Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests


*Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301; †Faculty of Chemistry, University of Gdańsk, Sobieskiego Str. 18, 80-952 Gdańsk, Poland; ‡Instituto de Matemática Aplicada San Luis, Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina, Facultad de Ciencias Físico Matemáticas y Naturales, Universidad Nacional de San Luis, Ejército de los Andes 950, 5700 San Luis, Argentina; §Cornell Theory Center, Cornell University, Ithaca, NY 14853-3801; and ¶Department of Chemistry, Chungbuk National University, Cheongju, Chungbuk 361-763, Korea

Contributed by H. A. Scheraga, March 31, 2005

Recent improvements in the protein-structure prediction method developed in our laboratory, based on the thermodynamic hypothesis, are described. The conformational space is searched extensively at the united-residue level by using our physics-based UNRES energy function and the conformational space annealing method of global optimization. The lowest-energy coarse-grained structures are then converted to an all-atom representation and enerminimized with the ECEPP/3 force field. The procedure was assessed in two recent blind tests of protein-structure prediction. During the first blind test, we predicted large fragments of proteins of the sequence under study (such as, e.g., the handedness of loops, their assemblies; no filters, based on the similarity to known proteins). The search procedure is, therefore, fully physics-based. To locate global energy minima, we developed the Conformational Space Annealing (CSA) (29, 33, 34) method, which is based on a genetic algorithm and local energy minimization. The energy is the only scoring function used in the search and to rank the resulting structures; no filters, based on the similarity to known structures or common features of folds characteristic of a given part of the sequence under study (such as, e.g., the handedness of loops, the directions and packing of β-strands in composite β-sheets, etc.), are used. The search procedure is, therefore, fully physics-based. To

methods based on this hypothesis require both the design of a reliable energy function to describe protein energy surfaces and implementation of a reliable method for global minimization.

At present, it is not possible to find the global minimum of a potential energy function at the all-atom level for proteins with sizes larger than ∼50 residues; the problem is even more complex, because solvent must be included. We, therefore, designed a hierarchical approach (11–13) in which the conformational space is searched extensively by using the physics-based united-residue (UNRES) force field (11–29) developed in our laboratory, which is the key component of the method. The UNRES force field has been derived (11, 18, 21, 25, 26) by averaging the all-atom energy of a system composed of protein plus solvent with implementation (21) of Kubo’s theory of cluster cumulants (30). Furthermore, it has been parameterized (24, 27, 28) to achieve a hierarchical structure of the potential-energy landscape, which is the fundamental feature of the energy landscapes of proteins (31) and the necessary condition for foldability (31, 32). Thus, the underlying principles of UNRES are the physics of the interactions in polypeptide chains and the physical features of protein energy landscapes. Consequently, this force field belongs to the physics-based, and not knowledge-based, force fields, although it was, in part, parameterized by using structural data as were all force fields developed for molecular mechanics and dynamics and even the semiempirical methods of quantum mechanics. Moreover, the ability of UNRES to reproduce regular protein-like structures depends entirely on the presence of the multibody terms derived based on the cluster-cumulant expansion of the restricted free energy of polypeptide chains (30). Knowledge-based force fields are derived based on structural databases exclusively both in functional form and parameterization (3) and usually incorporate structural information explicitly in the pseudoenergy (scoring) function; an extreme example is homology-based methods in which the scoring function reflects the similarity of a target sequence to the sequences of known proteins.

To locate global energy minima, we developed the Conformational Space Annealing (CSA) (29, 33, 34) method, which is based on a genetic algorithm and local energy minimization. The energy is the only scoring function used in the search and to rank the resulting structures; no filters, based on the similarity to known structures or common features of folds characteristic of a given part of the sequence under study (such as, e.g., the handedness of loops, the directions and packing of β-strands in composite β-sheets, etc.), are used. The search procedure is, therefore, fully physics-based. To

Abbreviations: CASP, Critical Assessment of Techniques for Protein Structure Prediction; CSA, conformational space annealing; GDT, TS, global distance test total score; PDB, Protein Data Bank; rmsd, rms deviation; SC, side chains; UNRES, united-residue.

