Protein Science (1993), *2*, 1715–1731. Cambridge University Press. Printed in the USA.

Prediction of protein conformation on the basis of a search for compact structures: Test on avian pancreatic polypeptide

A. LIWO,[1,4] M.R. PINCUS,[2] R.J. WAWAK,[1] S. RACKOVSKY,[3] AND H.A. SCHERAGA[1]

[1] Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853-1301
[2] Department of Pathology, Division of Clinical Pathology, State University of New York, Health Science Center, Syracuse, New York 13210
[3] Department of Biophysics, School of Medicine and Dentistry, University of Rochester, Rochester, New York 14642

(Received March 2, 1993; Accepted June 24, 1993)

Abstract

Based on the concept that hydrophobic interactions cause a polypeptide chain to adopt a compact structure, a method is proposed to predict the structure of a protein. The procedure is carried out in four stages: (1) use of a virtual-bond united-residue approximation with the side chains represented by spheres to search conformational space extensively using specially designed interactions to lead to a collapsed structure, (2) conversion of the lowest-energy virtual-bond united-residue chain to one with a real polypeptide backbone, with optimization of the hydrogen-bond network among the backbone groups, (3) perturbation of the latter structure by the electrostatically driven Monte Carlo (EDMC) procedure, and (4) conversion of the spherical representation of the side chains to real groups and perturbation of the whole molecule by the EDMC procedure using the empirical conformational energy program for peptides (ECEPP/2) energy function plus hydration. Application of this procedure to the 36-residue avian pancreatic polypeptide led to a structure that resembled the one determined by X-ray crystallography; it had an α-helix starting at residue 13, with the N-terminal portion of the chain in an extended conformation packed against the α-helix. Similar structures with slightly higher energies, but looser packing, were also obtained.

Keywords: compact conformations; conversion from a united-residue representation to an all-atom chain; hydrophobic-residue packing; Monte Carlo methods; multiple-minima problem; potential of mean force; protein folding; united-residue representation of a polypeptide chain

In our continuing effort to surmount the multiple-minima problem (Scheraga, 1989; Kostrowicki & Scheraga, 1992; Olszewski et al., 1992) in computing the structure of a protein, we have developed a procedure that takes advantage of the fact that the protein core tends to consist of tightly packed nonpolar residues, with the polar ones located on the surface (Kauzmann, 1959; Rackovsky & Scheraga, 1977; Richards, 1977; Wertz & Scheraga, 1978; Meirovich et al., 1980; Meirovich & Scheraga, 1980; Chan & Dill, 1990; Dill, 1990). Such an example of self-organization is reminiscent of early work by Onsager (1949) on the packing of tobacco mosaic virus particles and by Flory (1956) on the packing of rodlike polymers (including α-helices). These workers showed that, as the solution concentration increases, the system separates into two phases, an organized ordered one and a more dilute isotropic one. The self-organization of the more-concentrated ordered phase arises solely from entropic effects; i.e., in a concentrated solution it is easier to pack rods in an ordered anisotropic array than in a disordered isotropic one. Of course, while entropy alone can account for this phenomenon, possible attractive interactions between the ordered rods can also contribute to this self-organization. We thus consider the possibility that, if the available conformational space of a polypeptide is confined, organized structure will be promoted.

Recent extensive studies by Dill and coworkers (Chan & Dill, 1989a,b, 1990; Lau & Dill, 1990) with the use of a lattice model of the polypeptide chain have shown that compact structures constitute only a tiny fraction of the whole conformational space and that compactness alone can induce a considerable amount of ordered structure (helices, sheets, and turns). Covell and Jernigan (1990),

Reprint requests to: H.A. Scheraga, Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853-1301.

[4] On leave from the Department of Chemistry, University of Gdańsk, ul. Sobieskiego 18, 80-952 Gdańsk, Poland, 1990–1992. Contact A. Liwo for computer programs.

using another lattice model with interresidue energies of interaction calculated from the frequency of contacts (Miyazawa & Jernigan, 1985), demonstrated that the native structure is found among a very small fraction of low-energy structures in a larger set of compact structures. Hao et al. (1992), using an off-lattice treatment and an all-atom representation of a polypeptide backbone, found that confining the chain to a fixed volume and shape, as given by the three principal axes of the molecular ellipsoid, induces a large amount of sheet structure. It can, therefore, be concluded that the computational time to search the native structure can be reduced by confining the chain to a set of compact conformations. This is the basic principle of the method developed in this work, which involves the ultimate use of a real chain rather than a lattice model.

The second basic idea of our procedure is that a crude search of compact structures can be carried out initially not at the atomic level but at the amino acid-residue level, thereby reducing the time required for a complete search by several orders of magnitude. For this purpose, we have developed an approximate potential function that incorporates the tendency of nonpolar residues to form compact clusters in the early stage of the procedure; this is then followed by all-atom computations using the empirical conformational energy program for peptides (ECEPP/2) potential with hydration included.

We first describe the general strategy of the method, then the united-residue potential and the method to convert a virtual-bond united-residue structure to a real chain. Finally, we present the results of the application of this method to a small protein, the avian pancreatic polypeptide.

Procedure

The general strategy of our procedure can be summarized as follows:

1. The polypeptide chain is first described in a virtual-bond united-residue approximation, with the side chains represented by spheres. An appropriate potential is chosen to reflect the tendency for nonpolar residues to associate and for polar backbone peptide groups to optimize their hydrogen-bonding arrangement, and to take account of local interactions. Then, the conformational space of the simplified chain is explored to find low-energy conformations (that appear compact as a result of hydrophobic packing). We have used the Monte Carlo with energy minimization (MCM) method (Li & Scheraga, 1987, 1988) to carry out this exploration. A cluster analysis is then applied to divide the set of low-energy conformations into clusters (families) of similar conformations whose lowest-energy representatives are hereafter referred to as *structures*. Structures having energies within a chosen cut-off value above the lowest-energy structure are saved for further stages of the calculation.
2. The virtual-bond united-residue structures thus found are converted to an all-atom backbone. This is accomplished by optimizing the hydrogen-bonding arrangement between the backbone atoms but, at the same time, preserving the distances between α-carbons (virtual-bond structure) obtained in the united-residue treatment. This conversion was carried out with our recently developed dipole-path method, discussed in the accompanying paper (Liwo et al., 1993).
3. Generation of the backbone is completed by carrying out EDMC (electrostatically driven Monte Carlo) simulations (Ripoll & Scheraga, 1988, 1989; Williams et al., 1992) in a "hybrid" representation of the polypeptide chain, i.e., with an all-atom backbone and united side chains, still subject to the C^α-distance constraints following from the united-residue simulations. The latter means that some or even all the distances of the virtual-bond chain obtained in the united-residue simulation are substantially preserved. This, in turn, is aimed at conserving the shape of the molecule that was obtained in the previous stage. Therefore, the resulting structure will mainly acquire optimal peptide-group orientation, having almost unchanged virtual-bond geometry. In this stage, the distances between the α-carbons of the nonpolar residues that are in contact are maintained. Clustering is carried out on the conformations obtained from each structure, and those lowest-energy representatives of the resulting clusters of conformations are selected that have both a low energy and a low root mean square (rms) deviation with respect to the parent virtual-bond united-residue structure.
4. Full (all-atom) side chains are introduced with accompanying minimization of steric overlaps, allowing both the backbone and side chains to move. Then, all-atom EDMC simulations are carried out in order to explore the conformational space in the neighborhood of each of the low-energy structures found in the virtual-bond united-residue treatment. In this stage of the calculations, we used the ECEPP/2 potential (Momany et al., 1975; Némethy et al., 1983; Sippl et al., 1984) with solvation energy evaluated with the SRFOPT model of Vila et al. (1991; see also Williams et al. [1992]). Because this potential includes a solvation term, nonpolar residues should tend to remain packed in the interior; hence, it is no longer necessary to impose the C^α-distance constraints found in stage 1. Nevertheless, the final all-atom conformations satisfied these constraints.

The general procedure outlined above is illustrated in a flow chart (Fig. 1), while the details of the subsequent stages are presented on the following pages.

Virtual-bond united-residue model of a polypeptide chain and the associated force field

Description of model

In stage 1, a polypeptide chain is represented by a sequence of α-carbon atoms (C^{α}) with attached united side chains (SC) and peptide-group centers (p) centered between two consecutive α-carbons, as in Figure 2A. For convenience, we understand by C^{α} any atom initiating or terminating a peptide group. This means that we include here not only the actual C^{α} atoms, but also the terminal carbon atoms of such blocking groups as the acetyl or *N*-methylamide group, and the *trans* amide hydrogens

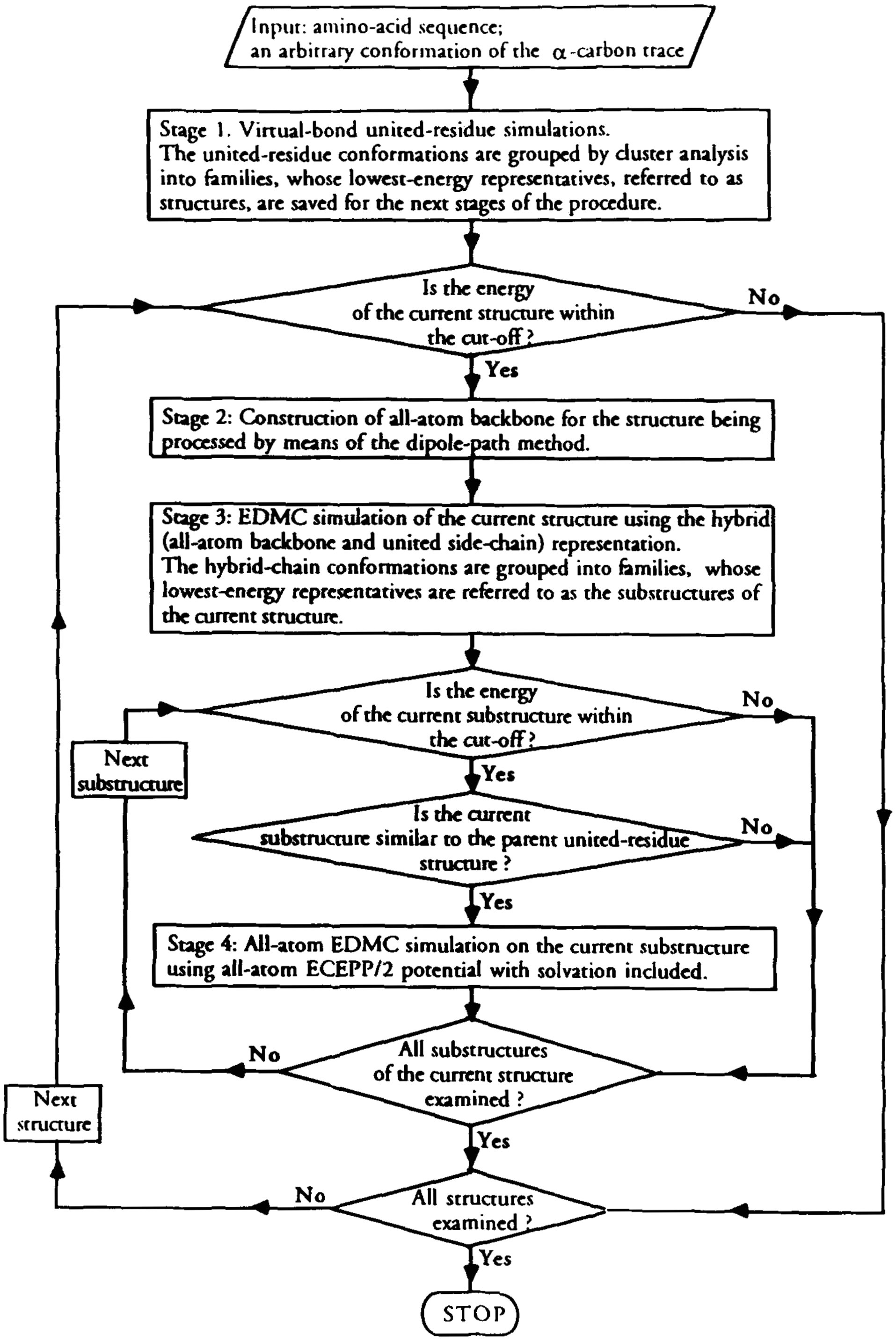
Input: amino-acid sequence;
an arbitrary conformation of the α-carbon trace
Stage 1. Virtual-bond united-residue simulations.
The united-residue conformations are grouped by cluster analysis into families, whose lowest-energy representatives, referred to as structures, are saved for the next stages of the procedure.
Is the energy of the current structure within the cut-off ?
No
Yes
Stage 2: Construction of all-atom backbone for the structure being processed by means of the dipole-path method.
Stage 3: EDMC simulation of the current structure using the hybrid (all-atom backbone and united side-chain) representation.
The hybrid-chain conformations are grouped into families, whose lowest-energy representatives are referred to as the substructures of the current structure.
Is the energy of the current substructure within the cut-off ?
No
Yes
Next substructure
Is the current substructure similar to the parent united-residue structure ?
No
Yes
Stage 4: All-atom EDMC simulation on the current substructure using all-atom ECEPP/2 potential with solvation included.
No
All substructures of the current structure examined ?
Yes
Next structure
No
All structures examined ?
Yes
STOP

