An Efficient Deformation-Based Global Optimization Method for Off-Lattice Polymer Chains: Self-Consistent Basin-to-Deformed-Basin Mapping (SCBDBM). Application to United-Residue Polypeptide Chains

Jaroslaw Pillardy,[†] Adam Liwo,[‡] Malgorzata Groth,[‡] and Harold A. Scheraga*[,†]

Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, and Department of Chemistry, University of Gdańsk, ul. Sobieskiego 18, 80-952 Gdańsk, Poland

Received: March 24, 1999; In Final Form: June 9, 1999

A new method to surmount the multiple-minima problem in protein folding is proposed. Its underlying principle is to locate a group of large basins containing low-energy minima (hereafter referred to as *superbasins*) in the original energy surface. This is achieved by coupling the superbasins in the original surface to basins in a highly deformed energy surface (which contains a significantly reduced number of minima, compared to the original rugged energy surface). The distance scaling method (DSM) and the diffusion equation method (DEM) have been implemented to carry out the deformation. The procedure consists of macroiterations in which the parameter a, that controls the deformation, changes between two extreme values, a_{max} and a_{min} (a=0 corresponds to the original energy surface). The first macroiteration is initialized by imposing a maximum deformation on the original surface and then selecting 10 randomly generated conformations in the maximally deformed surface, whose energies are then minimized, usually leading to less than 10 minima; the next macroiterations are fed with the results of the previous ones. Each macroiteration consists of the following steps: (i) reversal of the deformation from a_{max} to a_{min}; a limited search is carried out in the neighborhood of the minima at each stage of the reversal; (ii) collection of the new low-energy minima in the a_{min}-deformed energy surface; (iii) back-tracking these minima up to a_{max} while increasing the deformation. Steps i − iii are iterated until no new minima are found in the undeformed surface, or a predefined number of iterations is exceeded. In the initial macroiteration, a_{min} is greater than 0, and a_{max} is chosen so that the deformed energy surface has only a few minima. In each next macroiteration, the new a_{max} is set at a_{min} of the previous macroiteration, and a_{min} is decreased, to reach 0 in the last macroiteration. The method was applied to united-residue polyalanine chains with a length of up to 100 amino acid residues, and to locate low-energy conformations of the 10-55 fragment of the B-domain of staphylococcal protein A.

1. Introduction

The multiple-minima problem is encountered in essentially all areas of theoretical chemistry, physics, engineering, as well as in many other branches of science. Despite its importance and the many efforts to surmount it, this problem still remains unresolved. In particular, theoretical conformational analysis of macromolecules, investigations of the spatial structures of many-atom clusters, and theoretical prediction of crystal structures are at best very difficult to treat because of the huge number of local minima in the corresponding energy surfaces and lack of effective, sufficiently general global-optimization algorithms.[1−3] Recently, a very efficient global optimization method, conformational space annealing (CSA), has been introduced and applied successfully in many cases.[4−6]

Another promising approach to surmount the multiple-minima problem involves methods based on the deformation of the original rugged energy surface, thereby reducing the number of minima by orders of magnitude, at best even to a single minimum, and simplifying the conformational search greatly.[7−15] It was hoped that, by tracking the lowest energy minimum obtained in the maximally deformed energy surface back to the original energy surface, the global minimum of the original energy function would be located;[7] we will hereafter refer to this approach as the *single trajectory approach*. However, this approach works only for relatively simple systems; in more complex cases, the global minimum in the original energy surface corresponds to a higher-energy minimum in the deformed energy surface and vice versa. Moreover, during the reversal of the deformation (the *reversing procedure*), a single trajectory often branches, forcing the algorithm to track only one branch (tracking all of them is effectively impossible because of the exponential growth of the number of trajectories as the reversal progresses). This problem remains in a *multiple trajectory* approach, in which *all* minima in the deformed energy surface are tracked back to the original energy surface. An attempt to alleviate this problem was the multiple trajectory perturbation approach (MTPA).[16,17] In this approach, each structure encountered at a particular reversal step of the deformation is perturbed and then energy-minimized, and a predefined number of lowest energy structures is taken to the next step of the reversal; the addition of this perturbation step alleviates the problem of splitting the trajectories as the deformation decreases. This approach was by far more suc-

* To whom correspondence should be addressed. Phone: (607) 255-4034. Fax: (607) 254-4700. E-mail: has5@cornell.edu.
[†] Cornell University.
[‡] University of Gdańsk.

10.1021/jp991014y CCC: $18.00
Published on Web 08/10/1999

cessful than the single or multiple-trajectory approach and was applied in the theoretical prediction of crystal structures.[16,17] However, it does not work for highly demanding applications, such as very large Lennard-Jones clusters or polypeptide chains.

The above considerations led to the conclusion that tracing the minima in the deformed energy surface back to the the original energy surface is insufficient to achieve the necessary coupling of the minima in the deformed and original energy surfaces, which is of central importance in the successful application of deformation methods. In our view, achieving this coupling would enable one to find the global minimum in the original energy surface, even if it is represented by a higher energy minimum in the deformed energy surface. Because of this coupling, the global minimum in the undeformed surface cannot be located by considering only the energy relations in the deformed surface (without referring to the original, undeformed, energy surface) because the energy of the corresponding minimum in the deformed surface may be so high that it would never be taken as a starting point for the reversing procedure. This leads to the conclusion that it is even more important to design an efficient reversing procedere that can achieve the coupling between the minima in the original and deformed energy surfaces than to develop a deformation itself (although effort must also be devoted to the latter subject, as discussed in subsection 2.3). Unfortunately, this problem has not been addressed thus far. In the present paper, we propose an algorithm to achieve the coupling between the original and deformed energy surfaces. The idea is not only to carry out the reversing procedure but also, after the original energy surface is restored, to select the lowest energy minima in the original energy surface and deform them, in order to find their representations in the highly deformed energy surface. This selects the minima in the deformed energy surface with which it is worth starting another reversing procedure, again selecting the newly obtained energy minima in the original energy surface, and iterating the procedure until no new energy minima in the highest deformed surface are obtained (or, in other words, until self-consistency is achieved). In order to treat the trajectory-splitting problem, we carry out a local search in the vicinity of current minima during the reversing procedure.

Another important point in the application of the deformation method is to design an efficient deformation, the principle being that the height of the energy barriers between the minima should gradually diminish. This is not very difficult to achieve if the energy function is a sum of two-body terms that have identical functional form (e.g., the Lennard-Jones potential). However, the energy functions that describe even relatively simple systems (e.g., homogeneous polymer chains) in a sufficiently realistic manner are usually more complex and consist of contributions that have very different meanings and functional forms. Deformations that are expected to be of general application [such as the diffusion equation method[7] (DEM)] are very complicated mathematically and thus practically impossible to apply rigorously in real cases such as polypeptide chains. Our idea is to apply a different type of deformation to each type of energy component, with the distance scaling method[14] (DSM) and the DEM, thereby having in principle different deformation parameters for each kind of deformation, and to calibrate them by relating all of them to a single deformation parameter that scales the various types of deformation with respect to each other. This calibration is chosen so as to achieve a smooth transition between the original and the deformed energy functions.

In this paper, we describe in detail the method outlined above, viz., the design of an efficient deformation procedure and a coupling between the undeformed and deformed surfaces in the reversing procedure, and its application to finding the global energy minima of polypeptide chains considered at the united-residue level, with interactions described by the UNRES united-residue energy function recently developed in our laboratory.[18−23] It has been shown in earlier work that this energy function can distinguish native structures of small proteins from nonnative structures reasonably well.[6,20−24] We designate this procedure as the self-consistent basin-to-deformed-basin mapping (SCB-DBM) method.

2 Methods

2.1. The Algorithm. We consider a function $f(\mathbf{x})$ of several variables. Consider a mapping $F(\mathbf{x},a)$, where a defines the extent of deformation, such that $F(\mathbf{x},a)$ becomes smoother with a gradually smaller number of energy minima with increasing a; assume that $F(\mathbf{x},0) = f(\mathbf{x})$. In order to utilize the deformation F to locate the global minimum of f, we first note that, with increasing a, the number of minima gradually decreases, because some of the minima merge into one. As deformation proceeds, groups of individual minima are first merged into single minima, defining the *superbasins* of these groups of minima for certain values of the deformation parameter. As the deformation parameter a increases, the minima continue to merge, causing the superbasins from smaller deformation (*lower order superbasins*) to merge, constituting higher order superbasins (for larger deformation). Finally, for a very high deformation, only a few minima remain. A logical procedure to find the global minimum of $f(\mathbf{x})$ would, therefore, be first to locate the highest order superbasin related to this minimum and then to locate within it the superbasins of gradually lower order that still contain this minimum, until the deformation is fully reversed, i.e., the global minimum in the original energy surface is located. However, the major difficulty in proceeding in this manner is that there is no straightforward relation between the values of F at its minima and the corresponding minimum values of f. Therefore, one can never tell which superbasin corresponds to the global minimum of the original energy function, based only on the "energy" relations between superbasins. As a consequence, it is clearly insufficient to reverse the deformation only once, in order to find the global minimum of f, even if a multitrajectory search is carried out during the reversing procedure. To surmount this problem, we propose a self-consistent procedure that finds the coupling between the superbasins of different order, which is achieved by iterating the steps consisting of reversing the deformation and, subsequently, reintroducing the deformation.

The procedure consists of a series of *macroiterations*. Each macroiteration establishes the coupling between superbasins of consecutive order (and contains a self-consistent procedure within it). In macroiteration i, the parameter $a^{(i)}$, which controls the deformation, changes between two extreme values $a^{(i)}_{\max}$ and $a^{(i)}_{\min}$. For macroiteration $i+1$, $a^{(i+1)}_{\max} = a^{(i)}_{\min}$ and $a^{(i+1)}_{\min} = a^{(i)}_{\min}/\Delta$ (or 0 in the last macroiteration), where Δ is a logarithmic step length. The first macroiteration is initialized with randomly generated and minimized conformations in the highest deformed space, while each subsequent macroiteration is fed with the results of the previous one.

The coupling procedure within each macroiteration is carried out as follows (see Figure 1). Let $\mathbf{x}^0_i$ represent a vector of coordinates for the ith minimum at a given stage of the reversing procedure. At each stage, there is a maximum of p minima with coordinates $\{\mathbf{x}^\circ_1, \mathbf{x}^\circ_2, ..., \mathbf{x}^\circ_p\}$.

