Optimization of the UNRES Force Field by Hierarchical Design of the Potential-Energy Landscape. 2. Off-Lattice Tests of the Method with Single Proteins

Stanisław Ołdziej,[†,§] Adam Liwo,[†,§] Cezary Czaplewski,[†,§] Jarosław Pillardy,[†,‡] and Harold A. Scheraga*[,†]

Baker Laboratory of Chemistry and Chemical Biology and Cornell Theory Center, Cornell University, Ithaca, New York 14853-1301, and Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland

Received: April 30, 2004; In Final Form: August 12, 2004

We describe the application of our recently proposed method of hierarchical optimization of the protein energy landscape to optimize our off-lattice united-residue (UNRES) force field using single training proteins. First, the IgG-binding domain from streptococcal protein G (PDB code 1IGD) was treated; earlier attempts to use this protein to optimize the force field by optimizing the energy gap and *Z* score between the nativelike and non-native structures failed. The structure of this protein consists of an N-terminal antiparallel β-hairpin, a middle α-helix, and a C-terminal antiparallel β-hairpin, these elements being referred to as β_1, α_2, and β_3, respectively, with the two hairpins forming a parallel β-sheet packed against the α-helix. In our earlier study, one of these elements was assumed to form at level 1, two at level 2, and three at level 3, and higher levels corresponded to the proper packing of two or more elements. This approach resulted in a structure with the wrong packing of the β-sheet, and attempts at further optimization failed. We therefore tried a hierarchy scheme that corresponds to the sequence of folding events deduced from NMR experiments. In this scheme, level 1 corresponds to structures with either β_3 or α_2, level 2 to structures with both β_3 and α_2, level 3 to structures with β_3, α_2, and the N-terminal strand packed against α_2 (with β_1 still not fully formed), and level 4 to structures with β_1, α_2, and β_3, with β_3 being packed to β_1, which also implies the packing of β_1 and β_3 against α_2. This optimization was successful and resulted in a reasonably transferable force field that led to well-foldable proteins. This corroborates the conclusion from our model on-lattice studies (Liwo, A.; Arłukowicz, P.; Ołdziej, S.; Czaplewski, C.; Makowski, M.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16918) that a proper design of the structural hierarchy is of crucial importance to the foldability with the resulting potential-energy function. Moreover, in the off-lattice approach, the design of the hierarchy also appears to be important to the success of the optimization procedure itself. The next series of calculations was carried out with the LysM domain from the *E. coli* 1E0G ($\alpha + \beta$) protein, which is smaller than 1IGD. In this case, no experimental information about the folding pathway is available; nevertheless, we were able to deduce the appropriate hierarchy by a trial-and-error method. The resulting force field performed worse in tests on $\alpha + \beta$- and β-proteins than that derived on the basis of 1IGD with a correct hierarchy, which suggests that the structure of the 1IGD protein encodes more structure-determining interactions common to all proteins than the 1E0G protein does. For 1E0G, we also attempted to carry out a single energy gap and *Z*-score optimization; this effort resulted in an unsearchable force field. (The nativelike structures could not be found by a global search, although they were the lowest in energy). Technical details of the method, including the maintenance of proper secondary structure and a method to classify structure, are also described.

1. Introduction

Recently,[1] we proposed a novel method for optimizing protein potential-energy functions that is based on a hierarchical design of the potential-energy landscape such that the energy decrease follows the increase of nativelikeness. The structure of each training protein is described in terms of levels. Level 0 contains conformations with no native-structure elements; level 1 contains conformations with a single element of native secondary structure. The subsequent levels contain conformations with gradually more elements of native secondary structure gradually packed and arranged as in the native structure. The composition and arrangement of the levels is termed a structural hierarchy. It is important to note that not only do more nativelike elements appear with increasing level number but the elements also appear in a predetermined order. Preliminary applications[1] to the designed peptide 1FSD (a minimal 28-residue $\alpha + \beta$ motif) and to the third IgG-binding domain from streptococcal protein G (a 61-residue $\alpha + \beta$-protein referred to in the PDB as 1IGD[2])[1,3] showed that the method performs much better than the classical optimization of the energy gap and *Z* score between the nativelike and non-native structures. (The latter failed for 1IGD.)

In an accompanying paper,[4] we reported an extensive investigation of the properties of the method and compared it with energy-gap and *Z*-score optimization (see eqs 1 and 2 of ref 4 for definitions of these quantities) using simplified 12-

[*] Corresponding author. E-mail: has5@cornell.edu. Phone: (607) 255-4034. Fax: (607) 254-4700.

[†] Baker Laboratory of Chemistry and Chemical Biology, Cornell University.

[‡] Cornell Theory Center, Cornell University.

[§] University of Gdańsk.

10.1021/jp0403285 CCC: $27.50
Published on Web 10/02/2004

bead cubic-lattice models of proteins for which all conformations can be enumerated. We found that, even for such simple systems, the optimization of a single energy gap and *Z* score could result in both slow and fast folders for the same attained energy and *Z*-score gaps. Conversely, hierarchical optimization with an appropriately designed hierarchy always resulted in rapid folders, the value of the energy gap mainly affecting the stability of the native structure for thermodynamic reasons. Moreover, an analysis of the energy spectra indicated that successful nonhierarchical optimizations resulted in correspondence between energy ordering and nativelikeness whereas the energy ordering corresponding to optimizations that resulted in slow folders did not correspond to nativelikeness. This suggests that hierarchical optimization is the better method for designing protein energy landscapes. We also found that the design of the hierarchy is vital for the success of the procedure and that the correct hierarchy should follow the folding pathway.

In this paper, we apply the method to optimize our off-lattice UNRES potential-energy functions for two $\alpha + \beta$-proteins: 1IGD and the LysM domain from *E. coli* (PDB code 1E0G[5]). The sequence of folding events for 1IGD is known from experiment, but that for 1E0G is not. On the basis of the first example, we demonstrate that an appropriate design of the hierarchy of structure formation is critical for the success of the method and for the performance of the resulting force field. On the basis of 1E0G as an example, we demonstrate that it is possible to deduce the hierarchy even if there is no experimental information regarding the sequence of the folding events. In an accompanying paper,[6] we report the application of the approach to the simultaneous optimization of the force field using several test proteins.

2. Methods

2.1. UNRES Force Field. Because the algorithm was applied to optimize the UNRES force field derived in our laboratory, in this section we describe briefly the UNRES model of polypeptide chains and the corresponding force field. In the UNRES model,[1,3,7–15] a polypeptide chain is represented by a sequence of α-carbon (C^α) atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p). Each united peptide group is located in the middle between two consecutive α-carbons. Only these united peptide groups and the united side chains serve as interaction sites, the α-carbons serving only to define the chain geometry, as shown in Figure 1. All virtual bond lengths (i.e., $C^\alpha-C^\alpha$ and $C^\alpha-$SC) are fixed; the distance between neighboring C^α atoms is 3.8 Å corresponding to trans peptide groups, but the side-chain angles (α_{SC} and β_{SC}) and virtual-bond (θ) and dihedral (γ) angles can vary. The UNRES force field has been derived as a restricted free energy (RFE) function of an all-atom polypeptide chain plus the surrounding solvent, where the all-atom energy function is averaged over the degrees of freedom that are lost when passing from the all-atom to the simplified system (i.e., the degrees of freedom of the solvent, the dihedral angles χ for rotation about the bonds in the side chains, and the torsional angles λ for rotation of the peptide groups about the $C^\alpha\cdots C^\alpha$ virtual bonds).[11,12] The RFE is further decomposed into factors coming from interactions within and between a given number of united interaction sites.[12] An expansion of the factors into generalized Kubo cumulants[16] enables us to derive approximate analytical expressions for the respective terms,[11,12] including the multibody or correlation terms that are derived in other force fields from structural databases or on a heuristic basis.[17] The theoretical basis of the force field is described in detail in our earlier paper.[12]

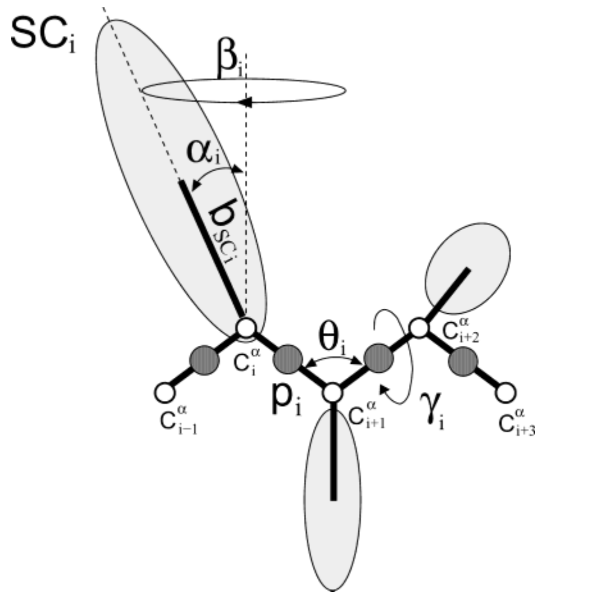

Figure 1. UNRES model of polypeptide chains. The interaction sites are side-chain centroids of different sizes (SC) and the peptide-bond centers (*p*) indicated by shaded circles, whereas the α-carbon atoms (small empty circles) are introduced only to assist in defining the geometry. The virtual $C^\alpha-C^\alpha$ 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}, and a 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 energy of the virtual-bond chain is expressed by eq 1.

$$U = \sum_{i<j} U_{SC_iSC_j} + w_{SCp}\sum_{i\neq j} U_{SC_ip_j} + w_{el}\sum_{i<j-1} U_{p_ip_j} + w_{tor}\sum_i U_{tor}(\gamma_i) + w_{tord}\sum_i U_{tord}(\gamma_i, \gamma_{i+1}) + w_b\sum_i U_b(\theta_i) + w_{rot}\sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i}) + w^{(3)}_{corr}U^{(3)}_{corr} + w^{(4)}_{corr}U^{(4)}_{corr} + w^{(5)}_{corr}U^{(5)}_{corr} + w^{(6)}_{corr}U^{(6)}_{corr} + w^{(3)}_{turn}U^{(3)}_{turn} + w^{(4)}_{turn}U^{(4)}_{turn} + w^{(6)}_{turn}U^{(6)}_{turn} \quad (1)$$

The term $U_{SC_iSC_j}$ represents the mean free energy of the hydrophobic (hydrophilic) interactions between the side chains, which implicitly contains the contributions from the interactions of the side chain with the solvent. The term $U_{SC_ip_j}$ denotes the excluded-volume potential of the side-chain−peptide-group interactions. The peptide-group interaction potential ($U_{p_ip_j}$) accounts mainly for the electrostatic interactions (i.e., the tendency to form backbone hydrogen bonds) between peptide groups p_i and p_j. U_{tor}, U_{tord}, U_b, and U_{rot} are the virtual-dihedral angle torsional terms, virtual-bond dihedral angle double-torsional terms, virtual-bond angle bending terms, and side-chain rotamer terms; these terms account for the local propensities of the polypeptide chain. The terms $U^{(m)}_{corr}$ represent correlation or multibody contributions from the coupling between backbone−local and backbone−electrostatic interactions, and the terms $U^{(m)}_{turn}$ are correlation contributions involving *m* consecutive peptide groups; they are, therefore, termed turn contributions. The correlation contributions were derived[11,12] from a generalized-cumulant expansion[16] of the restricted free energy (RFE) of the system consisting of the polypeptide chain and the surrounding solvent. The multibody terms are indispensable for the reproduction of regular α-helical and β-sheet structures.

The internal parameters of $U_{\mathrm{p}_i\mathrm{p}_j}$, U_{tor}, U_{tord}, $U_{\mathrm{corr}}^{(m)}$, and $U_{\mathrm{turn}}^{(m)}$ were recently derived by fitting the analytical expressions to the RFE surfaces of model systems computed at the MP2/6-31G** ab initio level,[3,15] and the parameters of $U_{\mathrm{SC}_i\mathrm{SC}_j}$, $U_{\mathrm{SC}_i\mathrm{p}_j}$, U_{b}, and U_{rot} were derived by fitting the calculated distribution functions to those determined from the PDB.[10] The *w*'s are the weights of the energy terms, and they can be determined (together with the parameters within each cumulant term) only by optimizing the potential-energy function, which is the subject of our present work.

The UNRES force field, together with the conformational space annealing (CSA)[18−21] global-optimization method developed in our laboratory, is now able to predict the structures of proteins containing both α-helical and β-sheet structures with a reasonable degree of accuracy, as assessed by tests on model proteins[22] as well as in the CASP3,[23] CASP4,[22] and CASP5[24] blind prediction experiments.

2.2. Hierarchical Optimization Algorithm for Off-Lattice Calculations. On the basis of the foundations of the methods given in our earlier work,[1] we developed the following hierarchical algorithm for off-lattice potential-function optimization:

(1) Define the elementary fragments of the molecule. These are usually secondary-structure elements. (See section 2.3.1.)