Fig. 1. Flow chart of the procedure for protein-structure prediction developed in this work.

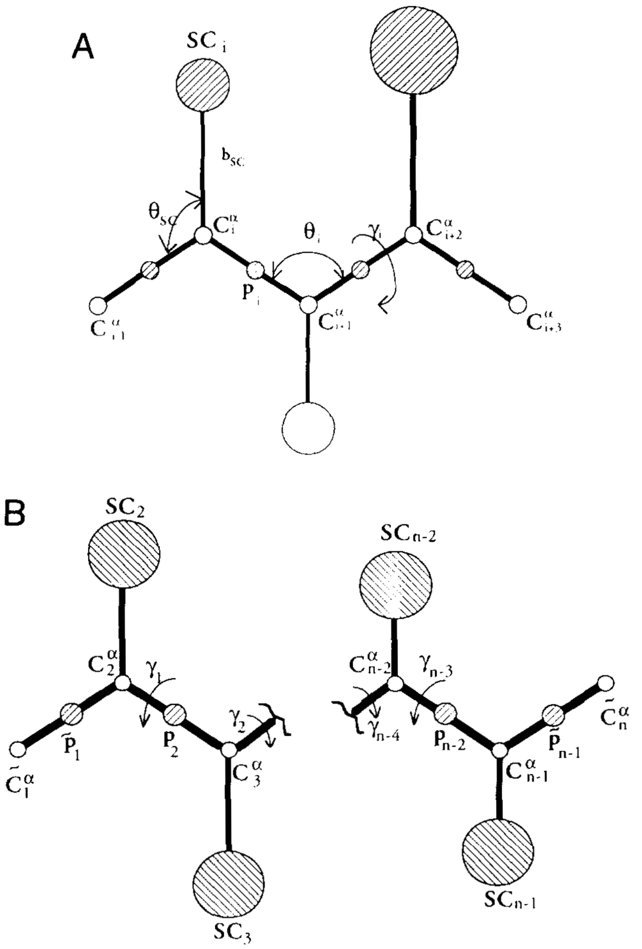

Fig. 2. **A:** United-residue representation of a polypeptide chain. The interaction sites are side-chain centroids (SC) and peptide-bond centers (p) indicated by dashed circles, while the α-carbon atoms (small empty circles) are introduced only to assist in defining the geometry. Each side chain is attached to the corresponding α-carbon with a fixed "bond length", b_{SC}; fixed "bond angle", θ_{SC}, formed by SC$_i$, C^{α}_i, and C^{α}_{i-1}; and fixed "dihedral angle" (Levitt, 1976), formed by SC$_i$, C^{α}_i, C^{α}_{i-1}, and C^{α}_{i+1}. The virtual-bond C^{α}-C^{α}-C^{α} valence angles are assigned a value of 90°. The only variables are thus the virtual-bond torsional angles, γ. **B:** The geometry of an α-carbon united-residue chain consisting of n C^{α} atoms. The variables defining the geometry of all sites are the virtual-bond torsional angles $\gamma_1, \gamma_2, \ldots, \gamma_{n-3}$. $\tilde{C}^{\alpha}_1$ ($\tilde{C}^{\alpha}_n$) is the position of either a real terminal C^{α}, if no side chain is attached to it, or a dummy atom, if the terminal C^{α} of the real chain has a side chain. Consequently, peptide-group centers $\tilde{p}_1$ and $\tilde{p}_{n-1}$ represent either real or dummy peptide groups. The position of SC$_1$ (SC$_n$) (if these sites are present, i.e., if the first [last] residue is glycine or a methyl group) coincides with that of $\tilde{C}^{\alpha}_1$ ($\tilde{C}^{\alpha}_n$).

of the amide groups, if they initiate or terminate the peptide backbone (Fig. 2B). Such blocking groups are identified as a glycine residue, because no side chain is attached to the terminal carbon atom.

Only the sites SC and p are considered as interaction sites in the energy evaluation (see the next section), while the positions of the C^{α}'s are used only to define the geometry. Each SC site represents all the atoms belonging to a real side chain plus the corresponding C^{α}. These atoms are collected together in the united-residue representation, when interaction energies are evaluated. Hence, for glycine the position of the SC coincides with that of its C^{α}. Each p site represents the corresponding C′O-NH group. The distance between successive C^{α} atoms is assigned a value of 3.8 Å, characteristic of a planar *trans* peptide group (a virtual-bond length [Nishikawa et al., 1974]). We also assume that the C^{α}-C^{α}-C^{α} virtual-bond angles (θ) have a fixed value of 90°, which is an approximate value corresponding to the global maximum in the frequency distribution of virtual-bond angles in protein crystal structures (Rackovsky, 1990). Accordingly, the united side chains also have fixed geometry with respect to the C^{α} frame, the geometric parameters being taken from protein crystal data (Levitt, 1976). They are summarized in Table 1.

To define the geometry of the simplified virtual-bond chain, we must distinguish between two cases, because the terminal α-carbons may or may not have side chains. If none of the terminal α-carbons has a side chain, i.e., if both residues are glycines or methyl groups, the only variables describing the geometry of the simplified chain are

Table 1. *Standard geometry and van der Waals radii of the side-chain centroids used in the calculations*[a]

Residue	b_{SC} (Å)	θ_{SC} (deg)	ϕ_{SC} (deg)	r^0 (Å)
Cys	1.38	120.7	−148.5	5.0
Met	2.34	120.5	−154.3	6.2
Phe	2.97	125.6	−154.2	6.8
Ile	1.83	125.2	−138.8	6.2
Leu	2.08	125.4	−152.1	6.3
Val	1.49	127.7	−135.9	5.8
Trp	3.58	125.8	−154.2	7.2
Tyr	3.56	117.8	−163.4	6.9
Ala	0.77	125.3	−111.9	4.6
Gly	0.00	—	—	3.8
Thr	1.43	122.5	−129.8	5.6
Ser	1.28	122.7	−124.1	4.8
Gln	2.58	125.3	−152.4	6.1
Asn	1.98	124.7	−150.6	5.7
Glu	2.63	124.9	−143.5	6.1
Asp	1.99	127.7	−141.7	5.6
His	2.76	124.3	−136.8	6.2
Arg	3.72	128.6	−150.7	6.8
Lys	2.94	128.9	−146.0	6.3
Pro	1.42	86.3	−123.3	5.6

[a] For a frame defined by C^{α}_{i-1}, C^{α}_i, and C^{α}_{i+1}, b_{SC} is the distance from SC$_i$ to C^{α}_i, θ_{SC} is the planar angle defined by atoms SC$_i$-C^{α}_i-C^{α}_{i-1}, and ϕ_{SC} is the dihedral angle defined by atoms SC$_i$-C^{α}_i-C^{α}_{i-1}-C^{α}_{i+1}; hence, the position of SC$_i$ is defined by the α-carbons of the closest residues, $i-1$, i, $i+1$ (see also Fig. 2). The r^0's are defined by Equation 2. Because the virtual-bond angles, θ, have a value of 90°, which is different from the standard value of 106.3° (Levitt, 1976), the values of θ_{SC} and ϕ_{SC} listed here were calculated from the values given by Levitt (1976), subject to the condition that the location of a side chain with respect to the plane bisecting the angle C^{α}_{i-1}-C^{α}_i-C^{α}_{i+1} does not change on changing θ_{i-1} from 106.3° to 90°.

the C^α-C^α-C^α-C^α virtual-bond torsional angles, γ. If either or both terminal residues contains side chains, we extend the chain by adding a terminal dummy residue(s) to the terminal amino and carbonyl groups. This process results in the addition of one (two) additional virtual-bond torsional angles, as shown in Figure 2B. This one (two) extra degree of freedom allows for the rotation of the side chains of the N-and of the C-terminal residues.

It should be noted that our model has much in common with other united-residue models (Levitt, 1976; Pincus & Scheraga, 1977; Crippen & Viswanadhan, 1985; Crippen & Ponnuswamy, 1987; Crippen & Snow, 1990; Seetharamulu & Crippen, 1991).

Use of the united-residue approach is equivalent to Boltzmann-averaging the intrachain interaction energy over all degrees of freedom except the virtual-bond torsional angles (Levitt, 1976), i.e., the averaging over all side-chain dihedral angles χ, and all angles of rotation λ of the peptide groups about respective C^α-C^α axes (see Nishikawa et al. [1974] and the accompanying paper [Liwo et al., 1993] for the definition of λ). Obviously, such an approach to evaluate the united-residue potential is unrealistic, when applied strictly, because of the large number of variables over which the average must be computed (which involves numerical evaluation of multidimensional integrals). Therefore, instead, we assume that the Boltzmann-averaged energy of a polypeptide chain can be approximated by a sum of Boltzmann-averaged site-site pairwise interaction energies and some local terms, also Boltzmann averages. To compute the average local terms, we consider short fragments (involving three consecutive peptide groups) of the chain.

