UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/775-10

# COLLEGE IN BIOPHYSICS: EXPERIMENTAL AND THEORETICAL ASPECTS OF BIOMOLECULES

## 26 September - 14 October 1994

*Miramare - Trieste, Italy*

## *Protein Folding Theory*

**E. Shakhnovich**
**Harvard University, Cambridge, USA**

# Theories of protein thermodynamics

We come to the conclusion that theory of coil-globule transitions in homopolymeric macromolecules, even in its most elaborate form does not describe the most important features of protein denaturation:
1) First-order character of denaturational transition.
2) The existence of Molten Globule state in proteins.

This implies that some very essential protein specific had been omitted in the theoretical treatment of proteins.

This specifics might be:
1) The heterogeneity of protein chains : both "Flory-type" Ptitsyn theory of coil-globule transitions and Lifshitz theory treated protein as a honopolymeric chains.
2) Complicated architecture of a protein chain: presence of different degrees of freedom: backbone ($\Phi$, $\Psi$ angles ("polymeric" degrees of freedom) and degrees of freedom connected with side-chains motions ($\chi$-angles).
3) The most fundamental feature of proteins: existence of unique 3D-structure deserves fundamental theoretic investigation and this investigation may shed light to the origin of mysterious denaturational transition.

In fact, the investigation of the nature of the peculiar thermodynamic behavior of proteins became one of the most challenging goals for theoreticians working in this field and the attack followed all the three directions outlined above.

Generally, the models of protein thermodynamics can be divided into two groups. To the first group belong models in which no special interactions other than known (hydrophobic, electrostatic, van-der-Waals, etc) were assumed to act in proteins. The second group of models involved assumption about som special "hidden" local or non-local interactions which lead to

formation of unique 3D structure of proteins. Both groups of models considered as a cornerstone and major goal the description of the nature of high cooperativity of protein structure. We will discuss all these exhisting models in logical rather than chronological order.

## Heteropolymer collapse

The natural way of search of the nature of cooperativity is to take into account the **heterogeneity** of a protein chain. This was done within different theoretical treatments: the extension of Lifshitz theory for the case of heteropolymer ( (A.Yu. Grosberg & E.I.Shakhnovich, Soviet Phys JETP **64** 3821 (1986), E.I. Shakhnovich & A.M.Gutin Biophys.Chem **34** 187-199 (1989)) & A.M.Gutin J.Physique (France) **50** 1843-1850 (1989)) and the extension of "Flory-type" theory for heteropolymers. (K.A.Dill, Biochemistry **24** p. 1501 (1985) ). It was shown within the extension of Lifshitz theory that the coil-globule transition to the globule **without unique structure** takes place in the same manner as in homopolymer with the only difference that it occurs at somewhat higher temperature. The physical reason of this result is quite clear: When the chain (homopolymer or heteropolymer) collapses to the globule without unique well defined conformation it it explores numerous conformations (it is worth mentioning that number of compact conformations of a polymeric globule is still very large: change of chain entropy in the coil-globule transition is very small, according to the Lifshitz theory). A chain behaves like a living snake with all contacts possible and explored and energy of interactions between different monomers becomes averaged so that heteropolymer in this regime looks like homopolymer. The detailed mathematic treatment of heteropolymer collapse problem done in (A.Yu. Grosberg & E.I.Shakhnovich, Soviet Phys JETP **64** 3821 (1986)) develops renormalisation group technique based on the idea of taking into account of all possible fluctuations of contacts on any scale:

1). Interactions are separated into local (between, say s neighboring along

the chain monomers) and non-local.
2). Contribution from local interactions are calculated explicitely
3) Each s monomers along the chain are united into new monomers; we come to a new chain of N/s new monomers and new interactions between new monomers $E'ij$ which are obtained on stage 2 when local interactions were excluded.

This procedure shows that "new" chains becomes more and more homopolymeric as size of "new" monomer grows. In the vicinity of transition point "new" chain main consist of only one monomer with completely smoothed differences between types of monomers (recall that polymeric coils are fractal, or scaling invariant objects!). This implies that heterogeneity of the chain does not change the character of the coil-globule transition in that chain, or, as it usually said, the transition belongs to the same universality class for heteropolymer as for homopolymer.

The coil-globule transition within "Flory-type" approach has been considered in the work (K.A.Dill, Biochemistry **24** p.1501 (1985)). The process of collapse is separated into two stages:

1) Collapse of the chain as if it is homopolymer. This process was treated in the "Flory-type" theory (with erroneous entropic term) and derivation repeated such given in previous works.
2) Rearrangement of the compact conformation to immerse unpolar groups inside the globule and polar groups outside it.
It was claimed that the total transition is of the first order since the free energy is increased after the first stage and thus a molecule encounters a barrier on its way to final folded state. However this conclusion is an obvious artifact of artificial division of the folding process into two parts: in reality the process of collapse and immersion of unpolar groups inside occur simultaneously and there is no barrier on the way from collapsed state to unfolded state. In the latest review (preprint available) K.Dill mentions



3

that his treatment does not address the question whether the transition is of the first order. Moreover, the chain connectivity was not taken into account within this treatment, i.e. it was assumed that hydrophobic and hydrophilic groups can choose their positions on the surface of a protein or inside it as if they were not restricted by chain conformations.

## Theory of protein "melting": the role of side-chains.

We see now that theory of polymer collapse does not explain main features of protein thermodynamics observed experimentally. Moreover, a question arises, what is the source of large latent heat in this process since one can expect that protein-protein interactions may be substituted by protein-solvent ones and therefore energies would be compensated. Following analogy with non-polymeric substances one could consider coil state as analog of gas phase, m olten globule as an analog of liquid, then native state should obviously the analog of solid. Then one should expect latent heat of denaturation in the same manner as it exists in the melting process.

However it should be emphasized that theory that considers intramolecular melting of a protein should differ drastically from such for crystal-liquid transitions. The reason is that in crystal structure the main effect is destruction of long-range order, crystal lattice. Obviously, there is no long-range order in placement of side-chains and backbone in proteins; therefore mechanism of the melting transition should differ drastically.

The consistent theory of protein melting must give answer to the key question: why destruction of a native state means transition from this native state through free-energy barrier rather than smooth evolution of the native state itself. In other words, why intermediate states (between native and denatured) are unfavourable?

The theory which addresses these questions was suggested in (E.I.Shakhnovich & A.V. Finkelstein Biopolymers **28** pp. 1667-1680 (for melting in vacuum) and in E.I.Shakhnovich & A.V. Finkelstein Biopolymers **28** pp. 1681-1694 (for melting in solvent).

The model of a protein globule suggested there was inspired by the experimental fact that in many cases denaturation does not disrupt secondary structure, i.e. the chain remains rather stiff. Therefore a native protein globule (and the globule in the barrier state: solvent does not penetrate!) can be visualised as a stack of few rigid structural segments to which side-chains bearing numerous rotational degrees of freedom are attached. (see fig. )

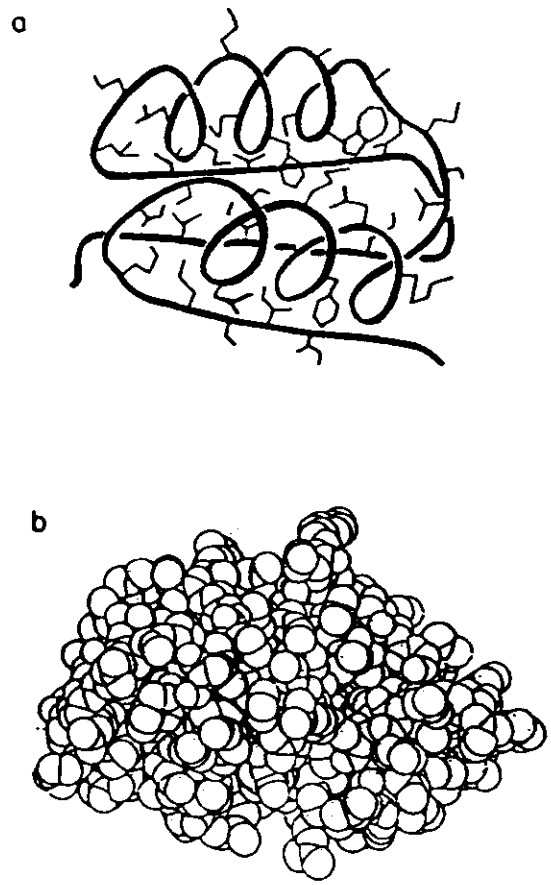



Fig. 1. Protein globule: The scheme (a) shows the backbone (solid line) forming secondary structure segments (here: two $\alpha$-helices and a $\beta$-sheet of three strands) connected by loops; the backbone is covered by the numerous side chains. The space-filling model (b) shows the compactness and tight packing of a protein globule.

