A Simplified Amino Acid Potential for Use in Structure Predictions of Proteins

A. Wallqvist and M. Ullner
Physical Chemistry 2, Chemical Center, University of Lund, 221 00 Lund, Sweden

ABSTRACT **A simplified description and a corresponding force field for polypeptides is introduced. Each amino acid residue is reduced to one interaction site, representing the backbone, and one or two side chain sites depending on its size and complexity. Site–site interactions are parameterized after a hydrophobicity criterium. The treatment of backbone sites is in addition designed to reproduce typical polypeptide hydrogen bonding patterns, as well as yielding conformations in accord with the allowed ϕ and ψ angles through an effective angle potential. There are no explicit charges in the model. The derived energy functions, which are based on thermodynamic data and sterical consideration of allowed backbone conformations, correspond to the introduction of an effective potential. The model is tested on two small proteins, avian pancreatic polypeptide and a parathyroid hormone-related protein, by simulating folding from an initially extended state using Monte Carlo methods. The reduced amino acid description is able to satisfactorily reproduce the experimentally determined native structures.** © 1994 Wiley-Liss, Inc.

Key words: hydrophobicity, effective backbone interactions, folding, avian pancreatic polypeptide, parathyroid hormone-related protein, Monte Carlo

INTRODUCTION

A long standing goal of biochemistry is to understand the apparently unique coding of the complete three-dimensional protein structure in the amino acid sequence.[1–5] The study of secondary and tertiary structures in proteins is complicated by the fact that the sequence contains not only the structural information but also the biological and chemical function. Function and structure are thus interconnected. The task is somewhat simplified in that some structural motifs, on both the secondary and tertiary levels, are quite common and appear to have been preserved by evolution. These observations have been used to formulate empirical rules or propensity parameters that predict participation in α-helices, β-sheet or turns for individual as well as sequences of amino acids. However, such rules does not contain the understanding of the molecular nature of forces that are important for protein structure, stability, and function.

Advances in computational chemistry have led to increased investigation using model systems[6–10] of the molecular nature of amino acids and its effect on protein folding and protein structure. The study of large assemblies of molecules using an atomistic description is often hampered by the sheer number of atoms involved and the computational inability to adequately sample the relevant degrees of freedom. This is definitely a real problem in the formation of secondary and tertiary structure elements in proteins. Experimentally it still remains to be clarified which structures are in a state corresponding to a global free energy minimum and which ones exist in a kinetically induced local minimum. This lack of clarity leads to an unresolved dilemma in constructing potential-based models of proteins. A free energy surface can be investigated with energy minimization techniques; the global minimum would then be the predicted native structure. Such an investigation sheds little light on the folding process itself, as it requires a correct description of both potential and entropic contributions at every stage in the process, and thus energy minimization cannot identify bottlenecks in folding space. A potential energy description only includes the entropic contributions through a molecular dynamics or a Monte Carlo simulation, and the search for free energy minima close to the native structure can be slow. Neither is it trivial to positively identify a true native state. Molecular dynamics simulations of fully solvated all-atom models have contributed tremendously to molecular biochemistry. However, due to the large separation of timescales involved in the global folding problem versus individual atoms or side chain dynamics, such simulations cannot address the folding of an arbitraty sequence of amino acids. Instead molecular dynamics simulations with all-atom models tend to focus on local unfolding events.[11] Often a

Received March 16, 1993; revision accepted November 2, 1993.

Address reprint requests to Dr. Anders Wallqvist at his present address: NCI-FCRDC, Building 1052, Room 238, P.O. Box B, Frederick, MD 21702.

mixture of potential energy and free energy terms is used, and energy minimization techniques are combined with simulation methods.

Effective models have been employed for many years in structure prediction and protein folding studies[12–20] and also recently in surfactant systems to mimic micelle formation from oil/water/surfactant mixtures.[21] Given the proper parameterization, such models are capable of yielding information on complex phenomena that cannot be made available from all atom models. The uniqueness of an effective potential is lost as different degrees of freedom are averaged over depending on the particular construction. Yet is is obvious from the literature[15, 20] that different effective potentials are capable of yielding sufficiently accurate structural information to positively identify a native configuration.

The efforts presented here are directed toward constructing an effective model potential built on our understanding of the important factors involved in protein folding and stability. In folding proteins from a random coil, or a denatured state, we would like to single out the forces that globally drive the folding and guide the polypeptide into its native conformation.[22–24] The main driving force is connected with the tendency for hydrophobic side chains to avoid aqueous surroundings.[25] This accounts for the formation of a molten globule state with a hydrophilic surface and a hydrophobic core, but without the fully developed secondary structure.[26] This is similar to the formation of hydrophobic aggregates and micelles in aqueous solutions. Secondly, interactions involving the polypeptide backbone, i.e., hydrogen bonds between peptide groups, give rise to more well-defined elements of secondary structure such as α-helices and β-sheets. The third and last contribution to the stability of the native state, and to the determining of the structure, is the close packing of individual atoms to include specific atom-atom interactions. This is a short-range interaction as compared to the two first forces which have a more intermediate- or long-range nature.[22]

Our aim is now to describe some more or less heuristic rules by which we can reduce the number of atoms and assign reasonable parameter values for all 20 amino acid and their interactions, yet still retain enough complexity to generate recognizable protein structures. Of course the stability of the close packing of side chain atoms is not available in a united atom model.

AMINO ACID DESCRIPTION AND INTERACTION

Generating protein structures from the primary sequence information using atomic models of polypeptides in solution requires a computational power that we do not possess. Neither is it obvious that current force fields for biomolecular molecules would be able to properly account for protein interactions in structures that are far from equilibrium and not in their native states, as most parameter values are obtained using empirical models only of equilibrium structures.[27] Instead it is more desirable to try to identify relevant degrees of freedom in the problem and build a model incorporating interactions conforming to this restricted space.

We would like to introduce a continuous model that combines an effective potential based on hydrophobic interactions between residues and a simplified description of the peptide backbone. Solvent molecules are not included; their effect is wholly incorporated into the effective residue–residue potential. Space is not discretized as in a lattice model and consequently molecules are able to take on any conformation, subject to the assigned force field. The reduction in the number of degrees of freedom compared with an all-atom model is two to three orders of magnitude. Leaving out the water molecules also means that the timescale of the problem has been shifted from the solvent to the intrinsic motion of the protein itself. This reduction in phase space is partly offset by the more stringent requirements of incorporating realistic interactions capable of reproducing protein structures in the effective potential. These requirements take the explicit form of additional potential terms not normally included in all-atom force fields nor explicitly considered in lattice models. The source of all these interactions is not information-based but relies on thermodynamic data and geometric considerations of allowed conformations. That the interactions are not information-based is in contrast to most other reduced protein representations that rely on data from structures already in the Protein Data Bank. The choice of the amino acid representation and the values assigned to the interaction parameters do not correspond to a unique selection, but rather reflects our bias in constructing a model that is sufficiently complex to reproduce common structural features in the protein, yet remains computationally managable.

The development of the simplified force field is given below. A complete treatment of the associated programs, simulations methods, analysis routines, etc. is given elsewhere.[28]

Amino Acid Representation

In the most primitive model of a protein each residue is represented by one interaction site and a bond between neighbors in the protein sequence. Such a model is not expected to be able to correctly reproduce the conformations proteins can assume, but can be used to study qualitative behavior of folding patterns, globularity, role of hydrophobicity etc.[29–32] Our efforts with single site representations, allowing for different size and interactions among the amino acids, did produce fluctuating, compact structures but with little or no secondary structure. Further extensive studies with the N-terminal EGF-like module of blood coagulation factor X (fX-EGF$_N$) lead us to divide the amino acid interactions and representations into two classes, one for the peptide backbone and one for the side chains. The solution conformation of fX-EGF$_N$ had previously been experimentally determined with 2D-NMR and a distance-constrained folding algorithm to generate three-dimensional structures.[33]

In the new representation the peptide backbone was modeled as a single interaction site (B), which coincides with the representation of glycine. To cap the protein, two special sites (BN, BC) were introduced at the beginning and end of the polypeptide chain to represent the effect of N- and C-terminal residues. Further, all amino acid side chains, except glycine, are represented with at least one site. The size of a side chain is derived from the solvent accessible area and each single or first site is connected to the α-carbon through a harmonic bond. The equilibrium distance of this bond corresponds to an average separation between the α-carbon and the center of mass of the atoms represented by the single or first side chain site. The single side chain site thus corresponds to an effective rotamer. The longer side chains are subdivided into two segments, partly to allow for better volume characterization and partly to allow for the introduction of amphiphilic chains. Double site side chains are used for the purely hydrophobic residues Leu, Ile, Phe, Tyr, Trp, and Met. A hydrophobic first residue and a hydrophilic second residue are assigned to Arg, Lys, Glu, Gln, and His. The introduction of a second site also means that there are now more than one side chain conformation, corresponding to an infinite number of effective rotamers, subject only to the internal constraints of the bond length and bond angle potentials. The division of the side chain atoms is shown in Table I and reflects the characteristic interactions associated with a group of atoms. Thus, e.g., the phenylalanine side chain is divided into two sites corresponding to the methylene group and the benzene ring. The division of lysine into a hydrophobic site comprising the chain —$(CH_2)_3$— and the hydrophilic head region —CH_2—NH_3^+ allows us to separate the interaction type, as well as regaining some of the rotational degrees of freedom lost in the reduction scheme. Figure 1 schematically illustrates the interaction sites used to represent a model protein. The special modelling of disulfide bridges and proline residues is discussed below.

