Optimization of Parameters in Macromolecular Potential Energy Functions by Conformational Space Annealing

ARTICLE *in* THE JOURNAL OF PHYSICAL CHEMISTRY B · JULY 2001

Impact Factor: 3.3 · DOI: 10.1021/jp011102u

6 AUTHORS, INCLUDING:

Jooyoung Lee
Korea Institute for Advanced Study
121 PUBLICATIONS **3,963** CITATIONS
SEE PROFILE

Daniel R Ripoll
Biotechnology High Performance Computing…
116 PUBLICATIONS **4,261** CITATIONS
SEE PROFILE

Cezary Czaplewski
University of Gdansk
159 PUBLICATIONS **2,581** CITATIONS
SEE PROFILE

Optimization of Parameters in Macromolecular Potential Energy Functions by Conformational Space Annealing

Jooyoung Lee,[†,‡] Daniel R. Ripoll,[§] Cezary Czaplewski,[†,||] Jarosław Pillardy,[†] William J. Wedemeyer,[†] and Harold A. Scheraga*[,†]

Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, Program of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-012, Korea, Cornell Theory Center, Ithaca, New York 14853-3801, and Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland

Received: March 22, 2001; In Final Form: June 4, 2001

A general protocol for refining the parameters of macromolecular potential energy functions by optimizing criteria that compare nativelike and nonnative conformations of one or more benchmark protein(s) is described. The protocol exploits the high efficiency of conformational space annealing (CSA) in finding the lowest-energy conformation of an isolated macromolecule. A novel form of the CSA method, local CSA, is introduced to provide better sampling of nativelike conformations. The computational expense of the protocol is reduced significantly by a linear approximation that estimates the energy of the (reminimized) native and nonnative conformations after every change of the force field parameters. The protocol is illustrated by optimizing the parameters of two force fields used in the CASP3 and CASP4 experiments, respectively. Another version of this general protocol (with different optimization criteria and optimization methods) was used to determine the parameters for the α, β and α/β force fields used in the CASP4 experiment, as reported in a companion publication (J. Pillardy et al. *J. Phys. Chem. B* **2001**, *105,* 7299).

I. Introduction

The prediction of the three-dimensional structure of a nonhomologous protein from its amino acid sequence is an important problem in computational biology, and its solution would have numerous applications, e.g., in interpreting the genomes that are becoming available. Several research groups have sought to solve this problem by minimizing a statistical (knowledge-based) potential that assesses protein conformations by their estimated probability of being observed in the database of protein structures.[1−3] By contrast, we have pursued an alternative strategy, that of minimizing a physics-based potential that assesses protein conformations by their estimated free energies.[4−7] This strategy is based on the thermodynamic hypothesis,[8] which postulates that proteins adopt a native structure that minimizes their free energy.

Physics-based potentials have an intrinsic advantage over knowledge-based potentials in that they may be applied even in atypical situations, such as simulating nonequilibrium states (e.g., protein folding intermediates) or predicting the structures of proteins under unusual conditions (e.g., in membranes or amyloid fibrils). Physics-based potentials are also more informative in assessing the relative contributions of the physical interactions that stabilize the native structures and structural motifs of proteins.

However, physics-based potentials have been less successful than knowledge-based potentials in predicting protein structure.[9] One key problem is that the physical energy landscape of proteins is riddled with local minima, making it difficult to search.[10] This multiple-minima problem has been largely overcome for typical single-domain proteins through the combination of united-residue models of proteins such as UNRES[11−15] and efficient search algorithms such as conformational space annealing (CSA).[16−18] Another obstacle to physics-based protein structure prediction is that accurate predictions of protein structure require high accuracy in the parameterization of the physics-based potentials; small errors in the force field parameters can produce large changes in the lowest-energy conformations. However, as described in this paper, efficient search methods such as CSA can also overcome this second problem.

Physics-based potentials are generally parameterized from ab initio quantum mechanical calculations and experimental data on model systems.[19−21] However, such calculations and experimental data do not determine the parameters with perfect accuracy. The residual errors in physics-based potentials may be relatively insignificant in simulations of uniform phases of small molecules, but such errors may have a significant effect in simulations of macromolecules where the total energy is obtained by summing the energies of a large number of specific interactions. Moreover, these interaction energies are known to cancel each other to a high degree in proteins,[22] making their systematic errors all the more significant. Thus, an iterative procedure is needed to refine the parameters determined from model systems to obtain the more accurate potentials necessary for macromolecular simulations. Such a procedure is described in this paper.

The essence of our general protocol is to tune the parameters of the force field to render the experimental structures of one or more benchmark proteins as the family of conformations with the lowest energy. Efficient search methods such as CSA are essential to this protocol for parameter refinement, since the

* To whom correspondence should be addressed. Tel: (607) 255-4034. FAX: (607) 254-4700. E-mail: has5@cornell.edu.
[†] Cornell University.
[‡] Korea Institute for Advanced Study.
[§] Cornell Theory Center.
[||] University of Gdańsk.

10.1021/jp011102u CCC: $20.00
Published on Web 07/03/2001

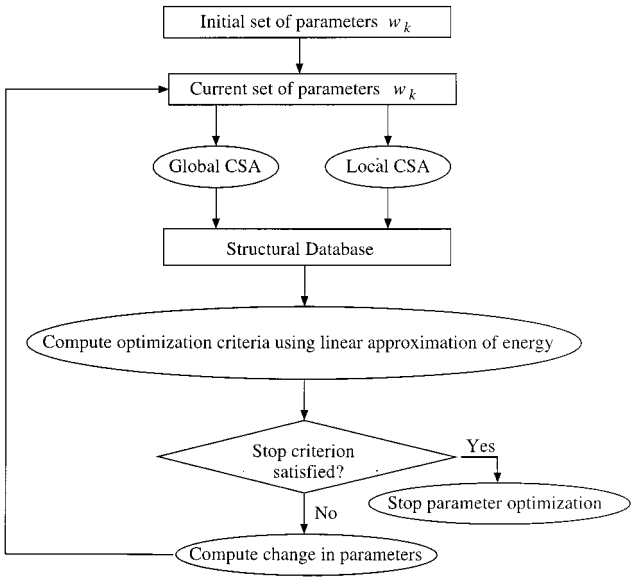

Figure 1. Flowchart for the general protocol for optimization of force field parameters.

