Acta Biochimica Polonica

Vol. 44 No. 3/1997

527–548

QUARTERLY

Design of a knowledge-based force field for off-lattice simulations of protein structure*,**

Adam Liwo, Stanisław Ołdziej, Rajmund Kaźmierkiewicz, Małgorzata Groth and Cezary Czaplewski

Department of Chemistry, University of Gdańsk, J. Sobieskiego 18, 80-952 Gdańsk, Poland

Key words: protein structure prediction, mean-field potential, united-residue representation of polypeptide chains, Monte Carlo methods

Prediction of protein structure from amino-acid sequence still continues to be an unsolved problem of theoretical molecular biology. One approach to solve it is to construct an appropriate (free) energy function that recognizes the native structures of some selected proteins (whose native structures are known) as the ones distinctively lowest in (free) energy and then to carry out a search of the lowest-energy structure of a new protein. In order to reduce the complexity of the problem and the cost of energy evaluation, the so-called *united-residue* representation of the polypeptide chain is often applied, in which each amino-acid residue is represented by only a few interaction sites. Once the global energy minimum of the simplified chain has been found, the all-atom structure can easily and reliably be constructed. The search of the lowest-energy structure is usually carried out by means of Monte Carlo methods, though use of more efficient global-optimization methods, especially those of deformation of original energy surface is potentially promising. Monte Carlo search of the conformational space can be accelerated greatly, if the chain is superposed on a discrete lattice (the on-lattice approach). On the other hand, the on-lattice approach prohibits the use of many efficient global-optimization methods, because they require both energy and its space derivatives. The on-lattice methods in which the chain is embedded in the continuous 3D space are, therefore, also worth developing. In this paper we summarize the work on the design and implementation of an off-lattice united-residue force field that is underway in our group, in cooperation with Professor H.A. Scheraga of Cornell University, U.S.A.

*This work was supported by a grant PB 190/T09/96/10 from the State Committee for Scientific Research (KBN).

**Computations were carried out on the IBM-SP2 computer at the Cornell National Supercomputer Facility, a resource of the Center for Theory and Simulation in Science and Engineering at Cornell University, which is funded by the National Science Foundation, New York State, the IBM Corporation, and members of its Corporate Research Institute, with additional funds from the National Institutes of Health, on the IBM-SP2 computer at the CI TASK at the Technical University of Gdańsk and on the CRAY YMP/4E and POWER CHALLENGE R-10000 computers at the Interdisciplinary Center for Mathematical and Computational Modelling of the University of Warsaw, Poland.

✉Correspondence: tel: (+058)+41-52-61 ext. 261; fax: 41-03-57; e-mail: adam@sun1.chem.univ.gda.pl.

Abbreviations: APP, avian pancreatic polypeptide; DEM, diffusion equation method; EDMC, electrostatically driven Monte Carlo method; MCM, Monte Carlo method with minimization; p, united peptide group; PDB, Protein Data Bank; r.m.s., root mean square; SC, united site chain; SCMTF, self consistent mean torsional field.

Proteins are an essential constituent of all known living organisms. Fibrous proteins, such as collagen or keratin, are main tissue building stuff, while globular proteins provide the necessary micro-environment for the chemical reactions of the cellular system (enzymes), serve as information relayers to and from the nervous system (neurotransmitters), or disclose intruder microorganisms and other alien potentially malignant bodies (the proteins of the immune system). All these diverse functions are associated with the ability of a protein to maintain its unique three-dimensional structure, the so-called *native structure*, under physiological conditions. This structure is, in turn, unequivocally determined by the amino-acid sequence. The knowledge of the native structure is a necessary condition to learn about the physiological role and the mechanism of the action of a protein.

X-ray crystallography, NMR spectroscopy, and other experimental methods of structure solving provide only about several hundred new structures a year, while in the same period of time ten thousands of new amino-acid sequences are revealed. The design of a reliable method of protein-structure prediction is therefore of vital importance. There are essentially two classes of approaches to this problem: the homology methods and the methods based on energetic criteria. In the first case, the unknown structure is constructed based on known structural motifs whose amino-acid sequences are similar to the sequence studied, taking advantage of empirical relationship between sequence and the 3D structure [1–5]. The methods of the second group are based on the *thermodynamic hypothesis* formulated by Anfinsen and co-workers [6], according to which the native structure of a protein is the global minimum of its free energy under given conditions.

An early criticism of the thermodynamic hypothesis was based on the fact that the time required to carry out a systematic search of the conformational space of even a small protein would be comparable to the age of the Earth, while in reality proteins fold in seconds or in minutes. This time-scale discrepancy is known as the Levinthal paradox [7]. Studies towards solving the Levinthal paradox were undertaken by Šali and coworkers [8], Wolynes and coworkers [9], as well as Hao & Scheraga [10, 11]. Based on simple polypeptide-chain and interaction-scheme models, these authors found that if the native structure is separated from the non-native structures by a sufficiently large energy barrier, the simulated folding process is very fast. Otherwise, the lowest-energy structure might never be reached. This behavior can be explained by substantial narrowing down of the number of conformational states with lowering the energy in the first case. Thus, after low-energy structures have been reached, the system needs to sample only a comparatively small number of states. If, however, the energy spectrum is quasi-continuous in the whole range, the number of conformational states does not narrow down significantly upon energy lowering and the system becomes frustrated. In view of this, Anfinsen's thermodynamic hypothesis should be formulated more rigorously: the native structure not only is the global minimum of protein's free-energy, but is also separated by a sufficiently large energy barrier from the non-native structures.* In conclusion, there are foldable and non-foldable amino-acid sequences and the former were chosen during the evolution as components of living organisms.

In order to make use of the thermodynamic hypothesis, reliable energy functions and efficient global-optimization methods are required to reproduce protein free energy surface and search its conformational space, respectively. These problems are strongly interrelated, because only with an efficient global-optimization method assessment can

*It should be kept in mind that because a protein molecule is subject to continuous fluctuations, the native structure should be regarded as a family of conformations very similar in geometry, rather than as a single well-defined conformation. Thus, the term *global minimum of protein's energy surface* should be regarded as an abbreviation for the "deepest depression" in energy surface that contains many small pits corresponding to interrelated conformations constituting the native structure.

be made, whether an energy function is a folding potential for an amino-acid sequence known experimentally as foldable and whether it leads to the experimentally known native structure.

Despite the progress made in the recent years in the design of the global-optimization methods [12–24], it is still beyond reach to search the global energy minimum of a protein at the all-atom resolution. Therefore, simplified representations of the polypeptide chain, in which each amino-acid residue is modeled by a few interaction sites, each of which comprises several real atoms (the so-called *united-residue* representations*) receive great attention, since the pioneering works of Levitt & Warshell [25] and Levitt [26]. Following these works, a considerable number of united residue force fields were designed [9, 27–63]. After the global energy minimum has been found for the simplified chain, it can be converted to the all-atom chain, and limited exploration of the conformational space of the all-atom chain can then be carried out in order to locate the global minimum in the all-atom representation [38, 39, 41, 59, 60]. Such a protocol has recently been developed and implemented with considerable success by Koliński, Skolnick and co-workers [38, 39, 41] in predicting the three-dimensional structures of model monomeric helical proteins crambin (which also contains a β-sheet section) [41], and the dimeric GCN4 leucine zipper [42, 45] and Liwo *et al.* [59, 60], who succeeded in predicting the three-dimensional structure of the avian pancreatic polypeptide.

There are two ways to explore the conformational space of polypeptide chains with the use of a united-residue potential: the on-lattice and the off-lattice approach. In the first case, the polypeptide chain is superposed on a discrete lattice, and the number of possible conformations is, therefore, finite. In the simplest approach, the interaction potential is reduced to a set of residue-residue contact free energies [30–34, 52–54]. The rationale for such an approach was based on the assumption that side-chain packing is the principal driving force in protein folding [64]; more recent studies, however, have shown that this assumption is probably not true [36]. The recent approach developed in Skolnick's group incorporates many different interactions that can be responsible for protein folding [64]: side-chain packing, local interactions, hydrogen bonding, surface energy, and cooperativity in side-chain packing and hydrogen bonding [37–40]. The parameters of the potentials for on-lattice simulations were determined from a statistical analysis of the distributions of interacting sites obtained from the crystal data of known proteins collected in the Brookhaven Protein Data Bank (PDB) [65].

Work on the off-lattice united-residue potentials was initiated even earlier than on the on-lattice ones [25–29, 47–51, 54–63]. These potentials have also been used with considerable success to predict the three-dimensional structure of known proteins [50, 51, 56–58, 60, 62]. In contrast to the on-lattice potentials, they are functions of continuous variables. Therefore, the off-lattice approach to protein folding enables using many powerful techniques for global-search minimization that require not only the energy, but also its spatial derivatives, e.g. Monte Carlo with Minimization (MCM) [16, 17], the Diffusion Equation Method (DEM) [19, 20], the Self Consistent Mean Torsional Field (SCMTF) [21], or the shift method [24]. In the last years we have undertaken the work on the design of a united-residue potential of this class and the associated procedure of the prediction of protein structure from sequence. This work is summarized in the present article.