(2) Define the hierarchy levels of the protein in terms of elementary fragments, which are identified as described in section 2.3.1. In our previous work,[1] we identified the levels with families of structures containing an increasing number of elementary fragments; however, in this work we demonstrate that the levels should follow the experimentally observed folding pathway.

(3) Carry out a global conformational search of the protein with current parameters of the energy function. Save the structures obtained at intermediate stages of the search (i.e., after predefined numbers of energy minimizations) and the final structures; these sets of structures are termed batches. The purpose of introducing various batches is to keep track of the process of simulated folding, and the batches can be identified with conformational ensembles from different points of the folding pathway. In this work, we use the CSA method[18−21] and save the batches after 8000, 16 000, and 64 000 energy minimizations for proteins with a size of 40−70 residues; this can, to some extent, be identified with the early, the intermediate, and the final folding stage. It should be noted that in the current work we consider only the free-energy relations within the batches and we postulate that within a given batch (a point on the folding pathway) the free energy decreases with nativelikeness. We do not consider the relations between the free energies of different batches. The free energy can be predicted to grow initially as folding progresses because of the presence of a free-energy barrier to folding. It would be possible, in principle, to include the free-energy relations between batches in optimization and use the experimental barriers as constraints. However, because we are using the CSA method, which does not provide canonical ensembles, we leave this issue to our future work.

4. Assign conformations of all batches obtained during the search to structural levels, as described in section 2.3.1.

5. For all batches and all structural classes (i.e., α, β, turn, etc.), carry out a cluster analysis of the sets of conformations (for each class and each batch separately). Form new sets from the leading conformations (those with the lowest energy) of the families. Add the new sets to the sets of conformations accumulated from the previous cycles. (Again, each batch and each class is considered separately.)

6. Check whether the free-energy relationships (eqs 2 and 3) between structural levels are satisfied. If they are, then the procedure ends. If not, then adjust the parameters of the potential-energy function to satisfy the following relationships between configurational free energies of the subsets of the database of conformations:

$$F_j^{(b)}(\beta) - F_i^{(b)}(\beta) \leq -\Delta_{ij}^{(b)}(\beta),\ i \in \{0, 1, ..., n_b\},\quad j \in \{0, 1, ..., n_b\},\ j \neq i \tag{2}$$

$$-\Delta_{IJ}^{(b)}(\beta) \leq \tilde{F}_I^{(b)}(\beta) - \tilde{F}_J^{(b)}(\beta) \leq \Delta_{IJ}^{(b)}(\beta),\quad (I, J) \in \{(I, J)^{(\mathrm{b})}\},\ J \neq I \tag{3}$$

where $F_i^{(b)}(\beta)$ is the configurational free energy of the ensemble consisting of the n_b conformations of structural class *i* in batch *b*, $\tilde{F}_I^{(b)}(\beta)$ is the configurational free energy of the conformations within any structural class containing nativelike fragment *I* in batch *b* at inverse temperature β, and the Δ's are target gaps between the free energies. The gaps $\Delta_{IJ}(\beta)$ should be on the order of a few kcal/mol; we use 1 or 2 kcal/mol. The gaps $\Delta_{ij}(\beta)$, on which structural levels are separated, are from several to several tens of kcal/mol. In our earlier work,[1] we always set $j = i + 1$ (i.e., the free-energy gaps were defined between consecutive levels). However, we realized that it is often convenient to set energy gaps between nonconsecutive levels. Suppose, for example, that a protein has an α-helix (α) and an β-hairpin (β). We might define the first structural level as a family of structures containing α or β and separate it from structures without any native-structure elements by a single energy gap. However, we might also want to set separate energy gaps between the non-native structures and structures with α as well as the non-native structures and structures with β. The free energies of eqs 2 and 3 are defined by eqs 4 and eq 5, respectively:

$$F_i^{(b)}(\beta) = -\frac{1}{\beta^{(b)}} \ln Z_i^{(b)}(\beta) \approx -\frac{1}{\beta^{(b)}} \ln \sum_{k \in \{i^{(b)}\}} \exp(-\beta^{(b)} E_k^{(b)}) \tag{4}$$

$$\tilde{F}_I^{(b)}(\beta) = -\frac{1}{\beta^{(b)}} \ln Z_I^{(b)}(\beta) \approx -\frac{1}{\beta^{(b)}} \ln \sum_{k \in \{I^{(b)}\}} \exp(-\beta^{(b)} E_k^{(b)}) \tag{5}$$

with $\{i^{(b)}\}$ and $\{I^{(b)}\}$ denoting the set of conformations of the *i*th structural level and that containing nativelike fragment *I*, respectively, in batch *b*, and *Z* denoting a configurational integral that is approximated by the sum over conformations. β can be identified with $^1/_{RT}$, *T* being the absolute temperature, or can be treated as a parameter of the method. With large β (low temperature), eqs 2 and 3 approach energy gaps, as in eq 1 of an accompanying paper.[4] The purpose of introducing the requirements in eq 2 is to push conformations with higher nativelikeness lower and lower in free energy, whereas the requirements in eq 3 cause all structural segments to be found with comparable probability. For example, for 1IGD, which contains an α-helix between two β-hairpins, we require there to be approximately the same number of structures with the middle α-helix and with the C-terminal β-hairpin summed over all levels. The lower-case index *i* in eq 2 runs through structural levels, and the upper-case indices *I* and *J* sum through structures with a given secondary-structure element.

The values of β (0.1, 0.2, 0.5, 1.0, and 2.0) were chosen on the basis of experience, increasing with maturing stages of CSA and varying depending on the batch structures. In principle, molecular dynamics or Monte Carlo simulations should run at different temperatures to reach equilibrium at each temperature to provide batches from structures obtained at each temperature after equilibration. This would simulate thermal denaturation/renaturation and provide thermodynamic data (e.g., the free-energy gaps at each characteristic temperature corresponding to the partial stages of folding) that could be inserted into eq 2 as the values of Δ_i. However, for efficiency, we use CSA to sample the conformational space. Although the advancement of CSA corresponds to the degree of thermal renaturation, this correspondence is not clear; hence the β's are treated as parameters to be selected on the basis of experience.

To accomplish the task of step 6, we minimize the following target function:

$$\Phi = \sum_b \sum_\beta \sum_{ij} w_i g[F_i^{(b)}(\beta) - F_j^{(b)}(\beta); -\infty, -\Delta_{ij}^{(b)}(\beta)] + \sum_{I,J \in \{(I,J)^{(b)}\}} w_{IJ} g[\tilde{F}_I^{(b)}(\beta) - \tilde{F}_J^{(b)}(\beta); -\Delta_{IJ}^{(b)}(\beta), \Delta_{IJ}^{(b)}(\beta)] + \Psi \tag{6}$$

with

$$g(x; x_{\min}, x_{\max}) = \begin{cases} {}^1/_4(x - x_{\min})^4 & x < x_{\min} \\ {}^1/_4(x - x_{\max})^4 & x > x_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and the w's being the weights of the respective terms in the target function. (It should be noted that they are different from the energy-term weights of eq 1.) The F's are defined by eqs 4 and 5, respectively, the Δ's are defined in the text following eqs 2 and 3, respectively, and Ψ is a penalty function to maintain the correct secondary structure as described in section 2.4. The components of the first sum in eq 6 correspond to the inequality relations defined by eq 2, and those of the second sum correspond to the inequality relations defined by eq 3. Each component is zero if the corresponding inequality (eqs 2 and 3) is satisfied. If Φ is zero, then all inequalities are satisfied, which is the case of a perfect solution. A nonzero solution means that the free-energy differences exceed the target gaps Δ (i.e., there is a compromise between contradictory requirements to satisfy the respective inequalities). By assigning different w's to different free-energy differences in eq 6, we can guide the optimization in favor of selected inequalities; however, at present, we usually set all of the w's to the same value. The adjustable parameters are, first of all, the energy-term weights in eq 1 and also other parameters of the UNRES energy function. In this work, we optimized the coefficients of the correlation terms and some of the parameters of the potential of the SC−SC interaction (see Results) in addition to optimizing energy-term weights. Minimization of the target function (eq 6) is carried out by the secant unconstrained minimization solver (SUMSL) method.[25]

It should be noted that Z-score differences can also be set between consecutive levels as in our on-lattice studies described in an accompanying paper (see eq 7 of ref 4). However, trial calculations showed that, with CSA used as a method of decoy generation, introducing Z-score differences does not make a qualitative difference; therefore, we did not include the Z score in the target function defined by eq 6.

(7) Iterate steps 3−6. Each iteration is termed a cycle.

The procedures developed to carry out the individual steps of the algorithm are described in the following subsections.

2.3. Defining Elementary Fragments and Evaluation of the Nativelikeness of Conformations. *2.3.1. Defining Elementary Fragments.* To assign conformations to structural levels, we first divide the native structure of a training protein into elementary fragments. Elementary fragments are those with defined secondary structure; they can consist of contiguous or noncontiguous parts of the chain. In this work, we distinguish the following elementary fragments:

(1) An α-helix. This is a fragment in which (i) all of the virtual-bond dihedral angles γ are within $30° \leq \gamma \leq 60°$ and (ii) every peptide group is in electrostatic contact with its third neighbor. To determine if the peptide groups are in electrostatic contact, we use the criterion developed in our earlier work:[7] two peptide groups are considered to be in electrostatic contact if their average electrostatic interaction energy ($U_{p_ip_j}$ in eq 1) is lower than −0.3 kcal/mol. It should be noted that this energy cutoff pertains to the old parametrization of $U_{p_ip_j}$ of ref 8.

(2) A two-stranded antiparallel β-sheet (this also includes a β-hairpin). This is a fragment in which (i) all virtual-bond dihedral angles γ except those at turn residues are greater in absolute value than 90° and (ii) an electrostatic-contact pattern characteristic of an antiparallel β-sheet is observed (i.e., if peptide group i is in electrostatic contact with peptide group j, then peptide group $i + 1$ is in electrostatic contact with peptide group $j - 1$, etc.). This type of element can involve either a contiguous part of the chain (a β-hairpin) or a noncontiguous part.

(3) A two-stranded parallel β-sheet. This is a fragment in which (i) all virtual-bond dihedral angles γ are greater in absolute value than 90° and (ii) an electrostatic-contact pattern characteristic of a parallel β-sheet is observed (i.e., if peptide group i is in electrostatic contact with peptide group j, then peptide group $i + 1$ is in electrostatic contact with peptide group $j + 1$, etc.). This type of fragment always involves two noncontiguous parts of the chain.

(4) A strand. This is a fragment in which (i) all virtual-bond dihedral angles γ are greater in absolute value than 90° and (ii) an electrostatic-contact pattern characteristic of a single chain in a parallel or antiparallel β-sheet is observed.

It should be noted that the above definitions do not exhaust all possibilities; one is free to define other types of structural elements such as, for example, a 3_{10}-helix, a β-helix, or a collagen helix. However, we did not need this in our present work.

According to the philosophy of the hierarchical optimization, we compare the database of conformations created during the course of optimization with the native structure at several levels of organization. First, the database structure is checked for the presence of elementary fragments; second, the packing of elementary fragments is checked, and then the geometry of packed triplets, quadruplets, and so forth is compared to that of the corresponding portions of the native structure. Finally, the whole geometry is compared to that of the native structure.

2.3.2. Comparing Elementary Fragments. The comparison is carried out to satisfy two criteria. First, the secondary structure of the part of the sequence in the conformation considered is compared to that of the native structure. The secondary structure is considered to match if at least 70% of the chain has the same secondary structure as in the native conformation. In the case of a two-stranded β-sheet elementary fragment, this means that at least 70% of the fragment is a strand (according to the definition given earlier in this section). Second, the native-contact pattern and geometry in terms of virtual-bond dihedral angles γ are checked as follows:

(1) The number of contacts between the peptide groups in the compared structures matching the native peptide-group contacts is computed. If this number is greater than 70% of the native contacts, then the fragment is considered to have the native peptide-group contact pattern. Shifting the sequence by ±3 residues is allowed to obtain a match. For example, in the case of a β-hairpin, such a shift corresponds to shifting the position of the β-turn.

(2) The average difference between the virtual-bond dihedral angles γ is computed between the native fragment and the corresponding fragment of the compared structure. If the difference is less than the cutoff value (45° for α-helices and 60° for the β-sheets), then the fragment is considered to be native with regard to the differences in the virtual-bond dihedral angles.

If conditions 1 and 2 are satisfied, then the fragment is considered to be native with regard to contact pattern and geometry; otherwise, it is considered to be non-native. The inclusion of condition 2 is necessary to prevent, for example, the consideration of left-handed helices, which have exactly the same peptide-group contact pattern as right-handed α-helices.

2.3.3. Comparing the Contact Pattern and Geometry of Pairs of Elementary Fragments. The packing of elementary fragments defines the assembly of larger elements of secondary structure (e.g., β-sheets) and supersecondary structure (e.g., helix-turn-helix motifs). The packing of fragments in a conformation as in the native structure is compared as follows:

(1) The number of peptide-group contacts (for β-strand packing) or the number of side-chain contacts (for helix-to-helix and helix-to-strand packing) corresponding to the native contact between a given pair of fragments is computed. If it is greater than 70% of the native contacts, then the packing is considered to be native. As in the case of elementary fragments, a sequence shift by up to ±3 residues is allowed to achieve a match.

