Optimization of the UNRES Force Field by Hierarchical Design of the Potential-Energy Landscape. 3. Use of Many Proteins in Optimization

Stanisław Ołdziej,[†,‡] Justyna Łągiewka,[‡] Adam Liwo,[†,‡] Cezary Czaplewski,[†,‡] Maurizio Chinchio,[†] Marian Nanias,[†] and Harold A. Scheraga*[,†]

Baker Laboratory of Chemistry and Chemical Biology, 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 report the application of the hierarchical optimization method of protein potential-energy landscapes described in the accompanying papers (Liwo, A.; Arłukowicz, P.; Ołdziej, S.; Czaplewski, C.; Makowski, M.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16918; Ołdziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16934) to optimize the UNRES potential energy function using two [1E0G ($\alpha + \beta$) and 1E0L (β)], three [1E0G, 1E0L, and 1GAB (α)], and, finally, four training proteins [1E0G, 1E0L, 1GAB, and 1IGD ($\alpha + \beta$)] simultaneously; these training sets and the resulting force fields are referred to as 2P, 3P, and 4P, respectively. The hierarchies of 1E0L and 1GAB were determined following the procedure applied to 1E0G described in an accompanying paper,[2] the hierarchies of 1IGD and 1E0G being taken from experiment and from the accompanying paper,[2] respectively. For all training sets, optimization was successful; in other words, (i) the target function composed of contributions from each set of training proteins could be optimized and (ii) the resulting force fields located the nativelike structures of each of the training proteins as the lowest energy by a global conformational search, which means that hierarchical optimization with multiple training proteins is feasible. Subsequently, the 3P and 4P force fields were tested on a set of 66 proteins (26 α-, 15 β-, and 25 $\alpha + \beta$-proteins with chain length from 28 to 144 amino acid residues). Both force fields perform comparably on the α proteins, but the 4P force field performs definitely better on the $\alpha + \beta$- and β-proteins. With the 4P force field, the average length of a continuous segment matching the corresponding segment of the experimental structure within 6 Å rmsd and the percentage of correctly predicted chain length are 54 (67%), 34 (45%), 42 (55%), and 45 (58%) for the α-, β-, $\alpha + \beta$-, and all proteins, respectively, and the length of the longest predicted continuous fragment is 96, 49, and 70 residues for the α-, β-, and $\alpha + \beta$-proteins, respectively; with the 3P force field, the longest predicted fragment within a 6-Å rmsd cutoff was 127 residues (for the 144-residue 1LPE α-protein). These results are a major step forward with respect to our earlier attempts at optimizing the UNRES force field by maximizing the energy gap and *Z* score between the nativelike structures and the lowest-energy non-native structure where, for feasibility, we earlier had to derive a separate force field for each structural class, and the predictive power of the force field derived to treat α-proteins was much greater than those derived for the $\alpha + \beta$ and the β structural classes. However, the 4P force field definitely performs better on the α- and $\alpha + \beta$- than on the β-proteins, which strongly suggests that further improvements are needed, the most significant issue being to differentiate the conformations within structural levels depending on their nativelikeness, not only those that belong to different structural levels.

1. Introduction

In the accompanying papers,[1,2] we described the theory of our recently proposed hierarchical method for protein potential-energy function optimization,[3] its tests with simple lattice models,[1] and an application to optimize our physics-based united-residue UNRES potential-energy function[3−14] with single proteins, namely, the IgG-binding domain from streptococcal protein G (PDB code 1IGD[15]) and the LysM domain from *E. coli* (PDB code 1E0G[16]). In one accompanying paper,[1] we demonstrated with simple 12-bead cubic-lattice protein models that good folders are obtained only when the energy levels are ordered according to nativelikeness, the best ordering following the pathway of simulated folding. Violation of the correspondence between energy and nativelikeness always resulted in poor folders with long folding times and low stability of the native structure. A wrongly designed hierarchy (i.e., one that does not follow the sequence of folding events, although it follows the increase in the degree of nativelikeness with decreasing energy) can also result in poor folders. These conclusions were extended in the other accompanying paper[2] to an off-lattice protein representation.

On the basis of 1IGD as an example (the native structure of which consists of N-, and C-terminal hairpins, respectively, packed into a β-sheet and a middle α-helix), we demonstrated that the success of the optimization procedure and the transferability of the force field depend critically on the choice of hierarchy. The first hierarchy, in which the number of nativelike elements gradually increased with level number, resulted in a

* Corresponding author. E-mail: has5@cornell.edu. Phone: (607) 255-4034. Fax: (607) 254-4700.
† Cornell University.
‡ University of Gdańsk.

10.1021/jp040329x CCC: $27.50
Published on Web 10/02/2004

reasonable force field. The lowest-energy structure of 1IGD produced by the force field had all nativelike elements; however, the packing of the hairpins was incorrect, and further optimization failed. Moreover, the force field had an overwhelming preference for the β-structure. We concluded that the latter undesirable feature was caused by the fact that optimization was aimed at the appearance of both the N- and the C-terminal hairpins in early folded structures, whereas it is known from experiment that the N-terminal hairpin appears later in folding when the C-terminal hairpin has already been formed and can stabilize it.[17,18] Designing the hierarchy according to the sequence of folding events deduced from experiment[17,18] 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. We also concluded that more than one training protein must be included simultaneously in the optimization to improve the transferability of the force field. In this paper, we demonstrate that hierarchical optimization can be applied with success to more than one training protein at the same time and that the resulting force fields are reasonably transferable.

2. Methods

The theoretical background of the hierarchical optimization method is described in the accompanying papers;[1,2] in the second paper, the details of the algorithm as applied to our united-residue UNRES force field[3−14] are given, and the UNRES force field itself is outlined. In refs 1 and 2, the hierarchical optimization method is also compared with the approaches developed by other authors. We therefore recall only the essentials here.

The structure of each training protein is described in terms of levels. Level 0 contains conformations with no native-structure elements, and level 1 contains conformations with a single element of native secondary structure. The secondary structure is defined in terms of helices, sheets, and strands; they are recognized on the basis of their hydrogen-bonding contact pattern and the local geometry of the chain.[2] Level 2 is defined in terms of the packing of pairs of secondary-structure elements; the packing is defined in terms of backbone hydrogen-bonding contacts (for strands packed into β-sheets) or side-chain contacts (for the remaining elements).[2] The subsequent levels are defined in terms of the appearance of larger clusters of fragments, and the arrangement of fragments is defined both in terms of packing and the root-mean-square deviation (rmsd) from the respective fragment of the experimental structure. The composition and arrangement of levels is termed a structural hierarchy. It should be noted that not only is the number of nativelike elements that increases with level number important but the order in which the fragments appear is also important; this order should follow the folding pathway.[1,2]

As mentioned in the Introduction, in this work we describe the hierarchical optimization of the UNRES force field using multiple training proteins simultaneously. We chose the LysM domain from *E. coli* (an $\alpha + \beta$-protein; PDB code 1E0G[16]), the Fbp28Ww domain from *Mus musculus* (a β-protein; PDB code 1E0L[19]), the albumin-binding GA module (an α-protein; PDB code 1GAB[20]), and the IgG-binding domain from streptococcal protein G (an $\alpha + \beta$-protein; PDB code 1IGD[15]). The experimental structures of these proteins, together with the partition into fragments, are shown in Figure 1. The selection of proteins to test our procedure was based on the following three criteria: (i) size 65 residues or less to be able to carry out computations in real time; (ii) different folds; we selected one α- (1GAB), one β- (1E0L), and two $\alpha + \beta$- (1IGD and 1E0G) proteins; and (iii) diversity in amino acid composition to cover as great a variety of side-chain−side-chain interactions as possible.

