BUTTERWORTH HEINEMANN

Simulating the folding of small proteins by use of the local minimum energy and the free solvation energy yields native-like structures

Robert Brasseur

Centre de Biophysique Moléculaire Numérique, Faculté Universitaire de Gembloux, 5030 Gembloux, Belgium

Assuming that the protein primary sequence contains all information required to fold a protein into its native tertiary structure, we propose a new computational approach to protein folding by distributing the total energy of the macromolecular system along the torsional axes.

We further derive a new semiempirical equation to calculate the total energy of a macromolecular system including its free energy of solvation. The energy of solvation makes an important contribution to the stability of biological structures. The segregation of hydrophilic and hydrophobic domains is essential for the formation of micelles, lipid bilayers, and biological membranes, and it is also important for protein folding. The free energy of solvation consists of two components: one derived from interactions between the atoms of the protein, and the second resulting from interactions between the protein and the solvent. The latter component is expressed as a function of the fractional area of protein atoms accessible to the solvent.

The protein-folding procedure described in this article consists of two successive steps: a theoretical transition from an ideal α helix to an ideal β sheet is first imposed on the protein conformation, in order to calculate an initial secondary structure. The most stable secondary structure is built from a combination of the lowest energy structures calculated for each amino acid during this transition. An angular molecular dynamics step is then applied to this secondary structure. In this computational step, the total energy of the system consisting of the sum of the torsional energy, the van der Waals energy, the electrostatic energy, and the solvation energy is minimized. This process yields 3-D structures of minimal total energy that are considered to be the most probable native-like structures for the protein.

This method therefore requires no prior hypothesis about either the secondary or the tertiary structure of the protein and restricts the input of data to its sequence. The validity of the results is tested by comparing the crystalline and computed structures of four proteins, i.e., the avian and bovine pancreatic polypeptide (36 residues each), uteroglobin (70 residues), and the calcium-binding protein (75 residues); the C_α–C_α maps show significant homologies and the position of secondary structure domains; that of the α helices is particularly close.

Keywords: solvation energy, protein folding, secondary structure, free energy

INTRODUCTION

Simulation of protein folding by computational approaches aims at the calculation of compact three-dimensional native structure from the linear amino acid sequence of the protein.[1,2] Various methods have been described in the literature, including Monte Carlo techniques,[3] simulated annealing, and the so-called genetic algorithms.[4,5] The efficiency of these approaches can be increased by using a lower resolution representation of proteins and the structures obtained by Monte Carlo simulations share several of the major structural motives of globular proteins.

Another strategy for solving protein folding has been to smooth the energy profile to remove local minima and make the global minimum more accessible.[6]

Considering the side chains of polar amino acid residues as amphiphilic molecules consisting of a hydrophilic head group and a hydrophobic segment of variable length,[7] we developed a new strategy for the ab initio folding of pro-

Address reprint requests to Dr. Brasseur at the Centre de Biophysique Moléculaire Numérique, Faculté Universitaire de Gembloux, Passage des Déportés, 2, 5030 Gembloux, Belgium.
Received 28 March 1995; revised 14 May 1995; accepted 23 May 1995.

Journal of Molecular Graphics 13:312–322, 1995

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(95)00052-8

teins. The amphiphilic character of amino acid side chains is, in many respects, similar to that of lipids and detergents. Amphiphilic molecules can spontaneously assemble and organize themselves to form compact structures, whose hydrophobic components are closely packed in the core, while hydrophilic head groups point toward the aqueous phase. The segregation between hydrophobic and hydrophilic domains is the basic principle underlying the formation of biological membranes, lipid bilayers and micelles.[8] The same principle can be applied to the folding of soluble proteins.[9–18] Indeed, the structure of globular proteins consists of a hydrophobic core surrounded by a hydrophilic surface layer made of the polar head groups of some of the amino acids. In the course of the folding process, a protein evolves toward a compact structure characterized by a minimal area for the interface between the hydrophobic core and the hydrophilic surface. A stepwise analysis of the changes occurring at this interface provides a good approach for the unraveling of the protein-folding process.

To describe the changes occurring in the shape and surface of the hydrophobic/hydrophilic interface, we developed an "ab initio" computational approach. This approach consists of iterative processes and requires only the primary sequence of the protein as input data.

Computational methodology

A classic approach to the calculation of stable native-like protein structures is based on the minimization of the total energy of the system. The total energy is mostly calculated as the sum of the torsional energy and the van der Waals, electrostatic, and hydrophobic energies. In this article we propose four basic modifications to this type of calculation:

1. The total energy of the system is considered as the sum of the torsional energies along all axes of the system, the van der Waals energy, the electrostatic energy, and the solvation energy. We replaced the classic hydrophobic energy term by the solvation energy, between protein atoms and between these atoms and the solvent. A new semiempirical equation describing the solvation energy is derived in this article (see the next section).

2. The total energy of the system is assigned to the torsional axes of the protein (see Calculation of the Total Energy of the Protein, below).

3. The initial secondary structure of the protein, used for the subsequent angular molecular dynamics, is obtained from the local energy minima calculated for each amino acid during the transition from an ideal α-helix to a β-sheet structure (see Distribution of the Total Energy to the Torsional Axes, below).

4. The native-like protein structure is finally selected as the structure with the minimal total energy obtained through an angular molecular dynamics process (see Calculation of the Initial Secondary Structure through the "Local Energy Minima," below). In these calculations, the total energy of the system is expressed according to Eq. (5) (see the next section), and a quantum of the total energy is then assigned to each torsional axis [Eq. (9)]. The angular molecular dynamics is performed on all torsional angles while keeping constant the length of the atomic bonds and the value of the atomic angles.

The different steps of our computational approach are illustrated in Figure 1a.

Calculation of the solvation energy of the protein The calculation method is based on the concept that each atom of a condensed system is completely surrounded by its neighbors. The definition of the "solvent-accessible surface" of an atom, first introduced by different authors,[19–21] proposes that the free energy of transfer of a molecule, defined as the free solvation energy, is a linear function of the solvent-accessible surface of its atomic constituents. Colonna-Cesari and Sander[22] estimated the solvent-accessible surface of several molecules by calculating the area covered by a sphere representing a molecule of solvent, moving along the surface of these molecules. The radius of a molecule of solvent water can be estimated as equal to 1.4 Å.[22–24]

On the basis of this concept, we developed a semiempirical equation to quantify the free energy of solvation between atoms i and j:

$$E^{ij}_{\text{sol.in}} = \delta_{ij}(|E_{tri}f_{ij}| + |E_{trj}f_{ji}|)\exp[(r_i + r_j - d_{ij})/(2r_{\text{sol}})] \quad (1)$$

Where δ_{ij} is equal to -1 when the atoms i and j are both either hydrophobic or hydrophilic and to $+1$ otherwise, E_{tri} and E_{trj} are the free energies of transfer from a hydrophobic to a hydrophilic phase,[25,26] r_i and r_j are the radii of atoms i and j, d_{ij} is the distance between the atomic centers, and r_{sol} is the radius of a solvent molecule.