2. The rmsd of the segment consisting of the compared pair of fragments is computed. If it is less than the threshold value (we assumed 0.1 Å per residue), then the segment is considered to be geometrically conformable with the native segment.

If conditions 1 and 2 hold, then the packing is considered to be nativelike; otherwise, it is considered to be non-native.

To determine if two side chains are in contact, we recently[26] developed a contact function that depends on both the distance and orientation of the side chains. This contact function is based on the Gay−Berne potential[27] of side-chain interactions used in UNRES.[9] The basis of the derivation was to achieve correspondence between contact assessment from the new contact function and the presence of a van der Waals contact between at least one pair of non-hydrogen atoms in terms of the minimum of the sum of false positives (when the new contact function indicates contact between the side chains but there is no pair of non-hydrogen atoms in van der Waals contact) and the minimum of the sum of false negatives (when the new contact function indicates no contact but there is a contact between non-hydrogen atoms). We used our earlier selected database of 195 nonhomologous high-resolution protein structures[9] to parametrize the new contact function. The maximum sum of false positives and false negatives over all 210 side-chain types was 6%; for comparison, when only the distance between side-chain centers was used, this sum was 15%. The contact function is given by eq 8.

$$C_{ij}(r_{ij}, \omega_{ij}^{(1)}, \omega_{ij}^{(2)}, \omega_{ij}^{(12)}) = -\ln \min\left\{\epsilon_{ij}^{(1)}\epsilon_{ij}^{(2)}\left[\frac{\sigma_{ij}^{\circ}}{r_{ij} - \sigma_{ij} + \sigma_{ij}^{\circ}}\right]^{6}, 10\right\} \tag{8}$$

where r_{ij} is the distance between the centers of side chains i and j, $\omega_{ij}^{(1)}$, $\omega_{ij}^{(2)}$, and $\omega_{ij}^{(12)}$ are the angles that define the orientation of these two side chains (Figure 2 in ref 9), $\epsilon_{ij}^{(1)}$ and $\epsilon_{ij}^{(2)}$ and σ_{ij} are the orientation-dependent components of the well-depth and the effective cross section of the Gay-Berne potential[27] (eqs 6 and 8 in ref 9), respectively, and σ_{ij}° is the reference cross section in the Gay−Berne potential (which is a constant characteristic of the types of side chains). The quantities $\epsilon_{ij}^{(1)}$, $\epsilon_{ij}^{(2)}$, and σ_{ij} depend on the side-chain-pair anisotropies[27] $\chi_{ij}^{(1)}$, $\chi_{ij}^{(2)}$ and $\chi_{ij}^{(12)}$, which are constants dependent on side-chain type. A contact between side chains occurs when C_{ij} is greater than the cutoff distances C_{ij}°, which depend on the types of side chains. Details of the parametrization and statistical evaluation of the new contact function will be published in a separate paper.[26]

2.3.4. Comparing Larger Fragments. For segments consisting of more than two elementary fragments (including the whole molecule), the only criterion for comparison is conformity in the rmsd from the corresponding part of the native structure (point 2 of section 2.3.3). Only those clusters are considered that are composed of such elementary fragments for which each of them makes contacts with at least one other fragment of the cluster. In this work, higher-order fragments comprised the whole molecule; this implies that we considered only three levels of structural organization, level 3 pertaining to a whole molecule.

2.3.5. Classifying the Conformations. The definitions of similarity given above enable us to define a class. The class is a binary number, and its consecutive bits define the similarity of fragments of a given conformation to those of the native structure. For different purposes, we arrange the bits structurewise or fragmentwise. In both arrangements, the first step is to arrange bits into groups corresponding to elementary fragments (group 1), pairs of elementary fragments (group 2), and larger clusters of elementary fragments (group 3 and higher), respectively.

In the structurewise arrangement (Figure 3A) used to define the structural levels, group 1 (elementary fragments) is further divided into subgroups corresponding to secondary structure (sec), contact pattern (cont), and sequence shift (shift) needed to achieve agreement with the corresponding fragment of the native structure. (See sections 2.3.1, 2.3.2, and 2.3.3 for definitions.) Group 2 has only the contact and shift fields, and groups composed of larger segments have only the rmsd (rms) and shift fields. Each subgroup is divided into fields corresponding to the respective fragments or group of fragments. A “1” in the sec, cont, or rms field means that the corresponding criterion of nativelikeness is satisfied, whereas a “0” means that it is not. A “1” in the shift field means that no shift is needed to satisfy a criterion, whereas a “0” means that a sequence shift is needed.

Let us consider the example of the 1IGD protein, which is one of the objects of the present work. The native structure of the 1IGD molecule consists of the N-terminal β-hairpin (hereafter referred to as β_1) and the C-terminal β-hairpin (hereafter referred to as β_3) packed N-to-C-end to form a parallel β-sheet and a middle α-helix (hereafter referred to as α_2) packed to the β-sheet, as shown in Figure 2A. The structure of the class number is shown in Figure 3A. The class number of the native structure consists of all 1’s. As other examples, the structure shown in Figure 2C has class number 111.111.010.010.000.1.1, which indicates that the structure has all three native fragments both as far as the secondary structure and contact pattern is concerned (111.111). However, β_1 and β_3 match the experimental fragment only after the sequence is shifted in the

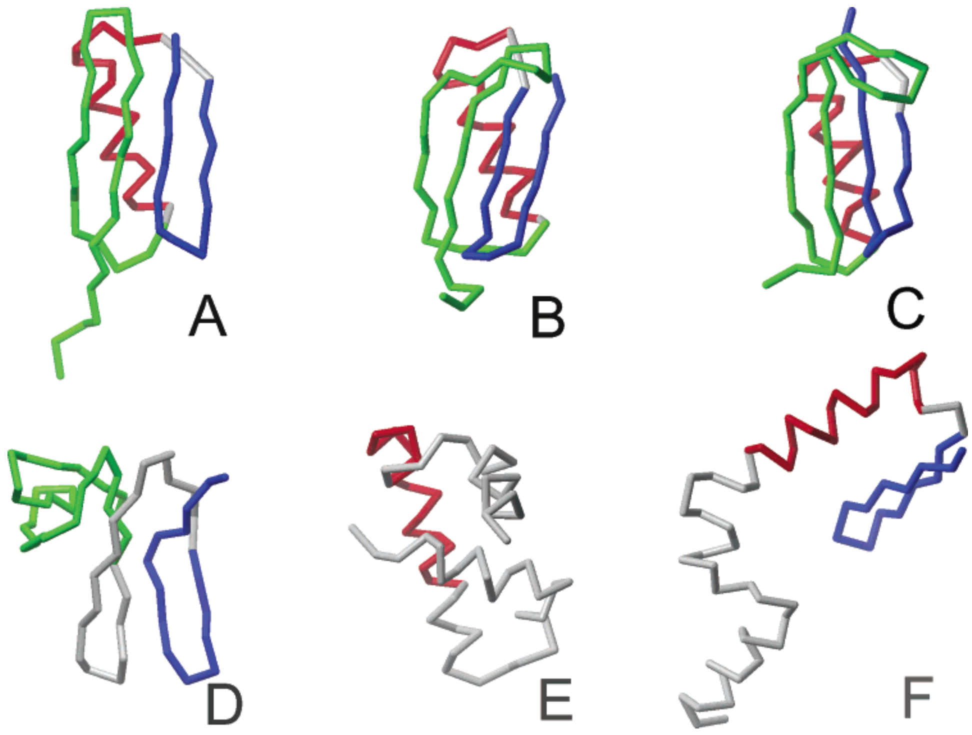

Figure 2. Structures of 1IGD with various degree of nativelikeness. The parts of the molecule are color coded as follows: green, the N-terminal β-hairpin (β_1); red, the middle α-helix (α_2); blue, the C-terminal β-hairpin (β_3). (A) Experimental structure. (B) Lowest-energy structure obtained with the F1 force field (resulting from optimization with hierarchy 1) with incorrectly packed hairpins, the rmsd from the experimental structure being 5.3 Å. (C) Lowest-energy structure obtained with the F2 force field (resulting from optimization with hierarchy 2); the hairpins are packed correctly, but the N-terminal hairpin is distorted in the turn region, which results in an rmsd of 5.9 Å. (D−F) Conformations with some native elements obtained in intermediate stages of force-field optimization.

Group 1									Group 2						Group 3	
sec			cont			shift			cont			shift			rms	shift
β_1	α_2	β_3	β_1	α_2	β_3	β_1	α_2	β_3	β_1-α_2	β_1-β_3	α_2-β_3	β_1-α_2	β_1-β_3	α_2-β_3	β_1-α_2-β_3	β_1-α_2-β_3
□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□

A

Group 1									Group 2						Group 3	
β_1			α_2			β_3			β_1-α_2		β_1-β_3		α_2-β_3		β_1-α_2-β_3	
shift	cont	sec	shift	cont	sec	shift	cont	sec	shift	cont	shift	cont	shift	cont	shift	rms
□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□

B

Figure 3. Illustration of the definition of a class for the example of the 1IGD protein (the native structure shown in Figure 2A). β_1, α_2, and β_3 denote the N-terminal β-hairpin, the middle α-helix, and the C-terminal β-hairpin, respectively, β_1-α_2, β_1-β_3, α_2-β_3, and β_1-α_2-β_3 denote the pairs and triplets of the segments, respectively. Shown are two ways of ordering the bits corresponding to specific match measures and fragments: structurewise (A) and fragmentwise (B). The □ symbols represent fields in which the bits (1 or 0) are placed. Vertical lines and double vertical lines were introduced to separate the groups of fields within a group and between the groups, respectively. See the text for more explanation.

corresponding regions. There is nativelike $\beta_1-\beta_3$ packing to the accuracy of shift (010.000), and the overall structure superposes on the native structure within the specified rmsd criterion (1.1). The structure shown in Figure 2F has class number 011.011.000.000.000.0.0, which indicates that it has nativelike α_2 and β_3 fragments (but not the β_1 fragment, in place of which an α-helix is formed) (011.011). The packing of these fragments is not similar to that in the native structure (000.000), and the structure is not geometrically similar to the native structure (0.0). The dots in the above examples that separate the groups of fields have been introduced for clarity.

As the above examples show, the binary class number provides essential qualitative information about the nativelikeness of a structure. On the basis of the class number, we can define structural families characterized by masks. A mask defines a set of binary numbers with certain common bits in the class number. It has the same structure as the class number, but its field can contain a 0, a 1, or a star (*) serving as a wildcard; both 0 and 1 can be substituted for a star in a particular field. Thus, a star in a given field of the mask characterizing the structural family under consideration indicates that the definition of that family ignores the particular feature described by the respective field (e.g., the respective fragment can have any secondary structure). If the corresponding feature or its lack is a part of the definition of a structural family, a 1 or 0, respectively, should be placed in the respective field of a mask. In other words, a mask defines a set of conformations with the required similarity to the native structure. For 1IGD for example, the structural family consisting of all structures with correct secondary structure in the α-helical part and any structure elsewhere is characterized by mask *1*.***. ***.***.***. ***.*.*. As another example, the structural family consisting of conformations with at least correct secondary structure in the second fragment, but with a conformation that does not have any more native-structure elements, is characterized by mask 010.0*0.0*0.000.000.000.0.0. Comparing the class number with

TABLE 1: Structural Similarity Associated with Specific Octal and Quaternary Numbers Corresponding to Elementary Fragments and Groups of Fragments[a]

number		
binary[b]	octal/ quaternary[c]	structural similarity
		Elementary Fragments
000	0	non-native fragment
001	1	native secondary structure
010	2	native H-bonding contacts, only after sequence shift
011	3	native secondary structure and H-bonding contacts after sequence shift
100	4	not used
101	5	not used
110	6	native H-bonding contacts only
111	7	native secondary structure and H-bonding contacts
		Group of Fragments
00	0	non-native arrangement of fragments
01	1	native packing/rmsd match after sequence shift
10	2	not used
11	3	native packing/rmsd match

[a] See Figure 3B. [b] Bits are ordered as in Figure 3B, the lowest bit corresponding to secondary structure and the highest bit, to sequence shift. [c] Computed from the binary number in the first column.

a mask enables one to assign a conformation immediately to a given structural family.

The structurewise ordering of bits of a class number (as described above) is convenient for the definition of structural families and, therefore, the structural levels in our hierarchical optimization procedure because the groups of bits are arranged in order of importance (secondary structure first, then packing, then shift). However, the resulting binary numbers are long and not particularly good for visual inspection. It takes some time to figure out from a binary class number how similar a conformation is to the experimental structure. For visualization purposes, it is better to rearrange the bits within groups to gather the bits corresponding to a given fragment, as shown in Figure 3B. The advantage of this is that for group 1 the bits corresponding to a given elementary fragment can be converted into an octal number whereas for higher groups (corresponding to associates of elementary fragments) they can be converted into quaternary numbers. The possible values taken by these numbers and the structural similarity associated with each of them are listed in Table 1. For example, the native structure of 1IGD has the octal/quaternary class code 777.777.3, the structure shown in Figure 2C has class code 373.010.3, and the structure shown in Figure 2F has class code 033.000.0.