global-minimum-energy conformation (GMEC) of the benchmark protein(s) must be determined for every trial set of force field parameters. In the following sections, we describe this protocol in more detail and use it to parameterize two united-residue force fields used in the third and fourth critical assessment of techniques for protein structure prediction (denoted as CASP3 and CASP4, respectively).[23,24] The CASP4 force field is denoted as α_0 and was most successful in predicting the tertiary structure of predominantly α-helical proteins. An alternative version of this general protocol (using different optimization criteria and a different optimization method) was used to optimize united-residue force fields[25] for three other structural classes of proteins (the α, β and α/β force fields), as described in a companion paper.[26]

II. Methods

General Protocol for Parameter Optimization. Our general protocol for refining protein force field parameters is as follows (Figure 1). We seek the set of parameters that gives the best agreement between the predicted and experimental structure of one or more benchmark proteins. The "best agreement" is defined in terms of one or more optimization criteria that compare the lowest energies and the distribution of energies in the native and nonnative families of conformations. To obtain a representative sampling of native conformations of each benchmark protein for a given set of parameters, we employ local CSA, a novel version of CSA,[16−18] which is described in the next section. We also carry out normal (global) CSA runs to obtain low-energy, nonnative (decoy) structures and to check whether the native family of conformations has been found by our global search procedure. The resulting low-energy native and nonnative conformations are stored in a structural database, from which the optimization criteria are computed. The force field parameters are then adjusted to improve these criteria, after which new local and global CSA runs are begun. The cycles of parameter adjustment and CSA sampling are continued until an acceptable set of parameters is obtained, as determined by some stopping criterion.

Two common criteria chosen for the assessment of force field parameters are the *Z*-score and energy gap of the nativelike versus nonnative conformations. The *Z*-score is defined here as the difference in the mean energies of the native and nonnative conformations, divided by the standard deviation of the nonnative conformation

$$Z = \frac{\langle E \rangle_{\mathrm{N}} - \langle E \rangle_{\mathrm{NN}}}{\sqrt{\langle E^2 \rangle_{\mathrm{NN}} - \langle E \rangle_{\mathrm{NN}}^2}} \tag{1}$$

where the angular brackets $\langle\ \rangle_{\mathrm{N}}$ and $\langle\ \rangle_{\mathrm{NN}}$ denote averaging over the native and nonnative conformational families, respectively. The energy gap is defined as the difference between the lowest energy of the native conformational family and that of the nonnative conformations.

These optimization criteria are computed from conformations produced in the present round of local and global CSA sampling and also from conformations produced in earlier rounds of CSA sampling (i.e., from conformations obtained by CSA sampling with different force field parameters). In other words, once a conformation has been admitted to the structural database in any round of CSA sampling, it is used in all subsequent rounds to compute the optimization criteria. A linear approximation (see Appendix) is used to estimate the energy obtained by reminimizing the earlier-round conformations using the force field parameters of the present round. This "recycling" of conformations allows the optimization criteria to be estimated accurately at low computational expense.

In the work described in this article, the stopping criterion was chosen to be a negative energy gap, i.e., with the energy of the native conformation being lower than that of any nonnative conformations; thus, the cycles of parameter adjustment and CSA sampling were continued until a negative energy gap (regardless of its magnitude) was obtained. In principle, it would have been possible to continue the optimization to produce larger (i.e., more negative) energy gaps, thus increasing the stability of the native family of conformations for the benchmark protein relative to its nonnative families. However, this was not done in this work for two reasons. First, the force field parameters should be transferable (i.e., applicable not only to the benchmark protein but to all proteins of the intended structural class), and it was judged that further tuning of the force field parameters would hamper their transferability. Second, it was desired to maintain the diversity of conformational families in the CSA runs, which likewise would have been hampered by further tuning of the force field parameters.

Local CSA Method. This parameter optimization protocol requires a representative sampling of the family of nativelike conformations for each benchmark protein. The local CSA method was developed to provide such a sampling.

To understand the local CSA method, it is helpful to review the normal (global) CSA method (Figure 2).[16−18] The global CSA method can be considered as a genetic algorithm that enforces a broad sampling in its early stages and gradually allows the conformational search to become focused in its later stages. The global CSA method maintains two banks of conformations: a fixed initial bank of randomly generated and energy-minimized conformations and an iteratively updated current bank that gradually converges on the lowest-energy conformations. In each iteration, new conformations are proposed by recombining a conformation of the current bank with segments of various length drawn from conformations in the initial and current banks. The recombined conformation is then energy-minimized and accepted or rejected depending on whether it is sufficiently different from the conformations in the current bank and whether its energy is lower than another conformation in the current bank. The "sufficiently different"