Formation of tight packing of side chains a protein hydrophobic core is

favourable energetically since short-range interactions (van-der-Waals forces) cause attraction between groups. This process is unfavourable entropically since many degrees of freedom connected with rotational isomerisation of side chains which, besides usual rotational isomerisation potential, become restricted due to mutual steric hindrances inside tightly packed hydrophobic core. (see fig. )

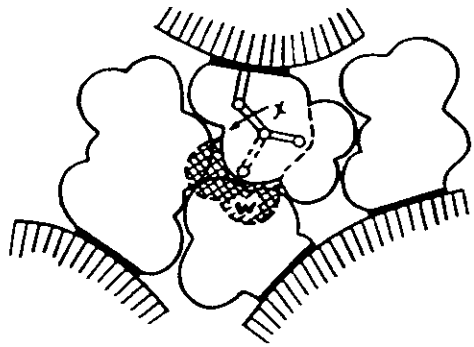


Fig. 2. Scheme of side-chain packing. Only a small part of the core is shown, including several chains attached to the structural segments. The shaded region is stands for the alternative of the central group ($\chi$ is the torsional angle). This rotamer is forbidden by the tight

Question arises how can destruction of tighly packed native conformation

occur? It is clear that intramolacular H-bonds (and, hence secondary structure) should be preserved during the initial stages of decompactisation while water molecules do not penetrate inside protein interior and form H-bond with polypetyde. Therefore first stages of denaturation in solvent should be indistingushable from such in vacuum. Therefore exit from the native state must occur via **uniform displacements of stable segments of secondary structure bearing numerous side- chains** .

The only terms of free energy which are changed during this process are van-der-Waals interactions and entropy of torsional motions of side-chains and loops. Consider these terms.

**The van-der-Waals interactions** have usual Lennard-Jones form:

$$E_w = \frac{1}{2}\sum_{i\neq j}\epsilon_{ij}\left[\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^0}{r_{ij}}\right)^6\right]$$

The **uniformity** of protein expansion implies that one macroparameter - volume of a globule V - governs all changes of pairwise vdW interactions via:

$$r_{ij} = r_{ij}^0\frac{V^{1/3}}{V_0^{1/3}}$$

with $V_0$ is the volume of most tightly packed protein.

This means that vdW energy depends on volume in a simple form:

$$E_w = \frac{1}{2}E_0\left[\left(\frac{V_0}{V}\right)^4 - 2\left(\frac{V_0}{V}\right)^2\right]$$

The plot of this dependence is shown in fig.



7

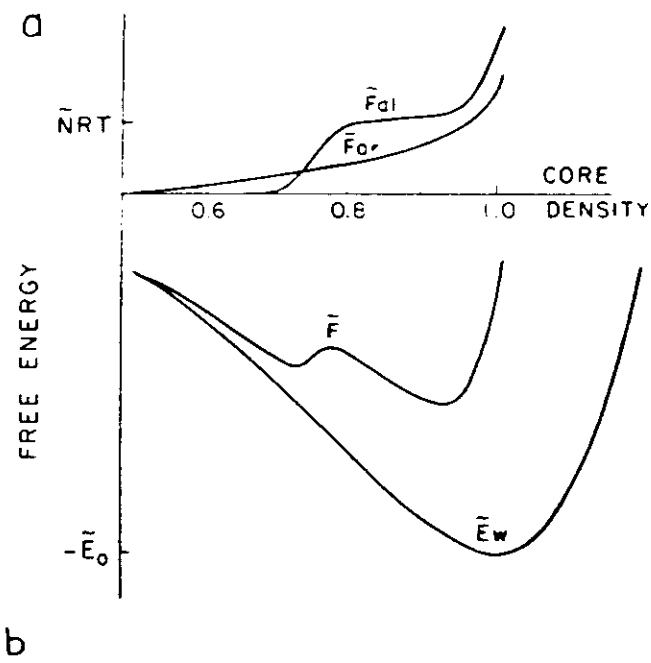



Fig. 3. The plot of free energy (a) and the main components of the internal pressure (b) vs the relative density $\tilde{V}_0/\tilde{V}$ of the hydrophobic core. The total free energy $\tilde{F}$ consists of the van der Waals energy $\tilde{E}_w$ and the free energy of torsional motions ($\tilde{F}_{al}$ for the aliphatic and $\tilde{F}_{ar}$ for the aromatic degrees of freedom). For the evaluation we have assumed that $\tilde{N} = 50$ side chains of the core have $\tilde{M}_{al} = 80$ aliphatic and $\tilde{M}_{ar} = 20$ aromatic degrees of freedom; for the other parameters, see the text.

**Torsional motions**

Each side-chain moves under action of two forces:

1) its own torsional potential $U_i(\chi)$ which is different for aromatic and aliphatic side-chains and

2) steric hindrances from surrounding side-chains and the backbone.

For aliphatic side-chains

$$U_{al}(\chi) = \frac{1}{2}U_0(1 - cos3\chi)$$

with $U_0 = 3kcal/m$ and for aromatic side-chains

$$U_{ar}(\chi) \ll RT$$

Each rotation is limited in some range

$$\chi_i^0 < \chi_i < \chi_i 1$$

by collisions with neighbouring segments. (see fig. )

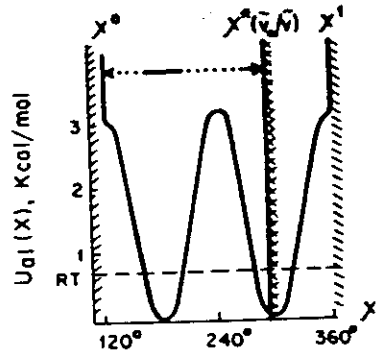



Fig. 4. The internal torsional potential $U$ as a function of the torsional angle $\chi$ for aliphatic degrees of freedom. The shaded regions are forbidden due to collision of the side chain with its own backbone. The sterically allowed (at a given $\tilde{V}_0/\tilde{V}$) region of $\chi$ is shown by the arrow; the dotted parts of the arrow correspond to the regions that are practically forbidden by the high torsional potential.

We may write the contribution to the free energy from torsional motions:

$$F_t(V) = \sum_{i=1}^{M} \ln z_i(V_0/V)$$

with

$$z_i(V_0/V) = \int_{\chi_i^0}^{\chi_i^*} \exp\left[-\frac{U_i(\chi)}{RT}\right]$$

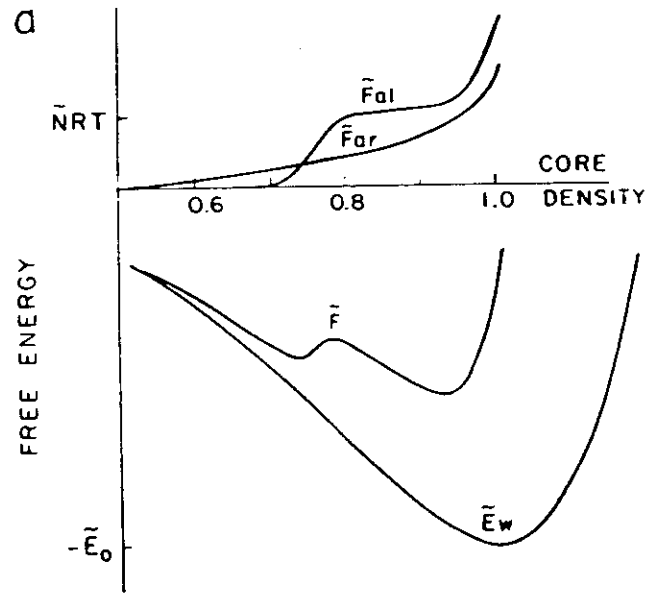
is a partition function of rotation of a side chain in torsional potential restricted by "steric walls".

Increase of volume corresponds to motion of the right steric wall to the right. There are three stages in changes of torsional free energy:

1) initial swelling with increase of amplitude and entropy of **librations**

2) Intermediate swelling when amplitude of motions is not increased being limited by potential of internal rotation

3) Further swelling when possibility to place second rotational isomere appeares and entropy increases drastically due to it. The plot of free energy of rotational motions is shown in fig.



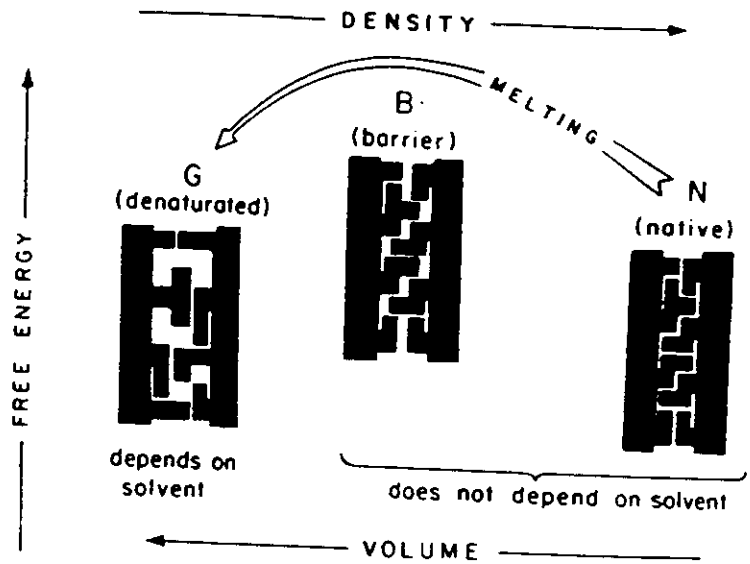



Fig. 6. Scheme of melting of a tightly packed protein core. The right part of this scheme is also valid for proteins in solution. The compactness of the denatured state depends on solvent.

The total free energy of van-der-Waals interactions and rotational motions $F = E_w + F_t$ is shown in the fig. above. It is clear from this fig. that indeed there exist two stable states native (N) and molten (D) separated by free energy barrier. This barrier corresponds to intermediate swelling when part of vdW interactions is lost but entropy of rotations, still hindered by rotational potential, is not gained. The volume of denatured state can be estimated from the condition that pores with free volume equivalent to volume of, say, $CH_2$ group should be formed to facilitate rotational isomerisation. This gives $V_D/VN \approx 1.3$. Obviously, barrier state B has smaller volume.