2.4. Maintaining the Correct Geometry of Secondary Structures. Minimization of the target function given by eq 6 in a given iteration of the algorithm does not guarantee that the resulting parameters of the force field will result in correct geometric details. This is due to the fact that minimization is carried out with a fixed database of conformations (defined in subsection 2.2) and therefore the energetically optimal local structure can become different from that in the database. Similarly, naturally occurring secondary structure elements (e.g., right-handed α-helices) should be lower in energy than the forbidden structure (e.g., left-handed α-helices) or less common structures (e.g., 3_{10} helices). To avoid forbidden structures, we added to the target function a penalty function constructed as follows. Energy minima were determined for two model systems: (i) the terminally blocked Ac−Ala_{19}−NHMe and a β-sheet composed of two antiparallel terminally blocked (Ac−Ala_{10}−NHMe) chains, each in a fully extended conformation (with all virtual-bond dihedral angles $\gamma = 180°$). The energy surface of the first system was expressed as a function of two variables: the virtual-bond valence angle θ and the virtual-bond dihedral angle γ, which had equal values for all residues. Three UNRES energy minima are usually present here: one corresponding to the right-handed α-helix ($0° < \gamma < 60°$), the second one, to either the 3_{10}-helix or the extended structure ($60° < \gamma < 270°$), and the third one, to the left-handed α-helix ($-90° < \gamma < 0°$). The energy surface of the second model system was expressed as a function of two variables: the distance d between the chains and the virtual-bond valence angle θ that took on the same values for all residues. Two minima were present on this surface, the first one corresponding to a correct β structure ($\theta > 90°$) and the second one, to a "compressed" structure ($\theta < 90°$) that can be identified with a sequence of C_7^{eq} conformations of each chain with local 1,7-hydrogen bonds and no interchain hydrogen bonds.

The form of the secondary-structure penalty function is given by eq 9.

$$\Psi = \Psi_{hel} + \Psi_{\beta} \tag{9}$$

where Ψ_{hel} and Ψ_{β} are the sums of all penalty terms for the helical and β structures, respectively; they are defined by eqs 10 and 11, respectively.

$$\begin{aligned}\Psi_{hel} = {} & g(\gamma_{\alpha-hel}; \gamma_{\alpha-hel}^{min}, \gamma_{\alpha-hel}^{max}) + g(\theta_{\alpha-hel}; \theta_{\alpha-hel}^{min}, \theta_{\alpha-hel}^{max}) \\ & + g(E_{\alpha-hel} - E_{3_{10}-hel}; -\infty, -\Delta E_{3_{10}-hel}^{min}) \\ & + g(E_{\alpha-hel} - E_{left-hel}; -\infty, -\Delta E_{left-hel}^{min})\end{aligned} \tag{10}$$

where $\gamma_{\alpha-hel}$ is the value of the virtual-bond dihedral angle γ for the right-handed α-helical minimum, $\gamma_{\alpha-hel}^{min}$ and $\gamma_{\alpha-hel}^{max}$ are the lower and the upper bounds, respectively, of this angle (we assumed $\gamma_{\alpha-hel}^{min} = 42°$ and $\gamma_{\alpha-hel}^{max} = 47°$), $\theta_{\alpha-hel}$ is the value of the virtual-bond valence angle θ for the right-handed α-helical minimum, $\theta_{\alpha-hel}^{min}$ and $\theta_{\alpha-hel}^{max}$ are the lower and the upper bounds, respectively, of $\theta_{\alpha-hel}$ (we assumed $\theta_{\alpha-hel}^{min} = 85°$ and $\theta_{\alpha-hel}^{max} = 95°$), $E_{\alpha-hel}$, $E_{3_{10}-hel}$, and $\Delta E_{3_{10}-hel}^{min}$ are the values of the energy minimum for the right-handed α-helix, the 3_{10}-helix, and the minimum allowed difference, respectively (we assumed $\Delta E_{3_{10}-hel}^{min} = 15$ kcal/mol), $E_{left-hel}$ and ΔE_{left}^{min} are the energy of the left-handed α-helix and the minimum allowed energy difference between the left-handed α-helix and the right-handed α-helix, respectively; we assumed $\Delta E_{left-hel}^{min} = 30$ kcal/mol, for the whole model molecule (Ac−Ala_{19}−NHMe), and g is the quartic penalty function defined by eq 7.

$$\Phi_{\beta} = g(d_{\beta}; d_{\beta}^{min}, d_{\beta}^{max}) + g(\theta_{\beta}; \theta_{\beta}^{min}, \theta_{\beta}^{max}) + g(E_{\beta} - E_{\gamma}; -\infty, -\Delta E_{\beta}^{min}) \tag{11}$$

where d_{β}, d_{β}^{min}, and d_{β}^{max} are the distances between the two Ac−Ala_{10}−NHMe chains at the minimum corresponding to the correct β-structure and the lower and upper bounds, respectively, of the distance for this minimum (we assumed $d_{\beta}^{min} = 4.5$ Å, $d_{\beta}^{max} = 6$ Å); θ_{β}, θ_{β}^{min}, and θ_{β}^{max} are the values of the virtual-bond angle θ for the "correct" β-sheet and the lower and the upper bounds, respectively, of θ (we assumed $\theta_{\beta}^{min} = 110°$, $\theta_{\beta}^{max} = 180°$); E_{β}, E_{γ}, and ΔE_{β}^{min} are the minimum-energy values for the minimum corresponding to the correct β-sheet, the sequence of residues in the C_7^{eq} state, and the lower bound to the energy difference, respectively, between these conformations (we assumed $\Delta E_{\beta}^{\min} = 10$ kcal/mol).

3. Results and Discussion

In this section, we describe the results of the hierarchical optimization of two proteins: 1IGD and 1E0G. For the description of the optimization procedure itself, the reader is referred to section 2.2; here, only the parameters of the procedure are listed. The hierarchies are described in detail.

3.1. Optimization of the UNRES Potential-Energy Function Using 1IGD. In our previous work,[1,3] we reported the results of the hierarchical optimization of 1IGD assuming the formation of structures with a growing number of native elements; this hierarchy is hereafter referred to as the "assembly" hierarchy or hierarchy 1. In our latest paper,[3] we reported the results of optimization and tests of the force field for which the electrostatic, torsional, double-torsional, and correlation terms were parametrized on the basis of the results of ab initio calculations on model systems. As announced in ref 3, we describe here in detail the assembly optimization procedure utilized there to optimize 1IGD and, subsequently, an improved procedure based on the hierarchy of folding events of that protein deduced from the experimental data;[28,29] this hierarchy is hereafter referred to as the "nucleation" hierarchy or hierarchy 2. The resulting force fields are hereafter referred to as F1 and F2, respectively.

Hierarchy 1

The assembly hierarchy (hierarchy 1) used in ref 3 is as follows:

Level 0: no native elements.
Level 1: β_1 **xor** α_2 present.
Level 2: α_2 **xor** β_3 present.
Level 3: β_1 **xor** α_2 **xor** β_3 present.
Level 4: β_1 **and** α_2 present.
Level 5: α_2 **and** β_3 present.
Level 6: β_1 **and** α_2 **xor** α_2 **and** β_3 present.
Level 7: β_1 **and** α_2 **and** β_3 present.

No experimental information was used to construct the above hierarchy. Energy gaps were set between levels 0 and 3, 1 and 4, 2 and 5, and 6 and 7. In the above description, xor denotes the "exclusive or" operator; for example, the notation β_1 xor α_2 denotes structures with either β_1 or α_2 but excludes structures with both β_1 and α_2. It can easily be realized that the assumed hierarchy corresponds to the formation of structures with a single nativelike structural element first (going from level 0 to level 3), then the formation of structures with two nativelike elements (going from level 1 to level 4 and level 2 to level 5, respectively), and finally, the formation of structures with all three elements present (going from level 6 to level 7). No distinction is made as to the order in which nativelike elements and, subsequently, their pairs appear except that we did not set energy gaps between structures with β_1 and β_3 and structures with lower organization because we found that doing so leads to an unacceptable predominance of all-β structures and, thereby, the failure to optimize the force field.

We used three batches of conformations obtained after approximately 8000 (batch 1), 16 000 (batch 2), and 64 000 energy minimizations (batch 3) in the CSA procedure. (In the parallel implementation of CSA, a cycle cannot be terminated exactly after a given number of energy minimizations.) As mentioned in Methods, this corresponds to the early, intermediate, and mature stage of a CSA run. Although including only the final CSA-optimized structures was sufficient in the case of simple folds such as that of 1FSD,[1] for 1IGD the goal of optimization could not be achieved when only the final batch was included. The CSA searches that were started from the energy parameters resulting from a given optimization cycle never produced structures with energies comparably as low as the energies of nativelike conformations already present in the database and calculated with parameters optimized in a given cycle. For batch 1, two β values equal to 0.1 and 0.2 were defined; for batch 2, $\beta = 0.1$, 0.2, and 0.5; and for batch 3, $\beta = 0.1$, 0.2, 0.5, and 2.

As mentioned in Methods (section 2.2), the values of β were chosen to increase with maturing stages of CSA. That is, for batch 1, which contains the least mature structures that can only be precursors of the native structure, we should not focus the optimization on a particular, of even a perfect, precursor conformation. Therefore, the highest selected value of β of batch 1 corresponds to promoting precursor conformations within a 5 (1/0.2) kcal/mol window. The energy window was narrowed down in batch 2, and finally, in batch 3, we wanted the best nativelike structure to be the lowest in energy. Lower β values were also considered for each batch in this optimization to force the high-energy structures of higher levels to be lower in energy than those of lower levels; however, later we found that this was not necessary if an appropriate hierarchy is chosen (which was not the case for the optimization described in this subsection). As mentioned in section 2.2, because CSA does not produce a canonical ensemble and does not incorporate physical temperature, the values of β can be considered only to be parameters of the optimization method. The free energies and the free-energy gaps were evaluated at each β. The increase of the largest β with batch number reflects the fact that the progress of the CSA procedure can, to some extent, be identified with the average "temperature" of the resulting conformational ensemble. By also using lower β values, even for the ensemble consisting of the final CSA structures, we wanted to counterbalance the effect of taking into account only a few structures that happened to have a remarkably lower energy than the other structures that occurred more frequently. Later, we realized that this effect can be achieved by using a single, moderately low β per batch. For batch 1 (least optimized structures), only the first energy gap (between levels 0 and 3) was optimized; for batch 2 (intermediate structures), all energy gaps between levels 6 and 7 were included; finally, for batch 3, all gaps were included. Because the structures containing potentially all native fragments (β_1, α_2, and β_3) were not explicitly specified in the definition of batches, for batch 1 the xor operator in level 3 and for batch 2 the xor operator in level 6 were replaced with the inclusive "or". The initial database of conformations consisted of 20 000 conformations produced in our earlier work[1] after applying a cluster analysis to remove similar structures. The cluster analysis was carried out for each structural class separately by means of the minimum-variance method.[30] The rmsd cutoff was 5 Å. The conformations obtained in the new CSA searches at the end of a given cycle were clustered in the same way (each structural class of each batch separately) and added to the database. For each new set of energy-function parameters, one to three CSA searches were carried out. The final size of the database was 42 932 conformations. The optimized parameters included energy-term weights and the coefficients of the second-order Fourier expansion of the local-interaction energy surfaces of glycine, alanine, and proline.[3]

The initial and final values of the energy gaps imposed in optimization (Δ_{ij}) and the resulting values of the free-energy gaps (ΔF_{ij}) for all batches and all values of β are collected in Table 2. The lowest-energy structure obtained with the optimized

TABLE 2: Initial and Final Free-Energy Gaps ($-\Delta F_{ij}$, kcal/mol) Obtained in Optimization Using the Assembly Hierarchy

Batch 1

		free-energy gaps ($-\Delta F_{ij}$)					
		$\beta = 0.1$			$\beta = 0.2$		
level *i* [a]	level *j* [a]	target	initial	final	target	initial	final
0	3	0.0	−4.5	2.8	1.0	−2.2	3.0
1	4						
2	5						

Batch 2

		free-energy gaps ($-\Delta F_{ij}$)								
		$\beta = 0.1$			$\beta = 0.2$			$\beta = 0.5$		
level *i* [a]	level *j* [a]	target	initial	final	target	initial	final	target	initial	final
0	3	2.0	17.6	23.6	5.0	10.0	13.0	10.0	7.0	5.2
1	4	1.0	−25.6	2.3	2.0	−15.3	2.3	5.0	−12.5	0.5
2	5	1.0	−20.7	1.4	2.0	−13.0	2.5	5.0	−9.3	2.8

Batch 3

		free-energy gaps ($-\Delta F_{ij}$)											
		$\beta = 0.1$			$\beta = 0.2$			$\beta = 0.5$			$\beta = 2.0$		
level *i* [a]	level *j* [a]	target	initial	final	target	initial	final	target	initial	final	target	initial	final
0	3	20.0	38.5	36.1	15.0	29.3	27.0	10.0	27.4	24.5	10.0	27.4	25.0
1	4	0.0	−20.3	13.5	1.0	−5.0	13.2	2.0	5.9	12.7	5.0	8.7	12.4
2	5	0.0	−9.8	9.3	1.0	−5.8	8.0	2.0	−2.5	6.1	5.0	−1.9	4.5
6	7	0.0	4.7	0.1	1.0	9.1	5.1	2.0	9.6	7.8	5.0	9.1	7.9

[a] Levels are numbered as in the text.

force field is shown in Figure 2B. In this Figure, other characteristic structures frequently occurring as the lowest-energy ones at various stages of optimization of the force field are also shown (Figure 2D−F). The plots of the free-energy gaps, compared with the target free-energy gaps imposed in optimization as functions of optimization cycle for the largest β of each batch, are shown in Figure 4A−C. The free-energy gaps shown here were calculated after including conformations obtained in the CSA search with the parameters resulting from a given cycle in the database of conformations. Figure 5 presents the rmsd of the lowest-energy conformation localized in the CSA searches with energy-function parameters determined in a given cycle.