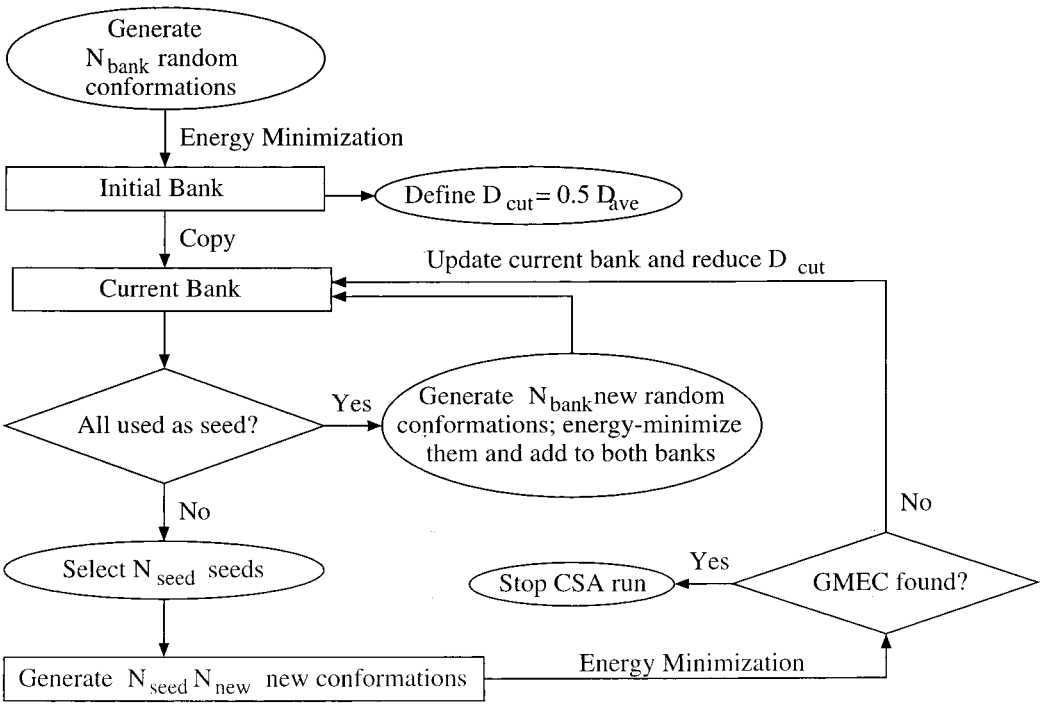

Figure 2. Flowchart of the global conformational space annealing (CSA) method. The quantities N_{bank}, N_{seed}, and N_{new} are typically taken as 50, 20, and 30, respectively. The quantity D_{ave} represents the average structural distance (here, a dihedral-angle distance) between the conformations of the initial bank.

criterion is assessed using a simple distance metric D in the dihedral-angle space. Specifically, the CSA algorithm computes the dihedral-angle distance between a new conformation (denoted as $\mathbf{x}_{new}$) and all conformations in the current bank; the conformation in the current bank with the smallest distance to the new conformation is denoted here as $\mathbf{x}_{nearest}$. If the dihedral-angle distance $D(\mathbf{x}_{new},\mathbf{x}_{nearest})$ is less than a cutoff D_{cut}, then the energies of $\mathbf{x}_{new}$ and $\mathbf{x}_{nearest}$ are compared; the lower-energy conformation is retained in the current bank and the other is eliminated. If, however, $D(\mathbf{x}_{new},\mathbf{x}_{nearest})$ is greater than the cutoff D_{cut} (i.e., if the new conformation is "sufficiently different" from all conformations in the current bank), the new conformation $\mathbf{x}_{new}$ is accepted into the current bank if its energy is lower than that of another conformation in the bank; the highest-energy conformation is then eliminated so that the number of conformations in the current bank remains constant. The dihedral-angle distance cutoff D_{cut} is gradually reduced as the CSA run continues, allowing the conformational search to focus on the low-energy regions of conformational space identified up to that point.

Local CSA differs from global CSA in two respects. First, the initial conformations are not generated randomly, but rather with the native conformation of the backbone and random side-chain conformations. These initial conformations are energy-minimized and are included in the initial bank, provided that they fall within a certain C^α rmsd of the native structure. Thus, the initial bank is composed of nativelike conformations (judging from their C^α rmsd) but with random side-chain dihedral angles. Second, the recombined structures are likewise subjected to the criterion that the C^α rmsd must also lie within a fixed cutoff of the native structure; this cutoff should be large enough to include a representative sampling of the native conformations but small enough to eliminate conformations with a grossly nonnative topology (4−5 Å rmsd for typical single-domain proteins). A new conformation is immediately rejected if it deviates from the native backbone structure by more than this C^α rmsd cutoff. Thus, local CSA samples low-energy conformations thoroughly while maintaining a nativelike structure, as assessed by the C^α rmsd.

Linear Approximation of Conformational Energy. By definition, the CSA method (whether local or global) produces protein structures that are local minima of the energy function. However, the force field parameters are altered in each round of local/global CSA sampling, and the structures obtained in earlier rounds are no longer local minima for the new set of parameters. This raises the question of whether such "legacy" structures should be reminimized using the new parameters before calculating the optimization criteria that assess the new force field parameters. Such reminimization would be computationally expensive, especially for a large number of structures, and this expense would inhibit our ability to parameterize force fields significantly.

However, we may sidestep the requirement of reminimization by exploiting three features of our optimization protocol, namely: (1) the force field parameters are being refined and, hence, their changes are relatively small; (2) the optimization criteria depend only on the energies of the conformations; and (3) the earlier-round structures are energy minima for a set of parameters that is close to the present-round set of parameters. As shown in the Appendix, small changes in the force field parameters generally produce first-order changes in the positions of the local minima, but these shifts in the minima produce only second-order shifts in their energies. Hence, the energy E of a reminimized structure for a new set of parameters w_m can be estimated to first order by the formula

$$E \approx E^0 + \sum_m V_m^0 (w_m - w_m^0) \qquad (2)$$

without having to actually reminimize the structures. Here w_m^0 and E^0 represent the parameters and corresponding energy for which the earlier-round conformation was a local minimum, and V_m^0 equals the first derivative of E with respect to the

parameters evaluated for the original parameters and conformation. Therefore, by storing the values of V_m^0 for each conformation produced by the local and global CSA runs, we may estimate the present-round energies of conformations produced in earlier rounds, even though the force field parameters have changed. These energies may then be used to calculate the present-round optimization criteria. The values of V_m^0 and E^0 are recalculated by reminimization using the new set of parameters to start the next iteration for the force field optimization.