As the amino acid side chain sites represent a contraction of the corresponding all-atom model, it is not possible to directly make a simple and unique back transformation to atomic coordinates in our model. What we can do is generate coordinates of the α-carbon atoms as well as identify hydrophilic regions, hydrophobic contacts, and specific amino acid residue interactions through the side chain site locations. However, high resolution experimentally determined α-carbon traces have been shown to contain enough information to regenerate all atom positions.[34]

TABLE I. Reduced Amino Acid-Side Chain Interaction Sites and Their Constitutent Atoms

Amino acid	Site label	Character	Atoms
Ala,A	A	Hydrophobic	—CH_3
Val,V	V	Hydrophobic	—$CH(CH_3)_2$
Pro,P	P	Hydrophobic	—$(CH_2)_3$—
Thr,T	T	Hydrophobic	—$CH_2(CH_3)$—OH
Ser,S	S	Hydrophilic	—CH_2—OH
Asn,N	N	Hydrophilic	—CH_2—$CO(NH_2)$
Asp,D	D	Hydrophilic	—CH_2—COO^-
Arg,R	R1	Hydrophobic	—$(CH_2)_3$—
	R2	Hydrophilic	—NH—$C(NH_2)_2^+$
Lys,K	K1	Hydrophobic	—$(CH_2)_3$—
	K2	Hydrophilic	—CH_2—NH_3^+
Glu,E	E1	Hydrophobic	—CH_2—
	E2	Hydrophilic	—CH_2—COO^-
Gln,Q	Q1	Hydrophobic	—CH_2—
	Q2	Hydrophilic	—CH_2—CO—NH_2
Leu,L	L1	Hydrophobic	—CH_2—
	L2	Hydrophobic	—$CH(CH_3)_2$
Ile,I	I1	Hydrophobic	—$CH(CH_3)(CH_2)$—
	I2	Hydrophobic	—CH_3
Phe,F	F1	Hydrophobic	—CH_2—
	F2	Hydrophobic	—C_6H_5
Tyr,Y	Y1	Hydrophobic	—CH_2—
	Y2	Hydrophobic	—C_6H_4—OH
Trp,W	W1	Hydrophobic	—CH_2—
	W2	Hydrophobic	—(C_8NH_6)
Met,M	M1	Hydrophobic	—$(CH_2)_2$—
	M2	Hydrophobic	—S—CH_3
Cys,C	C1	Hydrophobic	—CH_2—
	C2	Hydrophobic	—SH
His,H	H1	Hydrophobic	—CH_2—
	H2	Hydrophilic	—$(C_3N_2H_2)$
N-terminal	BN	Hydrophilic	—NH_3^+
C-terminal	BC	Hydrophilic	—COO^-

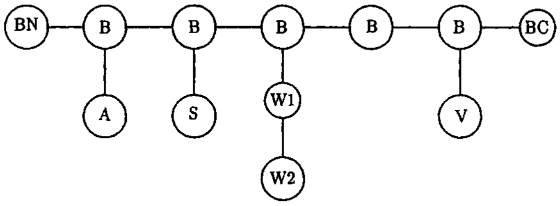

Fig. 1. The schematic representation of the pentapeptide ASWGV, which illustrates the reduction of amino acids into a few interaction sites. Note that the site denoted *B* is interchangable with the Gly residue, as it represents the peptide backbone site.

Hydrophobic Interactions

Without solvent molecules it is necessary to construct an effective potential between side chain sites, i.e., a potential of mean force averaged over

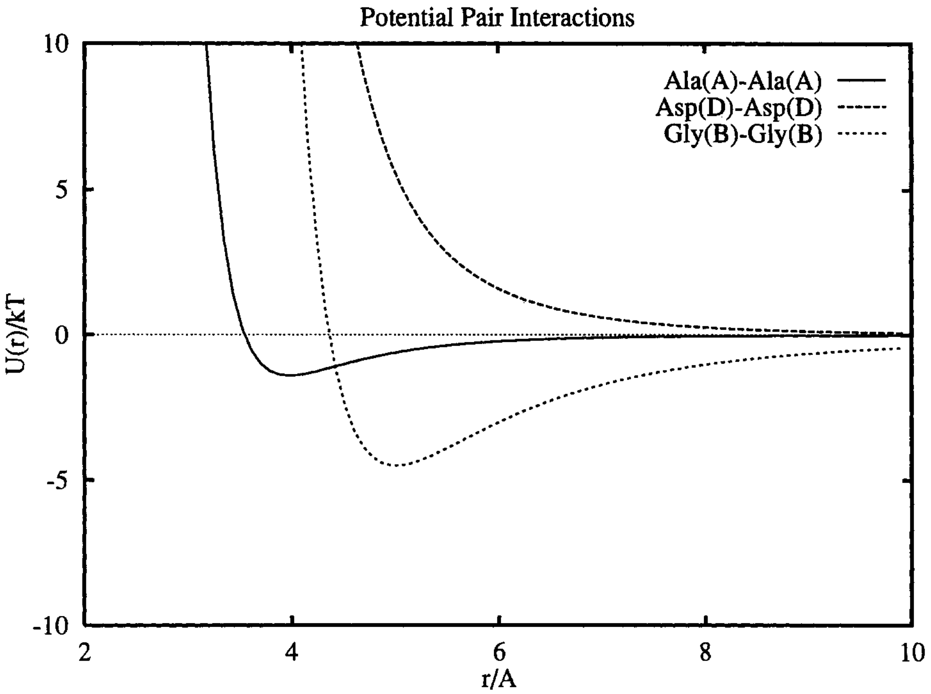

Fig. 2. The effective interactions between hydrophobic side chain sites of Ala-Ala, hydrophilic side chain sites of Asp-Asp, and between peptide backbone sites. The potential for the backbone sites is equivalent to the total Gly residue interaction. The alanine side chain potential resembles that of methane–methane interaction.

the solvent degrees of freedom. These interactions are based on a hydrophobicity criterium where the size of each residue site is related to the amount of exposed surface area. The potential form was taken to be either an attractive Lennard–Jones potential for purely hydrophobic side chains like alanine,

$$U_{AA}(r) = 4\epsilon_{AA}\left\{\left(\frac{\sigma_{AA}}{r}\right)^{12} - \left(\frac{\sigma_{AA}}{r}\right)^{6}\right\} \quad (1)$$

or a repulsive form for purely hydrophilic side chains like aspartate,

$$U_{DD}(r) = 4\epsilon_{DD}\left\{\left(\frac{\sigma_{DD}}{r}\right)^{12} + \left(\frac{\sigma_{DD}}{r}\right)^{6}\right\}. \quad (2)$$

The total interaction energy between two amino acids would be the sum of all side chain and backbone interactions. A sample of these functions is given in Figure 2, with the construction of the σ,ϵ-parameters discussed in the following sections.

No explicit charges are included in the model. Charge-charge interactions between residues are only taken into account by the fact that unlike charges attract and like charges repel each other. This is heuristicly done simply by using an attractive or repulsive Lennard–Jones interaction of the appropriate type, regardless of the degree of hydrophobicity of the side chains. Another approach, not employed here, is to add a $qq'/\epsilon_d r$-term and assign an effective value of the dielectric constant, ϵ_d.

