Optimization of Solvation Models for Predicting the Structure of Surface Loops in Proteins

Bedamati Das and Hagai Meirovitch*
School of Computational Science and Information Technology, Florida State University, Tallahassee, FL

ABSTRACT **A novel procedure for optimizing the atomic solvation parameters (ASPs) σ_i developed recently for cyclic peptides is extended to surface loops in proteins. The loop is free to move, whereas the protein template is held fixed in its X-ray structure. The energy is $E_{tot} = E_{FF}(\epsilon = nr) + \sum \sigma_i A_i$, where $E_{FF}(\epsilon = nr)$ is the force-field energy of the loop–loop and loop–template interactions, $\epsilon = nr$ is a distance-dependent dielectric constant, and n is an additional parameter to be optimized. A_i is the solvent-accessible surface area of atom i. The optimal σ_i and n are those for which the loop structure with the global minimum of $E_{tot}(n, \sigma_i)$ becomes the experimental X-ray structure. Thus, the ASPs depend on the force field and are optimized in the protein environment, unlike commonly used ASPs such as those of Wesson and Eisenberg (Protein Sci 1992;1:227–235). The latter are based on the free energy of transfer of small molecules from the gas phase to water and have been traditionally combined with various force fields without further calibration. We found that for loops the all-atom AMBER force field performed better than OPLS and CHARMM22. Two sets of ASPs [based on AMBER (n = 2)], optimized independently for loops 64–71 and 89–97 of ribonuclease A, were similar and thus enabled the definition of a best-fit set. All these ASPs were negative (hydrophilic), including those for carbon. Very good (i.e., small) root-mean-square-deviation values from the X-ray loop structure were obtained with the three sets of ASPs, suggesting that the best-fit set would be transferable to loops in other proteins as well. The structure of loop 13–24 is relatively stretched and was insensitive to the effect of the ASPs. Proteins 2001;43: 303–314.**
© 2001 Wiley-Liss, Inc.

Key words: proteins; atomic solvation parameters; surface loops; energy minimization; conformational search

INTRODUCTION

Surface loops of proteins in solution are relatively flexible, as found by multidimensional nuclear magnetic resonance (NMR) experiments. In many cases, this flexibility is also demonstrated in X-ray crystallography data in terms of large B-factors[1] and sometimes a complete disorder. However, numerous examples are known in which the structural flexibility of loops is essential to the function of proteins. Thus, the conformational change between a free antibody and a bound antibody demonstrates the flexibility of the antibody combining site, which typically includes hypervariable loops; this provides an example of *induced fit* as a mechanism for antibody–antigen recognition (e.g., Refs. 2 and 3). Alternatively, a mechanism called *selected fit* has been suggested, in which the free active site interconverts among different states, where one of them is selected upon binding[4]; the same also applies for loops. Dynamic NMR experiments[5] and molecular dynamics simulations[6] of HIV protease have found a strong correlation between the flexibility of certain segments of the protein and the movement of the flaps (that cover the active site) upon ligation.[7] Loops are known to form lids over active sites of proteins, and mutagenesis experiments show that residues within these loops are crucial for substrate binding or enzymatic catalysis; again, these loops are typically flexible (see the review by Fetrow[8]).

The interest in surface loops has yielded extensive theoretical work, where one avenue of research has been the classification of loop structures.[8–15] However, to understand various recognition mechanisms like those previously mentioned, we must be able to predict the structure (or structures) of a loop by theoretical/computational procedures. As discussed in detail later, this is not a trivial task because of the irregular structures of the loops, their flexibility, and their exposure to the solvent, which requires developing adequate modeling of solvation. In fact, the structure prediction of loops constitutes a challenge in protein engineering, where a loop undergoes mutations, insertions, or deletions of amino acids. The determination of the structure of large loops is still an unsolved problem in homology modeling.[16–18]

Loop structures are commonly predicted by a comparative modeling approach based on known loop structures from the Protein Data Bank (PDB), an energetic approach, or methods that are hybrid of these two approaches. To the first approach pertains the canonical structure method for hypervariable loops of antibodies.[19,20] Other methods in this category are based on matching segments from the

Grant sponsor: U.S. Department of Energy; Grant number: DE-FG05-95ER62070.

*Correspondence to: Hagai Meirovitch, School of Computational Science and Information Technology, Florida State University, Tallahassee, FL 32306-4052. E-mail: hagai@csit.fsu.edu

data base with the length of a target loop and the relative positions of its adjacent residues. Hence, these procedures are especially appropriate for homology studies, where the protein framework is not known exactly. However, only short loops of up to five residues can be treated effectively with this approach.[21–23] With a recent hybrid method,[24] the results for several three-target loops of seven and eight residues are comparable to the best predictions obtained in CASP3 (Critical Assessment of Methods of Protein Structure Prediction), but the root-mean-square deviation (RMSD) values from the X-ray structures are still relatively high, 1.75–2.80 Å, whereas only an RMSD less than 1 Å is considered to be satisfactory.[25,26] To date, statistical methods cannot handle loops of more than $n = 9$ residues because of the lack of sufficiently large data bases.[27]

With the energetic approach, loop structures are generated by a conformational search method subject to the spatial restrictions imposed by the known three-dimensional structure of the rest of the protein (the template). The quality of the prediction depends on the quality of the loop–loop and loop–template interaction energy and the extent of conformational search applied. With one method,[28,29] randomly generated backbone chains are tweaked to the template edges and treated by energy minimization and molecular dynamics simulations. Moult and James[25] carried out a systematic conformational search in internal coordinates based on various filters to reduce the searched space. Another set of methods is based on the ring-closure algorithm of Gō and Scheraga,[30] which has been used in a systematic search procedure (with filters) in dihedral-angle space.[26,31] Structures of loops were also generated by simulated annealing[32,33] and the bond-scaling relaxation algorithm,[34] which was enhanced by being combined with multiple-copy sampling techniques.[35] These methods are not expected to handle efficiently large loops because of the lack of conformational search capabilities, which in many cases is partially caused by complex construction procedures based on at least two stages in which the side-chains are added to an initially generated backbone.

In some of these studies, the solvation problem is not addressed at all, whereas most of them only use a distance-dependent dielectric constant ($\epsilon = r$). Better treatments of solvation were applied by Moult and James[25] and Mas et al.[36] A systematic comparison of solvation models was first carried out by Smith and Honig,[37] who tested the $\epsilon = r$ model against results obtained by the finite-difference Poisson–Boltzmann calculation, including a hydrophobic term; the implicit solvation model of Wesson and Eisenberg[38] with $\epsilon = r$ was also studied by them. More recently, the generalized Born/surface area (GB/SA) model[39] was applied to loops of ribonuclease (RNase) A.[40] Comparing the efficiency of these methods is not straightforward. However, as expected, the structure prediction improves as the ratio of loop length and distance between ends decreases, and the conformational restrictions imposed by the template increase. For example, for two five-residue loops of bovine pancreatic trypsin inhibitor, which are relatively constrained, average RMSD values of 0.43 and 0.68 Å were obtained for the backbone atoms.[32] Bruccoleri et al.[41] obtained RMSD values within the range 0.7–2.6 Å for backbone atoms and 1.4–4.1 Å for all-atoms for the 12 hypervariable loops (5–12 residues) of the antibodies McPC603 and HyHEL-5; these values are typical for such less restricted systems.[33] Although this imperfection has been attributed to the inadequate modeling of solvation, the better treatments of Smith and Honig[37] have also been found to be inconclusive. The results of Rapp and Friesner[40] have shown strong dependence on the force field. This supports our point of view, discussed later, that solvation parameters should be optimized together with the specific force field used. Thus far, none of these approaches has addressed the problem of loop flexibility in a systematic way.

The foregoing discussion indicates that to date the energetic approach is the best way for predicting the structure of large loops in homology modeling and protein engineering, and it constitutes the only alternative for studying the flexibility of loops. Recently, we developed a statistical mechanics methodology[42–45] for treating flexibility; it was used successfully to predict the solution structures and populations of cyclic peptides in dimethyl sulfoxide (DMSO).[46,47] This methodology relies on (1) a novel method for optimizing atomic solvation parameters (ASPs), (2) an extensive conformational search with our local-torsional-deformation (LTD) method,[44,48] and (3) Monte Carlo simulations and free-energy calculations with the local-state method.[49,50] Our long-range objective is to extend this methodology to loops of proteins in water. In this article, however, we only apply the first two stages of this project; that is, using LTD, we optimize ASPs for loops of RNase A, a process that requires taking flexibility into account and, therefore, involves some elements of the entire methodology. Thus, assuming for a moment that a perfect force field is available, we first provide a short discussion of stages 2–3, as applied to a loop free to move under the restrictions of the template; the treatment of solvation follows this discussion.