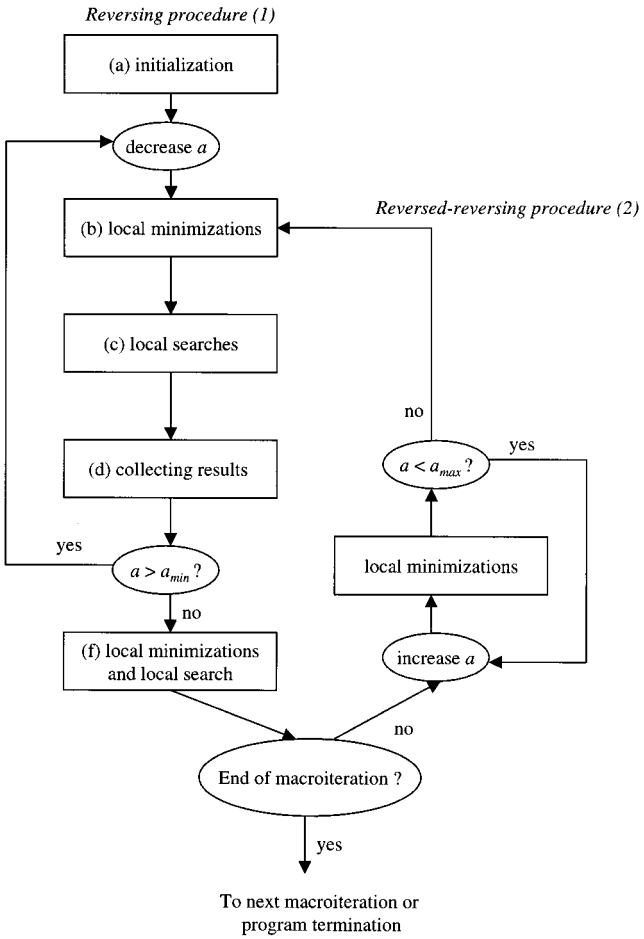

Figure 1. Block diagram of the reversing procedure coupled with the reversed-reversing procedure within a single macroiteration.

1. Reversing Procedure. (a) Set the deformation parameter $a = a_{\max}$ and set the initial values of the variables $\{\mathbf{x}_1^\circ, \mathbf{x}_2^\circ, ..., \mathbf{x}_p^\circ\}$. The initial vectors of the variables are selected at random in the first macroiteration or taken from the results of the previous one for each subsequent macroiteration. These variables are used later as starting points for local minimization of $F(\mathbf{x},a)$, with p being a predefined number of trajectories. (b) For the current value of the deformation parameter a, carry out local energy minimizations of $F(\mathbf{x},a)$ starting from $\{\mathbf{x}_1^\circ, \mathbf{x}_2^\circ, ..., \mathbf{x}_p^\circ\}$; this leads to the set of minima $\{\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_{n_1}^*\}$, where n_1 is $\leq p$. (c) Carry out n local conformational searches for the neighboring minima of each $\mathbf{x}_i^*$ (n being a pre-defined number). The local search algorithm is discussed later in this section. (d) Collect up to a predefined number ($n_2 \leq p$) of lowest-energy minima generated in steps 1b and c and store them as $\{\mathbf{x}_1^\circ, \mathbf{x}_2^\circ, ..., \mathbf{x}_{n_2}^\circ\}$. If $n_2 < p$, supplement this set with unused low-energy structures of previous iterations. (e) If the deformation parameter $a > a_{\min}$, decrease a and go to the next step of the reversing procedure (step 1b). (f) If the deformation parameter is already $a_{\min}$, store (up to the predefined number $P > p$) the remaining new higher energy structures, if any, for further back-tracking. Carry out local conformational searches on the undeformed surface to find neighboring minima. If the number of new structures is less than p, supplement the set with previously stored unused low-energy structures. If no new minima were discovered for $a = a_{\min}$ and all stored minima were already used for the reversed-reversing procedure, terminate. Otherwise go to the reversed-reversing procedure (step 2) using the updated and supplemented set $\{\mathbf{x}_1^\circ, \mathbf{x}_2^\circ, ..., \mathbf{x}_{n_3}^\circ\}$.

2. Reversed-Reversing Procedure. Starting from $\{\mathbf{x}_1^\circ, \mathbf{x}_2^\circ, ..., \mathbf{x}_{n_3}^\circ\}$, increase a and carry out local energy minimizations of $F(\mathbf{x},a)$; this gives the set of minima $\{\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_{n_4}^*\}$, $n_4 \leq p$. Iterate this step until $a_{\max}$ is reached. No local search is carried out during this procedure. When $a_{\max}$ is reached, go to the reversing procedure (step 1b), using $\{\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_{n_4}^*\}$ instead of $\{\mathbf{x}_1^\circ, \mathbf{x}_2^\circ, ..., \mathbf{x}_p^\circ\}$ as the starting set of minima. One cycle through the reversing and reversed-reversing procedure is considered as one iteration within the macroiteration.

The local search of step 1c plays a very important role in the algorithm by detecting branching of minima during the reversing procedure. However, this search should be carried out in the vicinity of a starting minimum; otherwise the relationship between minima may be lost (i.e., the newly found minimum may not be related to the previous one but may belong to a completely different "tree" of trajectories). The simplest but not the most efficient way to carry out the local search is of course random perturbation of a structure followed by a local energy minimization, as implemented in the multiple-trajectory perturbation approach.[16,17] If the perturbation is reasonably small the search is local, but quite often the subsequent energy minimization usually just restores the original structure. Another version of local search used in our algorithm is a linear search along randomly generated directions in multidimensional space; the search stops when a new basin is found or a predefined maximum number of steps is exhausted. This kind of search nearly always finds neighboring basins and rarely stays in the starting basin. Another kind of local search procedure uses the observation that geometrically neighboring structures may also be obtained by larger changes of a single variable rather than smaller changes of all of them. This is usually achieved by larger perturbations of a smaller number of variables (instead of small perturbations of all of them) followed by energy minimization.

Moreover, near the point at which a trajectory is branched, the daughter trajectories are formed by close energy minima that are separated by low-energy barriers. Thus, it is reasonable to search for those daughter energy minima that are connected to the parent minimum by gentlest ascent paths. The gentlest ascent method was first proposed by Crippen and Scheraga,[25] and by Cerjan and Miller,[26] and developed by Wales,[27,28] and by Tsai and Jordan.[29] We have tested various gentlest ascent approaches to the local search, but at the present time all of them are computationally too expensive and, therefore, they were not used in our program.

Instead, we have achieved the best results (i.e., the fastest convergence to the global minimum) when all kinds of local searches described above (except gentlest ascent) were used in our algorithm. The reason for this is, probably, that each kind of local search covers different regions around the starting minimum, and therefore, the collection of them searches the space around the starting minimum much more thoroughly. The local search on the undeformed energy surface (step 1f) is treated differently if $a_{\min} = 0$. In this case, after every predefined number of iterations within a macroiteration (usually three), a very short Monte Carlo with minimization (MCM)[30,31] search is carried out (until five structures are accepted by the Metropolis criterion), instead of using the local search procedures described above.

When the number of structures in the set $\{\mathbf{x}_i\}$ is equal to the maximum predefined number p (or P for the undeformed surface), and a new structure with low energy is found, the structure with the highest energy already present in the set is replaced. However, it very often happens that the energetic order of the minima changes with deformation, i.e., a structure having

low energy in the undeformed function becomes relatively higher in energy when deformation increases. This phenomenon may lead to elimination of potentially very low-energy structures, if the replacement rules are based on the order of the deformed energy. The alternative, used in our algorithm, is to track a newly obtained minimum back to the undeformed energy (by executing the reversing procedure for the single minimum without any local searches, in a single trajectory) and to use the energy of the corresponding undeformed structure as a criterion for replacement. Because all newly obtained minima are treated this way, the energies of their corresponding undeformed conformations are compared.

Another important point for increasing the efficiency of the procedure is to store the previously found structures during selected checkpoint steps of the reversing and reversed-reversing procedures and to compare them with the structures that are newly generated at this step. If a new structure is identical with a previously encountered structure, this indicates that the trajectory being followed has already been explored.

2.2. Representation of the Polypeptide Chain and Its Energy Function. In our model,[19,20] a polypeptide chain is represented by a sequence of α-carbon (C^α) atoms linked by virtual bonds with attached united-residue side chains (SC) and united-residue peptide groups (p) located in the middle between the consecutive α-carbons. Only the united peptide groups and united side chains serve as interaction sites, the α-carbons assisting only in the definition of the geometry (Figure 2). All the virtual bond lengths (i.e., $C^\alpha-C^\alpha$ and $C^\alpha-SC$) are fixed; the $C^\alpha-C^\alpha$ distance is taken as 3.8 Å which corresponds to *trans* peptide groups. In the current version of the force field, we allow, however, for variation of the side-chain positions with respect to the backbone (α_{SC} and β_{SC}) and for variation of the virtual-bond angles θ.

The energy (UNRES) of the virtual-bond chain in this united-residue representation is expressed by eq 1.

$$U = \sum_{i<j} U_{SC_iSC_j} + \sum_{i\neq j} U_{SC_ip_j} + w_{el}\sum_{i<j-1} U_{p_ip_j} + w_{tor}\sum_i U_{tor}(\gamma_i) + w_{loc}\sum_i [U_b(\theta_i) + U_{rot}(\alpha_{SC_i},\beta_{SC_i})] + w_{corr}U_{corr} \quad (1)$$

where $U_{SC_iSC_j}$, $U_{SC_ip_j}$, and $U_{p_ip_j}$ denote the energies of interactions between side chains, between side chains and peptide groups, and between peptide groups, respectively, $U_{tor}(\gamma_i)$ denotes the energy of variation of the virtual-bond dihedral angle γ_i, $U_b(\theta_i)$ denotes the "bending" energy of the virtual-bond angle θ_i, $U_{rot}(\alpha_{SC_i}, \beta_{SC_i})$ is the local energy of side chain i, U_{corr} includes cooperative terms (e.g., the four-body interactions considered by Skolnick and co-workers[32]), and the w's denote relative weights of the respective energy terms.