A secondary effect inherent in the potential of mean force between nonpolar particles is a non-uniform hydrophobic attraction stemming from the special clathrate-like water layering around solutes.[35,36] With the removal of explicit solvent molecules we cannot reproduce this type of solvent ordering, which is clearly seen in simulations of small hydrophobic molecules in water. Inclusion of an oscillating function is possible, but would then be restricted only to amino acid residues that interact with water molecules between them. As this is not the case for residue interactions within the protein, which form the bulk of all residue interactions, we have elected not to include this structuring effect. The effect on the folding process itself is harder to gauge as residues during the initial stages of folding may very well be fully hydrated and shed their sphere of waters only in the native state. Although the overall effect of water is correctly modeled via hydrophobic interactions, the details of the water structuring effect is lost at intermediate distances, i.e., for interactions where amino acids would be separated by one or two water molecules. Naturally the concept of bound or crystallographic waters is not applicable to the model.

ε-Values

The free energy of transfer of amino acids from water to octanol was used as a measure of the hydrophobicity of each side chain.[3, 37, 38] As the potential surface is a mix of free energy and potential energy terms, this choice restricts the validity of all parameters to the state point at which these energies were measured, with regards to temperature and pH.

The ϵ parameter for the interaction between side chain sites i and j was then chosen to be proportional to the sum of these energies,

$$\epsilon_{ij} = c_{ij} v_{\Sigma} (\Delta G_i + \Delta G_j). \tag{3}$$

These energies are measured relative to a zero of interaction between glycine residues, and the specific backbone interactions are treated separately below. If the interaction is purely hydrophilic we use the absolute value of ϵ combined with the repulsive Lennard–Jones potential interaction type. Thus it is the relative hydrophobicity of the pair that determines if the interaction between them is attractive or repulsive. In order to retain a hard-core volume, absolute values of ϵ less than 0.1 were set equal to 0.1. The factor multiplying the free energy measure, c_{ij}, is a parameter that determines how strongly we couple the potential energy to the free energy and is currently set to 1/3. The second factor v_{Σ} is either 1 or 1/3 depending on if the interaction is attractive or repulsive, i.e., it corrects for the overestimate of the closest approach of two repulsive amino acids.

The interaction energy needs to be further subdivided for amino acids whose molecular description includes two side chain centers. For purely hydrophobic double site residues we want the overall attractive interaction between a dimer to remain the same whether we think of the interaction as stemming from two or four sites, i.e.,

$$\epsilon_{ii} \approx \epsilon_{i1,i1} + \epsilon_{i2,i2} + 2\epsilon_{i1,i2} \tag{4}$$

where $i1$ and $i2$ denote the first and second site. Thus we have introduced the following condition on the free energies for the divided sites,

$$\Delta G_{i1} + \Delta G_{i2} = \Delta G_i/2. \tag{5}$$

For the division of the ampiphilic side chains we have to take into account the volume factor v_{Σ} multiplying the repulsive hydrophobic interaction,

$$\Delta G_{i2(\text{hydrophilic})} = \frac{1}{2}\left\{\Delta G_i - \left(1 + \frac{1}{v_{\Sigma}}\right)\Delta G_{i1(\text{hydrophobic})}\right\}. \tag{6}$$

The assignment of hydrophobic free energies for a double-sited amino acid side chain then proceeds from using the ΔG of a closely related single-site amino acid side chain, such as alanine for the first site of leucine etc., and then calculating the ΔG_{i2} values.

TABLE II. Assigned Free Energies and Radii for Interaction Sites Used in the Protein Potential

Amino acid	Site label	ΔG (kJ/mol)	Radius (Å)
Gly,G	B	0.00	2.18
Ala,A	A	1.76	1.77
Val,V	V	6.95	2.07
Pro,P	P	3.08	2.13
Thr,T	T	1.46	2.05
Ser,S	S	−0.21	1.64
Asn,N	N	−3.43	2.20
Asp,D	D	−4.39	2.05
Arg,R	R1	3.49	2.17
	R2	−5.19	2.18
Lys,K	K1	3.49	2.17
	K2	−5.15	2.80
Glu,E	E1	1.36	1.50
	E2	−2.73	2.50
Gln,Q	Q1	1.36	1.77
	Q2	−1.54	3.02
Leu,L	L1	1.36	1.77
	L2	3.49	2.85
Ile,I	I1	3.71	2.85
	I2	1.44	2.78
Phe,F	F1	1.36	1.77
	F2*	3.75	3.67
Tyr,Y	Y1*	1.36	1.77
	Y2	1.38	3.85
Trp,W	W1	1.36	1.77
	W2*	7.60	4.03
Met,M	M1	1.36	2.00
	M2	2.16	2.81
Cys,C	C1	1.36	1.77
	C2	1.45	2.78
His,H	H1	1.36	1.77
	H2*	−0.53	3.40
N-terminal	BN	−3.39	2.00
C-terminal	BC	−2.63	2.25

*Interaction sites containing an aromatic ring have their σ values reduced by 1.4 Å.

The values used for the free energies of each interaction site are given in Table II. From this table one can then construct all ϵ-parameters needed in the force field by applying Eq. (3).

σ-Values

The size of each interaction site was related to the solvent accessible area for each amino acid.[3, 37, 38] Glycine, which is choosen to be the basic unit of the simplified peptide backbone, was set to have a radius, r_G, of 2.18 Å. The σ value for a pair of side chain sites was constructed from the sum of the radii, e.g., for alanine and aspartate,

$$\sigma_{AD} = r_A + r_D. \tag{7}$$

Interaction sites containing benzene rings had their radius further reduced by 1.4 Å per residue and interaction to conform to the known sizes estimated from the potential of mean force calculation on aqueous benzene dimers.[39] This radius reduction diminishes the effect of the anisotropy of the benzene interaction by allowing side chain sites a closer approach, and thus affects the packing property of these side chains.

The excess surface area of a side chain was calculated by subtracting the idealized peptide backbone, i.e., the glycine, contribution from the area of the entire amino acid,

$$A_{\text{exc}} = A_{\text{amino acid}} - A_{\text{Gly}}. \tag{8}$$

We should take into account that two connected spheres will not have an additive surface exposed to the implicit solvent by further adding πr_{G}^2 per every half surface. This procedure compensates for the fact that two spheres, with radii giving effective distances of closest approach, overlap when they are joined by a bond. Thus the r-value for a single amino acid side chain site is determined from,

$$4\pi r^2 = A_{\text{exc}} + 2\pi r_{\text{G}}^2. \tag{9}$$

For a double side chain representation the separate areas were constrained to obey,

$$4\pi(r_1{}^2 + r_2{}^2) = A_{\text{exc}} + 4\pi r_{\text{G}}^2. \tag{10}$$

Again, the assignment of r values for side chains represented by two interaction centers proceeds from using known values of a closely related single-sited side chain. The values used in the construction of the σ-parameters are given in Table II.

Peptide Backbone Interactions

The peptide backbone has been singled out because of the specific considerations necessary of the intrinsic chain interactions of a protein. In our preliminary studies of the folding process of fX-EGF$_{\text{N}}$ without specific backbone interactions, it was obvious that the folding space was too large, as the compact states generated could not with certainty be assigned to the experimental structure. This was due to an avoidance of forming intermingled patches of hydrophobic and hydrophilic residues, leading to the formation of a molten globule state of micellar nature. Then we incorporated the sterical constraints imposed by the peptide groups on the α-carbon positions, by using an effective angle potential between connected backbone interaction sites (see below). In effect, this forces a folding process where the hydrophobic interaction of the side chains drives the protein toward the native structure, while still maintaining the sterically allowed backbone conformations. This reduces the allowed folding space, but at the same time introduces further potential barriers between conformations.

Since backbone interactions are not absolutely determined from the free energy measurements used to construct the side chain site potentials, we are at some liberty to choose these parameters. Thus the interaction between nonconnected backbone sites was slightly modified compared to side chain interactions in that the dispersive term was changed to a more long-ranged, r^{-4}, term. This reflects the hydrogen bonding between peptide groups,

$$U_{\text{BB}}(r) = \gamma_{\text{BB}}\left\{\left(\frac{\sigma_{\text{BB}}}{r}\right)^{12} - \left(\frac{\sigma_{\text{BB}}}{r}\right)^{4}\right\}. \tag{11}$$

This potential is given in Figure 2 with the assigned values $\gamma = 9.74$ kJ/mol and $\sigma = 4.36$ Å. Note that no explicit dihedral angle potential is used between backbone sites.