This same formula (2) can be used to determine the optimal change in the force field parameters for improving the energy gap between the nativelike and nonnative conformations. The variation in the energy gap with force field parameters is described by the equation

$$\text{gap} \approx [E^0(\text{lowest N}) - E^0(\text{lowest NN})] + \sum_m [V_m^0(\text{lowest N}) - V_m^0(\text{lowest NN})](w_m - w_m^0) \quad (3)$$

where the constants E^0 and V_m^0 are evaluated for the lowest-energy native (N) and nonnative (NN) conformations in the structural database for a given set of force field parameters. Hence, the most rapid decrease in the energy gap with the force field parameters w_m is obtained by changing the parameters proportionally to the gradient of the energy gap

$$\delta w_m \propto -[V_m^0(\text{lowest N}) - V_m^0(\text{lowest NN})] \quad (4)$$

where the step size should be chosen small enough so that the linear approximation holds.

The steepest-descent method was modified slightly for the optimization of the CASP3 and CASP4 UNRES potentials described in this paper. For brevity, let g_m represent the gradient of the energy gap with respect to weight w_m

$$g_m \equiv \frac{\partial \text{gap}}{\partial w_m} = [V_m^0(\text{lowest N}) - V_m^0(\text{lowest NN})] \quad (5)$$

assuming that the linear approximation holds. A normalization factor $g_{\max}$ may be defined as the largest absolute magnitude of any component of the gradient

$$g_{\max} \equiv \max_m |g_m| \quad (6)$$

For each component U_m of the energy (corresponding to the weight w_m), we also compute the spread S_m of that energy, which is defined as the maximum energy U_m observed in the structural database minus the minimum energy U_m observed in the database

$$S_m \equiv [\max_{\mathbf{x}} U_m(\mathbf{x})] - [\min_{\mathbf{x}} U_m(\mathbf{x})] \quad (7)$$

where the maximum and minimum are taken over all the conformations $\mathbf{x}$ in the structural database. In terms of these quantities, the weights were updated using the formula

$$\delta w_m = -q \times \left[\frac{g_m}{g_{\max}}\right]\left(\frac{1}{S_m}\right) \quad (8)$$

where the coefficient q is a simple scale factor set to 0.01. It should be noted that the vector of weight changes δw_m is not perfectly aligned with the gradient g_m, due to the presence of the energy spread S_m in the denominator. Nevertheless, provided that the linear approximation holds, this method for changing the weights always decreases the energy gap, since the dot product between the gradient g_m of the energy gap, and the weight changes δw_m is always negative

$$\sum_m g_m \delta w_m < 0 \quad (9)$$

because S_m is always a positive number. In general, this method produces weight changes $\delta w_m \equiv w_m - w_m^0$ of less than 10 percent. When the weights varied by more than 10 percent, all the minima of the structural database were reminimized explicitly using the new weights.

III. Results and Discussion

To illustrate this general protocol, we optimized the parameters of the UNRES force fields used in the CASP3 and CASP4 exercises[23,24] to predict the structures of α-helical proteins. The UNRES model is a low-resolution model of proteins in which the backbone is a simple polymer of virtual bonds and each side chain is represented by a single ellipsoid.[11−15] The UNRES model has only two interaction centers, the center of each virtual bond (corresponding to the dipole of the peptide group) and the center of each side-chain ellipsoid. The C^α atoms define the backbone geometry but are not interaction centers. The corresponding UNRES force field may be derived as a cumulant expansion in an all-atom potential, which has been carried out to high order.[25,26] For this study, we parameterized a relatively low-order (and computationally inexpensive) UNRES expansion of only seven terms

$$E = \sum_{i>j}\sum_j U_{\text{SCSC}}(i,j) + w_{\text{SCp}}\sum_{i\neq j}\sum_j U_{\text{SCp}}(i,j) + w_{\text{pp}}\sum_{i<j-1}\sum_j U_{\text{pp}}(i,j) + w_{\text{b}}\sum_i U_{\text{b}}(i) + w_{\text{tor}}\sum_i U_{\text{tor}}(i) + w_{\text{rot}}\sum_i U_{\text{rot}}(i) + w_{\text{el-loc}}^{(4)}\sum_i U_{\text{el-loc}}^{(4)}(i) \quad (10)$$

These terms correspond to the following physical interactions.[15] U_{SCSC} represents the mean free energy of the hydrophobic (hydrophilic) interactions between the side chains of residues i and j, which is expressed either by a spherical Lennard-Jones potential (in the CASP3 UNRES potential) or by an orientation-dependent Gay−Berne potential[27] (in the CASP4 UNRES potential). The potential U_{SCp} corresponds to the excluded-volume interactions between the side chain of residue i and peptide group of residue j. The potential U_{pp} accounts mainly for the electrostatic and hydrogen-bonding interactions between the peptide groups of residues i and j. The terms U_{tor}, U_{b}, and U_{rot} denote the energies of virtual-dihedral-angle torsions, virtual-angle bending, and side-chain rotamers, respectively, of residue i; these terms reflect the most local structural propensities of the polypeptide chain. Finally, the correlation term $U_{\text{el-loc}}^{(4)}$ results from the cumulant expansion of the restricted free energy of the polypeptide chain, as described in detail in ref 25.

The internal parameters of these energy terms have been determined by fitting to various oligopeptide data, as described in ref 15. However, the relative thermodynamic weights of these terms (corresponding to the coefficients w_{SCSC}, w_{SCp}, w_{pp}, etc. in the expansion above) are not determined by this procedure. Hence, our parameter-refinement protocol was employed to determine the optimal values of these weights.