The term $U_{SC_iSC_j}$ consists of the mean free energy of the hydrophobic (hydrophilic) interactions between the side chains. It, therefore, implicitly contains the contributions coming from the interactions with the solvent. Its functional form is expressed by eq 2.

$$U_{ij} = 4[|\epsilon_{ij}|x_{ij}^{12} - \epsilon_{ij}x_{ij}^{6}] \quad (2)$$

where ϵ_{ij} is the pair-specific van der Waals well-depth; $\epsilon > 0$ corresponds to hydrophobic−hydrophobic-type and $\epsilon < 0$ to hydrophobic−hydrophilic and hydrophilic−hydrophilic-type interactions (see Figure 3 for illustration). The quantity x_{ij} is the reciprocal of the reduced distance between side chains; it can depend on their distance alone for a radial-only potential (in this case $x_{ij} = \sigma_{ij}^\circ/r_{ij}$, r_{ij} being the distance between the side chains and σ_{ij}° a pair-specific constant that depends on the types of side chains i and j) or on both distance and orientation [i.e., $x_{ij} = x(r_{ij},\omega_{ij}^{(1)},\omega_{ij}^{(2)},\omega_{ij}^{(3)})$, where $\omega_{ij}^{(k)}$ defines the orientation of the united-atom side chains]; the same applies to ϵ_{ij}. In our work, we have considered and parameterized functional forms of both types. The functional forms of x and ϵ can be found in eqs 3−8 in the original paper.[20]

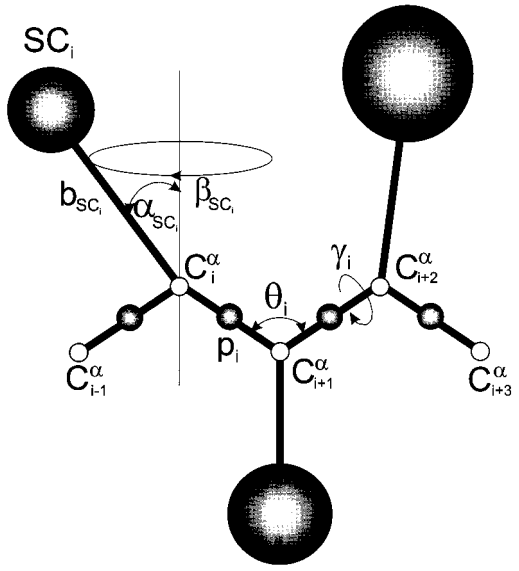

Figure 2. United-residue representation of a polypeptide chain. The interaction sites are side-chain centroids of different sizes (SC) and peptide-bond centers (p) indicated by solid circles, while the α -carbon atoms (small empty circles) are introduced only to assist in defining the geometry. The virtual $C^\alpha-C^\alpha$ bonds have a fixed length of 3.8 Å, corresponding to a trans peptide group; the virtual-bond (θ) and dihedral (γ) angles are variable. Each side chain is attached to the corresponding α-carbon with a fixed "bond length", b_{SC_i}, variable "bond angle", α_{SC_i}, formed by SC_i, and the bisector of the angle defined by C^α_{i-1}, C^α_i, and C^α_{i+1}, and with a variable "dihedral angle" β_{SC_i} of counterclockwise rotation about the bisector, starting from the right side of the C^α_{i-1}, C^α_i, C^α_{i+1} frame.

The term $U_{SC_ip_j}$ prevents too-close contacts between side chain of one residue with the backbone of another, and is described by a simple, excluded volume potential

$$U_{SC_ip_j} = \epsilon_{SC_ip_j}(r^\circ_{SC_ip_j}/r_{ij})^6 \quad (3)$$

The peptide-group interaction potential ($U_{p_ip_j}$) accounts mainly for the electrostatic interactions between them or, in other words, for their tendency to form backbone hydrogen bonds. In contrast to $U_{SC_iSC_j}$, its functional form was derived rigorously by averaging the simplified electrostatic-interaction energy of the peptide-group dipoles over the angles λ for their rotation about the corresponding $C^\alpha-C^\alpha$ virtual-bond axes, assuming that each peptide group is modeled by a point dipole located in the middle of the virtual bond, as proposed by Piela and Scheraga[33] (Figure 4). The potential is expressed by eq 4; the details of the derivation and parameterization can be found in the papers cited.[18,19]

$$U_{p_ip_j} = \frac{A_{p_ip_j}}{r_{ij}^3}(\cos\alpha_{ij} - 3\cos\beta_{ij}\cos\gamma_{ij}) - \frac{B_{p_ip_j}}{r_{ij}^6}[4 + (\cos\alpha_{ij} - 3\cos\beta_{ij}\cos\gamma_{ij})^2 - 3(\cos^2\beta_{ij} + 2\gamma_{ij})] + \epsilon_{p_ip_j}\left[\left(\frac{r^\circ_{p_ip_j}}{r_{ij}}\right)^{12} - 2\left(\frac{r^\circ_{p_ip_j}}{r_{ij}}\right)^6\right] \quad (4)$$

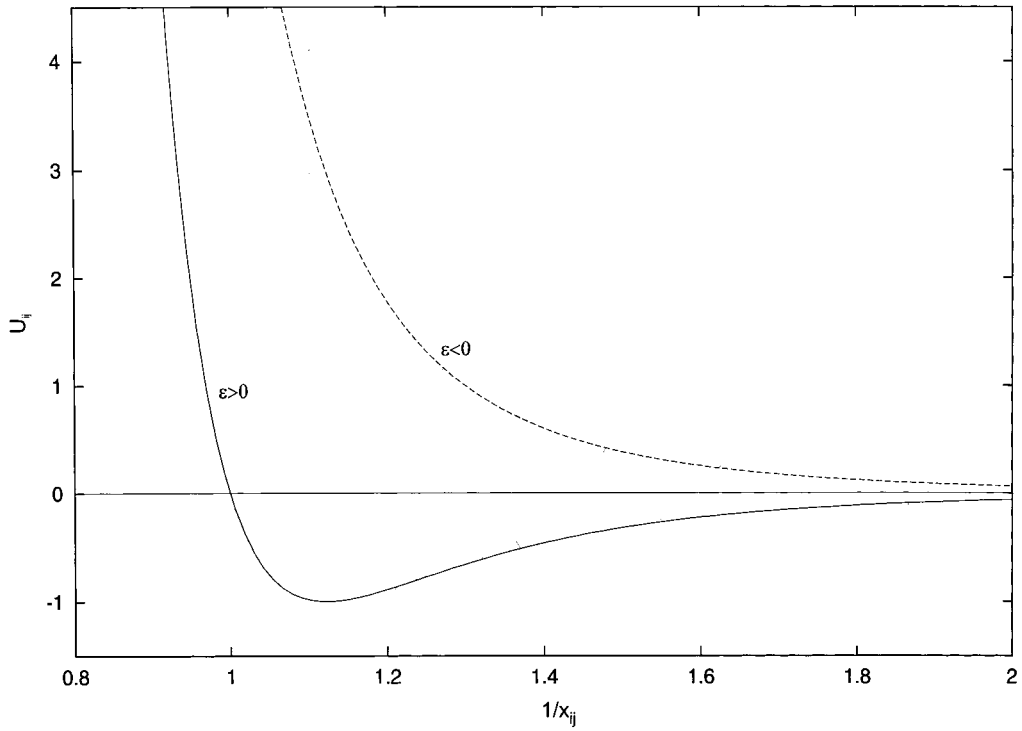

Figure 3. Plot of the reduced energy of interaction between the hydrophobic ($\epsilon > 0$) and hydrophilic ($\epsilon < 0$) side chains as a function of the reduced distance [eq 2].

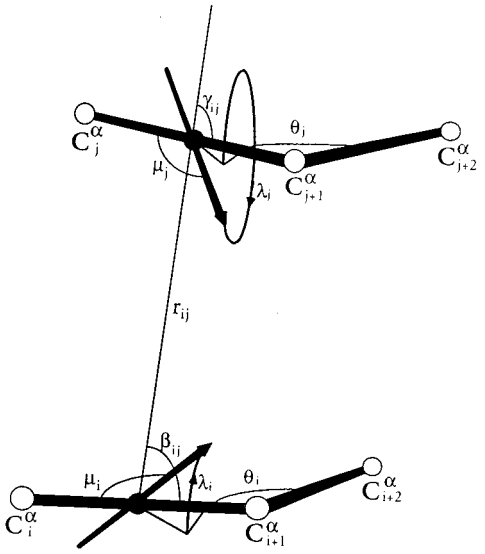

Figure 4. The relative orientation of the virtual bonds $C_i^\alpha - C_{i+1}^\alpha$ and $C_j^\alpha - C_{j+1}^\alpha$ is described by the angles α_{ij}, β_{ij} and γ_{ij}, defined by eq 4. The angle α_{ij} is not shown here because the two virtual bonds $C_i^\alpha - C_{i+1}^\alpha$ and $C_j^\alpha - C_{j+1}^\alpha$, are not necessarily coplanar. θ is the angle between two successive virtual bonds. The peptide-group dipole moments are represented by arrows (pointing from the carbonyl oxygen to the amide hydrogen of a peptide group), and the constant angles μ_i and μ_j between them and the virtual bonds are also shown, as well as the rotation angles λ_i and λ_j of the peptide-group dipoles. The two dipoles are separated by a distance r_{ij}.

with

$$\cos\alpha_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j \qquad \cos\beta_{ij} = \mathbf{v}_i \cdot \mathbf{e}_{r_{ij}} \qquad \cos\gamma_{ij} = \mathbf{v}_j \cdot \mathbf{e}_{r_{ij}}.$$

where $A_{p_ip_j}$, $B_{p_ip_j}$, $\epsilon_{p_ip_j}$, and $r^\circ_{p_ip_j}$ are constants characteristic of the type of interacting peptide groups, r_{ij} is the distance between the peptide-group centers, $\mathbf{v_i}$ is the unit vector pointing from C_i^α to C_{i+1}^α, and $\mathbf{e}_{r_{ij}}$ is the unit vector pointing from p_i to p_j (see Figure 4 for illustration). Two types of peptide groups were distinguished: ordinary and proline; the latter can act only as a hydrogen-bond acceptor and pertains to all *N*-methylated amino-acid residues (e.g., sarcosine). This gives a total of three sets of constants in eq 2; ordinary−ordinary, ordinary−proline, proline−proline. The angular part of eq 4 favors parallel and near-parallel, or antiparallel and near-antiparallel orientation of the virtual $C^\alpha - C^\alpha$ bonds, as encountered in hydrogen-bonded backbone peptide groups.

