Hierarchical Energy-Based Approach to Protein-Structure Prediction: Blind-Test Evaluation with CASP3 Targets

JOOYOUNG LEE,¹ ADAM LIWO,^{1,2} DANIEL R. RIPOLL,^{1,3} JAROSLAW PILLARDY,¹ JEFFREY A. SAUNDERS,¹ KENNETH D. GIBSON,¹ HAROLD A. SCHERAGA¹

¹*Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301*

²*Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland*

³*Cornell Theory Center, Ithaca, New York 14853-3801*

Received 16 August 1999; accepted 31 August 1999

ABSTRACT: A hierarchical approach based *exclusively* on finding the global minimum of an appropriate potential energy function, without the aid of secondary structure prediction, multiple-sequence alignment, or threading, is proposed. The procedure starts from an extensive search of the conformational space of a protein, using our recently developed united-residue off-lattice UNRES force field and the conformational space annealing (CSA) method. The structures obtained in the search are clustered into families and ranked according to their UNRES energy. Structures within a preassigned energy cutoff are gradually converted into an all-atom representation, followed by a limited conformational search at the all-atom level, using the electrostatically driven Monte Carlo (EDMC) method and the ECEPP/3 force field including hydration. The approach was tested (in the CASP3 experiment) in blind predictions on seven targets, five of which were globular proteins with sizes ranging from 89 to 140 amino acid residues. Comparison of the computed lowest-energy structures, with the experimental structures, made available *after* the predictions were submitted, shows that large fragments (~60 residues,

Correspondence to: H. A. Scheraga.

Contract grant sponsor: National Science Foundation.

Contract grant number: MCB95-13167.

Contract grant sponsor: National Institutes of Health.

Contract grant number: GM-14312.

Contract grant sponsor: NIH National Center for Research

Resources.

Contract grant number: P41RR-04293.

Contract grant sponsor: Polish State Committee for Scientific Research (KBN).

Contract grant number: DS 8245-4-0124-9.

Contract grant sponsor: National Foundation for Cancer Research.

representing 45–80% of the proteins) of those five globular proteins were predicted with the root mean square deviations (RMSDs) ranging from 4 to 7 Å for the C α atoms, with correct secondary structure and topology. These results constitute an important step toward the prediction of protein structure based solely on *global optimization of a potential energy function* for a given amino acid sequence. © 2000 John Wiley & Sons, Inc. Int J Quant Chem 77: 90–117, 2000

Key words: protein folding; global optimization; conformational search; potential energy function; structure prediction

Introduction

Knowledge of the native structure of a protein is necessary in order to learn about its physiological role and mechanism of action. It is a relatively easy task to compute the structures of fibrous proteins (e.g., myosin, keratin, collagen) which have periodic or quasi-periodic structures [1, 2]. A common feature of fibrous proteins is a small number of different structural motifs. Moreover, periodicity enables one to predict the structures of such systems by carrying out the conformational analysis of a single residue or a few consecutive residues; the geometry of the whole macromolecule can then be obtained by using the appropriate symmetry transformations.

By contrast, globular proteins, which constitute the majority of known proteins, are relatively small (~50 to ~5000 kDa; usually several hundred amino acid residues) and possess a variety of structural motifs; in recent classifications [3–6], there are several tens of different structural families. X-ray crystallography, NMR spectroscopy, and other experimental methods can provide only about several hundred new structures per year, while, in the same period, thousands of new actual or potential amino acid sequences are recorded. The design of a reliable method for protein-structure prediction is therefore of tremendous importance. Moreover, a primary goal of structure prediction is to gain an understanding of the interatomic interactions that lead to the folded native structure of a protein. There are essentially three classes of approaches to this problem: homology modeling methods, threading methods, and methods based on energetic criteria, usually referred to as *de novo* or *ab initio* methods.*

*It should, however, be noted that many approaches referred to as *ab initio* include a great portion of knowledge-based approaches, such as threading, homology modeling, or secondary-structure prediction. Throughout this article, *ab initio* methods

In homology modeling methods, the unknown structure is constructed on the basis of known structural motifs whose amino acid sequences are similar to the sequence studied, taking advantage of empirical relationships between the sequence and the 3D structure [8–13]. In threading methods, the native structure is sought in the database of known protein structures, using a designed sequence-structure fitness function (usually a pseudoenergy function) [9, 12, 14].

In *ab initio* methods, the native structure is sought as the global minimum of the energy surface of the protein, which is approximated by designed effective potentials of mean force. This class of methods is based on the physics of protein folding. According to the *thermodynamic hypothesis* formulated by Anfinsen [15], the native structure of a protein in a given environment is the structure that provides the lowest free energy of the system. A great advantage of *ab initio* methods is that the results are independent of any structural data bases.

To make use of the thermodynamic hypothesis, reliable energy functions and efficient global-optimization methods are required to reproduce the potential energy surface of the protein and search its conformational space. Each of these problems is extremely difficult. Moreover, they are strongly interrelated, because only with an efficient global-optimization method can it be decided that an energy function is a suitable potential for an amino acid sequence known experimentally as foldable and whether it leads to the experimentally known native structure. For these reasons, the consensus

will denote methods based only on energy, without the aid of knowledge-based information. The purpose of the *ab initio* approach is to provide an understanding of how pairwise and multibody (cooperative) interactions lead to the folded structure. Those interested in obtaining the structure itself have two options: (i) determine it experimentally with X-ray or NMR methods, or (ii) when this is not feasible, continue to develop heuristic approaches based on secondary-structure predictions, homology modeling, threading, use of fragments from the protein data bank (PDB) [7], etc.

of the protein-folding community has been that protein-structure prediction based on the *thermodynamic hypothesis* is hardly feasible now and perhaps in the foreseeable future [16]. This view became more apparent after the poor performances of the *thermodynamic* approach in CASP2 blind structure predictions of proteins [17]. Therefore, the latest methods for protein-structure prediction, such as the one developed by Skolnick et al. [18], make explicit use of the information provided by sequence homology, secondary-structure prediction, and/or threading.

In this article, we present a novel but straightforward *thermodynamic* approach toward the successful structure prediction of proteins. In recent years, several global-optimization methods have been designed in our laboratory [2, 19–30]. With current computer resources, some of these methods can treat polypeptide chains with up to about 20 residues (the *N*-terminal part of melittin) in an all-atom representation.

Despite this development of global optimization methods, and the rapid progress in computer technology, to treat real-size protein chains by the thermodynamic approach, it is necessary to use a simpler model than an all-atom representation of the polypeptide chain. A commonly applied solution is to represent each amino acid residue by a single or a few interaction sites [14, 31–42]; these are the so-called *united-residue* models. The corresponding force fields are mean-field force fields, whose parameters are determined by applying the Boltzmann principle to distribution and correlation functions calculated from protein-crystal data and/or by averaging all-atom potentials. After the low-energy conformations have been found in a virtual-bond united-residue representation, they can be converted to all-atom chains [9, 35, 43–46] and a limited conformational search can be carried out in an all-atom representation; this is the *hierarchical approach*. In the last few years, we have begun to develop such a method for protein-structure prediction [35, 36]. We have derived and parameterized a united-residue force field, hereafter referred to as UNRES, which represents the simplified potential energy surface of polypeptide chains reasonably well [37, 38, 47].

Recently, we developed a very efficient method of conformational search called the conformational space annealing (CSA) method [28, 29, 48]. One of the greatest advantages of the CSA method is that it can find many families of low-energy conformations that have distinct backbone structures.

This makes it possible to search thoroughly the conformational space of a protein. Only after such a thorough search has been accomplished can the global minimum-energy conformation (GMEC) be located at the desired confidence level.[†] Adaptation of CSA to the UNRES force field enabled us to explore the conformational space of chains with several tens of residues. Tests on helical proteins (the 48-residue fragment of the B domain of staphylococcal protein A and the 75-residue apo-calbindin D9K) showed that the near-native structure belongs to the lowest-energy family of structures or to the family with the second-lowest energies [49].

In the present article, we describe the performance of our approach in blind predictions that we made during the CASP3 test.[‡] We show that it has become possible to treat proteins up to about 140 residues long and that the CSA method with the state-of-the-art UNRES force field, followed by the remaining procedures of the hierarchical approach, can now predict 45–80% of the structure correctly.

Methods

THE HIERARCHICAL ALGORITHM

As pointed out in the Introduction, a global conformational search of real-size proteins is still not feasible using all-atom models (e.g., a 1 μ s molecular dynamics run on a 35-residue protein with solvent included requires months of CPU time on a supercomputer [50]). Therefore, in our approach, we start from a simplified united-residue representation of polypeptide chains, abbreviated as UNRES [36–38, 47] (see the section The UNRES Model of

[†]While there are deterministic methods for special cases, there is no deterministic method applicable to finding the global minimum of a multivariate function in a general case. The only generally applicable deterministic method would be to carry out a complete enumeration of the function values, which is infeasible in the case of proteins, even if one restricts the search to a coarse grid (the effort required grows exponentially with the dimensionality of the problem). In view of this, the methods that are currently applicable to proteins are probabilistic, and for a given method, the probability that the global minimum is found will depend on the length of the computations. If one is interested only in obtaining some representative conformations from the region of the global minimum, the length of the computations can be reduced; however, to increase the confidence that the GMEC has been located, many long runs are required [49].

[‡]Third Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction; <http://predictioncenter.llnl.gov/casp3/Casp3.html>.

Polypeptide Chains). UNRES is a reasonably accurate model of the energetics of protein interactions [49]. A global conformational search (for up to 140 residues) is achieved in this simplified representation, using the CSA method [28, 29, 48, 49] [see the section Conformational Search by CSA (Stage I)], which can identify many families of low-energy conformations with distinct backbone structures. This is the most crucial and computationally demanding stage in our approach [51]. At the end of stage I of our algorithm, a cluster analysis of the UNRES structures resulting from this CSA search is carried out, and the lowest-energy members of the conformational families within a preassigned energy cutoff (usually 5–10 kcal/mol) are selected. In stage II, the C^α -traces of the selected conformations are converted to all-atom backbones by optimal alignment of peptide-group dipoles [35]. In stage III, the backbone conformations are energy-optimized, subject to C^α — C^α -distance constraints from the parent UNRES structures (obtained in stage I). In stage IV, all-atom side chains are added, subject to the condition of nonoverlap; finally, in stage V, a limited exploration is carried out at the all-atom level, using the electrostatically driven Monte Carlo (EDMC) method [24, 52, 53]. All-atom calculations (stages III–V) are carried out with the ECEPP/3 potential function [54], including solvation with the SRFOPT surface-hydration model [55] in stages IV and V. The distance constraints from the UNRES structures are still used in stage IV, while in stage V, they are fully imposed only at the beginning and then gradually removed; the final all-atom structures correspond to energy minimizations without distance constraints.