To whom correspondence should be addressed. E-mail: has8@cornell.edu.

© 2005 by The National Academy of Sciences of the USA
facilitate the search, information from secondary-structure prediction can be used to construct starting structures, but the energy remains the only scoring function, and this modification does not, therefore, change the fact that the search is physics-based. The lowest-energy structures from the search are converted to an all-atom representation using a physics-based methodology (35, 36) and energy-minimized with the ECEPP/3 all-atom force field (37).

In this work, we present the performance of the hierarchical approach in two recent community-wide blind-prediction experiments, the fifth and sixth community wide experiments on the Critical Assessment of Techniques for Protein Structure Prediction, (CASP5 and CASP6, respectively).

Methods
The UNRES model of polypeptide chains and the associated energy function are described in our earlier papers (11–29), and only a brief summary is presented here. The polypeptide chain is represented as a sequence of backbone virtual-bond dihedral angle, and 

\[
\phi_i = \theta_i + \gamma_i + \lambda_i + \tau_i,
\]

is represented as a sequence of only a brief summary is presented here. The polypeptide chain and the associated

\[
U = \sum_{i<j} U_{SCSC} + w_{SCP} \sum_{i} U_{SCP} + w_1 \sum_{i} U_{p}\}

\[
+ w_{corr} \sum_{i} U_{corr}(\gamma_i, \gamma_i, \gamma_i, \gamma_i, \gamma_i),
\]

\[
+ w_0 \sum_{i} U_{\delta}(\theta_i) + w_0 \sum_{i} U_{\delta e}(\epsilon_i, \beta_i) + \sum_{i=3}^{6} w_{corr} U_{corr}^{(i)},
\]

where \(\theta_i\) is the \(i\)th backbone virtual-bond axis, \(\gamma_i\) is the \(i\)th backbone virtual-bond dihedral angle, and \(\alpha_i\) and \(\beta_i\) are two angles that define the location of the \(i\)th SC with respect to the \(C^\alpha\)... \(C^\alpha\)... \(C^\alpha\) frame (see, e.g., figure 1 in ref. 14). The parameters of \(U_{SCSC}, U_p,\) and \(U_{corr}\) were derived (14, 15) from the distribution and correlation functions calculated from the PDB (38), whereas those of \(U_{corr}^{(i)} - i = 3, 4, 5, 6\), \(U_{corr}, U_{corr}\), and \(U_{corr}\) were calculated by fitting the corresponding analytical expression to the restricted free energy surfaces of model all-atom systems calculated by using high-level quantum-mechanical \textit{ab initio} methods (25, 26).

Use of protein-structural data is not, however, necessary to derive the parameters of \(U_{SCSC}, U_p,\) and \(U_{corr}\), and was motivated only by lack of sufficient resources; these terms are currently being reparametrized based on the results of our simulation studies of the potentials of mean force of models of SC in water (39, 40) and quantum-mechanical calculations of the energy surfaces of terminally blocked amino acid residues (25), respectively.

The weights of the terms in Eq. 1 were determined, and the parameters of \(U_{corr}\) and \(U_{SCSC}\) were refined, by using our algorithm of hierarchical optimization of the potential-energy function (24, 26–28). The main assumption of this algorithm is that the energy landscape has a hierarchical structure with energy decreasing as the number of native-like elements increases in an ordered way resembling the sequence of folding events. Consequently, the conformational space is divided into “levels,” with each level containing conformations with a similar degree of native-likeness. Level 0 contains no native-like elements, level 1 contains single native secondary-structure elements, and higher levels contain gradually increasing native-like segments. Optimization of a potential-energy function is aimed at achieving the hierarchical structure of the energy landscape, by forcing appropriate free-energy gaps between hierarchy levels to place their energies in descending order. Initially, the procedure was used to determine only the weights of the energy terms in Eq. 1 (24), but subsequently it was used to optimize the \(U_{corr}^{(i)} - i = 3, 4, 5, 6\) terms in addition (26–28), and the current version of the procedure can be used to optimize the parameters of any energy term in Eq. 1. In the first tests of the optimization procedure, we used only PDB ID code 1IGD (an \(\alpha+\beta\) protein) to train the UNRES force field (24, 26, 27).