The values of E_{tri} and E_{trj} for the various atoms belonging to peptide residues are listed in Table 1. The solvation energy decreases exponentially with the ratio of the distance between the outer surface of atoms i and j and the diameter of the solvent molecule.[27,28] The parameter f_{ij} representing the portion of atom i covered by atom j is illustrated in Fig. 1a and is defined by Eq. (2):

$$f_{ij} = \frac{C_j}{S_{ij}}\left(1 - \frac{d_{ij} - r_i - r_j}{2r_{\text{sol}}}\right) = \frac{r_j^2}{4(r_i + r_j)^2}\left(1 - \frac{d_{ij} - r_i - r_j}{2r_{\text{sol}}}\right) \quad (2)$$

Where C_j is the surface of a circle of radius r_j, and S_{ij} is the surface of a sphere of radius $r_i + r_j$. In a condensed medium, the sum of the different fraction of atom i covered by its neighbors is equal to 1. The fraction of the surface covered equals f_{ij} $(0 < f_{ij} < 1)$ when the two atoms are in close contact. This parameter decreases to zero when the distance between the two atoms becomes large enough to accommodate a solvent molecule between them.

The calculation of the sum of all portions of the surface of an atom i covered by the other atoms (n) of the same

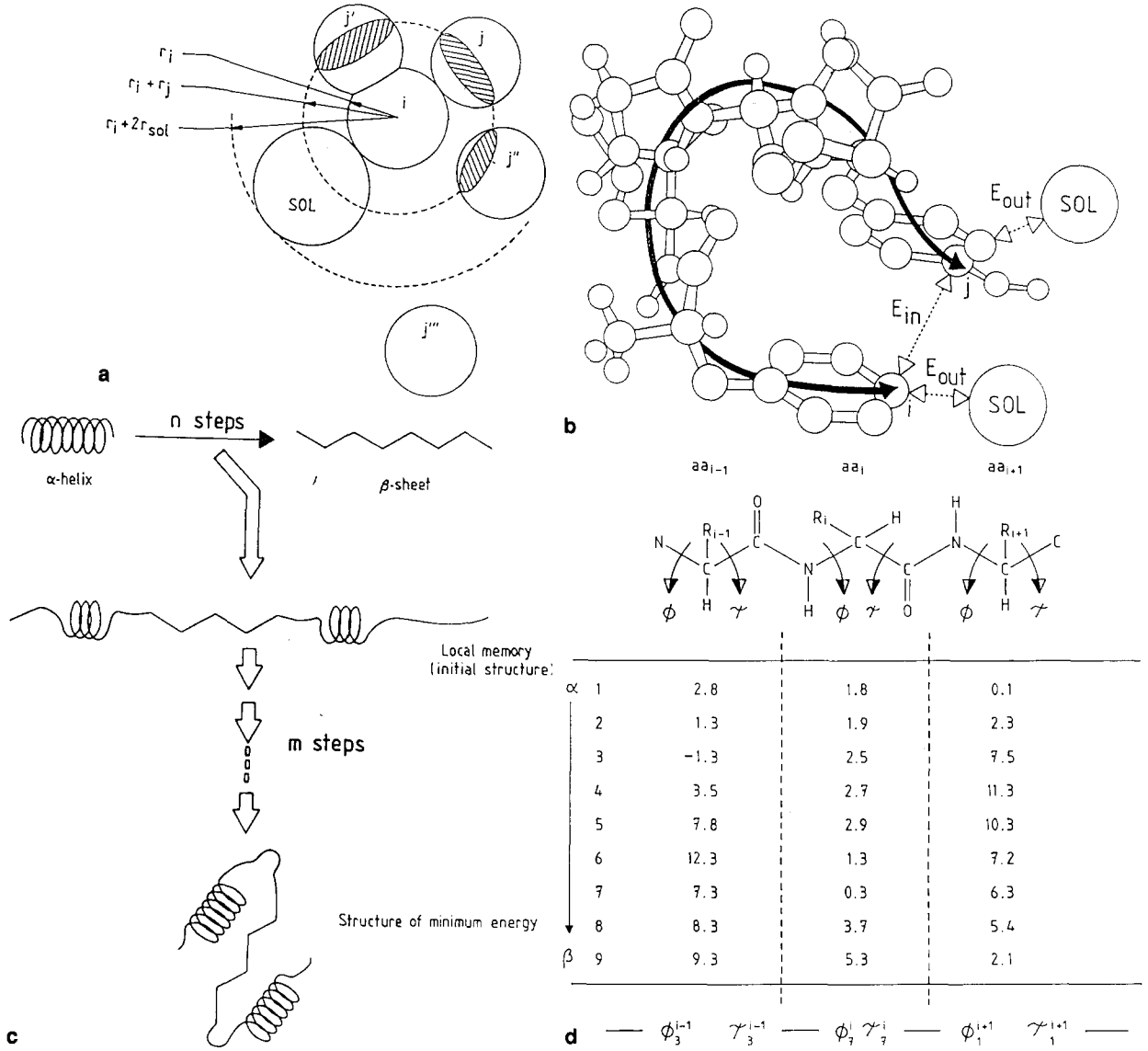

Figure 1. (a) Schematic representation of the covered fraction (hatched area) of the surface area of an atom. An atom i *interacting with an atom* j *can be either linked covalently (atom* j′*), in close contact (atom* j*), or at a distance either inferior (atom* j″*) or superior (atom* j‴*) to the diameter of a solvent molecule ($2r_{sol}$). (b) The interaction energy between atoms* i *and* j *(E_T) is equal to the sum of the van der Waals energy, the electrostatic energy, and the mutual solvation energies of atoms* i *and* j *(E_{in}) and of atom* i *and the solvent (E_{out}). The total energy of interaction is distributed along the backbone linking atoms* i *and* j. *An energy quantum, equal to the interaction energy, divided by the number of torsional axes between* i *and* j, *and supplemented by the torsional energy, is attributed to each torsional axis linking atoms* i *and* j. *Here the number of torsional axes is equal to 16. (c) Folding pathways for a protein molecule. During the first folding step, a conformational transition from an ideal helical structure to an ideal* β*-sheet structure is imposed on the protein in* n *step (*n *= 21). The initial structure for the molecular dynamics is hence obtained by assuming to the angle of each torsional axis the* ϕ *and* ψ *values corresponding to the local minimal energy during the* α*-helix →* β*-sheet transition. This can result in an excess of helical structure that gradually normalizes during the* M *steps of the molecular dynamics process (*M *is equal to three times the number of torsional axes). In the course of these* M *steps the structure with the minimal total energy is retained. (d) Schematic representation of how the initial structure used in molecular dynamics is constructed. The initial structure is built as a combination of the local minima reached during the 21-step transition from an ideal helical structure to an ideal* β *sheet (schematized as nine steps here). The* ϕ *and* ψ *angles of residues* i − *1,* i, *and* i + *1 are selected as those of steps 3, 7, and 1, respectively. The energy level calculated in (d) is the sum of the van der Waals and electrostatic energies and of the free energy of solvation associated with the angles* ϕ *and* ψ *of each amino acid residue.*

protein provides an estimation of the total surface covered. The solvent-accessible surface of this particular atom is obtained by subtraction of the surface covered from its entire surface. The ratio of the fraction of the solvent-accessible surface of atom *i* to the fraction of this atom surface covered by the solvent ($f_{j\,H_2O}$) yields the number of solvent molecules ($N^i_{H_2O}$) that are in close contact [Eq. (3)] and hence the solvation energy [Eq. (4)]

Table 1. Values of the free energy of transfer for individual atoms

Atoms	E_{tr} (KJ/mol)
C_{sp^2}	−9.40
C_{sp^3}	−10.20
H (q = 0)	−1.22
H (q/0)	4.31
O	11.83
S	−11.50
N	12.66
H_2O_{sol}	20.44

C_{sp^2} and C_{sp^3} are carbon atoms with a double and single bond, respectively. H (q = 0) and H (q/0) represent uncharged and charged hydrogen atoms. H_2O_{sol} represents the free energy of transfer for a solvent molecule.

$$N^i_{H_2O} = \frac{1 - \sum_{j=1}^{n} f_{ij}}{f_i\ H_2O} \tag{3}$$

$$E^i_{sol.out} = N^i_{H_2O} * E_{tr.H_2O} \tag{4}$$

where $E_{tr\ H_2O}$ is the free energy of transfer for the solvent molecule (Table 1).