The algorithm is illustrated by an example in Figure 1. Each stage is described in some detail in the next sections.

THE UNRES MODEL OF POLYPEPTIDE CHAINS

In the UNRES model [35, 37], 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) located in the middle between the consecutive α -carbons. Only the united peptide groups and united side chains serve as interaction sites, the α -carbons assisting in the definition of the geometry (Fig. 2). All the virtual bond lengths (i.e., C^α — C^α and C^α —SC) are fixed; the C^α — C^α distance is taken as 3.8 Å, which corresponds to *trans*-peptide groups. The angles α_{SC} and β_{SC} , which define side-chain orienta-

tion with respect to the backbone and virtual-bond angles θ , are variable. The energy of the virtual-bond chain is expressed by Eq. (1):

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

where $U_{SC_i SC_j}$, $U_{SC_i p_j}$, and $U_{p_i p_j}$ denote the pairwise energies of the interactions between side chains, between side chains and peptide groups, and between peptide groups, respectively; $U_{tor}(\gamma_i)$ denotes the energy of variation of the virtual-bond dihedral angle γ_i ; $U_b(\theta_i)$ denotes the “bending” energy of the virtual-bond angle θ_i ; $U_{rot}(\alpha_{SC_i}, \beta_{SC_i})$ is the lo-

FIGURE 1. Illustration of the hierarchical algorithm of protein-structure prediction for CASP3 target T0073: (a) the C^α trace; (b) the C^α trace with indicated dipole paths; (c) the all-atom backbone structure based on peptide-group dipole orientation followed by perturbation by the EDMC method to find a relaxed structure; (d) side chains attached with subsequent EDMC perturbation to give the all-atom structure.

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

cal energy of side chain i ; U_{corr} includes cooperative terms (e.g., the four-body interactions considered by Koliński and coworkers [56]); and the w 's denote relative weights of the respective energy terms.

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

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

where ε_{ij} is the pair-specific well depth; $\varepsilon > 0$ corresponds to hydrophobic–hydrophobic-type, and $\varepsilon < 0$, to hydrophobic–hydrophilic and hydrophilic–hydrophilic-type interactions. The quantity x_{ij} is the reciprocal of the reduced distance between side

chains. This can depend on the distance alone for a radial-only potential or on both distance and orientation, and the same applies to ε_{ij} . Currently, we are using a radial potential with

$$\varepsilon_{ij} = \varepsilon_{ij}^0 \quad (3)$$

$$x_{ij} = \frac{\sigma_{ij}^0}{r_{ij}}, \quad (4)$$

where r_{ij} is the distance between side chains i and j and ε_{ij}^0 and σ_{ij}^0 are pair-specific constants. Other functional forms for ε and x , as well as the numerical values of the parameters of the potential, can be found in the original article [37].

The interaction potential ($U_{p_i p_j}$) accounts mainly for the electrostatic interactions between peptide groups, that is, for their tendency to form backbone hydrogen bonds. Its functional form was derived by averaging the simplified electrostatic-interaction energy of the peptide groups over the angles λ of their rotation about the corresponding $C^\alpha-C^\alpha$ virtual-bond axes, assuming that each peptide group is modeled by a point dipole located in the middle of the virtual bond, as proposed by Piela and Scheraga [57]. The potential is expressed by Eq. (5); the details of the derivation and parameterization can be found in the original articles [35, 37]:

$$U_{p_i p_j} = \frac{A_{p_i p_j}}{r_{ij}^3} (\cos \alpha_{ij} - 3 \cos \beta_{ij} \cos \gamma_{ij}) - \frac{B_{p_i p_j}}{r_{ij}^6} [4 + (\cos \alpha_{ij} - 3 \cos \beta_{ij} \cos \gamma_{ij})^2 - 3(\cos^2 \beta_{ij} + \cos^2 \gamma_{ij})] + \varepsilon_{p_i p_j} \left[\left(\frac{r_{p_i p_j}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{p_i p_j}^0}{r_{ij}} \right)^6 \right], \quad (5)$$

with

$$\cos \alpha_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j$$

$$\cos \beta_{ij} = \mathbf{v}_i \cdot \mathbf{e}_{r_{ij}}$$

$$\cos \gamma_{ij} = \mathbf{v}_j \cdot \mathbf{e}_{r_{ij}},$$

where $A_{p_i p_j}$, $B_{p_i p_j}$, and $\varepsilon_{p_i p_j}$ are constants characteristic of the kind of interacting peptide groups; r_{ij} , the distance between the peptide-group centers; $\mathbf{v}_i$, the unit vector pointing from C_i^α to C_{i+1}^α ; and $\mathbf{e}_{r_{ij}}$, the unit vector pointing from p_i to p_j . Two types of peptide groups were distinguished: ordinary and proline; the second one can act as a hydrogen-bond acceptor only and also pertains to all *N*-methylated amino acid residues (e.g., sarcosine). This gives a total of three sets of constants in Eq. (5). The angular part of Eq. (5) favors parallel and near-parallel (or

antiparallel and near-antiparallel) orientation of the virtual $C^\alpha-C^\alpha$ bonds, as encountered in hydrogen-bonded backbone peptide groups.

The torsional energy, U_{tor} , is expressed in terms of a Fourier series in the virtual-bond dihedral angles γ , as given by Eq. (6):

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

This energy reflects the local propensities of the polypeptide chain, that is, to form right-handed rather than left-handed helices and left-handed rather than right-handed β -strands. It was natural to consider three torsional types of amino acid residues: glycine (because of the absence of the β -carbon), proline (because of the restriction caused by the presence of the pyrrolidine ring), and alanine (which is a model for the remaining amino acid residues). A detailed analysis of local propensities of the structures in the PDB confirmed this division [38].

The bending energy, U_b , and local side-chain energy, U_{rot} , are expressed as negative logarithms of sums of Gaussians. Because these expressions are lengthy, the reader is referred to the original article [38].

The multibody (cooperative) term U_{corr} arises from the fact that details of the all-atom chain are lost when it is converted into the simplified chain. Mathematically, this term can be considered to average the energy over some "less important" degrees of freedom [47]. In the current version of UNRES, we include the cooperativity of two pairs of adjacent peptide groups, such as are found in α -helices and β -sheets. In this case, U_{corr} is expressed by Eq. (7) [47]:

$$U_{corr}(i, i-1; k, k \pm 1) = -\frac{z^2}{4} (9\eta_{ik}\eta_{i-1, k \pm 1} + 4\bar{\eta}_{ik}\bar{\eta}_{i-1, k \pm 1}), \quad (7)$$

where η_{ik} is the energy of the interaction of two peptide-group dipoles lying on a line and pointing in the same direction (or two hydrogen-bonded peptide groups), and $\bar{\eta}_{ik}$ is the interaction of two *antiparallel* peptide-group dipoles (or two antiparallely stacked peptide groups); z^2 is a proportionality constant. These four-body terms have much in common with the terms introduced by Skolnick et al. on a heuristic basis [56, 58]. For the derivation and functional forms of the multibody terms, the reader is referred to the original article [47].

As in other works on structure-derived protein potentials [14, 56, 59, 60], the side-chain potential

(U_{SCSC}) and the components of the local-interaction potential (U_b and U_{rot}) have been parameterized using distribution and correlation functions determined from a set of 195 high-resolution nonhomologous structures from the PDB. The procedure was described by Liwo et al. [37]. The peptide-group interaction potential U_{pij} and the virtual-torsional potential U_{tor} were parameterized [35, 36] by averaging the all-atom ECEPP/2 [61] potential.

The relative weights of the components of the force fields were chosen so as to maximize the ratio of the energy gap between the native structure and the lowest-energy nonnative structure to the distribution of nonnative conformations. We developed a method [47] based on the approach of Wolynes and coworkers [62], Shakhnovich and coworkers [63], and Hao and Scheraga [64, 65], which is directed at achieving as negative a value as possible of the so-called Z-score function [Eq. (8)]:

$$Z = \frac{E_0 - (1/N) \sum_{i=1}^N E_i}{[(1/N) \sum_{i=1}^N E_i^2 - (1/N)^2 (\sum_{i=1}^N E_i)^2]^{1/2}}, \quad (8)$$

where N is the number of nonnative conformations; E_0 , the energy of the native conformation; and E_i , the energy of the i th nonnative conformation. The more negative the Z-score values, the more the native structure is distinguished from nonnative ones. To optimize the weights, we chose the phosphocarrier protein from *Streptococcus faecalis* (PDB code: 1PTF; 87 amino acid residues; 1.6 Å resolution). The set of nonnative conformations was constructed by superposing the 1PTF sequence on polypeptide-chain fragments of appropriate length randomly chosen from the PDB and minimizing the energies of such chains; this can be considered a version of the *inverse folding approach*. The procedure is described in detail in the original article [38].

CONFORMATIONAL SEARCH BY CSA (STAGE I)

The energy surface of a protein contains a huge number of local minima. The larger the protein, the more likely it is that there exist many low-energy local minima that correspond to very different structures (such as, e.g., the mirror image of a native structure with a helix-bundle fold). In reality, only the native structure exists, but because of the approximations and inaccuracies associated with existing potential functions, the energy relation between the native structure and other low-energy structures is not clear. It is not enough to consider only the lowest-energy structure as a possible candidate for the native structure; rather, one should take

account of many distinct low-energy conformations. Therefore, it is important that as large a portion of conformational space as possible be searched. The current version of the CSA method is intended to cover the whole conformational space in its early stages and then to narrow the search to smaller regions. Details of the CSA algorithm can be found in our earlier work [28, 29, 48]. Here, we provide only a brief description, and the relevant changes in the algorithm for its implementation with the UNRES force field.

In the CSA method, the exploration of conformational space in the early stages, followed by increasingly narrower searches of regions with low energy, is accomplished by reducing the distance cutoff, D_{cut} , which defines the similarity of two conformations (hence, the name conformational space annealing). As in our previous work [28], the distance between conformations i and j , D_{ij} is defined by Eq. (9), in terms of the differences between all variable angles that define the geometry of the united-residue chain (Fig. 2).

$$D_{ij} = \sum_{k=1}^n \left[|\theta_i - \theta_j| + \min[\text{mod}\{(\gamma_i - \gamma_j), 2\pi\}, \{2\pi - \text{mod}\{(\gamma_i - \gamma_j), 2\pi\}\}] + |\alpha_i - \alpha_j| + \min[\text{mod}\{(\beta_i - \beta_j), 2\pi\}, \{2\pi - \text{mod}\{(\beta_i - \beta_j), 2\pi\}\}] \right], \quad (9)$$

where n is the number of residues.

