Correlation between rate of folding, energy landscape, and topology in the folding of a model protein HP-36

Arnab Mukherjee and Biman Bagchi

Solid State and Structural Chemistry Unit, Indian Institute of Science, Bangalore-12, India

We explore the correlation between the energy landscape and topology in the folding of a model protein (chicken villin headpiece HP-36) by using a force-field which incorporates the effects of water through a hydropathy scale and the role of helical propensity of amino acids through a nonlocal harmonic potential. Each amino acid is represented by one side chain atom which is attached to the backbone C_{α} atom. Sizes and interactions of all the side chain residues are different and depend on the hydrophobicity of a particular amino acid, whereas helical propensities are incorporated in the interaction of C_{α} atoms. Simulations have been carried out by quenching from a fixed high temperature to two different low temperatures for many initial random configurations. The simulated structures resemble the real native state rather closely, with the root mean square deviation of the best structure being 4.5 Å. Moreover, the structure shows both the helices and bends at the appropriate positions of the model protein. The simplified model allows the study of energy landscape and also of the correlation between energy landscape with the dynamics of folding and topology. The initial part of folding is very fast, followed by two distinct slow stages, with the last stage being certainly the rate determining of the folding process. The initial fast dynamics is primarily due to hydrophobic collapse. The very slow last stage of folding is accompanied by a significant and sharp increase in the relative contact order parameter but relatively small decrease in energy. Analysis of the time dependence of the formation of the individual contact pairs show rich and complicated dynamics, where some contacts wait for a long time to form. This seems to suggest that the slow late stage folding is due to long range contact formation and also that the free energy barrier is entropic in origin. Results have been correlated with the theories of protein folding.

I. INTRODUCTION

Prediction of protein structure from its sequence is a most challenging and inquisitive problem of molecular biology. Many theoretical and experimental studies have been carried out in order to understand the mechanism or the path which leads an unfolded protein to the final folded state which is apparently its most stable state, known as the native state.^{1–20} The importance behind the quest of the nature of protein folding is that the native structure of a protein is closely related to its functionality and thus there is a prospect of tremendous potential scientific achievement in predicting the structure of the native state from the amino acid sequence.¹

Protein folding is a collective self-organization process which could occur by a multiplicity of routes down a folding funnel.^{3–5} A global statistical characterization of the folding funnel is fundamental to the understanding of protein folding. However, the funnel is not structurally featureless.⁶ The ensemble of structures contains ordering in different regions of the protein to various extent. These structural diversities in the energy landscape can be probed with the help of other topological characteristics such as hydrophobic topological contact, radius of gyration, etc.

Dynamics of protein folding is intimately connected with the issue of the Levinthal paradox which says that a protein can take an astronomically long time to fold if it needs to search all the possible configurations available to it.⁷ The search is clearly not random and the incorporation of correct energy bias reduces the astronomically large time to a biologically significant time.⁸ The slow dynamics of folding mediated by the internal motions in protein is the direct consequence of the complexity of the underlying energy landscape.⁹ The energy landscape of a foldable protein resembles a multidimensional rugged funnel with many local minima but overall free-energy gradient pointing toward the native structure.¹⁰ The connection between the energy landscape theory and real proteins is best established in the context of small fast folding proteins which fold on a millisecond or less time scale and have a single folding domain, i.e., they are two state folders with a well defined barrier.¹¹

Many theoretical studies have been carried out on models of small single domain proteins. The early statistical mechanical theories of Dill and co-workers^{2,12} and of Bryngelson and Wolynes⁵ were based on heteropolymer collapse and reordering among hydrophilic and hydrophobic residues. The main result of this class of theories is that proteins with funnel landscape have population dynamics that can be understood as a diffusion of an ensemble of configurations over a low dimensional free energy surface.^{5,4,13} These energy surface may be constructed by using several different order parameters such as topological contact, radius of gyration, etc.¹⁴ Zwanzig calculated folding kinetics in a simple model based on the "correctness" of the native contacts and showed that there is a large free energy barrier near the folded state and folding time has a maximum near the fold-ing transition temperature.¹⁵ More recently Wolynes and co-workers have presented a detailed microscopic theory on the rates of protein folding.^{16,17}

All atom simulation of protein folding is computationally expensive. Therefore, simple lattice and off-lattice models of protein were used in the early work of Levitt and Warshel.¹⁸ The lattice model, even being the simplest possible protein folding model, could still capture many of the essential characteristics of the folding problem and the prediction of the tertiary structure.^{12,19,20} With the help of offlattice models, both the dynamics and the structural aspects of protein folding are possible to monitor. A recent off-lattice model study of HP-36 based on hydrophobicity tried to correlate the folding with many important equilibrium properties.²¹ Honeycutt and Thirumalai showed the folding of a model protein into a β -barrel structure and the existence of many metastable minima having similar structural characteristics but different energetics.²² Earlier a detailed study of the protein structure was carried by Levitt using an off-lattice model.²³ Scheraga *et al.* extended the scope of the off-lattice model in mimicking the protein structure for more than 100 residue proteins with very good agreement with the real native structures using a complicated force field and applying the conformational space annealing technique to reach the global minimum.^{24,25} All atom molecular dynamics and Monte Carlo simulations also have been carried out to get rigorous detail of the structure and dynamics of protein folding. All atom molecular dynamics simulation study of Duan and Kollman on the HP-36 protein with explicit water showed that the folding process contains two distinct pathways.²⁶ The real native structure of the HP-36 protein has been mimicked with close agreement by Hansmann and Wille by an all atom Monte Carlo simulation with global optimization.²⁷ They showed that without the solvent accessible surface energy term, the native state may not be the lowest energy state.

Recently an interesting correlation between the rate of folding and relative contact order (RCO) has been discovered.²⁸ A study on the rates of a large number of proteins are found to depend exponentially on RCO, Rate $\sim \exp(-\text{RCO})$. This finding agrees with the earlier suggestion of Dill that local contact formation occurs first during protein folding, followed by the contacts which are distant along the contour.²⁹ This correlation between rate of folding and RCO is a clear indication of entropic nature of the large state of folding.

Although considerable progress has been made in understanding many aspects of protein folding, there are still many important questions that remain to be resolved. While it is clear that each protein can have its own unique pathway, there could be features which are common to many proteins. A relevant question is the nature of the free energy barrier that the folding pathway is supposed to face after it has completed the initial collapse. Is it entropic or energetic? If entropic, what is its precise origin and how is this barrier overcome? In this article, we explore these questions by simulating a model protein which allows explicit calculation of the relevant quantities.

We have used a simple off-lattice model to study the energy landscape and topology of a model of chicken villin headpiece (HP-36) protein. Construction of our model protein is motivated by the hydrophobicity of different amino acids and formation of the hydrophobic core in the folded state. Since the work of Kauzmann,³⁰ it is believed that the hydrophobic interactions play an important role in organizing and stabilizing the architecture of proteins. This is related to the relative insolubility of the nonpolar residues in water.³¹ It is now widely accepted that the hydrophobicity is the dominant force of protein folding.^{29,32} The hydrophobicity of different amino acids can be arranged along a hydropathy scale. There are many different hydropathy scales which come from different ways of calculating the hydrophobicity. Janin and Rose et al. constructed a hydropathy scale by examining proteins with known 3D structures and defining the hydrophobic character as the tendency for a residue to be found inside of a protein, rather than on its surface.^{33,34} Another way of construction of the hydropathy scale was done by Wolfenden et al.³⁵ and Kyte and Doolittle³⁶ from the physicochemical properties of amino acid side chains and, therefore, more clearly follow the trends that would be expected on the inspection of amino acid structures. According to the scale of Kyte and Doolittle, hydrophobicities of amino acids have been relatively quantified by a value called hydropathy index. In our model, this hydropathy index has been mapped linearly into the interaction of the amino acids in such a way that the most hydrophobic and most hydrophilic amino acids have highest and lowest interactions among themselves, respectively.