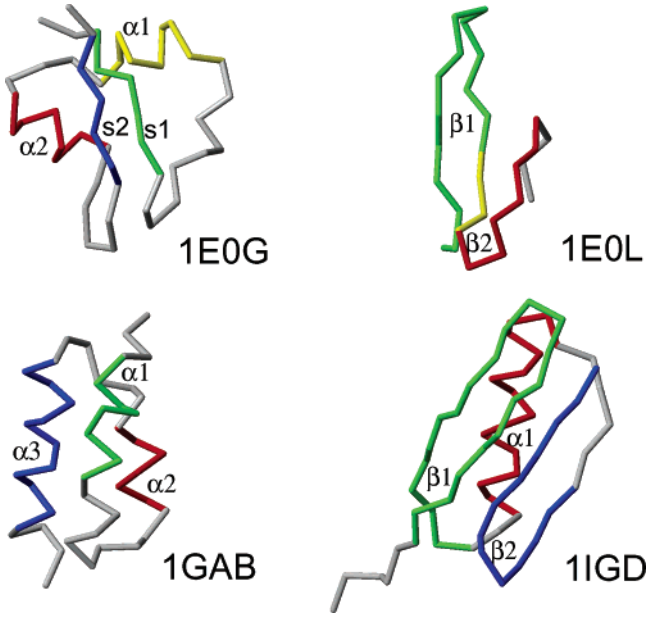

Figure 1. Experimental structures of 1E0G, 1E0L, 1GAB, and 1IGD. The nativelike elements considered in the hierarchical optimization are both color coded and marked with symbols used in the text. For 1E0L, the part of the chain shared by β_1 and β_2 is yellow.

To gain experience in force-field optimization with multiple training proteins, we gradually increased the number of proteins in the training set. Such a scheme enabled us to detect the relationship between the types of folds and amino acid compositions of the proteins from the training set and the predictive power of the resulting force field. In the first stage of our multiprotein UNRES energy function optimization, we used the 1E0G and 1E0L proteins. This choice was based on the following reasons. Of the two proteins (1IGD and 1E0G) used in our previous work to test the hierarchical optimization method with a single training protein,[2] the 48-residue 1E0G, is a smaller one. Of the two training proteins not treated in our previous work, 1E0L (28 residues), also is a smaller one. Therefore, such an initial set of proteins involves the least expensive computations, which was important because initially we had no experience with how the hierarchical optimization method behaves when applied to multiple training proteins. Moreover, from our previous studies,[2] we concluded that the force field obtained by hierarchical optimization using the 1E0G protein alone performs much better in the prediction of the structure of α-helical rather than β- and $\alpha + \beta$-proteins. Therefore, we thought it worthwhile to check how the addition of a β-protein to the training set would affect the predictive power of the resulting force field. This training set and the resulting force field are hereafter referred to as the 2P (two protein) set and force field, respectively. In the next step of our multiprotein parameter optimization, we added 1GAB (an α-helical protein) to the training set, obtaining a set that represents three basic types of protein folds [$\alpha + \beta$ (1E0G), β (1E0L), and α (1GAB)]. This set and the resulting force field will hereafter be referred to as the 3P set and force field, respectively. Finally, to test the influence of the number and size of proteins on the quality of the resulting force field, we added 1IGD, which is the largest

of the proteins considered; this set and the resulting force field will hereafter be referred to as the 4P set and force field, respectively. The addition of 1GAB and 1IGD to the final training set was motivated by the fact that the four selected proteins represent not only all basic types of structures but also the interactions between the majority of side-chain types (154 out of 210).

The optimized parameters included the energy-term weights (eq 1 in ref 2) and the well depths of the side-chain pairwise interaction potential; the orientation-dependent Gay−Berne[21] functional form was used for the latter. All parameters except for the well depths that were optimized here were taken from ref 6. The internal coefficients of the cumulant terms were assigned the values in the F2 force field of ref 2 and were not optimized further. The conformational space annealing (CSA) method,[22] with recent modifications to treat β-proteins,[23] was used to search the conformational space of united-residue chains both for decoy generation and test calculations.

Another objective of this study was to reduce the computational cost of the UNRES potential energy function. One of the main factors increasing the cost of the versions of the UNRES force field derived thus far,[2,3,12] except for its early versions that were unable to handle β-sheets,[6−8,24,25] is the presence of the fifth- and sixth-order terms arising from the cumulant expansion of the restricted free energy (RFE) function of the polypeptide chain.[9,13,14] These terms contain contributions from quadruplets of the peptide groups; therefore, they require a considerable computational effort even though only those peptide groups are considered that are not too far from each other. Initially, the weights of these high-order expansion terms were significant,[12] probably because the analytical expressions were parametrized[9] by using the ECEPP/3 all-atom force field[26] and were not optimized further. However, recently we reparametrized the cumulant terms based on high-level quantum-mechanical ab initio calculations of the energy surfaces of model systems;[13,14] moreover, the internal coefficients of these expressions were treated as variable parameters in hierarchical optimization.[2,14] When examining the weights of the high-order cumulant terms determined by hierarchical optimization in our recent work,[2,14] we concluded that they are smaller than the coefficients of the third- and fourth-order cumulant terms by about 2 orders of magnitude (Table 4 in ref 2); so are these respective contributions to the UNRES total energy. We therefore set these coefficients equal to zero in the recent study to reduce the computational cost.

For 1E0G, we used the same hierarchy as determined in one of the accompanying papers;[2] the hierarchy for 1IGD was a simplified hierarchy 2 of ref 2 (built on the basis of experimental information on the folding of that protein[17,18]). To determine the hierarchies for 1E0L and 1GAB and to build the initial databases, we first carried out hierarchical optimization for each of these proteins in a single-protein mode as described in the other accompanying paper.[2]

The simplified hierarchy for 1IGD used in this work was as follows:

Level 1: α_1 **or** β_2 formed.

Level 2: α_1 **and** β_2 formed.