As in genetic algorithms [66, 67], CSA starts with a preassigned number (usually 50) of randomly generated and subsequently energy-minimized conformations [68]. This pool of conformations is called the *first bank*. The first bank is copied to provide the initial conformations in the bank (see Fig. 3 for a flow chart of the CSA method). At the beginning, the bank is supposed to be a sparse representation of the entire conformational space. Subsequently, a number of dissimilar conformations (usually 20) are selected from the bank, excluding those that have already been used; they are called *seeds*. Each seed conformation is modified by changing from one to one-third of the total number of variables pertaining to a contiguous portion of the chain. The new variables are selected either from one of the remaining bank conformations or from the first bank, rather than picked at random. Each conformation is energy-minimized to give a trial conformation. Twenty to thirty trial conformations are generated for each seed (a total of 400–600 conformations). This is the most time-consuming stage of the computation (over 99%), but it is highly suitable for massively parallel computing, because the local minimizations are independent of each other (therefore, with the settings proposed above, it is possible to make efficient use of more than 100 processors [48]). For each trial conformation, α , the closest conformation A from the bank (in terms of the distance $D_{\alpha A}$) is determined. If $D_{\alpha A} < D_{\text{cut}}$ (D_{cut} being the current cutoff criterion), α is considered similar to A . In this case, α replaces A in the

FIGURE 3. Flow chart for the CSA algorithm to locate the GMEC.

bank if it is lower in energy; otherwise, it is discarded. If α is not similar to A (i.e., $D_{\alpha A} > D_{\text{cut}}$), but its energy is lower than that of B , the highest-energy conformation in the bank, α replaces B ; otherwise, it is discarded. Thus, the number of conformations in the bank remains unaltered. Narrowing the search regions is accomplished by setting D_{cut} to a large value initially and gradually diminishing it as the search progresses. Special attention is paid to selecting seeds that are far from each other [48]. One round of the procedure is completed when there is no seed left to select (i.e., all conformations from the bank have already been used). This round is repeated a predetermined number of times. If necessary, more random conformations are added to the bank and the whole procedure is repeated.

The CSA method has been successful in obtaining the GMEC of polypeptide chains. With the ECEPP/3 all-atom force field, the GMEC of peptides containing up to 20 amino acid residues, with variable dihedral angles up to 113 in number, have been located [28, 29, 48]. The average wall-clock time to find the GMEC of a polypeptide with 20 amino acid residues, based on 24 independent runs, was about 4.5 h with 32 processors of an IBM SP2 supercomputer. The corresponding wall-clock time for a pentapeptide was 36 s with 16 processors [48]. With the UNRES force field, the CSA method successfully located the nativelike conformations of two helical proteins [the *N*-terminal domain of staphylococcal protein A (46 amino acid residues) and apo-calbindin D9K (75 amino acid residues)] among the low-energy ones [49]. Alternative structures (which were mirror images of the native folds) were also found. The average wall-clock time to find the GMEC of the 46-residue fragment of protein A (based on seven independent runs) was about 14 h with 32 processors of an IBM SP2 supercomputer.

Generation of Random Conformations

To generate random polypeptide conformations at the united-residue level, we used the procedure of Hao et al. [69] that had already been built into UNRES [37, 38]. In brief, the algorithm is as follows:

1. Generate the first virtual-bond valence angle, θ_1 , and the angles defining the location of the first side chain, α_1 and β_1 , according to the distribution functions computed from the corre-

sponding part of the united-residue potential, U_b and U_{rot} (the distributions are computed as Boltzmann distributions with temperature $T = 298$ K and are residue-type specific, because the corresponding energy parameters are residue-type specific). Set the residue counter at $i = 2$. Compute the Cartesian coordinates of the α -carbon atoms C_1^α , C_2^α , and C_3^α and, subsequently, the peptide groups p_1 and p_2 , and the side chain SC_1 .

2. Generate the virtual-bond dihedral angle γ_{i-1} from a uniform distribution defined over the interval $[-\pi, \pi]$. Generate the virtual-bond valence angle θ_i and the angles of the i th side chain, α_i and β_i , based on the distributions computed from the corresponding energy terms in the force field assuming the Boltzmann law (as described in step 1). Compute the Cartesian coordinates of C_{i+1}^α , and, subsequently, p_{i+1} and SC_i .
3. Compute the distances between sites SC_i and p_{i+1} and the sites with already defined geometry: $SC_1 \dots SC_{i-1}$ and $p_1 \dots p_i$. If any of the distances is less than the pre-assigned (site-pair specific) overlap distance, repeat step 2. If the maximum number of 100 attempts is exceeded, generation is restarted from residue $i - 2$.
4. Chain generation is completed when the coordinates of p_{n+1} and SC_n , n being the number of full amino acid residues, have been generated subject to the nonoverlap condition. The procedure fails if chain generation is still incomplete and the predefined number of generation steps has been reached. If this happens, the procedure is repeated from the beginning.

Generation of Trial Conformations

Here, we describe the procedure to generate the trial conformations for a given seed conformation:

1. Generate a preassigned number of conformations (usually 15) by replacing one set of either (θ, γ) or (α, β) of the seed with the corresponding set from a randomly selected conformation from the first bank (which contains only those conformations that were obtained by local energy minimization of randomly generated conformations; see the subsection Generation of Random Conformations).
2. Generate a smaller number of conformations (usually three) by replacing a set of variable

angles pertaining to one residue, $(\theta, \gamma, \alpha, \beta)$ of the seed with the corresponding set from a randomly selected conformation from the bank.

3. Generate a preassigned number of conformations (usually 12) as above but replacing a larger segment of contiguous residues. The size of a segment is chosen at random from between two residues and one-third of the total number of residues [29].

CONSTRUCTING THE ALL-ATOM BACKBONE BY ELECTROSTATIC ALIGNMENT OF PEPTIDE GROUPS (STAGE II)

This stage is carried out using the *dipole path method* [35]. This method is based on the observation that the peptide groups in protein structures tend to form an extensive hydrogen-bond network, which can also be expressed in terms of their tendency to adopt optimal electrostatic interactions [57]. By taking advantage of the dipole model of the peptide group [35, 57], the problem of positioning the peptide groups, given the C^α trace, can be formulated as a problem of optimal alignment of the peptide-group dipoles. In regular helical or sheet structures, lines of aligned peptide-group dipoles or *dipole paths*, which pertain to all peptide groups in regular structures, can be distinguished (e.g., each α -helix contains three paths linking every third peptide group). The algorithm is as follows:

1. Find the *dipole paths*. A dipole path is defined as a sequence of n noncontiguous peptide groups $\{p_{i_1}, p_{i_2}, \dots, p_{i_n} \mid |i_k - i_{k-1}| > 1, k = 2, 3, \dots, n\}$, i_k being the position in the amino-acid sequence, with the following properties: (i) the energy of average electrostatic interactions, $U_{p_{i_k} p_{i_{k+1}}}$, $k = 1, 2, \dots, n - 1$ [Eq. (5)], is below a chosen cutoff limit (-0.5 kcal/mol in this study); (ii) the angle between peptide groups p_{i_k} , $p_{i_{k+1}}$, and $p_{i_{k+2}}$, $k = 1, 2, \dots, n - 2$, is greater than 90° (so that the peptide groups forming a dipole path lie on a line with a low curvature). The peptide groups that belong to a dipole path can form a hydrogen-bonded chain.
2. Align peptide-group dipoles along each path. Because alignment can be carried out in two directions, the correct direction is chosen by taking into account the electrostatic interactions of the dipoles of the neighboring paths

and the local interactions within the amino acid residues involved in the dipole paths.

3. Align "isolated" peptide groups in the electrostatic field of the determined dipole paths.
4. Based on the positions of the peptide-group dipoles, calculate the coordinates of all atoms in the peptide groups.

REFINEMENT OF BACKBONE CONFORMATION (STAGE III)

The all-atom-backbone structures from stage II are initially subjected to constrained energy minimization, using C^α -distance constraints from the parent united-residue structures [$n(n - 1)/2$ distance constraints for an n -residue chain]. Because the coordinates of the side-chain atoms are still undetermined, the polypeptide chain is treated as a "polyalanine" molecule, in which all amino acid residues except proline and glycine are replaced with alanine. Energy minimization is carried out initially by using a soft-sphere potential of atom-atom interaction (i.e., a truncated Lennard-Jones potential [70]), in order to obtain a nonoverlapping structure (the backbones obtained from the dipole-path method, which takes account of electrostatic interactions only, usually contain many overlaps). Subsequently, minimization of the energy of the polyalanine model is carried out by using the ECEPP/3 force field [54]. Then, the structures are refined further by using the EDMC method [24], still using C^α -distance constraints from stage I. This stage leads to structures with all-atom backbones in low-energy conformations and C^α trace geometries consistent with that of the parent united-residue structures.

GENERATION OF SIDE-CHAIN CONFORMATION (STAGE IV)

Once an all-atom backbone has been obtained, all-atom side chains are introduced, using the following algorithm [36]:

1. All side chains are initially assigned an extended conformation.
2. Consecutive dihedral angles χ are scanned over the whole interval $[-180^\circ, 180^\circ]$ (in increments of 30°), the total energy being calculated at each step, and the optimum position is kept for each χ . This step is repeated until the energy does not decrease by more than 1 kcal/mol, after all dihedral angles χ

have been adjusted. The ECEPP/3 force field augmented by the SRFOPT surface-hydration model is used in the energy minimization.

3. The low-energy structure obtained in step 2 is minimized subject to the C^α -distance constraints from the parent united-residue structure.
4. The cycle consisting of steps 2 and 3 is iterated until the energy does not decrease by more than 2 kcal/mol during one cycle.

REFINING THE ALL-ATOM STRUCTURES (STAGE V)

After conversion to an all-atom model, certain parts of the structure such as loop regions adopt conformations that are not commonly observed in real proteins (e.g., long left-handed α -helices, or conformations with many of the residues having positive ϕ values). In such cases, additional local searches, restricted to these specific regions, were carried out using the method developed by Palmer and Scheraga [71] which is based on the local deformation algorithm of Gō and Scheraga [72]. In this algorithm, deformation of a region of a polypeptide chain composed of n residues, whose dihedral angles ω are kept constant, involves the $2n$ dihedral angles ϕ and ψ along the backbone. $2n - 6$ of these dihedral angles may be specified in an arbitrary manner and the chain-closure equations can then be solved for the remaining six degrees of freedom. If a solution is found, the conformation of the n residues can be considered as a possible deformation. The method of Palmer and Scheraga used segments of five residues in which the dihedral angles ϕ and ψ of two of the residues are sampled systematically (grid search). For a complete description of this procedure, the reader is referred to the original publication.