Calculation of the total energy of the protein The total interaction energy [Eq. (5)] between atoms i and j (E_T) is equal to the sum of the torsional energy (E_{tor}), the van der Waals energy (E_{vdw}), the electrostatic energy (E_{elec}), and the solvation energy between atoms ($E_{sol.in}$) and between atoms and solvent ($E_{sol.out}$):

$$E_T = E_{tor} + E_{vdw} + E_{elec} + E_{sol.in} + E_{sol.out} \tag{5}$$

The London–van der Waals energy of interaction between all pairs of non-mutually-bonded atoms Buckingham's equation describing pairwise atom–atom interactions has been used to calculate the van der Waals energy:

$$E^{ij}_{vdw} = A_{ij} \exp(-B_{ij}d_{ij}) - C_{ij}d^{-6}_{ij} \tag{6}$$

where i,j = 1,2, . . . , n are the nonbonded atoms, d_{ij} their mutual distance, and A_{ij}, B_{ij}, and C_{ij} are coefficients assigned to atomic pairs. Values were previously reported for these coefficients.[29,30] They were successfully applied to the conformational analysis of molecular crystals, proteins, polypeptides, and lipids.[8] To compensate for the decrease of the E_{vdw} function over short distances d_{ij} ($d_{ij} < 0.1$ nm), we have imposed an arbitrary cutoff value of

$$E^{ij}_{vdw} = 418.4 \text{ kJ/mol} \quad \text{at} \quad d_{ij} < 0.1 \text{ nm}$$

Thanks to this correction the van der Waals energy calculated according to the classic Lennard–Jones function closely coincides with that calculated through Eq. (6).

The electrostatic interaction between atomic point charges within a distance range comparable to the size of a molecule the atomic point charge distribution should be used to calculate electrostatic interaction. The electrostatic energy can be written as follows:

$$E^{ij}_{elec} = 139.2 \left(\frac{e_i e_j}{d_{ij}\epsilon_{ij}} \right) \tag{7}$$

where e_i and e_j are expressed as electron charge units and d_{ij} as angstroms. ϵ_{ij} is the dielectric constant. To simulate the protein interface, we assumed a dielectric constant equal to 1.2 for the hydrophobic domain and a dielectric constant of 30 for the atoms more in contact with the aqueous phase. The dielectric constant was assumed to increase linearly between these two domains.

The potential energy of rotation of the torsional angles-The rotation around the C–C or C–O bonds was calculated according to Eq. (8):

$$E(k)_{tor} = \frac{U(k)}{2} \{1 + \cos[p\ x(k)]\} \tag{8}$$

where $U(k)$ corresponds to the energy barrier in the eclipsed conformation during the rotation of the k angle, p is the periodicity of the function, and $x(k)$ is the value of the torsional k angle. $U(k)$ is equal to 11.7 kJ/mol for the C–C bond and 7.5 kJ/mol for the C–O bond.

Distribution of the total energy to the torsional axes The total energy of interaction between atoms i and j is assigned to the shortest pathway between these two atoms, which represents the succession of the covalent bonds linking all atoms separating i and j in the sequence (Fig. 1b). An energy quantum (q), equal to the interaction energy between atoms i and j, divided by the number of torsional axes between i and j, is distributed among all torsional axes linking atoms i and j. The energy of each axis is supplemented by its own torsional energy [Eq. (8)]. The total energy associated with each torsional axis represents therefore the sum of the torsional energy, the interaction energy, and the solvation energy for the atoms at the extremities of this torsional axis. The energetic component $E(k)_N$, associated to the k angle of the torsional axis, between atoms w and w + 1, is defined as

$$E(k)_N = E(k)_{tor} + \sum_{i=1}^{w} \sum_{j=w+1}^{n} q(E^{ij}_{vdw} + E^{ij}_{elec} + E^{ij}_{sol.in}) + \sum_{i=1}^{w} qE^i_{sol.out} + \sum_{i=w+1}^{n} qE^i_{sol.out} \tag{9}$$

consisting of the torsional energy [$E(k)_{tor}$, Eq. (8)], a quantum of the van der Waals energy [qE_{vdw}, Eq. (6)], a quantum of the electrostatic energy [qE_{elec}, Eq. (7)], and a quantum of the solvation energy [$qE_{sol.in}$,$qE_{sol.out}$, Eqs. (1) and (4)].

Calculation of the initial secondary structure through the "local energy minima" The selection of the initial structure for the subsequent angular molecular dynamics was based both on theoretical and experimental considerations. Previous authors had proposed that both helices and turns can play a role in initiating the protein-folding process,[31–35] while hydrophobic cluster analysis[36,37] enabled the detection of hydrophobicity clusters within helical segments of a protein.

As the contact area between protein and solvent is min-

imal for an α helix and maximal for a β sheet, we imposed a theoretical transition to the protein, from an ideal α-helix to a β-sheet conformation. This conformational change was carried out in 21 steps, as this value represents a compromise between a too rapid transition and a too large number of transient conformations. At each step of the transition, we calculated the energy level for each amino acid residue as the sum of the van der Waals and electrostatic energies and of the free energy of solvation associated with the angles ϕ and ψ of each residue.

To build the initial secondary structure required for the subsequent angular molecular dynamics process, we then selected for each amino acid the minimal energy conformation observed during the 21-step α-helix → β-sheet transition. These conformations were selected for each residue separately and they correspond to the local minimal energy, the "local memory" associated with the α-helix → β-sheet transition (Fig. 1c). In this resulting secondary structure, the value assigned to the angles ϕ and ψ of each torsional axis is that of the lowest local energy encountered during the conformational transition.

The method used to build this initial structure is schematically represented in Fig. 1d for a tripeptide. This peptide consists of residues $i - 1$, i, and $i + 1$, which were randomly selected from the sequence of the calcium-binding protein. The values of the angles ϕ and ψ, with a local minimal energy were encountered at step 3 of the transition for residue $i - 1$, step 7 for residue i, and at step 1 for residue $i + 1$. The initial secondary structure for the molecular dynamics calculations represents therefore a combination of these three minima.

Using this approach, an initial secondary structure for the molecular dynamics calculations can thus be built for any protein sequence, by combining all local energy minima encountered during the α-helix → β-sheet transition.