THEORY AND METHODS

Methodology for Treating Flexibility

For a long nonstretched loop surrounded by a constant protein template, the number of energy-minimized structures is large, where around each minimum a *localized microstate* is defined, which is the ensemble of loop conformations pertaining to the basin of attraction of the minimum. The energy landscape of the loop also contains larger potential wells defined over regions called *wide microstates,* where each is decorated by many localized ones.[43] Molecular dynamics studies have shown that the molecule will visit a localized microstate only for a very short time (several femtoseconds), although it will stay for a much longer time within a wide microstate,[51–53] which means that the wide microstates have greater physical significance than the localized ones. In other words, structural and thermodynamic properties in solution obtained experimentally for a loop with a well-defined structure should be compared with theoretical values averaged over the most stable wide microstate, which is defined by the local loop fluctuations simulated, for example, by molecular dynamics. A large surface loop might also be a random coil or exhibit intermediate flexibility between these two extreme cases, where several wide microstates are populated significantly in thermodynamic equilibrium.

To determine the extent of flexibility, one should identify the most stable wide microstates i, that is, those with the largest contribution Z_i to the total partition function. From the relative populations, $p_i = Z_i/\sum Z_i$, one can obtain the statistical average $\langle G \rangle = \sum p_i G_i$ of a property G, where G_i is the contribution of wide microstates i. As for peptides, identification of the most stable wide microstates of a loop is carried out in two stages.[42–45] First, with LTD, which enables the crossing of energy barriers, an extensive conformational search is carried out for the global-energy-minimum (GEM) loop structure and low-energy-minimized structures within 2–3 kcal/mol above the GEM. For the pentapeptide Leu-enkephalin and cycloheptadecane, at 280k the corresponding localized microstates contribute 60–75%, respectively, of the total partition function; these values are expected to be typical for peptides and loops of similar size. Therefore, these minimized structures should reside within the most stable wide microstates, and a subgroup of them that are significantly different would represent the different wide microstates because, per definition, structures that pertain to the same wide microstate are similar.

A suitable criterion for the variance of two structures is that at least one dihedral angle differs by 60° or more. This angular criterion, which is based on energetic considerations, has been found to be suitable for a short peptide, whereas for a long peptide or loop, an additional criterion, such as the RMSD between structures, should be employed (see the discussion in Ref. 54). In the second stage, each selected structure becomes a seed for a Monte Carlo or molecular dynamics simulation that spans the related wide microstate. The free energies, F_i, of the most stable wide microstates are obtained with the local-state method applied to the corresponding samples. Criteria developed previously[43,45] enable one to check the structural distinctiveness and thermodynamic stability of the various samples; that is, they do not overlap and remain in their original conformational regions. Initially, a set of optimized ASPs was derived for a cyclic hexapeptide in DMSO based on NMR results of Kessler et al.[55] With this set, ab initio predictions of the solution structures (in DMSO) of a cyclic pentapeptide[46] and two cyclic heptapeptides were carried out.[47] Proton–proton distances and ^3J coupling constants obtained by NMR were reproduced with very good accuracy.

Modeling Solvation Effects

This methodology is useful if applied with a reliable force field that takes into account solvent effects. Although explicit water would probably provide the most accurate modeling, it is computationally time-consuming, and the conformational search is complicated. Also, defining the most stable structure would require comparing the free energies of significantly different wide microstates, which is extremely difficult to achieve with the commonly used perturbation and thermodynamic integration techniques.[56,57] Therefore, the work of McCammon's group[58] on loops of the anti-insulin antibody using explicit water is an exception in this field; in all the other studies (discussed previously), solvent effects are modeled implicitly. As for peptides, we use for loops the simplified implicit solvation model:

$$E_{\text{tot}} = E_{\text{FF}}(\epsilon = nr) + E_{\text{sol}} = E_{\text{FF}}(\epsilon = nr) + \sum_i \sigma_i A_i \quad (1)$$

where E_{FF} is the force-field energy (based only on the loop–loop and loop–template interactions), A_i is the structure-dependent solvent-accessible surface area of atom i, and σ_i is the corresponding ASP that should be optimized. One would expect the optimal ASPs to reasonably express the Born self-energies[59] and the hydrophobic interactions (see, however, later discussions). The screening of the electrostatic interactions by the surrounding water is modeled approximately by a distance-dependent dielectric constant ($\epsilon = nr$), where n is an additional parameter to be optimized together with the σ_i's. Notice that E_{tot} is a free-energy function that depends on the temperature (through the σ_i's) but is referred to as energy.

Equation 1 is not new and has been used in many previous studies where the ASPs for a protein have been commonly determined from the free energy of transfer of small molecules from the gas phase to water.[38,60] However, it is not clear to what extent ASPs derived for small molecules are suited for the protein environment. Also, these sets of ASPs were used with various force fields, in most cases without further calibration (see the discussions in Refs. 43 and 44 and references cited therein). This seems unjustified because the existing force fields are different, and probably none of them is expected to faithfully describe a protein in vacuo; therefore, even if a set of ASPs has been derived that describe correctly the first hydration shell of a protein, E_{tot} would still be inaccurate. In other words, the ASPs should be optimized with respect to the force-field energy used. Recent studies based on various solvation potentials, E_{sol}, support these reservations.[37,40] This problem was noticed first by Schiffer et al.[61] and more recently by Fraternali and van Gunsteren.[62]

Addressing this problem, we adopt here the same philosophy developed for peptides in DMSO, which constitute an alternative to the conventional parameterization described. Thus, the optimal ASPs and n are those for which the GEM structure with respect to $E_{\text{tot}}(n, \sigma_i)$ becomes the experimental X-ray loop structure. This optimization requires an extensive conformational search for the loop, which is carried out with our LTD method described in Appendix A. The optimal ASPs depend on the force field used, and they are based on the energy of the entire loop in the protein environment, in contrast to the conventional parameterization, which relies on free-energy data of small molecules.

RESULTS AND DISCUSSION

In this section, the optimal ASPs for eq 1 are derived on the basis of energy considerations, and their transferability and ability to lead to the correct loop structures are tested.

Loops Studied and Computer Programs Used

To be consistent with previous work,[40] our optimization is applied to loops of RNase A based on its X-ray structure, 1rat.pdb,[67] and the NMR structure, 2aas.pdb,[68] which were found by Rapp and Friesner[40] to be similar (RMSD 1.11 Å). To investigate the effect of the force field on the ASPs, we use the molecular mechanics/molecular dynamics package TINKER,[69] which enables one to apply various sets of force-field parameters. We tested the OPLS,[70] AMBER,[71] and CHARMM22[72] all-atom force fields, where Arg, Lys, His, Asp, and Glu are charged. The radius r_i of atom i, required for calculating the surface area, was determined from its Lennard–Jones parameter σ^i_{LJ}, where $r_i = (2^{1/6}\sigma^i_{\mathrm{LJ}}/2)$; the radius of hydrogen is 0.9 Å. For the calculation of the surface area only, CH, CH_2, and CH_3 were treated as united atoms, and for OPLS their radii were calculated from the corresponding σ_{LJ} values of the united atom OPLS force field; for the AMBER and CHARMM22 force fields, these three united atoms were given the same radius of 2.1 Å. As in our previous work, the surface area and its first derivatives were calculated by the program MSEED[73] incorporated in TINKER, where a water molecule is represented by a sphere of radius 1.4 Å. We also incorporated in TINKER the L-BFGS minimizer[74] and the LTD program.

We first investigated the performance of the three force fields as applied to the 8-residue loop 64–71 (loop 1), Ala^{64}-Cys^{65}-Lys^{66}-Asn^{67}-Gly^{68}-Gln^{69}-Thr^{70}-Asn^{71}, which has a well-defined structure. These studies (described in the next subsection) show that for loops the AMBER force field performs better than both OPLS and CHARMM22; therefore, only AMBER was used for studying the two additional loops of RNase A. One is the 12-residue loop, 13–24 (loop 2), Met^{13}-Asp^{14}-Ser^{15}-Ser^{16}-Thr^{17}-Ser^{18}-Ala^{19}-Ala^{20}-Ser^{21}-Ser^{22}-Ser^{23}-Asn^{24}, which was also studied before by Rapp and Friesner,[40] however, this loop was found to be unsuitable for checking the transferability of the ASPs because it is relatively stretched [length/(distance between ends) = 33.7/17.1 = 2.0; 21.6/6.7 = 3.2 for loop 1]. Therefore, we studied this loop only partially by applying to it ASPs derived for loop 1. However, a suitable candidate for transferability tests that was treated by us is the 9-residue loop 3 (89–97), Ser^{89}-Ser^{90}-Lys^{91}-Tyr^{92}-Pro^{93}-Asn^{94}-Cys^{95}-Ala^{96}-Tyr^{97}, which is less stretched than loop 2 [length/(distance between ends) = 24.2/8.8 = 2.8]; this loop was not treated by Rapp and Friesner.