Following the loop search, constrained energy minimization of the whole chain is carried out, with C^α -distance constraints derived from the parent UNRES structure. At this point, all C^α pair distances are used as constraints. Usually, two minimization runs are carried out, the first one with a larger weight of the distance-constraint term (usually 1.0), and the second one with a smaller weight (usually 0.1). Then, a series of EDMC runs is carried out with gradually decreasing weights for the distance constraints. First, the weights of all distances are set equal (usually to 0.1). In the next EDMC run, the weights corresponding to C^α distances greater than 10 Å are set at 0. Next, the weights of the dis-

tances corresponding to residues of the loop regions are set at 0. The last run is carried out with no distance constraints.

The conformations obtained in the last (unconstrained) run can be clustered into families, whose lowest-energy representatives, within a preassigned energy cutoff, can be considered as candidates for the predicted structure. However, for larger proteins, there is usually no substantial difference among the all-atom conformations derived from a given UNRES structure; therefore, in this study, we decided to retain only the lowest-energy all-atom conformation corresponding to each UNRES structure.

THE ELECTROSTATICALLY DRIVEN MONTE CARLO METHOD

The electrostatically driven Monte Carlo (EDMC) method [24, 52, 53] is an iterative method used for searching the conformational hypersurface of polypeptides and proteins. A search for the conformation with the lowest energy of a molecule, that is, the global energy minimum, proceeds as a "quasi-random walk" along a conformational pathway. In principle, this pathway is determined by a sequence of local energy minima of the potential encountered over an unbounded number of iterations. Practical considerations, however, require the number of iterations to be finite (the actual value is specified by the user at the beginning of the simulation). The underlying assumption behind the EDMC method is that the conformation of the polypeptide chain is primarily the result of two competitive phenomena: (a) the electrostatic interactions and (b) thermal fluctuations. Electrostatic interactions force the polypeptide chain to adopt conformations with lower electrostatic and total energies; thermal fluctuations, on the other hand, introduce disorder within the structure and can force the molecule to adopt higher-energy conformations. Because of the thermal fluctuations, the protein may be able to escape from stable local minima of relatively high energy.

Thermal effects are implemented in the algorithm as random changes in the molecular conformation, that is, random changes of a small set of variables selected randomly. As in the self-consistent electrostatic field (SCEF) method [57], the reordering effect of the electrostatic interactions is viewed as a tendency of all permanent dipole moments of the polypeptide to attain their best possible alignment in the local electric field produced by the rest of the molecule.

The energy of a newly generated conformation is minimized to relieve all possible atomic overlaps and is subsequently used as one of the elements involved in a set of selection criteria to determine the accepted conformations. The latter feature is one of the key elements of the Monte Carlo with minimization (MCM) method [22, 23]. The first accepted conformation on the EDMC conformational pathway usually corresponds to an unfolded state of the molecule, that is, a conformation in which the initial variables are assigned randomly and subsequently energy-minimized. The most recent EDMC implementation [53] includes a series of new features that accelerates the search and further optimizes the generation of new conformations.

The EDMC method has been applied successfully to surmount the multiple-minima problem associated with polypeptide sequences consisting of up to 20 amino acid residues [24, 52, 53, 73, 74]. It has also been applied to explore the conformational space of larger molecules; in these cases, the searches were restricted to explore defined regions of the conformational space close to the native conformations [75–77]. The method has also been applied to study the variation of the conformational properties of a series of oligo- and polypeptides with pH [78–80].

ECEPP/SRFOPT

The ECEPP/3 [54, 61, 81, 82] algorithm was used to generate the conformations of the all-atom models of the proteins. In these models, the bond lengths and bond angles of the polypeptide chain are fixed at their ECEPP/3 values, and the complete set of backbone and side-chain dihedral angles are considered as the independent variables.

Energy refinement was carried out by minimizing the sum of the ECEPP/3 potential plus a surface free-energy of solvation term (SRFOPT) [55]. For use with the solvation models, the solvent-exposed surface area was calculated with the MSEED algorithm [83]. The molecules were assumed to be neutral with all ionizable groups uncharged. ECEPP/3 partial charges for *neutral* residues were used at this stage. The secant unconstrained minimization solver (SUMSL) of Gay [68] was used for all minimizations.

CLUSTERING

Clustering was carried out with the minimal-spanning tree algorithm [84]. The distance between

two conformations (required in the clustering algorithm) was defined as the C^α root-mean-square deviation (RMSD) for united-residue chains and all nonhydrogen backbone and C^β atom RMSDs for all-atom chains.

RMSDs

RMSDs between pairs of structures were calculated from the positions of the selected atoms at optimal superposition, which was achieved with an algorithm that uses the singular-value decomposition of a correlation matrix formed from the atomic coordinates [85]. To compare the models with the experimental structures, the RMSD of C^α atoms was used.

FITTING X-RAY AND NMR STRUCTURES TO RIGID ECEPP GEOMETRY

Conformations generated by ECEPP/3 using the dihedral angles calculated directly from X-ray or NMR structures usually give large RMSDs from the experimental structure. To improve the fit, we carried out a series of minimizations of a target function composed of a pseudoenergy term (that prevents overlaps [70]) and an RMSD term. Once a structure with a reasonable RMSD from the experimental structure has been obtained, the pseudoenergy term is replaced with the ECEPP/3+SRFOPT energy and constrained minimizations are carried out again. Finally, only the ECEPP/3+SRFOPT energy (with no RMSD constraints) is minimized; the resulting conformation is considered as the regularized structure, with whose energy the energy of the calculated structure is compared. The RMSDs are calculated with respect to the experimental (non-regularized) structures. The experimental structures were revealed to us *after* we submitted the predicted coordinates.

Results and Discussion

The hierarchical energy-based approach described in the previous sections was evaluated in blind-prediction tests during the CASP3 experiment [86]. We chose seven targets of moderate chain length (from 37 to 140 residues), most of which were classified as not homologous to known structures. Another consideration was that these targets were thought to be mainly helical, as revealed by secondary structure predictions provided for each

TABLE I
Values of CSA parameters^a and wall-clock time of CSA runs for all seven CASP3 targets studied.

Model	No. residues ^b	No. variables	No. energy minimizations	No. energy evaluations $\times 10^{-6}$	Time (h) ^c	Range of energy in the final bank (kcal/mol)
T0056	114	443	52,250	104	67	$-501.3 < E < -489.8$
T0061	89	349	81,650	108	41	$-333.5 < E < -327.2$
T0061 ^d	89	349	65,450	92	34	$-336.5 < E < -330.3$
T0073	48	183	149,550	117	12	$-165.6 < E < -160.5$
T0074	104	413	66,650	123	65	$-490.2 < E < -480.4$
T0076	140	533	51,650	121	120	$-613.9 < E < -599.8$
T0079	129	505	52,850	114	95	$-535.5 < E < -525.2$
T0084	37	141	33,400	19	1	$-119.7 < E < -115.3$

^a CSA parameters common to all runs are as follows: The value of D_{cut} was reduced in the annealing schedule from $(1/2)D_{\text{ave}}$ to $(1/5)D_{\text{ave}}$ in 20,000 minimizations. D_{ave} is the average pair distance between conformations in the first bank. The number of seeds and the number of trial conformations per seed were 20 and 30, respectively.

^b This corresponds to the *full* sequence provided by CASP3.

^c Wall-clock time using 64 processors of an IBM SP2 supercomputer.

^d This run included a pseudoenergy term to force a disulfide bond between the only two cystines.

target by the CASP3 organizers. The reason for this was that, at the present stage of development, the UNRES force field can treat only α -helical proteins with sufficient accuracy. Although up to five models were allowed to be submitted for each target, we submitted fewer models in most cases. The targets and the number of predictions for each of them were T0056 (1), T0061 (4), T0073 (2), T0074 (2), T0076 (4), T0079 (4), and T0084 (5). Although targets T0073 and T0076 were classified as homologous to known structures (the names of the homologous proteins being kept secret by the CASP3 organizers), we included them as a further test of our approach. Predictions for the *entire* sequences were submitted. In our approach, the difficulty of the prediction depends mostly on the length of the amino acid sequence (i.e., the size of the conformational space) and not on the availability of sequence- or structural-homologous proteins in the databases. The parameters and timings of the CSA runs (stage I) are summarized in Table I. The models were ranked according to UNRES energies, except for target T0084 (see the target T0084 section for details).

Comparisons of all of our 22 predicted structures with the experimental ones are presented in Figure 4(a–g), where the lowest C^α -coordinate RMSDs of *contiguous* fragments of given length are plotted versus fragment length. Additionally, the best-matching models are presented in Figures 5–13, and their RMSDs from the corresponding fragments of

the experimental structures are presented in Table II. These RMSDs, as well as the RMSDs discussed in the text, refer to the all-atom models. The RMSDs corresponding to the *whole* chains of the parent UNRES structures (from stage I) are summarized in Table III for the sake of completeness only. The predictions for each of the seven targets are discussed in separate sections.

As shown in Table II and in Figures 4(a–g) and 5–13, large portions of the structures were predicted correctly, and in one case (target T0061, model 4), the experimental topology was also achieved (Fig. 7). The misalignment of parts of the models could originate from (a) the fact that some of the proteins exist as dimers (target T0061 and T0073) or tetramers (target T0084), whereas all our calculations were carried out for monomers, or (b) incompleteness of the search procedure and/or imperfectness of the UNRES force field.