The torsional energy U_{tor} is expressed in terms of a Fourier series in the virtual-bond dihedral angles γ of Figure 2, as given by eq 5.

$$U_{\text{tor}}(\gamma_i) = a_0 + \sum_{k=1}^{6}[a_k(\cos k\gamma_i + 1) + b_k(\sin k\gamma_i + 1)] \quad (5)$$

This energy reflects the local propensities of the polypeptide chain, i.e., to form the right- rather than left-handed α-helices and the left- rather than right-twisted β-strands (which results in right-twisted β-sheets). It was natural to consider three torsional types of amino-acid residues: glycine (because of the absence of the β-carbon), proline (because of the restriction caused by the presence of the pyrrolidine ring), and alanine (which represents all other amino-acid residues). Detailed analysis of local propensities of the structures contained in the PDB confirmed this division.[21]

There are two additional energy terms in the UNRES potential describing local interactions: U_b and U_{rot}. U_b represents the energy of virtual bond-angle bending and, therefore, introduces the correlation between virtual-bond angle θ and the virtual-bond torsional angles γ adjacent to it (see Figure 2). The term U_{rot}, in turn, describes the energy of side-chain rotamers. Because there is a correlation between the virtual bond angles θ and the angles α_{SC} and β_{SC} (see Figure 2) that belong to the same residue, U_{rot} is expressed using three-dimensional Gaussians in θ, α_{SC} and β_{SC}. The details about the functional forms of U_b and U_{rot} can be found in the original paper.[21]

2.3. Deformation. As mentioned in the Introduction, different types of deformation are applied to the different energy terms of eq 1. There are two kinds of energy terms: one depending only on intersite distances ($U_{SC_iSC_j}$, $U_{SC_ip_j}$, $U_{p_ip_j}$ and U_{corr}) and one depending explicitly on internal coordinates (U_{tor}, U_b, and U_{rot}). It should be noted that this situation is common for all molecular force fields. As to the first group of energy terms, it is reasonable to apply the distance scaling method (DSM) which is very simple to implement and was shown to perform very well in finding the global minimum of Lennard-Jones and water clusters[10,11,14] and in predicting the crystal structures of small molecules.[17] Because distances cannot be translated to internal coordinates unequivocally, in the case of the local-interaction terms (U_{tor}, U_b, and U_{rot}), we applied a deformation that is similar to the diffusion equation method (DEM).[7]

In the DSM,[14] the site−site distance r_{ij} is transformed to $\tilde{r}_{ij}$ as follows:

$$\tilde{r}_{ij} = \frac{r_{ij} + ar_{o,ij}}{1 + ba} \tag{6}$$

The parameter $r_{o,ij}$ in eq 6 is the position of the minimum in the undeformed pairwise-interaction term under consideration. On increasing the deformation parameter a, the original function of the site−site distance (e.g., the Lennard-Jones potential) is flattened, but the position of its minimum and the function value at the minimum remain the same, if the value of the parameter b is taken as 1 (as in the original formulation[14] of the DSM). The parameter b controls the position of the minimum and remains constant during the calculations. A value of $b > 1$ means that the position of the minimum of the deformed site−site function will shift to larger values, while for $b < 1$ it will shift towards zero, and a two-body potential will become totally attractive for $a = 1/(1 - b)$. Different possible values of b were considered for the calculations, as described later in this section. It is relatively easy to choose $r_{o,ij}$, if the function has a minimum (e.g., the Lennard-Jones potential). However, some of the $U_{SC_iSC_j}$ and all other long-range terms in eq 1 are monotonic functions (e.g., the elctrostatic term $1/r$). In this case, it is reasonable to choose $r_{o,ij}$ so large that the function value at this point is close to zero; thus, this energy contribution will effectively be eliminated for large deformation.

Equation 6 can be applied only to the terms that depend only on site−site distance. Thus, eq 2 becomes

$$\tilde{U}_{ij} = 4[|\epsilon_{ij}|[\sigma^\circ_{ij}/\tilde{r}_{ij}]^{-12} - \epsilon_{ij}[[\sigma^\circ_{ij}/\tilde{r}_{ij}]^{-6}] \tag{7}$$

and eq 3 becomes

$$U_{SC_ip_j} = \epsilon_{SC_ip_j}\left(\frac{r^\circ_{SC_ip_j}}{[r_{SC_ip_j} + aR]/[1 + ba]}\right)^6 \tag{8}$$

where R is a large value (e.g., 10 Å), at which $U_{SC_ip_j}$ becomes zero.

However, the terms $U_{p_ip_j}$ and U_{corr} depend on both site−site distance and orientation. To obtain appropriate formulas, one must refer to the physical origin of these energy contributions; they are effective average electrostatic potentials of interactions between peptide-group dipoles. Recalling that the interaction between electric multipoles can be represented as an asymptotic expansion of the energy of interaction of the point charges representing these multipoles we can derive formulas for deformation. Hence, writing eq 4 in functional form

$$U_{p_ip_j} = \frac{A}{r_{ij}^3} f(\alpha_{ij}, \beta_{ij}, \gamma_{ij}) + \frac{B}{r_{ij}^6} g(\alpha_{ij}, \beta_{ij}, \gamma_{ij}) + \epsilon_{p_ip_j}\left[\left(\frac{r^\circ_{p_ip_j}}{r_{ij}}\right)^{12} - 2\left(\frac{r^\circ_{p_ip_j}}{r_{ij}}\right)^6\right] \tag{9}$$

the transformation of eq 4 by the DSM becomes

$$U_{p_ip_j} = \frac{A}{\tilde{r}_{ij}^3} \tilde{f}(\alpha_{ij}, \beta_{ij}, \gamma_{ij}) + \frac{B}{\tilde{r}_{ij}^6} \tilde{g}(\alpha_{ij}, \beta_{ij}, \gamma_{ij}) + \epsilon_{p_ip_j}\left[\left(\frac{r^\circ_{p_ip_j}}{\tilde{r}_{ij}}\right)^{12} - 2\left(\frac{r^\circ_{p_ip_j}}{\tilde{r}_{ij}}\right)^6\right] \tag{10}$$

To obtain the functional forms for $\tilde{f}(\alpha_{ij}, \beta_{ij}, \gamma_{ij})$ and $\tilde{g}(\alpha_{ij}, \beta_{ij}, \gamma_{ij})$, we use a multipole expansion as

$$E_{el} = \sum_{ij}\frac{q_iq_j}{r_{ij}} = \sum_{ij}\frac{q_iq_j}{R + \delta_{ij}} = \sum_k \frac{f_k(\alpha, \beta, \gamma)}{R^k} \tag{11}$$

where R is the large distance between the centers of two systems of point charges and $\delta_{ij} = |\delta_i - \delta_j|$; δ_i is the vector pointing from the center of the first system of point charges to the charge q_i in this system, and δ_j is the vector pointing from the center of the second system of point charges to charge q_j in the second system.

After application of the DSM, the functional form of the deformed E_{el} is

$$\begin{aligned}
\tilde{E}_{el} &= \sum_{ij}\frac{q_iq_j}{\tilde{r}_{ij}} = (1 + ba)\sum_{ij}\frac{q_iq_j}{r_{ij} + ar_0} = (1 + ba)\sum_{ij}\frac{q_iq_j}{R + \delta_{ij} + ar_0} \\
&= (1 + ba)\sum_k \frac{f_k(\alpha, \beta, \gamma)}{(R + ar_0)^k} = \\
&\quad (1 + ba)\sum_k\left[\frac{1}{(1 + ba)^k}\frac{f_k(\alpha, \beta, \gamma)(1 + ba)^k}{(R + ar_0)^k}\right] \qquad (12) \\
&= (1 + ba)\sum_k\left[\frac{1}{(1 + ba)^k}\frac{f_k(\alpha, \beta, \gamma)}{\tilde{R}^k}\right] = \\
&\quad \sum_k\left[\frac{1}{(1 + ba)^{k-1}}\frac{f_k(\alpha, \beta, \gamma)}{\tilde{R}^k}\right] \\
&= \sum_k \frac{\tilde{f}_k(\alpha, \beta, \gamma)}{\tilde{R}^k}
\end{aligned}$$

where $\tilde{f}_k(\alpha, \beta, \gamma) = f_k(\alpha, \beta, \gamma)/(1 + ba)^{k-1}$.

The function f from eq 10 corresponds to the $k = 3$ term from eq 11. The function g from eq 10 corresponds to the square of the $k = 3$ term from eq 11; it is a part of the second-order term in the expression for the Boltzmann-averaged energy of the dipole−dipole interaction (averaged over the rotation of the peptide-group dipoles about the $C^\alpha - C^\alpha$ virtual bonds) and changes with deformation as the square of f. Therefore, their deformed functional forms are given by

$$\tilde{r}_{p_ip_j} = \frac{r_{p_ip_j} + ar_{o,p_ip_j}}{1 + ba}$$

$$\tilde{f}(\alpha_{ij}, \beta_{ij}, \gamma_{ij}) = \frac{f(\alpha_{ij}, \beta_{ij}, \gamma_{ij})}{(1 + ba)^2} \quad (13)$$

$$\tilde{g}(\alpha_{ij}, \beta_{ij}, \gamma_{ij}) = \frac{g(\alpha_{ij}, \beta_{ij}, \gamma_{ij})}{(1 + ba)^4}$$

The same type of distance scaling transformation is applied to the correlation term U_{corr} of eq 40 of ref 22, i.e.,

$$\tilde{U}_{corr;i,i-1,k,k-1} = \frac{z^2}{4}\Big[9(\tilde{\zeta}_{i-1,k} + \tilde{\bar{\zeta}}_{i-1,k})(\tilde{\zeta}_{i,k-1} + \tilde{\bar{\zeta}}_{i,k-1}) + 4(\tilde{\zeta}_{i-1,k} - \tilde{\bar{\zeta}}_{i-1,k})(\tilde{\zeta}_{i,k-1} - \tilde{\bar{\zeta}}_{i,k-1})\Big] \quad (14)$$

with

$$\tilde{\zeta}_{ik} = \frac{1}{\tilde{r}_{ik}^3}\tilde{\phi}(\alpha_{ij}, \beta_{ij}, \gamma_{ij}) \quad (15)$$

$$\tilde{\phi}(\alpha_{ij}, \beta_{ij}, \gamma_{ij}) = \frac{\phi(\alpha_{ij}, \beta_{ij}, \gamma_{ij})}{(1 + a)^2} \quad (16)$$

where ϕ is defined in eq 11 of ref 22.

For application of the DEM to the remaining terms of eq 1, the formulas for the deformation of the local-interaction terms are as follows. The torsional part of the potential is not distance dependent but is simply a sum of sines and cosines. The original DEM acts as a high-frequency Fourier filter, removing high frequencies with speed $\sim \exp(-n^2a)$, where n is the frequency. This would cause the torsional part of the potential to be too weak compared to the other contributions that are smoothed with the same value of a. Therefore, we applied a Fourier filter with coefficients $\sim \exp(-na)$, which diminishes the high-frequency terms much more slowly. Thus, after deformation by the DEM, eq 5 becomes

$$\tilde{U}_{tor}(\gamma) = a_0 + \sum_{n=1}^{6}\{a_i[\cos n\gamma \exp(-\kappa na) + 1] + b_i[\sin n\gamma \exp(-\kappa na) + 1]\} \quad (17)$$

The functional form of U_b is given by eqs 4−7 of ref 22. It is extremely difficult, if at all possible, to derive a DEM-deformed form of U_b. Therefore, we decided to change its weight exponentially with the deformation, eliminating this term for large deformations, especially since this term has only a very small influence on the conformation. From the physical point of view, this kind of deformation corresponds to smoothing of U_b by removing restrictions imposed on the virtual-bond angles. Hence, $\tilde{U}_b$ is given by

$$\tilde{U}_b = \exp(-\kappa' a)U_b \quad (18)$$

The numerical parameters κ and κ' in eqs 17 and 18, respectively, are chosen appropriately (both equal to 2) to achieve a balance between the different types and parameters of the deformation for different energy contributions (as discussed later in this section).