Another important aspect of our model is the incorporation of the helix propensity rule into the intermolecular potential. The α -helix plays an important role in the early stages of protein folding and it is the most prevalent type of secondary structure found in proteins.³⁷ It has been observed that there is a correlation with the frequency of amino acid at a particular position of protein helix. Chou and Fasman made this viewpoint clear by proposing that the location of the protein helix could be predicted from an amino acid sequence and helix propensities.³⁸ Helix propensity of a particular amino acid is a measure of how its side chain influences the conformation of the peptide backbone.³⁹ Note that in Zimm and Bragg theory, the peptide group is the basic helix forming unit.⁴⁰ In the model studied here, an amino acid is represented by two atoms, one mimicking the C_{α} atom and another representing the whole side chain residue. No explicit peptide group has been taken into account in this model. In this simple model, helix formation is taken into account by introducing a nonlocal harmonic potential which facilitates the formation of the α -helix. To form the helix preferentially, helix propensity has been mapped linearly to the strength of the helix forming harmonic potential. It has been observed that with the help of the helix forming potential with proper helix propensity, it is much easier to form the helix preferentially when the sequence is favorable.

In the absence of explicit water, Brownian dynamics simulations could be carried out for many different initial



FIG. 1. Basic construction of the model protein is shown. C_{α} atoms are numbered as 1, 2, 3, etc., whereas the side residues are shown by 1', 2', 3', etc. Note the varying size of the side residues.

configurations. The protein is first equilibrated at high temperature and then suddenly temperature is quenched to a very low value. Dynamics of folding is monitored for the decrease in energy and radius of gyration. We have used the conjugate gradient technique to find out the corresponding underlying minima for all the states obtained by quenching. An increase in the number of hydrophobic topological contact is shown to be very sharp and clearly follow the decrease in energy. Dynamics of relative contact order shows the formation of nonlocal contact with the progress of folding.

Final folded states are analyzed for the construction of statistical folding funnel and other different topological quantities. Probability distributions of energy and topological contact shows Gaussian distribution. Stability of the folded protein increases with the decrease in radius of gyration up to a certain value. The probability distribution of the radius of gyration shows a non-Gaussian distribution which peaks around the experimental value. The energy of the low energy states are in the right range, giving credence to the model potential employed (hydropathy scale and helix potential).

Perhaps the most important result is the finding of the correlation between a very slow late stage folding and the variation in the RCO parameter. The very slow last stage of folding is accompanied by a significant and sharp increase in the RCO parameter but relatively small decrease in energy. This seems to suggest that the slow late stage folding is due to long range contact formation. This in turn implies that the free energy barrier is entropic in origin because long range contact formation is of low probability.

The rest of the paper is arranged as follows: In Sec. II, the model is described in full detail. Section III contains the detailed description of the force-field. In Sec. IV simulation details are described. In Sec. V, statistical properties obtained from many different Brownian dynamics simulations is described. Section VI contains the dynamical studies. A high temperature quench study is discussed in Sec. VII. In Sec. VIII, a correlation is established with the recent theories. Finally, we close the paper with a few conclusions in Sec. IX.

II. DESCRIPTION OF THE MODEL

The model presented here consists of two atoms per amino acid residue along the sequence (see Fig. 1). In this figure, the smaller atom represents the backbone C_{α} atom

while the other atom mimics the whole side chain residue. Each C_{α} atom is connected to two other C_{α} atoms (except the end ones which are connected to only one C_{α} atom) and one side chain residue atom. All the bond lengths, bond angles, and torsional angles are flexible. Figure 1 represents the basic construction of the model protein. Backbone atoms are numbered as i's, where $i = 1, 2, 3, \ldots, 36$, whereas side residues are numbered as i''s, where $i' = 1', 2', \dots, 36$, etc. Each i and i' together represent one amino acid. It should be pointed out here that this type of model was first introduced by Levitt.¹⁸ Similar types of models have recently been considered by Scheraga et al.²⁴ also but, contrary to the above two types of models, here there are no peptide groups present in our model and the interactions between the amino acids are determined by the hydrophobicity and the helix propensity of the amino acids. Most of our potential parameters have been taken from Levitt.¹⁸

A. Backbone atoms

Backbone atoms represent the C_{α} atoms of the real protein. Sizes and interactions of all the backbone atoms are kept the same. The size of each C_{α} atom is 1.8 Å and the interaction is 0.05 kJ mol⁻¹. Equilibrium distance between the C_{α} atoms is 3.81 Å as in the case of the real proteins. The equilibrium bond angle between the C_{α} atoms is kept at 96°.

B. Side chain residue

In this model, the atoms representing the side chain residue carry the identity of a particular amino acid. Side chain residues all have different sizes and interactions between other amino acids. Equilibrium bond length and bond angles also differ for different amino acids. Sizes and equilibrium bond angles of the side chain residues are taken from the values given by Levitt.²³ Interactions among the side chain residues, on the other hand, are based on the hydrophobicity of the amino acids. The hydrophobic amino acids interact much less with the solvent but more strongly among themselves; so they are more correlated. On the other hand a hydrophilic group with polarity and charge is prone to be exposed to the solvent; so the effective interaction among two strongly hydrophilic groups each surrounded by the solvent (water), should be much less than that between two hydrophobic groups. In the latter case, the effective interaction will increase because of the repulsion of the solvent (water). In principle, the above discussion can be quantified by defining the effective potential through the radial distribution function,41

$$V_{ii}^{\text{eff}}(r) = -k_B T \ln g_{ij}(r). \tag{1}$$

Hydropathy scale arranges the amino acid in terms of their hydrophobicity and the measure of hydrophobicity is given in terms of the hydropathy index.³⁶ The interactions of the side chain residues are mapped from the hydropathy index to the values between 0.2 and 11.0 kJ mol⁻¹ using a linear equation as given below,

$$\boldsymbol{\epsilon}_{ii} = \boldsymbol{\epsilon}_{\min} + (\boldsymbol{\epsilon}_{\max} - \boldsymbol{\epsilon}_{\min}) * \left(\frac{H_{ii} - H_{\min}}{H_{\max} - H_{\min}} \right), \tag{2}$$

TABLE I. Sizes and equilibrium bond angle values for all the different amino acids. The Kyte–Doolittle hydropathy scale and its translation to the LJ interaction parameter.

Amino acid	Size(Å)	Bond angle(θ°)	H_{ii}	ϵ_{ii} (kJ mol ⁻¹)
ala	4.60	121.90	1.8	7.76
val	5.80	121.70	4.2	10.64
leu	6.30	118.10	3.8	10.16
ile	6.20	118.90	4.5	11.00
cys	5.00	113.70	2.5	8.60
met	6.20	113.10	1.9	7.88
pro	5.60	81.90	1.6	7.52
phe	6.80	118.20	2.8	8.96
tyr	6.90	110.00	-1.3	4.04
trp	7.20	118.40	-0.9	4.52
asp	5.60	121.20	-3.5	1.40
asn	5.70	117.90	-3.5	1.40
gln	6.10	118.00	-3.5	1.40
his	6.20	118.20	-3.2	1.76
glu	6.10	118.20	-3.5	1.40
ser	4.80	117.90	-0.8	4.64
thr	5.60	117.10	-0.7	4.76
arg	6.80	121.40	-4.5	0.20
lys	6.30	122.00	- 3.9	0.92
gly	3.80	109.50	-0.4	5.12

where ϵ_{ii} is the interaction parameter of the *i*th amino acid with itself. ϵ_{\min} and ϵ_{\max} are the minimum and the maximum value of the interaction strength chosen for the amino acids. H_{ii} is the hydropathy index of *i*th amino acid and H_{\min} and H_{\max} are the minimum and maximum hydropathy index among all the amino acids, where, H_{\max} =4.5 and H_{\min} =-4.5. Table I shows the hydropathy index for all the different amino acids.