Energy function for united-residue model

The complete energy function for the simplified chain is expressed as follows:

$$U = \sum_{i=ib}^{ie-1} \sum_{j=i+1}^{ie} U_{\mathrm{SC}_i\mathrm{SC}_j} + \sum_{i=ib}^{ie} \sum_{\substack{j=ib \\ j\neq i-1,i}}^{ie-1} U_{\mathrm{SC}_i\mathrm{p}_j} + \sum_{i=ib}^{ie-3} \sum_{j=i+2}^{ie-1} U_{\mathrm{p}_i\mathrm{p}_j} + \sum_{i=ib}^{ie-3} U_{tor}(\gamma_i), \tag{1}$$

where $ib = 1$ if the N-terminal group is not a dummy, and $ib = 2$ otherwise; similarly, $ie = n$ if the C-terminal group is not a dummy, and $ie = n - 1$ otherwise; n is the number of residues (including the terminal dummy residues).

The first term accounts for the interaction between united side chains and is determined mainly by hydrophobic/hydrophilic interactions. The second term is an excluded-volume potential preventing the collapse of side chains on to the peptide groups (for each side chain, it consists of the contributions from all peptide groups preceding and following it in the chain). The third term includes the *average* electrostatic interaction between the centers of the peptide groups, and accounts for hydrogen bonding within the backbone. Finally, U_{tor} consists of those local interactions involved with torsion around C^α-C^α virtual bonds.

The term $U_{\mathrm{SC}_i\mathrm{SC}_j}$, i.e., the average energy of the interaction within an isolated pair of side chains i and j (here C_i^α and C_j^α are included in SC_i and SC_j, respectively), is not calculated in this paper by the actual Boltzmann averaging; instead, we use a 12-6 potential for the hydrophobic interactions, and a potential proportional to $+r^{-6}$ for the hydrophilic interactions. The specific forms of these two potentials are arbitrary, but they satisfy the following two conditions: the energies of the contact interactions (i.e., at the distance equal to the sum of the van der Waals radii of the side chains) calculated by these formulas are equal to the energies of contact interactions estimated from protein crystal data (in this work, we calculate these contact energies from the data reported by Miyazawa and Jernigan [1985]); the asymptotic behavior of both types of potential is the same (except for the sign). Hence,

$$U_{\mathrm{SC}_i\mathrm{SC}_j} = \begin{cases} \epsilon_{ij}\left[\left(\dfrac{r_{ij}^0}{r_{ij}}\right)^{12} - 2\left(\dfrac{r_{ij}^0}{r_{ij}}\right)^{6}\right] \\ \qquad \text{for hydrophobic interactions } (\epsilon > 0) \\ -\epsilon_{ij}\left(\dfrac{r_{ij}^0}{r_{ij}}\right)^{6} \\ \qquad \text{for hydrophilic interactions } (\epsilon < 0), \end{cases} \tag{2}$$

with

$$r_{ij}^0 = (r_i^0 + r_j^0)/2.$$

The van der Waals radii, r^0, of the respective united side chains were taken from set C of Levitt (1976) and are summarized in Table 1. The ϵ's were calculated from the interresidue contact energies of Miyazawa and Jernigan (1985), MJ, using the following formula:

$$\epsilon_{\mathrm{GlyGly}} = 0.025 \text{ kcal/mol} \quad \text{(Levitt, 1976)}$$

$$\epsilon_{ij} = 0.6 \times (\mathcal{E}_{ij}^{MJ} - \mathcal{E}_{ij}^0) \quad \text{for } ij \neq \text{GlyGly}. \tag{3}$$

The factor of 0.6 converts the energy expressed by Miyazawa and Jernigan (1985) in kT units to kcal/mol, assuming $T = 298$ K; $\mathcal{E}_{ij}^{MJ}$ is the MJ contact energy for residue pair ij; $\mathcal{E}_{ij}^0$ is the estimated contact energy of the interaction between peptide groups p_i and p_j, and is calculated by:

$$\mathcal{E}^0_{ij} = \begin{cases} \mathcal{E}^{MJ}_{\text{GlyGly}}, & \text{if both } i \text{ and } j \text{ are not prolines} \\ \mathrm{f}_{\text{Pro}}\mathcal{E}^{MJ}_{\text{GlyGly}}, & \text{if one of the residues is proline} \\ \mathrm{f}_{\text{ProPro}}\mathcal{E}^{MJ}_{\text{GlyGly}}, & \text{if both residues are prolines,} \end{cases}$$

This equation was derived under the assumption that the contact energy of two glycine residues comes almost exclusively from the contact between their peptide groups. The values $f_{\text{Pro}} = 0.79$ and $f_{\text{ProPro}} = 0.41$ reflect the inability of the proline peptide group to act as a proton donor in hydrogen bonding, which diminishes the energy of interactions between two peptide groups if at least one of them is proline. The calculation of these parameters is discussed below (Equation 6).

Equation 3 is justified as follows. The original MJ contact energies include SC-SC, SC-p, p-SC, and p-p interactions. They are treated separately in our potential; hence, to calculate the contact energy of SC_i and SC_j (i.e., the interaction energy between SC_i and SC_j, when the *residues* i and j are in contact), we should subtract from the corresponding MJ energy, $\mathcal{E}^{MJ}_{ij}$, the contributions from SC_i–p_j, p_i–SC_j, and p_i–p_j. We neglect the *contact* interactions SC_i–p_j and p_i–SC_j, because they have been computed to be small, and recall that the p_i–p_j contact energies are equal to $\mathcal{E}^0_{ij}$. For the Gly–Gly pair the value of $\epsilon = 0.025$ kcal/mol was chosen from Levitt's set C (Levitt, 1976) to account for the effect of excluded volume.

When residues i and j are not in contact, we cannot neglect the average SC–p interactions, and we adopt a simple excluded-volume potential, given by Equation 4, to treat these interactions:

$$U_{\text{SC}_i\text{p}_j} = \epsilon_{\text{SCp}}\left(\frac{r^0_{\text{SCp}}}{r_{ij}}\right)^6, \qquad (4)$$

with $\epsilon_{\text{SCp}} = 0.3$ kcal/mol and $r^0_{\text{SCp}} = 4.0$ Å; i.e., this term is a penalty function that forbids too-close contacts of the side chain of one residue with the backbone of another. Because the peptide groups can be considered polar, the functional form of this equation (the same as that chosen for the hydrophilic interactions; Equation 2) is justified. The parameters ϵ_{SCp} and r^0_{SCp} were chosen based on several trial calculations (assigning r the values from the range of 3.0 to 5.0 Å and ϵ from 0.1 to 1.0 kcal/mol) on model systems, including β-turns, α-helices, and β-sheets, so as to avoid both collapse (i.e., an SC–p distance less than 4.0 Å) and too large a contribution from this artificial energy term. In fact, this contribution has proven to be small in all the calculations.

For the p–p interactions, we followed Piela and Scheraga (1987) in representing the electrostatic interactions between peptide groups by the interaction between the permanent dipoles of the peptide groups. We placed these dipoles in the middle of the C^α–C^α virtual bonds. To obtain the average electrostatic energy for a given orientation of a pair of virtual-bond peptide groups (i and j), we averaged the dipole–dipole interaction energy over the angles of the rotation of the peptide groups, λ_i and λ_j (see Fig. 1 of Liwo et al. [1993] and Fig. 3 for the definition of peptide-group arrangement). Since the dipole–dipole interaction energy is averaged over λ_i and λ_j for a given conformation of the virtual-bond chain, this energy depends only on the relative orientation of the C^α–C^α virtual bonds containing dipoles i and j. This averaged energy constitutes a mean-field approximation.

The derivation of the average energy of interaction between peptide-group dipoles is given in the Appendix of Liwo et al. (1993). The other (repulsion and dispersion) contributions to the energy are represented by a 6-12 term that also prevents the collapse of the peptide groups. The complete p–p interaction energy is given by Equation 5:

$$\begin{aligned}\bar{U}_{\text{p}_i\text{p}_j} = {} & \frac{A_{\text{p}_i\text{p}_j}}{r^3_{ij}}(\cos\alpha_{ij} - 3\cos\beta_{ij}\cos\gamma_{ij}) \\ & - \frac{B_{\text{p}_i\text{p}_j}}{r^6_{ij}}[4 + (\cos\alpha_{ij} - 3\cos\beta_{ij}\cos\gamma_{ij})^2 \\ & \qquad - 3(\cos^2\beta_{ij} + \cos^2\gamma_{ij})] \\ & + \epsilon_{\text{p}_i\text{p}_j}\left[\left(\frac{r^0_{\text{p}_i\text{p}_j}}{r_{ij}}\right)^{12} - 2\left(\frac{r^0_{\text{p}_i\text{p}_j}}{r_{ij}}\right)^6\right], \qquad (5)\end{aligned}$$

where α_{ij}, β_{ij}, and γ_{ij} are angles that define the relative orientation of the peptide groups (defined in Equation A6 of Liwo et al. [1993] and also shown in Fig. 1 of Liwo et al. [1993]) and r_{ij} is the distance between the centers of the peptide groups.

The energy surface of peptide-group interactions, calculated by averaging the ECEPP/2 energy for close distances, is qualitatively similar to the surface corresponding to the second term in Equation 5; at large distances, it is similar to the surface representing the first term in this equation. In particular, the location and shape of the minima remain the same in both surfaces in each case.