As seen from Figure 4C, in cycle 1 all Δ_{ij} values in batch 3 were set at 10 kcal/mol (solid horizontal line) except for Δ_{67}, which was set at 5 kcal/mol in cycle 1 and raised to 10 kcal/mol in cycle 2 (dotted horizontal line that merges with the solid line). The gaps were satisfied except for the one between levels 2 and 5. This means that structures containing both α_2 and β_3 (including the nativelike structures) are not sufficiently low in energy with respect to structures containing either α_2 or β_3. Simultaneously, structures with no native element are the lowest in energy in batch 1, and structures with at least two native elements are high in energy in batch 2. Consequently, CSA searches produced only non-native structures with a high rmsd (Figure 5); these structures were all-β structures with β_1 and β_3 formed and α_2 converted into an additional β-hairpin (Figure 2D). Running four cycles of optimization did not result in remarkable improvement; additionally, the free-energy gap between levels 1 and 0 decreased from about 30 to about 10 kcal/mol, although (except for cycle 2) still satisfying the target gap. However, the composition of the conformational ensemble obtained in the CSA searches did change from cycle to cycle; for example, in cycle 2, the lowest-energy conformations had predominantly α-helical structure (Figure 2E). We realized that this undesirable situation was due to our attempt to push level 1 too high in free energy with respect to the higher levels; therefore, in three subsequent cycles, we set $\Delta_{14} = \Delta_{25} = \Delta_{67} = 5$ kcal/mol. The free-energy gap between levels 2 and 5 immediately rose above zero (filled squares in Figure 4C), and finally, after cycle 7, all target free-energy gaps in all batches became satisfied or nearly satisfied. The progress of optimization can also be observed in the graph of the rmsd of the lowest-energy structures located in the CSA searches (Figure 5; the lowest-energy structures obviously correspond to batch 3). It should be noted that although all energy gaps ($-\Delta F_{ij}$) of batch 3 are positive since cycle 5 the native structures were not located in CSA searches until the end of cycle 7.

The lowest-energy structure located in the CSA searches after cycle 7 had a 5.3-Å rmsd from the native structure (Figures 2B and 5). It does belong to level 7; therefore, the goal of optimization was achieved. However, it is different from the native structure in that hairpin β_3 is flipped and therefore only its first second strand is packed to β_1 (Figure 2B). Further optimization in which higher levels corresponding to proper packing were included did not work. Such attempts resulted in the destruction of the already-obtained hierarchy. Including the parameters of the $U_{SC_iSC_j}$, $U_{SC_ip_j}$, and $U_{p_ip_j}$ components of the UNRES energy function (cf. eq 1) in addition to the energy-term weights and the Fourier coefficients of the correlation terms in optimization did not result in any improvement, which means that the failure to achieve the required similarity of the lowest-energy structure of the training protein to the native structure was most probably due to the wrong design of the hierarchy.

The resulting F1 force field has two undesirable features, the first of which is its “glassiness”. We carried out more CSA runs and found that the lowest-energy structure (Figure 2B) was obtained in an average of 20% of the runs and the most frequently obtained structure was a higher-energy β-structure or structures in which β_1 has a right-handed α-helical conformation, depending on the number of different recombination operations in the CSA method (Figure 2E). The second feature is the overpreference of the β-structure. This, and the failure of further optimization, means that the stepwise accumulation of

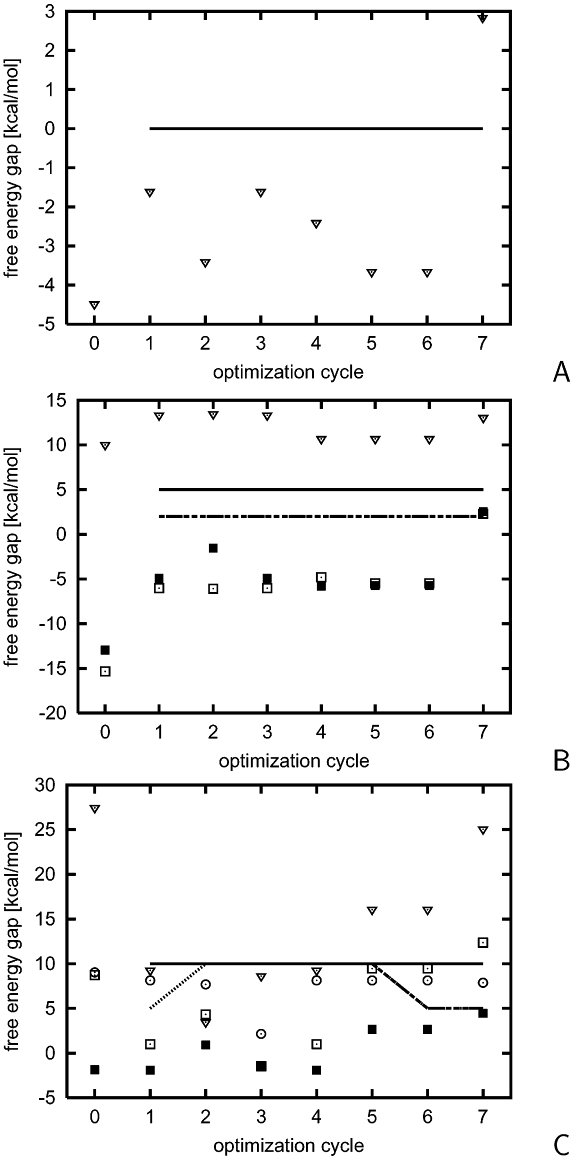

Figure 4. Variation of free-energy gaps ($-\Delta F_{ij}$) with the optimization cycle for assembly hierarchy batch 1 at $\beta = 0.2$ kcal/mol (A), batch 2 at $\beta = 0.5$ kcal/mol (B), and batch 3 at $\beta = 2$ kcal/mol (C). Symbols represent the calculated gaps, and lines represent the target gaps imposed in optimization. Empty inverted triangles and solid lines, hierarchy levels 0 and 3; empty squares and long-and-short-dashed lines, hierarchy levels 1 and 4; filled squares and long-and-short-dashed lines, hierarchy levels 2 and 5; empty circles and dotted line that merges into a solid line and then a long-and-short-dashed line, hierarchy levels 6 and 7, respectively. In batch 3, the target gaps between all levels are 10 kcal/mol in cycles 2−5; therefore, all lines overlap in this region; for cycles 6 and 7 the long-and-short-dashed line represents the target gaps between hierarchy levels 2 and 5 and 6 and 7.

secondary-structure elements does not result in a “natural” folding pathway (i.e., one in which structures of higher levels are easily accessible in the sense of geometry and conformational entropy from lower levels of the structural hierarchy).

Experimental studies of the folding of 1IGD[28,29] suggest that elements β_3 and α_2 are formed first, independently of each other, and hairpin β_1 is folded on the β_3 scaffold by developing hydrogen-bonding contacts between β_3 and the N-terminal part of the sequence of β_1 first, followed by the formation of the complete N-terminal β-hairpin packed against the C-terminal one. Because of geometric constraints the latter event already results in the formation of the complete native structure. Thus, the formation of β_1 is not independent of that of other elements as assumed in our “assembly” model. Although β_1 and α_2 seem to be stable even without the structural context, the unfolding experiments on 1IGD[28,29] indicate that hairpin β_1 cannot exist without being packed to β_3. Therefore, forcing its stability without a structural context could have resulted in exaggerating the stability of β structures in general.

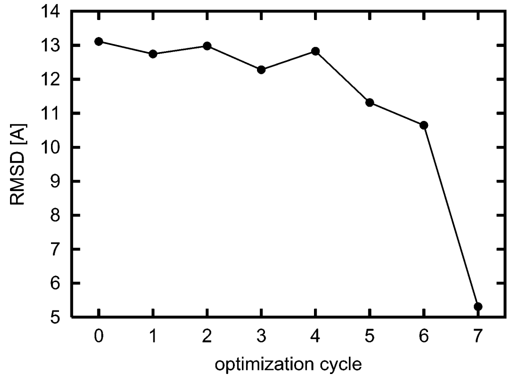

Figure 5. Variation of the rmsd of the lowest-energy structure (from batch 3) found by the CSA searches with the energy-function parameters calculated in a given cycle with cycle number for the assembly hierarchy.

Hierarchy 2

On the basis of this result, we assumed the “nucleation” hierarchy (hierarchy 2) of structural levels:

Level 0: no native elements.

Level 1: α_2 present.

Level 2: β_3 present.

Level 3: α_2 **xor** β_3 present.

Level 4: α_2 **and** β_3 present.

Level 5: α_2 **and** β_3 present but no correct structure in β_1.

Level 6: In addition to level 4, correct secondary structure in β_1.

Level 7: As level 6 but no correct contact pattern of β_1.

Level 8: In addition to level 6, correct packing of the first strand of β_1 to β_3.

Level 9: As level 8 but no complete β-hairpin formed in the β_1 part.

Level 10: β_1 **and** α_2 **and** β_3 present and β_1 packed against β_3.

Level 11: As level 10, but rmsd from the experimental structure exceeding the assigned cutoff.

Level 12: In addition to level 10, rmsd from the native structure below the cutoff value of 6.1 Å. (This follows from the value of 0.1 Å/residue quoted in Methods.)

Free-energy gaps were set between levels 0 and 1, 0 and 2, 3 and 4, 5 and 6, 7 and 8, 9 and 10, and 11 and 12. As in the previous optimization, batches of structures after 8000 (batch 1), 16 000 (batch 2), and 64 000 (batch 3) minimizations were collected. For batch 1 only the first three and for batch 2 only the first four free-energy gaps were considered. As opposed to the previous optimization, we used only a single β for each batch; the β values were 0.1, 0.2, and 0.5 for batchs 1, 2, and 3, respectively. In addition to the energy-term weights and the coefficients of the correlation terms, we also included the well depths of the Gay−Berne side-chain interaction potential (the quantities ϵ_{ij}° in ref 9) in the set of adjustable parameters. The initial and final free-energy gaps are presented in Table 3. The lowest-energy structure obtained with the optimized force field is shown in Figure 2C. The plots of free-energy gaps compared with the target energy gaps imposed in the optimization as functions of optimization cycle for each batch are shown in Figure 6A−C. Figure 7 presents the rmsd of the lowest-energy

TABLE 3: Initial and Final Free-Energy Gaps ($-\Delta F_{ij}$, kcal/mol) Obtained in Optimization Using the Nucleation Hierarchy

		Batch 1			Batch 2			Batch 3		
					free-energy gaps ($-\Delta F_{ij}$)					
		$\beta = 0.1$			$\beta = 0.2$			$\beta = 0.5$		
level *i* [a]	level *j* [a]	target	initial	final	target	initial	final	target	initial	final
0	1	10.0	9.3	12.4	20.0	5.8	26.7	20.0	2.0	49.7
0	2	10.0	9.0	15.0	20.0	5.1	27.4	20.0	−0.3	49.7
3	4	5.0	−10.0	−6.7	10.0	−2.1	8.2	15.0	−1.7	7.4
5	6				5.0	0.3	1.6	15.0	13.3	11.6
7	8							10.0	3.8	1.7
9	10							10.0	−5.8	12.0
11	12							5.0	−35.5	2.7

[a] Levels are numbered as in the text.

conformation localized in the CSA searches with the energy-function parameters determined in a given cycle.