Calculation of the "native-like" protein structure through angular molecular dynamics To analyze the folding process of a protein from the initial secondary structure defined above, toward its "native-like" structure, we carried out an angular molecular dynamics calculation. In this molecular dynamics computational method, both the length of the atomic bonds and the value of the atomic angles are kept constant. The molecular dynamics approach is applied to all torsional angles of the protein.

The angular molecular dynamics approach was performed in M steps, according to Eq. (10). This equation expresses the value of the torsional angle $x(k)$ as a function of the number of computational steps. Equation (10) was derived from that proposed by Verlet,[35] where the angle $x(k)$ was varied as a function of time. In Eq. (10) $x(k)$ varies in a discrete way ($\delta N = 1$) as a function of the number of calculation steps. For each calculation step N ($1 < N < M$), the value of the torsional angles is calculated as

$$x(k)_{N+\delta N} = 2x(k)_N - x(k)_{N-\delta N} + \delta N^2 \frac{1}{m_r(k)} \frac{[E(k)_{N-\delta N} - E(k)_N]}{[x(k)_N - x(k)_{N-\delta N}]} \quad (10)$$

where $E(k)_N$ is the energy associated to the angle k [Eq. (9)]. $m(k)$r is the reduced mass calculated from the mass of the atoms at the extremities of a particular torsional axis linking atoms w and $w + 1$ by Eq. (10):

$$m_r(k) = \frac{m_{1\ldots w} m_{w+1\ldots n}}{m_T} \quad (11)$$

where $m_{1\ldots w}$ and $m_{w+1\ldots n}$ are the atomic masses at the two extremities of this axis and m_T is the total mass of the protein made of n atoms.

Each step of the molecular dynamics process includes the calculation of the total energy of the system according to Eq. (5), as the sum of the energies for each atom of the protein. This total energy is further equal to the sum of the energies of all torsional axis [$E(k)_N$]:

$$E_T = \sum_{k=1}^{X} E(k)_N \quad (12)$$

Where X is the total number of torsional axes.

The angular molecular dynamics tends toward structures with a minimal total energy for the entire system, including atom–atom and atom–solvent interactions, which are optimized through the stepwise computational procedure.

During the angular molecular dynamics, the rotational velocity, expressed as degree per step, is calculated according to Equation (13).

$$V_r(k) = \frac{x(k)_{N+\delta N} - x(k)_N}{\delta N} \quad (13)$$

This rotational velocity [$V_r(k)$] is arbitrarily limited to 1.5° for each step of the angular molecular dynamics. As short-distance interactions between atoms are associated with high interaction energies, high velocities might cause the system to diverge during the calculations. Limiting the rotational velocity at each step simulates the viscosity of the system and the interactions of the solvent molecules with the protein.

Moreover, we observed that in this procedure a small random component with a value between −0.5 and 0.5° [$md(-0.5,0.5)$] had to be added to each torsional angle at every step $N + \delta N$ of the molecular dynamics calculation as

$$x(k)_{N+\delta N} = x(k)_{N+\delta N} + md(-0.5,0.5) \quad (14)$$

Taking into account this equation, the apparent dynamic velocity is directly related to the energy of the system while the random value arises from the thermal motion. This random parameter enables one to cross gaps between energy levels during the calculations, which would otherwise lead to unstable structures for the protein. When this random component is omitted, these calculations yield stable structures that are, however, only distantly related to the actual crystallographic structures. On the contrary, when the random value component exceeds 10°, disordered structures are obtained.

The number of computational steps (M) used in the molecular dynamics process was tested empirically and an optimal value equal to three times the number of torsional axes of the molecule was selected for all proteins tested. This value is a compromise between goodness of fit between computed and X-ray crystallographic structures and the duration of the molecular dynamics calculations.

The molecular dynamics calculations were carried out at least 100 times for the proteins illustrated in this article,

yielding different solutions at each step owing to the introduction of the random component for the angular velocity. After 100 successive calculations, the structure with the lowest energy is selected among the homologous structures as being the optimal solution. The goodness of fit between computed and crystallographic structures is estimated by calculating the root mean square deviation (RMS) between the coordinates of the C_αs in the two types of structure.[38]

RESULTS

We developed a new ab initio method for calculating protein folding. This method is applicable to proteins with fewer than 150 residues, consisting of a single amino acid chain. In this article we describe and analyze the results obtained for four proteins: the avian and protein pancreatic polypeptides (which in spite of differences in their sequences have homologous three-dimensional (3-D) structures), and uteroglobin and the calcium-binding protein (which both have a high percentage of α-helical structure).

The secondary structure of a protein is characterized by the value of the ϕ and ψ angles of the polypeptide backbone. A comparison of the experimental and computed values of the angles ϕ and ψ along the sequence of the avian pancreatic polypeptide and of the calcium-binding protein is shown in Figures 2 and 3. These two figures demonstrate that for 75% of the amino acid residues, the difference between the pairs of ϕ and ψ angles does not exceed 30°, thus stressing the similarity between computed and crystallographic structure. Larger differences between computed and measured structures were, however, observed for the N-terminal region and for the last C-terminal residues of the avian pancreatic polypeptide. Differences also appeared at the N-terminal residues of the three last helices of the calcium-binding protein (residues 21–25, 42–46, and 60–64). In spite of these differences, the agreement between calculated and measured 3-D structures of this protein is satisfactory (see below).

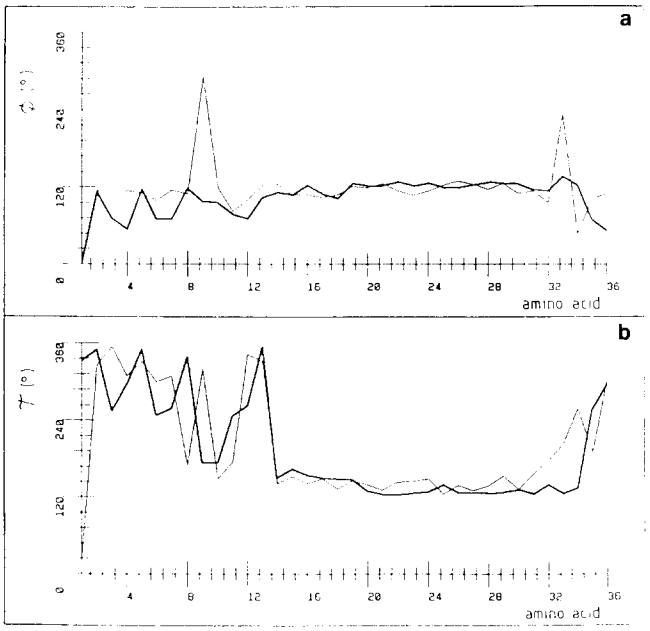

Figure 2. ϕ (a) and ψ (b) torsional angle values plotted along the amino acid sequence of the avian pancreatic polypeptide. The thin line represents the crystallographic structure, the thick line the computed structure.