As discussed above, the interaction between strongly hydrophilic groups is small because of screening by the water molecules. We assume the lowest interaction parameter, $\epsilon_{ii} = 0.2 \text{ kJ mol}^{-1}$ for the most hydrophilic group (arginine) and the highest interaction parameter $\epsilon_{ii} = 11.0 \text{ kJ mol}^{-1}$ is chosen for the most hydrophobic group (isoleucine) according to the hydropathy scale.³⁶ Next, Eq. (2) is used to calculate the interaction parameters for other amino acids.

Interaction parameters, sizes, and equilibrium bond angles used in this model are shown in Table I for for 20 different amino acids. All the amino acids are divided between two groups. Amino acids with positive hydropathy index is defined hydrophobic whereas those with negative hydropathy index are taken to be hydrophilic. So the first eight amino acids in Table I are hydrophobic and the rest are hydrophilic amino acids. An interaction between two different amino acids are governed by the Lorentz–Berthelot rule, $\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}$.

III. FORCE FIELD

The model protein studied here contains energy contributions from various degrees of freedom because all the bond lengths, bond angles, and torsional angles are flexible in this model. There are other potential contributions from nonbonding and helix potential. The complete energy function V_{Total} for the model protein is sum of the bonding (V_B) , bending (V_{θ}) , torsional (V_T) , nonbonding (V_{LJ}) , and helix forming potential (V_{helix}) ,

$$V_{\text{Total}} = V_B + V_\theta + V_T + V_{\text{LJ}} + V_{\text{helix}}.$$
(3)

A. Bonding potential

Bonding potential is the sum of the bonding energy between the C_{α} atoms (backbone atoms) and the side chain residues with attached C_{α} atoms,

$$V_{B} = (1/2) K_{r} \sum_{i=2}^{N} (r_{i,i-1} - r_{0})^{2} + (1/2) K_{r}^{s} \sum_{i=1}^{N} (r_{i,i}^{s} - r_{0}^{s}(i))^{2}, \qquad (4)$$

where N is the total number of amino acid units present in the model protein and each amino acid unit contains two atoms, one C_{α} atom and another side chain residue atom, where

$$r_{i,i-1} = |\mathbf{r}_i - \mathbf{r}_{i-1}| \tag{5}$$

and

$$\boldsymbol{r}_{i,i}^{s} = |\boldsymbol{\mathbf{r}}_{i}^{s} - \boldsymbol{\mathbf{r}}_{i}|, \tag{6}$$

where r_i and r_i^s are the position of the *i*th backbone atom and the *i*th side residue, respectively. r_0 is the equilibrium bond length between the C_{α} atoms and it is equal to 3.81 Å. $r_0^s(i)$ is the equilibrium bond length between the *i*th C_{α} atom and the *i*th side chain residue. Values of $r_0^s(i)$ depend on the size of the side chain residue atoms. K_r is the force constant of the bonds between backbone atoms and is equal to 43.0 kJ mol⁻¹ Å⁻², whereas K_r^s is the force constant of the bond between backbone and side chain residue atom. K_r^s is taken equal to 8.6 kJ mol⁻¹ Å⁻².

B. Bending potential

The bending potential around a central atom is the sum of three bending potential terms involving two other backbone atoms and one side chain residue. For example, in Fig. 1, when 2 is the central atom, the bending angles will involve backbone atoms 1 and 3 and side chain residue 2',

$$V_{\theta} = (1/2) K_{\theta} \sum_{i=2}^{N-1} (\theta_{i-1,i,i+1} - \theta_0)^2 + (1/2) K_{\theta} \sum_{i=2}^{N} (\theta_{i-1,i,i}^s - \theta_0^s(i))^2 + (1/2) K_{\theta} \sum_{i=1}^{N-1} (\theta_{i,i,i+1}^s - \theta_0^s(i))^2,$$
(7)

where $\theta_{i-1,i,i+1}$ is the angle formed by r_{i-1}, r_i , and r_{i+1} . $\theta_{i-1,i,i}^s$ is the angle formed by r_{i-1}, r_i , and $r_i^s \cdot \theta_{i,i,i+1}^s$ is the angle formed by r_i^s, r_i , and $r_{i+1} \cdot \theta_0$ is the equilibrium bond angle between any three backbone atoms, whereas $\theta_0^s(i)$ is the equilibrium bond angle containing the *i*th side chain resi-

FIG. 2. The sequence of HP-36 is shown in the one letter code. Solid circles indicate hydrophobic and the open circles indicate hydrophilic amino acids.

due atom. The values of $\theta_0^s(i)$ is given in Table I. K_{θ} is the force constant for the harmonic bending potential and is taken to be 10.0 kJ mol⁻¹ rad⁻².

C. Torsional potential

There are four torsional angles per bond between two C_{α} atoms except the terminal bonds which contains only two torsional angles. Total torsional potential is given by

$$V_T = \epsilon_T \sum_{\phi} \ (1/2) [1 + \cos(3\phi)], \tag{8}$$

where $\epsilon_T = 1 \text{ kJ mol}^{-1}$.

D. Nonbonding potential

It is assumed that the nonbonding potential is given by the sum of a pair of Lennard-Jones interactions,

$$V_{\rm LJ} = 4 \sum_{i,j} \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right], \tag{9}$$

where the sum goes for the 2N number of atoms constituting the model protein, and N is the number of amino acids in the model protein. Sizes and interactions are the same for all the C_{α} (backbone) atoms. Size is equal to 1.8 Å and the interaction is 0.05 kJ mol^{-1} . On the other hand, sizes and interaction parameters for the side chain residues are different. Sizes are taken from Levitt²³ and the interaction is determined by the hydrophobicity of a particular amino acids. As already mentioned, interactions for the side chain residues are mapped from standard hydropathy index to values between 0.2 to 11.0 kJ mol^{-1} . The sequence of amino acids is shown in Fig. 2 in one-letter code. The solid circles denote hydrophobic while the open circles denote hydrophilic amino acids. Table I gives the sizes, interaction parameters, hydropathy index (H_i) , and equilibrium bond angle value for side residues (θ_0^s) .

E. Helix potential

There are several driving forces towards helix formation in proteins. Perhaps the most important one is the formation of hydrogen bond between two peptide groups separated by sequence of four amino acids⁴⁰ and another one is the polar and/or dipolar interaction. Note that the helix is not found to be the ground state for interacting homopolymers. At high stiffness, rod and toroid are the most stable forms⁴² but, in the case of real proteins, the helix becomes the ground state for a certain sequence of amino acids. In the HP-36 protein, the helix content is about 52%, i.e., 19 residues out of 36 total residues are in helix configuration. The present model does not contain the peptide group explicitly and all the atoms are taken to be neutral. In order to take into account the propensity of an amino acid to form an α helix, we have introduced an effective helix potential (described below) to efficiently incorporate the helix forming tendency.

Note that in an α helix, the distances between *i* and i+2 and between *i* and i+3 remain nearly constant. This observation is exploited by putting a harmonic potential between the above mentioned atoms. The form of the potential is as below,

$$V_{\text{helix}} = \sum_{i=3}^{N-3} \left[\frac{1}{2} K_i^{1-3} (r_{i,i+2} - r_h)^2 + \frac{1}{2} K_i^{1-4} (r_{i,i+3} - r_h)^2 \right],$$
(10)

where r_h is the equilibrium distance and is equal to 5.5 Å. One more interesting observation about the helix formation in a real protein is that the amino acids have different propensity to form the helix. The reason is both entropic and steric in origin. Many studies on helix propensity show that alanine has the maximum propensity to form the helix, whereas glycine has the minimum except proline which rarely forms the helix. Again the formation of the helix is not only dependent on the the helix propensity of an individual amino acid but also on its neighboring amino acids.