The rest of this paper organized as follows. First, the united-residue representation de-

*As pointed out by one of the referees, the term *united residue* could misleadingly suggest that each amino-acid residue is represented by *one* interaction site, which is not always the case, and therefore the term *united atom* could be more appropriate. However, the literature usage of the latter term always means a non-hydrogen atom fused with the attached hydrogens, while the term *united residue* is used specifically to denote coarse-grained models of polypeptide chains. Therefore, in order to maintain consistency with the literature, we keep the wording *united residue* throughout this paper.

veloped in our group and the associated energy function are described and discussed. Then, the procedure of parameterization of the force field is described. Next, method used for the conversion of the simplified chain into an all-atom chain is presented. Finally, results of the application of the force field and whole procedure are described, including the inverse- and *de novo* folding tests.

REPRESENTATION OF POLYPEPTIDE CHAINS AND INTERACTION SCHEME

In our model [60, 61], 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. 1). 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. In the current version of the force field, we allow, however, for variation of the side-chain positions with respect to the backbone (α_{SC} and β_{SC}), and for the variation of the virtual-bond angles θ; in our earlier approach [59, 60] they were assumed fixed at the value of 90° (the most probable value as found by the analysis of the PDB [66]) and the average side-chain geometry relative to the three adjacent C^αs, as found by Levitt [26].

The energy of the virtual-bond chain is expressed by Eqn. (1).

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

where $U_{SC_iSC_j}$, $U_{SC_ip_j}$, and $U_{p_ip_j}$ denote the energies of the interactions between side

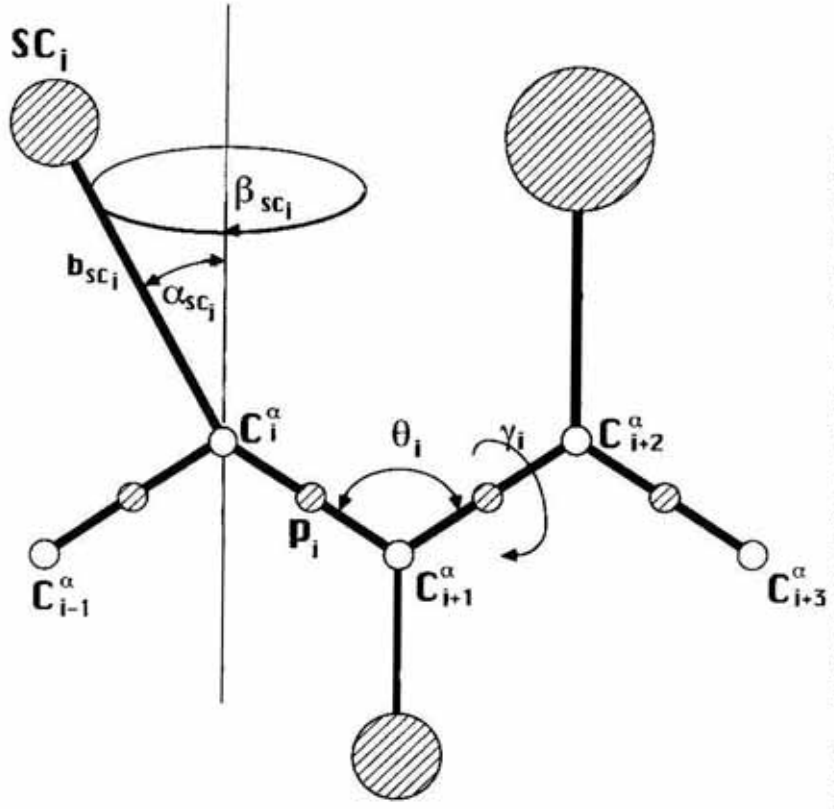

Figure 1. United-residue representation of a polypeptide chain.

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

chains, between side chains and peptide groups, and between peptide groups, respectively, $U_{tor}(\gamma_i)$ denotes the energy of variation of the virtual-bond dihedral angle γ_i, $U_b(\theta_i)$ denotes the "bending" energy of the virtual-bond angle θ_i, $U_{rot}(\alpha_{SC_i}, \beta_{SC_i})$ is the local energy of side chain i, U_{corr} includes cooperative terms (e.g. the four body interactions considered by Koliński, Skolnick and coworkers [38]), and the ω's denote relative weights of the respective energy terms.

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

$$U_{ij} = 4\left[\varepsilon_{ij}\left|x_{ij}^{12} - \varepsilon_{ij}x_{ij}^{6}\right.\right] \tag{2}$$

where ε_{ij} is the pair-specific van der Waals well-depth; $\varepsilon > 0$ corresponds to hydrophobic-hydrophobic-type and $\varepsilon < 0$ to hydrophobic-hydrophilic and hydrophilic-hydrophilic-type interactions (see Fig. 2 for illustration). The quantity x_{ij} is the reciprocal of the reduced distance between side chains; it can depend on their distance alone for radial-only potential (in this case $x_{ij} = \sigma_{ij}^{o} / r_{ij}$, r_{ij} being the distance between the side chains and σ_{ij}^{o} a pair-specific constant that depends on the types of side chains i and j) or on both distance and orientation; the same applies to ε_{ij}. In our work we have considered and parameterized functional form of both types. The functional forms of ε and x can be found in the original paper [61].

The peptide-group interaction potential ($U_{p_ip_j}$) accounts mainly for the electrostatic interactions between them or, in other words, for their tendency to form backbone hydrogen bonds. In contrast to $U_{SC_iSC_j}$, its functional form was derived rigorously by averaging the simplified electrostatic-interaction energy of the peptide groups over the angles λ of their rotation about the corresponding C^{α}–C^{α} virtual bond axes, assuming that each peptide group is modeled by a point dipole located in the middle of the virtual bond, as was proposed by Piela & Scheraga [15] (Fig. 3). The potential is expressed by Eqn. (3); the details of the derivation and parameterization can be found in the papers cited [59, 60]

$$\begin{aligned} U_{p_ip_j} = {} & \frac{A_{p_ip_j}}{r_{ij}^{3}}(\cos\alpha_{ij} - \\ & - 3\cos\beta_{ij}\cos\gamma_{ij}) - \\ & - \frac{B_{p_ip_j}}{r_{ij}^{6}}[4 + (\cos\alpha_{ij} - \\ & - 3\cos\beta_{ij}\cos\gamma_{ij})^2 - \\ & - 3(\cos^2\beta_{ij} + \cos^2\gamma_{ij})] + \\ & + \varepsilon_{p_ip_j}[(\frac{r^{o}_{p_ip_j}}{r_{ij}})^{12} - \\ & - 2(\frac{r^{o}_{p_ip_j}}{r_{ij}})^{6}] \end{aligned} \tag{3}$$

with

$$\cos\alpha_{ij} = \mathbf{v_i} \bullet \mathbf{v_j}$$
$$\cos\beta_{ij} = \mathbf{v_i} \bullet \mathbf{e}_{\mathbf{r}_{ij}}$$
$$\cos\gamma_{ij} = \mathbf{v_j} \bullet \mathbf{e}_{\mathbf{r}_{ij}}$$

where $A_{p_ip_j}$, $B_{p_ip_j}$, and $\varepsilon_{p_ip_j}$ are constants characteristic of the kind of interacting peptide groups, r_{ij} is the distance between the peptide-group centers, $\mathbf{v_i}$ is the unit vector pointing from C_i^{α} to C_{i+1}^{α}, and $\mathbf{e}_{\mathbf{r}_{ij}}$ is the unit vector pointing from p_i to p_j (see Fig. 3 for illustration). Two types of peptide groups were distinguished: ordinary and proline; the second one can act as hydrogen-bond acceptor only and also comprises all N-methylated amino-acid residues (e.g. sarcosine). This gives a total of three sets of constants in Eqn. (3). The angular part of Eqn. (3) favors parallel and near-parallel orientation of the virtual C^{α}–C^{α} bonds, as

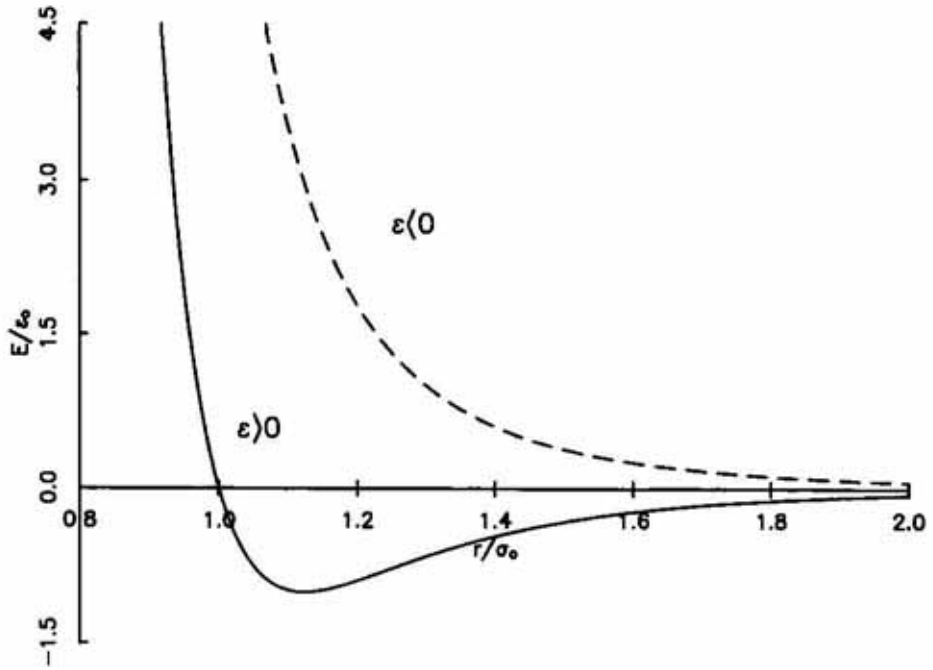

Figure 2. Plot of the energy of interaction between the hydrophobic ($\varepsilon > 0$) and hydrophilic ($\varepsilon < 0$) side chains (Eqn. (2)) in their reduced distance.