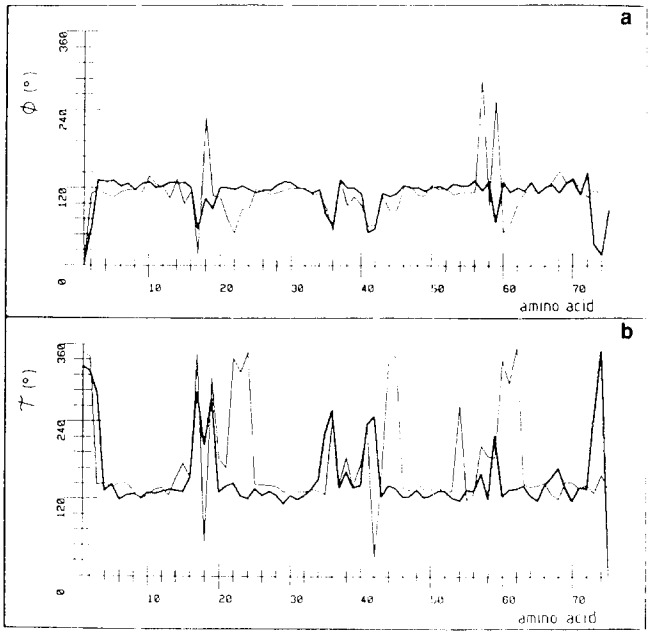

Figure 3. ϕ (a) and ψ (b) torsional angle values plotted along the amino acid sequence of the calcium-binding protein. The thin line represents the crystallographic structure, the thick line the computed structure.

The crystallographic representation of the polypeptide backbone[39,40] together with a ribbon drawing of the computed structure of the avian and bovine pancreatic polypeptides are depicted in Figures 4a and 5a. The crystallographic and computed C_α–C_α maps of these same proteins are compared in Figures 4b and 5b.

The ribbon and backbone stereorepresentations of the computed and crystalline structures (Figures 4a and 5a) demonstrate in both cases three major domains in the two proteins. The N-terminal domain (residues 2–8) is in an extended form, followed by a turn (residues 10–13), while the C-terminal domain (residues 14–32) has an helical conformation. The conformation of the C- and N-terminal residues in the computed structure is less regular than in the crystal. The interactions between the N- and C-terminal domains appear therefore less clearly in the computed structure of the avian pancreatic polypeptide. The RMS values between crystallographic and computed structures are, respectively, 2.4 and 2.6 Å for the avian and bovine pancreatic polypeptide, demonstrating the close similarity of the structures.

For the same polypeptides, a comparison of the C_α–C_α distances between amino acid residues in the crystallographic (upper left) and the computed structure (lower right) (Figures 4b and 5b) shows similar interactions between residues along the sequences. For the avian pancreatic polypeptide, residues 8 and 20, and residues 4 and 24, interact both in the crystal and in the computed structure. Same residues interact also in the bovine avian pancreatic polypeptide.

Figure 6a and b shows, respectively, the ribbon repre-

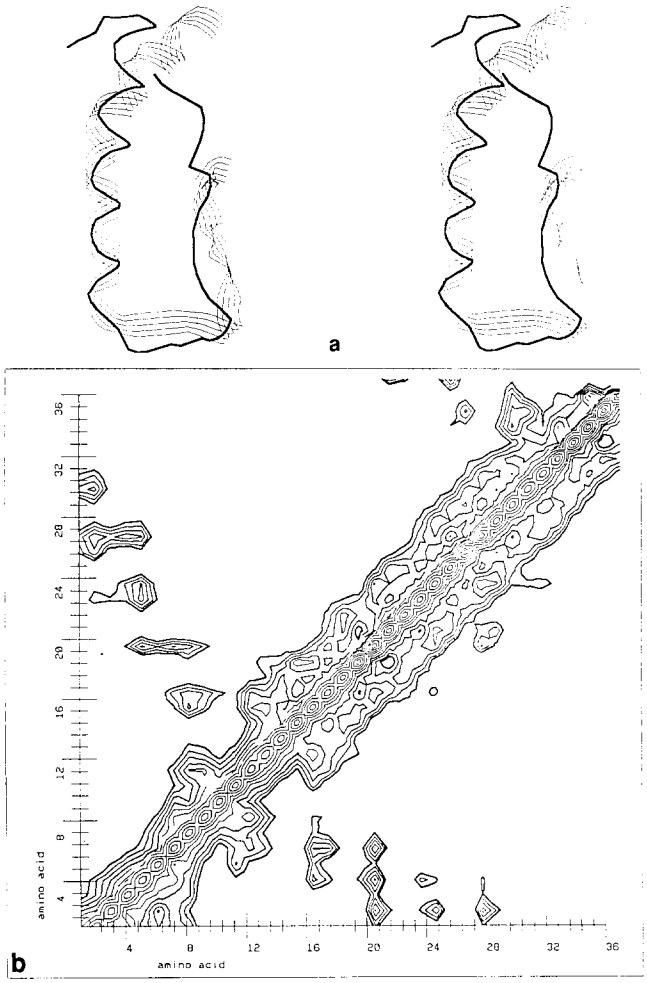

Figure 4. (a) Spatial fitting stereoview of the crystallographic backbone (thick full line) and the computed ribbon structures of the avian pancreatic polypeptide (RMS = 2.4 Å). (b) C_α–C_α contact maps of the avian pancreatic polypeptide: the crystallographic map is shown at upper left, and the computed map at lower right.

sentation and the C_α–C_α map of the crystallographic[41] and computed structure of uteroglobin (70 residues). The crystallographic structure of the dimeric form of crystallized uteroglobin is made of four interacting α helices separated by short extended sequences. The results of our calculations, carried out on the monomeric form of uteroglobin, show four helices interrupted by segments that tend to adopt an extended conformation. The structure computed for the uteroglobin monomer is therefore less compact but resembles that of the monomer in the crystalline dimeric structure. The RMS values between computed and crystalline structure amount to 5.6 Å.

The percentage of random coil structure in the N- and C-terminal domains of the uteroglobin molecule is lower in the computed than in the crystalline structure. Indeed, in the structure calculated by the local minimum energy approach, some residues located close to prolines (Pro-14, -18, -30, -49, and -67) take on a conformation intermediate between an α helix and a β sheet (Figure 7). During the local minimum energy calculation, steric contacts between residues close to prolines impair the formation of a stable helical structure and the prolines therefore appear as "helix breakers." Four helices are finally obtained: helix H1, spanning residues 6 to 15; helix H2, between residues 20 and 27; helix H3, encompassing residues 33 to 47; and helix H4, at residues 52 to 66.

To reach this final structure, two pathways can be proposed for the folding of uteroglobin: interactions can either occur first between helices H1 and H2, followed by close contacts with helix H3, and ultimately followed by contacts between helix H4 and the cluster formed by the three first helices. Alternatively, helices H2 and H3 come close to each other first, and thereafter H1 and H4 come simultaneously in contact with the cluster formed by H2 and H3.