Preliminary Results for Loop 1

As a first step, we energy-minimized the X-ray structure of RNase A (1rat.pdb; with hydrogens added by TINKER), using each force field [$E_{\mathrm{FF}}(\epsilon = 1)$] and applying harmonic restraints of 5 kcal/mol/Å^2 to each atomic position; this structure is called the *native optimized structure* (NOS), which deviates from the PDB structure by an all-heavy-atom RMSD of only about 0.14 Å. In the next step, we defined the template for loop 1, which includes any non-loop atom with a distance less than 10 Å from at least one loop atom (in NOS) together with all the other atoms pertaining to the same residue. Therefore, some loop–template distances, as well as loop–loop distances, are larger than 10 Å. For each loop structure, only the intraloop and loop–template interactions are considered, whereas the template–template interactions and the non-template atoms are ignored. This system consists of 108/614/1860 loop/template/protein atoms; a fixed template is necessary for comparing the minimized energies of different loop structures. We found that increasing the template radius from 10 to 13 Å changed the energy by less than 0.4 kcal/mol. The surface area is calculated for both the loop and the template.

TABLE I. Energy Gaps Δ^m_{FF} and $\Delta^m_{\mathrm{tot}}(\sigma^*)$ for Different Force Fields (FF) and Distance-Dependent Dielectric Constants nr (Where n is a Constant)†

$\epsilon = nr$/FF	r	$2r$	$3r$
OPLS	31 (18)	15.0 (8.8)	9.0 (8.3)
CHARMM22		11.2 (NA)	7.5 (5.3)
AMBER	14 (9.3)	6.8 (2.5)	6.3 (3.4)

†The energy gap (kcal/mol) is the difference between the minimized energies of NOS and the lowest energy structure obtained. Δ^m_{FF} is based on the force-field energy, whereas $\Delta^m_{\mathrm{tot}}(\sigma^*)$ (appears in parenthesis) is based on a single optimized ASP (σ^*). NA indicates that the ASP could not be optimized for the CHARMM22 force field for $\epsilon = 2r$.

Our aim is to find an optimal set of ASPs (denoted σ^*_i) and optimal n for which the minimized energy of NOS, $E^{\mathrm{NOS}}_{\mathrm{tot}}(n, \sigma^*_i)$, becomes the GEM, or at least $\Delta^m_{\mathrm{tot}}(n, \sigma^*_i) = E^{\mathrm{NOS}}_{\mathrm{tot}}(n, \sigma^*_i) - E^m_{\mathrm{tot}}(n, \sigma^*_i)$ is minimal and smaller than 2 kcal/mol; m denotes the lowest energy structure. In this context, the cyclic hexapeptide in DMSO studied previously[44] was modeled by the GROMOS 37D4 force field,[75] where all the residues are electrostatically neutral and $\epsilon = 1$. Using the GROMOS energy alone [i.e., $E_{\mathrm{FF}} = E_{\mathrm{GRO}}$, see eq 1), we found for one of the two experimental structures (called βI), $\Delta^m_{\mathrm{GRO}} = 15$ kcal/mol, which was decreased to $\Delta^m_{\mathrm{tot}} = 1.1$ kcal/mol with the optimized ASPs. For values of Δ^m_{FF} greater than 15, the optimized ASPs are not expected to reduce Δ^m_{tot} below the 2 kcal/mol threshold.

To check this aspect for the three force fields, we first calculated $\Delta^m_{\mathrm{FF}}(n)$ for different values of n. Thus, for each force field (and n), we generated with LTD a relatively small sample (several hundred structures) of significantly different energy-minimized structures and calculated $\Delta^m_{\mathrm{FF}}(n)$. The results, which appear in Table I, should be considered lower bounds because the correct GEM structures probably are not included in the corresponding samples. Next, we assigned the same ASP (σ) to all the atoms, optimized it by a procedure described in the next subsection (and Appendix B), and calculated for each case $\Delta^m_{\mathrm{tot}}(n, \sigma^*)$ as well; these results appear in parentheses in Table I. As expected, for each force field $\Delta^m_{\mathrm{FF}}(n)$ decreases as n is increased (because the electrostatic interactions weaken), where for every n the AMBER results are the lowest. However, for AMBER the minimum value of $\Delta^m_{\mathrm{tot}}(n, \sigma^*)$ is obtained for $n = 2$ rather than $n = 3$, and for OPLS the two values are very close [we could not optimize the ASPs ($n = 2$) for the CHARMM22 force field]. This suggests that at least for AMBER, $n = 2$ would be a better choice than both $n = 1$ and $n = 3$, even when the ASPs of all the atoms are considered; that is, $n = 2$ will lead to the lowest $\Delta^m_{\mathrm{tot}}(\sigma^*_i)$. This applies also for OPLS because, using $\epsilon = 3r$, we could not decrease $\Delta^m_{\mathrm{tot}}(\sigma^*_i)$ further by optimizing ASPs for different atoms. Finally, because of the difficulty of optimizing the ASP with CHARMM22 ($n = 2$), we decided to continue testing only the AMBER and OPLS force fields, using from now only $n = 2$.

Optimization Procedure: OPLS Versus AMBER

In the next step, we optimized complete sets of ASPs for loop 1 based on the OPLS and AMBER force fields; the optimization procedure is described in Appendix B. The $\Delta^m_{\mathrm{tot}}(\sigma^*_i)$ values (see Table II) obtained for the final sets of optimal ASPs (σ^*_i) are 6.8 and 2.1 kcal/mol for OPLS and AMBER, respectively, which are larger than the 2 kcal/mol target. However, it should be emphasized again that a protein structure in solution can only be defined up to the corresponding wide microstate, so NOS, used here, is not necessarily the preferred representative of the native wide microstate. In fact, using OPLS we generated several energy-minimized structures pertaining to this wide microstate with minimized energies E_{tot} that were up to 2 kcal/mol lower than the NOS energy; replacing NOS by these structures as references in the optimization process decreased the corresponding values of $\Delta^m_{\mathrm{tot}}(\sigma^*_i)$ while leading to exactly the same set of optimized ASPs. The same situation is expected for AMBER.

The optimized sets of ASPs for the two force fields, which appear in Table II, are discussed later. However, the fact that the energy gaps for loop 1, $\Delta^m_{\mathrm{FF}}(n = 2) = 15.0$ and 6.8 kcal/mol and $\Delta^m_{\mathrm{tot}}(\sigma^*_i) = 6.8$ and 2.1 kcal/mol (Table II), are significantly larger for OPLS than AMBER, respectively, suggests that the AMBER force field is more suitable than OPLS for handling loops. This is also supported by the results for the RMSD from NOS presented in Table III and discussed later. Therefore, loops 2 and 3 were studied only with the AMBER force field.

Optimizing the Positions of Polar Hydrogen

As has been pointed out, for loop 1 we obtained the NOS structure by adding hydrogens to the PDB structure and minimizing the force-field energy E_{FF} with strong harmonic restraints. However, the program TINKER assigns these hydrogens with a prescription that does not optimize their positions with respect to the energy. Thus, an OH vector of a serine side-chain, for example, will move only slightly in the restrained energy minimization, remaining very close to its initial direction, whereas other directions that may lead to lower energy will not be searched. This effect weakens for Asn and Lys, for example, as the number of symmetrically positioned NH groups (which are rotated together) increases to two and three, respectively. Therefore, the minimized energy of NOS (with and without ASPs) is relatively high compared with the low-energy structures generated by LTD, where the OH and NH vectors are rotated and their energy is optimized.

Therefore, one should optimize the positions of the polar hydrogens also for NOS. This optimization is particularly important for loops 2 and 3, which possess seven and four single OH groups, respectively, but less important for loop 1, which has only one such group. Indeed, preliminary application of the optimal ASPs of loop 1 to loops 2 and 3 without optimization of the hydrogens' positions has led to energy gaps $\Delta^m_{\mathrm{tot}}(\sigma^*_i)$ of 7.9 and 8.6 kcal/mol, respectively. This optimization was performed by the rotation of the angles of the OH and NH vectors of the loop and the template within the framework of a Monte Carlo minimization (MCM) procedure (eq A1, Appendix A), as described in Appendix C.

Optimizing the hydrogen network for loop 3 prior to the ASPs optimization indeed led to a low energy gap, $\Delta^m_{\mathrm{tot}}(\sigma^*_i) = 1.1$ kcal/mol (see Table II). To carry out an objective comparison between the sets of ASPs obtained for loops 3 and 1, we also optimized the hydrogen network of loop 1 and its template and reoptimized the ASPs for that loop. However, for loop 2, which was partially studied, only the positions of the polar hydrogens of the loop were optimized.