Level 3: α_1 **and** β_1 **and** β_2 formed, but β_1 and β_2 are not packed to each other.

Level 4: α_1 **and** β_1 **and** β_2 formed **and** β_1 and β_2 are packed to each other, but the rmsd from the experimental structure is above the cutoff value.

Level 5: α_1 **and** β_1 **and** β_2 formed **and** β_1 and β_2 are packed to each other and the rmsd from the experimental structure is above the cutoff value, but α_1 and β_2 are not packed to each other.

Level 6: α_1 **and** β_1 **and** β_2 formed **and** β_1 and β_2 are packed to each other, and the rmsd from the experimental structure is above the cutoff value **and** α_1 and β_2 are packed to each other.

The modification of the 1IGD hierarchy was focused on reducing the number of structural levels and achieving nativelike structures with as low as possible rmsd values from the experimental structure. The simplified hierarchy still corresponds to the folding hierarchy deduced from the available experimental data.[2] The quality of the force field (optimized using only 1IGD), obtained by using the hierarchy presented above, and the F2 force field (optimized by using the 1IGD protein and by assuming hierarchy 2 of ref 2 built according to the experimental data for 1IGD folding) reported in ref 2 are comparable, the rmsd of the lowest-energy structure from the experimental structure being 5.2 Å.

The 1E0L protein is a very small one with a very simple fold. In this case, the simplest fragment assembly scheme (all secondary structure elements form simultaneously and then assemble into the native structure) was proven to be the easiest and the most efficient folding scenario. The applied hierarchy was as follows:

Level 1: β_1 **or** β_2 formed.

Level 2: β_1 **and** β_2 formed, but there is an insufficient number of hydrogen-bonding contacts involving the middle β-strand, which is shared by both hairpins for the three-stranded antiparallel β-sheet to form.

Level 3: The three-stranded antiparallel β-sheet is formed, but the rmsd from the experimental structure is above the cutoff value.

Level 4: The three-stranded antiparallel β-sheet is formed, and the rmsd from the experimental structure is within the cutoff.

For 1GAB, which is a three-helix-bundle protein, the best working hierarchy was found to start with the formation of the C-terminal α-helix or with both the N-terminal and the middle α-helix formed first, followed by the formation of the remaining secondary-structure elements. When all secondary structure elements are present, they are packed in the nativelike fold. We therefore applied the following hierarchy:

Level 1: α_3 **or** α_1 **and** α_2 formed.

Level 2: α_1 **and** α_3 **or** α_2 **and** α_3 **or** α_1 **and** α_2 **and** α_3 formed but are not packed to each other.

Level 3: α_1 **and** α_2 **and** α_3 formed **and** α_2 and α_3 are not packed to each other.

Level 4: α_1 **and** α_2 **and** α_3 formed **and** α_2 and α_3 are packed to each other, but the rmsd from the experimental structure is above the cutoff value.

Level 5: α_1 **and** α_2 **and** α_3 formed **and** α_2 and α_3 **and** α_1 and α_2 are packed to each other, **and** the rmsd from the experimental structure is within the cutoff value.

It should be noted that the above hierarchy for 1GAB folding is identical to the hierarchy for folding found experimentally[27] and theoretically[28,29] for the B domain from *Staphylococcus aureus* protein A (PDB code 1BDD), which also is a three-helix-bundle protein. Both proteins fold into an identical 3D structure; however, their sequence similarity is lower than 30%, and folding is determined by the stability of the C-terminal helix that forms first and induces the formation of the other two helices.

3. Results and Discussion

The initial and final free-energy gaps corresponding to optimizations with the 2P, 3P, and 4P training sets are

TABLE 1: Initial and Final Free-Energy Gaps ($-\Delta F_{ij}$; kcal/mol) Obtained in the Optimization of the 2P UNRES Force Field with 1E0G and 1E0L as the Training Proteins

		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
					1E0G[b]					
0	1	30.0	18.5	45.3	30.0	8.9	34.8	30.0	26.1	43.4
1	2	25.0	14.5	24.1	25.0	12.0	24.7	25.0	17.9	24.3
2	3	20.0	23.9	26.0	20.0	15.9	19.5	20.0	11.8	38.6
3	4				15.0	7.2	14.2	15.0	3.3	21.4
4	5							10.0	10.3	9.8
5	6							5.0	11.5	4.6
					1E0L					
0	1	10.0	12.3	9.7	10.0	11.7	26.6	10.0	7.4	19.4
1	2	7.0	1.8	6.2	7.0	7.3	21.7	7.0	5.0	20.3
2	3				5.0	2.6	17.9	5.0	1.5	20.5
3	4							5.0	11.3	15.8

[a] Levels are numbered as in the text. [b] The initial gaps for the 1E0G proteins differ from the final gaps corresponding to hierarchy 3 reported in Table 8 of ref 2 because $w_{corr}^{(5)}$, $w_{corr}^{(6)}$, and $w_{turn}^{(6)}$ were set equal to zero in the 2P force field; therefore, the initial parameters are not exactly equal to the final parameters of the force field determined using the 1E0G protein and hierarchy 3 for this protein as reported in ref 2.

TABLE 2: Initial and Final Free-Energy Gaps ($-\Delta F_{ij}$, kcal/mol) Obtained in the Optimization of the 3P UNRES Force Field with 1E0G, 1E0L, and 1GAB as the Training Proteins

		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
					1E0G					
0	1	30.0	45.3	32.4	30.0	34.8	52.4	30.0	43.4	69.1
1	2	25.0	24.1	14.3	25.0	24.7	33.8	25.0	24.3	26.4
2	3	20.0	26.0	7.7	20.0	19.5	19.3	20.0	38.6	33.6
3	4				15.0	14.2	22.2	15.0	21.4	27.7
4	5							10.0	9.8	19.7
5	6							5.0	4.6	9.9
					1E0L					
0	1	10.0	9.7	7.7	10.0	11.7	14.2	10.0	19.4	20.5
1	2	7.0	6.2	5.0	7.0	7.3	16.7	7.0	20.3	15.4
2	3				5.0	2.6	14.2	5.0	20.5	13.2
3	4							5.0	15.8	15.4
					1GAB					
0	1	30.0	14.2	24.6	30.0	27.5	41.2	30.0	24.4	32.3
1	2	25.0	32.4	24.0	25.0	39.8	34.1	25.0	34.1	23.8
2	3				20.0	37.5	49.1	20.0	13.5	25.7
3	4							15.0	−4.3	10.4
5	6							5.0		5.7

[a] Levels are numbered as in the text.

summarized in Tables 1, 2, and 3, respectively. The C^{α} rmsd values of the lowest-energy conformations from the experimental structures are summarized in Table 4. The energy-term weights are summarized in Table 5, and the well depths of the side-chain pairwise interaction potential are summarized in Tables S1−S3 of the Supporting Information. It can be seen that all optimizations were successful (i.e., the target free-energy gaps were achieved or exceeded), and the resulting force fields were able to locate the nativelike structure of each of the training proteins as the lowest in energy in a global conformational search, which is demonstrated by the low rmsd values of the lowest-energy conformations from the respective experimental structures (Table 4). This demonstrates that hierarchical optimization can be applied to multiple training proteins. It can also be concluded that the removal of the fifth- and sixth-order correlation terms did not impair the performance of the force field. The third- and the fourth-order terms (both long-range terms and short-range-turn terms) responsible for the coupling between backbone-local and backbone-electrostatic interactions[9] appear to be sufficient to treat both α- and β-structure. In earlier versions of the UNRES force field,[6−8,24,25] the fourth-order terms were present, but the third-order terms were not; this was the reason that the force field could earlier not handle β-structure. Removal of the fifth- and the sixth-order terms in the present work resulted in a 55−58% reduction of computational cost on average.