The crystallographic backbone[42] and the computed ribbon structure of the calcium-binding protein are represented in a stereoview in Figure 8a. This protein consists of a bundle of four interacting helices. The location and length of these helices are comparable in the two structures: in the crystallographic structure the helices span residues 2–16,

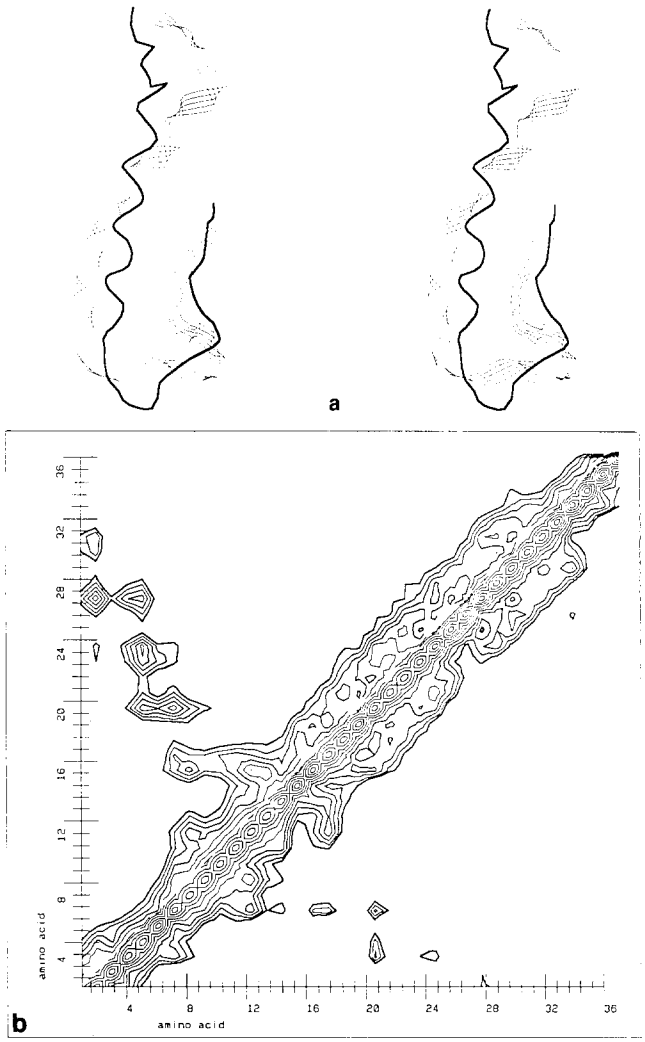

Figure 5. (a) Spatial fitting stereoview of the crystallographic backbone (thick full line) and the computed ribbon structures of the bovine pancreatic polypeptide (RMS = 2.6). (b) C_α–C_α contact maps of the bovine pancreatic polypeptide: the crystallographic map is shown at upper left, and the computed map at lower right.

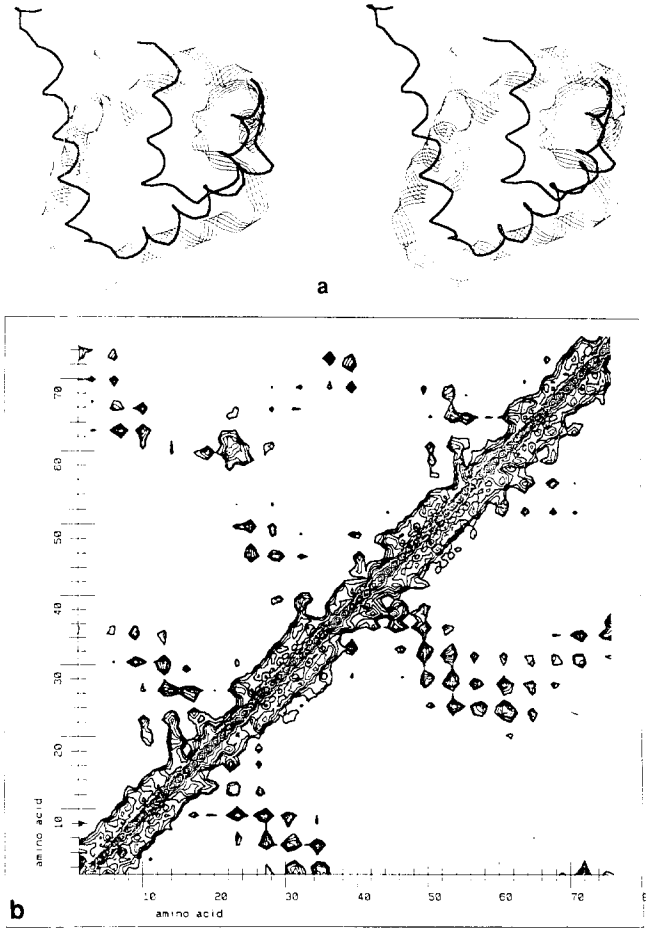

Figure 6. (a) Spatial fitting stereoview of the crystallographic backbone (thick full line) and the computed ribbon structures of uteroglobin (RMS = 5.6). (b) C_α–C_α contact maps of uteroglobin: the crystallographic map is shown at upper left, and the computed map at lower right.

24–36, 45–54, and 62–75, compared to the corresponding residues 2–16, 21–35, 43–54, and 60–73 in the computed structure. The interactions between these helices are illustrated in the C_α–C_α maps in Figure 8b. A comparison of the C_α–C_α distances between amino acid residues for the crystallographic (upper left) and computed structures (lower right) in Figure 8b shows good similarity. Residues 10 and 30, 1 and 74, and 54 and 68 interact in the crystal compared to residues 9 and 31, 1 and 72, and 56 and 68 in the computed structure. The differences observed between crystallographic and computed structures do not affect the general 3-D topology of the molecule, as the interacting domains are close in both structures. Moreover, the RMS values between the two structures are of the same order of magnitude as for the other proteins and are equal to 4.5 Å.

Besides the proteins illustrated in this article, we calculated the structure of a number of other proteins, including the L7 ribosomal protein, the amino-terminal domain of the 434 repressor, the cro protein, and the che*Y protein. This program is momentarily restricted to proteins of fewer than 150 residues, owing to the length of computation time on a computer equipped with a Pentium processor, as the computation time increase as N^3 residues. This method applies only to proteins without disulfide bridges, as the pathway between atoms cannot be non-unambiguously defined in these proteins with S–S bonds which form cyclic domains. We obtained a good agreement between crystallographic and computed structures for all computed proteins, with the exception of ubiquitin and other β sheet-rich proteins, suggesting that our method underestimates β-sheet secondary structures.

CONCLUSIONS

In this article we applied an angular molecular dynamics approach to the analysis of protein folding, in an ab initio method requiring the protein sequence as sole input data. In this method, the energy of solvation is calculated for all atomic constituents of the protein, and added to the van der Waals, electrostatic, and torsional energy to calculate the total energy of the system.