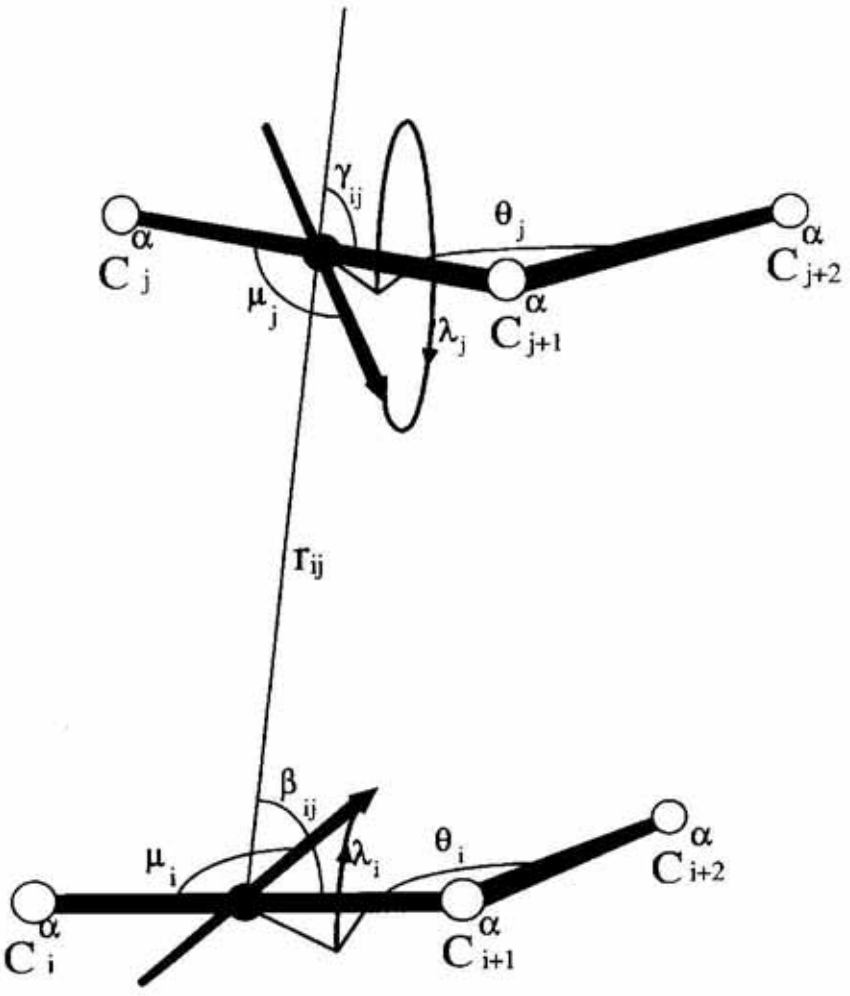

Figure 3. The relative orientation of the virtual bonds $C_i^\alpha - C_{i+1}^\alpha$ and $C_j^\alpha - C_{j+1}^\alpha$ is described by the angles α_{ij}, β_{ij} and β_{ji}, defined by Eqn. (3).

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

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 Eqn. (4).

$$U_{tor}(\gamma_i) = \sum_{k=1}^{6} [a_k(\cos k\gamma i + 1) + b_k(\sin k\gamma i + 1)] \quad (4)$$

This energy reflects the local propensities of the polypeptide chain, i.e. to form the right- rather than left-handed helices and the left- 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 comprises all other amino-acid residues). Detailed analysis of local propensities of the structures contained in the PDB confirmed this division [62].

The expressions of the bending energy, U_b, and local side-chain energy, U_{rot}, have a form of the negative of the logarithms of sums of Gaussians; in the second case their centers correspond to different rotameric states of the side chains. Because these expressions are lengthy, the reader is referred to the original paper [62].

The multibody (or cooperative) term U_{corr} arises from the fact that details of all-atom chain are lost when converting it into the simplified chain. Mathematically it can be expressed as averaging the energy over some "less important" degrees of freedom, as expressed by Eqn. (5) [63]:

$$\mathcal{U}(\mathbf{x}) = \mathcal{F}^{-1}\left\{\frac{\int_{\mathbf{y}} w[E(\mathbf{x};\mathbf{y})]\mathcal{F}[E(\mathbf{x};\mathbf{y})]dV_y}{\int_{\mathbf{y}} w[E(\mathbf{x};\mathbf{y})]dV_y}\right\} \quad (5)$$

where $\mathcal{F}$ is an unequivocal monotonic function in one variable, ω is a weight for the energy, $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$ and $\mathbf{y} = [y_1, y_2, \ldots, x_n]^T$ denote the "important" and the "less important" variable set, respectively.

The form of the averaged energy $\mathcal{U}$ will depend on the choice of the transformation $\mathcal{F}$ and the weight function ω. Usually, the energy is Boltzmann-averaged [26, 60], which means that $\mathcal{F}(E) = E$ and $\omega(E) = \exp(-E/k_BT)$, T being the absolute temperature and k_B the Boltzmann constant. However, because the local and two-body terms can be identified with free energies rather than with Boltzmann-averaged energies [35–37, 56, 61, 62], it is more appropriate to choose $\mathcal{F}(E) = \exp(-E/k_BT)$ and $\omega(E) = 1$ or, alternatively, $\mathcal{F}(E) = \exp(E/k_BT)$ and $\omega(E) = \exp(-E/k_BT)$. In this case, $\mathcal{U}(\mathbf{x})$ has the meaning of the free energy associated with a set of fixed values of the important variables $\mathbf{x}$, as expressed by Eqn. (6) [63].

$$\mathcal{U}(\mathbf{x}) = F(\mathbf{x}) = -k_BT \cdot \ln\left\{\frac{1}{V_y}\int_{\mathbf{y}} \exp[-E(\mathbf{x};\mathbf{y})/k_BT]dV_y\right\} \quad (6)$$

with

$$V_y = \int_{\mathbf{y}} dV_y.$$

Expanding Eqn. (6) into a power series in $\beta = 1/k_BT$ (the so-called cumulant expansion [67]), we obtain Eqn. (7).

$$F(\mathbf{x}) = U_1 - \frac{1}{2}(U_2 - U_1^2)\beta + \frac{1}{6}(U_3 - 3U_1U_2 + 2U_1^3)\beta^2 - \frac{1}{24}(U_4 - 3U_2^2 - 4U_1U_3 + 12U_1^2U_2 - 6U_1^4)\beta^3 + \cdots \quad (7)$$

where

$$U_k = \frac{1}{V_y}\int_{\mathbf{y}} E(\mathbf{x};\mathbf{y})^k dV_y \quad (8)$$

is the kth moment of the energy about $E = 0$.

The presence of the integrals of various powers of energy gives rise to the appearance of multibody terms in the average energy of the simplified system. Consider for example, three peptide groups that are in a close contact with each other, without the presence of the solvent. The forces acting between the individual atoms of the peptide groups are, to a good approximation, pairwise. If we consider only the peptide-group centers and average the interaction energy over the *internal* degrees of freedom of each peptide group according to Eqn. (7), U_1 will still include only the interactions of the pairs of the centers. We have shown elsewhere [59, 63] that in this specific case U_1 happens to be zero, to a very good approximation. The integrals contained in the second moment of the energy, U_2, will already contain products of the interactions between two pairs of peptide groups, which in our example will share a common peptide group, which could create three-body contributions to the average energy. However, in the specific case of the peptide groups, the three-body contributions to U_2 happen to vanish, to a very good approximation [59, 63]. Nevertheless, in the case of U_3 and higher energy moments we cannot avoid considering the terms that depend on the coordinates of more than two centers, which creates multibody contributions to the energy.

Based on the above general considerations and taking advantage of our dipole model of the peptide group, we have derived analytical expressions for the multibody terms in of peptide-group interactions [63]. We found that the cooperative contribution to the averaged energy is significant, if some of the peptide groups are adjacent in the chain. A frequent case is the cooperativity of two pairs of adjacent peptide groups; they are found in α-helices and β-sheets. In this case U_{corr} is expressed by Eqn. (9).

$$U_{corr}(i, i-1; k, k\pm1) = -\frac{z^2}{4}(9\eta_{ik}\eta_{i-1,k\pm1} + 4\bar{\eta}_{ik}\bar{\eta}_{i-1,k\pm1}) \quad (9)$$

where η_{ik} is the energy of the interaction of two *aligned* peptide-group dipoles (or two hydrogen-bonded peptide groups), while $\bar{\eta}_{ik}$ is the interaction of two *antiparallel* peptide-group dipoles (or two antiparallelly stacked peptide groups); z^2 is a proportionality constant. This is illustrated in Fig. 4. It should also be noted that these four-body terms have very much in common with the terms introduced by Skolnick and coworkers on a heurestic basis [37, 38]. For the derivation and the functional forms of the multibody terms the reader is referred to the original paper* [63].