To incorporate the effect of helix propensity, the force constants of the above harmonic potential (K_{helix}) is obtained by mapping the helix propensity of a particular amino acid to a value between 17.2 kJ mol⁻¹ (alanine) and 0.0 kJ mol⁻¹ (glycine) by using the linear equation given below,

$$\mathcal{K}_i = \mathcal{K}_{\text{alanine}} - \operatorname{Hp}_i(\mathcal{K}_{\text{alanine}} - \mathcal{K}_{\text{glycine}}), \tag{11}$$

where Hp_{*i*} is the helix propensity value of the *i*th amino acid taken from Scholtz *et al.*⁴³ $\mathcal{K}_{alanine}$ and $\mathcal{K}_{glycine}$ are the force constants for alanine and glycine, 17.2 and 0.0 kJ mol⁻¹, respectively. $\mathcal{K}_{proline}$ having the least helix propensity becomes negative. The values of \mathcal{K}_i are given in Table II.

Next, the influence of the neighboring amino acids has been incorporated by introducing an effective helix propensity by taking an average force constants for *i*th amino acids defined as below,

$$K_i^{1-3} = \frac{1}{3} [\mathcal{K}_i + \mathcal{K}_{i+1} + \mathcal{K}_{i+2}]$$
(12)

and

$$K_i^{1-4} = \frac{1}{4} [\mathcal{K}_i + \mathcal{K}_{i+1} + \mathcal{K}_{i+2} + \mathcal{K}_{i+3}].$$
(13)

With the condition that K_i^{1-3} , $K_i^{1-4} \ge 0$ as the force constant must remain positive. Incorporation of the effective helix propensity takes care of the environment of an amino acid. So if proline or glycine situates in between two alanine groups, helix formation will be hindered considerably. The above formulation of helix potential is motivated by the work of Chou and Fasman about the prediction of helix for-

TABLE II. Basic spring constant values of the individual amino acids used in the V_{helix} potential. Values obtained from a linear mapping from the helix propensities.

Amino acid	$\mathcal{K}_i \; (\mathrm{kJ} \; \mathrm{mol}^{-1} \; \mathrm{\AA}^{-2})$	
ala	17.20	
glu	14.45	
leu	13.59	
met	13.07	
arg	13.59	
lys	12.73	
gln	10.49	
ile	10.15	
asp	9.80	
ser	8.60	
trp	8.77	
tyr	8.08	
phe	7.91	
val	6.71	
thr	5.85	
his	7.57	
cys	5.50	
asn	6.02	
gly	0.00	
pro	- 37.15	

mation. The neighbors of an particular amino acids were considered rather than the helix propensity of a particular amino acid.³⁸

IV. SIMULATION METHOD

Simulation of the folding of a model protein consists of two steps. First the initial configuration is generated by the configuration bias Monte Carlo method for a certain high temperature (1000 K). Using the initial configuration generated by the Boltzmann sampling, Brownian dynamics simulation is carried out to monitor the dynamics of the folding process.

A. Generation of the initial configuration

There are several sophisticated Monte Carlo techniques such as configurational bias Monte Carlo,^{44,45} pivot algorithm,⁴⁶ recoil growth technique,⁴⁷ parallel rotation algorithm,⁴⁸ etc. to generate lattice and off-lattice polymer configurations. Configuration bias Monte Carlo (CBMC) has been improved recently by Siepmann *et al.* to coupled–decoupled CBMC growth in order to incorporate branching in the polymer.⁴⁹

In this work, we have constructed our model HP-36 protein by the CBMC technique. The model contains branching at every backbone atom as the one side residue is attached to each of the backbone atoms except for the terminal ones. Independent generation of the beads attached to a single one results in incorrect distribution of bond angles.⁵⁰ So we have generated both the side chain residue and the next backbone atom attached to a particular backbone atom simultaneously to have a correct angle distribution as discussed by Dijkstra⁵¹ and Smit *et al.*⁵⁰ The resulting polymer has the correct bond angle distribution.

B. The Brownian dynamics simulation and folding

Brownian dynamics simulations have been performed on the CBMC generated initial configurations. Time evolution of the model protein was carried out according to the motion of each bead as below,

$$\mathbf{r}_{i}(t+\Delta t) = \mathbf{r}_{i}(t) + \frac{D_{i}}{k_{B}T}\mathbf{F}_{i}(t)\Delta t + \Delta \mathbf{r}_{i}^{G}, \qquad (14)$$

where each component of $\Delta \mathbf{r}_{i\alpha}^{G}$ is taken from a Gaussian distribution with mean zero and variance $\langle (\Delta r_{i\alpha}^{G})^{2} \rangle = 2D_{i}\Delta t.^{52,41} \mathbf{r}_{i}(t)$ is the position of the *i*th atom (both the backbone atom and side chain residue) at time *t* and the systematic force on the *i*th atom at time *t* is $\mathbf{F}_{i}(t)$. Δt is the time step used in the integration of the equation of motion. D_{i} is the diffusion coefficient of the *i*th particle, where D_{i} is calculated from the solvent viscosity η and the size of the particle according to the Stokes–Einstein relation given below,

$$D_i = \frac{k_B T}{6\pi \eta R_i},\tag{15}$$

where R_i is the radius of the *i*th atom. k_B and *T* are the Boltzmann constant and temperature, respectively. The unit of length is $\sigma(3.41 \text{ Å})$ and the unit of time is $\tau = \sigma^2/D_0$. D_0 is the diffusion coefficient obtained by using σ as the diameter in the above equation. τ is approximately 1.2 ns in the real unit for the reduced viscosity $\eta = 10$. The time step Δt is taken equal to 0.001τ . At first, the protein is equilibrated at a high temperature of 1000 K for 0.5 million steps. Then the temperature is quenched suddenly from 1000 to 20 K at t=0. The simulation was continued at 20 K temperature for 9.5 million steps. With the decrease in temperature, the protein starts to fold with time and the dynamical properties such as energy, radius of gyration, topological contact, etc. were monitored. The simulations have been carried out for \mathcal{N} number of different initial configurations, where $\mathcal{N}=584$.

V. STATISTICAL PROPERTIES

Statistical properties discussed below have been obtained by performing folding studies through Brownian dynamics simulations for \mathcal{N} different initial configurations. The conjugate gradient technique is performed on each of the \mathcal{N} folded states to get the minimized structures corresponding to each folded state. The \mathcal{N} final folded configurations and minimized structures were analyzed for the study of the distribution of energy, the structures of the folded states, topological contacts, radius of gyration, root mean square deviation, and the relative contact order.

A. Energy distribution

The probability distribution of the total potential energy $P(E_N)$ of the final quenched states for \mathcal{N} different initial configurations is plotted in Fig. 3(a). The width of the energy bin is taken as 4.0 kJ mol⁻¹. The distribution shows the Gaussian nature. The solid line in Fig. 3(a) shows the Gaussian fitting. Similar probability distribution for the energy of the *minimized configurations* $P(E_N^{\min})$ with a energy bin of 4.0 kJ mol⁻¹ is plotted in Fig. 3(b). This distribution also



FIG. 3. (a) Probability distribution of final energy $P(E_N)$ is plotted for \mathcal{N} folded (quenched) configurations. (b) Probability distribution for the minimized energy $P(E_N^{\min})$ is plotted. The solid lines in the above figures show the Gaussian fit.

shows a Gaussian behavior which is fitted to a Gaussian function shown by the solid line in Fig. 3(b). The difference between the two Gaussian distributions for the quenched configurations and the corresponding minimized configurations is in the position of the mean of the distribution. So the relative depths of the local minima from different energy levels of the folded energy states can be regarded as more or less the same.