Hydrogen bond pattern correction

A specific feature of polypeptides is their property of forming regular structures stabilized by a network of hydrogen bonds interconnecting the peptide backbone. In particular for an α-helix, the geometry of the chain and the positions of the hydrogen bond donor and acceptor atoms result in a helix-pitch of 3.6 residues. An attractive pair potential between backbone sites, whose interaction centers coincide with the α-carbon atom, cannot reproduce this pattern. The helices generated from such a potential become squat, acquire an erroneous pitch, and a residue i tends to equalize its interactions with residues $i + 4$ and $i + 5$. This is due to the fact that the effective contact energies are optimized for such a conformation. This was evident from the helical structures generated by short, 16-residue long polyalanines. In reality the hydrogen bond occurs between the carbonyl oxygen, between residues i and $i + 1$, and the amide hydrogen in the peptide group between residues $i + 4$ and $i + 5$. Instead of introducing extra sites to represent the peptide group, we have used a hydrogen bond correction potential which enhances the typical α-carbon site position for hydrogen bonds between nonconnected residues,

$$U_{\text{hbpc}}(r_{i,i+4}, r_{i,i+5}) = Ae^{-c_1(c_2 - r_{i,i+4})^2}e^{-c_3(c_4 - r_{i,i+5})^2}. \tag{12}$$

The parameters are chosen so that the hydrogen bond pattern is preferentially stabilized for all residues except the proline residue. Thus for nonproline residues the parameters are $A = -5.0$ kJ/mol, $c_1 = 1.35$ Å^{-2}, $c_2 = 5.19$ Å, $c_3 = 1.09$ Å^{-2}, $c_4 = 6.29$ Å whereas for a proline residue they take on the following values, $A = -10.0$ kJ/mol, $c_1 = 0.180$ Å^{-2}, $c_2 = 8.5$ Å, $c_3 = 0.145$ Å^{-2}, $c_4 = 12.0$ Å.

The hydrogen bond pattern correction asserts the intrinsic propensity for a specific arrangement of backbone atoms in polypeptides due to the formation of hydrogen bonds. Of course, "hydrogen bonds" occur between all sites that have a possibility of doing so in real life, through the normal hydrophobic and peptide backbone interactions. It is the supression of the atomic details of the hydrogen bond that neces-

TABLE III. Parameters for the Stereospecific Potential

Type	k_e (kJ/mol/rad^4)	θ_α (rad)	θ_s (rad)	θ_β (rad)
Nonproline	50.0	1.571	1.134	2.185
Proline	75.0	1.825	0.700	2.020

sitates the hydrogen bond pattern correction. Proteins containing β-sheets in their experimentally determined native state do not show any undue tendency to denature when applying the force field constructed here. The formation of a specific type of secondary structure is thus determined by the side-chain interactions and not by backbone interactions, as in real molecules.[40–45]

Stereospecificity

Nature has specified that proteins are built up from only L-amino acids, giving an absolute orientation to the side-chain location relative to the direction of the polypeptide backbone. Since we do not have four explicit sites around a backbone interaction site we have constructed a potential to correctly orient the amino acid side chain relative to the backbone direction.

Thus the orientation of side chain j connected to backbone site i is fixed relative to the three backbone sites $i - 1$, i, and $i + 1$. The potential expression was chosen to be,

$$U(\theta_1,\theta_2)_{\text{stereo}} = k_e\{(\theta_1 - \theta_\alpha)^2 + (\theta_2 - \theta_s)^2\} \{(\theta_1 - \theta_\beta)^2 + (\theta_2 - \theta_s)^2\} \qquad (13)$$

where the angle θ_1 is the angle between the two vectors originating on site i and ending on sites $i - 1$ and $i + 1$. The parameters θ_α and θ_β are the preferred angles for triplets of C_α-sites in the helix or sheet conformation. θ_2 is the angle between the side-chain bond and the normal to the plane, formed by atoms $i - 1$, i, and $i + 1$, which has a preferred value of θ_s. These parameters are given in Table III. This potential ensures that there are no transformation between L- and D-amino acids during the simulation.

Bonds and Angles

Interaction sites connected via a bond, as specified in the amino acid representation, interact with each other only through a harmonic bond potential,

$$U_{\text{bond}}(r) = k_e'(r - r_e)^2. \qquad (14)$$

The bond lengths were set to coincide with center of mass separation of the interaction sites given in Table I. The assigned values and force constants are given in Table IV. Triplets of connected sites are associated with an angle potential that can be of either a harmonic or a bistable type, i.e., having two accessible angles. Sites connected via a specified angle potential do not use any other interaction between them. Triplets assigned a harmonic angle potential are given in Table V. The bistable potential consists of a broad quartic together with two Gaussians of variable width and depth to characterize the minima,

$$U_{\text{angle}}(\theta) = k_e''(\theta_{e0} - \theta)^4 - c_1 e^{-d_1(\theta_{e1} - \theta)^2} - c_2 e^{-d_2(\theta_{e2} - \theta)^2}. \qquad (15)$$

This potential operates between triplets of connected backbone sites (B–B–B) as well as between backbone/backbone/first-interaction sites (B–B–X1). It is introduced to effectively take into account the preferred α-carbon positions associated with the φ and ψ angles for α-helices and β-sheets found in a Ramachandran diagram. The angle that the α-β-carbon bond makes with respect to the backbone skeleton is accounted for by the bistable potential between backbone/backbone/first-interaction sites. The parameter values are given in Table VI.

In a real protein the peptide group can for all practical purposes be thought of as flat rigid unit. It is then possible to describe the entire backbone conformation in terms of the φ and ψ angles, i.e., the rotations around the N — C_α and the C_α — C′ bonds. These rigid rotations result in specific angles between triplets of connected C_αs and thus there exists a mapping of φ and ψ angles to a $C_{\alpha,i-1}$–$C_{\alpha,i}$–$C_{\alpha,i+1}$ angle. The mapping in the opposite direction is not single-valued and consequently we have chosen to restrict the angular space to reproduce right-handed α-helices and β-strands. The distance between two consecutive α-carbons is independent of φ and ψ and would be fixed if the bond lengths and angles were fixed. Thus the simplification of removing the explicit peptide group and having just one backbone site at the α-carbon site introduces no loss of conformational information per se.

Disulfide Bridges

In order to mimic the disulfide bridges formed in proteins, a strong interaction at the typical SS bond distance was added for pairs of cystein residues. Thus, it is possible for disulfide bonds to form as well as break during the simulation depending on the strength of the interaction. This can occur within a protein as well as between cysteins on different proteins. In order to allow for the bond formation, a deep attractive well was added to the cystein–cystein interaction by modifying the corresponding Lennard–Jones potential for the C2–C2 pair,

$$U_{\text{disulfide}}(r) = 4\epsilon_{C2C2}\left\{\left(\frac{\sigma_{C2C2}}{r}\right)^{12} - \left(\frac{\sigma_{C2C2}}{r}\right)^{6}\right\} \left(1 - \frac{1}{1 + e^{\sigma_{C2C2}(r - \sigma_{C2C2})}}\right) - c_1 e^{-c_2(r - c_3)^2}. \qquad (16)$$

The assigned values are $c_1 = 6.25$ kJ/mol, $c_2 = 3.0$ Å^{-2}, and $c_3 = 2.0$ Å. As this is a *pair* interaction

TABLE IV. Bond Lengths and Harmonic Force Constants*

Amino acid,X	Site label	$r_{B,X}/r_{X1,X2}$ (Å)	k_e (kJ/mol/Å^2)
Gly,G	B	3.80	150.00
Ala,A	A	1.50	62.25
Val,V	V	2.10	50.00
Pro,P	P	1.90	75.00
Pro,P†	P	4.00	125.00
Thr,T	T	2.10	50.00
Ser,S	S	1.90	50.00
Asn,N	N	2.70	37.50
Asp,D	D	2.50	37.50
Arg,R	R1	2.10	37.50
	R2	2.10	50.00
Lys,K	K1	2.10	50.00
	K2	1.90	50.00
Glu,E	E1	1.50	62.50
	E2	2.50	37.50
Gln,Q	Q1	1.50	62.50
	Q2	2.70	37.50
Leu,L	L1	1.50	62.50
	L2	2.10	50.00
Ile,I	I1	2.10	50.00
	I2	2.10	50.00
Phe,F	F1	1.50	62.50
	F2	3.00	37.50
Tyr,Y	Y1	1.50	62.50
	Y2	3.25	37.50
Trp,W	W1	1.50	62.50
	W2	4.00	37.50
Met,M	M1	1.90	50.00
	M2	1.90	50.00
Cys,C	C1	1.50	62.50
	C2	1.70	50.00
His,H	H1	1.50	62.50
	H2	2.60	37.50
N-terminal	BN	1.50	75.00
C-terminal	BC	1.50	75.00

*Bond lengths between the backbone site and the first interaction site, or between the side-chain sites in the amino acid in the case of a second interaction site.