PARAMETERIZATION OF THE FORCE FIELD

The following procedures are commonly used to parameterize united-residue potentials:

1. Direct averaging [Eqn. (5)] of the all-atom potentials over the "less important" degrees of freedom that are lost when passing from the all-atom to the united residue representation of the polypeptide chain [26–29].

2. Determination of the united-residue potentials so as to reproduce the single body, pair, and possibly triplet distribution functions, as well as contact free energies determined from protein crystal data [47–49, 54–56]. This approach is based on the following assumptions:

(a) The distribution functions obtained by using a sufficiently large number of protein crystal data (each of which corresponds to a system at a free-energy minimum) are sufficiently good approximations to those of a hypothetical "stochastic" mixture. This approximation is justified by the observation that, although a crystal structure is at equilibrium as the whole structure, its individual parts can be forced to assume geometries far from locally equilibrated, locally lower-energy conformations having, however, higher probability of occurrence in the whole structure [68]. For example, the distributions of X–H bond lengths obtained from large data bases of crystal structures are qualitatively

*The paper is available at http://chemik.chem.univ.gda.pl:8000/ECCC3/19/poster.html

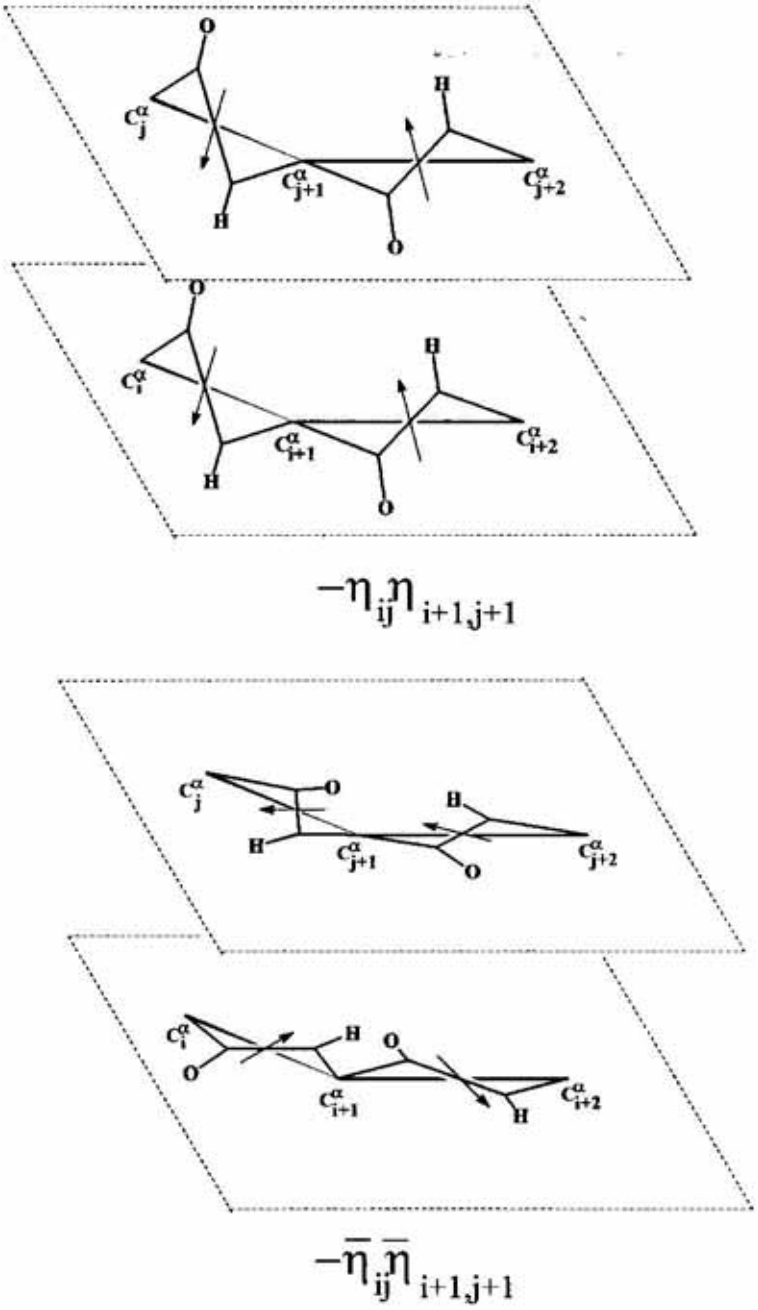

Figure 4. Graphical representation of the two components of the four-body-interaction energy of Eqn. (9).

similar to those calculated from potential-energy surfaces of proton transfer [68].

(b) Interactions can be described with sufficient accuracy by using the potential of mean force, $W(\mathbf{X})$, $\mathbf{X}$ denoting the degrees of freedom of the considered subsystem, which are related directly to the corresponding distribution functions, $\rho(\mathbf{X})$:

$$\rho(\mathbf{X}) = \rho_o(\mathbf{X})\exp[-\beta W(\mathbf{X})] \qquad (10)$$

where $\rho_0(\mathbf{X})$ is a known reference distribution function (e.g. the distribution function of non-interacting side chains tethered to the backbone).

3. A combination of the two preceding approaches in which some part of the potential is determined by direct averaging of the all-atom potential, e.g. the local and hydrogen-bonding interactions, and some estimated from protein crystal data. Such a division is motivated by the fact that, if direct averaging is computationally feasible, as in the case of the local and hydrogen-bonding interactions, the resulting potential will always be more accurate than that calculated from experimental distribution functions, whose accu-

racy is severely limited by the sparse number of protein crystal data. Conversely, obtaining the hydrophobic potential by direct averaging is in most cases not feasible, owing to the large number of degrees of freedom over which averaging must be carried out (i.e. the dihedral angles χ for each side chain) and possibly to the necessity of including explicit water molecules in the averaging.

4. Determination of the parameters of the potential so as to locate the native structures as global minima for a set of training proteins and, simultaneously, introducing a large energy gap between the near-native and non-native structures. For on-lattice simulations, such an approach based on spin-glass theory was developed by Wolynes and co-workers [9] and by Hao & Scheraga [10, 11]. This can be formulated in terms of the optimization of the so-called Z-score (Eqn. (11)):

$$Z = \frac{E_o - 1/N\sum_{i=1}^{N} E_i}{\sqrt{1/N\sum_{i=1}^{N} E_i^2 - 1/N^2(\sum_{i=1}^{N} E_i)^2}} \quad (11)$$

where N is the number of conformations, E_o is the energy of the native conformation, and E_i is the energy of the i-th non-native conformation.

The value of the Z-score is the normalized difference between the energy of the native conformation and the mean energy of the quasi-continuous energy distribution corresponding to non-native structures. The more negative the Z-score values, the more the native structure is distinguished from non-native ones. A similar method was developed for off-lattice simulations by Crippen and co-workers [50, 51].

Both in the first [59, 60] and in the second [61–63] generation of the force field we have implemented procedure 3 to determine the parameters of individual energy terms (the U 's). In the first version, the geometric and interaction parameters of the side chains were obtained based on the work of Levitt [26] and Miyazawa & Jernigan [31]. As mentioned in the preceding section, the peptide-group interaction potential, $U_{p_ip_j}$, was developed by averaging the simplified expression for the interaction energy of the peptide groups over the angles of the rotation of the peptide groups about the virtual-bond axis C^α–C^α [59, 60]. This energy function, as well as the virtual-bond torsional energy, was parameterized through averaging of the all-atom ECEPP/2 [69, 70] potential [59, 60]. Because rigid virtual-valence geometry was assumed, the terms U_b and U_{rot} were absent. The correlation term, U_{corr} was not considered either. Finally, the relative weights of the energy terms were estimated so as to achieve compatibility of some energies estimated from different parameterization procedures (e.g. the contact energy of the interaction of two glycine residues calculated from the PDB by Miyazawa & Jernigan [31] can be considered as the energy of the interaction of two peptide groups and should therefore be equal to the average energy of peptide-group interaction obtained by averaging the ECEPP/2 all-atom potential [60, 61]).