To evaluate the constants $A_{\text{p}_i\text{p}_j}$, $B_{\text{p}_i\text{p}_j}$, $\epsilon_{\text{p}_i\text{p}_j}$, and $r^0_{\text{p}_i\text{p}_j}$ in Equation 5, we calculated the Boltzmann-averaged ECEPP/2 energy for three model pairs of peptide groups: ordinary–ordinary, ordinary–proline, and proline–proline (an ordinary peptide group is any peptide group except the proline peptide group). Only the C′O-NH atoms were considered (i.e., without the bordering C^α atoms), half of the small residual charges at the C^α atoms being fused with the C′ and N atoms in order to achieve electroneutrality. For the "proline"-type peptide group, the amide hydrogen was simply replaced by a carbon atom, which makes the proline parameters also applicable for any other N-methylated residue, e.g., sarcosine. The average energy of the three model peptide groups was calculated for 20 different distances (from 4.4 to 8.2 Å) and 4 × 8 × 8 = 256 different values of the angles θ^1_{ij}, θ^2_{ij}, and ϕ_{ij}, which were chosen to define the relative orientation of the

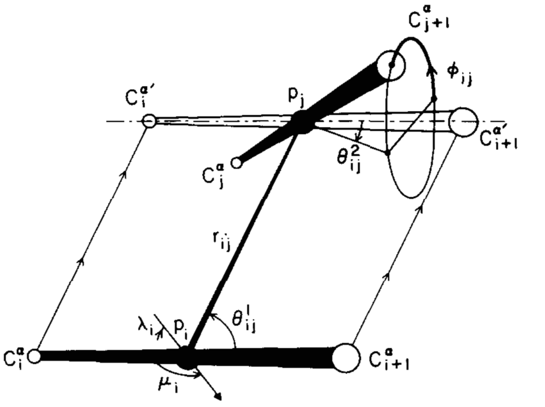

Fig. 3. Definition of the orientation of two peptide groups implemented in the calculation of the average energy. The points C^α_i and C^α_{i+1} (C^α_j and C^α_{j+1}) denote the C^α atoms of the peptide groups p_i (p_j). To define the orientation of the virtual bond C^α_j–C^α_{j+1} relative to the virtual bond C^α_i–C^α_{i+1}, we introduce an *imaginary* virtual bond $C^\alpha_i{}'$–$C^\alpha_{i+1}{}'$, which is parallel to the virtual bond C^α_i–C^α_{i+1} and whose center coincides with the center of the virtual bond C^α_j–C^α_{j+1}. The angle θ^1_{ij} is the angle between the virtual bond C^α_i–C^α_{i+1} and the line connecting the centers of the virtual bonds C^α_i–C^α_{i+1} and C^α_j–C^α_{j+1}; θ^2_{ij} is the angle between the *imaginary* virtual bond $C^{\alpha\prime}_i$–$C^\alpha_{i+1}{}'$ and the virtual bond C^α_j–C^α_{j+1}; ϕ_{ij} is the angle of anticlockwise rotation (when looking from $C^\alpha_{i+1}{}'$ toward $C^\alpha_i{}'$) of the virtual bond C^α_j–C^α_{j+1} about the $C^\alpha_i{}'$–$C^\alpha_{i+1}{}'$ axis. We define $\phi_{ij} = 0$ when the virtual bond C^α_j–C^α_{j+1} lies in the plane of C^α_i–C^α_{i+1} and $C^\alpha_i{}'$–$C^\alpha_{i+1}{}'$, and C^α_{j+1} is on the opposite side of $C^\alpha_i{}'$–$C^\alpha_{i+1}{}'$ from p_i; r_{ij} is the distance between p_i and p_j. For the peptide group p_i, its dipole-moment vector is shown as a small arrow, μ_i is the angle between the dipole-moment vector and the virtual-bond axis, and λ_i is the angle of rotation of the dipole-moment vector about the virtual-bond axis. The energy (averaged over λ_i and λ_j) was calculated for $0° < \theta^1_{ij} < 180°$, $-90° < \theta^2_{ij} < 270°$, and $0° < \phi_{ij} < 360°$, because of the symmetry of the energy surface.

virtual bonds (Fig. 3). This gives 20 × 256 = 5,120 "data points." For each point, the energy was Boltzmann-averaged over 36 values of λ_1 and 36 values of λ_2, which formed a uniform grid on the square $\{0° \leq \lambda_1 < 360°\} \times \{0° \leq \lambda_2 < 360°\}$, with a total of 1,296 points. The parameters $A_{p_ip_j}$, $B_{p_ip_j}$, $\epsilon_{p_ip_j}$, and $r^0_{p_ip_j}$ were then calculated by fitting Equation 5 to the set of average energies by using a linear least-squares procedure. The standard deviation of the energy calculated with Equation 5 from the ECEPP/2 average energy was about 0.06 kcal/mol for each of the three model pairs (compared to the maximum absolute value of energy of about 1.5 kcal/mol at a distance of 4.4 Å). The constants are summarized in Table 2.

In the calculations of the average ECEPP/2 energy, we used a dielectric constant of $D = 4$, since the totally averaged (i.e., over θ^1_{ij}, θ^2_{ij}, ϕ_{ij}, λ_i, and λ_j; see Fig. 3) interaction energy of two "ordinary" peptide groups (as calculated from Equation 5) is almost equal to the MJ energy of interaction between two glycine residues (1.35 kcal/mol vs. 1.3(MJ) kcal/mol) for this value of D. If D is taken as 2, the resulting average energy (2.23 kcal/mol) is closer to the value (3.9 kcal/mol) derived earlier by Tanaka and Scheraga (1976) for the glycine–glycine interaction energy; we use MJ values that are based on a later, more extensive set of X-ray structures.

Table 2. *Values of the constants in Equation 5 for the peptide bond-peptide bond interaction energy*[a]

Type of residue pair	ϵ (kcal/mol)	r^0 (Å)	A $\left(\frac{\text{kcal} \times \text{Å}^3}{\text{mol}}\right)$	B $\left(\frac{\text{kcal} \times \text{Å}^6}{\text{mol}}\right)$
Ordinary-ordinary	0.305	4.51	3.73	1,306
Ordinary-proline	0.365	4.54	0.00	1,129
Proline-proline	0.574	4.48	5.13	335

[a] Obtained from a least-squares fitting to the ECEPP/2 energy averaged over the angles λ_1 and λ_2 for the rotation of peptide groups about the virtual-bond axes.

The factors f_{Pro} and f_{ProPro} in Equation 3 are given by Equation 6:

$$f_{\text{Pro}} = \frac{U^{opt}_{\text{OrdPro}}}{U^{opt}_{\text{OrdOrd}}} \qquad f_{\text{ProPro}} = \frac{U^{opt}_{\text{ProPro}}}{U^{opt}_{\text{OrdOrd}}}, \tag{6}$$

where U^{opt}_{OrdOrd}, U^{opt}_{OrdPro}, and U^{opt}_{ProPro} are the lowest values of the average energies $\bar{U}$, calculated from Equation 5 for two ordinary (non-proline), one ordinary and one proline, and two proline peptide groups, respectively.

The last term in Equation 1, the torsional energy, was evaluated as follows (Levitt, 1976). First, we constructed nine model terminally blocked dipeptides of the type Ac-X-Y-NHMe, where X and Y represent one of the three residues Gly, Ala, or Pro. The alanine residue represents all the L-amino acid residues except proline, which has a different pattern of local interactions, and glycine, which is also a special case because it is not chiral. This is a "basic" set. For each pair, we varied the virtual-bond torsional angle γ from $-180°$ to 180° in increments of 20° and Boltzmann-averaged the ECEPP/2 energy over all angles λ (see Fig. 4). The number of degrees of freedom is reduced by one if one of the residues in the pair is proline and by two if both are prolines because the L-proline residue has a fixed value of the dihedral angle ϕ equal to $-75°$. In order to calculate the dihedral angles ϕ and ψ (which are used as actual input values for ECEPP/2) from γ, λ_1, λ_2, and λ_3 (defined in Fig. 4), including the restriction imposed by proline, we used the algorithm described in the Appendix of the accompanying paper (Liwo et al., 1993). The resulting averaged energy is denoted by $\bar{U}^{XY}_{\text{ECEPP/2}}$.

For each value of γ, U_{tor} was calculated from the Boltzmann-averaged ECEPP/2 energy, $\bar{U}^{XY}_{\text{ECEPP/2}}$, from the following formula (obtained by direct application of Equation 1 to the model shown in Fig. 4):

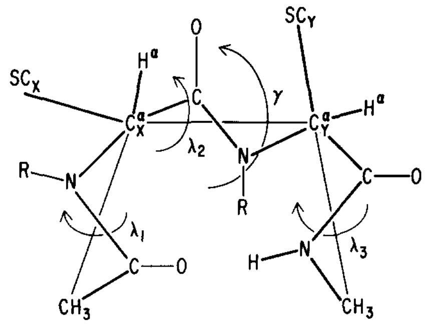

Fig. 4. Schematic representation of the model dipeptides used in the evaluation of the average local interaction energies. X and Y are Gly, Ala, or Pro side chains; R stands for hydrogens in the case of ordinary residues or the side chain δ carbon for proline. The angles λ_1 and λ_2 are identified with $(\lambda_1)_1$ and $(\lambda_1)_2$, while λ_3 is identified with $(\lambda_2)_2$ (see Nishikawa et al. [1974] and the accompanying paper [Liwo et al., 1993]). The average torsional energy is expressed as a function of the virtual-bond torsional angle, γ.

$$U^{XY}_{tor}(\gamma) = \bar{U}^{XY}_{ECEPP/2}(\gamma) - U_{SC_Xp_3}(\gamma) - U_{SC_Yp_1}(\gamma) - U_{(CH_3)_1p_3}(\gamma) - U_{(CH_3)_4p_1}(\gamma) - U_{p_1p_3}(\gamma). \quad (7)$$

Since ECEPP/2 does not include hydration, the interaction terms $U_{SC_XSC_Y}$, $U_{(CH_3)_1SC_Y}$, $U_{SC_X(CH_3)_4}$, and $U_{(CH_3)_1(CH_3)_4}$ are not subtracted from the total energy, $\bar{U}^{XY}_{ECEPP/2}$. The corresponding average ECEPP/2 12-6 energies were calculated to be negligible, in the absence of water.

The resulting data points, $U^{XY}_{tor}(\gamma)$, were fit to a Fourier series (Equation 8), with the coefficients summarized in Table 3:

$$U_{tor}(\gamma_i) = a_0 + \sum_{k=1}^{6} (a_k \cos k\gamma_i + b_k \sin k\gamma_i). \quad (8)$$

A Fourier expansion of order 6 was sufficient in order to cover all cases.

A search through the conformational space of the united-residue force field (stage 1 of the procedure)

The Monte-Carlo-with-energy-minimization (MCM) algorithm is used in stage 1 of the procedure for an extensive search through the conformational space of the simplified chain. Because we have developed a new potential function, the version of the MCM algorithm is slightly different from its original formulation by Li and Scheraga (1987, 1988) or from its more recent development (Ripoll & Scheraga, 1988, 1989; Williams et al., 1992).