The transferability of the resulting force fields was tested on a set of several proteins. For every test protein, we carried out three independent CSA runs 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 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.[23]

For the 2P force field, we ran test calculations on a limited set of only 14 proteins. (This set contains the same proteins as

TABLE 3: Initial and Final Free-Energy Gaps ($-\Delta F_{ij}$, kcal/mol) Obtained in the Optimization of the 4P UNRES Force Field with 1E0G, 1E0L, 1GAB, and 1IGD as the Training Proteins

		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
					1E0G					
0	1	30.0	32.4	30.7	30.0	52.4	20.2	30.0	69.1	28.2
1	2	25.0	14.3	18.9	25.0	33.8	7.0	25.0	26.4	16.9
2	3	20.0	7.7	17.0	20.0	19.3	21.1	20.0	33.6	15.9
3	4				15.0	22.2	22.5	15.0	27.7	19.1
4	5							10.0	19.7	23.8
5	6							5.0	9.9	11.2
					1E0L					
0	1	10.0	7.7	8.2	10.0	14.2	24.1	10.0	20.5	25.3
1	2	7.0	5.0	8.0	7.0	16.7	18.2	7.0	15.4	19.1
2	3				5.0	14.2	9.3	5.0	13.2	9.6
3	4							5.0	15.4	14.8
					1GAB					
0	1	30.0	24.6	16.2	30.0	41.2	19.0	30.0	32.3	75.9
1	2	25.0	24.0	14.2	25.0	34.1	19.4	25.0	23.8	29.8
2	3				20.0	49.1	12.2	20.0	25.7	29.8
3	4							15.0	10.4	6.4
4	5							5.0	5.7	3.7
					1IGD					
0	1	30.0	25.8	21.6	30.0	50.7	38.4	30.0	53.7	37.7
1	2	25.0	13.2	18.7	25.0	14.9	15.1	25.0	26.7	26.5
2	3	20.0	6.4		20.0	27.4	27.4	20.0	19.3	18.2
3	4				15.0	34.8		15.0	21.5	5.7
4	5							10.0	−14.8	9.1
5	6							5.0	−30.9	3.1

[a] Levels are numbered as in the text.

TABLE 4: C^α rmsd Values (Å) from the Native Structures of the Lowest-Energy Conformations of the Training Proteins Used to Derive the 2P, 3P, and 4P Force Fields, Respectively, Obtained by a Global Conformational Search with the Final Parameters of the Respective Force Field

	force field		
protein	2P[a]	3P[a]	4P
1E0G	5.11	5.00	4.06
1E0L	4.48	4.23	4.65
1GAB	*9.42*	3.88	2.93
1IGD	*12.95*	*6.33*	5.61

[a] For comparison with the 4P force field, which was optimized using the complete training set, the rmsd's of the lowest-energy structures of 1IGD and 1GAB obtained with the 2P force field and that of 1IGD obtained with the 3P force field are also given (in italics), although these proteins were not used in optimization.

that given in Table 9 of ref 2 except that 1E0L was excluded because it was one of the proteins used to optimize the 2P force field.) The results are summarized in Table 6. It can be seen that the performance of the 2P force field does not improve significantly compared to that obtained by using the 1E0G protein (Table 9 in ref 2). For two proteins (1GAB and 1QHK), remarkably worse and for two other proteins (1CLB and 1POU) remarkably better results, in terms of the length of the longest predicted fragment, were obtained; for the remaining proteins, the results are comparable. It should be noted that the prediction accuracy of β-proteins did not improve despite the fact that a β-protein was added to the training set. This is probably caused by the fact that 1E0L is a small protein with a rather trivial fold; it does not, therefore, make a significant addition to the structural patterns already present in 1E0G. Nevertheless, the exercise with the 2P training set has demonstrated that hierarchical optimization with more than one training protein is feasible.

TABLE 5: Comparison of the Energy-Term Weights Obtained by Hierarchical Optimization Using the 1E0G Protein Alone, Using Two Proteins 2P (1E0G + 1E0L), Three Proteins 3P (1E0G + 1E0L + 1GAB), and Four Proteins 4P (1E0G + 1E0L + 1GAB + 1IGD)

weight	value			
(dimensionless)	1E0G	2P	3P	4P
w_{SCp}	2.64146	2.80030	2.85111	2.73684
w_{el}	0.20803	0.36015	0.36281	0.06833
w_{tor}	1.88208	2.99132	3.00008	2.99546
w_{tord}	2.41019	2.99370	2.89863	2.89720
w_b	2.37760	2.56436	3.95152	4.15526
w_{rot}	0.04803	0.10524	0.15244	0.16761
$w_{loc-el}^{(3)}$	1.39959	1.53346	1.91423	1.98989
$w_{loc-el}^{(4)}$	1.62840	1.99829	1.72128	1.60072
$w_{loc-el}^{(5)}$	0.02730	0.00000	0.00000	0.00000
$w_{loc-el}^{(6)}$	0.00741	0.00000	0.00000	0.00000
$w_{turn}^{(3)}$	2.27520	2.99315	2.99827	2.36351
$w_{turn}^{(4)}$	1.07936	0.92238	0.59174	1.34051
$w_{turn}^{(6)}$	0.02391	0.00000	0.00000	0.00000

For the 3P and 4P force fields, we carried out more extensive tests using 66 (26 α, 15 β, and 25 $\alpha + \beta$) proteins with size ranging from 28 to 144 residues; this set included the proteins of the sets used to test the transferability of the 2P force field and the force fields derived in ref 14 and in one of the accompanying papers.[2] None of these proteins was used to optimize the 3P or the 4P force field. The results of the tests are summarized in Table 7, and the predicted structures of 10 selected proteins (3 α, 3 β, and 4 $\alpha + \beta$) are shown in Figure 2. On the basis of the data summarized in Tables 4, 6 and 7, it can be concluded that the 3P force field has a comparable prediction accuracy to the 2P force field (remarkably better predictions were obtained for 1KOY and 1IGD and remarkably worse for 1CLB and 1POU) and that the 4P force field performs significantly better (remarkably better predictions for 2PTL, 1UBQ, 1ED7, and 1WIU and remarkably worse only for 1POU).