In the second generation of the force field, the side-chain [61] and local-interaction [62] terms were parameterized based on pair-correlation and distribution functions calculated from 195 non-homologous high-resolution protein structures from the PDB. The electrostatic term ($U_{p_ip_j}$) was inherited in unchanged form from the preceding version of the force field. Further addition was the determination of the weights of the energy terms so as to optimize the Z-score of the phosphocarrier protein from *Streptococcus faecalis* (1PTF; 87 amino-acid residues; 1.6 Å resolution) by means of inverse-folding calculations [62]. Its structure is shown in Fig. 5. The procedure of weight determination was as follows [62]:

- 1. A database of contiguous C^α traces was selected from the PDB (a total of 502).
- 2. 1PTF sequence was superposed on a series of randomly chosen C^α traces (about 500) and on the native pattern of 1PTF.
- 3. For each pattern, the geometry was regularized by performing a series of energy minimizations with C^α distance constraints with gradually diminished weights of the distance constraints.
- 4. The energy of the virtual chain was minimized with no distance constraints and the resulting energy was considered

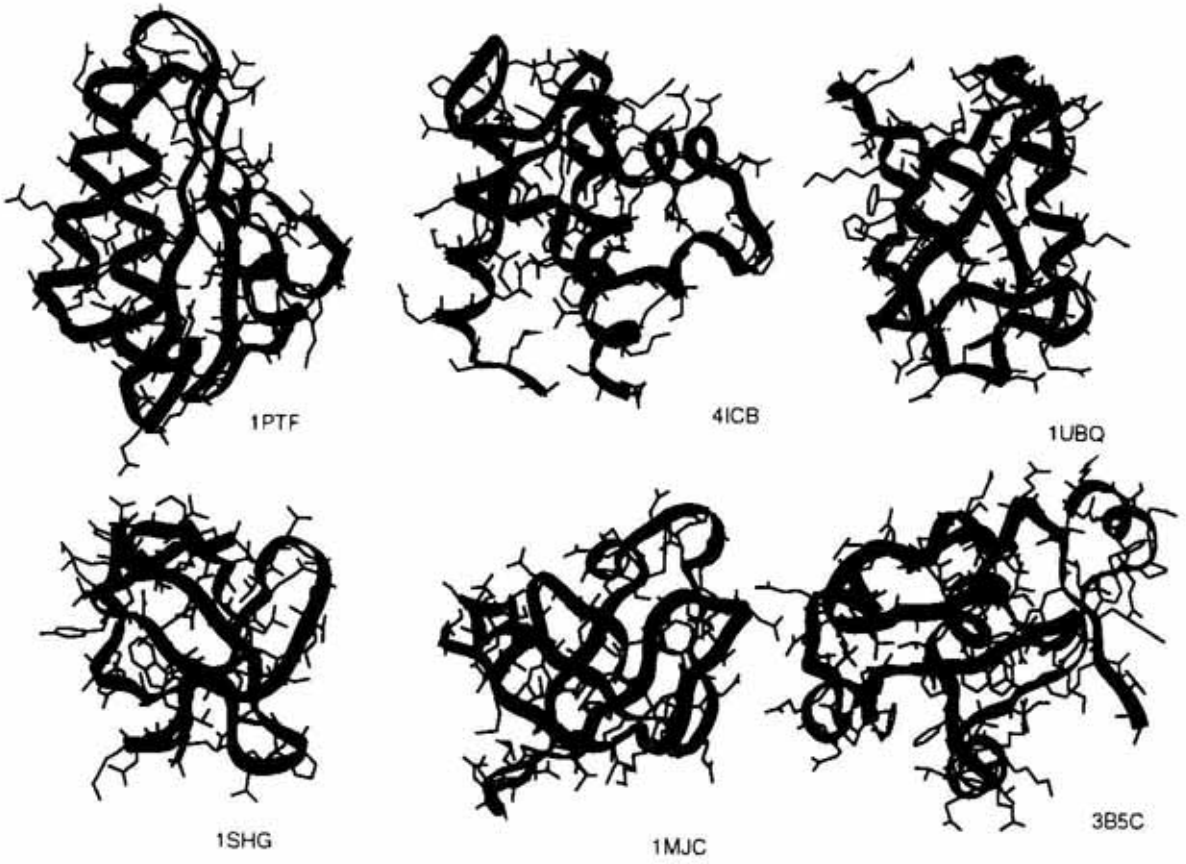

Figure 5. The structures of the proteins used in inverse-folding calculations: histidine-containing phosphocarrier protein (1PTF), calcium-binding protein (4ICB), ubiquitin (1UBQ), α-spectrin (1SHG), major cold-shock protein (1MJC), and cytochrome b5c (3B5C).

Table 1. Iterative determination of self-consistent weights of energy terms for 1PTF (data from [62])

Iter.	w_e^{oa}	w_l^{oa}	w_t^{oa}	ΔE_{nat}^b (kcal/mol)	Z	w_e^{*a}	w_l^{*a}	w_t^{*a}	ΔE_{nat}^b (kcal/mol)	Z
0	1.450	0.626	1.692	+20.0	−1.88	0.249	0.050	0.271	−25.8	−4.67
1	0.610	0.186	0.846	−16.8	−4.26	0.341	0.112	0.403	−21.0	−4.66
2	0.341	0.117	0.403	−23.5	−4.41	0.650	0.190	0.00	−26.4	−4.74
3	0.495	0.150	0.201	−14.4	−3.80	0.347	0.201	0.091	−12.7	−3.87
4	0.421	0.175	0.147	−13.4	−3.78	0.444	0.144	0.042	−14.0	−3.80

[a] w^o, w^* denote the initial weights used to carry out the threading-with-minimization calculations of a given iteration and the final weights optimized in this iteration, respectively. In order to avoid oscillations, the initial weights of the next iteration were arithmetic means of the weights optimized in two preceding iterations. [b] $\Delta E_{nat} = E_{nat} - \min_{i \neq nat}\{E_i\}$, where the latter term is the minimal element in the set of energies of all non-native structures.

as the final energy corresponding to the pattern.

- 5. The Z-score was calculated (Eqn. (11)) and then minimized as a function of energy-term weights.
- 6. The procedure was iterated from step 2 with new weights, until the weights calculated in two consecutive iterations were reasonably consistent.

The history of the determination of energy-term weights is summarized in Table 1. As shown, although the initial weights did not produce an energy function localizing the native structure of 1PTF as the lowest-energy structures, a folding potential was obtained already in the first iteration.*

SEARCH OF THE CONFORMATIONAL SPACE OF THE SIMPLIFIED AND OF THE ALL-ATOM CHAIN

In order to search the conformational space of the simplified chain, the Monte Carlo with Minimization Method of Li & Scheraga [16, 17] was implemented. In brief, the method consists of the following steps:

- 1. Choose an arbitrary starting conformation.
- 2. Minimize the energy; let the geometric parameters of the resulting conformation be contained in the vector Γ_0 and the corresponding energy as U_0.
- 3. Perturb Γ_0 according to a predetermined scheme.
- 4. Carry out energy minimization, obtaining the conformation Γ_1 and energy U_1.
- 5. If neither U_1 nor Γ_1 differs by more than a preassigned cut-off from the previous conformations, discard it and repeat the process beginning at step 3; otherwise apply a Metropolis test [71] in order to accept or reject the conformation.
- 6. If the new conformation is accepted, substitute Γ_1 for Γ_0, and U_1 for U_0, and repeat from step 3.
- 7. Iterate steps 3–7, until the requested number of accepted conformations is obtained.

In order to search the conformational space of the chains with all-atom backbone, the more efficient Electrostatically Driven Monte Carlo Method (EDMC) [18] was applied. This method is a modified version of MCM, in which a part of perturbations (step 3) is done so as to align the worst oriented peptide groups in the electrostatic field of the remaining part of the polypeptide chain. All-atom calculations were carried out using the ECEPP/2 [69, 70, 72] and subsequently the ECEPP/3 [73] force field, supplemented with the SRFOPT (surface) model of hydration [74].

*The parameters of the second-generation force field are available at http://chemik.chem.univ.gda.pl:8000/local/docs/adam; see the file readme.txt.