The Gaussian energy distribution is a common phenomenon in case of the minimalist models. If one assumes that the quenched states obtained here indeed are the representative of the energy distribution, then the Gaussian distribution can be used to obtain the energy surface of folding. The rate of folding can then be obtained by a mean fast passage time calculation⁵³ as given below,

$$\tau(E_0) = \frac{1}{D(E)} \int_{E_0}^{E^{\neq}} dy e^{\beta E(y)} \int_{b}^{y} dz e^{-\beta E(z)},$$
 (16)

where *E* denotes energy which can be obtained from Gaussian distribution. We have placed a reflecting boundary at *b* on the left of E_0 and an absorbing boundary at E^{\neq} . D(E) is the diffusion coefficients in the energy space. One can also write Eq. (16) in terms of an order parameter like RMSD, if the diffusion coefficient of the RMSD is known. Bryngelson and Wolynes derived an expression for D(E).

In Bryngelson and Wolynes's⁵ spin glass model of protein folding, the energy distribution is Gaussian. Entropy of a particular state can be obtained from the degeneracy of the distribution. Entropy is maximum near the peak of the distribution. So the peak in the energy distribution corresponds to the entropically favorable states. It creates an entropic bottleneck for the movement towards the low energy native state.

In addition, the distribution is useful in providing an estimation of the relative stability of the native state. It can be used to calculate the Z-score of a protein which is a good measure of the validity of a knowledge-based potential.⁵⁴ Z-score of the misfolded states is calculated from the following expression,⁵⁴

$$Z_{\text{misfolds}} = \frac{\langle E \rangle_{\text{misfolds}} - E_{\text{native}}}{\sigma_{\text{misfolds}}},$$
(17)

where E_{native} is the energy of the native state, $\langle E \rangle_{\text{misfolds}}$ is the mean energy, and σ_{misfolds} is the standard deviation of the Gaussian distribution. Z_{misfolds} of the model protein studied in this paper is around 3.3 which is close to the experimentally obtained Z-score values.

B. Hydrophobic topological contact

The contribution that a hydrophobic residue makes to the stability of a protein varies roughly with the extent of its burial.⁵⁵ So the number of hydrophobic topological contacts can be well correlated with the stability of a protein. As discussed before, in the model HP-36 studied here, the amino acids having a positive hydropathy index are called hydrophobic while the ones having negative hydropathy index are defined as hydrophilic. This way, 16 out of total 36 residues in the HP-36 model protein are hydrophobic residues. The C_{α} atoms have not been taken into account here as they are identical with respect to the interaction parameter. A hydrophobic topological contact is formed if two hydrophobic side chain residues come within a distance of 8.5 Å. Figure 4(a)shows the correlation between the number of topological contact N_{topo} with the potential energy E_N of the system for $\mathcal N$ different initial configurations. There is a clear average increase in N_{topo} with the decrease in energy. The overall behavior can be understood from the straight line fitting. It shows an increasing slope of -0.08 with a decrease in one unit of energy.

The probability distribution of the topological contacts $P(N_{topo})$ is plotted in Fig. 4(b). This distribution also shows a Gaussian behavior and thus fitted to a Gaussian function shown by the solid line in Fig. 4(b). This can be understood from the Gaussian distribution of energy and the relationship of energy with the topological contact.

C. Relative contact order

Relative contact order (RCO) is defined as the average *sequence distance* between all pairs of contacting residues normalized by the total sequence length,⁵⁶



FIG. 4. (a) Number of topological contact between the hydrophobic groups, N_{topo} , is plotted against total energy E_N for \mathcal{N} number of configurations. The solid line shows a linear fit with a slope of -0.08. (b) Probability distribution of topological contact $P(N_{\text{topo}})$ is plotted using \mathcal{N} folded configurations. The solid line shows the Gaussian fit.

$$\text{RCO} = \frac{1}{NL} \sum_{ij}^{N} \Delta S_{ij}, \qquad (18)$$

where *L* is the number of contacts formed by the protein. *N* is the number of hydrophobic residues in the protein and $\Delta S_{ij} = |j-i|$, where *i* and *j* form a contact. As mentioned earlier, a contact is defined to be formed if the two hydrophobic side residues come within a distance of 8.5 Å. In the model protein studied here, N = 16.

RCO obtained from \mathcal{N} folded states are plotted against energy E_N in Fig. 5(a). The solid line is a fitting which shows an average increase in RCO with the decrease in energy. As RCO denotes the average contour contact distance, Fig. 5(a) signifies the increase in stability with more nonlocal contact formation. The probability distribution of relative contact order P(RCO) is plotted in Fig. 5(b) which shows a wide spread in the distribution showing the ensemble of states having many different levels of ordering. Note that Fig. 5(a) correlates the RCO with stability, not with rate.



FIG. 5. (a) Variation of relative contact order (RCO) with total energy is plotted for N different configurations. The solid line shows the linear fit with a slope of -0.0015. (b) Probability distribution of relative contact order P(RCO) is plotted which shows a wide distribution.

D. Radius of gyration

Figure 6(a) shows the probability distribution of the radius of gyration (R_{γ}) for \mathcal{N} number of different configurations. The peak of the distribution is around 9.5 Å. The reported value for the radius of gyration of real native HP-36 is also 9.6 Å.²⁷ The spread of the distribution is not large, from 8.5 to 11.5 Å. Note that this distribution is not Gaussian and skewed towards larger values of R_{γ} .

To correlate the relation between the compactness of a structure and its stability, energy is plotted as a function of radius of gyration for \mathcal{N} configurations. Figure 6(b) shows the expected decrease in radius of gyration with the decrease in energy.

E. Root mean square deviation

The structure of the real HP-36 protein obtained from the protein data bank⁵⁷ with the pdb id 1VII (Ref. 58) is compared to the model protein by calculating the root mean square deviation. First the center of masses of both the model and real protein are superimposed. Then keeping the real protein fixed, model protein is rotated with respect to all



FIG. 6. (a) Probability distribution of the radius of gyration $P(R_{\gamma})$ is plotted for all the \mathcal{N} folded configurations. The peak of distribution is around 9.5 Å. (b) R_{γ} is plotted against the total energy E_N for \mathcal{N} configurations. The solid line shows the linear fit with a slope of 0.01.

orthogonal axes (by all Euler angles), and at each point the RMSD is calculated from the equation below,

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (|\mathbf{r}_i^{\text{model}} - \mathbf{r}_i^{\text{real}}|)^2}, \qquad (19)$$

where $\mathbf{r}_i^{\text{real}}$ is the position of the *i*th C_{α} atom of the real HP-36 protein in its native state. $\mathbf{r}_i^{\text{model}}$ is the position of the *i*th C_{α} atom in the model protein studied here. *N* is the number of C_{α} atoms present in the protein, where N=36.

The lowest RMSD obtained in this fashion is taken as the RMSD of the model protein. Figure 7(a) shows the RMSD of all the \mathcal{N} different structures ($\mathcal{N}=584$) obtained by Brownian dynamics simulations. The calculated RMSD shows a trend of decreasing energy with lower RMSD. However, we found that the folded state with the lowest RMSD is *not* the lowest energy state. We attribute this to the neglect of the solvation energy contribution to the solvent accessible surface area. A recent study by Hansmann and Wille²⁷ showed that neglect of this solvation contribution can lead to an error in the stabilization energy similar to the one observed here.