This investigation of protein denaturation in vacuum leads to the following results:

1) Denatured state of a protein is always compact.

2) Energy change (latent heat) $E_w(V_N) - E_w(V_D)$=100kcal/m.

2) Entropy change $\Delta S = RM$=200cal/m/deg.

3) Melting temperature $T_m = 500K$

## The role of solvent

Water molecule has volume close to such of $CH_2$. This implies that they cannot penetrate inside protein when it is in the barrier state. This is in correspondence with experimental results of Segawa and Sugikhara discussed in detail earlier. This means that presence of solvent does not influence on native and barrier state and hence all the qualitative features of the transition (and first of all the existence of the barrier itself!) are not changed due to the presence of solvent.

However, solvent is able to penetrate into denatured molecules and therefore it may change its density (or even swell it to coil) and may change energetic parameters of the denatured state and, hence of the transition. Therefore it is necessary to consider influence of solvent.

Generally speaking there may be two possibilities:
1) solvent does not penetrate into the interior of denatured protein: Molten Globule is "dry" and
2) solvent does penetrate into protein interior: Molten Globule is "wet". (see fig. ) In order to answer the question and find out how solvent influences the

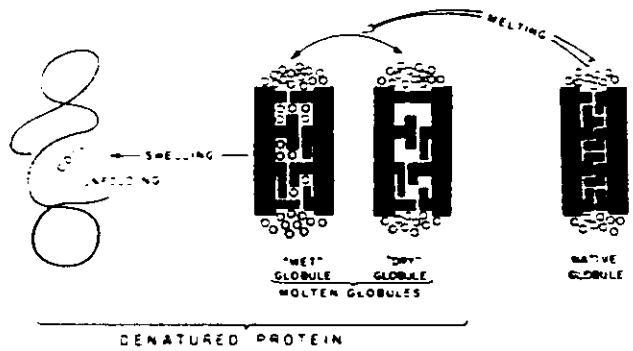



Fig. 1. Scheme of protein denaturation. The pores in the denatured globule may be either empty (dry globule) or may be filled with the solvent (wet globule).

denatured state we must include interaction of a protein with solvent. In fact this must take into account interaction of solvent molecules penetrated into

the protein interior, interaction between solvent molecules inside the protein, and entropy of placement of $N_{in}$ penetrated solvent molecules among $N_{pore}$ vacanvies existing inside.

Therefore the free energy of a denatured protein in a solvent looks like:

$$F(V, N_{in}) = F_{self} + N_{in}\left[\phi\frac{V_N}{V} + \frac{1}{2}\alpha\frac{N_{in}w}{V}\right] + RT\left[N_{in}\ln\frac{N_{in}}{N_{pore}} + (N_{pore} - N_{in})\ln\left(1 - \frac{N_{in}}{N_{pore}}\right)\right]$$

where $F_{self}$ is a free energy of a protein in vacuum, $V_N$ is a volume of a native protein, $w$ is a volume of a water molecule.

The equilibrium number of penetrated solvent molecules should be defined from the condition of equilibrium:

$$\mu_{in} = -\frac{\partial F(V, N_{in})}{\partial N_{in}} = \mu_{bulk}$$

where $\mu_{in,bulk}$ are chemical potentials of water inside protein and in a bulk solvent.

Parameter $\alpha$ characterizes interactions in a bulk solvent; it is connected with the chemical potential of the solvent via:

$$\alpha/2 = \mu_{bulk} - wP_0 = RT\ln\frac{\rho_{vap}}{\rho_{liq}}$$

with $P_0 = 1atm$ is an external pressure, $\rho_{vap,liq}$ are densities of saturating vapour and liquid solvent correspondingly; for water $\alpha \approx 12kcal/m$

The solution of equation of water equilibrium gives the result that the state of water inside protein is determined by the parameter

$$\Psi = \frac{\alpha}{2} - \phi$$

when $\Psi < 0$ $N_{in} \approx N_{pore}$, and globule is wet.

Of course, when solvent does not penetrate into protein it is globule. But what happens with protein when solvent starts to penetrate inside it? Will

it inevitably transfer globule to coil or there exists a possibility of we molten globule?

The answer is: yes, wet molten globule may (and must!) exist. Indeed, the self-interaction in the protein chain corresponds to repulsion: in vacuum protein will be compact. Therefore, additional swelling force from the penetrated solvent must overcome this attraction in order to convert globule to coil.

Addition of specific denaturants changes activity of a solvent either changing $\alpha$, i.e. destroing water structure or changing $\phi$-interactions between protein and solvent. The most suitable for the analysis is to follow how the curve F(V) changes when solvent is changed. This is shown in fig. for high and low temperatures.

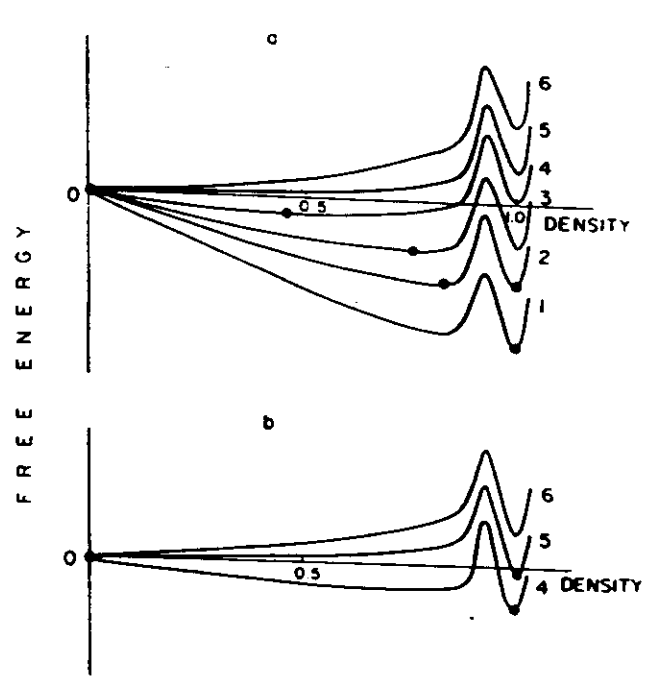



Fig. 3. Evolution of the plots of the free energy vs protein density. The plots correspond to two temperatures—(a) above and (b) below the triple point and to the different solvent qualities $\Psi_0/2 = \Psi_1 < \Psi_2 < \Psi_3 < \Psi_4 < \Psi_5 = \Psi_6 < \Psi_7$. The stable states are marked by solid circles.

Correspondingly we may follow what transitions occur when denaturant is added:

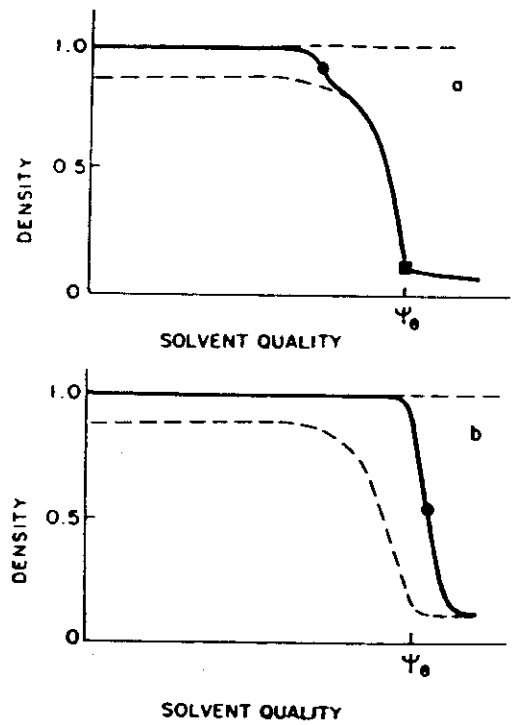



**Fig. 5.** Plot of the equilibrium density of a protein vs the solvent quality for two temperatures: (a) above and (b) below the triple point. The circles denote the midpoints of the first-order phase transitions. The square denotes the $\theta$ point, which is the point of the second-order phase transition (note that swelling is a gradual *pre*transitional effect). The dashed parts of the curves correspond to the metastable states. The density of the coil is $N^{-1/2}$.

All these results can be summarised in the phase diagram plotted in "theoretical" coordinates $T - \Psi$:

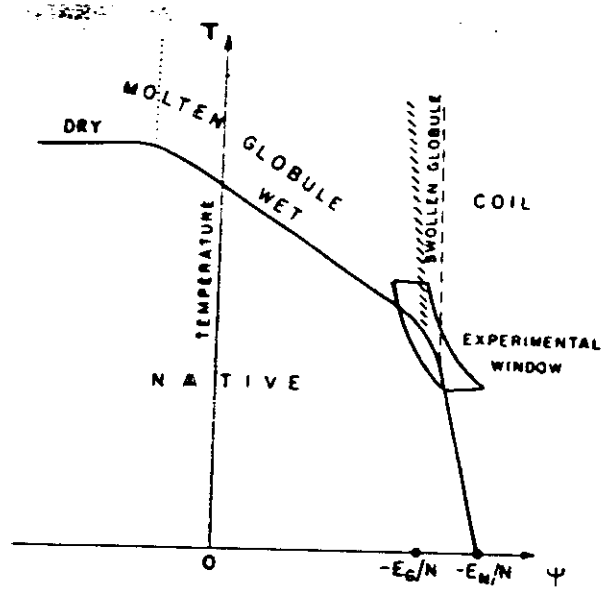



Fig. 4. Phase diagram of a protein molecule plotted in the coordinates of solvent quality $\Psi$ – temperature $T$. (———) Line of the first-order phase transitions from the native to any of the denatured states. (– –) Line of the $\theta$-points, the swollen globule → coil second-order phase transition. (//////) The crossover line between molten and swollen globules (without phase transition). (·····) The crossover line between the wet and dry molten globules (without phase transition). The insertion shows the experimental window of parameters corresponding to the available experimental conditions.