We start the MCM process with an arbitrarily chosen conformation (in terms of the virtual-bond torsional angles $\gamma_1, \ldots, \gamma_{n-3}$) and minimize the energy. Then, each step of this process consists of a random perturbation fol-

Table 3. *Coefficients of the Fourier expansion (Equation 8) for the torsional energy about the virtual-bond dihedral angle γ (kcal/mol)*[a]

Residues	a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4	a_5	b_5	a_6	b_6
Gly-Gly	0.2358	0.0000	−0.0231	0.0000	0.1092	0.0000	0.0002	0.0000	−0.0660	0.0000	−0.0908	0.0000
Ala-Gly	−0.0664	0.1956	−0.3537	0.1232	−0.0296	0.0110	−0.1626	0.0854	−0.0928	0.0444	−0.0991	−0.0070
Pro-Gly	0.0067	0.1432	−0.3496	0.2297	0.2691	0.1251	−0.1589	−0.0090	−0.0608	0.0057	−0.0410	−0.0109
Gly-Ala	−0.0518	−0.6427	−0.4497	−0.1558	−0.1150	−0.0611	−0.2401	0.0307	−0.0272	−0.0048	−0.0379	−0.0435
Ala-Ala	0.5361	−0.4189	−0.0855	−0.3691	−0.1473	−0.2484	−0.2823	0.1008	−0.0952	0.1487	−0.0750	0.0532
Pro-Ala	0.5853	−0.5579	−0.1328	−0.1436	0.0579	−0.1318	−0.2751	0.0284	0.0261	0.0682	−0.0833	−0.1180
Gly-Pro	−0.3479	0.9326	−0.5485	−0.3216	0.9967	−0.0065	−0.2566	0.4761	−0.0892	−0.2608	0.3671	−0.0289
Ala-Pro	8.1793	5.7038	−0.8413	0.0401	3.3488	−3.2615	1.5991	0.5468	−1.2076	−0.6290	0.0894	−2.1917
Pro-Pro	3.2742	−1.2914	1.2540	−2.6647	1.0881	−1.9609	−0.3420	−0.6935	−0.5555	−0.7774	0.1951	−0.2442

[a] The coefficient a_0 is not included because it would change the torsional energy only by a constant.

lowed by energy minimization; we apply a Metropolis test in order to accept or reject the new local-minimum conformation, unless both the difference in the energy and the difference in shape between the new and the old conformations are less than preassigned values (in this work $0.03 \times n$ kcal/mol for energy and 10° for dihedral angles). The difference in shape is understood here as the average difference between the two sets of dihedral angles corresponding to two conformations k and l, and is given by the norm in Equation 9:

$$\|\Delta\Gamma^{kl}\| = \sqrt{\frac{\sum_{i=1}^{n-3} w_i |\Delta\gamma_i^{kl}|^2}{\sum_{i=1}^{n-3} w_i}}, \qquad (9)$$

where

$$w_i = 1 - \left(\frac{2i}{n-3} - 1\right)^2$$

$$|\Delta\gamma_i^{kl}| = \min\{|\gamma_i^l - \gamma_i^k|,\ ||\gamma_i^l - \gamma_i^k| - 360°|\}. \qquad (10)$$

The weights w_i, $i = 1, 2, \ldots, n - 3$ were chosen to reflect the fact that changing dihedral angles in the middle of the chain has a greater effect on the conformation than changing dihedral angles near the chain ends. The second of Equations 10 reflects the periodicity of the values of the γ's. Equation 9 provides a better estimate of the differences between conformations than, for example, the maximum value of the difference in dihedral angles, while avoiding the relatively time-consuming computation of the actual rms deviation between conformations.

In those rare occasions in which the rms deviation between the Cartesian coordinates of two structures, k and l, is small, but the value of the norm $\|\Delta\Gamma^{kl}\|$ is large (due to compensating changes in dihedral angles, γ), this will not invalidate the use of Equation 9. For example, when this norm is used for clustering, we will simply have to consider an additional cluster.

Conversion of the α-carbon chain to an all-atom backbone (stage 2 of the procedure)

The dipole-path algorithm used to convert the united-residue chain to the all-atom backbone is described in the accompanying paper (Liwo et al., 1993). This method generates polypeptide backbone structures with optimized backbone–backbone hydrogen bonds in each generated structure. The choice of the dipole-path method to carry out the conversion of the α-carbon trace to the all-atom chain is motivated by the fact that this method uses only the γ-angles and inter-C^α distances as input values, while the "optimum" values of the virtual-bond angles θ are obtained as a result of the procedure. As pointed out earlier, in our united-residue calculations (stage 1), we assumed all θ's to be fixed at a common "average" value of 90°, and this value cannot be maintained for an all-atom backbone. Therefore, it is not possible to use conversion algorithms based on geometric criteria, such as those developed by Purisima and Scheraga (1984) or by Rey and Skolnick (1992). Another reason for using the dipole-path method is that it finds the optimum hydrogen-bond network, preserving the given virtual-chain geometry. This results in a good approximation of a low-energy conformation of the all-atom backbone, subject to the restrictions imposed by the predefined virtual-chain geometry (Liwo et al., 1993).

Refinement of peptide-group orientation by means of the EDMC method with the use of a hybrid force field (stage 3 of the procedure)

Once a set of low-energy structures has been obtained in the united-residue simulations (stage 1) and subsequently converted to all-atom backbones (stage 2), a backbone-conformation refinement stage is introduced. Each of the structures obtained in the preceding stages is subjected to Monte Carlo perturbation using the EDMC method (Ripoll & Scheraga, 1988, 1989; Williams et al., 1992). At this stage of the calculations the all-atom representation of the polypeptide backbone is maintained, while the side chains are still represented by single interaction sites.

For interactions involving the backbone atoms only, the ECEPP/2 force field is applied. In addition, the ECEPP/2 force field is applied for the local interactions (i.e., within a unit comprising an α-carbon with the attached side chain and the two adjacent peptide groups). Because no side-chain degrees of freedom are taken into account, the side chains of all amino acids except glycine and proline are treated as alanines (but with the C^β hydrogens neglected) when evaluating the local interactions. For glycine and proline, with no side-chain degrees of freedom, all atoms are considered in evaluating the local interactions.

For the calculation of the energies of interaction between the united side chains, the side-chain centroids are attached to the α-carbons along the directions of the C^α-C^β bonds, at the distances and with van der Waals radii given in Table 1. The energy of interaction is calculated from Equation 2, using the ϵ values calculated from the respective MJ energies (Equation 3).

For the long-range interactions between the backbone atoms and side chains, the van der Waals radii are calculated as arithmetic means of the van der Waals radii of the respective side chains (Table 1) and those of the corresponding backbone atoms, while the ϵ's are assigned the geometric means of ϵ of an ECEPP/2 type-6 aliphatic carbon (Momany et al., 1975) and that of the corresponding backbone atom.

In order to maintain the shape of the structure as obtained in the united-residue calculations, a Braun and Gō (1985) type function is added to the energy:

$$g(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Omega}) = \frac{1}{4} \sum_{i=1}^{ndist} \begin{cases} (d_i^2 - d_{i0}^2)^2 \\ \quad \text{if } |d_i - d_{i0}| > 0.3 d_{i0} \\ 0 \quad \text{otherwise,} \end{cases} \tag{11}$$

where d_i is the ith C$^\alpha$-C$^\alpha$ distance, d_{i0} is the corresponding distance in the united-residue structure, *ndist* is the number of distances considered in g, and $(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Omega})$ is shorthand for all the dihedral angles ϕ, ψ, and ω in the polypeptide chain.

In the function g, we do not consider the 1,2-or 1,3-adjacent C$^\alpha$'s (if these distances were included, the function g would force the virtual-bond angles of the all-atom backbone to be close to 90° and the virtual-bond lengths to be close to 3.8 Å, as in the parent united-residue chain; these are, however, average values and are not conserved for an all-atom backbone). Moreover, we include in g only the distances between the α-carbons of the hydrophobic residues whose side chains are in contact in the corresponding united-residue structure. Here, a residue is considered hydrophobic if it is positioned to the left of glycine in the MJ contact energy table (Miyazawa & Jernigan, 1985) or is proline. Side chains i and j are considered to be in contact if the following relation holds for $d_{\text{SC}_i\text{SC}_j}$, the distance between the side chains:

$$d_{\text{SC}_i\text{SC}_j} \leq \tfrac{1}{2}(r_i^0 + r_j^0) + 1 \text{ Å}, \tag{12}$$

where r_i^0 and r_j^0 are the van der Waals radii of the side chains (Table 1). Since the topology of the clusters of hydrophobic residues is the most invariant characteristic of a protein (Dill, 1990), preserving the above-mentioned distances as obtained in the united-residue calculations is of primary importance. If we were to include all (except the 1,2 and 1,3) α-carbon distances, the time required for obtaining low-energy conformations would increase greatly, while the rms deviation with respect to the parent united-residue structure would not be lower than when considering hydrophobic contact distances only.

The complete function to be minimized is, therefore:

$$f(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Omega}) = E(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Omega}) + wg(\boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Omega}), \tag{13}$$

where E is the energy of the hybrid chain and w is the weight of the penalty function.

The presence of the penalty function g in the function f to be minimized greatly complicates the EDMC simulations. If we were to sample the dihedral angles randomly, then the minimization of f is likely to be influenced mainly by a possibly large g-term, resulting in a high energy and diminishing the conformation-acceptance rate. Therefore, in this work, the introduction of the penalty function g in Equation 11 is accompanied by a procedure for sampling the backbone dihedral angles in such a way that the distance constraints included in g are not violated. In other words, the perturbations of the variables $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Omega}$ during the Monte Carlo sampling must conserve the distance constraints included in g. The procedure for such sampling of the dihedral angles is described in the Appendix of Liwo et al. (1988).

The hybrid-force-field EDMC calculations, with an acceptance or rejection criterion based on energy alone, were carried out (with f of Equation 13 and $w = 1$ as the weight of the penalty function) for all conformations obtained in stage 2. However, before starting these calculations, three consecutive minimizations of f were carried out (to remove possible steric overlaps gradually) with w equal 100, 10, and finally 1, taking the conformation resulting from the previous minimization as a starting point for the next one. The value $w = 1$ was chosen on the basis of several trial runs so as to achieve a reasonable acceptance rate while avoiding too high an rms deviation from the respective parent united-residue structure.

All-atom force field and procedure for calculations (stage 4 of the procedure)

The all-atom calculations are carried out by using the ECEPP/2 force field with the SRFOPT model for hydration included (Vila et al., 1991). All the side chains and the N-terminal group were taken in their neutral form. However, in order to prevent the formation of hydrogen bonds that do not occur in reality, the carboxyl hydrogens of the glutamic and aspartic acid side chains (the C-terminal group being -CONH$_2$) are assigned ECEPP/2 type-1 (aliphatic carbon hydrogen; Momany et al., 1975), as in the earlier study by Lambert and Scheraga (1989).

In order to obtain starting conformations free of side-chain overlaps, partial optimization of the internal degrees of freedom of the side chains is first carried out, using a procedure similar to that described by Gibson and Scheraga (1987):

1. All the side chains are initially assigned an extended conformation (i.e., χ_1 is set at $-120°$ and the remaining dihedral angles χ at 180°).
2. The consecutive dihedral angles χ are scanned over the whole interval [− 180°, 180°] (an increment of 20° was used in this work), the total ECEPP/2 + solvation energy being calculated at each step, and the optimum position kept for each χ. This step is iterated until the energy does not decrease more than 1 kcal/mol after the scanning of all the dihedral angles χ has been completed.
3. The target function defined by Equation 13, with the hybrid-chain energy replaced by the sum of the all-atom-chain ECEPP/2 energy plus the solvation

free energy, and $w = 1$, is minimized, taking all the dihedral angles as variables.

4. The cycle consisting of steps 2 and 3 is iterated until the energy does not decrease by more than 2 kcal/mol, when compared with its value at the beginning of the cycle.

An unconstrained EDMC run (i.e., without a Braun and Gō [1985] type term in the function to be minimized) is then carried out for each of the starting conformations obtained in this way.

Clustering

A cluster analysis was applied to group conformations (obtained in stages 1, 3, and 4 of the procedure) into clusters of similar conformations. The united-residue structures (stage 1) were clustered taking the difference in dihedral angles γ (Equation 9) as a measure of the distance between two conformations. Conformations with differences in dihedral angles of $\leq 30°$ were grouped into one cluster. This measure of the distance describes the folding-pattern differences better than the rms deviation (at the best superposition of two conformations) when very diverse shapes are to be classified. It also keeps the number of clusters of conformations at a reasonable level, for purposes of clustering.