Results for the ASPs

The various sets of ASPs appear in Table II, where the optimal sets obtained without optimization of the hydrogens' positions are denoted AMBER$^{\mathrm{no}}$ and OPLS$^{\mathrm{no}}$ (no means not optimized), and the results denoted AMBER were obtained after hydrogen optimization. The best-fit set of ASPs based on the optimized sets for loops 1 and 3 appear under the title AMBER$^{\mathrm{bf}}$ (discussed later). For comparison, we also provide the ASPs derived by Wesson and Eisenberg[38] based on the free energy of transfer of small molecules from the gas phase to water. The table reveals that all the ASPs calculated by us are negative (i.e., hydrophilic), meaning that exposed loop structures get lower solvation energy than the less exposed ones. The fact that the rest of the protein is frozen means that the hydrophobic interaction has already been taken into account in the folding process, becoming ineffective for a loop; therefore, the ASP of the carbon groups, like those of the hydrophilic atoms, is also negative, in contrast to the positive value obtained by Wesson and Eisenberg. In other words, with E_{FF} the loop structure with lowest energy is relatively unexposed, collapsing on the template; however, for the optimal negative ASPs, the minimal $E_{\mathrm{tot}}(\sigma^*_i)$ is obtained for a loop structure that, like NOS, is relatively exposed to the solvent.

The OPLS$^{\mathrm{no}}$ and AMBER$^{\mathrm{no}}$ results for loop 1 (obtained without optimization of the hydrogen network) show that the ASPs, as expected, are force-field-dependent, which is mainly demonstrated by the values of σ^*_{H} and σ^*_{O}. Also, because for both loops 1 and 3 only a few atoms of S (one belongs to the loop and the others to the template) are involved in the optimization, their effect is very small, and their ASP could not be determined; the values in the table

TABLE II. Optimal and Best-Fit (bf) ASPs [cal/(mol · Å^2)] for the OPLS and AMBER Force Fields for Loops 1 and 3†

Atom/Method	C	N	O	H	S	One ASP	ΔE^m_{FF}	ΔE^m_{tot}
				Loop 1				
OPLS^{no}	−80	−100	−120	−240	−110	−110	15.0	6.8
AMBER^{no}	−80	−110	−70	−120	−100	−90	6.8	2.1
AMBER	−80	−100	−60	−120	−90	−90	6.1	1.9
AMBER^{bf}	−80	−120	−72	−125	−90		6.1	2.1
				Loop 3				
AMBER	−80	−180	−85	−130	−90	−80	11.5	1.1
AMBER^{bf}	−80	−120	−72	−125	−90		11.5	1.9
WE	+12	−120	−120		−20			
WE*		−180	−190					

†The results are based on the distance-dependent dielectric constant $\epsilon = 2r$. The ASPs for AMBER^{no} and OPLS^{no} were obtained without optimization of the positions of the hydrogens of the template and loop of NOS. The best-fit ASPs are based on the individual optimal sets of ASPs (AMBER) obtained for loops 1 and 3. The ASP values used for S were not optimized. ΔE^m_{FF} and ΔE^m_{tot} (kcal/mol) are the energy gaps without and with ASPs, respectively. WE are ASPs obtained by Wesson and Eisenberg;[38] the values under WE* are for the charged atoms.

are those used in the calculations. We also provide results for a single ASP defined for all the atoms. Trying to assign different ASPs to loop and template atoms, to side-chain and backbone atoms, and to highly charged atoms did not lead to a further decrease in $\Delta^m_{tot}(\sigma^*_i)$. This is an important result, suggesting that the ASPs are not very specific, so optimizing ASPs for different amino acids is not necessary.

For loop 1, which possesses only a single OH group, optimizing the positions of the polar hydrogens, as expected, affected the results only slightly, where Δ^m_{FF} and ΔE^m_{tot} decreased (i.e., improved) from 6.8 to 6.1 and 2.1 to 1.9 kcal/mol in going from the original set of ASPs (AMBER^{no}) to the new set (AMBER) obtained with the optimization of the hydrogen positions. However, despite the latter optimization and the fact that the samples used for obtaining ASPs (AMBER) are larger than those used for ASPs (AMBER^{no}) (~300 vs 150, respectively), the two sets of ASPs are very similar. This indicates that the ASPs are not very sensitive to the specific loop and its environment and might, therefore, be transferable to other loops. The same conclusion can be drawn from the similarity between the optimal ASPs (AMBER) of loops 1 and 3, although these loops have different sizes, templates, and numbers of polar hydrogens (134/595 atoms for loop 3/template 3); in particular, the ASP of C, which is the most effective atom (because carbon constitutes the majority) is equal for the two sets. A significant difference is observed only for σ^*_N, which is −100 for loop 1 and −180 cal/mol/Å^2 for loop 3. However, for loop 3 σ^*_N affects the results only weakly; that is, the decrease in the energy gap, ΔE^m_{tot}, in going from $\sigma^*_N = -100$ to −180 cal/mol/Å^2 is relatively small.

One does not expect sets of ASPs derived for loops of different sizes, sequences, and templates to be exactly the same. However, if the difference is not significant, a best-fit set can be obtained from the individual sets, which performs reasonably well (even though not optimally) for each loop and is thus expected to be transferable also to other loops. We, therefore, devised a best-fit (bf) set with ASPs that (besides that of N, discussed later) are averages of the optimal ASPs (AMBER) of loops 1 and 3 (see Table II). Two LTD runs (~3000 minimizations) for loops 1 and 3 based on this best-fit set led to a slight increase in the energy gaps for both loops but to RMSD values that are almost unchanged compared with those obtained for the optimal sets (see Table III). Because of the weak effect of σ^*_N on the energy of loop 3, the best-fit value, $\sigma^*_N = -180$, could be replaced by $\sigma^*_N = -120$ cal/mol/Å^2, which increased the energy gap from 1.9 to 2.1 and 1.1 to 1.9 kcal/mol for loops 1 and 3, respectively.

Prediction of Loop Structures

To check the ability of our optimized solvation models to predict the correct loop structures, we identified the significantly different energy-minimized structures t within the 2 kcal/mol range above the GEM, calculated for each the RMSD$_t$ from NOS, based on the heavy atoms and without superposition on NOS. These values are averaged according to their Boltzmann probabilities [based on the minimized energies $E^t_{tot}(\sigma^*_i)$]:

$$\mathrm{RMSD} = \sum_t \mathrm{RMSD}_t \exp - [E^t_{tot}/k_BT] / \sum_t \exp - [E^t_{tot}/k_BT] \qquad (2)$$

where $T = 300$ K. The RMSD results presented in Table III were calculated for the heavy atoms of the backbone (BB), side-chains (SC), and the entire loop (TOT). For each case, results were calculated for E_{FF} ($\epsilon = 2r$) and E_{tot} ($\epsilon = 2r$) (eq 1), and the number of structures averaged is provided as well. The best-fit results are based on structures generated by two LTD runs applied to loops 1 and 3 with the best-fit ASPs of Table II. In each case, results are also presented for the lowest energy structure.

The table reveals that for loops 1 and 3 the RMSD values obtained with the optimal and best-fit ASPs are significantly better (i.e., smaller) than those obtained with ASPs = 0, meaning that our energy-based optimization is effective. Also, for loop 1 the RMSD results obtained without optimization of the coordinates of the hydrogens of NOS are significantly better for AMBER^{no} than for

OPLS^{no}, suggesting, as pointed out earlier, that for loops the AMBER force field performs better than OPLS. Therefore, loops 2 and 3 were studied only with the AMBER force field. The reason for this difference in performance is not clear. We minimized several structures with both potentials (i.e., using E_{FF}) and found a significant difference only in the torsional energy, which indeed is defined differently in these force fields with a single term in AMBER but with several terms in OPLS.

The NMR study of Santoro et al.[68] and X-ray crystallography studies[79,80] have found some of the side-chains to be disordered or to populate multiple conformations. To this category belong Lys^{66} of loop 1; Ser^{15}, Ser^{18}, and ASN^{24} of loop 2; and Lys^{91} and Asn^{94} of loop 3. These side-chains populated multiple conformations also in our low-energy structures; therefore, a more realistic measure of performance would be an RMSD of the side-chains (and the related total RMSD) calculated by the omission of the contribution of the disordered ones; these results appear with an asterisk in Table III. Multiple side-chain conformations typically occur for residues exposed to the solvent, and this intermediate flexibility is sensitive to crystal packing, temperature, mutations, and other parameters. Thus, in the high-resolution crystal structure (0.87 Å at 100 K) of RNase,[81] where Asn^{67} was replaced by an isoaspartyl residue, 15% of the side-chains (19) populated multiple conformers; however, only a few of them are specified, including Ser^{90} of loop 3, which was found in the NMR study to have a well-defined structure. Esposito et al.[81] argued that "atomic resolution can reveal, in some cases, degree of side-chain flexibility that is not found in NMR solution studies".

Structural Results for Loop 1

Let us discuss first the results for loop 1. The Boltzmann-averaged RMSDs for the backbone, 0.50, 0.78, and 0.38 Å for AMBER^{no}, AMBER, and AMBER^{bf}, respectively, are very good. We checked the backbone dihedral angles to find that actually all the ϕ and ψ are located within 60° of the corresponding NOS values, meaning that these structures belong to the same backbone wide microstate of NOS. An insignificant deviation from this picture occurs for the AMBER^{no} and AMBER results, where for a single structure (with a small Boltzmann probability of 1%) ϕ and ψ, and ϕ, respectively, of one residue deviate from the NOS value by about 70° (i.e., only slightly beyond our 60° criterion).