It is seen directly that molten globule in proteins must be wet: it follows from the fact that denaturation occurs in the interval 300-380 K and that denaturation temperature depends on solvent strongly.

All the above was discussed in general terms. However the solvent quality $\Psi$ depends on temperature (temperature dependent hydrophobic effect) and on addition of denaturant. For example it has the following empirical dependence on GuHCl concentration:

$$\Psi(C) = \Psi(0) - 0.111C + 0.057C^2$$

It makes it possible to present the phase diagram in "experimental" coordinates C-T:

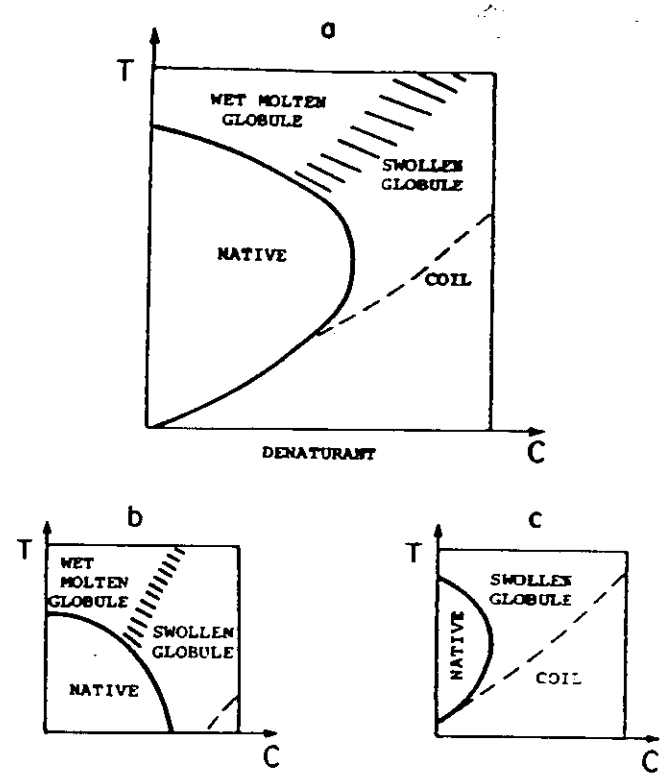



Fig. 6. The protein phase diagrams plotted in the experimental coordinates of concentration of denaturant $C$ – temperature $T$. The generalized diagram (a) corresponds to the experimental window in Fig. 4; (b) corresponds to a less thermostable, and (c) corresponds to a less hydrophobic protein.

We see that this theory gives some answers about the nature of denatured state and possible driving mechanisms of denaturation. It formulates the minimal, necessary structural changes which must occur in a protein during denaturation.

However it does not address the very important problem: what happens with the backbone upon denaturation? Indeed, the treatment of the backbone conformations in this theory is rather vague: it is treated as a homopolymer, and Lifshitz theory of coil-globule transitions was applied here.

However, a key question emerges: does molten globule have unique backbone conformation or it is rather a mixture of different conformations as collapsed homopolymer?

In order to answer this question a more sophisticated theory which considers formation of unique structure is required as well as more sophisticated experiment.

# Kinetics of protein folding: theoretical approach and computer simulations.

The ultimate goal of theoretical investigation of protein folding is to resolve the Levinthal paradox, i.e. to find a model in which a protein chain folds to the stable state at a reasonable time. These investigations have also a practical goal to work out a reliable algorithm which will provide native 3D structure from a protein sequence - algorithm which will fold a protein starting from disordered (random coil) conformation.

The first attempt to simulate protein folding was done in 1975 (M.Levitt, A.Warshel Nature **253**, p.694 (1975)) where "renaturation" of BPTI was done. First of all, a simplified model was introduced where each aminoacid is represented by two points: centroid of $C_\alpha$ atom and centroid of side-chain (see fig.)

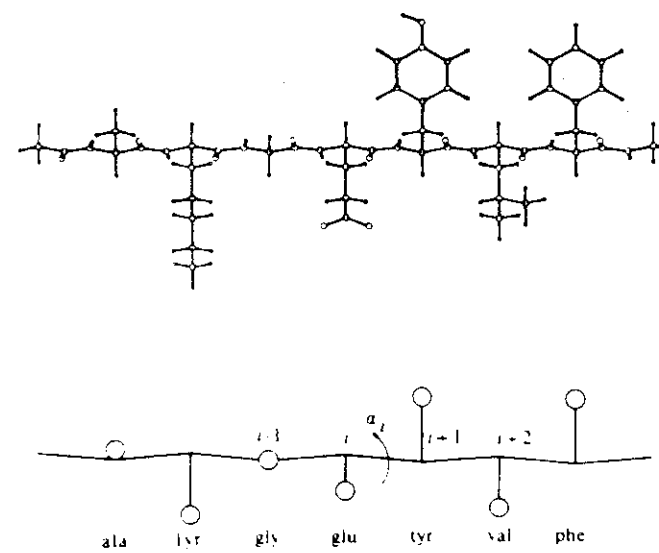


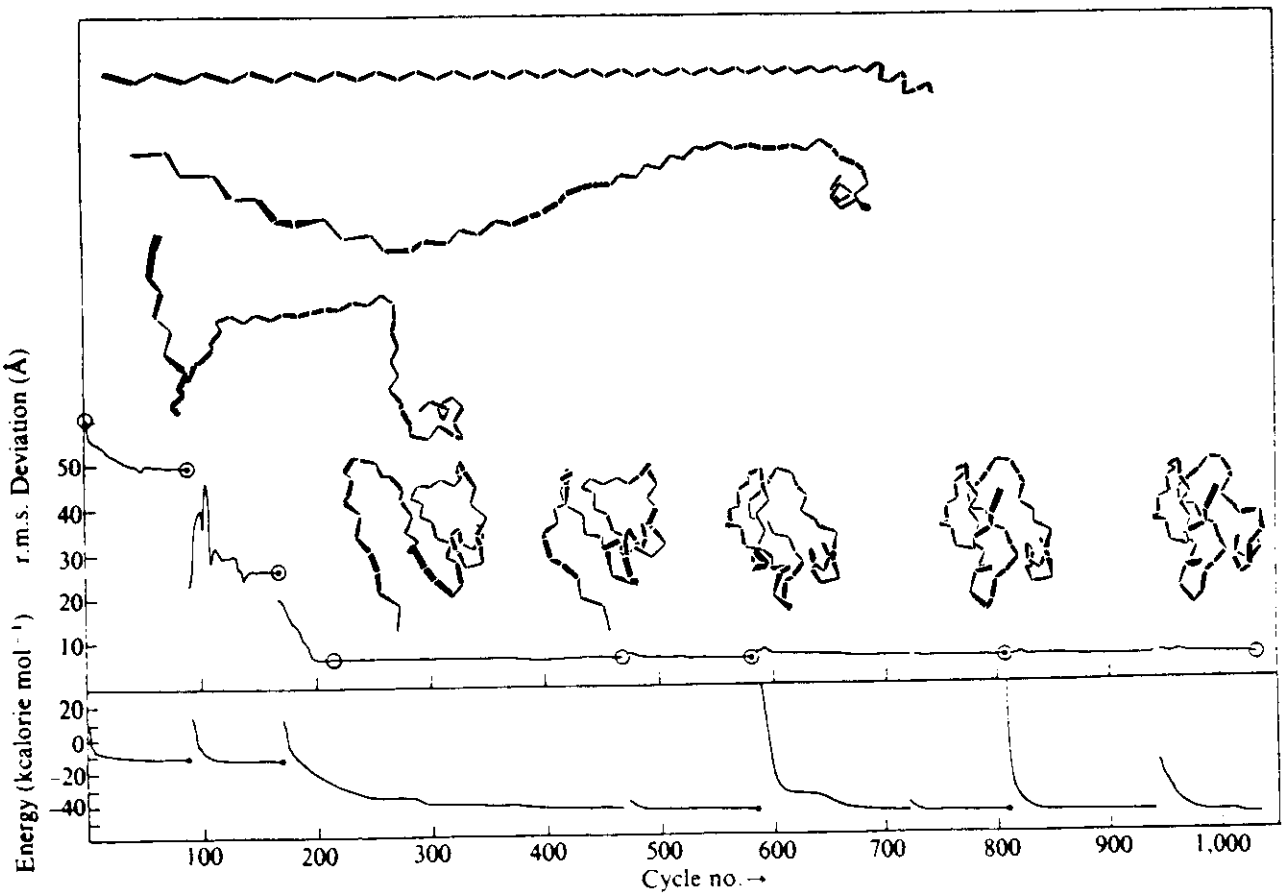

**Fig. 1** Relationship between the simplified model of protein structure introduced here and the real all-atom structure of proteins. The two reference points for each residue in the simplified model correspond to the centroid of the side chain and the $C^\alpha$. Each residue is only allowed one degree of freedom: the torsion angle $\alpha$ between the 4 successive $C^\alpha$s of residues $(i-1, i, i-1, i-2)$. All the side chains of a given type have the same simplified geometry. The bond lengths, bond angles, and torsion angles used to define the geometry of the simplified molecule were taken as the average values found in eight protein conformations, though they could just as well have been taken from amino-acid model compounds.