In order to further quantify the structure of the folded configuration of the model protein, pair contact deviation (P_{RMSD}) is defined as



FIG. 7. (a) Solid circles show (a) RMSD [as defined in Eq. (19)] and (b) pair contact RMSD P_{RMSD} [defined in Eq. (20)] for \mathcal{N} folded states against the configuration number. The solid lines are to guide the eye.

$$P_{\text{RMSD}} = \sqrt{\frac{1}{NC_2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (r_{ij}^{\text{model}} - r_{ij}^{\text{real}})^2}, \qquad (20)$$

where $r_{ij}^{\text{model}} = |\mathbf{r}_j^{\text{model}} - \mathbf{r}_i^{\text{model}}|$ and $r_{ij}^{\text{real}} = |\mathbf{r}_j^{\text{real}} - \mathbf{r}_i^{\text{real}}|$. The quantity P_{RMSD} provides additional quantification of the spectrum of deviation of the folded protein structures from the internal structure of the real protein. Figure 7(b) shows the P_{RMSD} for \mathcal{N} different folded states.

F. Characterization of the folded structures

Many of the folded states have close similarity with the real native protein as can be seen from the RMSD values reported for all the \mathcal{N} folded structures in Fig. 7(a). However, there are many states, mostly high energy states, that have considerably higher RMSD values (more than 6 Å). These high RMSD states arise due to entanglement and less correlation between the hydrophobic residues. Folded states with lower RMSD values show a considerably high helix content and higher hydrophobic topological contacts and relative contact order.

A representative backbone structure of the folded states with the lowest RMSD amongst all the folded structures is shown in Fig. 8(a). Figure 8(b) shows the backbone structure of the real HP-36 protein. The model structure shown in Fig. 8(a) has the 4.5 Å RMSD as defined in Eq. (19). In spite of many shortcomings of the model protein such as the absence of all the atoms, charge, explicit water or even the peptide atoms, there is a good agreement between the model and the real protein structures. The model structure shows very high helix content as that of the real protein and the formation of helices and bends occur nearly at the appropriate positions. This can be attributed to the introduction of the helix potential with an environment dependent helix propensity.

Another interesting observation is shown in Fig. 8(c), where the folded structure of the model protein is shown



FIG. 8. (a) Backbone structure of the model protein with the lowest RMSD amongst all the N folded states. (b) Backbone structure of the real HP-36 in the native state as obtained from its pdb file. (c) Complete structure of the model protein for the lowest energy folded state. The dark shaded atoms denote hydrophobic while the light shaded atoms denote the hydrophilic side residues. Smaller atoms denote the C_{α} atoms. See the text for detail.

with both the backbone and side chain residue. The structure shows clearly that the hydrophobic beads (dark shaded atoms) are inside the structure forming the hydrophobic core while the hydrophilic beads (light shaded atoms) are outside. These observation could not be obtained in case of a simpler polymer bead model of protein.²¹

Comparison of the energy distribution with the distribu-



FIG. 9. The solid line shows the variation of energy with time. The circles show the minimized energies corresponding to a particular energy value at a certain time. Inset shows the magnified plot of the change in the minimized energy during folding.

tions of contact order and topological contact, etc. reveals that there are *different levels* of ordering present in the potential energy landscape of the protein. The degeneracies in the misfolded states due to different extent of contact formations (which show the deviation in the linear fit) imply that both the backbone topology and ordering of the side chain residue differ in different misfolded sates. One needs all the three quantities to characterize the landscape and the pathway (shown later in Fig. 17).

VI. DYNAMICAL STUDIES

The \mathcal{N} initial configurations generated by the CBMC were subjected to Brownian dynamics simulations to study the pathways of folding. The model protein was equilibrated first at a high temperature of 1000 K. Then at t=0, temperature was quenched from 1000 to 20 K and the folding dynamics of the protein was monitored. The dynamical studies discussed below are obtained from the dynamical evolution of the initial configuration which leads to the lowest energy state among \mathcal{N} folded configurations. Time dependence of various dynamical quantities reveals a *multistep* folding phenomenon. There is an initial fast hydrophobic collapse which is followed by slower decay. The final stage of folding occurs after a long plateau.

A. Time dependence of the potential energy variation

Figure 9 shows the different stages of decrease in the potential energy of the model protein (E_N) with time. An ultrafast initial decay is observed which corresponds to the hydrophobic collapse of the protein. This ultrafast stage is over within 10τ . Subsequently, a comparatively slower evolution follows until time 500τ . At the final stage of folding, a very slow rearrangement takes place among the side residues which causes the final decay to the lowest energy at around 2400τ . Thereafter, only thermal oscillations are observed around the lowest energy. So the maximum time (about 2000τ) is spent in the final stage.

The underlying evolving (that is, dynamic) energy landscape of the folding pathways has been analyzed by mini-



FIG. 10. The dynamics of energy for three different trajectories (a), (b), and (c) are plotted. In the inset, (a),(b),(c) show the corresponding minimized energies obtained for a particular value of energy for the trajectory (a), (b), and (c), respectively.

mizing the configurations of the model protein formed during its natural time evolution. Solid circles in Fig. 9 represent the energy of minimized structures E_N^{\min} corresponding to the time evolved configurations at a particular time. As the potential energy E_N decreases, energy of the underlying minima also changes to a lower value. At the initial time of collapse, the system goes over many minima, both low and high, but, at the latter stage of folding, *the system goes over the monotonically decreasing minima to reach the folded state*. The inset of the above figure shows the magnified diagram for E_N^{\min} for a shorter time window. It shows the change in the energy minima with time. The energy minima corresponding to the energy states after folding remain unchanged.

The dynamics of energy for a few more representative folding trajectories is shown in Fig. 10. The inherent structures along the folding trajectories are plotted in the inset of the same figure. As before, when the energy of the system decreases very fast in the short time, all the trajectories show fluctuations in the inherent structure energy. The slow relaxation part of the energy corresponds to a large waiting time at a particular local minimum before it changes to the another lower minimum.

B. Time dependence of the radius of gyration

The compactness of a structure can be understood quantitatively by monitoring the radius of gyration R_{γ} . So the time evolution of the radius of gyration can be regarded as a good measure of the dynamics of collapse of a protein. In the



FIG. 11. The solid line shows the decrease in radius of gyration (R_{γ}) with time. The circles with the dashed line show the R_{γ}^{\min} for the corresponding minimized configurations.



FIG. 12. Increase in N_{topo} with time is shown by the circles. The solid line is to guide the eye.

model protein studied here, the change in the radius of gyration with time is plotted in Fig. 11. The solid line shows the time evolution of R_{γ} , whereas the symbols with the dashed line show the R_{γ} for the corresponding minimized configurations. After quenching of temperature at t=0, there is a sudden decrease in R_{γ} which can be correlated with the huge fall in energy due to the hydrophobic collapse. A slower rate of decrease again follows after the initial impact which continues until 500 τ . Thereafter, the radius of gyration does not change for a long time until 2400τ , where a sudden decrease is observed again. The dynamical behavior of R_{γ} is consistent with that of the energy except that there is a small decrease in energy even in the plateau region of R_{γ} . This can be attributed to the detailed dynamical motions and the rearrangement of the atoms. R_{γ} leading to the locally minimized structures (inherent structures) also shows many oscillations, but there is an overall trend towards the more compact structure with folding.

C. Dynamics of contact formation

Hydrophobic topological contact N_{topo} as defined in the previous section, is formed when two hydrophobic side residues come within a distance of 8.5 Å. So the dynamics of N_{topo} can furnish a more detailed and microscopic aspect of folding than energy and radius of gyration. Figure 12 shows the increase in topological contact N_{topo} with time. N_{topo} also shows the similar dynamical behavior consistent with E_N and R_{γ} . There is a stiff initial increase in N_{topo} followed by a slower rate of formation which continues until 500 τ . Oscillations in terms of formation and breaking of contact continue until 2400τ when another overall increase in topological contact is observed. Note that, topological contact is formed by only the hydrophobic beads. So it can be concluded from the similar dynamical behavior of N_{topo} with that of energy and R_{γ} that mainly the hydrophobic residues govern the nature of protein folding. Formation and breaking of the contacts in the folded state due to thermal oscillations is observed in the long time.