ASSESSMENT OF THE COMPLETENESS OF THE CONFORMATIONAL SEARCH

At the present stage of development of our energy-based approach, a reliable search procedure that can retrieve *all* families of low-energy conformations is essential. Only with a thorough search procedure can the force field be reliably tested and further improved. If the parameters of a force field are optimized (to locate the experimental structure as the lowest in energy) using an incomplete set

(a)

(b)

FIGURE 4. Plots of C^α RMSD between calculated (all-atom models from stage IV) and (nonregularized) experimental best-matching fragments of given length (ordinates) versus fragment length (abscissae), for all targets and all models. The references to the experimental structures are as follows: T0056: [86], T0061: [89], T0073: [86], T0074: [92], T0076: [88], T0079: [93], T0084: [86].

of conformations from an inefficient search procedure, the GMEC can still be missing in the set of decoys [87]. In previous work [49], we carried out exhaustive conformational searches, using the CSA

method with the UNRES force field, on two proteins whose experimental structures were known, namely, the *N*-terminal domain of staphylococcal protein A and apo-calbindin D9K. In both cases,

(c)

(d)

FIGURE 4. (Continued)

the regularized experimental structures[§] belonged

[§]By *regularized experimental structure*, we mean an energy-minimized conformation close to the experimental structure. In the case of all-atom ECEPP/3 models, regularization is carried out as described in the section Fitting X-ray and NMR Structures to Rigid ECEPP Geometry. In the case of UNRES structures, regularization involves a series of minimizations of a target function composed of UNRES energy and a C^α distance-constraint

to the first or the second family (ranked according to energy) of conformations. Moreover, the energies of the regularized experimental structures

term. The distance constraints are calculated from the experimental structure and the weight of the distance-constraint term is gradually reduced to zero in consecutive minimizations [38]. This assures that minimization will not drift too far away from the experimental structure, because of possible bad initial overlaps.

(e)

(f)

FIGURE 4. (Continued)

were reasonably close to those of the calculated GMECs. The regularized experimental structure of protein A was 10 kcal/mol higher in energy compared to its GMEC and the regularized experimental structure of 1CLB was 16 kcal/mol higher in energy than its GMEC. This suggests that minimization of the energy of the regularized experi-

mental structure should be an appropriate test of the completeness of the search, in those cases in which the experimental-like structure is not among the structures obtained by the global conformational search. If the energy of the lowest-energy conformation found in the search is higher than that of the regularized experimental structure, the

(g)

FIGURE 4. (Continued)

search is clearly incomplete. We therefore compared the UNRES energies for the seven regularized experimental structures with those of our models. For most of the structures, the set of experimental coordinates was incomplete (some *N*- and/or

C-terminal residues were missing). Therefore, for comparison, we energy-minimized the experimentally observed fragments and the corresponding portion of our models. The results of UNRES energy minimization are included in columns 3 and 4

FIGURE 5. Stereoview of the superposition of the predicted structure of the *N*-terminal domain of DNAB helicase (target T0056; gray) on the nonregularized experimental structure (PDB code: 1JWE; black). The parts of the chains for which the predicted structure matches the experimental one (with 5.8 Å RMSD) are shown as ribbons; misaligned parts are shown as thin lines. The RMSD between the all-atom structure and the (nonregularized) experimental structure for whole chains is 15.7 Å.

FIGURE 6. Stereoviews of superpositions of two fragments of model 1 of HDEA (target T0061; yellow) on the corresponding parts of experimental structure [PDB code: 1bg8 [89]; red]. The parts of the chains for which the predicted structure matches the experimental one are shown as ribbons; misaligned parts are shown as thin lines: (a) the D25–I85 fragment superposed (with coordinate RMSD 4.2 Å); (b) the W16–K42 fragment superposed (with RMSD 2.9 Å). The C^α coordinate RMSD for the whole structure superposed is 8.9 Å and the C^α -distance RMSD is 5.4 Å, which is defined by $DRMSD = \{[1/N(N-1)] \sum_{i < j} [d_1(C_i^\alpha C_j^\alpha) - d_2(C_i^\alpha C_j^\alpha)]^2\}^{1/2}$, where N is number of the residues and $d_1(C_i^\alpha C_j^\alpha)$ and $d_2(C_i^\alpha C_j^\alpha)$ are the distances between the i th and the j th C^α atoms of the first and the second of the compared structures, respectively. (This figure appeared in Proc Natl Acad Sci USA 1999, 96, 5482–5485, copyright 1999 National Academy of Sciences, U.S.A., and is reproduced here by permission from the N.A.S.)

of Table III. As shown there, the RMSDs of the regularized experimental structures from the experimental coordinates are reasonably low (in the range of 0.8–3.4 Å). The UNRES energies of our models were all lower than the energies of the regularized experimental structures, even when truncated sequences were considered. This indicates that the conformational search by the CSA method seems to have found all low-energy conformations, while the UNRES force field, which is able to predict large portions of the structure correctly, needs further development.

Table III (columns 6 and 7) also includes ECEPP/3+SRFOPT energies of all-atom structures.

The energy-minimization procedure was carried out as described in the section Fitting X-ray and NMR Structures to Rigid ECEPP Geometry. These values are presented here for completeness, but they are not very useful to determine which structure is more probable, because the conformational search at the all-atom level was mostly carried out using distance constraints to maintain the C^α trace obtained in stage I.

TARGET T0056

The *N*-terminal domain of *E. coli* DNAB helicase consists of six α -helices. For this protein, the loca-

FIGURE 7. Stereoview of the superposition of model 4 of HDEA (target T0061; gray ribbon) on the experimental structure (black ribbon). All 76 residues are superposed with an RMSD of 7.3 Å.

tion of the six helical segments in the amino acid sequence was publicly available [88]. The protein represents a new type of fold [86]. The predicted model (the only model that we submitted for this target) and the NMR structure (PDB code: 1JWE) are shown in Figure 5. First, we carried out a simulation without using this secondary-structure information, as was done with all the other target proteins. Next, we carried out a biased conformational CSA search in which all the initial conformations were generated by assigning values of the

variable angles γ_i and θ_i that produce helical segments in the appropriate regions of the sequence. All the other angles were generated randomly, and no energy constraints were applied to maintain the helical fragments. Consequently, these helical portions were able to grow, split, shorten, or disappear during the course of the simulation. Except for some minor differences, the final conformations of both runs were similar; however, the lowest-energy conformation from the second (biased) run was lower in energy (by 4 kcal/mol; not shown

FIGURE 8. Stereoview of model 1 of T0061 (blue ribbon) superimposed on monomer A (red ribbon) of the experimental *dimer* of HDEA. The entire *N*-terminal part is shown, including the residues corresponding to the unresolved part of the crystal structure (which form a short α -helix in model 1). Misaligned helices H-1 and H-2, as well as the additional *N*-terminal α -helix of model 1, occupy the place of helix H-3 of monomer B (yellow ribbon) in the native dimer. The *C*-termini are indicated by the letter "C."

FIGURE 9. Stereoview of model 1 of target T0073 (the experimental structure, which was made available to us *after* we submitted our predictions, has not been published yet). The part of the chain for which the predicted structure matches the experimental one (with 1.2 Å RMSD) is shown as a ribbon; the misaligned part is shown as thin lines. The RMSD from the experimental structure for whole chains superposed is 12.0 Å.

in Table III) than the corresponding conformation from the unbiased run. This conformation was submitted after refinement. The final conformations composed a single family, within a 3.5 Å C^α RMSD cutoff.

The C^α -RMSD plot for model 1 of T0056 and the (nonregularized) experimental structure [Fig. 4(a)] shows a plateau in a residue range 23–35, with the RMSD below 3.0 Å. This corresponds to good superposition of two C-terminal helices, starting from L83. Superposition continues to be fairly good, until fragment length 59 (RMSD of 5.8 Å). Within this range, three C-terminal helices superpose (Fig. 5, Table II). After this fragment length, there is a sudden jump in the RMSD plot caused

by a different orientation of the N-terminal fragment with respect to the experimental structure (Fig. 5).

TARGET T0061

The experimental structure of HDEA (PDB code: 1BG8) is a five-helix bundle with a long loop between the third (H-3) and the fourth (H-4) helix [89]. The experimental structure has two cysteine residues, C18 and C66, which form a disulfide bond, but this information was not available to predictors. Consequently, we carried out two CSA calculations considering sequences with and without a disulfide bond. In the CSA run without a disulfide bond, we obtained four families of conformations, from

FIGURE 10. Stereoview of the superposition of the predicted structure (model 2) of the N-terminal domain of the EH2 domain of human EPS15 (target T0074; gray) on the experimental structure (PDB code: 1EF2 [92]; black). The parts of the chains for which the predicted structure matches the experimental one (with 5.8 Å RMSD) are shown as ribbons; misaligned parts are shown as thin lines. The RMSD from the experimental structure for whole chains superposed is 10.0 Å. (The experimental structure is reproduced by permission from M. Overduin.)

FIGURE 11. Stereoview of model 3 of target T0076 (the experimental structure, which was made available to us *after* we submitted our predictions, has not been published yet). The part of the chain for which the predicted structure matches the experimental one (with 6.8 Å RMSD) is shown as a ribbon; misaligned parts are shown as thin lines. The RMSD from the experimental structure for whole chains superposed is 16.5 Å.

which models 1, 2, and 4 were derived (ranked according to UNRES energy). We submitted three models without a disulfide bond and one with a disulfide bond; comparison of energies among these two groups of models is not possible because of their different covalent structure (both UNRES and ECEPP/3 include only the conformational energy and not the disulfide-bond energy). The CSA run with a disulfide bond found two dominant families of conformations: one from which model 3 was de-

rived, and the second having the same fold as the structure submitted as model 4. A family of structures equivalent to model 3 was also obtained in the first CSA run without a disulfide bond. In models 1 and 2, the cysteine residues are far apart; hence, no disulfide bond can be formed in these models. All final conformations had an α -helical structure with the cysteine residues in helical regions. Experimentally, S—S bonds are rarely observed between helices; therefore, we submitted only one model

FIGURE 12. Stereoview of superposition of the predicted structure of MarA (target T0079; gray) on the experimental structure (PDB code: 1BL0 [93]; black). The parts of the chains for which the predicted structure matches the experimental one (with 6.0 Å RMSD) are shown as ribbons; misaligned parts are shown as thin lines. The RMSD from the experimental structure for whole chains superposed is 11.0 Å.

FIGURE 13. Stereoview of model 1 of target T0084 (the experimental structure, which was made available to us *after* we submitted our predictions, has not been published yet). The part of the chain for which the predicted structure matches the experimental one (with 0.9 Å RMSD) is shown as ribbons; misaligned parts are shown as thin lines. The RMSD from the experimental structure for whole chains superposed is 11.1 Å.

(model 3) with a disulfide bond and all others with no such bond.

The RMSD plots of the four submitted models are shown in Figure 4(b). For model 1, good alignment continues until fragment length 61 (RMSD of 4.2 Å). This corresponds to the alignment of three C-terminal helices (H-3, H-4, and H-5). The sudden jump in the RMSD plot after length 61 is caused by misalignment of the *N*-terminal helix. This is illustrated in Figure 6(a) for model 1. It should be noted

that the 61-residue fragment that matches within a C α RMSD of 4.2 Å comprises 80% of the structure (Table II). Superposing the *N*-terminal helices H-2 and H-3 (residues W16 to K42), we obtain an RMSD for the C α atoms of 2.9 Å (Fig. 7), which means that the misaligned helix in this model could be placed in the correct position by a single rotation about one C α —C' bond (ψ of A32); the overall RMSD after this transformation would change from 8.9 Å to about 6.0 Å. The fact that only a minor portion of the

TABLE II
Summary of the RMSDs between the best-matching fragments of our models of the seven targets submitted to CASP3 and the corresponding parts of the (nonregularized) experimental structures.