Instead of using the DSM or DEM, the side-chain rotamer energy U_{rot} is deformed by application of explicit averaging of its Gaussian terms. U_{rot} is the negative of the logarithm of a sum of Gaussians in the (α_{SC}, β_{SC}) space. Each term of U_{rot} also depends on the corresponding virtual-bond angle θ. From eqs 8 and 11 of ref 21, U_{rot} can be expressed as follows:

$$U_{rot}(\theta, \alpha_{SC}, \beta_{SC}; \mathbf{h}, \mathbf{x}^\circ, \mathbf{D}) = -RT \log\sum_{i=1}^{nc} h_i \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}_i^\circ)^T\mathbf{D}_i(\mathbf{x} - \mathbf{x}_i^\circ)\right] \quad (19)$$

where nc is the number of Gaussians, h_i is the height of the ith Gaussian; for the sake of uniqueness, we set $h_1 = 1$. $\mathbf{x} = (-\cot\theta, \alpha_{SC}, \beta_{SC})^T$ and $\mathbf{x}_i^\circ = (-\cot\theta_i^\circ, \alpha_{SC;i}, \beta_{SC;i}^\circ)^T$ ($i = 1, 2, ..., nc$) are the vectors of the variables $\cot\theta$, α_{SC}, β_{SC} and the vectors of the coordinates of the centers of the Gaussians $\cot\theta_i^\circ$, $\alpha_{SC;i}$, $\beta_{SC;i}^\circ$, respectively, and

$$\mathbf{D}_i = \begin{pmatrix} d_{i;11} & d_{i;12} & d_{i;13} \\ d_{i;21} & d_{i;22} & d_{i;23} \\ d_{i;31} & d_{i;32} & d_{i;33} \end{pmatrix} \qquad i = 1, 2, ..., nc \quad (20)$$

are the symmetric dispersion matrices of the Gaussians, and the superscript T denotes the transpose of a matrix or a vector.

Because these local energy terms usually have many minima, it is reasonable to choose a deformation in which the minima coalesce gradually into one broad minimum. Let a_{max} denote the value of the deformation parameter for which there is only a single "average" Gaussian in eq 19, "single" because it leads to a single minimum. The parameters $\bar{h}$, $\bar{x}^\circ$, and $\bar{D}$ of this average Gaussian are expressed by eqs 21−23, respectively.

$$\bar{h} = \sum_{i=1}^{nc} h_i/\sqrt{\det \mathbf{D}_i} \quad (21)$$

$$\bar{\mathbf{x}}^\circ = \sum_{i=1}^{nc} x_i^\circ h_i/\sqrt{\det \mathbf{D}_i} \quad (22)$$

$$\bar{\mathbf{D}} = \sum_{i=1}^{nc} \mathbf{D}_i h_i/\sqrt{\det \mathbf{D}_i} \quad (23)$$

After imposing the deformation parameter a, the parameters of the deformed Gaussians are expressed as follows:

$$\tilde{h}_i = \alpha\bar{h} + (1 - \alpha)h_i \quad (24)$$

$$\tilde{\mathbf{x}}_i = \alpha\bar{\mathbf{x}} + (1 - \alpha)\mathbf{x}_i \quad (25)$$

$$\tilde{\mathbf{D}}_i = \alpha\bar{\mathbf{D}} + (1 - \alpha)\mathbf{D}_i \quad (26)$$

with $\alpha = a/a_{max}$.

Finally,

$$\tilde{U}_{rot}(\theta, \alpha_{SC}, \beta_{SC}; \mathbf{h}, \mathbf{x}^\circ, \mathbf{D}) = \exp(-\kappa' a)U_{rot}(\theta, \alpha_{SC}, \beta_{SC}; \tilde{\mathbf{h}}, \tilde{\mathbf{x}}^\circ, \tilde{\mathbf{D}}) \quad (27)$$

There are many different ways to achieve a balance between the different types and parameters of deformation for the different energy contributions, and a proper choice is based on the demands of the particular algorithm used. For example, when using an old, one-trajectory approach without any local searches (as described in the Introduction), the number of local minima has to decrease to one, and the remaining minimum has to be the representation of the global minimum on the undeformed energy surface; these requirements are impossible to meet for such a complicated function. For the SCBDBM method, however, there are only two important requirements: (a) the

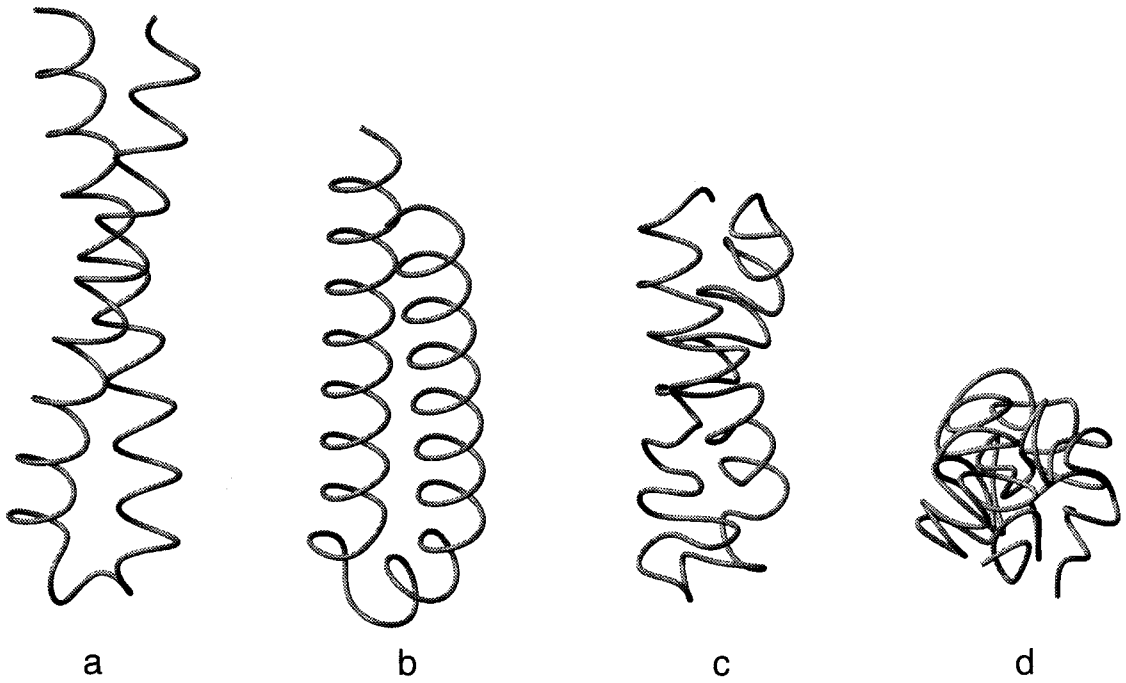

Figure 5. Changes of energy-minimized conformation with increasing deformation for the parameter b of eq 6 equal to 1 in the case of $(Ala)_{70}$. (a) No deformation, (b) small deformation, (c) intermediate deformation, (d) high deformation.

number of minima must decrease significantly when the deformation increases, (b) local minima should merge while the deformation increases, and the minima should disappear geometrically close to one another (i.e., trajectories should merge instead of vanishing, especially for low-energy minima, because otherwise local search would be useless).

There are generally three qualitatively different choices for different deformations based on the value of the parameter b in eq 6. All of them were investigated by using the reversed-reversing procedure (described in section 2.1) starting from a set of preselected low-energy structures of polyalanine $(Ala)_{70}$. This set of structures consisted of a single-helix conformation and a few structures obtained from the single-helix conformation by randomly perturbing selected torsional angles followed by energy minimization. The first choice of b was to use the standard DSM, i.e., set $b = 1$. In this case, the energy-minimized structure gradually collapsed to a coil (see Figure 5) as the deformation increased, because of the dominant long-range deformed Lennard-Jones interactions for a large deformation parameter. In this case, even residues far from one another in space significantly attract one another, while other parts of the deformed UNRES potential are nearly completely flattened. Therefore, the optimal structure is the one that is preferred by deformed Lennard-Jones interactions, i.e., the structure in which the distances between residues are as close to the Lennard-Jones equilibrium distance as possible. The resulting deformation did not satisfy requirement (a) described above, because the number of minima did not decrease significantly for larger deformations; i.e., there were still many available coil conformations. Therefore, this type of deformation cannot be used in the SCBDBM method.

Setting $b < 1$ will worsen this situation, causing even earlier coil formation, because residues will attract each other even at smaller distances, because the Lennard-Jones potential shifts toward zero distance. Thus, a reasonable choice is to set $b > 1$. For $1 < b < 1.5$, however, there is still a coil-like intermediate structure present before the whole polypeptide unfolds to an extended conformation because of the shift of the Lennard-Jones equilibrium distances toward higher values. A coil formation occurs when non-Lennard-Jones contributions to the UNRES energy are already flattened, and they cannot prevent unordered close packing of residues caused by the deformed Lennard-Jones long-range attraction. However, for larger values of the deformation parameter, the Lennard-Jones equilibrium distances shift toward higher values and the structure unfolds.