As shown in Figure 7, the CSA searches located structures with low rmsd's as the lowest-energy structures already in cycle 6. However, an inspection of the energy gaps of batch 3 (Figure 6C) indicates that level 10 (structures with completely formed β_1 packed against β_3) is higher in energy than level 8 (structures with incompletely formed β_1 packed against β_3) and that other energy gaps, including those separating levels 1 and 2 from level 0, are quite low. It is particularly disturbing that level 0 is the lowest in free energy for batches 1 and 2, which means that structures with nativelike elements are not favored in energy until the latest stages of CSA, which impairs the search considerably. The lowest-energy structure obtained in cycle 6 did not have nativelike packing, and the result of the search depended very much on the number of different recombination operations in CSA. Nativelike packing in the lowest-energy structure was attained in cycle 7, although the rmsd was higher. However, the energy gaps were still low, and structures with nativelike elements were still high in energy until the latest stages of CSA. In subsequent cycles, nativelike structures became favored in batches 1 and 2. In cycle 14, we obtained the best results in terms of the rmsd. However, low free-energy gaps corresponding to batch 3 again caused a strong dependence of the results of the search on the particular settings of the CSA procedure, so the force field was not a good folder. Therefore, in cycle 17, we raised the energy gaps corresponding to the separation of lower levels, and finally, after cycle 37, we obtained structures with nativelike topology as the lowest-energy structures in the CSA searches and the free-energy gaps rose to positive values, although not all of them were satisfied. In particular, level 0 became separated from the higher levels by a large energy gap. The rmsd of the lowest-energy structure, obtained in cycle 37 in batch 3, was 5.9 Å, which was due to deformation of the turn of the N-terminal β-hairpin (Figure 2C). We tried to optimize the force field further to diminish the rmsd; however, this was unsuccessful. We therefore concluded that further improvement can be achieved only by introducing a quantitative measure of similarity into the native structure within each structural level in optimization. The final energy-term weights, and the coefficients of the expressions for the correlation terms of the F2 force field, are compared with the corresponding parameters of the F1 force field reported in our earlier work[3] in Tables 4 and 5, respectively, and the well-depths (ϵ_{ij}°) of the Gay−Berne potential are summarized and compared with values determined from protein-crystal data[9] in Table S1 in the Supporting Information.

TABLE 4: Comparison of the Energy-Term Weights Obtained by Hierarchical Optimization Using the 1IGD Protein with the Hierarchy 1 (F1 Force Field) and Hierarchy 2 (F2 Force Field) Schemes and Those Obtained by Hierarchical Optimization Using the 1E0G Protein with Hierarchy 3

weight (dimensionless)	value		
	F1	F2	1E0G
w_{SCp}	1.54864	2.79405	2.64146
w_{el}	0.20016	0.14581	0.20803
w_{tor}	1.70537	2.04698	1.88208
w_{tord}	1.24442	1.69624	2.41019
w_{b}	1.00572	1.95684	2.37760
w_{rot}	0.06764	0.17010	0.04803
$w_{\mathrm{loc-el}}^{(3)}$	1.51083	1.21837	1.39959
$w_{\mathrm{loc-el}}^{(4)}$	0.91583	1.84615	1.62840
$w_{\mathrm{loc-el}}^{(5)}$	0.00607	0.02730	0.02730
$w_{\mathrm{loc-el}}^{(6)}$	0.02316	0.00741	0.00741
$w_{\mathrm{turn}}^{(3)}$	2.00764	2.91386	2.27520
$w_{\mathrm{turn}}^{(4)}$	0.05345	0.73178	1.07936
$w_{\mathrm{turn}}^{(6)}$	0.05282	0.02391	0.02391

TABLE 5: Fourier Coefficients (kcal/mol) of Approximate RFE Expressions Pertaining to the Correlation between Local and Electrostatic Interactions $U_{\mathrm{corr}}^{(\mathrm{m})}$ and $U_{\mathrm{turn}}^{(\mathrm{m})}$ of Equation 1[a] up to Sixth Order for the Gly and Ala Type[b] Obtained by Hierarchical Optimization Using the 1IGD Protein with the Assembly (F1 Force Field) and Nucleation (F2 Force Field) Schemes

coefficient[b] (kcal/mol)	Gly	Ala	Gly	Ala
	F1[c]		F2	
b_{11}	0.164513	−0.499500	0.206069	−0.233786
b_{12}	0.000000	1.044194	0.000000	1.501220
b_{21}	0.887388	0.224897	0.791965	0.500262
b_{22}	0.000000	−0.469470	0.000000	−0.878021
c_{11}	−1.175519	2.606108	−0.927963	2.302366
c_{12}	0.000000	−1.348282	0.000000	−1.596422
d_{11}	2.351567	−2.169758	2.373462	−2.089734
d_{12}	0.000000	−0.021834	0.000000	−0.532502
e_{11}	1.700467	1.136668	0.958846	1.676245
e_{12}	0.000000	−0.340904	0.000000	−0.143300
e_{21}	0.000000	−0.234006	0.000000	−0.944257
e_{22}	−1.055494	−1.134231	−1.699998	−0.876359

[a] Equations 43, 46, 48−50, and 56−58 in ref 12. [b] The coefficients corresponding to proline were not optimized because 1IGD contains only one proline residue close to the N terminus that is only weakly involved in long-range interactions. [c] From ref 3.

Tests of Derived Force Fields. To test the derived force fields, we selected small proteins with α, α + β, and β structure (five from each group). These proteins are characterized briefly in Table 6 in terms of decomposition into elementary fragments, which were defined for each protein according to the procedure described in section 2.3.1, and the packing of pairs of these fragments, as defined in section 2.3.3. For each protein, we carried out several CSA searches using three different sets of genetic operators. In the first set, we used an increased number of operations that exchange the β-hairpins and portions of nonlocal β-sheets and only a small number of operations that exchange α-helical fragments; the proportion of these operations

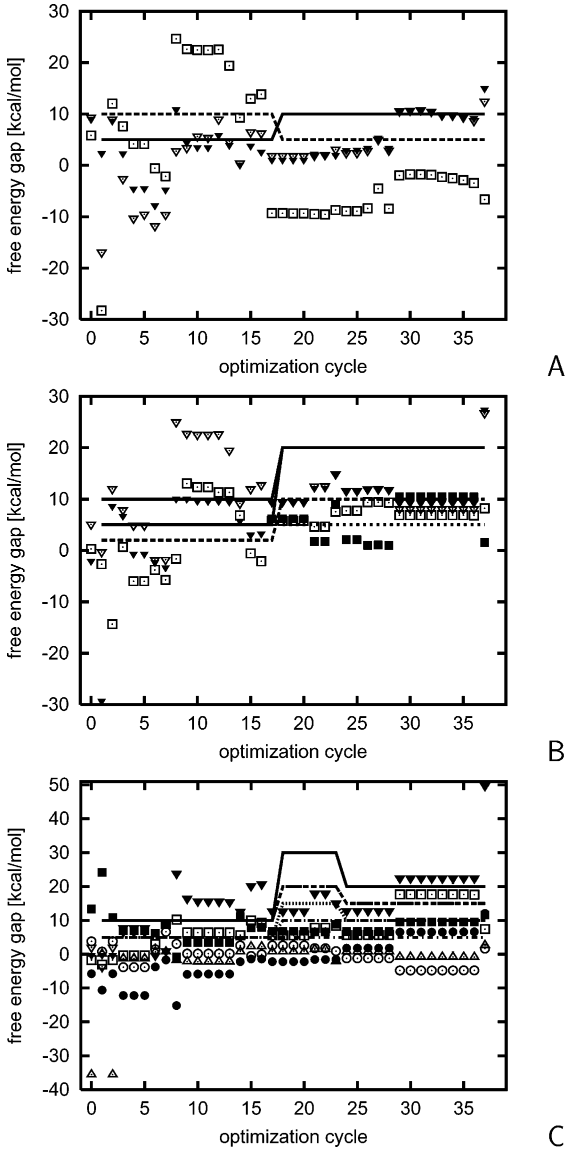

Figure 6. Variation of free-energy gaps with optimization cycle for the nucleation hierarchy for batch 1 (A), batch 2 (B), and batch 3 (C). Each batch was optimized at only one β value; these values are summarized in Table 3. Symbols represent the calculated gaps, and lines represent the target gaps imposed in optimization. Empty inverted triangles and solid lines, levels 0 and 1; filled inverted triangles and solid lines, levels 0 and 2; empty squares and dashed lines, levels 3 and 4; filled squares and dotted lines, levels 5 and 6; empty circles and dot-long-dashed lines, levels 7 and 8; filled circles and dot−long-dashed lines, levels 9 and 10; empty triangles and dot−short-dashed lines, levels 11 and 12.

was reversed in the third set, and the number of both types of operations was equalized in the second set. We found that this procedure greatly increases the chances of finding the global minimum. The genetic operators mentioned here are described in detail in our recent paper.[31] We looked for the nativelike structures or for structures with the largest contiguous nativelike fragments within a 10 kcal/mol energy cutoff from the lowest-energy structure of the respective protein. The results are summarized in Table 7. It can be seen that the force field obtained with the "nucleation" hierarchy scheme is better transferable, whereas, as mentioned earlier in this section, the force field obtained with the "assembly" hierarchy scheme seems to be biased toward the β-structure. This is best illustrated by the fact that the F1 force field predicted grossly wrong secondary structure for 1GAB and 1KOY, which are α-helical proteins.

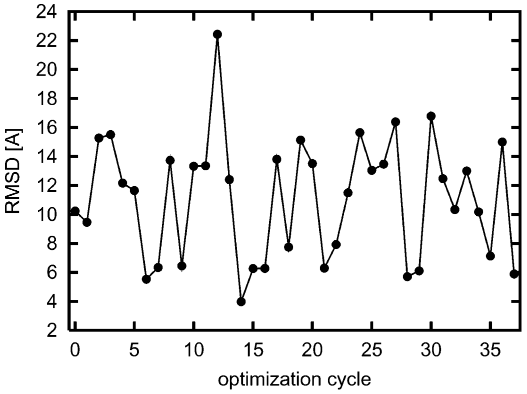

Figure 7. Variation of the rmsd of the lowest-energy structure (from batch 3) found by the CSA searches with the energy-function parameters calculated in a given cycle with the cycle number for the nucleation hierarchy.

As opposed to this, the F2 force field predicted correct secondary structure for all proteins considered, and for protein A (1BDD), 1POU, and 1CLB, the fragments within 6 Å cover almost the entire protein for structures within the 10 kcal/mol energy cutoff. Generally, for the α- and α + β-proteins, the length of the largest matching segment is substantially greater with the F2 force field compared to that with the F1 force field. For the β-proteins, the F1 force field seems to perform slightly better; however, except for simple β-proteins (1E0L and 1ED7), the largest matching segments of structures within the 10 kcal/mol energy cutoff are short even at the 6-Å rmsd cutoff, and it must be concluded that neither of the two force fields performs well for the β-proteins.

3.2. Optimization of the UNRES Potential-Energy Function Using 1E0G. The results presented in the preceding section as well as the results of our on-lattice study of an accompanying paper[4] demonstrate that the choice of hierarchy is essential for the success of optimization. For 1IGD, experimental information on the sequence of events during folding exists, but such information is not available for all potentially interesting training proteins. Nevertheless, on the basis of our lattice studies,[4] it can be supposed that an appropriate hierarchy can be found by a trial-and-error method. To check this, we selected another α + β-protein, 1E0G (48 residues). The structure of this protein consists of a helix-turn-helix motif and an N-terminal and a C-terminal strand packed into a two-stranded antiparallel β-sheet. Consequently, we divided the native structure into four fragments—s_1, α_2, α_3, and s_4—as given in Table 6 and illustrated in Figure 8A.

We tried three hierarchies, the levels being composed as follows (level 0 contains no native-structure elements in all cases):

Hierarchy 1

Level 1: s_1 **or** s_4 formed.

Level 2: s_1 **and** s_4 **or** α_2 **or** α_3 formed.

Level 3: s_1 **and** s_4 **and** α_2 **and** α_3 formed.

Level 4: As in level 3 and additionally s_1 packed to s_4 **or** α_2 packed to α_3.

Level 5: As in level 3 and additionally s_1 packed to s_4 **and** α_2 packed to α_3, but the rmsd from the experimental structure was above the cutoff value.

Level 6: As in level 5 and additionally the rmsd from the experimental structure was below a cutoff value.

Hierarchy 2

Level 1: s_1 and s_4 formed and packed to form a β-sheet **or** α_2 **or** α_3 formed.