The Boltzmann-averaged RMSD values for the side-chains without Lys^{66}, 0.84, 1.28, and 1.46 Å for AMBER^{no}, AMBER, and AMBER^{bf}, respectively, are very satisfactory because of the difficulty in handling the side-chains. The corresponding total RMSD values, 0.66, 1.04, and 0.96 Å, are very good as well. We also checked the deviation of the side-chain dihedrals for these three sets of ASPs. In the 9 lowest energy structures of AMBER^{no} (within 1.55 kcal/mol above the GEM, which contribute 95% of the Boltzmann probability), all the χ values (excluding Lys^{66}) pertain to the same wide microstate of NOS; the only exception is χ^3 of Gln^{69}, which in 6 out of the 9 structures deviates from the NOS value by about 80° (i.e., only slightly beyond our 60° limit). For the 5 structures with the highest energy (1.55–1.99 kcal/mol; 5%), the χ values of Gln^{69} and Thr^{70} deviate significantly from the corresponding NOS values. These results should be considered exceptionally good. For each of the other sets of ASPs (AMBER and AMBER^{bf}), the number of structures is much larger than 14 (40 and 31), and the number of deviations increases. Thus, for the AMBER results significant deviations occur for Gln^{69} in 18/40 structures, Asn^{67} (5/40), Thr^{70} (6/40), and Asn^{71} (2/40). For AMBER^{bf}, 18/31 deviations occur for Gln^{69}, 7/31 occur for Asn^{67}, and 2/31 occur for Thr^{70}. The deviations occurred for Asn^{67} are not totally unexpected because this side-chain has been found in the X-ray crystallography studies (but not in the NMR study) to populate multiple conformers.

Finally, the TOT* results for loop 1 for the lowest energy structure are always smaller than their Boltzmann-averaged counterparts, suggesting that the RMSD is correlated with the energy. This correlation is demonstrated in Figure 1, where the RMSD (TOT*) is plotted against the energy for structures generated with the optimal ASPs (AMBER) of loop 1. The figure also shows that the GEM structure does not have the lowest RMSD and that the structures with energy within the 2 kcal/mol range above the GEM populate two regions of low and higher RMSD values, around 0.5 and 1.3 Å, respectively. However, the fact that low-energy structures pertain to the latter group is not necessarily an indication for the imperfection of E_{tot}; for example, the four structures with lowest minimized energy all belong to the wide microstate of NOS, whereas two of them have large RMSDs of about 1.3 Å. Thus, this RMSD just reflects the relatively large conformational space defined by the wide microstate of NOS. A similar figure (not shown) has also been obtained for the best-fit results of loop 1.

Structural Results for Loop 3

For loop 3, the Boltzmann-averaged RMSD results for the backbone, 0.16, and 0.25 Å for AMBER and AMBER^{bf}, respectively, are excellent. For the AMBER results (based on 132 structures), none of the backbone dihedral angles violates the 60° criterion, whereas for AMBER^{bf}, in two structures the same four dihedrals deviate from the NOS values, two of them only slightly by about 70°, whereas the other two deviate significantly. The RMSD values for the side-chains without Lys^{91} and Asn^{94} are excellent as well, 0.78 and 0.88 Å, where the TOT* values are 0.55 and 0.63 Å, respectively. For both sets of ASPs, significant deviations of side-chain dihedral angles (besides Lys^{91} and Asn^{94}) occur only for Ser^{89} and Cys^{95}; again, these are very good results. For this loop, the TOT* result of the lowest energy structure obtained for the optimal ASPs (AMBER) (but not for AMBER^{bf}) is smaller than the related Boltzmann-averaged value. Figure 2 shows that for loop 3 the structures within 3 kcal/mol above the GEM have an RMSD (TOT*) smaller than 1 Å, whereas most of the higher energy structures are of RMSD (TOT*) >2 Å. A similar figure has been obtained for the best-fit results.

TABLE III. RMSD From NOS (Å) of Low-Energy Structures of Loops 1–3†

	ASPs ≠ 0						ASPs = 0					
	# str.	BB	SC	TOT	SC*	TOT*	# str.	BB	SC	TOT	SC*	TOT*
						Loop 1						
OPLS^{no}	10	0.94	2.23	1.64	2.13	1.51	5	2.00	2.49	2.23	1.95	1.98
	1	0.71	1.74	1.26	1.72	1.19	1	1.99	2.45	2.20	1.91	1.96
AMBER^{no}	14	0.50	1.49	1.06	0.84	0.66	8	1.61	2.42	2.01	1.73	1.66
	1	0.52	1.04	0.78	0.77	0.62	1	1.76	2.47	2.10	1.90	1.81
AMBER	40	0.78	1.95	1.45	1.28	1.04	5	1.60	2.38	1.97	1.73	1.65
	1	0.27	1.99	1.32	0.97	0.65	1	1.61	2.36	1.97	1.71	1.68
AMBER^{bf}	31	0.38	2.12	1.43	1.46	0.96						
	1	0.27	1.99	1.32	0.97	0.65						
						Loop 3						
AMBER	132	0.16	1.34	0.95	0.78	0.55	7	1.16	3.22	2.40	1.43	1.22
	1	0.20	1.86	1.31	0.48	0.35	1	1.21	3.31	2.48	1.54	1.30
AMBER^{bf}	72	0.25	1.31	0.94	0.88	0.63						
	1	0.16	1.27	0.90	1.25	0.87						
						Loop 2						
AMBER^{no}	33	0.58	1.01	0.77	0.93	0.77	30	0.60	1.07	0.81	1.03	0.83
	1	0.62	0.88	0.73	0.74	0.72	1	0.59	0.93	0.74	0.84	0.73

†Results are given for the heavy atoms of the backbone (BB), the side-chains (SC), and all the atoms (TOT). For each force field (FF), the first row presents the Boltzmann averages of the low-energy structures (their number is provided under # str.) found within the 2 kcal/mol range above the lowest energy structure (eq 2); the RMSD results for the lowest energy structure appear in the second row. Results with an asterisk were calculated without the side-chain atoms of Lys^{66} for loop 1; Ser^{15}, Ser^{18}, and Asn^{24} for loop 2; and Lys^{91} and Asn^{94} for loop 3 (see text). Results denoted *no* (not optimized) were obtained with the set of ASPs obtained for loop 1 without optimization of the hydrogen positions. These ASPs were used for loop 2, where only the positions of the hydrogens of the loop (but not the template) were optimized.

Notice that the larger RMSD values obtained in Figure 2 than in Figure 1 (even though loop 3 is more stretched than loop 1) stem from the rigidity of loop 1 due to a stabilizing network of hydrogen bonds.[81]

Structural Results for Loop 2

Chronologically, after calculating the ASPs for loop 1 without hydrogen optimization (AMBER^{no}), we applied them also to loop 2 to test their transferability. However, because of the relatively large number of polar groups in loop 2 (specified earlier), the energy gaps were high (8.8 and 7.9 kcal/mol for Δ^m_{FF} and Δ^m_{tot}, respectively), which led us to the conclusion that the positions of the polar hydrogens of NOS should be optimized; still, we applied this optimization only to the hydrogens of the loop, whereas those of the template remained intact. Then, LTD runs were carried out for E_{FF} ($\epsilon = 2r$) (~8000 minimizations) and E_{tot} ($\epsilon = 2r$) (~6000 minimizations) with the optimal ASPs (AMBER^{no}); that is, no reoptimization of ASPs was performed for loop 13–24. We obtained $\Delta^m_{FF} = 4.5$ kcal/mol and $\Delta^m_{tot} = 3.1$ kcal/mol, which are still relatively high,

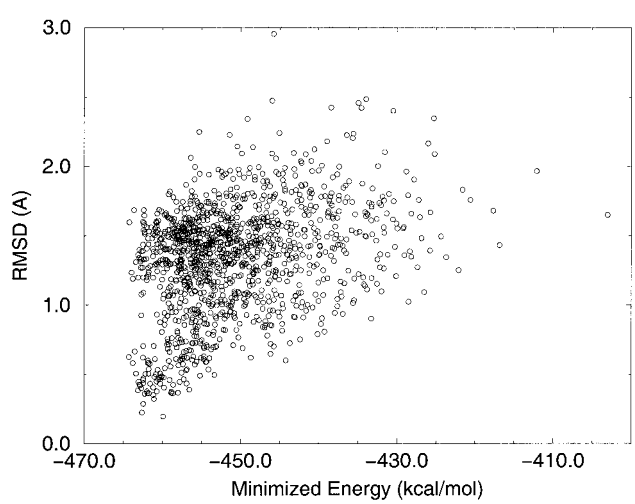

Fig. 1. RMSD (TOT*) from NOS (Å) versus the minimized energy $E_{tot}(\sigma_i^*)$ of structures generated by LTD with the optimal ASPs (AMBER) (σ_i^*) obtained for loop 1 (64–71). The RMSD was calculated for all the heavy atoms of the loop besides the side-chain atoms of Lys^{66} (see text).