For $b \geq 2.0$, unfolding occurs smoothly with increasing deformation, because the shift is large enough to counteract the increased attraction (see Figure 6). At the beginning, the secondary structure melts and the tertiary structure with a simplified (deformed) chain conformation remains (Figure 6c). Finally, this structure unfolds to a completely extended conformation (Figure 6d). The number of minima decreases with increasing deformation parameter a in this case, and finally very few extended structures remain, so that requirement (a) is satisfied. However, because the secondary structure vanishes much before the tertiary structure starts to change, the geometrically neighboring structures on the undeformed surface may appear in positions quite far from each other on the intermediately-deformed surface. Therefore, this deformation may not satisfy requirement (b). Indeed, test global optimization runs (both with larger polyalanine chains as well as with protein A) showed that, after an initial fast convergence to reasonably low-energy structures, further improvements in the energy were very slow. The reason for the fast vanishing of the secondary structure is repulsion (for nonzero deformation) between residues located close to each other in the polypeptide chain. When $b > 1$, the position of the equilibrium distances of the Lennard-Jones interactions shift to larger values, but the geometrical constraints of the polypeptide chain allow close residues to move away from each other only by destroying the secondary structure; the potential energy preferences due to torsional and correlational contributions from neighboring residues are overcome by increasing Lennard-Jones repulsion, forcing residues to move as far from each other in space as possible. By contrast, residues that are not close to each other may simply shift away by

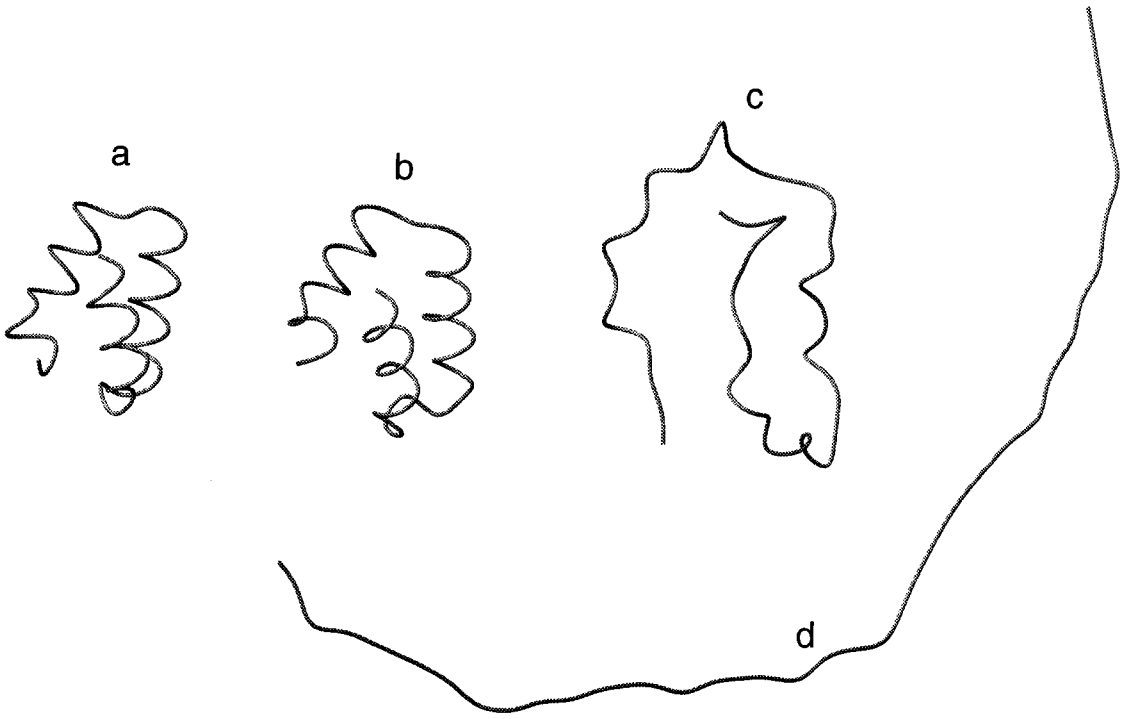

Figure 6. Changes of energy-minimized conformation with increasing deformation for the parameter b of eq 6 equal to 2 in the case of $(Ala)_{70}$. (a) No deformation, (b) small deformation, (c) intermediate deformation, (d) high deformation.

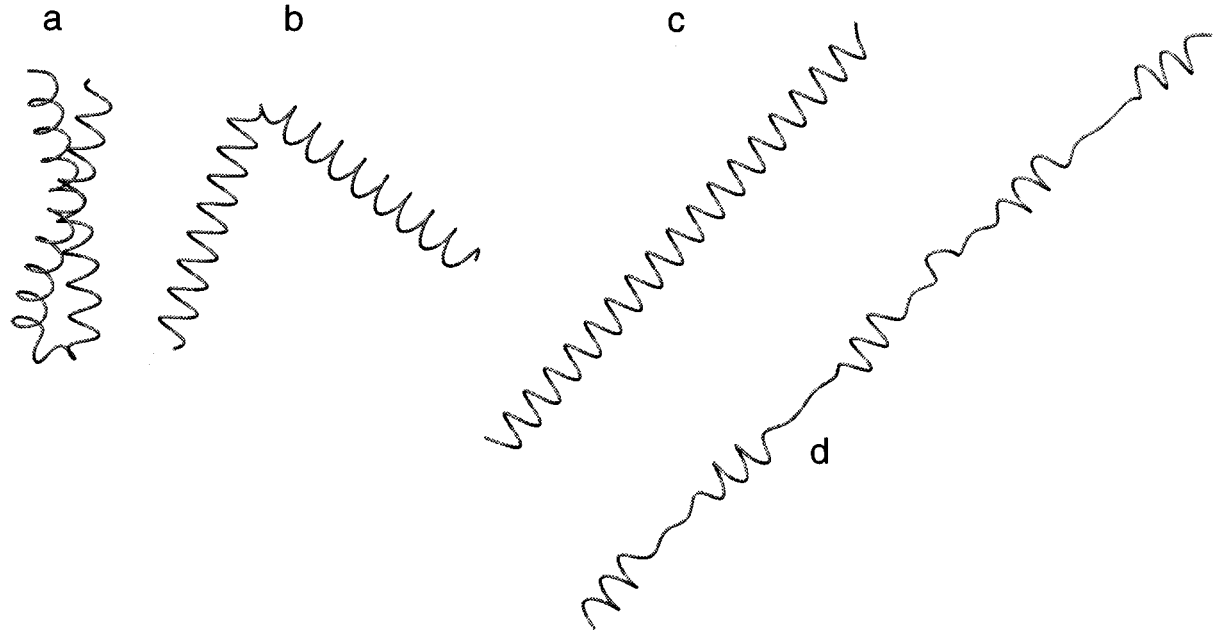

Figure 7. Changes of energy-minimized conformation with increasing of deformation for the parameter b of eq 6 equal to 2 for residues separated by more than 12 residues in a polypeptide chain [$(Ala)_{70}$ in this case] and $b = 1$ otherwise, with a repulsive unfolding term (a/r_{ij}^2) included in the potential. (a) No deformation, (b) small deformation, (c) intermediate deformation, (d) high deformation.

loosening the tertiary structure (unfolding to an extended conformation). This clearly suggests that the parameter b should be different for interactions between residues located close to each other in the chain (closer than 8-12 residues) than for those located far from each other.

The simplest solution is to set the value of the parameter $b = 1$ for all Lennard-Jones interactions between residues closer than 12 residues in the chain and $b = 2$ for all other Lennard-Jones interactions. In this case, the initial conformation unfolds very slowly, and before full unfolding starts to melt the secondary structure. This is a result of the strong flattening of the potential, which causes very weak repulsion. The solution is to increase the value of b further (to 2), and to add a new, repulsive term to the energy function; this term, however, is also calculated only if the interacting residues are separated by more than 12 other residues in a chain. Because this new term should contribute to the interaction between remote parts of the chain, it should also be of a long-range functional form. There are many possible formulas, all of them working similarly; in our approach, we have chosen a/r_{ij}^2. After all adjustments, including the new energy term, the structure unfolds gradually (see Figure 7), while the secondary structure is initially preserved, and finally melts in a highly deformed extended conformation. This behavior allows for a smooth transition of

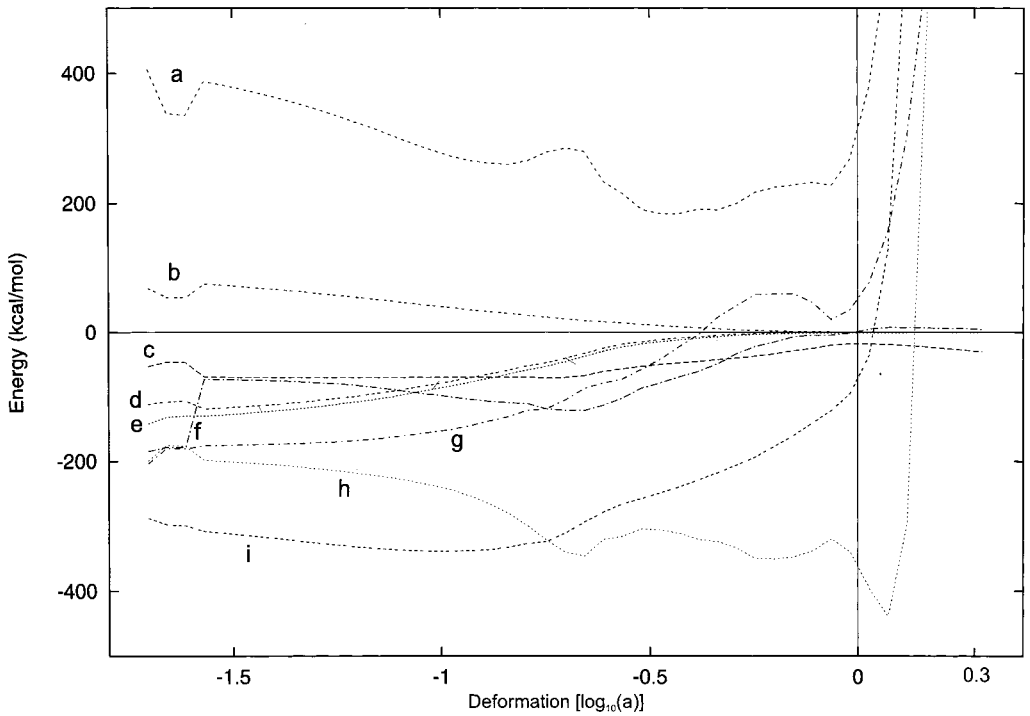

Figure 8. Variation of the energy contributions and the total energy with the deformation parameter a for the Ala_{70} united-residue chain. (a) U_{tor}, (b) $U_{SC_ip_j}$, (c) $U_{p_ip_j}$, orientation-dependent terms, (d) U_{corr}, (e) $U_{p_ip_j}$, distance-dependent terms, (f) U_b, (g) $U_{SC_iSC_j}$, (h) U_{rot}, (i) total potential energy (sum of all contributions).

the geometry during the reversing procedure while satisfying requirements (a) and (b) described above.

A very important feature of the deformation procedure described above is that the values of the parameters (b, and the distance in the chain) have qualitative, rather than quantitative, meaning, i.e., they may change in certain intervals without changing the results.

As mentioned in the Introduction, the deformation should be carried out in such a way that each energy contribution to the total energy will change with the deformation parameter a with approximately the same speed, i.e., the *relative* contributions to the total energy from different terms should remain approximately constant for a particular value of the deformation parameter a (i.e., they should change similarly with the deformation). Otherwise, a situation can arise in which a particular energy term rapidly outweighs the other terms at some value of a, and the value of the deformed function would then depend only on this term. However, for larger values of the deformation parameter, the relative contributions from the different energy terms will change because of their different functional form, and therefore will change in a different way with the deformation. Some contributions simply vanish when the deformation parameter becomes large (e.g., the electrostatic terms; see eq 13); others, such as Lennard-Jones type interactions, become larger, but mostly flat.