TABLE 6: Results of Tests of the 2P Force Field Obtained by Hierarchical Optimization of 1E0G and 1E0L

protein[a]	length	type	E[b]	rms[c]	n4[d]	n5[e]	n6[f]
1BDD	46	α	0.0	3.2	46	46	46
1GAB	49	α	0.0	9.4	32	35	40
1KOY	62	α	0.0	10.1	25	28	32
			9.5	10.4	27	30	47
1CLB	75	α	0.0	5.3	33	58	75
1POU	71	α	0.0	9.9	34	39	45
			9.8	5.4	34	54	71
1FSD	28	α + β	0.0	4.8	25	28	28
1IGD	61	α + β	0.0	13.0	18	21	24
			4.9	10.8	19	31	40
2PTL	62	α + β	0.0	9.8	19	33	43
1UBQ	76	α + β	0.0	11.5	24	27	44
1QHK	47	α + β	0.0	8.3	28	33	35
1ED7	45	β	0.0	7.6	22	26	32
1BK2	57	β	0.0	11.5	15	18	24
1FYN	62	β	0.0	11.9	14	17	22
1WIU	93	β	0.0	13.5	12	19	24

[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 1FSD, which has less 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 1FSD, which is 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] Relative energy. [c] rmsd from the native structure. [d] Length of the largest contiguous segment within a 4-Å rmsd from the native structure. [e] Length of the largest contiguous segment within a 5-Å rmsd from the native structure. [f] Length of the largest contiguous segment within a 6-Å rmsd from the native structure.

When comparing the 2P and the 3P force fields, 1GAB was excluded, and when comparing the 3P and 4P force fields, 1IGD was excluded because these proteins were used in the optimization of the 3P and the 4P force fields, respectively. It can be seen that the improvement of the 4P force field with respect to 2P (and also 3P) has been achieved for the α + β- and β-proteins. This allows us to conclude that the addition of the 1IGD protein to the training set was essential to improve the quality of the force field.

Table 8 presents a summary of the comparison of the 3P and 4P force fields based on the extended set of 66 proteins mentioned above. It can be seen that the results are generally consistent with those obtained with a smaller set of 12 proteins; however, the prediction accuracy on an extended set is poorer in terms of the length of the matching segment for both the 3P and the 4P force fields. When structures within 10 kcal/mol are considered, the 3P and 4P force fields perform comparably on α-proteins, but the 4P force field is definitely better in predicting the structure of β- and α + β-proteins. However, when only the lowest-energy conformations are considered, the two force fields appear to have comparable predictive power. The 3P force field is, on average, able to predict a 42-residue (55% of the average chain length) and the 4P force field is able to predict a 45-residue (58% of the everage chain length) continuous chain fragment within a 6-Å rmsd, respectively (Table 8). The longest predicted fragment within a 6-Å rmsd cutoff exceeds 70 amino acid residues for the α-proteins, is about 70 residues for the α + β-proteins, but is only 49 residues for the β-proteins. An outstandingly good resolution of 2.3 Å was obtained with the 4P force field for the 67-residue 1LQ7 protein (Table 7 and Figure 2B); this is the best resolution that we ever achieved for a protein of this size. The longest predicted segment for both the 3P (127 residues within the 10 kcal/mol energy cutoff) and the 4P (96 residues within the 10 kcal/mol energy cutoff) force fields corresponds to the 1LPE 144-residue protein, which is a four-helix bundle. The 3P force field found a longer fragment, but first, structures with correctly predicted fragments of this size were also obtained with the 4P force field being, however, about 15 kcal/mol higher in energy than the lowest-energy conformation. Second, the lowest-energy structure found with the 4P force field has a longer correctly predicted fragment (85 residues) than that obtained with the 3P force field (48 residues; Table 7).

It is interesting that the average length of fragments matching the corresponding fragments of a native structure of the test proteins within a 6-Å rmsd cutoff (Table 8) is comparable to the average length of the proteins used to parametrize the 3P (41 residues) and 4P (46 residues). It is therefore tempting to conclude that the size of the proteins used in force-field optimization limits the prediction power of the resulting force field (as far as the size of a correctly predicted fragment is concerned); however, a more extensive study with larger training proteins is needed to prove this. It should also be noted that in our protein training sets the content of residues involved in α-helical structure is 55 and 45% for the 3P and 4P force fields, respectively. In contrast, the number of amino acid residues involved in the β-structure is 30 and 37% for the 3P and 4P force fields, respectively. Moreover, it should be noted that the only β-protein of the training set is 1E0L, which has a very simple β-fold. It is therefore not surprising that both the 3P and the 4P force fields perform best in the prediction of the structure of α-helical proteins and worst in the prediction of the structure of the β-proteins. Increasing the content of the β-structure in the training set used to parametrize the 4P force field (by addition of 1IGD) with respect to the 3P training set resulted in better performance in the prediction of the structure of α + β- and β-proteins. However, most probably, because of the lack of nontrivial β-folds in the training set, the 4P force field still does not perform satisfactorily in the prediction of the structure of β-proteins.

We tried to include more complicated β-folds in the training set with the present approach. Our target was the β-meander folds, an example of which is the 1BK2 protein.[30] However, even with this protein alone, the optimization method failed; we were able to produce a force field that located partially native structures as both the lowest in energy and accessible by a global search, but continuing optimization to locate the completely nativelike structures as the lowest in energy resulted in an unsearchable force field (with which the nativelike structures were the lowest in energy but could not be located by a global conformational search). Trying different hierarchies did not make a significant difference. We found that the most probable cause of the failure of the optimization with nontrivial β-proteins is that the division of the set of conformations into discrete structural levels in which each of which the conformations is not differentiated according to nativelikeness is too crude an approximation of a folding pathway. The conformations within a given structural level can still differ significantly in their nativelikeness; this difference is probably not as important for α- and α + β-proteins where the structural elements can be clearly distinguished. Therefore, ignoring the correspondence between the energy and the nativelikeness within a given level, as done in our present approach, just impairs the detailed resolution of the resulting force field (e.g., we could not achieve a resolution better than 5−6 Å for 1IGD), but the examples