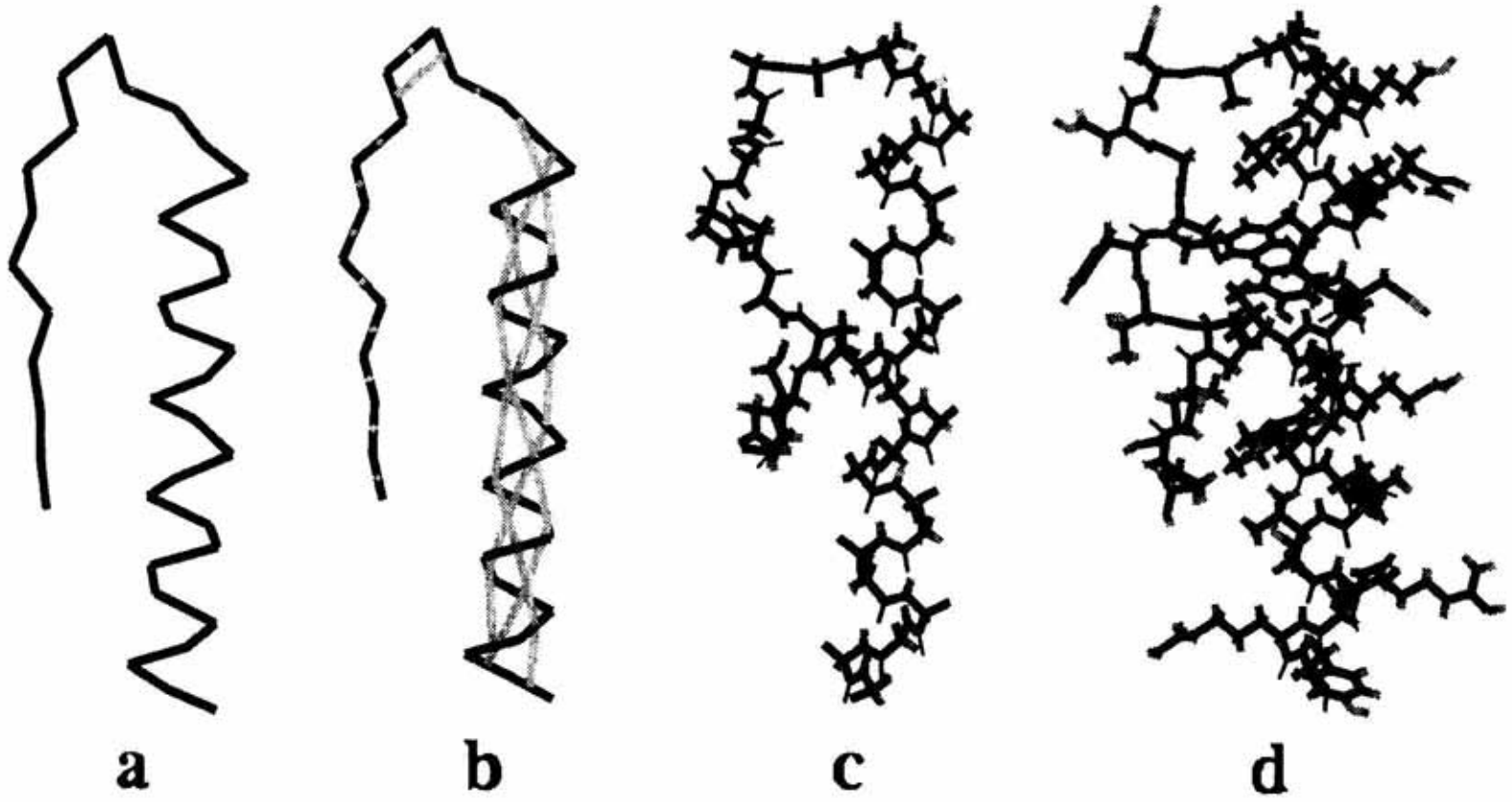

Figure 6. Illustration of the conversion of the lowest-energy united-residue structure of the avian polypeptide (APP) to the all-atom structure: a) the C^{α} trace; b) C^{α} trace with indicated dipole paths; it should be noted that there are three paths, each one connecting every third peptide groups of the α-helical part running from residue 13 to 36; c) all-atom backbone constructed based on peptide-group dipole orientation followed by perturbation of the structure by means of the EDMC method [18] to find a relaxed structure; d) side chains attached with subsequent EDMC perturbation to give the all-atom structure.

RECONSTRUCTION OF THE ALL-ATOM CHAIN FROM THE SIMPLIFIED CHAIN

The method consists of two stages: i) positioning the peptide groups between the α-carbons given the α-carbon trace, which gives the complete all-atom backbone and ii) positioning the side chains for on the all-atom backbone [59, 60].

The approach used to solve problem i) was based on the observation that the peptide groups in protein structures tend to form an extensive hydrogen-bond net work, which can also be expressed in terms of their tendency to optimal electrostatic interactions [15]. Taking advantage of the dipole model of the peptide group [15, 59] the problem of positioning the peptide groups given the C^{α} trace can be formulated as the problem of optimal alignment of peptide-group dipoles. In regular helical or sheet structures, lines of aligned peptide-group dipoles or *dipole paths* can be distinguished, which comprise all peptide groups in regular structures (e.g. each α-helix contains three paths linking every third peptide group). We therefore proposed the following algorithm for positioning of the peptide groups [59], which we named the *dipole path* method:

- 1. Find the chains of non-contiguous peptide groups, such that the average energy of the electrostatic interactions of the neighboring peptide groups is below a chosen cut-off limit and the peptide groups lie on a line with a low curvature.
- 2. Align the peptide-group dipoles along each path. Because alignment can be carried out in two directions, the correct direction is chosen 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 the "isolated" peptide groups in the electrostatic field of the determined dipole paths.
- 4. Based on the position of peptide-group dipoles, calculate the coordinates of all atoms in the peptide groups. The side chains are still represented by single interaction sites.
- 5. Complete the determination of backbone geometry by carrying out short MCM runs (or even just a single energy minimization) with C^{α} distance constraints corresponding to the parent united-residue structure.

Once the all-atom backbone has been obtained, structure generation is completed by putting on all-atom side chains in such a manner that collisions between the atoms are avoided and carrying out EDMC simulations with and subsequently without the C^{α} distance constraints [60]. The process of all-atom chain building is illustrated in Fig. 6

RESULTS OBTAINED WITH THE FIRST-GENERATION FORCE FIELD

The first test of the united-residue force field and the whole protocol for protein structure prediction was the avian pancreatic polypeptide (APP) — a small 36-residue protein. Its crystal structure was determined by Blundel *et al.* [75] at 1.4 Å and then by Glover *et al.* [76] at the 0.98 Å resolution. The structure consists of a polyproline-like or collagen-like helix running from residue 1 to 8, packed against the hydrophobic face of an α-helix which extends from residue 13 to 31. The C terminus does not participate in the α-helix. Although the molecule forms a dimer both in solution [77, 78] and in the crystal phase [75, 76], there is some evidence that even the monomer is sufficiently stabilized by hydrophobic interactions between the two domains to retain the X-ray conformation [77, 78, 75].

The lowest-energy structure of APP obtained with the united-residue force field had an r.m.s. deviation from the C^{α}-trace of the 1PPT structure of 3.8 Å [60]. After applying the conversion procedure, the r.m.s. deviation increased to 4.1 Å; however the structure still resembled the native structure (see Fig. 7 for superposition of both structures). This means that the quality of the united-residue force field is critical for the success of the procedure; perturbing the converted structure using the all-atom force field does not bring it closer to the native structure. This feature of the hierarchical protocols for pro-

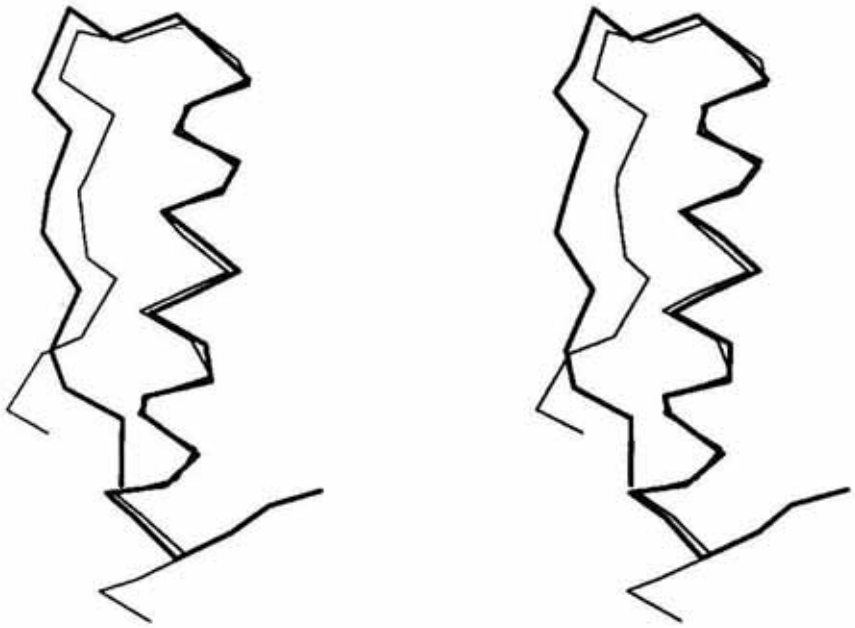

Figure 7. The α-carbon trace of the lowest-energy all-atom conformation of APP (thin lines) superposed on the 1PPT crystal structure (heavy lines). Residues 13–33 were used in the superposition.

tein-structure prediction was also observed by Skolnick and coworkers [37, 38].

The procedure was subsequently applied in *de novo* prediction of the structure of the 29-residue neuropeptide galanin isolated from a number of mammal species [79]. No X-ray structure is available for this peptide. Our calculations had suggested that even in aqueous solution this peptide should contain a considerable fraction of helical structure [80]; this was later confirmed by NMR measurements [81].

The performance of the first-generation force field was, however, significantly poorer with more complicated structural motifs that included β-sheets, such as the epidermal growth factor (EGF), ferredoxin, and crambin, although it was able to distinguish the native-like structures from alternative structures as low-energy ones. The resulting best structures were, however, too much distorted with respect to the native structures and resembled molten globules rather than the native structures, though they essentially had the native-like packing of the side chains (Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S. & Scheraga, H.A., unpublished results). This could be caused both by the as-

Table 2. Summary of control threading experiments using all test proteins with weights determined using the phosphocarrier protein (1PTF); initial weights of iteration 4 of Table 1 (data from [62])

Protein[a]	N[b]	Type	Cofactor	Initial[c]		Optimized[d]		r.m.s.[e]
				ΔE_{nat} (kcal/mol)	Z-score	ΔE_{nat} (kcal/mol)	Z-score	(Å)
4ICB	76	α	Ca^{2+}	−24.6	−4.84	−27.6	−5.08	4.5
1UBQ	76	β + α	none	−15.5	−3.29	−10.1	−3.68	3.1
3B5C	85	α + β	heme, Fe^{2+}	−13.5	−3.56	−14.8	−3.99	3.1
1SHG	57	β	none	−5.2	−3.06	−5.2	−3.78	2.5
1MJC	69	β	none	−3.9	−3.40	−7.8	−3.82	2.5

[a]See Fig. 5 for the names of these proteins. [b]The number of amino-acid residues. [c]Values calculated using the final weights obtained for the phosphocarrier protein (1PTF) (w^a of iteration 4 from Table 1). [d]Values calculated using the weights optimized using the energies obtained in threading-with-minimization calculations for a given protein. [e]r.m.s. deviation from the native structure.

sumption of rigid virtual-valence geometry and the absence of multibody interactions in the force field.