The most recent version of the force field was derived by using four training proteins with PDB ID codes 1BDD (\(\alpha\), 1E0L (\(\beta\), 1E0G (\(\alpha+\beta\)), and 1IGD (\(\alpha+\beta\)) (28); this force field was termed 4P. Intensive tests of the 4P force field using a set of 66 benchmark proteins with various structural classes (\(\alpha\), \(\beta\), and \(\alpha+\beta\)) and sizes (from 28 to 140 residues) (28) demonstrated that it predicts up to 96-residue fragments of proteins correctly without ancillary information from structural databases, thus surpassing all previous versions of UNRES in terms of predicting ability (28). Only \(\beta\)-protein predictions are limited to very simple folds at present (28). Moreover, while optimizing the 4P force field, we found that the weights of the fifth and sixth order correlation terms \(U_{corr}^{(i)} - i = 3, 4, 5, 6\) decrease to zero during optimization (28), and, consequently, we removed the fifth- and the sixth-order correlation terms from the force field. This modification did not impair the performance of the force field, but the computational cost of the energy function decreased by \(=40\%\), thereby enabling us to treat larger proteins.

In parallel with energy-function development, we improved our conformational search technique. To search conformational space efficiently, we developed a global optimization method, CSA (33). The CSA method is based on a genetic algorithm in which populations are generated by a set of crossover operators. We found that, in the case of proteins with sizes of \(>80\) residues, the CSA method very often fails to locate low-energy structures. The correlation of the success rate with size points to the deficiency of the search method. In our earlier work (17, 34), we found that the success rate in the CSA search depends strongly on the set of operators that are used to create new conformations. In the previous version of CSA, the crossover operators copied secondary structure elements (\(\alpha\)-helices or \(\beta\)-hairpins) from one parent conformation to another one, and we found that such simple crossover operators are usually efficient to predict structures of proteins with very simple folds (usually only helical proteins). To be able to treat all structural classes of proteins, we introduced more complicated crossover operators, which are able to exchange some patterns of nonlocal contacts (for example, nonlocal \(\beta\) structures) between parent conformations (34). Introduction of this type of crossover operators now enables us to predict more complicated folds. We also introduced a procedure for dynamic formation/breaking of disulfide bonds in the CSA algorithm. This attempt predicts disulfide bond patterns based on energetic criteria (29). The first version of the algorithm was implemented and tested on several proteins with disulfide bonds with some success. Moreover, we improved parallelization of the CSA code to be able to treat large proteins. At present, our global-optimization CSA procedure scales very well up to 1,000 processors with a peak performance of 90% and an average performance of 75% with 1,000 processors (34), which enables us to predict the structures of proteins containing up to 250 residues.
CASP5 when three different force fields were used depending on ever, for all targets, the searches also were always started from Ołdziej (41) definitely pointed to /H9251 the double-torsional terms to Eq. The CASP4 force field was improved for CASP5 by introduction of quantum-mechanical torsional, correlation, and backbone-electrostatic terms based on the five lowest-energy families were selected as candidate models every protein, at least three runs were carried out with different possibilities of using one force field for all structural classes in CASP5 was a great step forward. Additionally (mainly in the beginning of the clustering procedure. In the CASP6 experiment, we used only the 4P force field (28) (see Methods), which is the most developed version of UNRES at present. During the CASP5 and CASP6 exercises, we used almost identical protocols for predicting the 3D structures of the target proteins. The global conformational search was performed by using the CSA method with the UNRES force field (see Methods). For every protein, at least three runs were carried out with different search parameters to assure good exploration of conformational space. The final conformations from all searches were collected and subjected to a cluster analysis. The lowest-energy conformations of the five lowest-energy families were selected as candidate models and converged to an all-atom representation (35, 36). Because of the high computational demands of energy-based prediction, we did not consider targets with lengths of >200 amino acid residues, unless they potentially could be new folds. In both the CASP5 and CASP6 exercises, information from secondary structure prediction by PSIPRED (41) was used to speed up the search for larger proteins. In CASP6, the secondary structure information was used only to generate the initial structures for unrestricted CSA searches; however, for all targets, the searches also were always started from random structures, and the resulting structures from all searches were ranked based on energy alone. Because the force field available during the CASP5 exercise was not yet mature, in that exercise we additionally performed searches with the secondary structure constrained by using information from PSIPRED. The only element of human intervention/decision in the procedure described above is choosing the rms deviation (rmsd) cut-off value in the clustering procedure.