†For the proline side-chain site there are two bonds connecting this site to the peptide backbone chain. See the section of the special treatment of proline residues.

TABLE V. Bond Angles and Harmonic Force Constants

Amino acid	Sites	Angle (rad)	k_e (kJ/mol/rad^2)
Arg,R	B–R1–R2	3.141	0.1
Lys,K	B–K1–K2	3.141	0.1
Glu,E	B–E1–E2	2.094	1.0
Gln,Q	B–Q1–Q2	2.094	1.0
Leu,L	B–L1–L2	2.094	5.0
Ile,I	B–I1–I2	2.094	5.0
Phe,F	B–F1–F2	2.094	15.0
Tyr,Y	B–Y1–Y2	2.094	15.0
Trp,W	B–W1–W2	2.094	15.0
Met,M	B–M1–M2	2.094	1.0
Cys,C	B–C1–C2	2.094	15.0
His,H	B–H1–H2	2.094	15.0
N-terminal	B–B–BN	2.094	1.0
C-terminal	B–B–BC	2.094	1.0

there is no restriction on how many disulfide bridges any one cystein residue can partake in. For reasonable values of the bond strength this leads to unphysical results where cystein residues cluster together. In order to prevent clustering a repulsive term parameterized for triplets of sulfide pair distances r_1, r_2, and r_3 was added to the total potential energy. The form of this three-body correction was chosen as

$$U_{\rm sss}(r_1,r_2,r_3) = A\prod_{i=1}^{3} e^{-c(r_i - r_{C2C2})^2} \qquad (17)$$

with parameters $A = 5000.0$ kJ/mol, $c = 0.7$ Å^{-2}, and $r_{C2C2} = 2.5$ Å. The values of these parameters are connected to the choice of the strength of the bond interaction. In essence this potential allows two cystein residues to approach each other within a typical bonding distance, whereas a third cystein residue will be repelled by the pair. It is also possible to select other strategies for simulating disulfide bonds depending on the reducing conditions. They can be explicitly included in the connectivity of the protein as a covalent bond or they can be ignored altogether as a special interaction and just assigned a hydrophobic free energy in analogy with other amino acids.

From our initial calulation on the folding of fX-EGF$_N$ it became clear that while the three disulfide bridges were not structure determining, it was possible to fall into deep kinetic traps resulting from wrongly connected disulfide bonds. Even though bond dissociation did occur, the time spent generating conformations with wrong bond order was substantial. In essence this confirms that the complexity of our surface mimics the complexity of real systems, where it is found that only a fraction of fIX-EGF$_N$ molecules expressed in *E. coli* are fully active[46] due to a variance in disulfide bond connectivity. Folding intermediates with nonnative disulfide bonds has also been found experimentally in studies of the folding of BPTI, though the population of these states has been disputed.[47, 48]

Proline Residues

As the proline side chain binds directly to the peptide backbone, an extra bond between the interaction site representing the proline side chain and the previous backbone site was introduced. This equilibrium bond distance is then set to 4.0 Å with a harmonic force constant of 125.0 kJ/mol/Å^2. The sterically allowed ϕ and ψ angles between the proline backbone site and adjoining sites are incorporated in the corresponding angle potentials (see Tables III and VI). These angular restrictions allow the proline

TABLE VI. Bistable Angle Potential for Triplets of Connected Interaction Sites

Type	k''_e (kJ/mol/rad^4)	θ_{e0} (rad)	c_1 (kJ/mol)	d_1 (rad^{-2})	θ_{e1} (rad)	c_2 (kJ/mol)	d_2 (rad^{-2})	θ_{e2} (rad)
Nonproline*	50.0	1.920	−6.0	−35.0	1.571	−6.0	−35.0	2.185
Proline*	85.0	1.825	0.0	—	—	−6.0	−50.0	2.020
B–B–X1[†]	50.0	2.182	−3.0	−35.0	1.920	−3.0	−35.0	2.443

*These potentials operate only between triplets of connected backbone sites (B) and depends on whether these sites include a proline residue or not.

[†]This potential operates between triplets of backbone/backbone/first-interaction sites and assures the alignment of the α-β-carbon bond with respect to the backbone geometry.

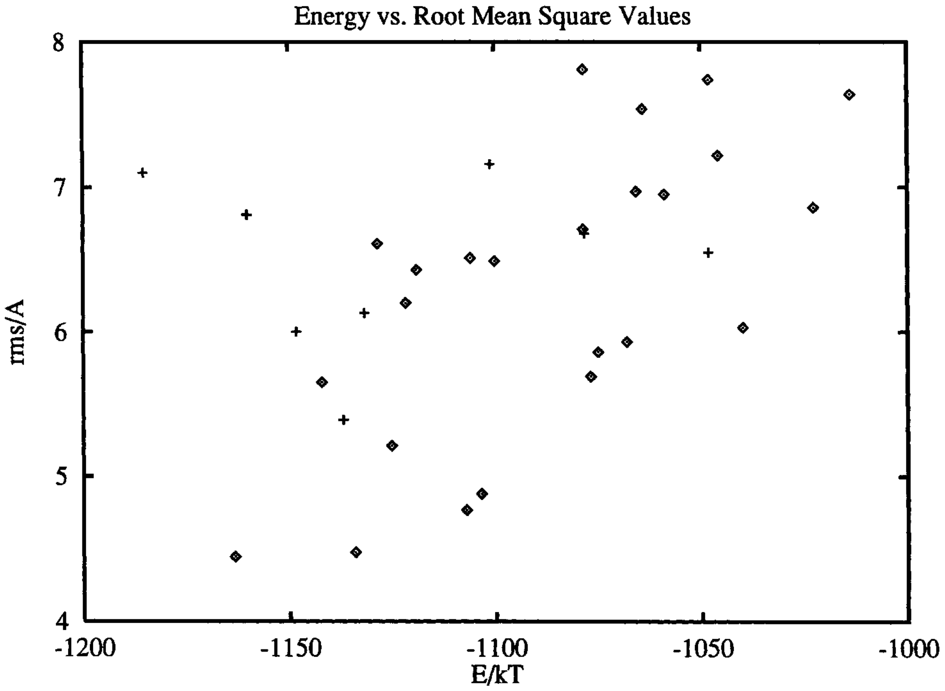

Fig. 3. A scatter plot of potential energies generated for a set of Monte Carlo runs of PPT from extended configurations. (+) Compact structures with distorted helices; (◇) conformations with a general agreement of just the helix region, i.e., with an rms deviation from the X-ray structure of less than 2.5 Å for residues 14–32.

representation to cause the typical hindrance of the normal α-helix structures while still retaining the possibility of forming proline helices.

PROTEIN STRUCTURE DETERMINATION

The actual folding of a protein in vivo is a complex process that may depend on many specific factors in the environment. Partially misfolded proteins have to be taken care of, kinetic traps have to be overcome, and the search for the native state may be advanced through chaperonins.[24] For a large number of proteins it is, however, still possible to perform renaturing experiments in vitro, obviating the need to reproduce a specific intracellular environment. For the small proteins studied here the folding process itself is rather unencumbered, and does not present a great challenge to study. For larger proteins the number of local minima will increase drastically and the folding process itself will become more relevant. Consequently for the molecules presented below an ordinary Monte Carlo procedure[49] suffices to generate conformations from an extended state. The conformations generated will then correspond to equilibrium fluctuations around native states. If the native state is well-defined in our potential model, these fluctuations will visit only those states that are closely related to the native structure. Trapped structures with potential energies deviating far from each other can then be excluded as unlikely candidates for the native state.