This approach was applied to a number of small proteins (fewer than 150 residues), yielding computed structures with a high level of homology with the crystallographic structure, as illustrated in this article for four different pro-

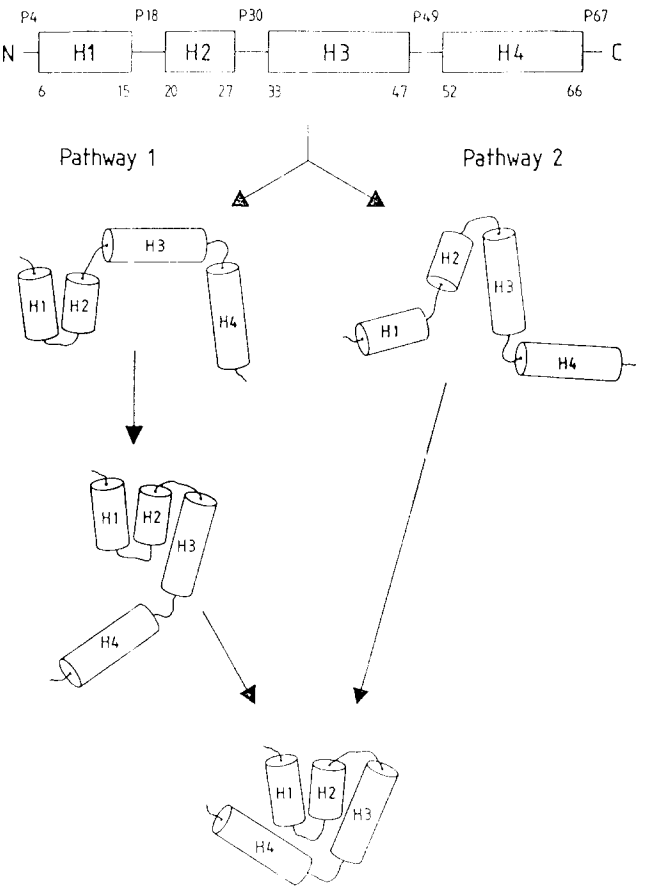

Figure 7. Schematic representation of the putative pathways for uteroglobin folding. Top: *Structure calculated from the initial step. Four helices (H1 to H4) are separated by stable nonhelical structures owing to the presence of prolines. Helix H1 spans residues 6 to 15, helix H2 spans residues 20 and 27, helix H3 encompasses residues 33 to 47 and helix H4 spans residues 52 to 66. According to the calculations, protein folding can occur along two possible pathways (see text).*

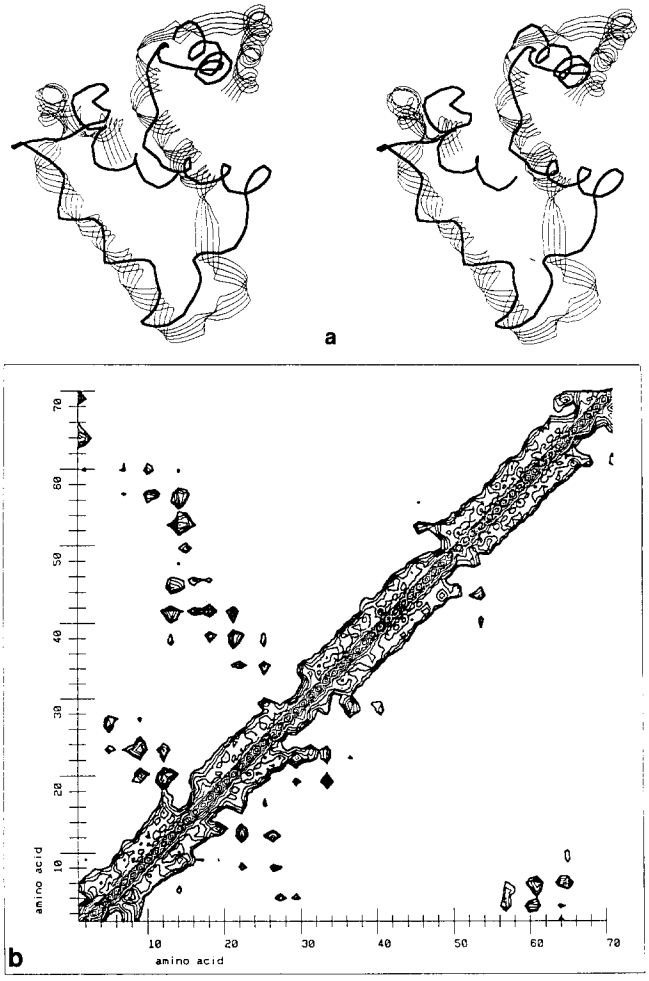

Figure 8. (a) Spatial fitting stereoview of the crystallographic backbone (thick full line) and the computed ribbon structures of the calcium-binding protein (RMS = 4.5). (b) C_α–C_α contact maps of the calcium-binding protein: the crystallographic map is shown at upper left, and the computed map at lower right.

teins. This method is, however, restricted to proteins containing no disulfide bridge and with a low β-sheet content.

This method consists of two steps: first the calculation of the local minimal energy for each amino acid along the sequence during an α-helix → β-sheet transition, followed by an angular molecular dynamics step. This enables the visualization of possible folding pathways for the protein, including the formation of intramolecular clusters and domains. The computational approach finally yields the tertiary structure of the protein, consisting of separate domains with a defined secondary structure. These secondary structure domains are mainly helical and their positions match closely those of the corresponding regions detected by X-ray crystallography. Moreover, short-distance contacts between amino acids, separated by a large number of residues in the primary sequence, appear in the interdistance C_α–C_α maps, suggesting that the simulation process yields native-like tertiary structures.

Several attempts to simulate protein folding have been carried out by an exhaustive search of the lattice conformation of small proteins.[43] This was done by perturbing the conformational surface energy of the ϕ and ψ angles of the peptides, using a force field,[44,45] and by Monte Carlo techniques, either standard[46] or modified according to the method of Metropolis.[47] This last method seems more efficient than the Metropolis strategy, as extensively described by Dill.[10] Several groups have furthermore used the inverse folding approach to examine the chances for a given sequence to adopt a given 3-D structure. A generalization of the inverse folding algorithm to predict superstructural elements, including β and helical hairpins, and α–β–α fragments, has been proposed.[48] Previous attempts[23,48] to calculate protein folding did not include, or only partially included, the calculation of the energy of solvation for all atoms of the system, that is, for both those of the protein and of the solvent. In our expression, when calculations were performed without the free solvation energy term or when using a randomized primary sequence for the protein, no agreement could be reached between computed and crystalline structures.

This method has, moreover, two significant advantages in terms of computational time and power required for this protein-folding analysis. In the first step, the rapid search for a stable secondary structure consists of a combination of the local energy minima. This approach therefore reduces the search for a single energy minimum in a hyperspace with N dimensions to that of a combination of separate energy minima in N single-dimension spaces. The degrees of freedom for the ϕ and ψ angles can therefore vary only between the α-helical and β-sheet values. This technique significantly decreases the size of the calculation space and the duration of the calculation time.

The same holds true for the subsequent angular molecular dynamics step. Here, only the torsional angles are varied whereas the position of the atoms and the length of the atomic bonds are varied in a classic molecular dynamics approach.

The present approach, using the solvation energy term, was developed to mimick as closely as possible the available experimental observations describing protein folding.[33] These include a rapid appearance of defined secondary structures at the beginning of the folding process, followed by interdomain interactions. Our method is currently developed to include a classic molecular dynamics step after the angular molecular dynamics, so as to refine further the computed structure and increase the level of comparability with the crystallographic data.

ACKNOWLEDGMENTS

R.B. is Directeur de Recherche of the Belgian Fonds National de la Recherche Scientifique. I am grateful to Drs. M. Rosseneu, A. Burny, A. Schanck, J.M. Ruysschaert, A. Soumarnon, L. Lins, and J.P. Mornon for helpful suggestions about the manuscript. I thank E. and Y. Saey, B. Delplace, P. Duquenoy, and C. Stil for their support. I am grateful to the Association Française de Lutte contre la Mucoviscidose for financial assistance.