Model ^a	Length ^b	Size ^c	Fragment ^d	RMSD ^e	% Matching ^f
T0056(1)	114/114	59	(71–129)	5.8	52
T0061(1)	89/76	61	(25–85)	4.2	80
T0061(1)	89/76	27	(16–42)	2.9	36
T0073(1)	48/48	32	(1–32)	1.2	67
T0074(2)	98/95	53	(160–212)	5.8	56
T0076(3)	140/140	62	(76–137)	6.8	44
T0079(1)	129/116	61	(9–69)	6.0	53
T0084(1)	37/30	13	(21–33)	0.9	42

^a The number in parentheses indicates the model number of the target.
^b Total chain length/length of the fragment, whose structure was determined experimentally.
^c Size of the matching fragment.
^d First and last residues of the matching fragment.
^e RMSDs (in Å) between the matching fragments and the corresponding parts of the (nonregularized) experimental structures. It should be noted that some of the RMSDs reported by Orengo et al. [86] are based on a different superposition criterion.
^f The definition of % matching is % matching = (chain length of the matching fragment/chain length of the experimental structure) $\times$ 100%.

TABLE III
UNRES and ECEPP/3+SRFOPT energies of the regularized experimental structures and models of seven CASP3 targets.^a

Target	Structure	UNRES			ECEPP/3+SRFOPT		
		Energy (kcal/mol)		C α RMSD ^c (Å)	Energy (kcal/mol)		C α RMSD ^d (Å)
		Full	Truncated ^b		Full	Truncated ^b	
T0056	Exp. ^e	−220.3	—	2.9	−995.9	—	1.5
	Model 1	−503.3	—	15.2	−1292.0	—	15.7
T0061	Exp. (no SS) ^{e,f}	—	−258.5	3.4	—	−656.5	1.2
	Exp. (with SS) ^{e,f}	—	−268.6	3.4	—	−807.5	1.1
	Model 1 (no SS)	−333.5	−289.9	8.8	−805.6	−592.4	8.9
	Model 2 (no SS)	−329.3	−293.0	9.1	−796.9	−600.3	9.1
	Model 3 (with SS)	−336.5	−293.1	8.8	−871.2	−687.9	9.2
	Model 4 (no SS)	−328.0	−286.8	7.4	−774.9	−628.5	7.3
T0073	Exp. ^e	−152.6	—	2.6	−364.9	—	0.1
	Model 1	−165.6	—	12.0	−444.4	—	12.0
	Model 2	−164.3	—	11.2	−412.1	—	11.2
T0074	Exp. ^e	—	−285.7	0.3	—	−819.1	1.2
	Model 1	−490.2	−425.2	12.5	−732.7	−631.2	12.9
	Model 2	−484.7	−423.0	9.6	−908.9	−745.6	10.0
T0076	Exp. ^e	−380.4	—	0.8	−857.0	—	4.9
	Model 1	−613.9	—	15.6	−1233.5	—	15.7
	Model 2	−603.0	—	17.3	−1241.4	—	17.7
	Model 3	−601.0	—	16.6	−1297.6	—	16.5
	Model 4	−599.8	—	15.5	−1284.2	—	15.7
T0079	Exp. ^e	—	−418.0	3.1	—	−1252.8	0.8
	Model 1	−535.5	−486.5	10.8	−1459.1	−1260.3	11.0
	Model 2	−534.4	−480.9	11.1	−1278.0	−1223.4	11.3
	Model 3	−533.8	−485.3	10.8	−1420.7	−1247.8	11.0
	Model 4	−529.9	−470.2	11.6	−1436.2	−1233.5	11.7
T0084	Exp. ^e	—	−81.2	1.0	—	−421.8	0.9
	Model 1	−119.7	−84.4	10.9	−450.7	−361.6	11.1
	Model 2	−117.4	−88.5	10.0	−483.1	−381.3	10.4
	Model 3	−119.7	−87.9	10.4	−466.8	−365.8	10.5
	Model 4	−118.9	−86.8	10.2	−460.6	−384.3	10.7
	Model 5	−115.5	−86.5	10.2	−464.2	−344.6	10.6

^a In contrast to Table II, which pertains only to superposed fragments, the data in this table pertain to the whole experimental structure.

^b Some *N*- and *C*-terminal residues were missing in the experimental structures of these targets; therefore, in order to compare the energies, the corresponding residues of the models were also removed and the energy was reminimized.

^c The RMSD values listed in the “experimental structure” entries of this column are the C α RMSDs of the regularized conformations (with the UNRES force field) from the (nonregularized) experimental structures (see text). The values listed in the “model” entries are C α RMSDs of the models from the corresponding (nonregularized) experimental structures.

^d The RMSD values listed in the “experimental structure” entries of this column are the C α RMS deviations of the regularized (using standard ECEPP/3 geometries; see the section Fitting X-ray and NMR Structures to Rigid ECEPP Geometry) and subsequently energy-minimized conformations from the (nonregularized) experimental structures. The values listed in the “model” entries are C α RMSDs of the models from the corresponding (nonregularized) experimental structures.

^e This is the regularized experimental structure.

^f The experimental structure of T0061 was minimized with no disulfide bond to compare the energy with those of models 1, 2, and 4, and with a disulfide bond, to compare the energy with that of model 3.

structure is misaligned is reflected in the small value of the average difference between the $C^\alpha \dots C^\alpha$ distances of the model and the crystal structure, which is 5.4 Å.

The three-helix fragment composed of helices H-3, H-4, and H-5 constitutes a rare fold, which was not recognized by any threading approach in CASP3, although one such fragment exists in the PDB [86]. Instead, the mirror-image fold is dominant in the PDB, and this was retrieved by threaders.

Model 2 has a similar topology for the C-terminal three-helix fragment, but the RMSD from the (nonregularized) experimental structure is greater [Fig. 4(b)]. Model 3 has the mirror-image topology of the three-helix fragment, which shows up in the RMSD plot as a steady increase of RMSD with fragment length [Fig. 4(b)]. Model 4 is the only one with the correct topology (Fig. 7) of all helices (defined in terms of interhelix contacts [90]), which is manifested as a modest increase in RMSD after fragment length 34. However, model 4 has an RMSD of 7.3 Å for all 76 residues observed in the crystal structure. This is caused by the fact that the helices are almost parallel, which is different from the arrangement in the experimental structure [Fig. 6(a)]. Models 1, 2, and 4 have essentially the same kind of fold of the C-terminal part, differing only in the angles at which the helices pack against each other. This result strongly suggests that, in this particular case, our method has located a core of the structure (formed by three C-terminal helices corresponding to 80% of the structure), and different families of conformations can arise from alternative packing of the remaining helices. The best-matching conformation from the family of model 4 has an RMSD of 6.9 Å and an UNRES energy 5 kcal/mol higher than that of model 4 and was therefore not submitted to CASP3. The range of the RMSDs between the experimental structure and the set of *all* final 100 UNRES conformations from the two CSA runs was 6.9 to 10.5 Å.

Matching between the secondary structure elements of our four models and the experimental structure is quite good {79.5, 59.7, 52.7, and 72.4% according to the SOV score [91] and 73.7, 63.2, 65.8, and 78.9% according to the Q3 score [percentage of residues predicted correctly for all three conformational states (helix, strand, coil)][†] for models 1–4, respectively; values for models 1 and 4 are among

the top SOV and Q3 values obtained in the CASP3 experiment, including *ab initio* secondary structure predictions}.

The misalignment of the *N*-terminal helix in our model 1 could be caused by two factors: First, the experimental structure contains a disulfide bond (which we considered unlikely at the time the prediction was carried out), which forces the correct orientation of helix H-2. Second, the experimental structure is a dimer [89]. The misaligned helices H-1 and H-2, as well as a short *N*-terminal helix which corresponds to the part of the chain not present in the crystal structure [89], pack against H-3 and H-4 in model 1, which, to some extent, substitutes for the packing of helix H-3 of the second molecule in the dimer (Fig. 8). The packing of the misaligned helices H-1 and H-2 in model 1 is probably a contributory reason for the geometry of the H-3–H-5 fragment, which is closer to that of the experimental structure than model 4, although the latter has the correct helix topology. Both in the experimental dimer and in model 1, helix H-3 is forced to pack almost orthogonally against helix H-4 (Fig. 8).

TARGET T0073

The experimental structure is a dimer of two α -hairpins packed head-to-tail to form a four-helix bundle [86]. The first helix extends from residue 1 to 14 and the second one from residue 21 to 48. Without being given the information that the structure is a dimer, our prediction was carried out under the assumption that it is monomeric. Two families of conformations resulted from the CSA runs; therefore, we submitted two models, each of which is a three-helix bundle. The positions of the first and a part of the second helix of our models match the positions of the native helices (Fig. 9). The third helix in our model, extending from residue 34 to residue 48, arises from a break in the second native helix at residue 32. This break is caused by the absence of the second monomer molecule in our model, which would therefore leave the hydrophobic face of the monomer exposed. In model 1, the C-terminal portion of the broken helix packs against the same face of the *N*-terminal portion of the hairpin as in the native dimer. Moreover, in model 1, the first two helices are packed at an angle very close to that of the experimental structure, with very accurate matching of the turn region. This can be observed on the corresponding RMSD plot [Fig. 4(c)]. The alignment is very good until residue 32 [Fig. 4(c) and Table II], with an RMSD of 1.2 Å; then, there is a sudden jump

[†]See also <http://PredictionCenter.llnl.gov/casp3/results/ab-measures.html> for quantities used to evaluate prediction results.

in the RMSD caused by the break in the second helix. Model 2 is the mirror image of model 1, which shows up in the RMSD plot as very good matching of a single helix followed by a jump in the RMSD corresponding to the *N*-terminal α -hairpin portion, where the two helices of model 2 are packed at a different angle than in the experimental structure.