For the structures with all-atom backbones obtained in stages 3 and 4 of the algorithm, the virtual-torsional-angle norm defined by Equation 9 was not useful as a measure of the distance between conformations because we had to distinguish between structures with similar C^α geometry but different orientations of peptide groups. Because the same folding patterns are compared in this case, the rms deviation at the best superposition of the selected atoms of two conformations was used as a measure of distance. Superposition was carried out by means of the singular value decomposition algorithm (Golub & Van Loan, 1985), using the N, C^α, C', and C^β atoms for the hybrid chain and all non-hydrogen atoms for the all-atom-chain conformations. To define the clusters in stages 3 and 4, the value of 1.0 Å was chosen for the rms deviation cutoff, based on several trial calculations with the cut-off value ranging from 4.0 to 0.1 Å.

In all three stages, the minimal-tree algorithm (Späth, 1980) was used for clustering.

Results and discussion

Test protein

In order to assess the applicability of our algorithm to treat the protein folding problem, we chose the avian pancreatic polypeptide (APP) – a small 36-residue protein. Its crystal structure was determined by Blundell et al. (1981) at 1.4 Å resolution and then by Glover et al. (1983) at a resolution of 0.98 Å. The structure consists of a polyproline-like or collagen-like helix running from residues 1 to 8, packed against the hydrophobic face of an α-helix that extends from residues 13 to 31. The C-terminus does not participate in the α-helix. Although the molecule forms a dimer both in solution (Chang et al., 1980; Noelken et al., 1980) and in the crystal phase (Blundell et al., 1981), there is some evidence that even the monomer is sufficiently stabilized by hydrophobic interactions between the two domains to retain the X-ray conformation (Chang et al., 1980; Noelken et al., 1980; Blundell et al., 1981). Moreover, the structure of this protein has already been computed correctly by Lambert and Scheraga (1989), using the pattern recognition importance-sampling minimization (PRISM) method.

United-residue simulations (stage 1)

The simplified APP chain consists of 37 "α-carbons," starting from C^α of Gly[1] and ending at the *trans* amide hydrogen of the C-terminal $CONH_2$ group (which is therefore considered as C^α_{37}), giving a total of 36 peptide-group centers. Thus, the number of backbone peptide groups in our model is 36, as in a real APP molecule.

The united-residue MCM run started from the completely extended chain (i.e., all dihedral angles γ set at 180°) at $T = 1{,}000$ K. This temperature was chosen after several trial runs at different temperatures. At lower temperatures, the molecule became trapped in one of the low-energy regions of the conformational space, while increasing temperature did not allow the fragments of secondary structure being formed to "survive" long enough to assemble into the complete tertiary structure. The simulation was terminated after 1,000 MCM iterations. Cluster analysis revealed 37 clusters of conformations. The lowest-energy representatives of the five lowest-energy clusters had energies within 10 kcal/mol above the lowest energy. These structures, together with their energies and the rms deviations from the α-carbon trace of the crystal 1PPT structure, are shown in Figure 5A–E. The deviations of the α-carbon trace of the lowest-energy united-residue structure from that of the 1PPT crystal structure are shown in Figure 6. These five structures were selected to carry out the next stages of the algorithm.

Additional MCM runs started from initial conformations other than the fully extended one (all α-helical, partially extended [until residue 13], and partially helical [from residue 13 to the C-terminus]) were also carried out. They did not reveal any conformation lower in energy than that found in the first run. Moreover, all of the classes of structures with sufficiently low energies were found in these additional runs.

As shown in Figure 5A, the lowest-energy united-residue structure has the lowest rms deviation from, and is qualitatively similar to, the α-carbon trace of the 1PPT structure (Blundell et al., 1981). In particular, the α-helix

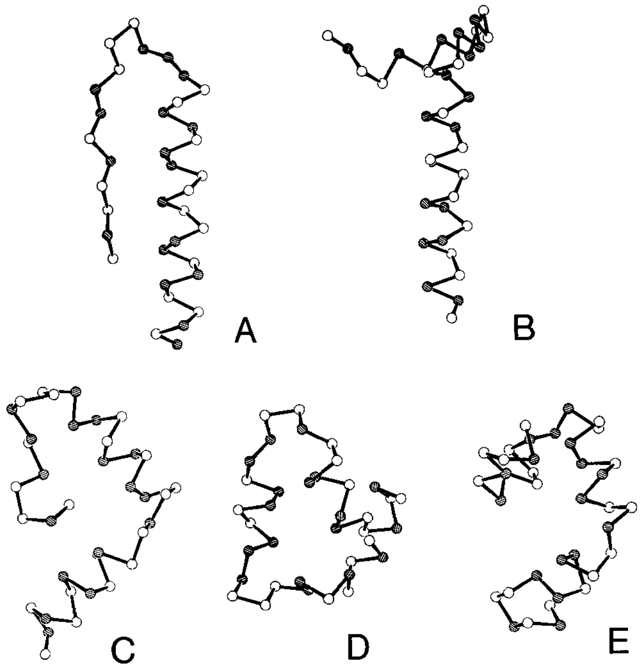

Fig. 5. The lowest-energy representatives of the five lowest-energy clusters of conformations obtained in the united-residue MCM run for APP. Hatched balls indicate the α-carbon atoms of hydrophobic residues, and white balls indicate the α-carbon atoms of hydrophilic residues. **A:** Structure 1; MCM iteration 515; $U = -151.2$ kcal/mol; rms deviation (rmsd) = 3.8 Å. **B:** Structure 2; MCM iteration 586; $U = -148.3$ kcal/mol; rmsd = 5.8 Å. **C:** Structure 3; MCM iteration 921; $U = -144.9$ kcal/mol; rmsd = 6.1 Å. **D:** Structure 4; MCM iteration 92; $U = -143.7$ kcal/mol; rmsd = 7.5 Å. **E:** Structure 5; MCM iteration 366; $U = -143.5$ kcal/mol; rmsd = 6.4 Å.

begins in the correct position and the extended N-terminal part of the chain is packed against its correct side. The main difference is that, in our structure, the α-helix extends as far as the end of the chain, as in the case of the best conformation of APP computed by Lambert and Scheraga (1989). It can also be seen that, in our structure, the two domains are oriented in an antiparallel manner, while in the crystal structure they make an angle of about 150° (Blundell et al., 1981). Finally, there are some minor differences in the loop region, although the shape of the loop agrees qualitatively with that in the crystal structure, in that there is a turn conformation at Asp^{10}-Asp^{11}. The differences pointed out here are shown in Figure 6.

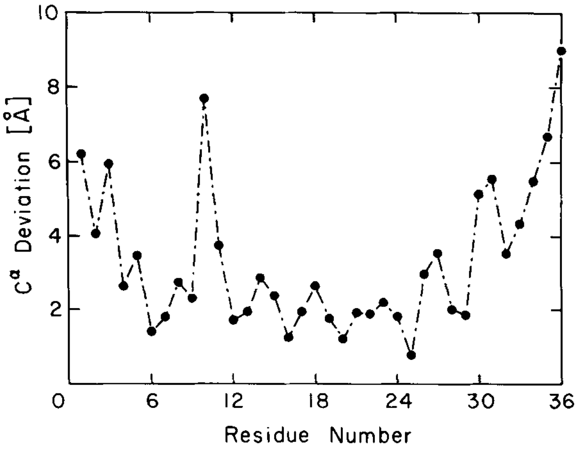

Fig. 6. Plot of the deviations between individual α-carbon atoms for the best superposition of the crystal and lowest-energy united-residue structure of APP (Fig. 5A) along the chain.

The next (higher-energy) structure (Fig. 5B) has an almost correctly formed α-helix, except that it starts one turn later, but the N-terminal part is not packed against it. This structure does not have a much higher energy than the lowest-energy one. Next are structures in which the helix is broken in the middle, which results in a greater number of contacts between the hydrophobic side chains (Fig. 5C–E).

Conversion of united-residue structures to all-atom backbone and hybrid-chain simulations (stages 2 and 3)

The five low-energy structures obtained in the united-residue simulation (Fig. 5A–E) were first converted to all-atom backbones by the dipole-path method (Liwo et al., 1993). The cut-off value of the electrostatic energy for considering two peptide-group centers as participating in a dipole-path network was 0.3 kcal/mol. Because of a high content of helical structure in the united-residue structures, more than 2/3 of the peptide groups participated in the dipole-path network. It is, therefore, very likely that the initial conformational states of at least that fraction of residues were located either in the correct or neighboring (ϕ, ψ) regions of Zimmerman et al. (1977) (Liwo et al., 1993). Then, EDMC simulations with the use of the hybrid force field were carried out on each of the five hybrid-chain structures, using the algorithm described above (section on Refinement of peptide-group orientation).

Each of the five EDMC runs, which led to many intermediate local-minima energy conformations, was terminated after about 200 iterations, no further lower-energy conformations having been obtained in the last 100 iterations.

The numbers of resulting clusters corresponding to each of the five united-residue runs (see section on Clustering) were 10, 20, 13, 3, and 16, respectively. The representatives (the lowest energy conformation within each cluster) of each cluster are hereafter referred to as *substructures*. The lowest-energy substructures corresponding to the five united-residue structures had energies of −96.7, −99.6, −85.5, −76.2, and −77.6 kcal/mol, respectively.

In all cases but structure 2, the α-carbon geometries of the low-energy substructures were similar to those of the parent united-residue structures, the rms deviation being of the order of 3–4 Å (by low-energy substructures within each structure we understand here those substructures whose energies are not higher by more than 10 kcal/mol than the energies listed above). However, in the case of structure 2, whose N-terminal segment was almost unconstrained in the hybrid-force-field run, because of low compactness of this structure, the resulting lowest-energy substructure turned out to be more similar to the united-residue structure 1 (rms = 3.3 Å) than to the parent united-residue structure 2 (rms = 4.7 Å).

All substructures obtained by cluster analysis constituted the set of possible conformations used later to generate starting approximations for the all-atom EDMC simulations of stage 4. In order to select appropriate substructures, we applied the following criteria:

1. First, those substructures were selected whose energy was within 5 kcal/mol above the lowest-energy conformation found in the corresponding hybrid-force-field EDMC run.
2. From the substructures selected in step 1, those with an rms deviation from the parent united-residue structure not greater than 4 Å were selected. Substructures with a higher rms deviation from the parent united-residue structure usually had a folding pattern different from that of the united-residue structure (e.g., the helix present in the parent united-residue structure was broken).
3. From the substructures selected in steps 1 and 2, those substructures were eliminated whose rms deviation with respect to any other united-residue structure was lower than that with respect to the parent one. This happened only for structure 2, for which some of the substructures were closer to united-residue structure 1 than to structure 2, as stated above.
4. In addition to the substructures selected according to criteria 1–3, one more substructure of a given structure was accepted, if it satisfied criteria 2 and 3 and criterion 1 with a 10-kcal/mol instead of 5-kcal/mol energy cut-off, and if it had the lowest rms deviation from the parent united-residue structure of all the conformations belonging to the same structure.