TABLE 7: Results of Tests of the 3P and 4P Force Fields

			3P force field					4P force field								3P force field					4P force field				
protein[a]	length	type	E[b]	rms[c]	n4[d]	n5[e]	n6[f]	E[b]	rms[c]	n4[d]	n5[e]	n6[f]	protein[a]	length	type	E[b]	rms[c]	n4[d]	n5[e]	n6[f]	E[b]	rms[c]	n4[d]	n5[e]	n6[f]
1A1W	83	α	0.0	7.9	42	68	73	0.0	10.1	38	43	51	1LPE	144	α	0.0	29.2	42	46	48	0.0	18.3	57	67	85
1ACX	108	β	0.0	15.2	23	24	28	0.0	15.2	22	27	32				8.9	14.2	84	105	127	6.8	13.8	85	89	96
1ADR	76	α	0.0	13.6	23	35	41	0.0	13.4	14	19	25	1LQ7	67	α	0.0	9.0	46	48	51	0.0	2.3	67	67	67
1AH9	71	β	0.0	13.1	31	33	35	0.0	11.8	26	31	35	1MJC	69	β	0.0	12.7	19	23	32	0.0	12.7	16	21	27
1BBY	69	α + β	0.0	6.0	53	55	68	0.0	6.4	43	52	55									8.2	10.1	15	20	37
1BDD	46	α	0.0	3.5	46	46	46	0.0	5.5	22	42	46	1NKL	78	α	0.0	7.6	34	44	50	0.0	12.5	34	41	45
								3.8	3.3	46	46	46				8.7	11.4	31	36	59					
1BK2	57	β	0.0	11.1	13	17	23	0.0	12.8	14	17	23	1O6X	81	α + β	0.0	11.2	18	30	38	0.0	12.7	29	37	46
1BPI	58	α + β	0.0	10.9	16	28	33	0.0	13.0	15	25	27	1ONC	103	α + β	0.0	13.2	26	30	35	0.0	14.9	20	25	26
								3.3	9.5	19	30	41	1OPD	85	α + β	0.0	15.4	21	25	34	0.0	12.7	24	34	37
1CLB	75	α	0.0	11.7	31	38	44	0.0	5.1	32	58	75									4.6	10.0	51	64	70
1COO	81	α	0.0	10.9	40	52	65	0.0	11.4	42	45	60	1POH	85	α + β	0.0	16.0	26	34	35	0.0	13.0	28	31	35
1CRO	66	α + β	0.0	17.2	18	21	23	0.0	11.9	18	24	37	1POU	71	α	0.0	10.3	29	34	44	0.0	12.2	19	27	30
			7.6	12.9	20	25	33														4.9	11.1	22	50	57
1CSP	67	β	0.0	17.7	12	16	31	0.0	12.4	19	23	34	1PRQ	125	α + β	0.0	13.0	27	30	33	0.0	12.7	23	31	47
1DNY	76	α	0.0	10.7	30	40	44	0.0	11.7	26	29	31				0.9	15.5	29	36	38	6.4	12.8	27	33	52
				8.5	29	42	46	3.1	8.6	31	35	52	1PRU	56	α	0.0	11.8	17	20	24	0.0	12.4	12	15	17
1DVC	98	α + β	0.0	14.1	22	29	33	0.0	12.2	26	35	40				8.9	8.8	28	33	46					
1ED7	45	β	0.0	7.5	19	26	30	0.0	6.6	25	28	35	1PTF	87	α + β	0.0	19.3	22	26	34	0.0	11.9	18	23	25
			7.3	6.6	23	28	38	9.4	5.2	34	43	45				0.6	15.1	22	27	35	7.2	12.9	22	28	34
1EH2	95	α	0.0	13.4	24	38	48	0.0	9.7	36	42	58	1QHK	47	α + β	0.0	9.2	22	27	35	0.0	9.1	23	31	35
			6.0	13.7	48	52	55									8.2	7.1	34	37	43					
1FSD	28	α + β	0.0	5.3	24	26	28	0.0	6.6	22	24	26	1QQV	67	α	0.0	10.4	22	34	36	0.0	11.2	19	23	27
			9.6	3.6	28	28	28	8.3	4.9	25	28	28	1RES	43	α	0.0	14.8	19	23	26	0.0	11.1	14	20	24
1FYN	62	β	0.0	11.8	13	17	24	0.0	10.6	18	23	27				7.1	7.0	20	32	38	9.3	6.0	23	36	43
1G6P	66	β	0.0	11.2	17	20	28	0.0	11.8	15	24	31	1SAP	66	α + β	0.0	15.0	22	24	27	0.0	11.0	29	31	33
1GAK	137	α	0.0	22.0	37	40	41	0.0	22.6	36	39	49				8.4	15.8	25	34	38	9.2	10.2	37	42	51
			1.6	18.5	34	43	57	0.5	21.9	45	68	76	1STU	68	α + β	0.0	14.0	22	23	25	0.0	12.1	16	19	21
1GH5	87	β	0.0	13.5	12	15	24	0.0	12.6	25	30	34				0.8	13.3	22	23	26	3.0	13.1	24	32	34
1GHH	81	α + β	0.0	13.7	33	39	41	0.0	9.6	33	35	41	1TEN	89	β	0.0	13.2	12	17	22	0.0	12.8	25	30	32
1GPT	47	α + β	0.0	18.0	19	21	23	0.0	14.0	20	23	26				7.6	15.3	14	18	31	6.2	12.3	28	31	33
1H40	69	α	0.0	12.8	28	35	39	0.0	14.2	12	14	17	1THX	108	α + β	0.0	15.9	24	29	35	0.0	13.8	29	39	42
			4.7	7.1	40	62	65	9.1	11.3	31	49	54	1TNS	76	α + β	0.0	10.5	20	23	27	0.0	12.4	16	21	23
1HS7	97	α	0.0	9.8	44	60	67	0.0	9.4	61	67	71	1TPM	50	β	0.0	11.9	10	13	16	0.0	12.8	21	28	30
			2.5	9.5	61	66	70									5.7	9.7	15	22	24	7.8	6.1	41	46	49
1HYP	75	α	0.0	9.8	22	27	40	0.0	9.7	37	43	51	1UBQ	76	α + β	0.0	14.1	19	26	39	0.0	13.0	26	29	41
			8.3	8.7	28	39	47	6.5	9.5	32	46	57									1.8	11.8	35	41	54
1IMQ	86	α	0.0	8.1	40	47	75	0.0	8.9	38	44	54	1UXC	50	α	0.0	7.8	19	25	33	0.0	9.2	30	33	35
1IYU	79	β	0.0	17.2	11	16	21	0.0	12.8	17	22	25				9.8	8.4	30	33	36	1.4	7.7	31	35	44
1J7O	76	α	0.0	7.2	38	42	52	0.0	8.7	26	35	48	1VCC	77	α + β	0.0	11.6	22	34	41	0.0	11.4	19	27	35
			5.7	6.2	41	43	71	9.4	5.7	48	60	76	1VIG	71	α + β	0.0	10.8	27	33	34	0.0	11.8	26	33	37
1J7P	67	α	0.0	10.4	24	28	40	0.0	10.0	26	33	49	1WIU	93	β	0.0	12.4	20	25	34	0.0	14.0	32	36	41
			2.1	11.3	21	30	43	7.7	9.3	45	55	60	2ACY	98	α + β	0.0	14.7	20	27	34	0.0	14.2	21	32	37
1JWE	114	α	0.0	14.2	32	36	42	0.0	13.7	33	38	44	2FMR	65	α + β	0.0	12.6	19	27	32	0.0	11.7	25	28	31
1K40	126	α	0.0	19.6	42	43	45	0.0	16.5	34	43	46				7.1	10.6	26	34	37	9.4	7.6	25	42	47
			9.0	10.6	52	57	60						2PTL	62	α + β	0.0	11.6	34	38	41	0.0	12.7	17	19	26
1KOY	62	α	0.0	11.8	27	31	36	0.0	11.0	21	24	32									7.0	7.2	42	51	59
			2.0	6.7	27	42	54	4.6	10.4	27	30	41	2YGS	92	α	0.0	13.9	33	40	45	0.0	14.8	46	63	67
1LEA	72	α + β	0.0	8.9	26	45	58	0.0	11.1	29	35	42				8.1	15.6	39	53	63	7.3	14.4	49	66	70
								0.8	10.7	30	48	54	3NCM	92	β	0.0	14.5	17	24	30	0.0	13.6	16	22	27
													4AIT	74	β	0.0	13.1	17	26	30	0.0	13.7	19	22	30