Avian Pancreatic Polypeptide, PPT

The structure of avian pancreatic polypeptide has been elucidated from high-resolution (1.4 Å) X-ray experiments,[50] designated as 1PPT in the Protein Data Bank. In the crystal structure the peptide forms dimers, which are further crosslinked by zinc ions. The monomer consists of one short α-helix with a polyproline-like strand that is bent on top of the helix. The helix segment has one hydrophobic and one hydrophilic side. Thus, in spite of its small size, PPT is able to form a nonpolar core between one side of the helix and the hydrophobic residues of the polyproline-like segment.

In a previous work Crippen and Snow[15] used PPT as a test case in constructing a protein potential that has a global minimum at the native PPT structure. Although the potential is specific for the molecule, it did show that it was possible to achieve a low root mean square (rms) deviation of 1.8 Å for an effective potential.

Recently Sun[20] obtained an even lower rms of 1.3 Å with a simplified protein model consisting of the backbone atoms and single united atoms for the side chains and an effective potential based on information in the Protein Data Bank. The effective potential also contains a term that uses the known radius of gyration. A population of conformations is optimized by a genetic algorithm that uses conformational dictionaries containing short segments from 110 structures in the Protein Data Bank, including the PPT molecule itself.

In our description the 36-residue protein was reduced to 89 interaction sites. The main events in the folding process from an extended state are the initial formation of local contacts between neighboring residues initiating helix structure, followed by the folding of the entire, proline-rich strand (1–8) on top of the hydrophobic side of the helix (14–32). Energetically nonoptimal alignment of the proline strand resulted in many structures that were discarded.

In the effort to predict a structure a number of conformations are generated. These differ in potential energy. From those with low energy at least one has to be chosen as a representation of the prediction. A good correlation between energy and structure is possible only if the native structure is located in a very deep and well-defined minimum. In tackling the problem of deriving a three-dimensional structure from 2D-NMR experiments, the set of all distance constraints can be used to define such a minimum. Thus in principle, protein folding utilizing the experimental information correlates structure with a minimum in the energetic constraint penalties. In practise the experiments are not capable of yielding all distance constraints necessary to uniquely define a structure. Folding in this less well-defined conformation space then encompasses some of the problems encountered here in assigning a structure prediction, i.e., a choice must be made to discard structures that contain erroneous structure elements yet are intermediate or low in energy.[33] In Figure 3 we give the potential energy values versus the rms deviation of the α-carbon atoms for a set of Monte Carlo runs. The structures can roughly be divided into a class with a well-defined helix (◇) and one in which the helix formation is incomplete (+).

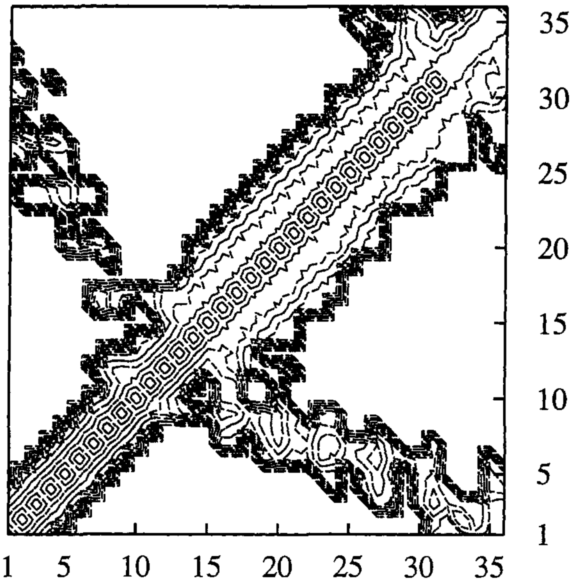

Fig. 4. α-Carbon distance map of PPT compared with the experimental X-ray structure (top portion of the graph). An element of the distance map $R_{N,N'}$ is defined as $|r_{B(N)} - r_{B(N')}|$, where $r_{B(N)}$ is the α-carbon coordinate of residue N. Only $R_{N,N'}$ values of less than 10 Å are plotted. Contours represent 2 Å intervals. The sequence of PPT is GPSQP TYPGD DAPVE DLIRF YDNLQ QYLNV VTRHR Y.

The structures with a well-developed helix show rough linear behavior between energy and structure. In the incomplete helix structure there is no such dependency, and the structures correspond to more compact states, where the side chains have maximized their contacts. These states are similar to the non-native states found by Crippen and Snow[15] which deviated by only 2% in energy from the native conformation. The structure discussed below corresponds to the lowest energy conformation with a well-developed helix as selected from Figure 3.

The α-carbon distance map for PPT is given in Figure 4 and compares our results with the experimental structure. The main features of the molecule in the polyproline (1–9), turn (9–14), and helix (14–32) region are qualitatively reproduced. The roughly antiparallel nature of the polyproline-like strand is also clearly shown.

Figure 5 shows a simulated structure and the X-ray structure. The rms deviation is 4.5 Å. This is larger than previous studies, but Crippen and Snow have constructed their potential from the known structure and it is unclear how much Sun's results are biased by the X-ray structure through the known radius of gyration and the conformational

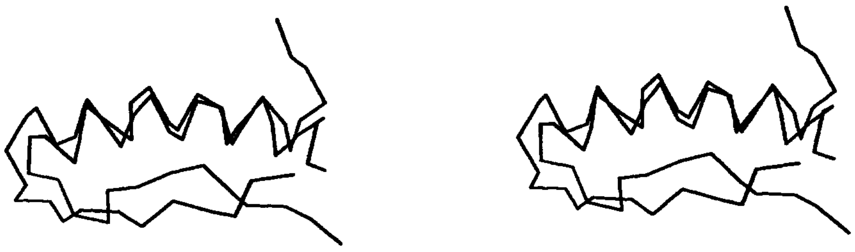

Fig. 5. Stereoplot of a simulated conformation compared with the experimental structure of PPT. The overlayed predicted structure deviates from the crystal structure with an rms of 4.5 Å. The matching in the figure is done over the α-helix to facilitate viewing.

dictionaries. The fact that the rms between any of Sun's structures is less than 0.3 Å indicates that there is either a very deep potential minimum or limited conformational freedom.

It should be kept in mind that the rms deviation is a single number used to describe something as complex as a comparison between two three-dimensional structures and should not be the sole criterium for judging theoretical predictions. Classifications of proteins based on their secondary structure show that there is insight to be gained from qualitative features. It is much more straightforward to give a number, however, and our rms value is still quite large. It may be partly explained by a conformational freedom that is too large despite the corrections in the effective potential. Figure 5 shows that the largest discrepancies occur in the beginning and at the end of the protein. This is also evident from the contact map in Figure 4 between the beginning and end residues.

From the crystal structure[50] we know that one zinc ion coordinates the Asp-23, Gly-1, and His-34 residues from three different monomers, making the position and orientation of the proline strand sensitive to the packing conditions of the crystal. There is also a slight overall bend in the simulated helix.

True differences between the conditions in the crystal and in our simulation may thus contribute to the deviation of the predicted structure from the crystal structure. We have predicted the structure of one molecule of the polypeptide in solution, whereas in reality it tends to self-associate. The question thus remains as to what extent the dimerization affects the structure as observed in the crystal, and what the influence of the coordinating zinc ions is.

Parathyroid-Hormone-Related Protein, PTHrP (1–34)

The native conformation in solution of this protein has been characterized by CD and NMR experiments by Barden and Kemp.[51] At the physical conditions of the experimental study (pH = 4.5) the five histidine residues are positively charged, whereas negative glutamic acid and aspartate side chains have a reduced net charge. Unfortunately no coordinates are given to facilitate the comparison, except in the form of photographs of a CPK model. A very low helical content of about 20% was extrapolated from the CD measurements, located mainly between residues 3 and 9. The rest of the protein is assigned to a random coil state, yet it forms a compact shape with many close contacts.