Our optimization protocol was first applied to the 10−55 fragment of the B-domain of staphylococcal protein A, with the UNRES force field used in the CASP3 exercise.[4,5] This earlier version of UNRES differs from the present CASP4

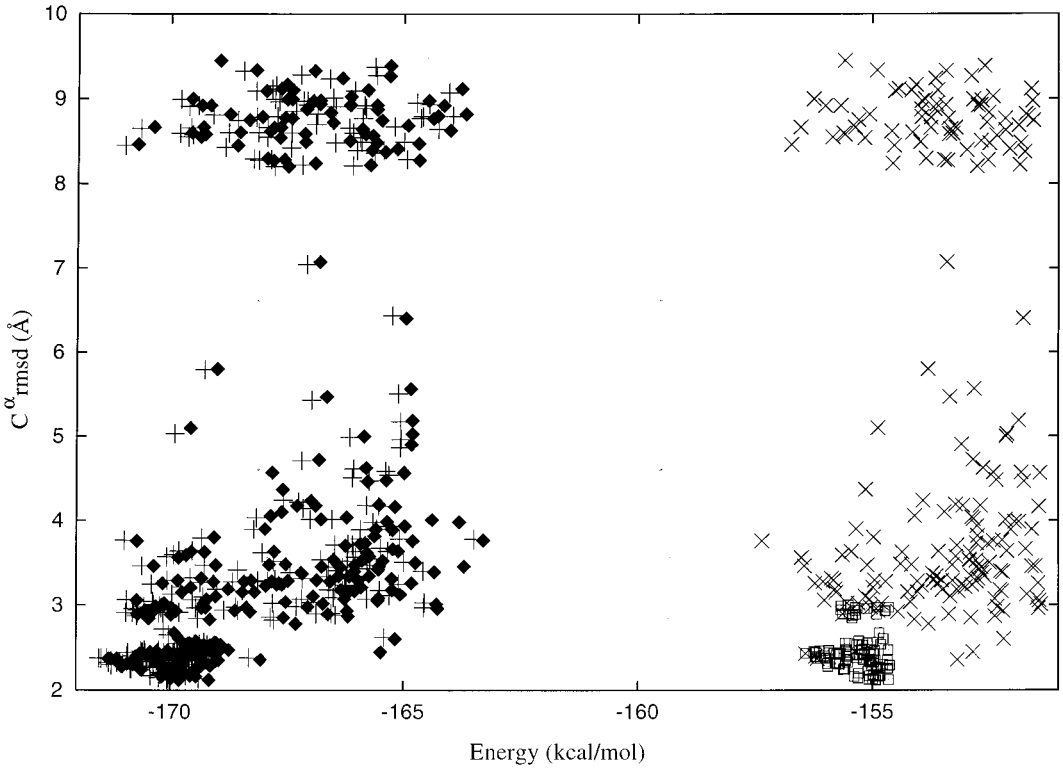

Figure 3. Accuracy of the linear approximation of conformational energy (see Appendix) for the 10−55 fragment of the B-domain of staphylococcal protein A. The open squares and Xs correspond to an original set of conformations sampled in local and global CSA runs, respectively, for a single set of force field parameters using the CASP3 UNRES potential (first row of Table 1). The plus signs correspond to the conformations obtained by reminimizing the original set of conformations using different force field parameters (second row of Table 1). The filled diamonds correspond to the energy values obtained by the linear approximation. The generally close agreement between the plus signs and filled diamonds indicates that the linear approximation estimates the energies of the reminimized conformations accurately and, hence, may be used to calculate the optimization criteria. For a small fraction of the data points, the linear approximation does not estimate the energy well, due to higher-order terms and (in some cases) the elimination of the local minimum with parameter changes. However, this fraction has a relatively small effect on the optimization criteria.

version[25] in two respects: the side-chain/side-chain potential U_{SCSC} is described by a spherical Lennard-Jones potential (not the Gay−Berne potential) and the torsional potential U_{tor} has a slightly different parameterization that does not disfavor left-handed α-helices.[15] Protein A was chosen as a benchmark protein, since it has been well studied and exhibits only two well-defined families of low-energy structures, the native fold and a "mirror-image" fold.[5] The energy gap was chosen as the optimization criterion for the force field parameterization, although other runs carried out using both the *Z*-score and the energy gap yielded similar results.

In each cycle of CSA sampling and parameter adjustment, the weights were allowed to vary by no more than 10%. A snapshot of the distribution of energies after one such cycle indicates that the linear energy approximation (Appendix) is generally valid (Figure 3). A negative energy gap was obtained after only six cycles (Table 1). Moreover, for these parameters, the global-minimum-energy conformation (GMEC) had a C^α rmsd of only 2.2 Å from the native structure.[28] This GMEC was found in seven independent global CSA runs, suggesting that it is indeed the global energy minimum, with a (CASP3) UNRES energy of −166.88 kcal/mol (Figure 4). On the average, the global CSA method required only 17800 local minimizations to reach this GMEC, indicating that its search of the conformational space of this 46-residue protein is efficient.

The weights for the new CASP4 version of the UNRES potential were computed by applying the same protocol to protein A (last line of Table 1). Once again, the energy gap alone was chosen as the optimization criterion, although other runs carried out using both the *Z*-score and the energy gap yielded similar results. In this case, the GMEC deviated from the native structure by only 2.1 Å C^α rmsd. The final set of weights is denoted as the α_0 force field and was used in the CASP4 exercise to predict the structure of predominantly α-helical proteins.[29]

TABLE 1: Relative Weights of the UNRES Energies in the CASP3 and CASP4 Force Fields[a]

iteration	w_{SCp}	w_{pp}	w_b	w_{tor}	w_{rot}	$w_{el-loc}^{(4)}$
0	1.00000	1.50000	0.10380	0.08617	0.10380	1.50000
1	0.96296	1.54011	0.15279	0.11133	0.11107	1.52180
2	0.99724	1.52081	0.16526	0.12616	0.06854	1.45778
3	0.96902	1.54604	0.18674	0.12406	0.04399	1.45766
4	0.97735	1.54589	0.17567	0.14434	0.06407	1.42960
5	0.97763	1.56392	0.17133	0.19504	0.06546	1.40583
6	0.98365	1.58463	0.19183	0.22648	0.06754	1.38512
CASP4	1.08229	1.51329	0.79839	0.74637	0.95818	1.69580