REFERENCES

1 Chan, H.S. and Dill, K.A. The protein folding problem. *Physics Today* February 1993, 24–32

2 Chothia, C. Principles that determine the structure of proteins. *Annu. Rev. Biochem.* 1984, **53,** 357–372

3 Skolnick, J., Kolinski, A., and Yaris, R. Monte Carlo simulations on an equilibrium globular protein folding model. *Proc. Natl. Acad. Sci. U.S.A.* 1989, **86,** 1229–1233

4 Dandekar, T. and Argos, P. Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.* 1994, **236,** 844–861

5 Jones, D.T. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.* 1994, **3,** 567–574

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

7 Creighton, T.E. In: *Proteins: Structures and Molecular Properties*. Freeman, New York, 1993

8 Brasseur, R. In: *Molecular Description of Biological Membranes by Computer Aided Conformational Analysis* (Brasseur, R., ed.), Vols. I and II. CRC Press, Boca Raton, Florida, 1990

9 Creighton, T.E. Stability of folded conformations. *Curr. Opin. Struct. Biol.* 1991, **1,** 5–16

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

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

12 Kim, P.S. and Baldwin, R.L. Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.* 1990, **59,** 631–660

13 Moult, J. and Unger, R. An analysis of protein folding pathways. *Biochemistry* 1991, **30,** 3816–3824

14 Pace, C.N. Contribution of the hydrophobic effect to globular protein stability. *J. Mol. Biol.* 1992, **226,** 29–35

15 Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. *Science* 1985, **229,** 834–838

16 Tanford, C. In: *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*. John Wiley & Sons, New York, 1973

17 Teeter, M.M. Water–protein interactions: Theory and experiment. *Annu. Rev. Biophys. Chem.* 1991, **20,** 577–600

18 Zaccai, G. and Eisenberg, H. Halophilic proteins and the influence of solvent on protein stabilization. *Trends Biochem. Sci.* 1990, **15,** 333–337

19 Eisenberg, D. and McLachlan, A.D. Solvation energy in protein folding and binding. *Nature* (*London*) 1986, **319,** 199–203

20 Jones, D.T., Taylor, W.R., and Thornton, J.M. A new approach to protein fold recognition. *Nature* (*London*) 1992, **358,** 86–89

21 Lee, B.K. and Richards, F.M. The interpretation of protein structure: Estimation of static accessibility. *J. Mol. Biol.* 1971, **55,** 379–400

22 Colonna-Cesari, F. and Sander, C. Excluded volume approximation to protein–solvent contact model. *Biophys. J.* 1990, **57,** 1103–1107

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

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

25 Brasseur, R. Differentiation of lipid-associating helices by use of three-dimensional hydrophobicity potential calculations. *J. Biol. Chem.* 1991, **266,** 16120–16127

26 Brasseur, R., Lins, L., Vanloo, B., Ruysschaert, J.M., and Rosseneu, M.Y. Molecular modeling of amphipathic helices of the plasma apolipoproteins. *Proteins Struct. Funct. Genet.* 1992, **13,** 246–257

27 McIntosh, T.J. and Simon, S.A. Hydration forces and bilayer deformation: A revaluation. *Biochemistry* 1986, **25,** 4058–4066

28 Israelachvili, J. and Pashley, RM. The hydrophobic interaction is long range, decaying exponentially with distance. *Nature* (*London*) 1982, **300,** 341–342

29 Giglio, E., Liquori, A.M., and Mazzarella, L. van der Waals interactions and the packing of molecular crystals. IV. Orthorhombic sulfur. *Nuovo Cimento* 1968, **56,** 57–59

30 Liquori, A.M., Giglio, E., and Mazzarella, L. van der Waals interactions and the packing of molecular crystal. II. Adamantane. *Nuovo Cimento B* 1968, **55,** 475–477

31 Brooks, C.L. Molecular simulations of peptide and protein unfolding: In quest of a molten globule. *Curr. Opin. Struct. Biol.* 1993, **3,** 92–98

32 Creighton, T.E. Up the kinetic pathway. *Nature* (*London*) 1992, **356,** 194–195

33 Dobson, C.M. Resting places on folding pathways. *C.M. Curr. Biol.* 1992, **2,** 343–345

34 Radford, S.E., Dobson, C.M., and Evans, P.A. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* (*London*) 1992, **358,** 302–307

35 Verlet, L. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lenard–Jones molecules. *Phys. Rev.* 1967, **159,** 98–105

36 Gaboriaud, C., Bissery, V., Benchetrit, T., and Mornon, J.P. Hydrophobic cluster analysis: An efficient new way to compare and analyse amino acid sequences. *FEBS Lett.* 1987, **224,** 149–155

37 Lemesle-Varloot, L., Henrissat, B., Gaboriaud, C., Bissery, V., Morgat, A., and Mornon, J.P. Hydrophobic cluster analysis: Procedures to derive structural and functional information from representation of protein sequence. *Biochimie* 1990, **72,** 555–574

38 Lesk, A. In: *Protein Architecture: A Practical Approach* (Rickwood, D. and Harnes, B. eds.). IRL Press, Oxford, 1991

39 Glover, I., Hanneef, I., Pitts, J., Wood, S., Moss, Tickle, I., and Blundell, T. Conformational flexibility in a small globular hormone. X-Ray analysis of avian pancreatic polypeptide at 0.98 Å resolution. *Biopolymers* 1983, **22,** 293–304

40 Li, X., Sutcliffe, M., Schwartz, T., and Dobson, C.M. Sequence-specific 1H-NMR assignments and solution structure of bovine pancreatic polypeptide. *Biochemistry* 1992, **31,** 1245–1248

41 Morize, I., Surcouf, E., Vaney, M.C., Epelboin, Y., Buehner, M., Fridlansky, F., Milgrom, E., and Mornon, J.P. Refinement of the C222 1 crystal form of oxidized uteroglobin at 1.34 Å resolution. *J. Mol. Biol.* 1987, **194,** 725–731

42 Szebenyi, D. and Moffat, K. The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine. *J. Biol. Chem.* 1986, **261,** 8761–8763

43 Covell, D. and Jernigan, R. Conformations of folded proteins in restricted spaces. *Biochemistry* 1990, **29,** 3287–3294

44 Kostrowicki, J. and Scheraga, H.A. Application of the diffusion equation method for global optimization in oligopeptides. *J. Phys. Chem.* 1992, **96,** 7442–7449

45 Stillinger, F.H. Role of the potential-energy scaling in the low-temperature relaxation behavior of amorphous materials. *Phys. Rev.* 1985, **32,** 3134–3141

46 Bouzida, D., Kumar, S., and Swendsen, R. Efficient Monte Carlo methods for the computer simulation of biological molecules. *Phys. Rev.* 1992, **45,** 8894–8901

47 O'Toole, E. and Pangiotopoulos, A.Z. Monte Carlo simulation of folding conditions of simple lattice model proteins using a chain growth algorithm. *J. Chem. Phys.* 1992, **97,** 8644–8651

48 Godzik, A., Skolnick, J., and Kolinski A. Simulations of the folding pathways of triose phosphate isomerase-type α–β barrel proteins. *Proc. Natl. Acad. Sci. U.S.A.* 1992, **89,** 2629–2633