Results and Discussion Participation in the CASPS Experiment. The CASP5 experiment took place between June and September 2002. Sixty-seven targets, whose structures were in the process of being determined during the exercise, were made available for prediction; of those targets, five were later assessed as new folds (42). We submitted the predictions for 27 targets, 5 models per each. A summary of our predictions, including relative energies, similarity to the corresponding experimental structures, and ranking of our prediction for two targets classified as new folds, is presented in Table 1. For most of the targets, we correctly predicted the structures of 50–79 residue fragments irrespective of the type of structures (α, α+β, or β). This result was a clear improvement over our results in the CASP4 exercise in which we predicted only some very short fragments with β-structure components (22); in particular, we made an outstanding prediction of the 115-residue C-terminal domain of target T0149 (the YJIA protein from Escherichia coli, PDB ID code 1NJF; α+β structure) that was classified as a new fold. A 74-residue fragment (64% of the whole sequence length) of our model 2 (which was our best prediction of this target) matched the experimental structure within 6-Å rmsd (see Fig. L4 and Table 1). Our model 2 of T0149 (Fig. L4) was ranked as the second best model submitted by all predictors for this target when CASPS 5 assessors used “visual inspection.” When, as is typical for CASP exercises, the global distance test total score (GDT. TS) (43) measure was used, our model 2 of T0149 (Fig. L4) was ranked as the first over all of the predictions (44). Despite the very good geometrical match between the experimental structure and our model, it should be mentioned that the overall topology of our prediction is especially different in the packing of β-strands. Nevertheless, this prediction is, to our knowledge, the first example that a carefully optimized UNRES force field is capable of predicting a large continuous fragment of a protein with β structure. Another successful prediction from the CASP5 experiment is that of the C-terminal domain of target T0129

Table 1. Summary of the performance of the hierarchical protocol for protein-structure prediction in CASPS

<table>
<thead>
<tr>
<th>Target</th>
<th>Length*</th>
<th>Type</th>
<th>ΔEf, kcal/mol</th>
<th>C&lt; sup&gt;4 &lt;/sup&gt; rmsd, Å</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>GDT. TS, %</th>
<th>Rank</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T0129</td>
<td>170</td>
<td>α</td>
<td>0.0</td>
<td>16.6</td>
<td>43</td>
<td>52</td>
<td>58</td>
<td>24.70</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
<td>13.3</td>
<td>21.0</td>
<td>35</td>
<td>46</td>
<td>57</td>
<td>16.91</td>
<td>239</td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
<td>15.9</td>
<td>14.4</td>
<td>56</td>
<td>62</td>
<td>79</td>
<td>22.65</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td></td>
<td></td>
<td>0.0</td>
<td>14.2</td>
<td>53</td>
<td>57</td>
<td>83</td>
<td>23.97</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Model 4</td>
<td></td>
<td></td>
<td>9.3</td>
<td>18.4</td>
<td>32</td>
<td>36</td>
<td>44</td>
<td>15.73</td>
<td>293</td>
<td></td>
</tr>
<tr>
<td>T0149.2</td>
<td>126</td>
<td>α+β</td>
<td>0.0</td>
<td>13.3</td>
<td>29</td>
<td>44</td>
<td>80</td>
<td>28.88</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
<td>0.4</td>
<td>11.9</td>
<td>54</td>
<td>68</td>
<td>74</td>
<td>31.25</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
<td>2.4</td>
<td>12.8</td>
<td>34</td>
<td>38</td>
<td>43</td>
<td>23.28</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td></td>
<td></td>
<td>3.8</td>
<td>10.8</td>
<td>34</td>
<td>41</td>
<td>45</td>
<td>25.22</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Model 4</td>
<td></td>
<td></td>
<td>4.5</td>
<td>12.9</td>
<td>34</td>
<td>45</td>
<td>85</td>
<td>29.74</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Model 5</td>
<td></td>
<td></td>
<td>4.5</td>
<td>12.9</td>
<td>34</td>
<td>45</td>
<td>85</td>
<td>29.74</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