A proper calibration is achieved by monitoring the changes of all the energy terms with deformation and choosing an effective *separate* deformation parameter a_i for each of them; a_i is related to the common deformation parameter a by the relation $a_i = c_i a$. Figure 8 illustrates the final effect of this procedure for the energy function considered in the present paper. From this figure it can also be inferred that the most pronounced changes in the energy occur close to $a = 0$. Thus, a should be changed according to a logarithmic and not a linear scale in the reversing and reversed-reversing procedure. As can also be seen, increasing a beyond the value of 2 [$\log(a) = 0.3$] results only in an increase of the total energy, but not in any further changes in the relations between the energy terms. Moreover, no geometrical changes in conformation occur beyond this point. This value should, therefore, be chosen as the first value of a_{max} to start the procedure.

TABLE 1: Results of Global Optimization with SCBDBM Method for Polyalanine Chains

no. of residues	energy (kcal/mol)	energy per residue (kcal/mol)	no. of helical segments	residues in loop regions
10	−30.45	−3.04	1	
20	−73.33	−3.67	1	
30	−116.22	−3.87	1	
40	−159.12	−3.98	1	
50	−202.02	−4.04	1	
60	−244.92	−4.08	1	
70	−288.22	−4.12	2	33−37
80	−334.05	−4.18	3	26−30; 52−56
90	−381.66	−4.24	3	28−32; 58−62
100	−427.87	−4.28	3	32−36; 65−69

3. Results and Discussion

Polyalanine Chains. The SCBDBM method has been applied to find the global minimum for polyalanine chains with length varying from 10 to 100 amino acids, and for the fragment of Staphylococcal protein A consisting of residues 10−55, all in the united-residue representation.

All calculations were carried out using numerical parameters controlling the deformation described in the previous section. The number of trajectories has been chosen as 20 and the number of macroiterations as 2; for four different values of the deformation a, a local search was carried out during the reversing procedure. The maximum deformation parameter a_{max} was 2.0 (with b held constant at the value 2.0), and the deformation parameter a was changed logarithmically during

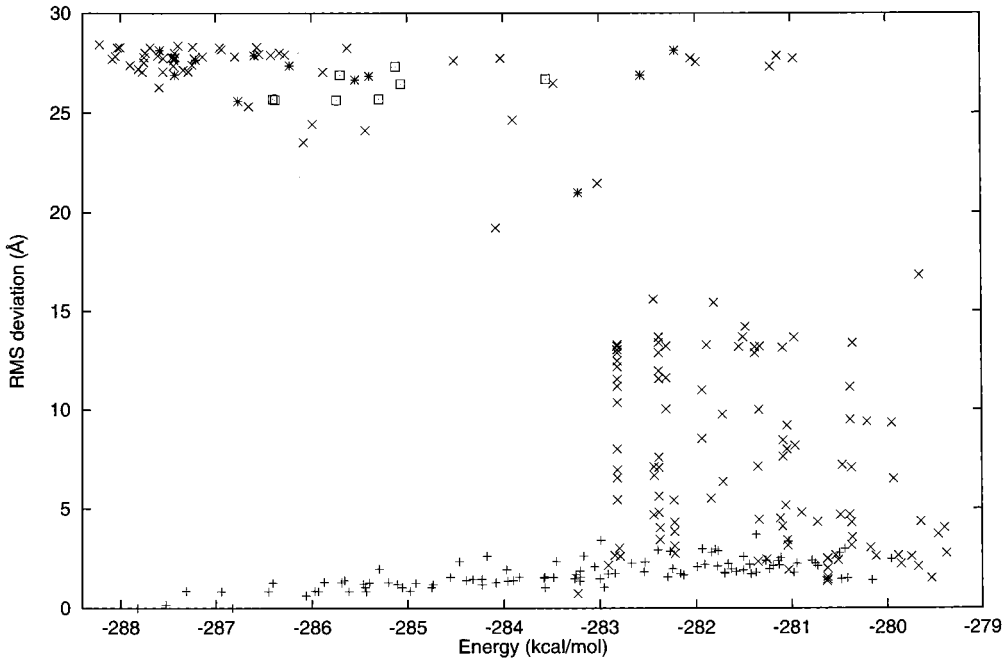

Figure 9. Map showing RMS deviations from the full α-helix versus energies for the 300 lowest energy conformations obtained using the SCBDBM method for $(Ala)_{70}$. (+) Single helix, (×) double helix, (*) triple helix, (□) other (mostly disorganized).

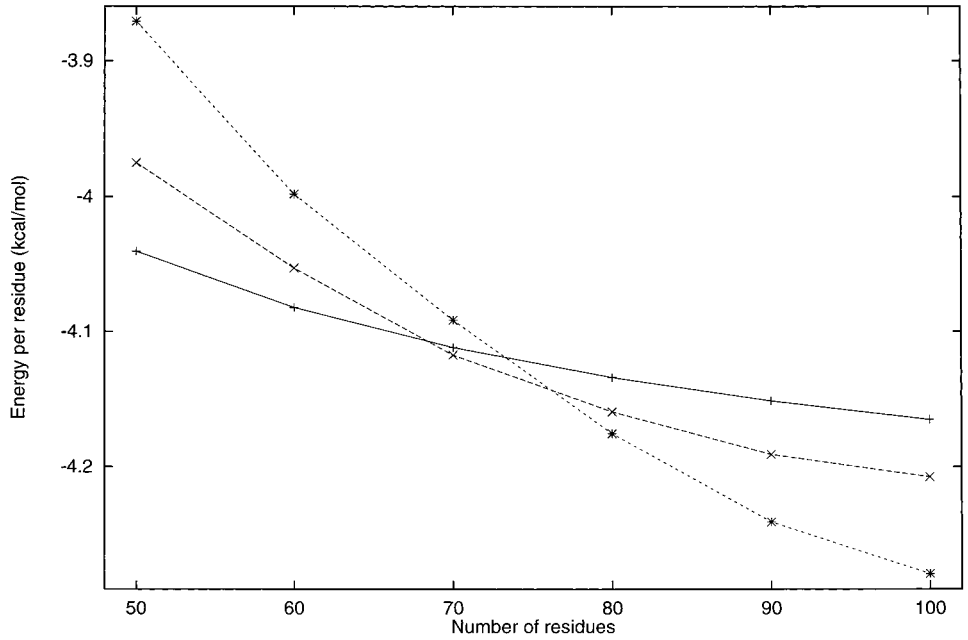

Figure 10. Dependence of the energy per residue on the polyalanine chain length for the lowest energy single, double, and triple helix structures. (+) Single helix, (×) double helix, (*) triple helix.

the reversing procedure. The maximum number of iterations within one macroiteration (one iteration being a pair of reversed and reversed-reversing procedures) was set to 10. The program has been parallelized on a coarse grain level, i.e., local searches and trajectory tracking were carried out in parallel. All calculations were carried out on an IBM SP2 supercomputer at the Cornell Theory Center, and the resources consumed varied greatly with the chain length. For $(Ala)_{70}$, full global optimization required 36 h, using 31 processors of the SP2 supercomputer.

It might be expected that the most stable structure for small and medium-size polyalanine chains is the single α-helix. However, for longer chains, the energy increase caused by breaking the single helix could be compensated by an even larger energy decrease because of helix−helix and hydrophobic interactions (when the resulting helices are properly packed), and therefore a two-helix structure would become the most stable conformation.[34]

Our results confirm this observation. For $(Ala)_n$, $n = 10, ..., 60$, the most stable structure found by global optimization is the single α-helix; all other structures, including bent ones, are higher in energy. The first chainlength at which the double-helix structure with a hairpin-like bend in the middle of the chain was the lowest-energy one is $(Ala)_{70}$. For longer chains, starting at $(Ala)_{80}$, the most stable structure is the three-helix bundle (triple-helix structure). These results are shown in Table 1, where the energy, the number of helical segments and the residues in loops are reported. In all the lowest-energy, structures, there are always five residues involved in forming each loop.

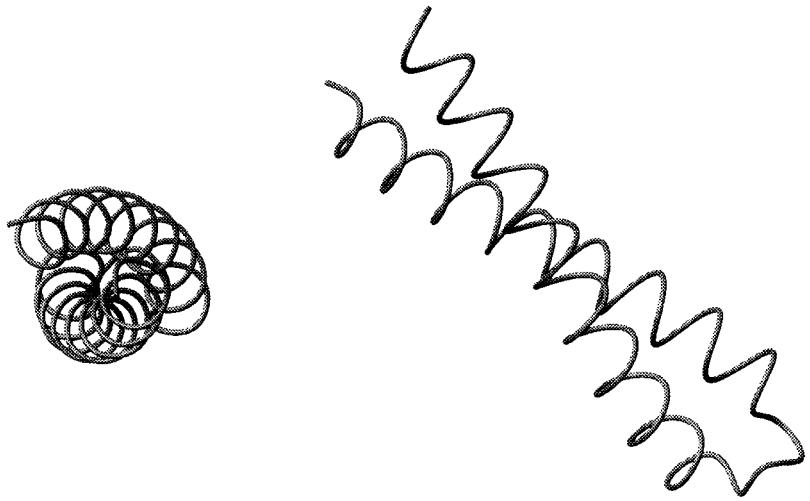

Figure 11. Top and side view of the lowest energy structure for $(Ala)_{70}$ showing the tight packing of two helical parts forming a coiled coil.