TABLE 6: Summary of the Proteins Used to Test the Force Field

protein[a]	type	N_{res}[b]	elementary fragments[c]	packing
1BDD[33]	α	46	$\alpha_1(1-6)$ $\alpha_2(17-30)$ $\alpha_3(33-45)$	α_1/α_3 α_2/α_3
1GAB[34]	α	47	$\alpha_1(5-12)$ $\alpha_2(21-27)$ $\alpha_3(33-44)$	α_2/α_3
1KOY[35]	α	62	$\alpha_1(1-6)$ $\alpha_2(19-28)$ $\alpha_3(30-35)$ $\alpha_4(41-62)$	α_2/α_4 α_3/α_4
1CLB[36]	α	75	$\alpha_1(4-13)$ $\alpha_2(25-36)$ $\alpha_3(47-55)$ $\alpha_4(63-75)$	α_1/α_2 α_1/α_4
1POU[37]	α	76	$\alpha_1(1-20)$ $\alpha_2(25-34)$ $\alpha_3(40-49)$ $\alpha_4(55-70)$	α_1/α_3 α_1/α_4 α_2/α_4 α_2/α_4
1FSD[38]	$\alpha+\beta$	28	$\beta_1(1-13)$ $\alpha_2(14-26)$	β_1/α_2
2PLT[39]	$\alpha+\beta$	61	$\beta_1(2-22)$ $\alpha_2(27-38)$ $\beta_3(47-58)$	β_1/α_2 β_1/β_3 α_2/β_3
1UBQ[40]	$\alpha+\beta$	75	$\beta_1(1-17)$ $\alpha_2(23-35)$ $\beta_3(43-50)$ $s_4(65-72)$	β_1/α_2 β_1/s_4 α_2/β_3
1E0G[5]	$\alpha+\beta$	48	$s_1(2-7)$ $\alpha_2(13-20)$ $\alpha_3(24-32)$ $s_4(41-46)$	s_1/α_2 s_1/s_4 α_2/s_4 α_3/s_4
1QHK[41]	$\alpha+\beta$	45	$s_1(1-7)$ $s_2(12-18)$ $\alpha_3(18-26)$ $s_4(33-37)$ $\alpha_5(39-47)$	s_1/s_2 s_1/α_3 s_1/s_4 s_1/α_5 s_2/α_3 s_2/α_5 s_4/α_5
1E0L[42]	β	28	$\beta_1(2-18)$ $\beta_2(14-25)$	β_1/β_2
1ED7[43]	β	45	$s_1(6-10)$ $\beta_2(12-23)$ $s_3(25-29)$ $s_4(41-45)$	s_1/s_3 β_2/s_4
1BK2[44]	β	57	$s_1(2-7)$ $s_2(8-12)$ $s_3(18-22)$ $s_4(24-29)$ $\beta_5(35-50)$ $s_6(52-56)$	s_1/s_4 s_1/s_6 s_2/s_3 s_4/β_5
1FYN[45]	β	61	$s_1(4-9)$ $s_2(10-14)$ $s_3(20-24)$ $s_4(27-31)$ $s_5(37-44)$ $s_6(48-54)$ $s_7(56-60)$	s_1/s_4 s_1/s_7 s_2/s_3 s_4/s_5 s_5/s_6
1WIU[46]	β	93	$s_1(3-6)$ $s_2(10-14)$ $s_3(18-26)$ $s_4(25-28)$ $s_5(32-39)$ $\beta_6(49-62)$ $\beta_7(69-92)$	s_1/s_4 s_2/β_7 s_3/β_6 s_5/β_7

[a] PDB codes and literature references. [b] Number of residues. [c] α − α-helix; s − β-strand; β − β-hairpin.

TABLE 7: Results of Tests of the Two Force Fields Obtained by Hierarchical Optimization

	force field											
	F1[b]						F2					
protein[a]	E[c]	rms[d]	n4[e]	n5[f]	n6[g]	class[h]	E[c]	rms[d]	n4[e]	n5[f]	n6[g]	class[h]
1BDD	0.0	8.9	31	34	38	727.00.0	0.0	3.2	46	46	46	717.20.3
	4.4	3.7	46	46	46	777.10.3						
1GAB	0.0	10.0	12	14	19	000.0.0	0.0	11.1	24	28	32	707.0.0
	6.9	8.5	12	17	22	000.0.0	5.1	8.7	35	38	41	737.0.0
1KOY	0.0	14.8	12	14	17	0000.00.0	0.0	11.6	27	30	36	7777.00.0
	1.1	11.5	20	26	34	0007.00.0	9.5	10.0	24	29	43	7730.00.0
1CLB	0.0	10.4	26	30	49	0177.00.0	0.0	8.5	40	54	57	0377.00.0
	5.4	9.0	27	40	59	7377.10.0	0.3	7.4	46	51	64	0377.00.0
1POU	0.0	12.8	13	16	24	0000.0000.0	0.0	9.9	34	36	39	7031.0101.0
							4.8	6.0	37	52	71	7001.0000.3
1FSD	0.0	3.2	28	28	28	77.0	0.0	4.9	25	28	28	17.0
							1.8	2.7	28	28	28	77.0
2PTL	0.0	11.0	19	23	26	103.00.0	0.0	11.6	29	38	41	070.000.0
	9.1	10.4	24	27	30	007.00.0	8.6	7.2	49	57	59	173.000.0
1UBQ	0.0	13.1	19	21	23	0107.000.0	0.0	15.7	24	27	36	3000.000.0
	8.9	13.8	21	23	26	0107.000.0	3.6	14.0	31	35	51	3720.000.0
1E0G	0.0	10.7	18	22	32	7007.0000.0	0.0	11.8	18	20	26	7007.0000.0
	7.3	6.8	19	26	36	7007.0100.0	8.1	5.3	36	45	48	7777.0101.0
1QHK	0.0	10.3	17	20	26	37070.1000000.0	0.0	7.5	31	36	40	77707.0100200.0
	3.0	6.7	21	42	45	77070.1010000.0	1.6	7.4	33	37	42	77707.0100200.0
1E0L	0.0	3.9	28	28	28	77.3	0.0	4.2	27	28	28	33.1
1ED7	0.0	8.3	14	17	25	7330.00.0	0.0	8.1	21	25	32	7730.00.0
	4.1	5.9	27	37	45	3760.00.0	7.9	6.0	25	35	45	7320.00.0
1BK2	0.0	9.9	15	18	29	007717.0000.0	0.0	10.8	14	18	24	000000.0000.0
	1.9	10.8	29	35	38	777717.0001.0	5.1	12.2	14	19	27	000000.0000.0
1FYN	0.0	11.1	19	34	38	7377706.00010.0	0.0	12.5	15	23	28	0020076.00000.0
	2.4	9.8	20	30	32	7277766.00010.0	7.2	10.8	21	26	31	0020066.00000.0
1WIU	0.0	13.5	20	24	29	7700703.0000.0	0.0	12.2	19	23	30	1773003.1000.0
	7.1	14.7	23	26	32	7336233.0000.0	7.7	15.9	25	33	36	0002070.0000.0

[a] Proteins are identified by PDB codes. The first line of each entry is the lowest-energy structure, and the second line is a structure within a 10 kcal/mol energy cutoff for which one of the following holds: (i) it has the lowest rmsd and the rmsd is within 0.1 Å per residue or within 4 Å for proteins with fewer than 40 amino acid residues; (ii) the longest fragment within a 4-, 5-, or 6-Å rmsd cutoff is not shorter than 40, 50, or 60 residues, respectively; or (iii) it has the longest fragment within a 6-Å cutoff. Only the lowest-energy structure is reported if its rmsd is lower than 0.1 Å per residue or 4 Å for proteins shorter than 40 residues or if there is no other structure within a 10 kcal/mol energy cutoff remarkably more similar to the native structure than the lowest-energy structure. [b] Except for the 1KOY and 1FYN data, which were gathered from the results of calculations reported in ref 3. [c] Relative energy. [d] rmsd from the native structure; [e] Length of the largest contiguous segment within a 4-Å rmsd from the native structure. [f] Length of the largest contiguous segment within a 5-Å rmsd from the native structure. [g] Length of the largest contiguous segment within a 6-Å rmsd from the native structure. [h] Segment-wise class numbers in the octal/quaternary forms, as described in section 2.3.5. See Table 6 for the order of elementary fragments and pairs of packed fragments. Three groups of fragments were defined, the third groups comprising the whole molecule, except 1FSD and 1E0L, which consist of two fragments and for which, consequently, only two groups were defined.

Level 2: s_1 and s_4 formed and packed to form a β-sheet **and** (α_2 **or** α_3 formed) **or** α_2 **and** α_3 formed.

Level 3: α_2 **and** α_3 formed **and** (α_2 packed to α_3 **or** the β-sheet formed from s_1 and s_4).

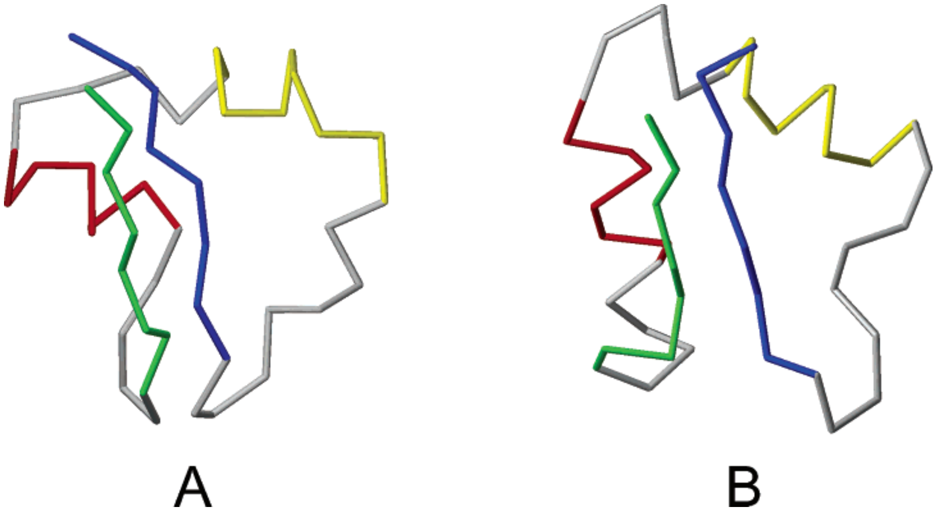

Figure 8. Experimental structure of 1E0G (A) and the lowest-energy structure obtained by hierarchical optimization with hierarchy 3 (B); the rmsd from the experimental structure is 4.2 Å. The parts of the molecule are color coded as follows: green, the N-terminal β-strand (s_1); red, the first α-helix (α_1); yellow, the second α-helix; blue, the C-terminal β-strand (s_2).

TABLE 8: Initial and Final Free-Energy Gaps ($-\Delta F_{ij}$, kcal/mol) Obtained in the Optimization of the UNRES Force Field with 1E0G as the Benchmark Protein

		Batch 1			Batch 2			Batch 3		
		free-energy gaps ($-\Delta F_{ij}$)								
		$\beta = 0.1$			$\beta = 0.2$			$\beta = 0.5$		
level *i* [a]	level *j* [a]	target	initial	final	target	initial	final	target	initial	final
					Hierarchy 1					
0	1	5.0	3.7	3.9	5.0	−8.5	−8.5	5.0	3.5	12.5
1	2	5.0	−7.0	−7.2	5.0	−6.0	−6.0	5.0	−0.8	−2.5
2	3	5.0	−11.3	−11.2	5.0	−16.8	−16.8	5.0	−1.1	−10.8
3	4				5.0	7.4	7.3	5.0	11.6	33.3
					Hierarchy 2					
0	1	20.0	12.6	12.6	20.0	13.5	30.7	20.0	−2.5	14.6
1	2	15.0		−11.4	15.0		13.9	15.0	−10.8	5.8
2	3	10.0		2.7	10.0		5.1	10.0		3.4
3	4	5.0			5.0			5.0		−24.7
4	5							5.0		11.2
					Hierarchy 3					
0	1	25.0	12.6	21.7	25.0	14.9	21.8	25.0	−2.5	51.9
1	2	20.0	−2.8	14.3	20.0	−10.8	15.5	20.0	−10.8	23.7
2	3	15.0	−3.0	26.8	15.0	−17.8	11.8	15.0	9.8	33.8
3	4				10.0		3.2	10.0		6.1
4	5				5.0		5.4	5.0		13.1
5	6							5.0		17.7

[a] Levels are numbered as in the text.

Level 4: α_2 **and** α_3 formed **and** α_2 packed to α_3 **and** the β-sheet formed from s_1 and s_4, but the rmsd from the experimental structure was above the cutoff.

Level 5: As in level 4 and additionally the rmsd from the experimental structure was below a cutoff value.

Hierarchy 3

Level 1: α_1 **or** α_2 formed.

Level 2: α_1 **and** α_2 formed, but not packed to each other.

Level 3: α_1 **and** α_2 formed and packed to each other **and** (s_1 **or** s_2 formed and packed α_1 or α_2, respectively).

Level 4: α_1 **and** α_2 formed and packed to each other **and** the β-sheet formed with strands packed to the respective helices, but the rmsd from the experimental structure was above the cutoff value.

Level 5: α_1 **and** α_2 formed and packed to each other **and** the β-sheet formed with strands packed to the respective helices, and the rmsd from the experimental structure was within the cutoff.