The numbers of substructures selected for all-atom simulations according to the above criteria were 4, 6, 2, 2, and 1, for structures 1–5, respectively, giving a total of 15 substructures.

All-atom simulations (stage 4)

In using EDMC in stage 4, no hydration was included in the electrostatic sampling part of the procedure. The starting conformations for the EDMC simulations were obtained from the 15 hybrid-chain substructures selected from the results obtained in the hybrid-force-field runs (see the preceding section) by introducing all-atom side chains, followed by the side-chain optimization procedure described earlier. In each of the 15 all-atom runs, about 100 EDMC iterations were carried out. As a condition for termination, it was required that no conformation lower in energy than those found so far would appear in the last 30 iterations.

Cluster analysis of each of the 15 sets of conformations obtained in the EDMC simulations was carried out, and the lowest-energy conformation from each cluster was selected as a representative of the whole cluster. As a result of this procedure, 128 representative conformations were obtained, of which 28 were within 10 kcal/mol above the lowest-energy all-atom conformation found in this study. Most of them were derived from united-residue structures 1 and 2. Structure 3 gave only two low-energy all-atom conformations, while no low-energy all-atom conformations were found for structures 4 and 5. The total (ECEPP/2 + hydration), ECEPP/2, and hydration energies of the low-energy all-atom conformations, as well as the lowest-energy all-atom conformations found for structures 4 and 5, together with other selected properties, are summarized in Table 4. It can be seen that the lowest-energy conformation is the native-like one, as in the united-residue simulations. In the case of structure 2, many conformations, in particular the lowest-energy all-atom conformation obtained for this structure, are more similar to the united-residue structure 1 than to structure 2, as also observed in the hybrid-force-field simulations. Therefore, they were finally assigned to structure 1 (Table 4). It is also noteworthy that the energy ranking of the lowest-energy conformations corresponding to each structure is the same as obtained in the united-residue calculations (Table 4 and legend to Fig. 5A–E), which shows that our united-residue force field is compatible with the all-atom one. The conformational states of the lowest-energy all-atom conformation obtained for each structure, together with the conformational states of the 1PPT structure, are summarized in Table 5.

The comparison of the lowest-energy conformation obtained in this study with the 1PPT crystal structure reveals the same differences as appeared in the case of the lowest-energy united-residue structure: a longer helix and antiparallel packing of the extended and helical domains. However, in comparison with the lowest-energy united-residue structure, the loop region is closer to that of the crystal structure (Figs. 7, 8).

Although the native-like conformation was the lowest-energy one found in this study, the data in Table 4 indicate that it is not separated from the conformations derived from structure 2 by a large energy gap. Figure 9, in which are superposed the low-energy conformations from both clusters, indicates that the N-terminal part may be quite flexible in a solution of monomeric APP, although it is either more loosely or more tightly packed against the α-helix.

Table 4. *Summary of the low-energy conformations of APP obtained in the final all-atom EDMC simulations*[a]

No.	Ns	Nf	E_{tot}	E	E_{hyd}	r_{36}^{cryst}	r_{33}^{cryst}	r^u	r_1^u	r_{gyr}
1	1	1	−565.5	−373.3	−192.2	4.10	3.26	2.48		11.40
2	2	1	−565.4	−388.5	−176.9	5.86	5.43	5.10	3.95	11.86
3	2	2	−564.9	−368.7	−196.2	6.61	6.33	4.83	5.27	12.94
4	1	1	−563.8	−378.2	−185.6	4.33	3.45	2.22		11.61
5	2	2	−563.6	−362.4	−201.2	6.40	6.12	5.05	5.48	12.56
6	2	1	−563.5	−378.8	−184.7	5.43	5.13	5.67	4.20	11.41
7	3	3	−563.5	−370.6	−192.9	8.42	7.22	5.08	7.33	10.05
8	2	1	−563.3	−371.0	−192.3	5.57	5.16	4.56	3.86	12.52
9	2	2	−562.8	−365.5	−197.3	6.82	6.59	5.47	5.57	13.03
10	1	1	−562.5	−367.9	−194.6	5.12	4.47	3.05		11.35
11	2	2	−561.5	−371.0	−190.5	6.11	5.77	3.56	5.07	12.42
12	2	2	−561.0	−383.8	−177.2	5.58	5.21	3.43	4.34	12.14
13	2	2	−560.7	−375.7	−185.0	7.14	7.02	4.98	5.82	12.66
14	2	1	−560.7	−377.9	−182.6	7.58	7.54	7.15	5.77	11.58
15	2	1	−560.6	−363.2	−197.4	6.69	6.67	6.19	5.52	12.80
16	2	1	−560.4	−369.7	−190.7	5.97	5.60	4.42	4.30	12.58
17	1	1	−559.9	−371.1	−188.8	5.19	4.74	2.83		11.58
18	2	2	−559.6	−370.6	−189.0	6.89	6.82	5.77	5.84	12.44
19	2	2	−559.3	−379.2	−180.1	6.10	5.72	4.25	4.68	12.00
20	1	1	−558.3	−375.4	−182.9	4.20	3.58	2.08		11.72
21	1	1	−557.4	−372.7	−185.2	4.07	3.18	3.48		11.42
22	2	2	−557.2	−373.3	−183.9	5.42	5.18	3.98	4.64	11.75
23	2	2	−557.2	−358.9	−198.3	6.24	5.93	4.16	5.00	12.57
24	1	1	−556.8	−360.7	−196.1	5.20	4.46	3.16		11.27
25	3	3	−556.4	−371.7	−184.7	7.96	6.85	4.05	6.44	10.18
26	2	2	−556.1	−370.2	−185.9	5.46	5.14	3.58	4.98	11.87
27	2	2	−555.9	−369.1	−186.8	6.13	5.96	4.07	5.60	12.15
28	2	1	−555.7	−371.7	−184.0	7.38	7.18	6.26	5.27	12.92
29	4	4	−546.3	−355.1	−191.2	9.31	9.06	5.24	8.12	10.19
30	5	5	−536.4	−357.5	−178.9	6.58	5.09	4.40	6.39	9.67

[a] No., the number of a conformation according to energy ranking; Ns, the parent united-residue structure; Nf, the structure to which a conformation obtained in the all-atom simulations is finally assigned (based on the rms deviation from the parent united-residue structure and from the other united-residue structures); E_{tot}, total energy (ECEPP/2 + solvation); E, ECEPP/2 without hydration energy; E_{hyd}, hydration-free energy (kcal/mol); r_{36}^{cryst}, the α-carbon rms deviation with respect to the crystal 1PPT structure, for all 36 C^{α}'s superposed; r_{33}^{cryst}, the same, but for only the first 33 C^{α}'s superposed; r^u, the rms deviation with respect to the parent united-residue structure; r_1^u, the rms deviation with respect to the first (lowest-energy) united-residue structure (Å) for the conformations originating in the united-residue structures other than 1; r_{gyr}, the radius of gyration (Å).

Table 5. *Conformational states of the 1PPT crystal structure and the lowest-energy all-atom conformations*[a]

Structure	Conformational code[b]
1PPT	F F F F C C B E* A B F F A A A A A A A A A A A A A A A A A A B A* D B
1	A D F C A D A C A D E F A A A A A A A A A A A A A A A A A A A A A A
2	F G F A A D A B* F A D F C A A A A A A A A A A A A A A A A A A A A A
3	A* E F G D C A F C D A A A A A A A A A A A A* C C A A A A A A A A A A
4	A* F F A D C F C A* D F A A A A A A B A A A C C D A B A A A A A A A D
5	C D D B A A* B A A A D C E A A A A A A A A A C A A A A A A A A* C* D A*

[a] Conformational states of residues 2–35 for each of the five lowest-energy all-atom structures of APP.

[b] The conformational codes of the individúal residues assigned according to Zimmerman et al. (1977).

Computational details

The united-residue and hybrid-force-field calculations were carried out on a Stardent series 3000 computer. A total of of 4,432 energy-minimized conformations were generated per 1,000 iterations, which required about 9.5 CPU hours. A total of about 10,000 energy-minimized conformations were generated during each of the hybrid force-field EDMC runs, the execution time being about 70–100 CPU hours. The all-atom EDMC simulations were carried out on either a Stardent 3000 or an IBM 3090 supercomputer, making use of full vectorization of the code for the IBM computations. A total of 2,000–4,500 energy-minimized conformations were generated in each run, the execution time being about 80–180 CPU hours on the Stardent 3000 and about 40–90 CPU hours on the IBM 3090. Energy minimization was carried out with the SUMSL (secant unconstrained minimization solver) routine (Gay, 1983). The MSEED algorithm was used to compute the solvent-exposed surface area and its gradient, to evaluate the solvent contribution to the energy (Perrot et al., 1992).

Conclusions

The method proposed here provides, in stages 1–3, estimates of the low-energy conformations of a protein that can later be processed (stage 4) by Monte Carlo or molecular dynamics methods. The actual prediction is made at the united-residue level (stage 1), which requires a relatively short computation time. The purpose of the two further stages is to make a smooth transition to the all-atom structure. For the avian pancreatic polypeptide, the lowest-energy conformations were native-like in both the united-residue and all-atom representations. It should be noted that APP was not included in the set of proteins used to evaluate the MJ contact energies (Miyazawa & Jernigan, 1985). Also, no other information regarding the native structure of this protein was incorporated into the united-residue force field. Of course, both the united-residue force field and the whole procedure must be tested on a larger number of known protein structures in order to assess the predictive power of the method.

It should also be noted that for *all* of the starting all-atom conformations of APP derived from the lowest-

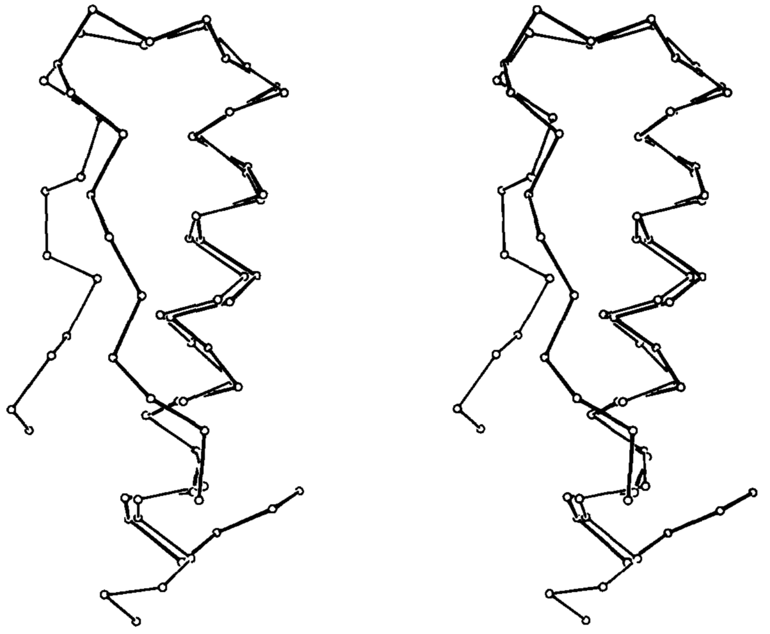

Fig. 7. The α-carbon trace of the lowest-energy all-atom conformation of APP (thin lines) superposed on the 1PPT crystal structure (heavy lines). Residues 13–33 were used in the superposition.

energy united-residue structure, the EDMC simulations found at least one cluster of conformations with energy less than 10 kcal/mol above the energy of the best conformation found during all runs. Moreover, the lowest-energy conformations obtained in *any* EDMC run for which the starting conformation was selected from structure 1 are in reasonable agreement with the crystal structure of APP (Table 4). Therefore, a reasonable structure could be estimated by taking just one of the structures obtained from the united-residue structure of lowest energy.