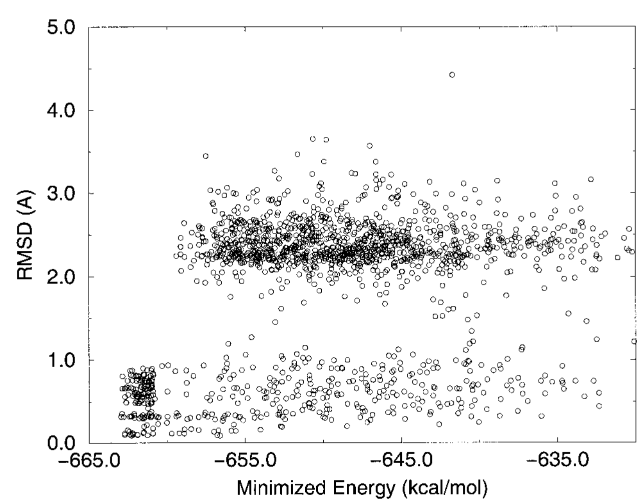

Fig. 2. RMSD (TOT*) from NOS (Å) versus the minimized energy $E_{tot}(\sigma_i^*)$ of structures generated by LTD with the optimal ASPs (AMBER) (σ_i^*) obtained for loop 3 (89–97). The RMSD was calculated for all the heavy atoms of the loop besides the side-chain atoms of Lys^{91} and Asn^{94} (see text).

probably reflecting our incomplete hydrogen optimization. The results presented in Table III for loop 2 are very good for both potentials, where RMSD (TOT*) for E_{tot} is only slightly smaller than for E_{FF}, 0.77 versus 0.83 Å, respectively. Further analysis has shown that all the backbone dihedral angles satisfy the 60° criterion, besides significant deviations of ψ (Ala[20]) and ϕ (Ser[21]), which occur in 2/30 and 25/33 structures obtained with E_{FF} and E_{tot}, respectively.

The NMR study of Santoro et al.[68] found Ser[16], Ser[18], and Asn[24] to be disordered, correspondingly, the first two residues exhibit double occupancy of side-chain conformers in our samples for both potentials: for Asn[24], χ^1 and χ^2 deviate from the NOS values by 77 and 90°, respectively. We also find double occupancy of χ angles for Asp[14], Ser[21], Ser[22], and Ser[23], but the conformational change of Asp[14] is small, involving only χ^2. Therefore, only 3 side-chains out of 12 should actually be considered as deviating significantly from the corresponding side-chain conformers of NOS. The close RMSD values obtained for E_{FF} and E_{tot} for loop 2 (but not for loops 1 and 3) stem from the fact that loop 2 is more stretched than these loops, as discussed earlier. Indeed, trying to shake loop 2 by increasing the temperature parameter in the LTD procedure to $T^* =$ 2000K resulted in RMSD values not larger than 2 Å. Because of this insensitivity of loop 2 to the effect of the ASPs, we did not attempt to study it further; however, the previous discussion of this loop is important for emphasizing the effect of loop stretchability on the RMSD results, a point ignored in most loop studies in the literature. Also, the good results obtained for the side-chains of loop 2 demonstrate the quality of the AMBER force field. Finally, loop 2 was studied by Rapp and Friesner,[40] and it is interesting to compare our results to theirs.

Rapp and Friesner[40] studied loops 1 and 2 with the same AMBER force field combined with the generalized Born (GB/SA) solvation model.[39] For the backbone heavy atoms of the lowest energy structure of loop 1, they obtained RMSD = 1.46 Å, in comparison with our lower values of 0.27 and 0.52 Å for AMBER and AMBER^{no}, respectively. The corresponding results for loop 2 are 0.8 (Ref. 40) and 0.62 Å (our result). However, a fair comparison between the two solvation models is difficult because Rapp and Friesner used a flexible template bound by a region of atoms locally restrained by harmonic potentials. Also, our conformational search runs are 1–1.5 orders of magnitude more extensive than those of Rapp and Friesner.

SUMMARY

The present energetic approach based on a conformational search with LTD is convenient because the entire loop (i.e., backbone and side-chains) is treated at once, in contrast to other methods discussed in the introduction. We showed that the optimization of ASPs for loops should rely on known loop structures in the protein environment rather than on thermodynamic data of small molecules. In particular, with the latter derivation carbon is a hydrophobic atom with a positive ASP, whereas our optimization for loops leads to a negative ASP of the carbon groups. This hydrophilicity of carbon has wider implications, suggesting, for example, that simulation of the entire folded protein based on E_{tot} with positive ASP for carbon might be inadequate.

These results demonstrate that the ASPs are force-field-dependent; therefore, the common practice of using the same set of ASPs with different force fields is unwarranted. We found that the all-atom AMBER force field performs better for loops than the all-atom OPLS and CHARMM22 force fields.

Most importantly, performing independent optimizations (with the AMBER force field and $\epsilon = 2r$) for loops 1 and 3 led to very similar sets of ASPs. The fact that these loops differ significantly in size (108 vs 134 atoms), number of polar groups, and templates suggests that similar sets would be obtained for loops in other proteins as well; one can then devise a best-fit set of ASPs based on the individual sets that performs reasonably well for each loop. Indeed, our individual optimal sets and the best-fit set based on them all led to very good RMSD results, not only for the backbone but also for the side-chains. These results are based on a stricter than usual analysis, which considers, in addition to the commonly used RMSD criterion, the deviation of dihedral angles from the wide microstate of NOS. Moreover, the conformational search runs performed here are significantly more extensive than those carried out in previous studies of loops.

The fact that an ASP is not sensitive to the partial charge and the atom location (on the side-chain, backbone, loop, or the template) makes the derivation of ASPs feasible because treating these specific cases is not necessary. In other words, the ASPs perturb the force-field interactions in a mean-field manner, increasing the preference of the exposed conformations. Therefore, further improvements in this solvation model are expected with the advent of better force fields, which, however, will require reoptimizing the ASPs.

Our analysis showed that all the loops studied populate a single backbone wide microstate, whereas intermediate flexibility was demonstrated only for specific side-chains, where multiple occupancy of different conformers was observed both experimentally and in our results. In the future, we shall apply the entire methodology, based on the best-fit set of ASPs, mainly to long loops of antibodies and enzymes whose flexibility is essential for recognition and catalysis processes. We shall also reoptimize ASPs for specific loops to verify the transferability of our best-fit set or to improve it. Optimizing the ASPs is a time-consuming serial process that cannot benefit from parallel computing; however, the application of the methodology, which is based on conformational search and Monte Carlo runs with a given set ASPs, can be carried out in parallel.

We also intend to treat problems for structure prediction addressed by CASP 5, where the protein's template will be determined by homology modeling and the loop structures will be determined by our procedures. Obviously, the accuracy of the loop structures depends on the quality of the template; the extent of this dependence will be studied.

Finally, our optimization procedure for loops is not limited to E_{tot} (eq 1) but can, in practice, be applied to several of the relatively large number of parameters defining other solvation models, such as the GB/SA model[39,82] or the model suggested recently by Lazaridis and Karplus,[83] our present and future calculations will provide benchmark results for comparison with results obtained with these models.