The results of optimization with these three hierarchies are summarized in Table 8. We optimized the energy-term weights and the well depths of the $U_{SC_iSC_j}$ potentials, and the Fourier coefficients of the correlation terms were taken from the F2 force field because this force field is reasonably transferable. As for 1IGD, we used three batches of conformations corresponding to 8000, 16 000, and 64 000 energy minimizations, respectively, in a CSA run. It can be seen in Table 8 that optimization with the first hierarchy, in which strand formation was assumed first, got stuck, and we were unable to place level 2 above level 1 and level 3 above level 2, although even with initial parameters the free-energy gap between level 3 and level 4 was fairly large. Moreover, level 1 was above level 0 in batches 1 and 2. This situation did not improve after five iterations, although we eliminated levels 5 and 6 to focus on the correct energy ordering of levels 0−4. We therefore tried hierarchy 2, in which structure formation starts from either one of the α-helices or the β-sheet. As can be seen from Table 8, this hierarchy was more successful, but we were unable to

TABLE 9: Results of Tests of the Force Fields Obtained by the Hierarchical Optimization of 1E0G with Hierarchy 3

protein[a]	E[a]	rms[a]	n4[a]	n5[a]	n6[a]	class[a]
1BDD	0.0	3.3	46	46	46	717.10.3
1GAB	0.0	9.9	35	38	40	777.0.0
	9.3	3.5	47	47	47	777.1.3
1KOY	0.0	9.2	30	34	37	7177.00.0
	6.6	7.2	27	38	52	7777.11.0
1CLB	0.0	6.4	39	55	69	7377.00.3
	2.2	5.0	49	75	75	0777.00.3
1POU	0.0	10.2	34	36	39	6003.0100.0
1FSD	0.0	6.7	20	24	25	06.0.0
	3.5	2.9	28	28	28	77.1.3
1IGD	0.0	12.3	28	29	32	070.000.0
	6.2	11.8	27	31	39	070.000.0
2PLT	0.0	11.6	31	40	42	071.000.0
1UBQ	0.0	11.0	28	35	38	0700.000.0
	9.1	12.3	29	34	45	3700.000.0
1QHK	0.0	6.4	24	36	45	20102.0001000.0
1E0L	0.0	8.9	11	14	19	00.0.0
	2.6	4.4	27	28	28	33.1.3
1ED7	0.0	7.3	20	23	33	0320.00.0
1BK2	0.0	12.1	14	18	26	00000.0000.0
1FYN	0.0	11.4	15	19	21	0020003.00000.0
1WIU	0.0	14.5	14	20	25	0002000.0000.0

[a] See the legend to Table 7 for an explanation of the symbols.

achieve the target free-energy gap between levels 3 and 4; in other words, it was possible to locate low-energy structures with the central helix-turn-helix motif or the parallel β-sheet but not with both. We concluded that the fallacy of this hierarchy was ignoring the role of helix-strand packing. Therefore, we tried hierarchy 3, in which the helix-turn-helix motif is formed first, which is followed by formation and packing of a single strand to the respective α-helix, and the parallel β-sheet between the N- and the C-terminal strands forms only in the last stage when one of the strands is already present and stabilized through interaction with the respective α-helix. It can be seen from Table 8 that optimization with this hierarchy was successful. The lowest-energy structure has a 4.2-Å rmsd from the experimental structure (Figure 8B). The energy-term weights obtained by optimization with hierarchy 3 are summarized in Table 4. The well depths of the $U_{SC_iSC_j}$ potential (ϵ_{ij}°) are summarized in Table S2 of the Supporting Information.

We subsequently tested the force field based on 1E0G using the same set of proteins as for 1IGD (including 1IGD). The results, summarized in Table 9, show that overall the force field does not perform as well in terms of the longest predicted segment within the 4-, 5-, and 6-Å rmsd cutoffs, respectively, as the F2 force field does (trained using the 1IGD protein). Definitely better results compared to the F2 force field were obtained for 1KOY and 1CLB, where the longest predicted segment within a 6-Å rmsd cutoff increased by more than 10 residues, and better results in terms of the overall rmsd were obtained for 1GAB. However, the F2 force field definitely gives better results for 1POU, 2PLT, 1ED7, 1FYN, and 1WIU and still better for 1UBQ. It should be noted that this last list contains mostly $\alpha + \beta$- and β-proteins. This means that, generally, the force field derived on the basis of 1E0G is able to handle the α proteins reasonably well but is not as transferable to $\alpha + \beta$- and β-proteins. Conversely, the F2 force field (trained using the 1IGD protein) handles the α- and $\alpha + \beta$-proteins reasonably well, its performance deteriorating (but still better than that of the force field trained using the 1E0G protein) for the β-proteins. It can therefore be concluded that the structure of 1IGD is determined by a greater number of essential interactions responsible for the structure formation of proteins than that of 1E0G.

Because 1E0G is smaller than 1IGD and, at the same time, contains sufficiently complex structural motifs, we also carried out nonhierarchical optimization runs for 1E0G to compare with hierarchical optimization. We started from the database of conformations resulting from hierarchical optimization; therefore, the database contained structures with various degree of nativelikeness. However, after 13 iterations we achieved a free-energy gap of 13.4 kcal/mol between nativelike structures and the lowest-energy non-native structures (this value was obtained after the database of conformations was updated following CSA searches with the final parameters), the nativelike structure could not be located in any CSA run, the lowest-energy structure found by a global CSA conformational search being a full β-sheet.

4. Conclusions

Here, we further developed the method of hierarchical optimization of the UNRES potential-energy function for off-lattice proteins that was introduced in our earlier work.[1] On the basis of 1IGD as an example, we demonstrated that the success of the optimization procedure and the transferability of the force field depends critically on the choice of hierarchy. The assembly scheme resulted in a glassy force field, which had a bias toward β-structures. Designing the hierarchy according to the sequence of folding events deduced from experiment resulted in a force field with good foldability properties, which was much more transferable to other proteins than the former one. On the basis of 1E0G as another example, we demonstrated that it is possible to deduce a reasonable hierarchy without having experimental information. On the basis of this example, we also demonstrated that ignoring structural hierarchy in optimization leads to a nonsearchable potential, even though the database of conformations implemented in optimization contained conformations with different degrees of nativelikeness.

It should be noted, however, that although introducing (though qualitatively) some experimental information about the folding process resulted in qualitative improvement of the force field (cf. the F2 vs the F1 force field obtained using 1IGD as the benchmark protein) the F2 force field still does not perform satisfactorily in the sense of a reliable prediction. It is clear that the use of a single protein is not sufficient, and more proteins representing different classes of folds and containing different interaction patterns should be considered. This issue is addressed in an accompanying paper.[6] Moreover, the force fields obtained by the present procedure of hierarchical optimization find only medium-resolution structures as low-energy structures even in the case of 1IGD, which was used in the optimization. The reason for this is most probably that the structural classes were described only qualitatively; there is no way to differentiate structures within a level on the basis of their similarity to the native structure. Introducing a quantitative similarity measure within structural levels and incorporating it into the optimization should rectify this shortcoming. This research is currently being carried out in our laboratory.

The hierarchical method of optimizing the energy landscape was connected in this work with the conformational search using the CSA method, which searches the space of energy minima and, therefore, produces a noncanonical distribution of conformational states which, moreover, depends on the particular set of CSA operations implemented in the simulation. (Hence, we carried out several conformational searches to compensate for this bias.) Therefore, in contrast to the on-lattice studies carried out in an accompanying paper[4] where we used the complete set of conformations in the optimization, it could not be expected in principle that the resulting energy functions will lead to the native structures in canonical simulations, where the entropic factor matters. However, recently[32] we applied Langevin molecular dynamics (LMD) to the UNRES force field and ran simulations on selected proteins with the F2 force field derived in this work. Preliminary results for simple helical proteins, such as protein A, show that LMD simulations converge to the respective native structure. The results will be published in our forthcoming paper.[32]

Acknowledgment. This work was supported by grants from the National Institutes of Health (GM-14312), the National Science Foundation (MCB00-03722), the NIH Fogarty International Center (TW1064), and grant 3 T09A 032 26 from the Polish State Committee for Scientific Research (KBN). This research was conducted by using the resources of (a) our 392-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, (b) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, (c) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk, (d) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and (e) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

Supporting Information Available: Well depths (ϵ_{ij}°) of the Gay–Berne[27] potential of side-chain interactions,[9] which correspond to the energy-term weights of the F2 and the 1E0G-trained force field summarized in Table 4. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Liwo, A.; Arłukowicz, P.; Czaplewski, C.; Ołdziej, S.; Pillardy, J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1937–1942.

(2) Derrick, J. P.; Wigley, D. B. *J. Mol. Biol.* **1994**, *243*, 906–918.

(3) Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kozłowska, U.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 9421–9438.

(4) Liwo, A.; Arłukowicz, P.; Ołdziej, S.; Czaplewski, C.; Makowski, M.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16918–16933.

(5) Bateman, A.; Bycroft, M. *J. Mol. Biol.* **2000**, *299*, 1113–1119.

(6) Ołdziej, S.; Łągiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nanias, M.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16950–16959.

(7) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1697–1714.

(8) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1715–1731.

(9) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849–873.

(10) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874–887.

(11) 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*, 259–276.

(12) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Chem. Phys.* **2001**, *115*, 2323–2347.

(13) Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7291–7298.

(14) Pillardy, J.; Czaplewski, C.; Liwo, A.; Wedemeyer, W. J.; Lee, J.; Ripoll, D. R.; Arłukowicz, P.; Ołdziej, S.; Arnautova, Y. A.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7299–7311.

(15) Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. A* **2003**, *107*, 8035–8046.

(16) Kubo, R. *J. Phys. Soc. Jpn.* **1962**, *17*, 1100–1120.

(17) Kolinski, A.; Skolnick, J. *J. Chem. Phys.* **1992**, *97*, 9412–9426.

(18) Lee, J.; Scheraga, H. A.; Rackovsky, S. *J. Comput. Chem.* **1997**, *18*, 1222–1232.

(19) Lee, J.; Scheraga, H. A.; Rackovsky, S. *Biopolymers* **1998**, *46*, 103–115.

(20) Lee, J.; Scheraga, H. A. *Int. J. Quantum Chem.* **1999**, *75*, 255–265.

(21) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025–2030.

(22) Pillardy, J.; Czaplewski, C.; Liwo, A.; Lee, J.; Ripoll, D. R.; Kaźmierkiewicz, 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*, 2329–2333.

(23) Orengo, C. A.; Bray, J. E.; Hubbard, T.; LoConte, L.; Sillitoe, I. *Proteins: Struct., Funct., Genet.* **1999**, *Supplement 3*, 149–170.

(24) Czaplewski, C.; Ripoll, D. R.; Ołdziej, S.; Kazmierkiewicz, R.; Vila, J. A.; Liwo, A.; Pillardy, J.; Saunders, J. A.; Chinchio, M.; Nanias, M.; Khalili, M.; Arnautova, Y. A.; Jagielska, A.; Kang, Y. K.; Gibson, K. D.; Scheraga, H. A. Physics-Based Protein-Structure Prediction Using the UNRES and ECEPP/3 Force Fields – Tests on CASP5 Targets. In *Fifth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction 2002*. http://predictioncenter.llnl.gov/casp5/Casp5.html.

(25) Gay, D. M. *ACM Trans. Math. Software* **1983**, *9*, 503–524.

(26) Liwo, A.; Czaplewski, C.; Ołdziej, S.; Scheraga, H. A. To be submitted for publication.

(27) Gay, J. G.; Berne, B. J. *J. Chem. Phys.* **1981**, *74*, 3316–3319.

(28) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584–590.

(29) Kuszewski, J.; Clore, G. M.; Gronenborn, A. M. *Protein Sci.* **1994**, *3*, 1945–1952.

(30) Späth, H. *Cluster Analysis Algorithms*; Halsted Press: New York, 1980.

(31) Czaplewski, C.; Liwo, A.; Pillardy, J.; Ołdziej, S.; Scheraga, H. A. *Polymer* **2004**, *45*, 677–686.

(32) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. A. To be submitted for publication.

(33) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. *Biochemistry* **1992**, *31*, 9665–9672.

(34) Johansson, M. U.; de Chateau, M.; Wikstrom, M.; Forsen, S.; Drakenberg, T.; Bjorck, L. *J. Mol. Biol.* **1997**, *266*, 859–865.

(35) Fukushima, K.; Kikuchi, J.; Koshiba, S.; Kigawa, T.; Kuroda, Y.; Yokoyama, S. *J. Mol. Biol.* **2002**, *321*, 317–327.

(36) Svensson, L. A.; Thulin, E.; Forsen, S. *J. Mol. Biol.* **1992**, *223*, 601–606.

(37) Assa-Munt, N.; Mortishire-Smith, R. J.; Aurora, R.; Herr, W.; Wright, P. E. *Cell* **1993**, *73*, 193–205.

(38) Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82–87.

(39) Wikstrom, M.; Drakenberg, T.; Forśen, S.; Sjobring, U.; Bjorck, L. *Biochemistry* **1994**, *33*, 14011–14017.

(40) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531–544.

(41) Evans, S. P.; Bycroft, M. *J. Mol. Biol.* **1999**, *291*, 661–669.

(42) Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H. *Nat. Struct. Biol.* **2000**, *7*, 375–379.

(43) Ikegami, T.; Okada, T.; Hashimoto, M.; Seino, S.; Watanabe, T.; Shirakawa, M. *J. Biol. Chem.* **2000**, *275*, 13654–13661.

(44) Martinez, J. C.; Pisabarro, M. T.; Serrano, L. *Nat. Struct. Biol.* **1998**, *5*, 721–729.

(45) Musacchio, A.; Saraste, M.; Wilmanns, M. *Nat. Struct. Biol.* **1994**, *1*, 546–551.

(46) Fong, S.; Hamill, S. J.; Proctor, M.; Freund, S. M.; Benian, G. M.; Chothia, C.; Bycroft, M.; Clarke, J. *J. Mol. Biol.* **1996**, *264*, 624–639.