TARGET T0074

The NMR structure of the EH2 domain of human EPS15 (PDB code: 1EF2) [92] consists of two EF-hand motifs with calcium ions bound to each of the loops. The calcium ions were not included in our calculations. We carried out one CSA run, resulting in two families of structures, and we submitted two models differing in the orientation of the *N*-terminal domain. The RMSD plot [Fig. 4(d)] shows that the agreement between our models and the experimental structure is reasonable until the fragment length reaches about 50 (RMSD less than 6 Å; Table II). This is caused by the fact that the topology of the second EF-hand (extending from L160 to R215) is essentially correct in our models (Fig. 10), while the first EF-hand did not come out correctly in our predictions. As shown [Fig. 4(d)], model 2 is in better agreement with the experimental structure than is model 1.

TARGET T0076

The NMR structure of CDC4P consists of two separate domains [88], as in other proteins of the kinase family. Two domains are also observed in all our models; it should be noted that the information about the two-domain structure of the protein was not known to us a priori. The RMSD plot shows that 63 and 62 residue fragments of models 3 and 4, respectively, match the NMR structure within 6.8 Å RMSD [Fig. 4(e); Table II]. These fragments correspond to the second (C-terminal) calcium-binding domain that contains an EF-hand motif (Fig. 11). The discrepancy between the three-helix C-terminal domain of our models 3 and 4 and the NMR structure is caused by close packing of the last C-terminal helix, extending from H131 to N140, against the other two helices of the domain in our structures. By contrast, the C-terminal helix is quite exposed in the NMR structure, leaving large hydrophobic patches exposed to the solvent. Our prediction is less accurate for the *N*-terminal domain than for the C-terminal domain. However, a two-helix 32-residue fragment extending from S5

to Q36 of model 3 has an RMSD of 4.3 Å from the corresponding part of the experimental structure. As opposed to the NMR structure, in our models, the two domains are tightly packed, resulting in large overall RMSDs. The loose arrangement of the two domains, each of which has significantly exposed hydrophobic patches in the NMR structure, may indicate that the structure is a dynamic average of quite different conformers (in terms of the orientation and packing of the domains). Although the overall tertiary structure of CDC4P is different in our predictions from the NMR structure, the secondary structure of our models is in good agreement with the experimental secondary structure, the average Q3 value for models 1–4 being 75%.

TARGET T0079

In the crystal structure of a complex with DNA (PDB code: 1BL0), MarA appears as a bipartite helix-turn-helix protein [93]. It contains two domains composed of four helices each, each of which binds to separate parts of DNA. Our simulations resulted in low-energy conformations containing two domains, a feature that was not known to us a priori. Our predicted structure has the same topology of the *N*-terminal domain as in the crystal [93] (residues 1–8 are not observed in the X-ray structure). The RMSD between the predicted and experimental structures for the fragment containing residues D9 to A69 is 6.0 Å for the C α atoms [Figs. 4(f) and 12; Table II]. The major difference between our prediction and the crystal structure is that the C-terminal domain in the predicted structure is a mirror image of that part of the experimental structure. Furthermore, the long helix, H-4, connecting the two domains in the X-ray structure is significantly kinked in the middle in our predicted structure. This leads to tight packing of the two domains in our model, while in the crystal structure the *N*- and the C-terminal domains each bind separately to DNA and are not packed against each other. Our calculations were carried out on this protein in the absence of the DNA.

TARGET T0084

In the crystal structure, this peptide forms a tetramer, resulting in a tightly packed four-helix bundle [86]. Each monomer is a single α -helix. Not having been informed that the structure is a tetramer, we predicted it as a monomer. We divided the models into three classes: models 1 and 2, models 3 and 4, and model 5. The two models of the

second class are mirror images of their counterparts of the first class. The RMSD plot [Fig. 4(g)] shows that only 12- to 14-residue fragments match reasonably well for models 1–5, and for model 1, the RMSD is less than 1 Å (Table II). These fragments correspond to the C-terminal part of the native helix; the N-terminal part of the native helix is largely disrupted and packed against the hydrophobic face of the C-terminal helix (Fig. 13).

TIME REQUIREMENTS OF DIFFERENT STAGES OF THE HIERARCHICAL APPROACH

As shown in Table I, a global conformational search at stage I requires from 1 h (T0084, 37 residues) to 120 h (T0076, 140 residues) wall-clock time on 64 processors of an IBM SP2 supercomputer. This use of CPU, although significant, still enables the key stage of the prediction to be accomplished in reasonable time.

The time required for stage II ranges from a few seconds to several minutes per structure on one processor of an IBM SP2 supercomputer and can therefore be considered negligible. Refining backbone conformation in the polyalanine model in stage III requires, on average, several hours wall-clock time with up to 20 processors of an IBM SP2 supercomputer. Stage IV (introducing the side chains at nonoverlapping conformations) requires up to 20 h on one processor of an IBM SP2 supercomputer; this task cannot easily be parallelized. Refining all-atom conformations in stage V requires 30–50 h on 20 processors of an IBM SP2 supercomputer. For stage V, the timing depends strongly on the desired refinement level of the all-atom structure: A reasonable all-atom structure can be obtained after several EDMC iterations which requires only a few hours on up to 10 processors of an IBM SP2 supercomputer. Therefore, stage I is the most time-consuming step of our approach.

Except for stage IV, every stage of the hierarchical approach has been efficiently parallelized. Thus, with a larger parallel system, the wall-clock time required for the most important stage I can be reduced to several hours.

Conclusions

Until recently, protein-structure prediction based solely on the *thermodynamic hypothesis* was considered to be infeasible [16, 17]. The reasons for this point of view were the inaccuracy of the potential

energy functions employed to represent the energy of the protein and the lack of powerful methods for global optimization. The results of blind predictions presented in this article, however, show that large portions (~60 amino acid residues) of the three-dimensional structures of globular proteins with sizes ranging from 89 to 140 amino acid residues can now be predicted with an accuracy of 4–7 Å C α RMSD using our state-of-the-art energy-based approach. For fragments of this size, RMSD values below 7 Å indicate that the predicted topology of the fragment is correct. Recent work showed that the probability of obtaining a purely chance prediction with a 6 Å RMSD is extremely low [94]. The secondary and super-secondary structure of our models were predicted well for all targets, the average value of Q3 over all 22 models being about 72%. There are two reasons for this success: First, although our UNRES force field is approximate, it is based on physical grounds, both in the derivation of the energy terms and in their parameterization [37, 38, 47]. Second, the CSA method of conformational search appears to locate all significant low-energy structures in a reasonable time. This is very important because, given the approximations inherent in the force field, the lowest-energy structure should not be considered as the only possible candidate for the prediction, but a few distinct low-energy structures should be taken into account.

It should be noted that the predicted structures were more compact than were the experimental ones. For small synthetic proteins (targets T0073 and T0084), this might be due to the difference between monomeric and multimeric structures. For example, the experimental structure of target T0073 has hydrophobic residues that are greatly exposed in the monomer. The crystal structure of T0079 was obtained in a complex with DNA, which was not present in our simulations; this too could lead to greater compactness of our structure. The NMR structure of T0076 is probably a dynamic average and, therefore, in our prediction, which corresponds to a *static* energy-minimized structure, the two domains are tightly packed, resulting in screening of hydrophobic residues from the solvent. In the NMR structure, the domains are far apart, allowing significant exposure of hydrophobic residues. However, even taking into account the above explanations, another possible cause of greater compactness of our models may be that the current version of the UNRES force field does not stabilize the secondary structure sufficiently strongly; this is also the reason why this force field cannot yet treat β -sheet proteins

properly. This is currently being improved by introducing additional multibody terms which were missing in UNRES at the time when the predictions were carried out and by fine-tuning the energy parameters.

Once again, we stress that our approach uses only energy criteria to select the best prediction, without the need of secondary structure prediction, multiple-sequence alignment, or threading. In principle, these tools can be used with our approach, as they are in other *ab initio* approaches [18, 95, 96]. However, while this might save time required for conformational search, it introduces a bias coming from the dependence of the method on a structural database. As shown by the example of target T0061, the fact that a particular fold occurs once or a few times in the database used for threading does not guarantee that a threading approach will be able to retrieve it; it might, instead, retrieve a more common, but incorrect, fold. Problems of this type will remain even if the database is expanded. With our approach, which is based on the physical principles of protein folding, it can be expected that more accurate force fields will lead to even better results. In addition to this, a force field good enough for protein-structure prediction along with a powerful conformational search method could also be used to study the mechanisms of protein folding.

The key stage of our approach is the complete conformational search at a coarse-grained (united-residue) level. Wrong energy relations and major errors in the structure at this level cannot be corrected when passing to all-atom resolution. Therefore, two elements are vital for the performance of the whole procedure: a good united-residue force field and a powerful conformational-search method. Because of errors and approximations inherent in the united-residue force field, the conformational search should retrieve not only the lowest-energy conformation, but also all geometrically different structures lying within a reasonable energy cutoff. In our current approach, the UNRES force field appears to represent the approximate potential energy surface of proteins that are mainly helical reasonably well, and the CSA method has shown itself to be an appropriate tool for efficient conformational search. With the current tools, and with state-of-the-art computer power, structural predictions can be made in reasonable time: For example, for target T0076 (140 amino acid residues), a complete conformational search took about 5 days of computations on 64 processors of an IBM SP2 supercomputer (Table I).

ACKNOWLEDGMENTS

We thank D. Sverdlik and D. Dwyer of the Cornell Theory Center for technical assistance. This research was supported by grants from the National Science Foundation (MCB95-13167), the National Institutes of Health (GM-14312), the NIH National Center for Research Resources (P41RR-04293), and the Polish State Committee for Scientific Research (KBN) (DS 8245-4-0124-9). Support was also received from the National Foundation for Cancer Research. The computations in this work were carried out at the Cornell Theory Center which receives funding from Cornell University, New York State, the National Center for Research Resources at the National Institutes of Health, the National Science Foundation, the Department of Defense Modernization Program, and members of the Corporate Partnership Program. Part of the computations were carried out at the Interdisciplinary Center for Mathematical and Computational Modeling (ICM) in Warsaw, and at the Informatics Center of the Metropolitan Academic Network (ICMAN) in Gdansk, Poland.