REFERENCES

1. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. Naturwissenschaften 1985;72:212–213.
2. Getzoff ED, Geysen HM, Rodda SJ, Alexander H, Tainer JA, Lerner RA. Mechanisms of antibody binding to a protein. Science 1987;235:1191–1196.
3. Rini JM, Schulze-Gahmen U, Wilson IA. Structural evidence for induced fit as a mechanism for antibody–antigen recognition. Science 1992;255:959–965.
4. Constantine KL, Friedrichs MS, Wittekind M, Jamil H, Chu C-H, Parker RA, Goldfarb B, Muller L, Farmer II T. Backbone and side chain dynamics of uncomplexed human adipocyte and muscle fatty acid-binding proteins. Biochemistry 1998;37:7965–7980.
5. Nicholson LK, Yamazaki T, Torchia DA, Grzesiek S, Bax A, Stahl SJ, Kaufman JD, Wingfield PT, Lam PYS, Jadhav PK, Hodge CN, Domaille PJ, Chang C-H. Flexibility and function in HIV-1 protease. Struct Biol 1995;2:274–280.
6. Collins JR, Burt SK, Erickson JW. Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. Struct Biol 1995;2:334–338.
7. Wagner G. The importance of being floppy. Struct Biol 1995;2:255–257.
8. Fetrow JS. Omega loops: nonregular secondary structures significant in protein function and stability. FASEB J 1995;9:708–717.
9. Leszcynski JF, Rose GD. Loops in globular proteins: a novel category of secondary structure. Science 1986;234:849–855.
10. Ring CS, Kneller DG, Langridge R, Cohen FE. Taxonomy and conformational analysis of loops in proteins. J Mol Biol 1992;224:685–699.
11. Fechteler T, Dengler U, Schomburg D. Prediction of protein three dimensional structures in insertion and deletion regions: a procedure for searching data bases of representative protein fragments using geometric scoring criteria. J Mol Biol 1995;253:114–131.
12. Donate LE, Rufino SD, Canard LHJ, Blundell TL. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a data base for modeling and prediction. Protein Sci 1996;5:2600–2616.
13. Oliva B, Bates PA, Querol E, Avilés FX, Sternberg JE. An automated classification of the structure of protein loops. J Mol Biol 1997;266:814–830.
14. Kwasigroch J-M, Chomilier J, Mormon J-P. A global taxonomy of loops in globular proteins. J Mol Biol 1996;259:855–872.
15. Martin AC, Toda K, Stirk HJ, Thornton JM. Long loops in proteins. Protein Eng 1995;8:1093–1101.
16. Mosimann S, Meleshko R, James MNG. A critical assessment of comparative molecular modeling of tertiary structures of proteins. Proteins 1995;23:301–317.
17. Ŝali A. Modeling mutations and homologous proteins. Curr Opin Struct Biol 1995;6:437–451.
18. Bates PA, Sternberg MJE. Model building by comparison at CASP3: using expert knowledge and computer automation. Proteins 1999;Suppl 3:47–54.
19. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 1987;196:901–917.
20. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff F, Padlan AA, Davies D, Tulip WR, Colman PM, Spinelli S, Alzari PM, Poljak RJ. Conformations of immunoglobulin hypervariable regions. Nature 1989;342:877–883.
21. Summers NL, Karplus M. Modeling of globular proteins. A distance-based data search procedure for the construction and insertion/deletion regions and Pro $\leftrightarrow$ non-Pro mutations. J Mol Biol 1990;216:991–1016.
22. Sudarsanam S, Dubose RF, March CJ, Srinivasan S. Modelling protein loops using a ϕ_{i+1}, ψ dimer data base. Protein Sci 1995;4:1412–1420.
23. Fidelis K, Stern PS, Bacon D, Moult J. Comparison of systematic search and data base methods for constructing segments of protein structure. Protein Eng 1994;7:953–960.
24. Wojcik J, Mormon J-P, Chomilier J. New efficient sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. J Mol Biol 1999;289:1469–1490.
25. Moult J, James MNG. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1986;1:146–163.
26. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 1987;26:137–168.
27. van Vlijmen HWT, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267:975–1001.
28. Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. Predicting antibody hypervariable loop conformations II: minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. Proteins 1986;1:342–362.
29. Shenkin PS, Yarmush DL, Fine RM, Wang H, Levinthal C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 1987;26:2053–2085.
30. Gō N, Scheraga HA. Ring closure and local conformational deformations of chain molecules. Macromolecules 1970;3:178–187.
31. Dudek MJ, Scheraga HA. Protein structure prediction using a combination of sequence homology and global energy minimization I. Global energy minimization of surface loops. J Comput Chem 1990;11:121–151.
32. Carlacci L, Englander SW. Loop problem in proteins: developments on the Monte Carlo simulated annealing approach. J Comput Chem 1996;17:1002–1012.
33. Higo J, Collura V, Garnier J. Development of extended simulated annealing method: application to the modeling of complementary determining regions of immunoglobulins. Biopolymers 1992;32:33–43.
34. Zheng Q, Rosenfeld R, Vajda S, DeLisi C. Loop closure via bond scaling and relaxation. J Comput Chem 1993;14:556–565.
35. Rosenfeld R, Zheng Q, Vajda S, DeLisi C. Computing the structure of bound peptides: application to antigen recognition by class I MHCs. J Mol Biol 1993;234:515–520.
36. Mas MT, Smith KC, Yarmush DL, Aisaka K, Fine RM. Modeling of anti-CEA antibody combining site by homology and conformational search. Proteins 1992;14:483–498.
37. Smith KC, Honig B. Evaluation of conformational free energies of loops in proteins. Proteins 1994;18:119–132.
38. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1:227–235.
39. Qiu D, Shenkin PS, Hollinger FP, Still WC. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. J Phys Chem 1997;101:3005–3014.
40. Rapp CS, Friesner RA. Prediction of loop geometries using generalized Born model of solvation effects. Proteins 1999;35:173–183.
41. Bruccoleri RE, Haber E, Novotný J. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. Nature 1988;335:564–568.
42. Meirovitch H, Meirovitch E, Lee J. New theoretical methodology for elucidating the solution structure of peptides from NMR data. I. The relative contribution of low energy microstates to the partition function. J Phys Chem 1995;99:4847–4854.
43. Meirovitch H, Meirovitch E. New theoretical methodology for elucidating the solution structure of peptides from NMR data. III. Solvation effects. J Phys Chem 1996;100:5123–5133.
44. Baysal C, Meirovitch H. Determination of the stable microstates of a peptide from NOE distance constraints and optimization of atomic solvation parameters. J Am Chem Soc 1998;120:800–812.
45. Baysal C, Meirovitch H. Populations of interconverting microstates of a cyclic peptide that are based on free energy simulations lead to experimental NMR data. Biopolymers 1999;50:329–344.
46. Baysal C, Meirovitch H. *Ab initio* prediction of the solution structures and populations of a cyclic pentapeptide in DMSO based on an implicit solvation model. Biopolymers 2000;53:423–433.

47. Baysal C, Meirovitch H. On the transferability of atomic solvation parameters. *Ab initio* structural prediction of cyclic heptapeptides in DMSO. Biopolymers 2000;54:416–428.
48. Baysal C, Meirovitch H. Efficiency of the local torsional deformations method for identifying the stable structures of cyclic molecules. J Phys Chem 1997;101:2185–2191.
49. Meirovitch H. Calculation of entropy with computer simulation methods. Chem Phys Lett 1977;45:389–392.
50. Meirovitch H, Koerber SC, Rivier J, Hagler AT. Computer simulation of the free energy of peptides with the local states method: analogues of gonadotropin releasing hormone in the random coil and stable states. Biopolymers 1994;4:815–839.
51. Karplus M, Kushick JN. Method for estimating the configurational entropy of macromolecules. Macromolecules 1981;14:325–332.
52. Stillinger FH, Weber TA. Packing structures and transitions in liquids and solids. Science 1984;225:983–989.
53. Elber R, Karplus M. Multiple conformational states of proteins—a molecular dynamics analysis of myoblobin. Science 1987;235:318–321.
54. Baysal C, Meirovitch H. Efficiency of simulated annealing for peptides with increasing geometrical restraints. J Comput Chem 1999;20:1659–1670.
55. Kessler H, Matter H, Gemmecker G, Kottenhahn M, Bats JW. Structure and dynamics of a synthetic o-glycosylated cyclopeptide in solution determined by NMR spectroscopy. J Am Chem Soc 1992;114:4805–4818.
56. Beveridge DL, DiCapua FM. Free energy via molecular simulation: applications to chemical and biomolecular systems. Annu Rev Biophys Biophys Chem 1989;18:431–492.
57. Meirovitch H. Calculation of the free energy and entropy of macromolecular systems by computer simulation. In: Lipkowitz KB, Boyd DB, editors. Reviews in computational chemistry. New York: Wiley-VCH; 1998. Vol. 12, p 1–74.
58. Tanner JJ, Nell LJ, McCammon JA. Anti-insulin antibody structure and conformation. II. Molecular dynamics with explicit solvent. Biopolymers 1992;32:23–31.
59. Gilson MK, Honig B. The inclusion of electrostatic hydration energies in molecular mechanics calculations. J Comput-Aided Mol Des 1991;5:5–20.
60. Ooi T, Oobatake M, Némethy G, Scheraga HA. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci USA 1987;84:3086–3090.
61. Schiffer CA, Caldwell JW, Kollman PA, Stroud RM. Protein structure prediction with a combined solvation free energy–molecular mechanics force field. Mol Simul 1993;10:121–149.
62. Fraternali F, van Gunsteren WF. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. J Mol Biol 1996;256:939–948.
63. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci USA 1987;84:6611–6615.
64. Von Freyberg B, Braun W. Efficient search for all low energy conformations of polypeptides by Monte Carlo methods. Comput Chem 1991;12:1065–1076.
65. Meirovitch H, Vásquez M. Efficiency of simulated annealing and the Monte Carlo minimization method for generating a set of low energy structures of peptides. J Mol Struct (Theochem) 1997;398–399:517–522.
66. Baysal C, Meirovitch H. Efficiency of simulated annealing for peptides with increasing geometrical restraints. J Comput Chem 1999;20:1659–1670.
67. Tilton RF Jr, Dewan JC, Petsko GA. Effects of temperature on protein structure and dynamics: X-ray crystallography of the protein ribonuclease-A at nine different temperatures from 98 to 320 K. Biochemistry 1992;31:2469–2481.
68. Santoro J, González C, Bruix M, Neira JL, Nieto JL, Herranz J, Rico M. High-resolution three dimensional structure of ribonuclease A in solution by nuclear magnetic resonance spectroscopy. J Mol Biol 1993;229:722–734.
69. Ponder JW. TINKER—software tools for molecular design, Version 3.7. St. Louis: Washington University; 1999.
70. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
71. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
72. MacKerell AD, Jr., Bashford D, Bellott M, Dunbrack RL, Jr., Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
73. Perrot G, Cheng B, Gibson KD, Palmer KA, Nayeem A, Maigret B, Scheraga HA. MSEED: a program for the rapid analytical determination of accessible surface areas and their derivatives. J Comput Chem 1992;13:1–11.
74. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Technical Report NAM03; Evanston, IL: Department of Electrical Engineering and Computer Science, Northwestern University; 1988.
75. van Gunsteren WF, Berendsen HJC. Groningen Molecular Simulations (GROMOS) library manual; Biomos, Nijenborgh 16 9747 AG, Groningen NL; 1987.
76. Brünger AT, Karplus M. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. Proteins 1988;4:148–156.
77. Bass MB, Hopkins DF, Andrew W, Jaquysh N, Ornstein RL. A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. Proteins 1992;12:266–277.
78. Glick M, Goldblum A. A novel energy-based stochastic method for positioning polar protons in protein structures from X-rays. Proteins 2000;38:273–287.
79. Svenson LA, Sjölin L, Dill J, Gilliland GL. The conformation flexibility of surface residues of bovine pancreatic ribonuclease A at 1.1A. In: Cuchillo CM, de Llorens R, Nogués MV, & Parés X, editors, Structure, mechanism and function of ribonucleases, Proceedings of the 2nd International Meeting, Universitat Autonoma de Barcelona, Barcelona. p 31–38.
80. Kuriyan J, Ösapay K, Burley SK, Brünger AT, Hendrickson WA, Karplus M. Exploration of disorder in protein structures by X-ray restrained molecular dynamics. Proteins 1991;10:340–358.
81. Esposito L, Vitagliano L, Sica F, Sorrentino G, Zagari A, Mazzarella L. The ultrahigh resolution crystal structure of ribonuclease A containing an isoaspartyl residue: hydration and stereochemical analysis. J Mol Biol 2000;297:713–732.
82. Haukins GD, Liotard DA, Cramer CJ, Truhlar DG. OMNISOL: fast prediction of free energies of solvation and partition coefficients. J Org Chem 1998;63:4305–4313.
83. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.