This representation is based on idea of averaging over displacement of side-chains on small time and space scales so that protein structure can be presented via coarse-grained low-resolution model.

Each residue is only allowed one degree of freedom: the torsion angle between 4 successive $C_\alpha$ atoms: (i-1,i,i+1,i+2).

The folding of such idealised protein can be simulated by solving the equations of molecular dynamics (in this work simple minimization of energy is performed in this stage). After reaching a minimum a thermal fluctuations are reinforced and the conformation is considered to be vibrating about the minimum so that each normal mode has energy kT. At some time the motion is stopped and each normal mode is displaced randomly by a value $(RkT/\lambda)^{1/2}$ where $\lambda$ is the eigenvalue of energy second derivative matrix corresponding to some particular mode.

The process of folding by this algorithm is shown in fig.

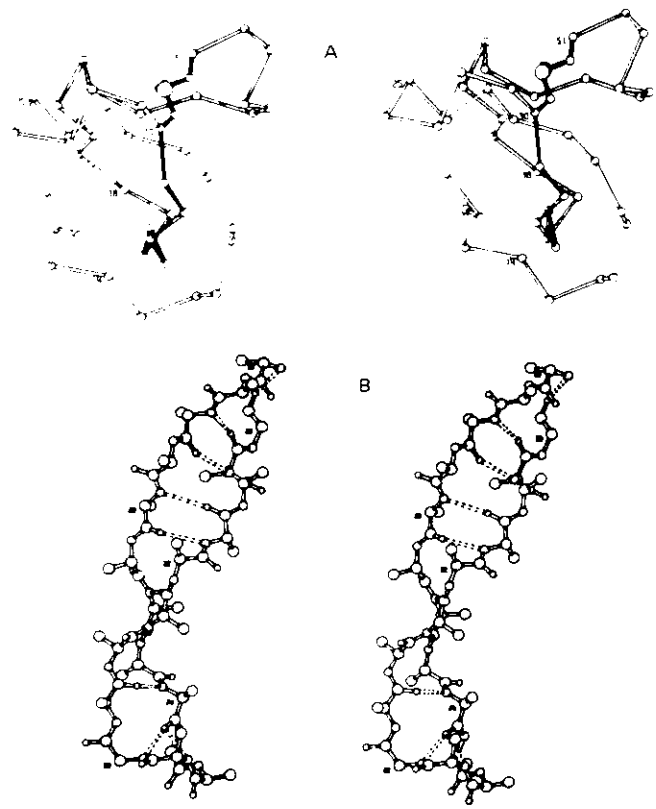


As a result of this simplified treatment folding proceeded quite rapidly and lead to the structure which had r.m.s. deviation of 3.36 A from the native one.

On the basis of this investigation authors claimed that the plausible model of folding is maybe that at the initial stage only effective time-averaged forces between residues play role, folding a chain to a compact structure with most groups close to its final positions (say, within 5A) then subsequent "fine-tuning" brings the structure to its final form. We see however that this scenario does not differ very much from one proposed earlier (e.g. framework model).

This approach was revisited, however, in the work of A.Hagler and B.Honig (PNAS, **75** pp. 554-558 (1978)).

They criticized the previous approach to simulations of folding claiming that criteria of obtaining folded structure in the work, say, of Levitt and Warshel are too permissive (e.g. r.m.s. deviation of the simulated "native" structure from X-ray structure is 5-6 A.). More important is that major topological fetaures of the BPTI structure were not reproduced in the simulations, namely threading of the loop 30-51 by the chain and $180^0$ twist in the $\beta$-structure (see fig.)



Therefore these authors conclude that "folding" within 5A reproduces only very general features of a protein structure such as, say type and sometimes location of secondary structure elements but not the important topological features of the chain.

In the further discussion they introduced an oversimplified model of a "BPTI" which contains only glycine and alanine in the sequence. Glycine was chosen to account for places in the sequence which give probable bends.

The same minimization procedure as such of Levitt and Warshel gave similar results, i.e. $\beta$-sheet 16-36 and even the C-end helix, however this helix was left-handed rather than right-handed.

The relation of simulation results and experimental results of Creighton on BPTI folding was discussed there. It was mentioned that very subtle features of kinetic intermediates, e.g. formation of structures with incortrect S-S bonds may account for "threading" of a loop 30-51 by chain.

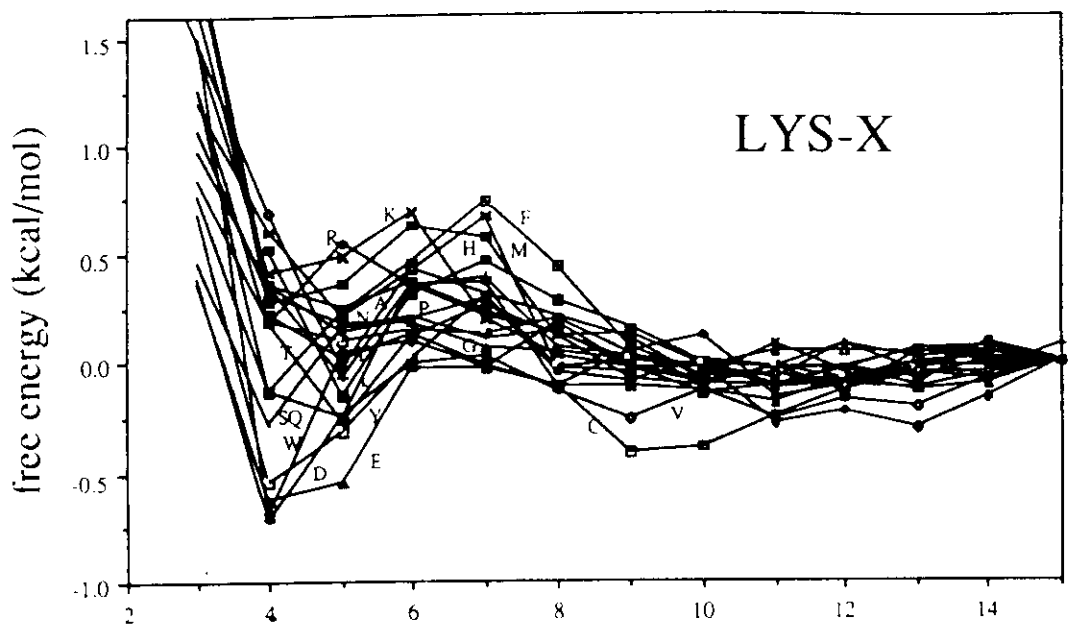
However these (and others at that time) simulations didi not answer an important question, whether a well-concieved, unbiased simplified model can contain enough structural and energetic information to truly predict protein structure.

This question was addressed in the recent work of Wilson and Doniach (Proteins, **6**, p. 194 (1989)). They investigated folding of Crambin - small globular protein consisting of 46 amino acids. Important issues - choice of parameters and Monte-Carlo technique combined with simulated annealing scheme were discussed in some detail there.

The choice of parameters- interaction energies between different aminoacids - was done following the method first proposed by Myasawa and Jernigan (Macromolecules, **18**, p 534 (1985). These potentials were derived directly from the distribution of residues observed in known protein structures. $C_\alpha - C_\alpha$ distances of all residue pairs in 100 PDB proteins were analysed and used to generate histograms of the number of occurences versus distance for

each amino acid pair. The observed distributions of reidue-residue distances were converted into free energies assuming Boltsmann distribution.

Example of resulting potential is shown in fig.




The folding kinetics was simulated by MC algorthm of Methropolis. The procedure can be described as following:

1) Store the energy of the old configuration $E_{old}$.

2) Choose randomly a monomer to be displaced.



25

3) Choose randomly one of a possible (permitted by covalent bonds) displacement of this monomer.

4) Calculate the energy of a new configuration $E_{new}$.

5) If $E_{new} < E_{old}$ accept this motion.

6) If $E_{new} > E_{old}$ accept this motion with probability $\exp(E_{new} - E_{old})/T$.

In some versions of MC algorithm also a large-scale trial motions (such as e.g. simultaneous motion of a large piece of a chain) are also permitted.

It was shown long ago that such algorithm provides Boltsmann distribution on large time and, hence may reproduce thermodynamic features of the transition.

Annealing scheme was proposed by Kirkpatrick and Gelatt (Science, **220**, pp.671 (1983)) and its idea is to decrease temperature slowly as simulation progresses.

As a result of simulations structures with RMS deviation 4.7 A from the native structure were the best obtained. However, the contact matrices were reproduced reasonably well (see fig.)