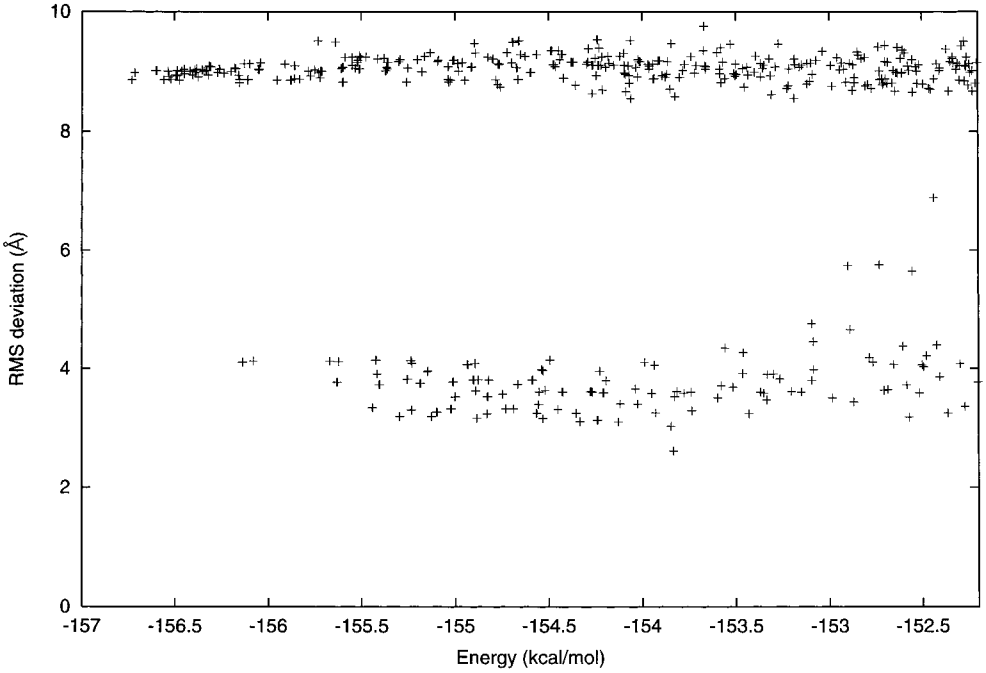

Figure 12. Map showing RMS deviations from the native structure versus energies for 300 lowest energy conformations obtained using the SCBDBM method for the fragment of Staphylococcal protein A consisting of residues 10−55.

A very interesting feature of the SCBDBM method is that it provides not only one, hopefully the lowest energy, structure, but also accumulates a large set of conformations obtained during the macroiterations. This is especially important when differences in energy between close structures are small, and the conformations are grouped into families. This behavior can be described by analyzing a map of root-mean-square deviations (RMSD) from a reference structure versus energies for the whole set of conformations. Figure 9 shows an RMSD-energy map for a set of 300 conformations of $(Ala)_{70}$. The straight α-helix has been chosen as the reference for measuring the RMSD deviations. There are clearly three main distinguishable groups of structures in this map. The first group lies at the bottom of the map where the structures have low values of the RMS deviation, and the energies vary greatly from very low to very high. This group consists of conformations of single helices, possibly disorganized (melted) at the C- or N-terminus. The lowest-energy member of this family is the full, straight α-helix.

The second family is located just above the single-helix set of conformations, having high energies and low-to-moderate RMS deviations. These are double-helix structures without proper loops formed, i.e., the helix is broken in the middle of the chain leaving some disorganized part, but two helical fragments are usually far from one another and, therefore, not interacting with each other. This results in high energy because the energy increases because helix breaking is not compensated. Additionally, the terminal parts of these conformations may also be disorganized.

The symbols representing the third group of structures form a triangle in the upper left corner of the map (see Figure 9); it contains structures of low energies (including the lowest energy structure) and high RMS deviations from the straight α -helix. These conformations are "true" double-helix and triple-helix structures, in which the helical parts are properly connected by loops and interact strongly with each other. One might also expect to see a large number of completely disorganized structures, which are essentially not present on the map in Figure 9. Such structures were present during the early stages of the

calculations (long before self consistency was reached) but then were replaced by lower-energy (and better organized) conformations.

Figure 9 clearly shows that, for all values of the energy, there is a nearly equal density of conformations, without any significant gaps, both for low, as well as for high-RMS deviations.

The topology of the lowest-energy conformation for a given chain length changes qualitatively with the length of the chain from a single helix (10−60 residues) to a three-helix bundle (80 residues and above). To check this preference quantitatively the lowest-energy structure belonging to each class (single helix, double helix or triple helix) was identified for each chain length. These results are shown in Figure 10 as the dependence of the energy per residue of a structure from each of the three classes on the number of amino acids. The first change of energy order of structures (where the double helix becomes the preferable structure) occurs just before, but very close to $(Ala)_{70}$, and the second change (where the triple helix becomes the preferable structure) occurs between $(Ala)_{70}$ and $(Ala)_{80}$. For polyalanine chains longer than 80 residues, the differences between the energies (per residue) of the three classes of structures rapidly become large.

In all the lowest-energy structures with a bent helix, the helical segments are packed very close to one another. As shown in Figure 11, using as an example the lowest energy structure of $(Ala)_{70}$, the two helical segments are slightly deformed from a straight helix to maximize the helix-to-helix surface contact, resulting in a two-helical coiled-coil structure.

Protein A

The SCBDBM method has also been applied to the fragment of Staphylococcal protein A consisting of residues 10−55. The numerical parameters for the SCBDBM method were the same as those used in polyalanine chains, and full global optimization required 25 h; using 31 processors of the IBM SP2 supercomputer. All structures found could be divided into two families (see Figure 12), the first with RMS deviations from the native structure in the range of 2−4 Å showing the native three-helix bundle fold and the second with RMS deviation between 8 and 10 Å being a mirror image fold. Both families spread horizontally on the RMS deviation vs. energy map. The lowest energy structure found belongs to a mirror-image family and has an energy of −156.73 kcal/mol. These results are in agreement with those obtained by the CSA method for the same protein;[6] however, comparison with a similar map in reference[6] shows, that the native-fold family of structures has not been properly searched by the SCBDBM method, and some low-energy low-RMSD structures are missing. The RMSD map from ref 6 shows the native family of structures extended to lower energies than those in Figure 12, with the number of conformations in the native family similar to the number of conformations in the mirror-image family. The structure closest to the native structure has an RMSD of 2.61 Å and an energy of −153.83 kcal/mol. The reason for this insufficient search is that structures are selected for local search and tracking based only on a value of potential energy. Therefore, these structures may all be members of the same family. When lower energy structures from the second family are found first, then the search would be focused mostly on the second family, resulting in underrepresentation of members of the first family. This problem applies only to local searches carried out for relatively small deformations because all important families are always located early in highly deformed surfaces; they are not, however, searched later with equal attention. In the case of protein A, there are two energetically close (but distant in RMSD) families of structures, and the mirror-image family was found first by the SCBDBM method. The global minimum belongs to the native family of structures, which unfortunately was explored much less extensively than the mirror-image family because of the underrepresentation described above. This may be rectified in the future by adding an additional criterion to choose (for local searches) not only minima with the lowest energy, but also those with higher RMS deviations from each other; the same modification should be applied in the procedure deciding which structure to remove from the set of stored structures described in section 2.1.

Acknowledgment. This research was supported by grants from the National Science Foundation (Grant MCB95-13167) and from the National Institutes of Health (Grant GM-14312). The computations in this work were carried out at the Cornell Theory Center which is funded in part by NSF, New York State, the NIH National Center for Research Resources (Contract P41RR-04293), the IBM Corporation, and the CTC Corporate Partnership Program.

References and Notes

(1) Scheraga, H. A. *Int. J. Quant. Chem.* **1992**, *42*, 1529−1536.
(2) Vásquez, M.; Némethy, G.; Scheraga, H. A. *Chem. Rev.* **1994**, *94*, 2183−2239.
(3) Scheraga, H. A. *Biophys. Chem.* **1996**, *59*, 329−339.
(4) Lee, J.; Scheraga, H. A.; Rackovsky, S. *J. Comput. Chem.* **1997**, *18*, 1222−1232.
(5) Lee, J.; Scheraga, H. A.; Rackovsky, S. *Biopolymers* **1998**, *46*, 103−115.
(6) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025−2030.
(7) Piela, L.; Kostrowicki, J.; Scheraga, H. A. *J. Phys. Chem.* **1989**, *93*, 3339−3346.
(8) Kostrowicki, J.; Piela, L.; Cherayil, B. J.; Scheraga, H. A. *J. Phys. Chem.* **1991**, *95*, 4113−4119.
(9) Kostrowicki, J.; Scheraga, H. A. *J. Phys. Chem.* **1992**, *96*, 7442−7449.
(10) Pillardy, J.; Olszewski, K. A.; Piela, L. *J. Phys. Chem.* **1992**, *96*, 4337−4341.
(11) Pillardy, J.; Olszewski, K. A.; Piela, L. *J. Mol. Struct. (Theochem)* **1992**, *270*, 277−285.
(12) Amara, P.; Hsu, D.; Straub, J. E. *J. Phys. Chem.* **1993**, *97*, 6715−6721.
(13) Orešič, M.; Shalloway, D. *J. Chem. Phys.* **1994**, *101*, 9844−9857.
(14) Pillardy, J.; Piela, L. *J. Phys. Chem.* **1995**, *99*, 11805−11812.
(15) Pappu, R. V.; Hart, R. K.; Ponder, J. W. *J. Phys. Chem. B* **1998**, *102*, 9725−9742.
(16) Wawak, R. J.; Gibson, K. D.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 1743−1746.
(17) Wawak, R. J.; Pillardy, J.; Liwo, A.; Gibson, K. D.; Scheraga, H. A. *J. Phys. Chem.* **1998**, *102*, 2904−2918.
(18) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Science* **1993**, *2*, 1697−1714.
(19) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Science* **1993**, *2*, 1715−1731.
(20) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849−873.
(21) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874−887.
(22) Liwo, A.; Kaźmierkiewicz, R.; Czaplewski, C.; Groth, M.; Oldziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Comput. Chem.* **1998**, *19*, 259−276.
(23) Liwo, A.; Pillardy, J.; Kaźmierkiewicz, R.; Wawak, R. J.; Groth, M.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A. *Theor. Chem. Acc.* **1999**, *101*, 16−20.

(24) Liwo, A.; Oldziej, S.; Ciarkowski, J.; Kupryszewski, G.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Prot. Chem.* **1994**, *13*, 375−380.
(25) Crippen, G. M.; Scheraga, H. A. *Arch. Biochem. Biophys.* **1971**, *144*, 462−466.
(26) Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800−2806.
(27) Wales, D. J. *J. Chem. Phys.* **1989**, *91*, 7002−7010.
(28) Wales, D. J. *J. Chem. Soc., Faraday Trans.* **1990**, *86*, 3505−3517.
(29) Tsai, C. J.; Jordan, K. D. *J. Phys. Chem.* **1993**, *97*, 11227−11237.
(30) Li, Z.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611−6615.
(31) Li, Z.; Scheraga, H. A. *J. Molec. Struct. (Theochem)* **1988**, *179*, 333−352.
(32) Godzik, A.; Koliński, A.; Skolnick, J. *J. Comput.-Aided Mol. Design* **1993**, *7*, 397−438.
(33) Piela, L.; Scheraga, H. A. *Biopolymers* **1987**, *26*, S33−S58.
(34) Silverman, D. N.; Scheraga, H. A. *Arch. Biochem. Biophys.* **1972**, *153*, 449−456.