[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 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 and that fragment is at least 10 residues longer than the longest fragment of the lowest-energy conformation. 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] Relative energy. [c] rmsd from the native structure. [d] Length of the largest contiguous segment within a 4-Å rmsd from the native structure. [e] Length of the largest contiguous segment within a 5-Å rmsd from the native structure. [f] Length of the largest contiguous segment within a 6-Å rmsd from the native structure.

discussed in this and in one of the accompanying papers[2] strongly suggest that with an appropriately chosen hierarchy optimization can be accomplished. Conversely, the β-proteins contain a significant amount of weakly organized structure; therefore, failure to carry out optimization to ensure not only the decrease of the energy with increasing level number but also the decrease of energy within a level with increasing nativelikeness results in the failure of the procedure. We are now upgrading the hierarchical optimization method to address this issue.

4. Conclusions

In this work, we used our recently developed method of hierarchical optimization of protein potential-energy landscapes to optimize the UNRES force field[3−7,9−14] using multiple training proteins. We demonstrated that the method can be

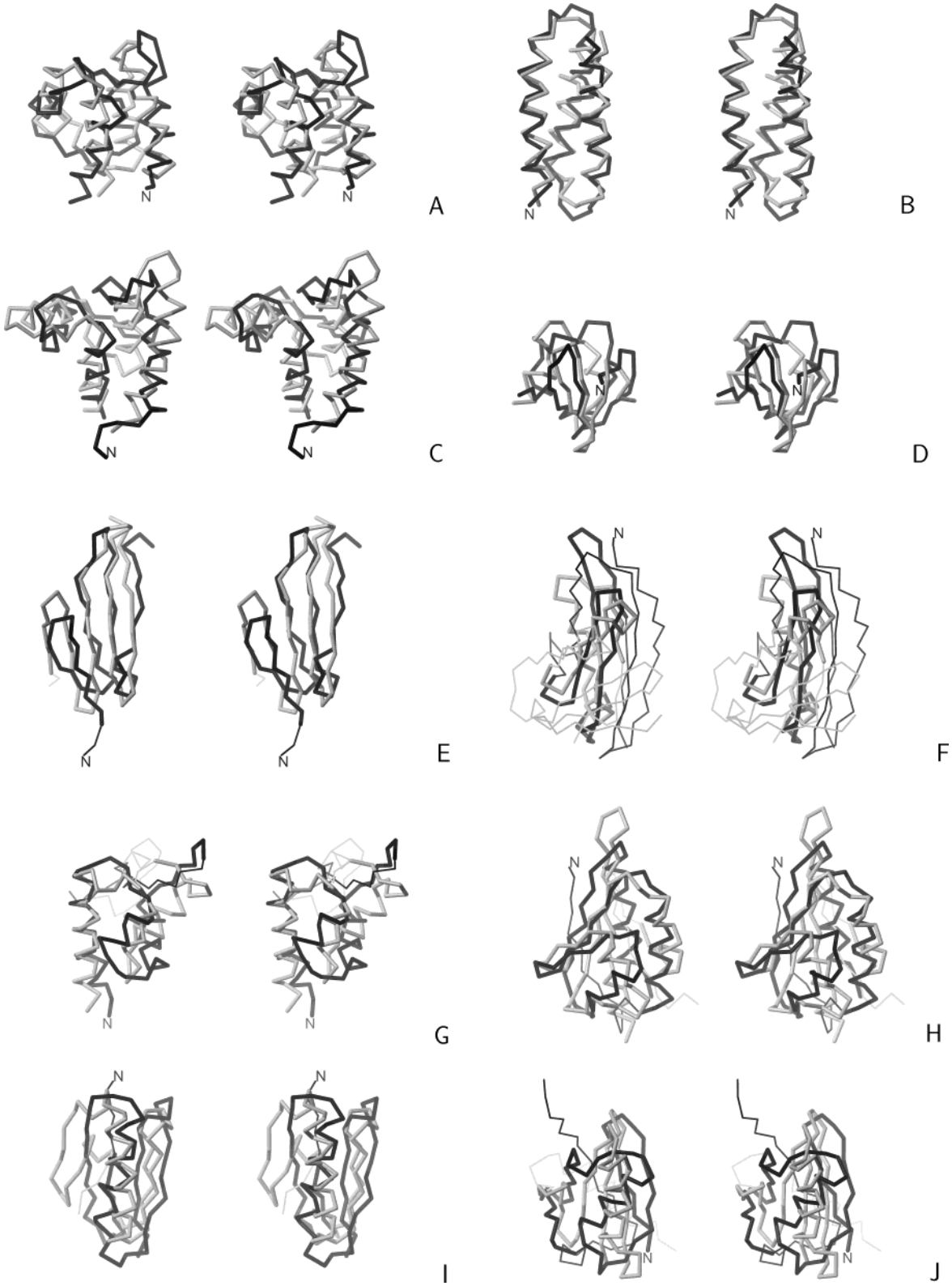

Figure 2. Stereoviews of the superposition of the experimental (black) and best predicted within 10 kcal/mol energy cutoff (gray) structures of selected proteins used to test the force fields obtained in this work. The predicted structures were obtained with the 4P force field. Cylinders mark the C^{α} traces of the fragments of the experimental and computed structures matching within the 6-Å rmsd cutoff, and lines mark the C^{α} traces of the remaining fragments. See Table 7 for energies, rmsd values, and the lengths of matching segments. (A) 1CLB (an α-protein, lowest-energy conformation); (B) 1LQ7 (an α protein, lowest-energy conformation); (C) 1J7O (an α-protein, 9.4 kcal/mol above the global minimum); (D) 1ED7 (a β-protein, 9.4 kcal/mol above the global minimum); (E) 1TPM (a β-protein, 7.8 kcal/mol above the global minimum); (F) 1WIU (a β-protein, lowest-energy conformation); (G) 1BBY (an α + β-protein, lowest-energy conformation); (H) 1OPD (an α + β-protein, 4.6 kcal/mol above the global minimum); (I) 2PTL (an α + β-protein, 7.0 kcal/mol above the global minimum); (J) 1UBQ (an α + β-protein, 1.8 kcal/mol above the global minimum).