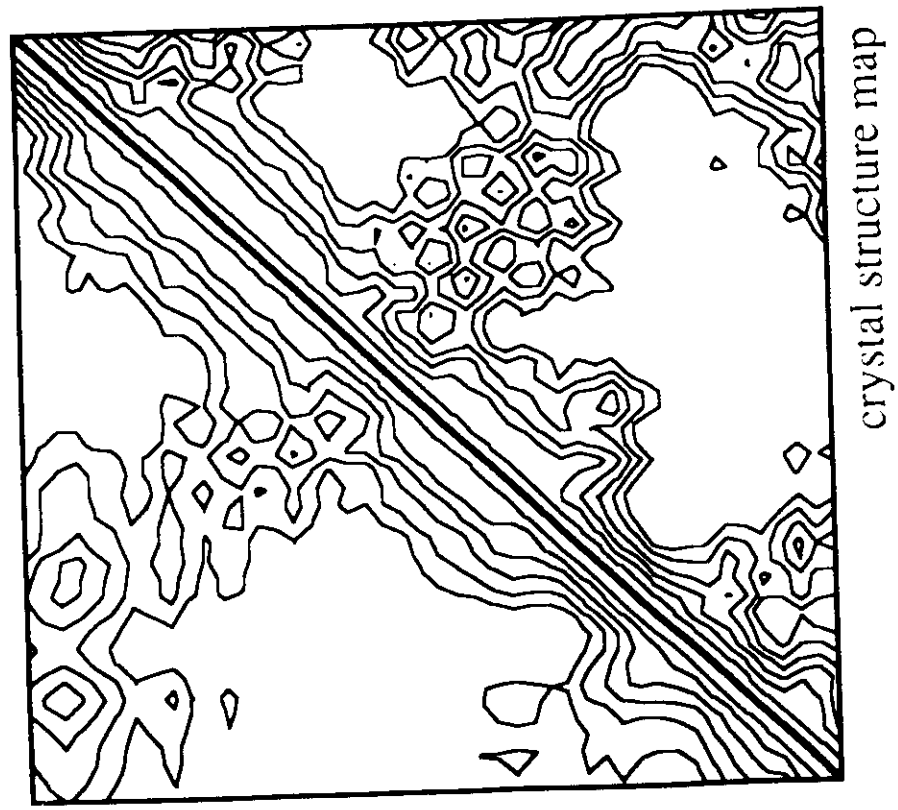


**Table II. Characteristics of Fixed Temperature and Annealed Simulations***

| Temperature | Energy (kT) | Radius of gyration (Å) | Total contacts | rms native (Å) | DME native (Å) | DME internal (Å) | DME average (Å) |
|---|---|---|---|---|---|---|---|
| 0.5 | 2531.6 | 9.68 | 781.2 | 8.15 | 5.34 | 5.03 | 4.06 |
| 1.0 | 2507.2 | 9.90 | 755.6 | 7.95 | 5.54 | 5.03 | 4.30 |
| 2.0 | 2528.5 | 9.65 | 784.6 | 8.01 | 5.36 | 4.94 | 4.14 |
| 5.0 | 2529.8 | 9.77 | 754.9 | 7.72 | 5.16 | 4.95 | 3.86 |
| 10.0 | 2509.7 | 9.99 | 727.0 | 8.36 | 5.51 | 5.35 | 4.08 |
| 20.0 | 2690.8 | 12.24 | 590.5 | 8.34 | 7.01 | 7.04 | 5.29 |
| 100.0 | 3357.3 | 17.47 | 469.7 | 12.42 | 14.28 | 7.91 | 13.46 |
| Simulated annealing | 2349.7 | 9.15 | 828.1 | 7.54 | 4.76 | 4.21 | 3.76 |
| Crystal structure | 3192.9 | 9.70 | 876.0 | — | — | — | — |

Monte Carlo dynamics was run at the indicated fixed temperatures. For the simulated annealing run, the temperature was initially set to 100, and gradually lowered. rms native refers to the average rms deviation of each model structure from the PDB crystal structure. DME indicates the average distance matrix errors between each model structure and the crystal structure (native) and between each model structure and every other structure (internal). The average DME is the distance matrix error calculated by comparing the contact map of the crystal structure to the average contact map generated from all the model structures of a given simulation.

The produced structures were sequence-dependent: randomly "shiffled" sequences did not produce reasonable agreement with X-ray structures.

The conclusions drawn by the authors of this work are the following:

1) Starting from random conformations secondary structure forms in correct places.

2) The formation of secondary structure is influenced by long-range interactions.

3) With the secondary structures assigned $\alpha$-helices and $\beta$-strands associate as they do in the true structure.

The last conclusion seems to be the most important and meaningful. The thing is that the intrinsic propensity to the formation of different local secondary structure elements is the underlying feature of this model and all other models of simulations of protein folding.

A very extensive simulations of protein folding by MC algorithm were done in a series of papers of Scolnick and co-workers. (see, e.g. PNAS **85**, p.5057 (1988), J.Mol.Biol. **212** p.787, *ibid*, p.819) and references therein.

The basic model studied in these works was a chain composed of monomers of two types (hydrophobic, hydrophilic) positioned in sites of a diamond lattice. The conformational state is given by a sequence of n-3 rotational states for the bonds, each of which may be in either the planar (trans) or one of the two out of plane gauche states. Intrincic energies $\epsilon_g$ were ascribed to each A.a. as well as energies of hydrophobic interactions. $\epsilon_h$.

The primary structure for a $\beta$-protein was defined as following. $B_i(k)$ is a i-th stretch in the primary sequence that consists of k residues; all these residues have preference for $\beta$-conformation. b denotes position of turn and L denotes loops. Then typical sequence can be presented in the form $B_1(i_1)b_1B_2(i_2)b_2B_3(i_3)b_4...B_k(i_k)L_1$.

This model was investigated (as well as models where $\alpha$-helical bundles were formed by special primary structures).

The equilibrium transition is described as an all-or-none transition between coil and native globule (see figs).

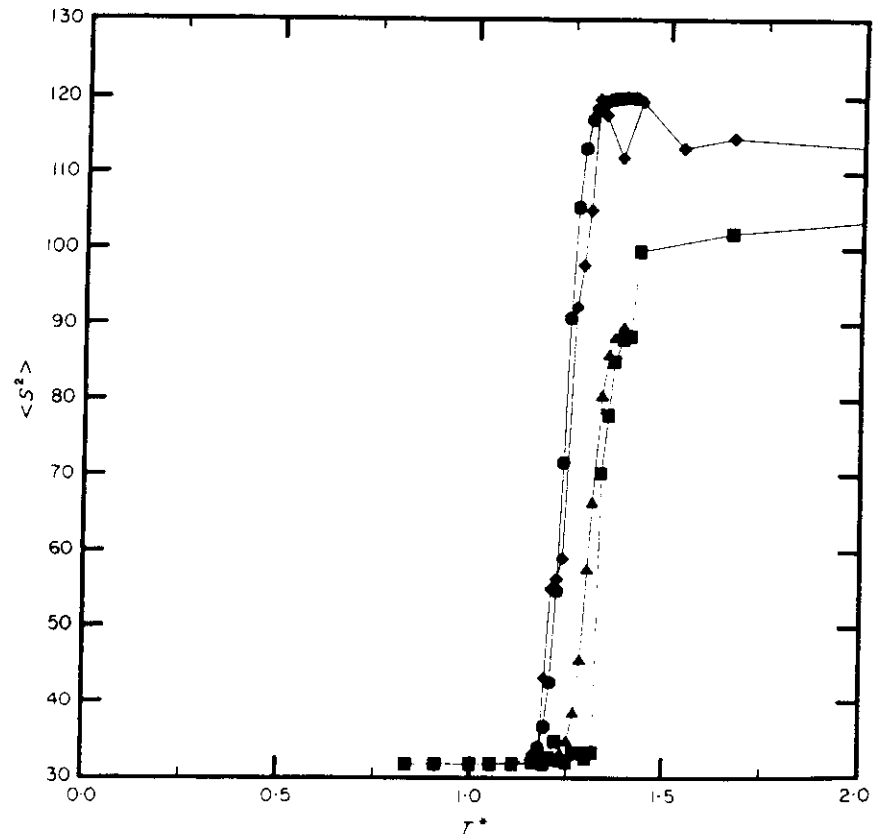



**Figure 4.** Plot of mean square radius of gyration $\langle S^2 \rangle$ *versus* reduced temperature $T^*$ for model A in the curves denoted by the (a) filled diamonds and (b) filled squares, respectively. (b) In the curves denoted by the filled circles (triangles) $< S^2 >$ *versus* $T^*$ is calculated *via* eqns (9) and (10) for model A.