References

1. Scheraga, H. A.; Némethy, G. In *Molecules in Natural Science and Medicine—An Encomium for Linus Pauling*; Maksic, Z. B.; Maksic, M. E.; Eds.; Ellis Horwood: Chichester, 1991; pp. 141–176.
2. Vásquez, M.; Némethy, G.; Scheraga, H. A. *Chem Rev* 1994, 94, 2183.
3. Holm, L.; Sander, C. *J Mol Biol* 1993, 233, 123.
4. Holm, L.; Sander, C. *Nucleic Acids Res* 1997, 25, 231.
5. Holm, L.; Sander, C. *Nucleic Acids Res* 1999, 27, 244.
6. Orengo, C. *Curr Opin Struct Biol* 1994, 4, 429.
7. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B., Jr.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J Mol Biol* 1977, 112, 535.
8. Warme, P. K.; Momany, F. A.; Rumball, S. V.; Tuttle, R. W.; Scheraga, H. A. *Biochemistry* 1974, 13, 768.
9. Jones, T. A.; Thirup, S. *EMBO J* 1986, 5, 819.
10. Clark, D. A.; Shirazi, J.; Rawlings, C. *J Prot Eng* 1991, 7, 751.
11. Rooman, M. J.; Wodak, S. J. *Biochemistry* 1992, 31, 10239.
12. Johnson, M. S.; Overington, J. P.; Blundell, T. L. *J Mol Biol* 1993, 231, 735.
13. Fischer, D.; Rice, D.; Bowie, J. U.; Eisenberg, D. *FASEB J* 1996, 10, 126.
14. Sippl, M. J. *J Comput-Aid Mol Des* 1993, 7, 473.
15. Anfinsen, C. B. *Science* 1973, 181, 223.
16. Mirny, L. A.; Shakhnovich, E. I. *J Mol Biol* 1998, 283, 507.
17. Jones, D. T. *Curr Opin Struct Biol* 1997, 7, 377.

18. Skolnick, J.; Koliński, A.; Ortiz, A. R. *J Mol Biol* 1997, 265, 217.
19. Scheraga, H. A. *Int J Quantum Chem* 1992, 42, 1529.
20. Scheraga, H. A. *Biophys Chem* 1996, 59, 329.
21. Piela, L.; Kostrowicki, J.; Scheraga, H. A. *J Phys Chem* 1989, 93, 3339.
22. Li, Z.; Scheraga, H. A. *Proc Natl Acad Sci USA* 1987, 84, 6611.
23. Li, Z.; Scheraga, H. A. *J Mol Struct (Theochem)* 1988, 179, 333.
24. Ripoll, D. R.; Scheraga, H. A. *Biopolymers* 1988, 27, 1283.
25. Kostrowicki, J.; Piela, L. *J Optimiz Theory Appl* 1991, 69, 269.
26. Kostrowicki, J.; Scheraga, H. A. *J Phys Chem* 1992, 96, 7442.
27. Olszewski, K. A.; Piela, L.; Scheraga, H. A. *J Phys Chem* 1992, 96, 4672.
28. Lee, J.; Scheraga, H. A.; Rackovsky, S. *J Comput Chem* 1997, 18, 1222.
29. Lee, J.; Scheraga, H. A.; Rackovsky, S. *Biopolymers* 1998, 46, 103.
30. Pillardy, J.; Liwo, A.; Groth, M.; Scheraga, H. A. *J Phys Chem* 1999, 103, 7353.
31. Levitt, M.; Warshell, A. *Nature* 1975, 253, 694.
32. Pincus, M. R.; Scheraga, H. A. *J Phys Chem* 1977, 81, 1579.
33. Crippen, G. M.; Viswanadhan, V. N. *Int J Peptide Protein Res* 1984, 24, 279.
34. Skolnick, J.; Koliński, A. *Science* 1990, 250, 1121.
35. Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci* 1993, 2, 1697.
36. Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci* 1993, 2, 1715.
37. Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J Comput Chem* 1997, 18, 849.
38. Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J Comput Chem* 1997, 18, 874.
39. Liwo, A.; Lee, J.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proc Natl Acad Sci USA* 1999, 96, 5482.
40. Liwo, A.; Pillardy, J.; Kazmierkiewicz, R.; Wawak, R. J.; Groth, M.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A. *Theor Chem Acc* 1999, 101, 16.
41. Koliński, A.; Skolnick, J. *Proteins Struct Funct Genet* 1994, 18, 338.
42. Koliński, A.; Skolnick, J. *Proteins Struct Funct Genet* 1994, 18, 353.
43. Purisima, E. O.; Scheraga, H. A. *Biopolymers* 1984, 23, 1207.
44. Bassolino-Klimas, D.; Brucoleri, R. E. *Proteins* 1992, 14, 465.
45. Rey, A.; Skolnick, J. *J Comput Chem* 1992, 13, 443.
46. Wang, Y. L.; Huq, H. I.; de la Cruz, X. F.; Lee, B. K. *Fold Des* 1998, 3, 1.
47. Liwo, A.; Kazmierkiewicz, R.; Czaplewski, C.; Groth, M.; Oldziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J Comput Chem* 1998, 19, 259.
48. Lee, J.; Scheraga, H. A. *Int J Quantum Chem* 1999, 75, 255.
49. Lee, J.; Liwo, A.; Scheraga, H. A. *Proc Natl Acad Sci USA* 1999, 96, 2025.
50. Duan, Y.; Kollman, P. A. *Science* 1998, 282, 740.
51. Lee, J.; Liwo, A.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proteins Colm Struct Funct Genet* 1999, Suppl. 3, 204.
52. Ripoll, D. R.; Scheraga, H. A. *J Protein Chem* 1989, 8, 263.
53. Ripoll, D. R.; Liwo, A.; Scheraga, H. A. *Biopolymers* 1998, 46, 117.
54. 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.
55. Vila, J.; Williams, R. L.; Vázquez, M.; Scheraga, H. A. *Proteins Struct Funct Genet* 1991, 10, 199.
56. Godzik, A.; Koliński, A.; Skolnick, J. *J Comput-Aid Mol Des* 1993, 7, 397.
57. Piela, L.; Scheraga, H. A. *Biopolymers* 1987, 26, S33.
58. Koliński, A.; Godzik, A.; Skolnick, J. *J Chem Phys* 1993, 98, 7420.
59. Hinds, D. A.; Levitt, M. *J Mol Biol* 1994, 243, 668.
60. Crippen, G. M. *J Mol Biol* 1996, 260, 467.
61. Némethy, G.; Pottle, M. S.; Scheraga, H. A. *J Phys Chem* 1983, 87, 1883.
62. Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc Natl Acad Sci USA* 1992, 89, 9029.
63. Sali, A.; Shakhovich, E. I.; Karplus, M. *Nature* 1994, 369, 248.
64. Hao, M.-H.; Scheraga, H. A. *J Phys Chem* 1994, 98, 9882.
65. Hao, M.-H.; Scheraga, H. A. *Proc Natl Acad Sci USA* 1996, 93, 4984.
66. Goldberg, D. E. *Genetic Algorithms in Search, Optimization & Machine Learning*; Addison-Wesley: Reading, MA, 1989.
67. Rabow, A. A.; Scheraga, H. A. *Protein Sci* 1996, 5, 1800.
68. Gay, D. M. *ACM Trans Math Software* 1983, 9, 503.
69. Hao, M.-H.; Rackovsky, S.; Liwo, A.; Pincus, M. R.; Scheraga, H. A. *Proc Natl Acad Sci USA* 1992, 89, 6614.
70. Vázquez, M.; Scheraga, H. A. *J Biomol Struct Dyn* 1988, 5, 757.
71. Palmer, K. A.; Scheraga, H. A. *J Comput Chem* 1991, 12, 505.
72. Gö, N.; Scheraga, H. A. *Macromolecules* 1970, 3, 178.
73. Ripoll, D. R.; Vázquez, M. J.; Scheraga, H. A. *Biopolymers* 1991, 31, 319.
74. Liwo, A.; Tempczyk, A.; Oldziej, S.; Shenderovich, M. D.; Hruby, V. J.; Talluri, S.; Ciarkowski, J.; Kasprzykowski, F.; Lankiewicz, L.; Grzonka, Z. *Biopolymers* 1996, 38, 157.
75. Ripoll, D. R.; Piela, L.; Vázquez, M.; Scheraga, H. A. *Proteins Struct Funct Genet* 1991, 10, 188.
76. Liwo, A.; Gibson, K. D.; Scheraga, H. A.; Brandt-Rauf, P. W.; Monaco, R.; Pincus, M. R. *J Protein Chem* 1994, 13, 237.
77. Ashkenazi, G.; Ripoll, D. R.; Lotan, N.; Scheraga, H. A. *Biosen Bioelect* 1997, 12, 85.
78. Ripoll, D. R.; Vorobjev, Y. N.; Liwo, A.; Vila, J. A.; Scheraga, H. A. *J Mol Biol* 1996, 264, 770.
79. Vila, J. A.; Ripoll, D. R.; Villegas, M. E.; Vorobjev, Y. N.; Scheraga, H. A. *Biophys J* 1998, 75, 2637.
80. Vila, J. A.; Ripoll, D. R.; Vorobjev, Y. N.; Scheraga, H. A. *J Phys Chem B* 1998, 102, 3065.
81. Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. *J Phys Chem* 1975, 79, 2361.
82. Sippl, M. J.; Némethy, G.; Scheraga, H. A. *J Phys Chem* 1984, 88, 6231.
83. Perrot, G.; Cheng, B.; Gibson, K. D.; Vila, J.; Palmer, K. A.; Nayeem, A.; Maigret, B.; Scheraga, H. A. *J Comput Chem* 1992, 13, 1.

84. Späth, H. Cluster Analysis Algorithms; Halsted: New York, 1980.
85. Golub, G. H.; Van Loan, C. F. Matrix Computations; Johns Hopkins University: Baltimore, MD, 1985.
86. Orengo, C. A.; Bray, J. E.; Hubbard, T.; LoConte, L.; Sillitoe, I. Proteins: Struct Funct Genet 1999, Suppl. 3, 149.
87. Snow, M. E. J Comput Chem 1992, 13, 579.
88. CASP3, Third Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction; <http://predictioncenter.llnl.gov/casp3/Casp3.html>; Asilomar Center: Pacific Grove, CA, 1998.
89. Yang, F.; Gustafson, K. R.; Boyd, M. R.; Wlodawer, A. Nature Struct Biol 1998, 5, 763.
90. Chothia, C. Nature (Lond) 1975, 254, 304.
91. Zemla, A.; Venclovas, C.; Fidelis, K.; Rost, B. Proteins Struct Funct Genet 1999, 34, 220.
92. De Beer, T.; Carter, R. E.; Lobel-Rice, K. E.; Sorkin, A.; Overduin, M. Science 1998, 281, 1357.
93. Rhee, S.; Martin, R. G.; Rosner, J. L.; Davies, D. R. Proc Natl Acad Sci USA 1998, 95, 10413.
94. Reva, B. A.; Finkelstein, A. V.; Skolnick, J. Fold Des 1998, 3, 141.
95. Bystroff, C.; Baker, D. J Mol Biol 1998, 281, 565.
96. Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D. Proteins Struct Funct Genet 1999, 34, 82.