[a] The iteration number refers to the round of CSA optimization for the CASP3 force field. The weights are normalized relative to $w_{SCSC} = 1.0$. The weights differ between CASP3 and CASP4 because additional terms were included in the CASP4 cumulant expansion of UNRES.[25]

The parameterization of macromolecular force fields has also been carried out by threading methods.[30−32] Such methods generally employ a fixed set of decoy structures that do not depend on the force field parameters; hence, the optimization of linear weights (so that the native structure has the lowest energy) becomes a relatively simple problem in linear programming, albeit in a high-dimensional space.[31] A potential drawback of a fixed decoy set is that the variation of force field parameters

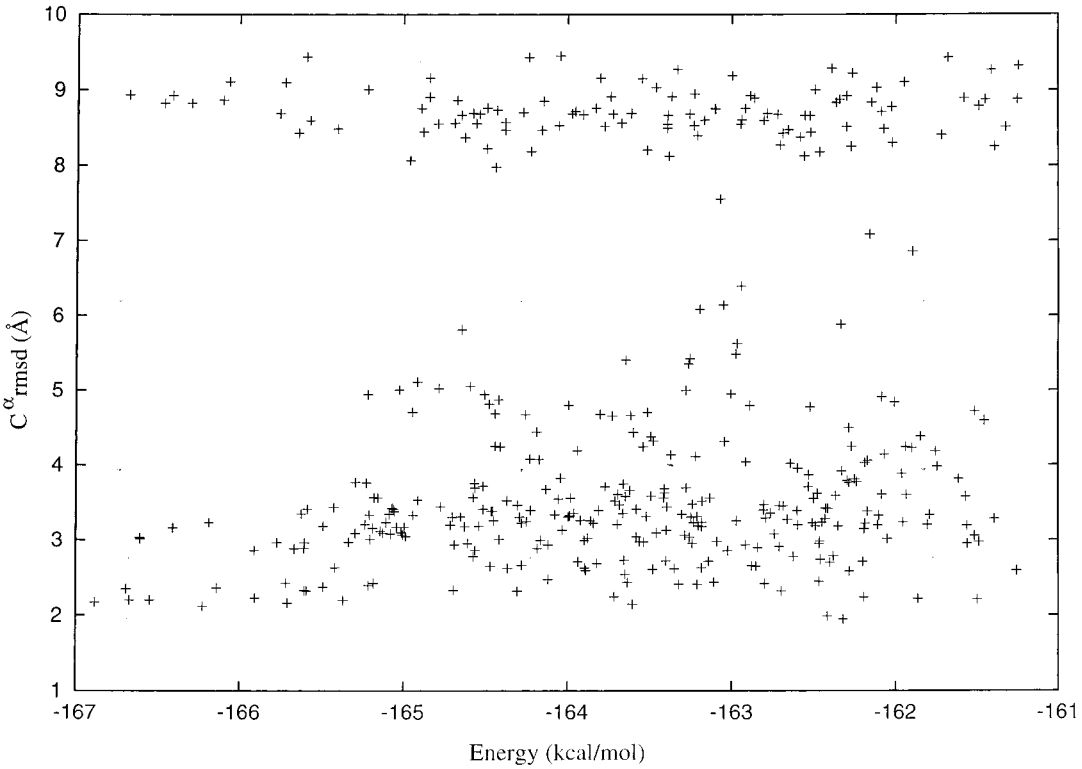

Figure 4. Plot of UNRES energy and C^α rmsd (from the experimental structure) for conformations of protein A obtained in local and global CSA sampling using the final optimized set of parameters for the CASP3 UNRES potential (sixth row of Table 1). The global minimum-energy conformation (GMEC) has an UNRES energy of −166.88 kcal/mol and differs from the experimental structure by only 2.2 Å C^α rmsd.

may cause some decoy structures to adopt a high energy that could be relieved by minimization. Hence, optimization criteria such as the energy gap and *Z*-score may not reflect the quality of the force field parameters. In particular, the optimization of force field parameters by fixed-decoy methods may not guarantee that the nativelike structures of the benchmark proteins are truly the global minimum structures for those force-field parameters; although nativelike structures will be preferred relative to the decoy structures, there may yet be other structures that are preferred relative to the nativelike structures. By contrast, our protocol carries out an unrestricted search of the full continuum of protein conformations (rather than a fixed, discrete set of decoys) and, hence, may allow us to have more confidence that the nativelike structures of the benchmark proteins are indeed global minima of the force field.

IV. Summary

We have described here a general protocol for optimizing force field parameters and illustrated one version of this protocol on united-residue force fields designed for the prediction of medium-resolution structures of predominantly α-helical proteins. An alternative version of this general protocol was used to develop UNRES force fields for structural predictions of proteins of other structural classes, as described in a companion paper.[26] Similar techniques are now being applied to the refinement of our all-atom potential, ECEPP.[33−35]

Acknowledgment. This work was supported by grants from the National Institute of General Medical Sciences (GM-14312), the National Science Foundation (MCB95-13167), the Fogarty Foundation (R03 TW1064), the NIH National Center for Research Resources (P41RR-04293), and the Polish State Committee for Scientific Research (PB 1244/T09/99/17). Support was also received from the National Foundation for Cancer Research. The computations described in this work were carried out primarily on our own array of 55 dual-processor PC computers. Additional computational resources were provided by: (a) the Cornell Theory Center, which receives funding from Cornell University, New York State, the NIH National Center for Research Resources, and members of the Theory Center's Corporate Partnership Program; (b) the National Partnership for Advanced Computational Infrastructure at the San Diego Supercomputer Center, which is supported in part by the NSF cooperative agreement ACI-9619020; and (c) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdánsk. W.J.W. was supported by an NIH postdoctoral fellowship (GM-19399).