FIG. 13. The circles show the increase in the relative contact order (RCO) with time. The solid line is to guide the eye.

D. Dynamics of growth in relative contact order

RCO is already defined in Eq. (18). Although relative contact order was introduced to correlate it with the rate of folding,⁵⁶ time dependence of the relative contact order serves the purpose of depicting the dynamics of folding with time. Figure 13 shows that the relative contact order increases with time. As the contact order is calculated only from the sequence separation, the increase in RCO with time signifies the progressive participation of residues *far from one another along the contour length* to form a contact with time. Time dependence of the RCO also shows a multistep process. A comparison of the RCO with N_{topo} suggests that the model protein studied here form a single hydrophobic core which is the characteristic of a globular protein.

E. Dynamics of folding through RMSD

Folding of the model protein can be quantitatively depicted by monitoring the RMSD of the structure with its native state. Here RMSD of the model structure is calculated with respect to the experimentally obtained native structure



FIG. 14. Variation of RMSD is plotted against time. Note that, RMSD decreases with time. The solid line is to guide the eye.



FIG. 15. Time dependence of the average contact pair correlation function $\overline{C}_{P}(t)$ is plotted. The solid line denotes the dynamics of $\overline{C}_{P}(t)$ for the hydrophobic side residues and the dashed line shows the same for the backbone atoms attached to the hydrophobic side residues.

of the real protein following Eq. (19). Figure 14 shows the decrease of RMSD along with time. Although the RMSD of the model protein does not reach a very low value, a significant decrease in RMSD with time is observed. The dynamical behavior of the RMSD is consistent with the time dependence of energy and other topological parameters which signify that the chosen model and the Brownian dynamics simulations are consistent.

F. Dynamics of distance pair correlation

Dynamics of folding can be probed microscopically by monitoring the dynamics of pair separation between different amino acids along the sequence. Dynamics will be different for different pairs. For many amino acid pairs, the distance of separation is expected to decrease as the folding occurs. The widely different time scales of movement of all the different pairs together give rise to an overall dynamics of folding which is reflected in the macroscopic quantities. Here, the effective dynamics of pair separation can be described by introducing a new pair correlation function as defined below.⁵⁹

$$C_{P}^{ij}(t) = \frac{d^{ij}(t) - d^{ij}(\infty)}{d^{ij}(0) - d^{ij}(\infty)},$$
(21)

where $d^{ij}(t) = r_i(t) - r_j(t)$. r_i and r_j are the positions of the *i*th and *j*th atom, respectively. *i* and *j* can be the indices of either backbone or side residue atom. Detailed dynamics of the contact pair correlation function $C_P^{ij}(t)$ will be described elsewhere.⁶⁰ Here we present an analysis of an average contact pair correlation function \bar{C}_P defined below as,

$$\bar{C}_{P}(t) = \frac{\sum_{j>i} C_{P}^{ij}(t)}{\sum_{j>i} C_{P}^{ij}(0)}.$$
(22)

Figure 15 shows the dynamics of the average contact pair correlation function $\overline{C}_P(t)$ for the hydrophobic side residues and the backbone atoms attached to them separately by the solid and dashed line. Time dependence of $\overline{C}_P(t)$ for both



FIG. 16. Time variation of energy is shown for two different trajectories (a) and (b) for a quench to a high temperature of 100 K. In the inset, (a) and (b) show the minimized energy values obtained for a particular energy for the trajectory (a) and (b), respectively.

the backbones atoms shows multistep relaxation processes, whereas the dynamics of $\overline{C}_P(t)$ for side residues shows more oscillations as a result of different dynamical behavior among different types of amino acids.

VII. EFFECT OF QUENCH-TEMPERATURE

A similar Brownian dynamics study has been carried out by quenching the initial configurations to a higher quench temperature (T_q) of 100 K, keeping all other parameters unchanged. The high T_q folding studies show similar dynamical behavior (the initial collapse and the slower long time decay) except that the intensity of fluctuations is large here and can be seen in Fig. 16. This is due to the high temperature. Even at the temperature 100 K, many structures resemble the real protein with low RMSD values (lowest observed is 4.2 Å). As expected, the magnitude of the change in the radius of gyration and contact order parameter is less (results not shown). Figure 16 shows the variation of energy with time for two different trajectories [(a) and (b)]. The inset shows the time variation of the minimized energy corresponding to the time evolution of potential energy of the two different trajectories. The system seems to explore many fluctuating local minima in the initial time of folding. However, here at high temperature, more frequent crossover of the local minima is observed (even in the very long time) leading the system towards the deeper minima with time (see the inset of Fig. 16).

VIII. CORRELATION WITH THE THEORY OF PROTEIN FOLDING

The dynamics of the relevant quantities discussed above can be correlated with the theories of protein folding. Early statistical mechanical theories by Dill and co-workers² and by Bryngelson and Wolynes⁵ were based on the heteropolymer collapse and reordering of the residues. Both the theories are based on two order parameter model of protein folding where the order parameters are packing fraction η and fraction ρ of residues in the native state. A simplified version



FIG. 17. Variation of the energy with radius of gyration R_{γ} and topological contact N_{topo} is plotted. (a), (b), (c) show three different folding trajectories. The increase in time along the curve is shown by the arrows.

of the the free energy function for the collapse and ordering transition can be obtained in terms of the two order parameters as given below,⁶¹

$$\frac{F\{\rho,\eta\}}{NT} = -\frac{1+\rho^2}{T}\eta + \rho\log\rho + (1-\rho)\log\frac{(1-\rho)}{N} + 1$$
$$+ \left(\frac{1}{\eta} - 1\right)\log(1-\eta) + \frac{3}{2}N^{-4/3}\eta^{-2/3}$$
$$+ \frac{2}{3N}\log\eta.$$
(23)

This free energy functional $F\{\rho, \eta\}$ shows two minima against ρ and η which are separated by a free energy barrier. The minimum for the low ρ and low η correspond to the extended state and the minimum for high ρ and high η correspond to the folded state.

The initial hydrophobic collapse observed here, followed by the slow ordering, is in good agreement with the above theories. The energy of the hydrophobic collapse has a steep decrease and hardly involves a barrier. The folding process encounters a barrier in the later stage where the ordering and rearrangement start to build up. So, there is a long waiting time before another drift occurs in the value the macroscopic variables such as energy, R_{γ} , topological contact, etc. As discussed before, from the microscopic point of view, the dynamics of pair separation has a very complicated dynamics spanning different time scales. Figure 17 shows the 3D diagram of energy with respect to R_{γ} and topological contact N_{topo} for three different folding trajectories. The arrows indicate the increase in time. Figure 17 shows an elegant description of the folding dynamics.

IX. CONCLUSION

Extensive Brownian dynamics simulations have been carried out using an off-lattice model for HP-36 protein on many different configurations to capture the structural, statistical, and dynamical aspects of protein folding. The structures obtained by the folding of the model protein resemble quite well the structure of the native state of real HP-36. The radius of gyration was also found to be close to the experimental value. The absence of water has been incorporated through the hydrophobicity of the amino acids and has been used extensively for all different amino acids. The absence of the peptide bond and H-bonding have been attempted to overcome by incorporating a new nonlocal harmonic potential with helix propensities.

The Gaussian distribution obtained through statistical sampling from many quenching simulation can be correlated with the underlying landscape, though the real dynamics may not reflect the characteristics of the landscape as the protein goes through many different saddles to avoid the minima. Hydrophobicity of the amino acids correlates very well with the stability of the protein. The dynamics of folding shows a very fast decay in energy initially which slows down considerably after some time. The underlying minima shows many oscillations in the energy value. Increase in number of topological contact and relative contact order shows the building up of nonlocal contacts with time and the decrease in radius of gyration shows the folding nature.