*Number of amino acid residues in the experimental structure.
1UNRES relative energy.
Length of the longest continuous fragment that fits the corresponding fragment of the experimental structure within 4, 5, and 6 Å, respectively.
GDT.TS = (GDT.P1 + GDT.P2 + GDT.P4 + GDT.PB)/4, where GDT.Pn denotes percent of residues under distance cutoff ≤ n Å (43).
Ranking (according to GDT.TS) among all models of that target submitted by all groups participating in CASPS; see http://predictioncenter.llnl.gov/casp5/pubResults/CASP.BROWSER.
These models were produced with the α force field developed earlier (20) to treat α-helical proteins. Models 4 and 5 were produced with the CASPS force field (26).
(protein ygfB from Haemophilus influenzae, PDB ID code 1IZM), an all-α-helical, 182-residue protein, classified as a new fold. The protein is composed of seven α-helices and can be treated as a tightly packed two-domain protein. The first four α-helices form a distorted up-and-down four-helix bundle, whereas the rest form a left-handed three-helix bundle. In all five submitted models, we predicted the structure of the C-terminal part of the protein correctly. Judging by the similarity to the topology of the experimental structure, our best prediction was model 3 in which fragment 95–171 (79 residues) matched the experimental structure within 6-Å rmsd (see Fig. 1B and Table 1). It should be noted, however, that models 1 and 4 are better than model 3 when judging by the GDT.TS measure (Table 1), and, consequently, these models are ranked better. In the previous CASP experiments (12, 18, 22), we were quite successful in predicting the structures of α-helical proteins; however, our model 3 of target T0129 contained the longest continuous fragment of the protein predicted within a 6-Å rmsd, until CASP6 when further improvements were made.

**Participation in the CASP6 Experiment.** CASP6 took place between April and September 2004. Sixty-two targets whose structures were in the process of being determined during the exercise were made available for prediction. We submitted predictions for 32 targets, with 5 models each. A summary of selected predictions discussed below is presented in Table 2.

The most successful prediction was that of target T0215 (a hypothetical membrane protein from Thermoplasma acidophilum, PDB ID code 1X9B; an α-helical 53-residue protein). This protein possesses a very common and simple fold of a three-helix bundle with a kinked middle α-helix. Our model 1 matched the experimental structure within 3.5-Å rmsd over all C\(^\alpha\)s, and this prediction was the best over all predictions submitted by all groups participating in the CASP6 experiment (see Fig. 1C and Table 2). The CASP assessors classified this target as a fold recognition/analogy target, which means that a good nonhomologous template exists among structures deposited in the PDB. However, with this template, other groups obtained structures with rmsd of only \(\approx 5\) Å from the native structure, and the homology was so weak that many groups, using fold recognition methods, missed the correct template and submitted completely wrong predictions. The prediction for target T0215 (a very small, all-α-helical protein with a very common fold) is not a critical test of significant progress of our protein structure prediction methodology; however, this example shows that our physics-based method can perform much better than comparative modeling or fold recognition methods even for a known fold.