Appendix: Derivation of the Linear Approximation of Local Minima Energies

In this Appendix, we derive a first-order (linear) approximation to the energies of the reminimized conformations that does not require them actually to be reminimized. We derive this approximation first for the case when the force field parameters are linear weights, and then demonstrate its validity for the general case.

The Case of Linear Weights. Let E be a weighted linear combination of several independent energy functions $U_m(\mathbf{x})$ of the macromolecular conformation $\mathbf{x}$

$$E(\mathbf{x};w_m) = \sum_m w_m U_m(\mathbf{x}) \tag{A1}$$

and let $\mathbf{x}_{\min}$ be a local minimum of E for one set of weights w_m^0

$$\frac{\partial}{\partial \mathbf{x}} E(\mathbf{x};w_m^0) = 0 \tag{A2}$$

The goal is to estimate the energy that would be obtained if the

macromolecule were reminimized starting from $\mathbf{x}_{\min}$ with a new set of weights w_m that differ only slightly from w_m^0, i.e., for which $\delta w_m \equiv w_m - w_m{}^0$ is small.

The individual energy functions $U_m(\mathbf{x})$ may be expanded in a Taylor expansion in the macromolecular conformation $\mathbf{x}$ about the local energy minimum $\mathbf{x}_{\min}$

$$U_m(\mathbf{x}) = V_m + \mathbf{D}_m\cdot\delta\mathbf{x} + \frac{1}{2}\delta\mathbf{x}\cdot\mathbf{J}_m\cdot\delta\mathbf{x} + ... \quad \text{(A3)}$$

where $V_m \equiv U_m(\mathbf{x}_{\min})$, $\mathbf{D}_m \equiv \partial/\partial\mathbf{x}\, U_m(\mathbf{x}_{\min})$, and $\mathbf{J}_m \equiv \partial^2/\partial\mathbf{x}\partial\mathbf{x}\, U_m(\mathbf{x}_{\min})$ are the energies, derivatives, and Hessian matrices of the energy functions at the local minimum $\mathbf{x}_{\min}$ and where $\delta\mathbf{x} \equiv \mathbf{x} - \mathbf{x}_{\min}$ represents the deviation from the local minimum. Thus, the energy E can be expanded similarly

$$E = E^0 + \sum_m \delta w_m V_m + \sum_m \delta w_m \mathbf{D}_m\cdot\delta\mathbf{x} + \frac{1}{2}\sum_m w_m^0 \delta\mathbf{x}\cdot\mathbf{J}_m\cdot\delta\mathbf{x} + ... \quad \text{(A4)}$$

where $E^0 \equiv \sum_m w_m^0 V_m$. It should be noted that all terms with $\sum_m w_m{}^0\mathbf{D}_m$ have been neglected because the local minimum condition $\partial/\partial\mathbf{x}E(\mathbf{x}_{\min},w_m^0) = 0$ implies that $\sum_m w_m{}^0\mathbf{D}_m = 0$. The gradient can be likewise expanded in $\delta\mathbf{x}$ and δw_m

$$\frac{\partial}{\partial\mathbf{x}}E = \sum_m \delta w_m \mathbf{D}_m + \sum_m w_m^0 \mathbf{J}_m\cdot\delta\mathbf{x} + ... \quad \text{(A5)}$$

The condition $\partial/\partial\mathbf{x}E(\mathbf{x}_{\min},w_m) = 0$ describes how the local minimum moves in response to changes in the weights δw_m; to first order in the weight changes, the local minimum is displaced by an amount $\delta\mathbf{x}_{\min}$ defined by the equation

$$[\sum_m w_m^0 \mathbf{J}_m]\cdot\delta\mathbf{x}_{\min} \approx -\sum_m \delta w_m \mathbf{D}_m \quad \text{(A6)}$$

Thus, the displacement $\delta\mathbf{x}_{\min}$ of the local minimum is first-order (i.e., scales linearly) in the weight changes δw_m. (This assumes that the matrix $\sum_m w_m^0\mathbf{J}_m$ can be inverted, i.e., that $\mathbf{x}_{\min}$ is a true local minimum and not an inflection point.) However, the displacement $\delta\mathbf{x}_{\min}$ of the local minimum produces only a second-order change in the energy

$$E = E^0 + \sum_m \delta w_m V_m + \frac{1}{2}\sum_m \delta w_m \mathbf{D}_m\cdot\delta\mathbf{x} + ... \quad \text{(A7)}$$

where the term involving the Hessian matrices $\mathbf{J}_m$ has been eliminated using the local minimum condition given in the previous equation. Therefore, the energies of the local minima can be estimated to first order in the weight changes δw_m without recalculating the positions of the local minima

$$E \approx E^0 + \sum_m \delta w_m V_m \quad \text{(A8)}$$

This linear approximation of local minimum energies accelerates the optimization of force field parameters significantly, as described in the main text.

The General Case. The linear approximation of energy does not depend on the specific functional form assumed above, i.e., does not require that E vary linearly in the parameters w_m. Any function $E(\mathbf{x};w_m)$ can be expanded in a double Taylor expansion in $\mathbf{x}$ and w_m about $\mathbf{x}_{\min}$ and w_m^0

$$E = E^0 + \frac{1}{2}\delta\mathbf{x}\cdot\mathbf{J}^0\cdot\delta\mathbf{x} + \sum_m \delta w_m V_m^0 + \sum_m \delta w_m \mathbf{D}_m{}^0\cdot\delta\mathbf{x} + ... \quad \text{(A9)}$$