PRELIMINARY RESULTS OBTAINED WITH THE SECOND-GENERATION FORCE FIELD

Using the weights determined from the inverse-folding calculations on the phosphocarrier protein (1PTF), we checked the ability of the potential to locate the native structures of other proteins correctly, using the inverse-folding approach [62]. In these calculations, the force field did not include the correlation term U_{corr}. Table 2 summarizes the results of these calculations for a number of monomeric proteins of length exceeding 50 amino-acid residues and the structures are shown in Fig. 5. As shown, in each case the native structure is the lowest in energy and is separated from non-native structures by a significant energy gap. The Table also includes weights optimized specifically for each protein after the appropriate series of threading with-minimization calculations (cf. Section "Parameterization of the force field"). As shown, optimization of weights in this manner did not result in major decreases of the Z-score values or the energy differences between the native and lowest-energy non-native conformation, which means that weights obtained by determining the weights of the energy terms using the phosphocarrier protein (1PTF) are also relevant for other proteins. It should be noted that none of the above proteins was used in parameterization of the potential.

Use of energy minimization in our threading calculations gives a possibility that the procedure will find a structure close to the native pattern of the target protein, even if the structural fragments from the data base are distant from its native structure. Table 3 summarizes the results of threading-with-minimization calculations for the 10–58 fragment of the B-domain of staphylococcal protein A. The native pattern of protein A was not present in the data base. As shown, all but one of the five lowest-energy patterns found in the data base are close to the native structure of protein A. Structures 4 and 5 were moved by energy minimization very far apart from the starting PDB structure, but they had become close to the native structure of protein A.

DIRECTIONS OF FURTHER WORK

Although the energy-term weights determined using the inverse-folding calculations of 1PTF proved to be able to locate the native structures of other proteins as the lowest-energy ones, it is clear that for *de novo* folding this is insufficient to obtain reliable energy-term weights. The structures from the PDB

Table 3. Summary of the results of the calculations on the 10–58 fragment of staphylococcal protein A (data from [63])

Structure[a]	Start[b]	Energy (kcal/mol)	r.m.s.p[c] (Å)	r.m.s.n[d] (Å)	%nc[e]
1BAB:D	23	−146.5	4.4	3.8	55
1CSC	144	−143.7	5.1	5.8	43
1ECA	2	−143.6	9.1	9.5	43
2HMZ:C	48	−143.3	7.4	4.1	48
1CPC:B	1	−141.7	9.5	4.4	52
Native	10	−149.7	2.5	2.5	89

[a]4-digit code of the PDB entry followed by the chain code, if applicable; [b]the residue of the PDB structure onto which the first residue of protein A was superposed; [c]r.m.s. deviation of the energy-minimized structure from the original PDB pattern; [d]r.m.s. deviation from the NMR structure of protein A; [e]percentage of native contacts.

usually correspond to favorable local interactions and also consist of regular patterns. It can therefore be supposed that while they can serve to determine the relative weights of the hydrophobic and electrostatic term, the weights of the local and correlation term will be estimated poorly. At this moment, we are working on extending the determination of the energy-term weights (Eqn. (1)) on the structures outside the PDB, applying the entropic-sampling [82] and other efficient methods to search the energy space.

For the force field to be used efficiently in *de novo* folding, an efficient method of global optimization must be applied. We are now working on the application of the combination of the methods of deformation of the energy surface, such as the diffusion-equation and the shift method [19, 20, 24] with Monte Carlo methods. Such an approach has resulted in an efficient methods for finding the global energy minima of crystal structures of small molecules [83].

REFERENCES

1. Jones, T.A. & Thirup, S. (1991) Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822.
2. Clark, D.A., Shirazi, J. & Rawlings, C.J. (1991) Protein topology prediction through constraint-based approach and the evaluation of topological folding rules. *Protein Eng.* **7**, 751–760.
3. Rooman, M.J. & Wodak, S.J. (1992) Extracting information on folding from the amino acid sequence: Consensus regions with preferred configuration in homologous proteins. *Biochemistry* **31**, 10239–10249.
4. Johnson, M.S., Overington, J.P. & Blundel, T.L. (1993) Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.* **231**, 735–752.
5. Fischer, D., Rice, D., Bowie, J.U. & Eisenberg, D. (1996) Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J.* **10**, 126–136.
6. Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. *Science* **181**, 223–230.
7. Levinthal, C. (1968) Are there pathways for protein folding? *J. Chem. Phys.* **65**, 44–45.
8. Šali, A., Shaknovich, E.I. & Karplus, M. (1994) How does a protein fold? *Nature* **369** 248–251.
9. Goldstein, R.A., Luthey-Schulten, Z.A. & Wolynes, P.G. (1992) Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9029–9033.
10. Hao, M.-H. & Scheraga, H.A. (1994) Statistical thermodynamics of protein folding: Sequence dependence. *J. Phys. Chem.* **98**, 9882–9893.
11. Hao, M.-H. & Scheraga, H.A. (1996) How optimization of potential functions affects protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 4984–4989.
12. Scheraga, H.A. (1992) Some approaches to the multiple-minima problem in the calculation of polypeptide and protein structures. *Int. J. Quant. Chem.* **42**, 1529–1536.
13. Vásquez, M., Némethy, G. & Scheraga, H.A. (1994) Conformational energy calculations on polypeptides and proteins. *Chem. Rev.* **94**, 2183–2239.
14. Scheraga, H.A. (1996) Recent developments in the theory of protein folding: Searching for the global energy minimum. *Biophys. Chem.* **59**, 329–339.
15. Piela, L. & Scheraga, H.A. (1987) On the multiple-minima problem in the conformational analysis of polypeptides. I. Backbone degrees of freedom for a perturbed α-helix. *Biopolymers* **26**, S33–S58.
16. Li, Z. & Scheraga, H.A. (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6611–6615.
17. Li, Z. & Scheraga, H.A. (1988) Structure and free energy of complex thermodynamic systems. *J. Mol. Struct. (Theochem.)* **179**, 333–352.

18. Ripoll, D.R. & Scheraga, H.A. (1988) On the multiple-minima problem in the conformational analysis of polypeptides. II. An electrostatically driven Monte Carlo method — tests on poly(L-alanine). *Biopolymers* **27**, 1283–1303.

19. Piela, L., Kostrowicki, J. & Scheraga, H.A. (1989) The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method. *J. Phys. Chem.* **93**, 3339–3346.

20. Kostrowicki, J. & Scheraga, H.A. (1992) Application of the diffusion equation method for global optimization to oligopeptides. *J. Phys. Chem.* **96**, 7442–7449.

21. Olszewski, K.A., Piela, L. & Scheraga, H.A. (1992) Mean field theory as a tool for intramolecular conformational optimization. 1. Tests on terminally-blocked alanine and Met-enkephalin. *J. Chem. Phys.* **96**, 4672–4676.

22. Oresić, M. & Shalloway, D. (1994) Hierarchical characterization of energy land scapes using Gaussian packet states. *J. Chem. Phys.* **101**, 9844–9857.

23. Amara, P., Hsu, D. & Straub, J.E. (1993) Global energy minimum search using an approximate solution of the imaginary time Schrödinger equation. *J. Phys. Chem.* **97**, 6715–6721.

24. Pillardy, J., Olszewski, K.A. & Piela, L. (1992) Theoretically predicted lowest-energy structures of water clusters. *J. Mol. Struct.* **270**, 277–285.

25. Levitt, M. & Warshell, A. (1975) Computer simulation of protein folding. *Nature* **253**, 694–698.

26. Levitt, M. (1976) A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

27. Pincus, M.R. & Scheraga, H.A. (1977) An approximate treatment of long-range interactions in proteins. *J. Phys. Chem.* **81**, 1579–1583.

28. Gerber, P.R. (1992) Peptide mechanics: A force field for peptides and proteins working with entire residues as smallest units. *Biopolymers* **32**, 1003–1017.

29. Wallqvist, A. & Ullner, M. (1994) A simplified amino acid potential for use in structure predictions of proteins. *Proteins* **18**, 267–280.

30. Tanaka, S. & Scheraga, H.A. (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structure of proteins. *Macromolecules* **9**, 945–950.

31. Miyazawa, S. & Jernigan, R.L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* **18**, 534–552.

32. Gregoret, L.M. & Cohen, F.E. (1990) Novel method for rapid evaluation of packing in protein structures. *J. Mol. Biol.* **211**, 959–974.