applied successfully by this procedure. The resulting force field trained on a set of four proteins gives good medium-resolution predictions of the structures of most of the α- and the α + β-proteins, but it still cannot handle nontrivial β-motifs suf-

TABLE 8: Comparison of the Shortest, Longest, and Average Segment Lengths and Percentage of the Chain Length Predicted within a 4 (n4)-, 5 (n5)-, and 6 (n6)-Å rmsd, Respectively, with the 3P and 4P Force Fields On the Basis of the Set of 66 Test Proteins of Table 7

	best predictions within 10 kcal/mol						lowest-energy conformations					
	3P			4P			3P			4P		
	n4	n5	n6	n4	n5	n6	n4	n5	n6	n4	n5	n6
						α-Proteins (26 Proteins)						
shortest	20	30	36	12	15	17	17	20	24	12	14	17
longest	84	105	127	85	89	96	46	68	75	67	73	85
average	36	46	55	37	47	54	32	39	46	32	40	47
ave % length	45	56	68	45	57	67	39	48	57	40	50	59
						β-Proteins (15 Proteins)						
shortest	11	15	21	14	17	23	10	13	16	14	17	23
longest	31	33	38	41	46	49	31	33	35	32	36	41
average	17	22	29	23	28	34	16	21	27	21	26	31
ave % length	23	29	39	31	38	45	22	28	37	28	35	42
						α + β-Proteins (25 Proteins)						
shortest	16	21	23	16	21	23	16	21	23	15	19	21
longest	53	55	68	51	64	70	53	55	68	43	52	55
average	25	32	37	28	35	42	24	30	35	24	30	35
ave % length	33	42	49	36	47	55	32	40	47	31	39	46
						All Proteins (66 Proteins)						
shortest	11	15	21	12	15	17	10	13	16	12	14	17
longest	84	105	127	85	89	96	53	68	75	67	73	85
average	28	35	42	30	38	45	25	32	38	27	33	39
ave % length	36	45	55	39	49	58	33	41	49	34	43	50

ficiently well. Attempts to include such β-proteins (e.g., 1BK2) in optimization failed. The most probable reason for this is that many proteins, particularly β-proteins, do not contain well-defined secondary structural elements. To address this problem, we are currently extending the hierarchical optimization method to include quantitative measures of the correlation of energy with the nativelikeness of both individual native-structure elements or their groups that constitute hierarchy levels and the whole structure. Preliminary results suggest that introducing such correlations both increases the resolution of the force field and extends the treatment to proteins with poorly defined secondary structure or such structure in which secondary structural elements are not stable alone but their folding is induced by the formation of other elements.[31] (In the current version of the method, we can define only topological features of protein fragments or the whole protein, and we do not have control over the quantitative similarity to the native structure.)

As mentioned in the Methods section, in the present work we optimized only the energy-term weights (eq 1 in ref 2) and the well depths of the side-chain pairwise interaction potentials. Because of the amino acid composition of the proteins used as a training set, we are able to optimize only 154 out of 210 side-chain interactions. Also, of the 154 optimized side-chain parameters, 78 of the interactions appear only once. We can expect that optimizing all 210 parameters with better statistics will lead to a force field with better prediction power.

Despite continuing problems with the β-proteins, it should be noted that the work reported in this and in the two accompanying papers[1,2] is a great step forward with respect to our earlier efforts to optimize the UNRES force field by using the energy gap and *Z*-score optimization.[10−12] With our former approach, we were unable to derive a single force field good for all structural classes; instead, we had to derive separate force fields for the α, β, and α + β structural classes.[12] This was because only very simple β- or α + β-folds could be considered when optimizing the energy gap and *Z*-score alone, which impaired the transferability. Therefore, the force field derived for α-proteins in ref 12 performed within this structural class significantly better than those derived for the β- and the α + β-proteins within their respective classes. With the hierarchical optimization, we were able to include more complicated α + β-folds together with α-folds and simple β-folds. As a result, for the first time, the 4P force field derived in this work is able to treat all structural classes without ancillary information from structural databases.

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 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 of the side-chain pairwise interaction potential. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

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

(2) Ołdziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16934−16949.

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

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

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

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

(7) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874−887.
(8) 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.
(9) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Chem. Phys.* **2001**, *115*, 2323−2347.
(10) Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7291−7298.
(11) 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.
(12) 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.
(13) Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. A* **2003**, *107*, 8035−8046.
(14) Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kozłowska, U.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 9421−9438.
(15) Derrick, J. P.; Wigley, D. B. *J. Mol. Biol.* **1994**, *243*, 906−918.
(16) Bateman, A.; Bycroft, M. *J. Mol. Biol.* **2000**, *299*, 1113−1119.
(17) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584−590.
(18) Kuszewski, J.; Clore, G. M.; Gronenborn, A. M. *Protein Sci.* **1994**, *3*, 1945−1952.
(19) Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H. *Nat. Struct. Biol.* **2000**, *7*, 375−379.
(20) Johansson, M. U.; de Chateau, M.; Wikstrom, M.; Forsen, S.; Drakenberg, T.; Bjorck, L. *J. Mol. Biol.* **1997**, *266*, 859−865.
(21) Gay, J. G.; Berne, B. J. *J. Chem. Phys.* **1981**, *74*, 3316−3319.
(22) Lee, J.; Scheraga, H. A. *Int. J. Quantum Chem.* **1999**, *75*, 255−265.
(23) Czaplewski, C.; Liwo, A.; Pillardy, J.; Ołdziej, S.; Scheraga, H. A. *Polymer* **2004**, *45*, 677−686.
(24) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025−2030.
(25) Liwo, A.; Lee, J.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5482−5485.
(26) Némethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. *J. Phys. Chem.* **1992**, *96*, 6472−6484.
(27) Myers, J. K.; Oas, T. G. *Nat. Struct. Biol.* **2001**, *8*, 552−558.
(28) Alonso, D. O. V.; Daggett, V. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 133−138.
(29) Ghosh, A.; Elber, R.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10394−10398.
(30) Martinez, J. C.; Pisabarro, M. T.; Serrano, L. *Nat. Struct. Biol.* **1998**, *5*, 721−729.
(31) Ołdziej, S.; Liwo, A.; Łągiewka, J.; Czaplewski, C.; Scheraga, H. A. To be submitted for publication, 2004.