Recently Ota and Saito predicted the structure using an island model.[19] The amino acids are modelled as single atoms. The procedure starts with a determination of secondary structure. In this case the α-helix reported by Barden and Kemp was used. From a distance map of the structure, with the assigned secondary structure and the rest of the molecule in an extended conformation, close hydrophobic pairs are chosen. These are allowed to interact through a Lennard–Jones potential and the energy is minimized. New hydrophobic pairs are chosen from the distance map of the minimized structure and the energy is reminimized. The cycle is repeated until no more hydrophobic pairs can drive the folding. By using different random numbers, different conformations are obtained and finally the best one is refined with a more extensive force-field, including hydrogen bonding and electrostatic interactions.

The conformations generated from our model were all of a rather compact form and did not deviate much from this pattern. The conformation selected was the one with the lowest energy. The corresponding α-carbon contact map is given in Figure 6 and a schematic drawing is presented in Figure 7. The bends occuring between residues 9–12 and 18–21 (an arginine rich region) are indicated by the contact map. The characteristic close contact between the N- and C-terminal region is also seen. In Figure 7 we have further emphasized the predicted helical region with a ribbon skeleton. A hydrophobic pocket is formed in the space between the three helices. This region is comprised to a large extent of leucine residues (7,8,18,24,27), and forms a stable core. Although there is a tendency for hydrophobic residues to be at the bottom of the protein, as seen in Figure 7, this division does not constitute a complete separation of hydrophobic and hydrophilic residues.

In comparison with the rough dimensions of the

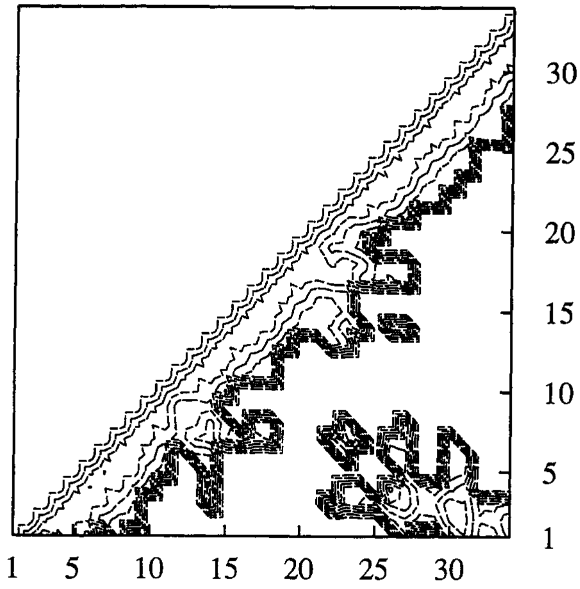

Fig. 6. α-Carbon distance map as defined in Figure 4 of PTHrP (1–34). The sequence of this protein is AVSEH QLLHD KGKSI QDLRR RFFLH HLIAE IHTA.

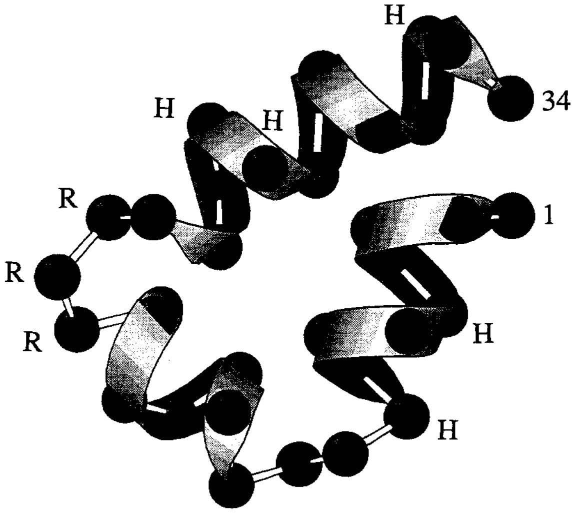

Fig. 7. Predicted structure of PTHrP (1–34) drawn as a combined ribbon and stick-and-ball cartoon. Of the total of 93 interaction sites for clarity only the α-carbon sites are indicated. The positively charged Arg residues (19–21) and the His (5,9,25, 26,32) sites are specially labeled.

molecule as presented by Barden and Kemp as 12 × 22 × 35 Å, our structure compares favorably with a dimension of 12 × 20 × 27 Å. Our model does however show a tendency for helix formation, not present in the experimental structure. The structure of Ota and Saito may also have three helices although the part of the structure that corresponds to the second and third helix is in that case rather perturbed. As our model does not explicitly include charges, charge effects as such are only modelled indirectly via the hydrophobicity criteria at neutral pH. At lower pH histidine residues become fully charged, a phenomenon not included in our model. Thus the added stability of salt bridges, whose presence was postulated by Barden and Kemp, cannot be correctly modelled. The conversion of His^0 to His^+ is also a direct cause of helical instabilities in short polypeptides.[41,44] As PTHrP (1–34) contains five histidines these effects must be significant, especially as the third predicted helix between residues 24 and 34 contains three histidines.

A set of trial runs was carried out where this charge effect was introduced ad hoc. This was done by considering the histidines to be positively charged in the assignment of a repulsive Lennard–Jones interaction to residues with like charges and an attractive interaction to unlike charges. No other parameter changes were made. Configurations generated with this parameter set did not fall into any specific category, but rather showed a wide dispersal in secondary and tertiary structures. Some looked more like the data from Barden and Kemp, others displayed even more compactly packed helices, but no measure could be given which selected these conformations as being native. This could indicate that the helical structure is destabilized at low pH. The difference between the predicted structures with charged and uncharged histidines confirms that proteins whose stability depend on conditions different from those for which the parameter set was built cannot be adequately represented.

CONCLUSIONS

This work has primarily been concerned with the development of a simplified amino acid representation together with a corresponding force field. A protein molecule is built up from connected segments representing groups of atoms. Thus there is one segment corresponding to the peptide backbone unit, and one or more segments to model the amino acid side chains. The main interactions between amino acid residues are based on the hydrophobic character of the side chains and are parameterized from thermodynamic data. A further set of potential functions is introduced to account for geometric and sterical constraints that amino acids must satisfy in polypeptides. Otherwise the conformational repertoire of the model will have a too large part that is unphysical in the sense that a real protein cannot adopt the corresponding conformations, e.g., conformations intermediate between the L- and D-forms and conformations resulting in severe steric overlap in an all-atom model. In a lattice model the need for such corrections is partly disguised by the fact that the lattice itself constrains the conformational freedom and dictates what a certain element of secondary structure should look like. In our model both secondary and tertiary structures are predicted. There is no prior knowledge of sequence propensities for any given structure element, nor are any potential functions parameterized from structures in the Protein Data Bank. No assumptions are made of the folding process itself.

In order to test the force field using only the amino acid sequence we tried to predict the secondary and tertiary structure of two small proteins, one whose coordinates are known and one for which there exist predictions but where the coordinates are not yet available, using only the amino acid sequence. Thus from folding studies of PPT and PTHrP (1–34) it is shown that the ideas incorporated into the model are capable of giving a satisfactory representation of the overall molecular structure.

Applications to the folding process of larger molecules will require refined methods in order to more rapidly generate significant conformations with a high probability. The current Monte Carlo scheme is fraught with the danger of getting trapped in intermediate states. The contributing entropic effects from a finite temperature simulation help to preclude the collapse of the molecule into a completely phase-separated micelle, with a hydrophobic core surrounded by hydrophilic residues on the surface. Thus part of the predicted protein structure is actually due to the folding process and not solely a function of the energy parameters evaluated at zero temperature. It should be emphasised that we do not claim that our simulations correspond to the true folding event as it occurs in nature. The natural folding process is, to our minds, partly stochastic and partly guided. Thus, we believe that if we have managed to construct an effective potential that contains the most important factors determining protein structures, some of the elements that guide proteins to their native states will also be present in the Monte Carlo simulation.

Protein folding studies for molecules, where either the physical environment deviates too far from the parameterization conditions or where a specific type of interaction dominates the protein stability with a different weight than was assigned here, will not yield quantitative results. This is especially true for cases involving direct charge stabilization between residues, e.g., salt bridges. Denaturing conditions are not included in the parameter space and thus absolute stabilities cannot be estimated.

Yet the simplifications introduced are tractable in that they contain a physical representation of the main ideas responsible for protein stability. The predicted structures presented here show an overall agreement with the experimentally determined native structures and thus so far corroborate our assumptions on protein folding. Deviations from the known structure data were small in the avian pancreatic polypeptide. The differences may have as much to do with the experimental conditions as with the construction of our reduced protein model. The results for the parathyroid-hormone-related protein, point toward the inadvisability of representing a protein with our model that appears to be dominated by charge stabilization effects. The qualitative structure elements are however present. Obviously there is room for further adjustment of parameter space, especially through the incorporation of information from known structures into the parameter set, even though this would limit the predicting power of the model.