Another example of a very good prediction is target T0281 (a hypothetical protein from Thermus thermophilus, PDB ID code 1WHZ; a 70-residue \(\alpha+\beta\) protein). The structure of this protein is composed of a central three-stranded β-structure flanked by one N-terminal α-helix and two C-terminal α-helices and was classified by CASP assessors for a fold recognition/analogy target. Our model 1 matches the experimental structure with 5.5-Å rmsd over all C\(^\alpha\)s (see Fig. 1D and Table 2). The topology of our model almost matches the topology of the native structure; the only difference is a large deformation of the short C-terminal α-helix. It also is very surprising that our model is much less tightly packed than the native structure, whereas energy minimization used in search procedures usually produces very tightly packed structures with very regular secondary-structure elements. This example, to our knowledge, is the first in which UNRES predicted a complete structure of an \(\alpha+\beta\) protein, and, consequently, this result can be considered a significant step forward, even though our prediction is ranked only as 41 according to the GDT.TS measure (Table 1) because of the presence of good templates for this protein in the PDB, which gave an advantage to the knowledge-based methods.

For targets T0215 and T0281, our correct predictions were the first (lowest-energy) models. Below, we also will present correct predictions for other targets; however, for these proteins, the best models are not the lowest-energy structures.

Target T0223.2 (the C-terminal domain of a putative nitroreductase from Thermotoga maritima; PDB ID code 1VKW; a 92-residue \(\alpha+\beta\) protein) was classified by the CASP assessors as a fold recognition/homology target, which means that the target sequence is homologous to the sequence of a structure deposited in the PDB database. The structure is composed of two β-hairpin/α-helix motifs, two β-hairpins being packed together to form a four-stranded β-sheet with both α-helices packed against it from one side. Our best prediction for this target was model 3. The topology of the native structure was correctly reproduced; however, in our model, the N-terminal part of the structure formed an additional α-helix instead of an unstructured region. This model matches the experimental structure with 8.8-Å rmsd over all C\(^\alpha\) carbons (see Fig. 1E and Table 2), and a 67-residue fragment (73% of the sequence) fits a \(< 6.0\) Å rmsd cut-off. The main reason for the high rmsd values for this model is the presence of a nonnative N-terminal α-helix and a correct but very loose packing of the secondary-structure elements (a similar situation was observed for target T0281). Despite a high rmsd value, the prediction should be considered as very good, taking into account the correctly reproduced topology of nonsequential β-hairpin packing, which in the past caused many problems for the UNRES potential and search method.
Target T0230 is a 102-residue \( \alpha + \beta \) protein from *T. maritima*. The structure of this protein was solved independently by two NMR laboratories, so there are two PDB ID codes, 1WCI and 1UWD. This protein was classified by the CASP assessors as a fold recognition/analogy target. The core of the protein is formed by a \( \beta \)-hairpin/\( \alpha \)-helix/\( \beta \)-strand motif. Both the \( \beta \)-hairpin and the \( \beta \)-strand are packed together to form a three-stranded \( \beta \)-sheet. Short N- and C-terminal \( \alpha \)-helices flank the core part of the protein and are loosely packed to it. Our best prediction for this target is model 3 with rmsd 7.3 Å over all C\(^\alpha\)s (see Fig. 1F and Table 2), and a 90-residue fragment (88% of the sequence) fits a \(<6.0\)-Å rmsd cut-off. The topology of the native structure is correctly reproduced. The high overall rmsd value for model 3 is caused mainly by formation of \( \alpha \)-helical structures in the unstructured regions around the N and C termini and by the wrong orientation of the C-terminal \( \alpha \)-helix with respect to the rest of the structure. Nevertheless, as for T0230, our prediction for target T0230 marks another step forward in our physics-based approach to protein structure prediction, because it is the largest \( \alpha + \beta \) protein predicted qualitatively correctly so far.