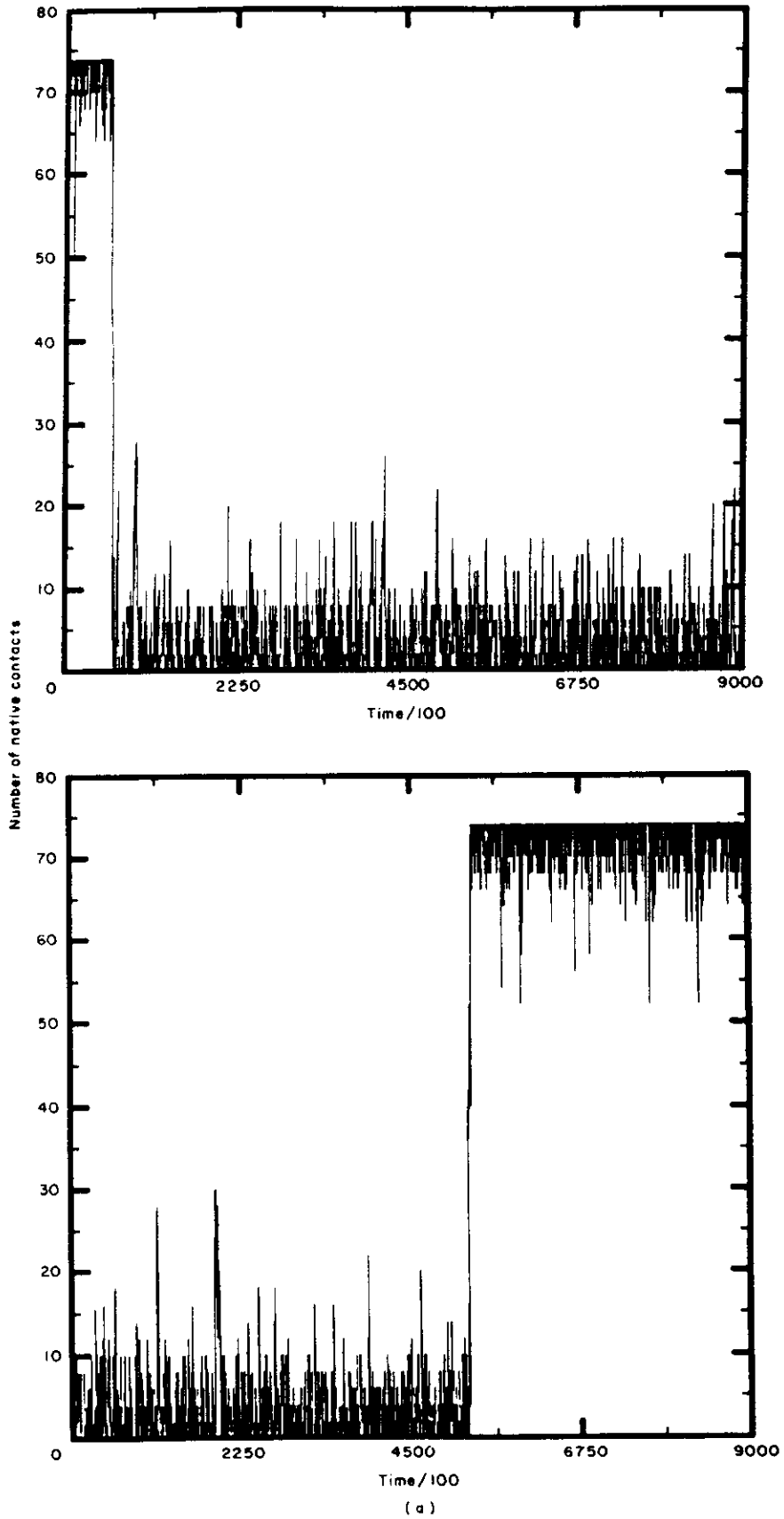

Number of native contacts
80
70
60
50
40
30
20
10
0
2250
4500
6750
9000
Time/100
80
70
60
50
40
30
20
10
0
2250
4500
6750
9000
Time/100
(a)

Fig. 5.

The folding trajectory was also analyzed:

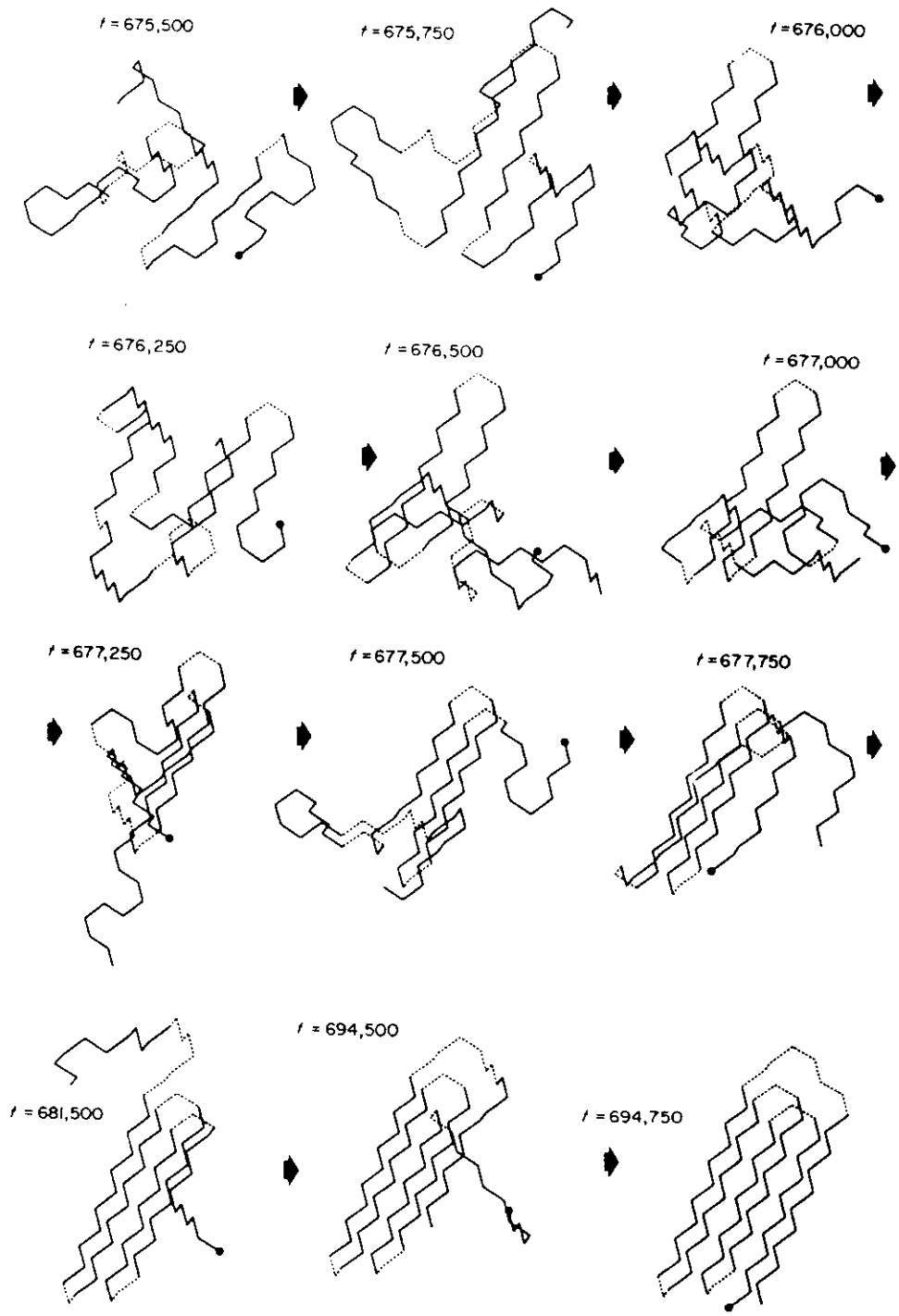



**Figure 10.** Representative folding trajectory extracted from run 4 of model B. The circle denotes the N terminus.

It was also claimed that kinetic intermediates were observed in the course of folding; this intermediate has $N_C = 18 - 26$ native contacts and corresponds to a broad free energy valley at which 4 of 6 $\beta$-strands are assembled and fluctuations are strong.

According to these simulations the transition state of folding involves 29 of 37 native contacts and is close to the native conformation.

Further extension of this study was proposed in the recent paper (Science, 23 November 1990, p.1121). The new lattice was applied in which side-chains are taken into account explicitely. This is the cubic lattice but adjacent $C_\alpha$ are connected by a vector of the type $(\pm 2, \pm 1, 0)$ - by some generalization of a "knight's" walk in chess.

Folding of plastocyanin was investigated - a protein containing 99 residues with topology of antiparallel 6-$\beta$-strand Greek key.

Again intrinsic propensities for the native $\beta$-structure were assumed: total energy of native state was -239kT while contribution from secondary structure to this value was -179.5kT. However authors of this work claim that secondary structure itself does not determine native structure.

As a result of simulations folding as "all-or-none" transition was observed to the structure which was topologically equivalent to the best lattice fit to the structure of crambin but had rms deviation of 6.1 A.

The time course of folding in this model is shown in fig.