APPENDIX A: THE LTD METHOD

The LTD[44,48] method is a conformational search procedure for cyclic molecules and protein loops modeled by a force field with flexible bond lengths and angles. An LTD simulation starts from an arbitrary energy minimized loop structure, i, with energy E_i^0; i is then distorted by a single or several local torsional rotations along the chain followed by energy minimization. The resulting conformation j (with E_j^0) is accepted according to the Metropolis transition probability, p_{ij};

$$p_{ij} = \min(1, \exp[-(E_j^0 - E_i^0)/k_B T^*]) \qquad \text{(A1)}$$

where the accepted structure is deformed again and the process continues. In contrast to a usual Metropolis procedure, the generated conformations are not distributed according to the Boltzmann probability because the minimized energies (rather than the energies themselves) appear in eq A1. However, this MCM procedure[63] is a selection procedure that efficiently directs the search toward the low-energy region in the conformational space. Therefore, T^* is not a usual temperature but a parameter that affects the efficiency of the process.[64] In most of our runs, T^* was changed every 50 MCM steps by 10 K from 200 to 1000 K and vice versa. The coordinates and energies of all the energy-minimized structures, including those which were rejected through eq A1, were stored in a file for further analysis.[44]

With a local rotation around the bond connecting atoms $k + 1$ and $k + 2$, only atom $k + 3$ is moved (and the entire side-chain connected to it); that is, only the bond connecting $k + 3$ and $k + 4$ is disrupted, whereas the rest of the molecule is unaffected. Typically, in each LTD step several independent but significant local rotations (determined randomly) are carried out along the chain; therefore, energy barriers are crossed efficiently. These local conformational changes are especially important in a dense protein environment to reduce the chance for creating undesired loop–template entanglements. Notice that together with the backbone angles, side-chain dihedrals are randomly selected as well, and they are changed at random but not locally. Thus, the whole loop is treated at once, in contrast to procedures used by others and discussed in the introduction. Our implementation of LTD is exactly the same as that applied to the cyclic hexapeptide described in detail in Ref. 44. LTD (MCM) is significantly more efficient than simulated annealing.[65,66]

APPENDIX B: THE OPTIMIZATION PROCEDURE FOR THE ASPS

We describe here the optimization procedure as applied initially to loop 1 with the OPLS and AMBER force fields. The same procedure was used in the other optimizations as well, where only some parameters, such as the sample size, were changed.

The optimization was carried out in several stages. First, with E_{FF} ($n = 2$) a set of about 5000 energy-minimized structures was generated with LTD, from which a smaller set of structures that were significantly different (based on the 60° criterion) was extracted. From this set, we retained only about 300 structures with energy within 30 kcal/mol above the lowest energy structure found, where NOS was added to the set as well. Then, the same ASP, σ, was assigned to all the atoms; the energy of each structure t (including that of NOS without the harmonic restraints) was minimized [becoming $E^t_{tot}(\sigma)$, where for simplicity $n = 2$ is omitted]; and the difference $\Delta^m_{tot}(\sigma) = E^{NOS}_{tot}(\sigma) - E^m_{tot}(\sigma)$ between the minimized energy of NOS, $E^{NOS}_{tot}(\sigma)$, and the lowest minimized energy of the set, $E^m_{tot}(\sigma)$, obtained for structure m was calculated. The minimizations were carried out with the L-BFGS program with an accuracy of about 0.1 kcal/mol for E_{tot} and a slightly higher accuracy for E_{FF}; the minimization of E_{tot} of a structure requires 30–50 CPU s on an Alpha workstation with the 21264 processor.

This process was repeated for various values of σ, where the optimal σ^* is defined as that leading to the lowest value of $\Delta^m_{tot}(\sigma)$ (we verified that the structure of NOS changed only very slightly in these minimizations). Then, an LTD run based on $E_{tot}(\sigma^*)$ was carried out, and if structures with energies lower than $E^m_{tot}(\sigma^*)$ were obtained, they were added to the set. At that point, we retained only about 150 structures within about 10 kcal/mol above the lowest energy structure and continued the process until $\Delta^m_{tot}(\sigma^*)$ did not change.

In the next stage, a different ASP was assigned to the carbon groups (C, CH, CH_2, and CH_3), it was optimized in the same way followed by another LTD run, and the process continued for three ASPs and more; when a set of ASPs was obtained, a new cycle of optimization followed to correct ASP values determined in the early stages of the process. The whole optimization process required generating up to about 20,000 energy-minimized structures by LTD. Because of our extensive conformational search, the lowest energy structure is identified in this process with the GEM structure; obviously, the validity of this assumption is not guaranteed.

APPENDIX C: A PROCEDURE FOR OPTIMIZING THE POSITIONS OF POLAR HYDROGENS

This optimization was performed by the rotation of the angles of the OH and NH vectors within the framework of an MCM procedure (eq A1) described in Appendix A. Some elements of our method are borrowed from methods developed previously.[76–78]

1. The polar H atoms of the side-chains of Tyr, Ser, Thr, Asn, Gln, and Arg were identified on the loop and the template, and for each one of them, a group of *neighbor polar hydrogens* was defined as those located within a radius of 10 Å. These polar hydrogens were free to move, whereas each of the other template + loop atoms was restrained to its PDB position by a harmonic potential with a prefactor of 0.15 kcal/mol/Å^2. In the MCM simulation, the dihedral angles of the OH and NH vectors of the polar hydrogens were changed.
2. In each MCM step, a hydrogen was selected at random, and the number of neighbor polar hydrogens to be treated was determined at random as well. Then, the specific neighbor hydrogens were chosen at random, their angles were rotated at random within the range (180°, −180°), and the energy [E_{FF} ($\epsilon = 2r$) + the harmonic potentials] of this trial structure was minimized, becoming E^0_j, and was accepted or rejected with eq A1. Every 10 MCM steps, the temperature parameter T^* was changed by 50 K from 200 to 500 K and vice versa.
3. The template + loop structure with the lowest energy obtained after about 3000 MCM steps was defined as the NOS structure; thus, the coordinates of the template's atoms were frozen, and the loop structure became the reference structure against which the RMSD of other loop structures was calculated. The NOS structures obtained for loops 3 and 1 deviate by RMSD ~ 0.2 Å from the corresponding PDB loop structures.