The last successful prediction presented in this work is for target T0198. Target T0198 is a large 235-residue \( \alpha \)-helical protein (phosphate transport system regulator PhoU from *T. maritima*; PDB ID code 1SUM), classified by CASP assessors as a fold recognition/analogy target. The protein is composed of six \( \alpha \)-helices, which form a bundle and a small C-terminal \( \beta \)-hairpin. Our best prediction is model 5 with overall rmsd 9.8 Å over all C\(^\alpha\)s, but with correct topology of the bundle (see Fig. 1G and Table 2). Also, 139 residues (62% of the sequence; the first three \( \alpha \)-helices of the six-helix bundle) fit a \(<6.0\)-Å rmsd cut-off and 203 residues that constitute the whole protein, except that the C-terminal \( \beta \)-hairpin. Our best prediction is model 5 with overall rmsd 9.8 Å over all C\(^\alpha\)s, but with correct topology of the bundle (see Fig. 1G and Table 2). Also, 139 residues (62% of the sequence; the first three \( \alpha \)-helices of the six-helix bundle) fit a \(<6.0\)-Å rmsd cut-off and 203 residues that constitute the whole protein, except that the C-terminal \( \beta \)-hairpin. Our best prediction is model 5 with overall rmsd 9.8 Å over all C\(^\alpha\)s, but with correct topology of the bundle (see Fig. 1G and Table 2).
using the UNRES potential (M.C. and H.A.S., unpublished data). The structures obtained from this two-stage procedure were clustered and energy-ranked together with structures resulting from the regular UNRES/CSA search. One model resulting from the two-stage procedure and named model 5, was submitted as a prediction; this model had the correct topology with 203 residues fitting A <8-Å rmsd cut-off and is thus far the largest protein predicted correctly by our physics-based approach.

Conclusions

The results reported in this work demonstrate the significant progress made in our approach to predict 3D structures of proteins based on the thermodynamic hypothesis. During the last 4 years, we extended the capability of our prediction methodology significantly in terms of size and the type of protein fold that can be treated. During the CASP3 and CASP4 experiments, we were able to predict fragments or, rarely, whole structures of 60- to 100-residue all-α-helical proteins and, in CASP4, only started to predict the structures of α+β- and β-proteins, because of the introduction of cumulant-based correlation terms to the UNRES force field. In CASP5, we extended the applicability of our methodology to predict 70-residue fragments of α+β proteins; however, there were problems in achieving the correct packing topology of the β-strands. During CASP6, we were able to predict 70–80 residue fragments or whole structures of α+β proteins and up to 140-residue fragments (within 6-Å rmsd cut-off) of all-α-helical proteins with correct secondary-structure element packing. The main reason for this progress was improvement of the UNRES force field by deriving parameters based on quantum-mechanical calculations (25, 26) rather than on PDB-related statistical data and development of an optimization procedure based on a hierarchical design of the potential energy landscape (24, 26–28). However, we found that the global optimization method, which we used so far to search conformational space and is based on a genetic algorithm and potential energy local minimization, reached the limit of its usefulness because of computational cost related to larger proteins. We successfully reduced the computational cost of global conformational search by using secondary-structure prediction to generate the starting structures. However, this modification meant introducing some elements of knowledge-based information, which we want to avoid. To overcome problems related to a global search procedure, we recently developed Langevin dynamics (46) for the UNRES potential, which is currently being introduced as our main conformational search method. Molecular dynamics facilitates extending our prediction capabilities not only to predict protein 3D structure more accurately but also to predict protein-folding pathways.

This work was supported by National Institutes of Health Grant GM-14312, National Science Foundation Grant MCB00-03722, National Institutes of Health Fogarty International Center Grant TW7193, and Polish Ministry of Scientific Research and Information Technology Grants 3 T09A 032 26 and 6 T11 2003 C/06098. This research was conducted by using the resources of our 392-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology (Cornell University); the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center; our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdansk; the Informatics Center of the Metropolitan Academic Network in Gdansk; and the Interdisciplinary Center of Mathematical and Computer Modeling at the University of Warsaw.