Bats, J. W. *J. Am. Chem. Soc.* **1992**, *114*, 4805.
(27) Baysal, C.; Meirovitch, H. *J. Chem. Phys.* **1996**, *105*, 7868.
(28) Baysal, C.; Meirovitch, H. *J. Phys. Chem. A* **1997**, *101*, 2185.
(29) Meirovitch, E.; Meirovitch, H. *Biopolymers* **1996**, *38*, 69.
(30) Meirovitch, H.; Meirovitch, E. *J. Phys. Chem.* **1996**, *100*, 5123.
(31) Meirovitch, H. *Chem. Phys. Lett.* **1977**, *45*, 389.
(32) Meirovitch, H.; Koerber, S. C.; Rivier, J.; Hagler, A. T. *Biopolymers* **1994**, *34*, 815.