33. Covell, D.G. (1992) Folding protein α-carbon chains in to compact forms by Monte Carlo methods. *Proteins* **14**, 409–420.

34. Hinds, D.A. & Levitt, M. (1994) Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* **243**, 668–682.

35. Skolnick, J. & Koliński, A. (1990) Simulations of the folding of a globular protein. *Science* **250**, 1121–1125.

36. Koliński, A. & Skolnick, J. (1992) Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides. *J. Chem. Phys.* **97**, 9412–9426.

37. Koliński, A., Godzik, A. & Skolnick, J. (1993) A general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *J. Chem. Phys.* **98**, 7420–7433.

38. Godzik, A., Koliński, A. & Skolnick, J. (1993) *De novo* and inverse folding predictions of protein structure and dynamics. *J. Comput.-Aided Mol. Design* **7**, 397–438.

39. Skolnick, J., Koliński, A., Brooks, C.L., Godzik, A. & Rey, A. (1993) A method for predicting protein structure from sequence. *Curr. Biol.* **3**, 414–424.

40. Koliński, A. & Skolnick, J. (1994) Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* **18**, 338–352.

41. Koliński, A. & Skolnick, J. (1994) Monte Carlo simulations of protein folding. II. Application to protein A and crambin. *Proteins* **18**, 353–366.

42. Vieth, M., Koliński, A., Brooks, C.L. & Skolnick, J. (1994) Prediction of the folding pathways and structure of the GCN4 leucine zipper. *J. Mol. Biol.* **237**, 361–367.

43. Koliński, A., Milik, M. & Skolnick, Z. (1995) A reduced model of short range interactions in polypeptide chains. *J. Chem. Phys.* **103**, 4312–4323.

44. Godzik, A. Koliński, A. & Skolnick, J. (1995) Are proteins ideal mixtures of amino acids? Analysis of energy parameter set. *Protein Sci.* **4**, 2107–2117.

45. Hirst, J.D., Vieth, M., Skolnick, J. & Brooks III, C.L. (1996) Predicting leucine zipper structures from sequence. *Protein Eng.* **9**, 657–662.

46. DeBolt, S.E. & Skolnick, J. (1996) Evaluation of atomic level mean force potentials *via* inverse folding and inverse refinement of protein structures: Atomic burial position and pairwise non-bonded interactions. *Protein Eng.* **9**, 637–655.

47. Crippen, G.M. & Wiswanadhan, V.N. (1984) A potential function for conformational analysis of proteins. *Int. J. Peptide Protein Res.* **24**, 279–296.

48. Crippen, G.M. & Viswandhan, V.N. (1985) Sidechain and backbone potential function for conformational analysis of proteins. *Int. J. Peptide Protein Res.* **25**, 487–509.

49. Crippen, G.M. & Ponnuswamy, P.K. (1987) Determination of an empirical energy function for protein conformational analysis by energy embedding. *J. Comput. Chem.* **8**, 972–981.

50. Crippen, G.M. & Snow, E. (1990) A 1.8 Å resolution potential function for protein folding. *Biopolymers* **29**, 1479–1489.

51. Seetharamulu, P. & Crippen, G.M. (1991) A potential function for protein folding. *J. Math. Chem.* **6**, 91–110.

52. Maiorov, V.N. & Crippen, G.M. (1992) Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.

53. Crippen, G.M. (1996) Easily searched protein folding potentials. *J. Mol. Biol.* **231**, 467–476.

54. Nishikawa, K. & Matsuo, Y. (1993) Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Protein Eng.* **8**, 811–820.

55. Sippl, M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.

56. Casari, G. & Sippl, M.J. (1992) Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725–732.

57. Sippl, M.J. (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput.-Aided Mol. Design* **7**, 473–501.

58. Sun, S. (1993) Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.

59. Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S. & Scheraga, H.A. (1993) Calculation of protein backbone geometry from α-carbon coordinates based on peptide-group dipole alignment. *Protein Sci.* **2**, 1697–1714.

60. Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S. & Scheraga, H.A. (1993) Prediction of protein conformation on the basis of a search for compact structures: Test on avian pancreatic polypeptide. *Protein Sci.* **2**, 1715–1731.

61. Liwo, A., Ołdziej, St., Pincus, M.R., Wawak, R.J., Rackovsky, S. & Scheraga, H.A. (1997) A united-residue force field for off-lattice protein-structure simulations. I: Functional

forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.* **18**, 849–873.

62. Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S., Ołdziej, St. & Scheraga, H.A. (1997) A united-residue force field for off-lattice protein-structure simulations. II: Parameterization of local interactions and determination of the weights of energy terms by Z-score optimization. *J. Comput. Chem.* **18**, 874–887.

63. Liwo, A., Ołdziej, St., Czaplewski, C., Groth, M., Kaźmierkiewicz, R., Pincus, M.R., Wawak, R.J., Rackovsky, S. & Scheraga, H.A. (1996) A knowledge based united-residue force field for off-lattice calculations of protein structure that recognizes native folds. *International Symposium on Theoretical and Experimental Aspects of Protein Folding*, San Luis, Argentina, 17–21 June, 1996; Liwo, A., Kaźmierkiewicz, R., Czaplewski, C., Groth, M., Ołdziej, St., Wawak, R.J., Rackovsky, S., Pincus, M.R. & Scheraga, H.A. (1996) Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *Third Electronic Computational Chemistry Conference, ECCC-3, Northern Illinois University*, 1–30 November 1996, Poster No. 19.

64. Dill, K.A. (1990) Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155.

65. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

66. Rackovsky, S. (1990) Quantitative organization of the known protein X-ray structures. I. Methods and short-length-scale results. *Proteins* **7**, 378–402.

67. Reichl, R.E. (1980) *Modern Course in Statistical Mechanics*; pp. 142–144, University of Texas Press, Austin.

68. Bürgi, H.B. & Dunitz, J.D. (1983) From crystal statics to chemical dynamics. *Acc. Chem. Res.* **16**, 153–161.

69. Momany, F.A., McGuire, R.F., Burgess, A.W. & Scheraga, H.A. (1975) Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. Phys. Chem.* **79**, 2361–2381.

70. Némethy, G. & Scheraga, H.A. (1983) Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bonding interactions for the naturally occurring amino acids. *J. Phys. Chem.* **87**, 1883–1891.

71. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

72. Sippl, M.J., Némethy, G. & Scheraga, H.A. (1984) Intramolecular potentials from crystal data. 6. Determination of empirical potentials for O–H...O=C hydrogen bonds from packing configurations. *J. Phys. Chem.* **88**, 6231–6233.

73. Némethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S. & Scheraga, H.A. (1992) Energy parameters in polypeptides.10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm with application to proline-containing peptides. *J. Phys. Chem.* **96**, 6472–6484.

74. Vila, J., Williams, R.L., Vásquez, M. & Scheraga, H.A. (1991) Empirical solvation models can be used to differentiate native from near-native conformations for bovine pancreatic trypsin inhibitor. *Proteins* **10**, 199–218.

75. Blundell, T.L., Pitts, J.E., Tickle, I.J., Wood, S.P. & Wu, C.-W. (1981) X-ray analysis (1.4 Å resolution) of avian pancreatic polypeptide: Small globular protein hormone. *Proc. Natl. Acad. Sci. U.S.A.* **78**, 4175–4179.

76. Glover, I., Haneef, I., Pitts, J., Wood, S., Moss, D., Tickle, I. & Blundell, T. (1983) Conformational flexibility in a small globular hormone: X-ray analysis of avian pancreatic polypeptide at 0.98 Å resolution. *Biopolymers* **22**, 293–304.

77. Noelken, M.E., Chang, P.J. & Kimmel, J.R. (1980) Conformation and association of pan-

creatic polypeptide from three species. *Biochemistry* **19**, 1838–1843.

78. Chang, P.J., Noelken, M.E. & Kimmel, J.R. (1980) Reversible dimerization of avian pancreatic polypeptide. *Biochemistry* **19**, 1844–1849.

79. Bersani, M., Johnsen, A.H., Hojrup, P., Dunning, B.E., Andreasen, J.J. & Holst, J.J. (1991) Human galanin: Primary structure and identification of two molecular forms. *FEBS Lett.* **283**, 189–194.

80. Liwo, A., Ołdziej, St., Ciarkowski, J., Kupryszewski, G., Pincus, M.R., Wawak, R.J., Rackovsky, S. & Scheraga, H.A. (1994) Prediction of conformation of rat galanin in the presence and absence of water with the use of Monte Carlo methods and the ECEPP/3 force field. *J. Protein Chem.* **13**, 375–380.

81. Morris, M.B., Ralstone, G.B., Biden, T.J., Browne, C.L, King, G.F. & Iismaa, T.P. (1995) Structural and biochemical studies of human galanin: NMR evidence for nascent helical structure in aqueous solution. *Biochemistry* **34**, 4538–4545.

82. Lee, J. (1993) New Monte Carlo algorithm: Entropic sampling. *Phys. Rev. Lett.* **71**, 211–214.

83. Wawak, R.J., Gibson, K.D., Liwo, A. & Scheraga, H.A. (1996) Theoretical prediction of a crystal structure. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1743–1746.