where E^0 equals the energy for the original weights w_m^0 at the original local minimum $\mathbf{x}_{\min}$. The zero superscript on the constants of this expansion $\mathbf{J}^0 \equiv \partial^2/(\partial\mathbf{x}\partial\mathbf{x})E(\mathbf{x}_{\min},w_m^0)$, $V_m^0 \equiv \partial/\partial w_m E(\mathbf{x}_{\min},w_m^0)$, and $\mathbf{D}_m{}^0 \equiv \partial^2/(\partial w_m\partial\mathbf{x})E(\mathbf{x}_{\min},w_m{}^0)$ indicates that these constants now depend on the original parameters w_m^0 (unlike in the previous expansion). Nevertheless, the similar forms of the expansions (A4) and (A9) indicate that the same arguments may be applied to demonstrate the validity of the local approximation for the general case. Specifically, to first order in the weight changes δw_m, the local minimum is displaced by an amount $\delta\mathbf{x}_{\min}$ defined by the equation

$$\mathbf{J}^0\cdot\delta\mathbf{x}_{\min} \approx -\sum_m \delta w_m \mathbf{D}_m^0 \quad \text{(A10)}$$

which is the general-case analogue of eq A6 and which again assumes that the matrix $\mathbf{J}^0$ can be inverted. However, the displacement $\delta\mathbf{x}_{\min}$ of the local minimum produces only a second-order change in the energy

$$E = E^0 + \sum_m \delta w_m V_m^0 + \frac{1}{2}\sum_m \delta w_m \mathbf{D}_m^0\cdot\delta\mathbf{x} + ... \quad \text{(A11)}$$

which is analogous to eq A7. Hence, to first order in the weight changes, the energy can be estimated as

$$E \approx E^0 = \sum_m \delta w_m V_m^0 \quad \text{(A12)}$$

thus verifying the linear approximation for the general case.

References and Notes

(1) Sippl, M. J. *J. Comput.-Aid. Mol. Des.* **1993**, *7*, 473.
(2) Skolnick, J.; Koliński, A.; Ortiz, A. R. *J. Mol. Biol.* **1997**, *265*, 217.
(3) Bystroff, C.; Baker, D. *J. Mol. Biol.* **1998**, *281*, 565.
(4) Lee, J.; Liwo, A.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proteins Struct. Funct. Genet.* **1999**, *Suppl. 3*, 204.
(5) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025.
(6) Liwo, A.; Lee, J.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5482.
(7) Lee, J.; Liwo, A.; Ripoll, D. R.; Pillardy, J.; Saunders, J. A.; Gibson, K. D.; Scheraga, H. A. *Int. J. Quantum Chem.* **2000**, *77*, 90.
(8) Anfinsen, C. B. *Science* **1973**, *181*, 223.
(9) Orengo, C. A.; Bray, J. E.; Hubbard, T.; LoConte, L.; Sillitoe, I. *Proteins Struct. Funct. Genet.* **1999**, *Suppl. 3*, 149.
(10) Scheraga, H. A. *Biophys. Chem.* **1996**, *59*, 329.
(11) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1715.
(12) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849.
(13) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874.
(14) Liwo, A.; Kaźmierkiewicz, R.; Czaplewski, C.; Groth, M.; Ołdziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Comput. Chem.* **1998**, *19*, 259.
(15) Liwo, A.; Pillardy, J.; Czaplewski, C.; Lee, J.; Ripoll, D. R.; Groth, M.; Rodziewicz-Motowidło, S.; Kaźmierkiewicz, R.; Wawak, R. J.; Ołdziej, S.; Scheraga, H. A. In *RECOMB 2000, Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*; Shamir, R., Miyano, S., Istrail, S., Pevzner, P., Waterman, M., Eds.; ACM: New York, 2000, 193.
(16) Lee, J.; Scheraga, H. A.; Rackovsky, S. *J. Comput. Chem.* **1997**, *18*, 1222.
(17) Lee, J.; Scheraga, H. A.; Rackovsky, S. *Biopolymers* **1998**, *46*, 103.
(18) Lee, J.; Scheraga, H. A. *Int. J. Quantum Chem.* **1999**, *75*, 255.
(19) Maple, J. R.; Dinur, U.; Hagler, A. T. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5350.

(20) Maxwell, D. S.; Tirado-Rives, J.; Jorgensen, W. L. *J. Comput. Chem.* **1995**, *16*, 984.
(21) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
(22) Honig, B.; Yang, A.-S. *Adv. Protein Chem.* **1995**, *46*, 27.
(23) Third Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction; Asilomar Conference Center, December 13−17, 1998; http://predictioncenter.llnl.gov/casp3/Casp3.html.
(24) Fourth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, Asilomar Conference Center December 3−7, 2000; http://predictioncenter.llnl.gov/casp4/Casp4.html.
(25) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Chem. Phys.* **2001** in press.
(26) Pillardy, J.; Czaplewski, C.; Liwo, A.; Wedemeyer, W. J.; Lee, J.; Ripoll D. R.; Arlukowicz, P.; Oldziej, S.; Arnautova, Y. A.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7299.
(27) Gay, J. G.; Berne, B. J. *J. Chem. Phys.* **1981**, *74*, 3316.
(28) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. *Biochemistry* **1992**, *31*, 9665.
(29) Pillardy, J.; Czaplewski, C.; Liwo, A.; Lee, J.; Ripoll, D. R.; Kazmierkiewicz, R.; Oldziej, S.; Wedemeyer, W. J.; Gibson, K. D.; Arnautova, Y. A.; Saunders, J.; Ye, Y.-J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2329.
(30) Maiorov, V. N.; Crippen, G. M. *J. Mol. Biol.* **1992**, *227*, 876.
(31) Tobi, D.; Elber, R. *Proteins Struct. Funct. Genet.* **2000**, *41*, 40.
(32) Vendruscolo, M.; Mirny, L. A.; Shakhnovich, E. I.; Domany, E. *Proteins Struct. Funct. Genet.* **2000**, *41*, 192.
(33) Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. *J. Phys. Chem.* **1975**, *79*, 2361.
(34) Némethy, G.; Pottle, M. S.; Scheraga, H. A. *J. Phys. Chem.* **1983**, *87*, 1883.
(35) Némethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. *J. Phys. Chem.* **1992**, *96*, 6472.