The main results are the multistage folding dynamics and the correlation of folding rate with topology.

Even being a simple model, the study could capture some of the structural and dynamical aspects of protein folding, though future work on this line should be done to get a clearer view.

ACKNOWLEDGMENTS

One of the authors (A.M.) thanks Ashwin S. Sampangiraj (JNC), Kausik Chakraborty (IISC,MBU), and Prasanth P. Jose (SSCU,IISC) for many helpful and valuable discussions both in the technical and conceptual level. He also thanks S. Roychowdhury and R. K. Murarka for discussion and help.

- ¹C. B. Anfinsen, Science **181**, 223 (1973); M. Sela, F. H. White, and C. B. Anfinsen, *ibid*. **125**, 691 (1957).
- ²K. A. Dill, D. O. V. Alonso, and K. Hutchinson, Biochemistry 28, 5439 (1989).
- ³P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
- ⁴J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).
- ⁵J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. 84, 7524 (1987); J. Phys. Chem. 93, 6902 (1989).
- ⁶B. A. Shoemaker, J. Wang, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 777 (1997).
- ⁷C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems*, in Proceedings of a Meeting held at Allerton House, Monticello, Illinois, edited by P. Debrunner, J. C. M. Tsibris, and E. Munck (University of Illinois Press, Urbana), p. 22.
- ⁸R. Zwanzig, A. Szabo, and B. Bagchi, Proc. Natl. Acad. Sci. U.S.A. **89**, 20 (1992).
- ⁹H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).
- ¹⁰ J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626 (1995).
- ¹¹J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, and P. G. Wolynes, Folding Des. 1, 441 (1996).
- ¹²K. Yue and K. A. Dill, Phys. Rev. E 48, 2267 (1993); H. S. Chan and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. 87, 6368 (1990).
- ¹³K. A. Dill and H. S. Chan, Nat. Struct. Biol. 4, 10 (1997).
- ¹⁴ H. Nymeyer, N. D. Socci, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. 97, 634 (2000).
- ¹⁵R. Zwanzig, Proc. Natl. Acad. Sci. U.S.A. **92**, 9801 (1995).

- ¹⁶J. J. Portman, S. Takada, and P. G. Wolynes, Phys. Rev. Lett. **81**, 5237 (1998).
- ¹⁷J. J. Portman, S. Takada, and P. G. Wolynes, J. Chem. Phys. **114**, 5069 (2001); **114**, 5082 (2001).
- ¹⁸M. Levitt and A. Warshel, Nature (London) 253, 694 (1975).
- ¹⁹J. Skolnick and A. Kolinski, Science **250**, 1121 (1990).
- ²⁰A. Kolinski, W. Galazaka, and J. Skolnick, Proteins 26, 271 (1996).
- ²¹G. Srinivas and B. Bagchi, J. Chem. Phys. **116**, 8579 (2002).
- ²²J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. 87, 3526 (1990); Biopolymers 32, 695 (1992).
- ²³ M. Levitt, J. Mol. Biol. 104, 59 (1976).
- ²⁴ A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, J. Comput. Chem. 18, 850 (1997).
- ²⁵ J. Pillardy, C. Czaplewski, A. Liwo *et al.*, Proc. Natl. Acad. Sci. U.S.A. **98**, 2329 (2001); J. Lee, A. Liwo, and H. A. Scheraga, *ibid.* **96**, 2025 (1999).
- ²⁶Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
- ²⁷U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. 88, 068105 (2002).
- ²⁸ V. Grantcharova, E. J. Alm, D. Baker, and A. L. Horwich, Curr. Opin. Struct. Biol. 11, 70 (2001).
- ²⁹ K. A. Dill, Biochemistry **29**, 7133 (1990).
- ³⁰W. Kauzmann, Adv. Protein Chem. **14**, 1 (1959).
- ³¹C. Tanford, *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, 2nd ed. (Wiley, New York, 1980).
- ³²E. S. Huang, S. Subbiah, and M. Levitt, J. Mol. Biol. 252, 709 (1995).
- ³³J. Janin, Nature (London) 277, 491 (1979).
- ³⁴G. Rose, A. Geselowitz, G. Lesser, R. Lee, and M. Zehfus, Science 229, 834 (1985).
- ³⁵ R. Wolfenden, L. Andersson, P. Cullis, and C. Southgate, Biochemistry 20, 849 (1981).
- ³⁶J. Kyte and R. F. Doolittle, J. Mol. Biol. 157, 105 (1982).
- ³⁷D. F. Sickle, L. G. Presta, K. A. Dill, and G. D. Rose, J. Mol. Biol. 226, 1143 (1992).
- ³⁸ P. Y. Chou and G. Fasman, Biochemistry **13**, 211 (1974).
- ³⁹A. Chakrabartty and R. L. Baldwin, Adv. Protein Chem. 46, 141 (1995).
- ⁴⁰B. H. Zimm and J. K. Bragg, J. Chem. Phys. **31**, 526 (1959).
- ⁴¹J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Academic, New York, 1986).
- ⁴²(a) H. Noguchi and K. Yoshikawa, J. Chem. Phys. **113**, 854 (2000); (b) **109**, 5070 (1998).
- ⁴³C. N. Pace and J. M. Scholtz, Biophys. J. 75, 422 (1998).
- ⁴⁴G. C. A. M. Mooij, D. Frenkel, and B. Smit, J. Phys.: Condens. Matter 4, L255 (1992); D. Frenkel and B. Smit, Mol. Phys. 75, 983 (1992).
- ⁴⁵ J. J. De Pablo, M. Laso, J. I. Siepmann, and U. Suter, Mol. Phys. **80**, 55 (1993).
- ⁴⁶N. Madras and A. D. Sokal, J. Stat. Phys. **50**, 109 (1988).
- ⁴⁷ S. Consta, T. J. H. Vlugt, J. W. Hoeth, B. Smit, and D. Frenkel, Mol. Phys. **97**, 1243 (1999); S. Consta, N. B. Wilding, D. Frenkel, and Z. Alexandrowicz, J. Chem. Phys. **110**, 3220 (1999).
- ⁴⁸S. Santos and U. W. Suter, J. Chem. Phys. **14**, 9772 (2001).
- ⁴⁹M. G. Martin and J. I. Siepmann, J. Phys. Chem. B 103, 4508 (1999).
- ⁵⁰ T. J. H. Vlugt, R. Krishna, and B. Smit, J. Phys. Chem. B **103**, 1102 (1999).
- ⁵¹M. Dijkstra, J. Chem. Phys. **107**, 3277 (1997).
- ⁵²D. L. Ermak and J. A. McCammon, J. Chem. Phys. 69, 1352 (1978).
- ⁵³R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, Oxford, 2001).
- ⁵⁴ J. U. Bowie, R. Luthy, and D. Eisenberg, Science **253**, 164 (1991); L. Zhang and J. Skolnick, Protein Sci. **7**, 1201 (1998).
- ⁵⁵D. Shortle, W. E. Stites, and A. K. Meeker, Biochemistry **29**, 8033 (1990).
- ⁵⁶K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. 277, 985 (1998).
- ⁵⁷ F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).
- ⁵⁸C. J. McKnight, D. S. Doering, P. T. Matsudaria, and P. S. Kim, J. Mol. Biol. **260**, 126 (1996).
- ⁵⁹G. Srinivas and B. Bagchi, Theor. Chem. Acc. (to be published).
- ⁶⁰A. Mukherjee and B. Bagchi, Phys. Rev. Lett. (to be published).
- ⁶¹R. Zwanzig and B. Bagchi (unpublished work).