ACKNOWLEDGMENTS

A.W. thanks Drs. Bo Jönsson and G. Karlström for an extended period of financial support.

REFERENCES

1. Schulz, G. E., Schirmer, R. H. "Principles of Protein Structure." New York: Springer-Verlag, 1979.
2. Ghelis, C., Yon, J. "Protein Folding." New York: Academic Press, 1982.
3. Creighton, T. E. "Proteins." New York: Freeman, 1984.
4. Branden, C., Tooze, J. "Introduction to Protein Structure." New York: Garland, 1991.
5. Lesk, A. M. "Protein Architecture." Oxford: IRL Press, 1991.
6. Kim, P. S., Jernigan, R. L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. Ann. Rev. Biochem. 51:459–489, 1982.
7. Scheraga, H. A. Calculations of stable conformations of polypeptides, proteins and protein complexes. Chem. Scripta 29A:3–13, 1989.
8. Levitt, M. Protein folding. Curr. Opinions Struct. Biol. 1:224–229, 1991.
9. Thomas, D. J. Concepts in protein folding. FEBS 307:10–13, 1992.
10. Dill, K. A. Folding proteins: Finding a needle in a haystack. Curr. Opinions Struct. Biol. 3:99–103, 1993.
11. Brooks, C. L. Molecular simulations of peptide and protein unfolding: In quest of a molten globule. Curr. Opinions Struct. Biol. 3:92–98, 1993.
12. Hagler, A. T., Honig, B. On the formation of protein tertiary structure on a computer. Proc. Natl. Acad. Sci. U.S.A. 75:554–558, 1978.
13. Miyazawa, S., Jernigan, R. L. Equilibrium folding and unfolding pathways for a model protein. Biopolymers 21: 1333–1363, 1982.
14. Wilson, C., Doniach, S. A computer model to dynamically simulate protein folding: Studies with crambin. Proteins 6:193–209, 1989.
15. Crippen, G. M., Snow, M. E. A 1.8 Å resolution potential function for protein folding. Biopolymers 29:1479–1489, 1990.
16. Sippl, M. J. Calculations of conformational ensembles from potential of mean force. An approach of the knowledge-based prediction of the local structure in globular proteins. J. Mol. Biol. 213:859–883, 1990.
17. Covell, D. G. Folding protein α-carbon chains into compact forms by Monte Carlo methods. Proteins 14:409–420, 1992.
18. Godzik, A., Kolinski, A., Skolnick, J. Topology fingerprint approach to the inverse folding problem. J. Mol. Biol. 227: 227–238, 1992.
19. Ota, M., Saito, N. Prediction of the tertiary structure of the parathyroid-hormone-related protein (residues 1–34) by the island model. J. Prot. Chem. 11:623–628, 1992.
20. Sun, S. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. Protein Sci. 2:762–785, 1993.
21. Smit, B., Esselink, K., Hilbers, P. A. J., van Os, N. M., Rupert, L. A. M., Szleifer, I. Computer simulation of surfactant self-assembly. Langmuir 9:9–11, 1993.
22. Go, N., Taketomi, H. Respective roles of short- and long-range interactions in protein folding. Proc. Natl. Acad. Sci. U.S.A. 75:559–563, 1978.
23. Dill, K. A. Dominant forces in protein folding. Biochemistry 29:7133–7155, 1990.

24. Ptitsyn, O. B. How does protein synthesis give rise to 3D-structure? FEBS Letts. 285:176–181, 1991.
25. Sharp, K. A. The hydrophobic effect. Curr. Opinions Struct. Biol. 1:171–174, 1991.
26. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. Proteins 6:87–103, 1989.
27. Rullmann, J. A. C., and van Duijnen, P. Th. Potential energy models of biological macromolecules: A case for ab initio quantum chemistry. Rep. Mol. Theory 1:1–21, 1990.
28. Wallqvist A. WIGGLE 1.0: Polymer and Protein Simulation Program. University of Lund, 1993.
29. Honeycutt, J. D., Thirumalai, D. Metastability of the folded states of globular proteins. Proc. Natl. Acad. Sci. U.S.A. 87:3526–3529, 1990.
30. Guo, Z., Thirumalai, D., Honeycutt, J. D. Folding kinetics of proteins: A model study. J. Chem. Phys. 97:525–535, 1992.
31. Honeycutt, J. D., Thirumalai, D. The nature of folded states of globular proteins. Biopolymer 32:695–709, 1992.
32. Shortle, D., Chan, H. S., Dill, K. A. Modeling the effects of mutations on the denatured states of proteins. Protein Sci. 1:201–215, 1992.
33. Ullner, M., Selander, M., Persson, E., Stenflo, J., Drakenberg, T., Teleman, O. Three-dimensional structure of the Apo form of the N-terminal EGF-like module of blood coagulation factor X as determined by NMR spectroscopy and simulated folding. Biochemistry 31:5974–5983, 1992.
34. Levitt, M. Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 226:507–533, 1992.
35. Wallqvist, A., Berne, B. J. Hydrophobic interaction between a methane molecule and a paraffin wall in liquid water. Chem. Phys. Lett. 145:26–32, 1988.
36. Wallqvist, A. Molecular dynamics study of a hydrophobic aggregate in an aqueous solution of methane. J. Phys. Chem. 95:8921–8927, 1991.
37. Eisenberg, D., Wesson, M., Yamashita, M. Interpretation of protein folding and binding with atomic solvation parameters. Chem. Scripta 29A:217–221, 1989.
38. Wesson, L., Eisenberg, D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci. 1:227–235, 1992.
39. Linse, P. Stacked or T-shaped benzene dimers in aqueous solution? A molecular dynamics study. J. Am. Chem. Soc. 114:4366–4373, 1992.
40. Marqusee, S., Robbins, V. H., Baldwin, R. L. Unusually stable helix formation in short alanine-based peptides. Proc. Natl. Acad. Sci. U.S.A. 86:5286–5290, 1989.
41. Fairman, R., Armstrong, K. M., Shoemaker, K., York, E. J., Stewart, J. M., Baldwin, R. L. Position effect on apparent helical propensities in the C-peptide helix. J. Mol. Biol. 221:1395–1401, 1991.
42. Finkelstein, A. V., Badretdinov, A. Y., Ptitsyn, O. B. Physical reasons for secondary structure stability: α-Helices in short peptides. Proteins 10:287–299, 1991.
43. Scholtz, J. M., Qian, H., York, E. J., Stewart, J. M., Baldwin, R. L. Parameters of helix-coil transition theory for alanine-based peptides of varying chain lengths in water. Biopolymers 31:1463–1470, 1991.
44. Scholtz, J. M., Baldwin, R. L. The mechanism of α-helix formation by peptides. Annu. Rev. Biophys. Biomol. Struct. 21:95–118, 1992.
45. Stellwagen, E., Park, S.-H., Shalongo, W., Jain, A. The contribution of residue ion pairs to the helical stability of a model peptide. Biopolymers 32:1193–1200, 1992.
46. Handford, P. A., Baron, M., Mayhew, M., Willis, A., Beesley, T., Brownlee, G. G., Campbell, I. D. The first EGF-like domain from human factor IX contains a high-affinity calcium binding site. EMBO J. 9:475–480, 1990.
47. Creighton, T. E., Weissman, J. S., Kim, P. S. The disulfide folding pathway of BPTI. Science 256:111–114, 1992.
48. Goldenberg, D. P. Native and non-native intermediates in the BPTI folding pathway. TIBS 17:257–261, 1992.
49. Allen, M. P., Tildesley, D. J. "Computer Simulation of Liquids." Oxford: Clarendon, 1987.
50. Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P., Wu, C.-W. X-ray analysis (1.4-Å resolution) of avian pancreatic polypeptide: Small globular protein hormone. Proc. Natl. Acad. Sci. U.S.A. 78:4175–4179, 1981.
51. Barden, J. A., Kemp, B. E. NMR study of a 34-residue N-terminal fragment of the parathyroid-hormone-related protein secreted during humoral hypercalcemia of malignancy. Eur. J. Biochem. 184:379–394, 1989.