In the case of APP, the energy ranking of the lowest-energy united-residue structures was also maintained at the all-atom level. The united-residue force field therefore seems to be compatible with the all-atom force field at least in this case. This feature probably results from the fact that the united-residue force field was derived on the basis of energetic considerations (averaging the energy of the all-atom chain over some of the degrees of freedom). This confirms our assumption that the choice of united-residue structures to carry out further stages of the algorithm should be based on their relative energies. The first stage in the conversion of the virtual-bond backbone to the all-atom backbone, the dipole-path method, is also based on energetic criteria, namely optimization of the electrostatic energy of the interactions between the peptide groups of the backbone.

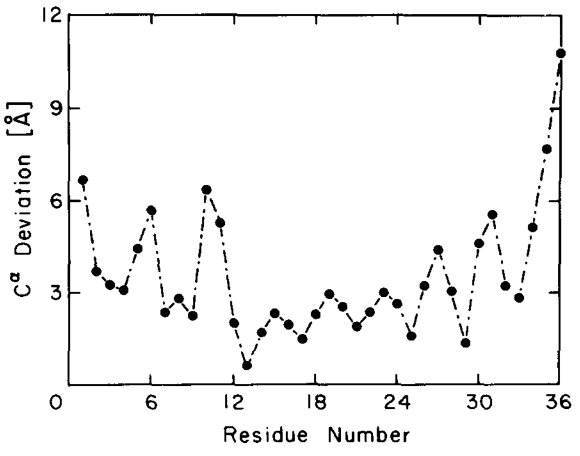

Fig. 8. Plot of the deviations between individual α-carbon atoms for the best superposition of the crystal and the lowest-energy all-atom structure of APP along the chain.

Acknowledgments

This work was supported by grant DMB 90-15815 from the National Science Foundation, grant GM-14312 from the National Institute of General Medical Sciences of the National Institutes of Health, NIH grant CA 42500 (to M.R.P.), and Office of Naval Research grant N00014-91-J-1943 (to S.R.). Support was also received from the National Foundation for Cancer Research.

We thank K.D. Gibson for providing the FORTRAN code for the singular-value-decomposition algorithm, as well as for helpful discussions on the problems encountered in this work; J. Vila and A. Nayeem for providing the FORTRAN code and for help with the EDMC algorithm; and M.J. Dudek for providing the ECEPP-refined coordinates of 1PPT and for helpful comments.

Part of the computations were carried out with one processor of the IBM-3090 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.

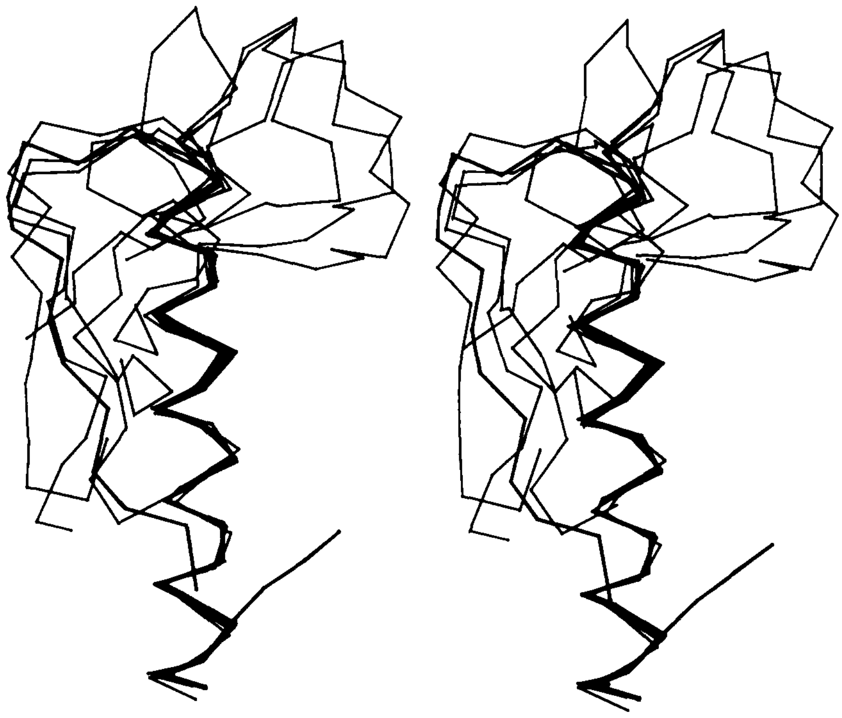

Fig. 9. All conformations derived from structures 1 and 2 of energy within 3-kcal/mol cut-off with respect to the lowest-energy conformation from Table 4 (thin lines), superposed on the 1PPT structure (thick line) (residues 14–33 used in the superposition).

References

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. USA 78*, 4175–4179.

Braun, W. & Gō, N. (1985). Calculation of protein conformations by proton–proton distance constraints. A new efficient algorithm. *J. Mol. Biol. 186*, 611–626.

Chan, H.S. & Dill, K.A. (1989a). Intrachain loops in polymers: Effect of excluded volume. *J. Chem. Phys. 90*, 492–509.

Chan, H.S. & Dill, K.A. (1989b). Compact polymers. *Macromolecules 22*, 4559–4573.

Chan, H.S. & Dill, K.A. (1990). Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA 87*, 6388–6392.

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

Covell, D.G. & Jernigan, R.L. (1990). Conformations of folded proteins in restricted spaces. *Biochemistry 29*, 3287–3294.

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.

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

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

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

Flory, P.J. (1956). Phase equilibria in solutions of rod-like particles. *Proc. R. Soc. Lond. A234*, 73–89.

Gay, D.M. (1983). Algorithm 611. Subroutines for unconstrained minimization using a model/trust-region approach. *Assoc. Comput. Math. Trans. Math. Software 9*, 503–524.

Gibson, K.D & Scheraga, H.A. (1987). Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization. *J. Comput. Chem. 8*, 826–834.

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.

Golub, G.H. & Van Loan, C.F. (1985). *Matrix computations*, pp. 425–426. The Johns Hopkins University Press, Baltimore, Maryland.

Hao, M.H., Rackovsky, S., Liwo, A., Pincus, M.R., & Scheraga, H.A. (1992). Effects of compact volume and chain stiffness on the conformations of native proteins. *Proc. Natl. Acad. Sci. USA 89*, 6614–6618.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem. 14*, 1–63.

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

Lambert, M.H. & Scheraga, H.A. (1989). Pattern recognition in the prediction of protein structure. III. An importance-sampling minimization procedure. *J. Comput. Chem. 10*, 817–831.

Lau, K.F. & Dill, K.A. (1990). Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA 87*, 638–642.

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

Li, Z. & Scheraga, H.A. (1987). Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. USA 84*, 6611–6615.

Li, Z. & Scheraga, H.A. (1988). Structure and free energy of complex thermodynamic systems. *J. Mol. Struct. (Theochem.) 179*, 333–352.

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.

Liwo, A., Tempczyk, A., & Grzonka, Z. (1988). Molecular mechanics calculations on deaminooxytocin and on deamino-arginine-vasopressin and its analogues. *J. Comput.-Aided Mol. Design 2*, 281–309.

Meirovich, H., Rackovsky, S., & Scheraga, H.A. (1980). Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids. *Macromolecules 13*, 1398–1405.

Meirovich, H. & Scheraga, H.A. (1980). Empirical studies of hydrophobicity. 2. Distribution of the hydrophobic, hydrophilic, neutral, and ambivalent amino acid residues in the interior and exterior of native proteins. *Macromolecules 13*, 1406–1414.

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

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.

Némethy, G., Pottle, M.S., & 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*, 1883-1887.

Nishikawa, K., Momany, F.A., & Scheraga, H.A. (1974). Low-energy structures of two dipeptides and their relationship to bend conformations. *Macromolecules 7*, 797-806.

Noelken, M.E., Chang, P.J., & Kimmel, J.R. (1980). Conformation and association of pancreatic polypeptide from three species. *Biochemistry 19*, 1838-1843.

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.

Onsager, L. (1949). The effect of shapes on the interaction of colloidal particles. *Ann. N.Y. Acad. Sci. 51*, 627-659.

Perrot, G., Cheng, B., Gibson, K.D., Vila, J., Palmer, K.A., Nayeem, A., Maigret, B., & Scheraga, H.A. (1992). MSEED: A program for the rapid analytical determination of accessible surface areas and their derivatives. *J. Comput. Chem. 13*, 1-11.

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.

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

Purisima, E.O. & Scheraga, H.A. (1984). Conversion from a virtual-bond chain to a complete polypeptide backbone chain. *Biopolymers 23*, 1207-1224.

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

Rackovsky, S. & Scheraga, H.A. (1977). Hydrophobicity, hydrophilicity, and the radial and orientational distribution of residues in native proteins. *Proc. Natl. Acad. Sci. USA 74*, 5248-5251.

Rey, A. & Skolnick, J. (1992). Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. *J. Comput. Chem. 13*, 443-456.

Richards, F.M. (1977). Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng. 6*, 151-176.

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.

Ripoll, D.R. & Scheraga, H.A. (1989). The multiple-minima problem in the conformational analysis of polypeptides. III. An electrostatically driven Monte Carlo method: Tests on enkephalin. *J. Protein Chem. 8*, 263-287.

Scheraga, H.A. (1989). Calculations of stable conformations of polypeptides, proteins, and protein complexes. *Chem. Scripta 29A*, 3-13.

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

Sippl, M.J., Némethy G., & Scheraga, H.A. (1984). Intermolecular 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.

Späth, H. (1980). *Cluster analysis algorithms*, pp. 170-194. Halsted Press, New York.

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.

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

Wertz, D.H. & Scheraga, H.A. (1978). The influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules 11*, 9-15.

Williams, R.L., Vila, J., Perrot, G., & Scheraga, H.A. (1992). Empirical solvation models in the context of conformational energy searches. Application to bovine pancreatic trypsin inhibitor. *Proteins Struct. Funct. Genet. 14*, 110-119.

Zimmerman, S.S., Pottle, M.S., Némethy, G., & Scheraga, H.A. (1977). Conformational analysis of the twenty naturally occurring amino acid residues using ECEPP. *Macromolecules 10*, 1-9.