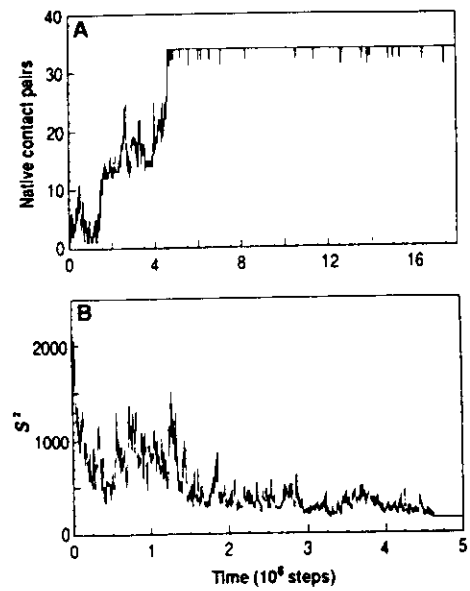
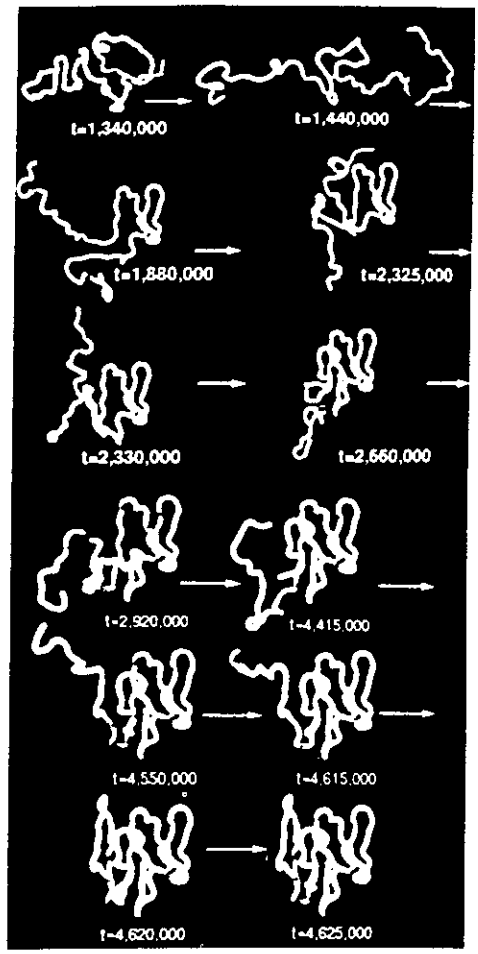



**Fig. 3.** For a representative folding trajectory, (**A**) a plot of the number of contacts between pairs of side chains versus time and (**B**) the instantaneous value of the square of the radius of gyration $S^2$ versus time.

The sequence of events in the folding process is shown in the next fig.




However the above approach to the protein folding problem raises some serious questions:

1) It is not clear whether the initial assumptions such as, e.g. intrinsic propensities of the native secondary structure does not create bias towards already known 3D structure.

2) It is also important to note that these simulations demonstrate "all-or-

none" coil-globule transitions. Experimental data give evidence in favor of "all-or-none" transitions from native to molten globule ststes (at least in some cases).

3) MG state was not observed in these experiments (probably more detailed account for side-chains is necessary to clarify this point)

One should await very rapid development of events in the near future.

## Analytical investigations of the folding kinetics

1. Diffusion-Collision-Model.

This model was proposed by Karplus and Weaver (Nature, **260**, p.404 (1976)). The idea of this approach is to consider regions of unfolded chain to fluctuate between helical and condensed random coil states. These nascent helices move diffusively until they meet and coalesce to give rise to mutual stabilization (see fig.)



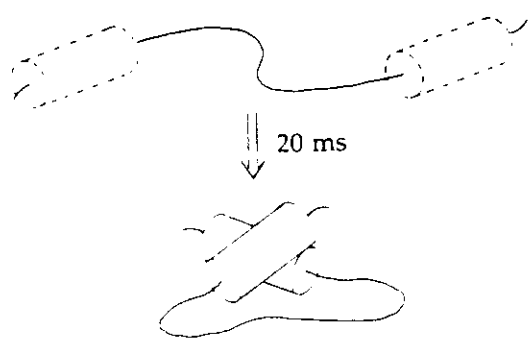



**Fig. 1.** The elementary step in the diffusion-collision model. Two nascent helices move diffusively until a collision results in their coalescence and mutual stabilization. The nascent helices in the upper figure are drawn with dashed lines to indicate a rapid equilibrium between helical and condensed random coil states.

The diffusion equation is written for two components: $a$ corresponding to the helical state which can coalesce and $r$, the non-helical state which do not coalesce:

$$\frac{\partial}{\partial t}\begin{bmatrix} \rho_a(r,t) \\ \rho_r(r,t) \end{bmatrix} = D\nabla^2 \begin{bmatrix} \rho_a(r,t) \\ \rho_r(r,t) \end{bmatrix} + \begin{bmatrix} -K_{ar} & K_{ra} \\ K_{ar} & -K_{ra} \end{bmatrix} \begin{bmatrix} \rho_a(r,t) \\ \rho_r(r,t) \end{bmatrix}$$

The diffusion takes place in the space interval $R_{min} < R < R_{max}$ (see fig.)

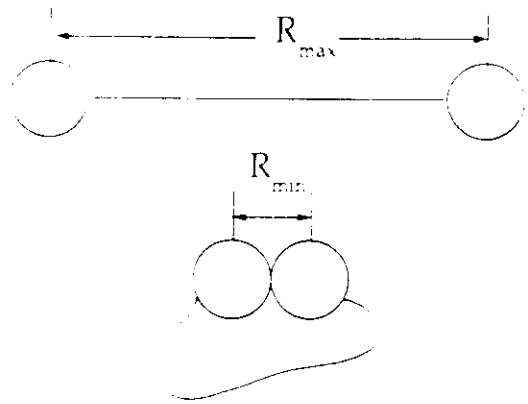



**Fig. 2.** The nascent helices of Fig. 1 are idealized as spheres and the connecting polypeptide as a string

As a result the characteristic time of helix coalescence can be found:

$$\tau_b = \frac{R_{max}^3}{3DR_{min}} + \frac{1-\beta}{\beta}\frac{V}{DA}(D\tau_r)^{1/2}$$

where

$$\beta = \frac{k_{ra}}{k_{ra}+k_{ar}} \qquad \tau_r = \frac{1}{k_{ra}+k_{ar}}$$

are the helix stability and characteristic time of fluctuations between helix and coil states.

The calculation gives for for $\tau_b$ the value 28 msec which is in reasonable correspondence with the experiment.

The attempt to construct a whole passway of folding of apo-myoglobin was also done within a model where native-like helices can form only native-like contacts. The general scheme arising from these evaluation is shown in fig.

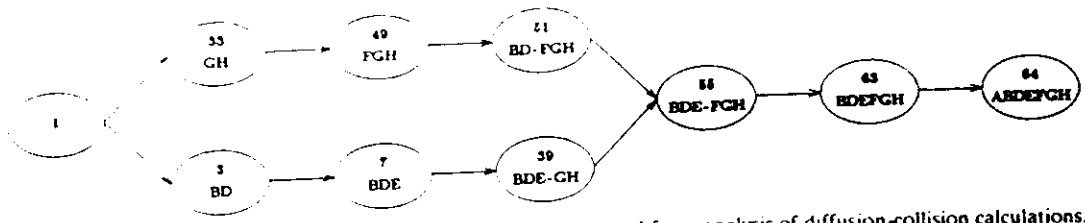



**Fig. 6.** The general pathway for the folding of apomyoglobin derived from analysis of diffusion-collision calculations. Only states rising above a probability of 0.02 are shown

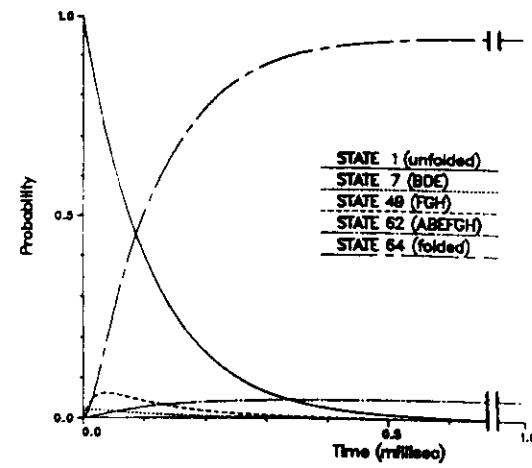



**Fig. 5.** The results of a diffusion-collision calculation for apomyoglobin using the choice of B values described in the text. Notation of the form "FGH" indicates that helices F, G, and H have coalesced into a single micro-domain as in the fourth state shown in Fig. 4, but no other helix-helix contacts are formed.

The obvious limitation of the diffusion-collision model is that only native localisation and native-like contacts of helices are considered there and thus this model a'priori avoids the main difficulty of protein folding problem. It is also known from experiment that some non-native conformations develop in the course of folding.

# Proteins with Selected Sequences Fold into Unique Native Conformation

E. I. Shakhnovich
*Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138*
(Received 1 December 1993)

We design sequences of 80-monomer model protein which provide very low energy in the target ("native") structure. Then the designed sequence is subjected to lattice Monte Carlo simulation of folding. In all runs model protein folded from random coil to the unique native conformation, effectively "solving" the multiple minima problem. These results suggest that thermodynamically oriented selection of sequences which makes the native conformation a pronounced deep minimum of energy solves the problem of kinetic accessibility of this conformation as well.

PACS numbers: 87.15.Da, 61.43.–j, 64.60.Cn, 64.60.Kw

The complexity of the protein folding problem is in the fact (often referred to as Levinthal paradox [1]) that unique, native conformation should be chosen in the folding process without scanning the astronomic number of possible conformations. The important question is whether this kinetic ability of natural proteins to fold is due to evolutionary selection of their sequences and, if yes, how can this feature be encoded in a protein sequence?

The straightforward approach (tried, e.g., in [2]) would be to take natural amino-acid sequence and simulate a (simplified) model of a protein expecting convergence to the native 3D conformation. However, the difficulty with this approach is that protein sequences could have been evolutionary designed to fold to their native structures with some "exact" set of potentials while simulations necessarily use approximate energetics [3] for which the native structure may be neither a global nor a pronounced local minimum. It is then hard to expect any folding algorithm to converge to a "native" state which may not be distinguished by energy from many other conformations.

This suggests the idea of using protein design to study folding of model one-domain proteins of realistic size. The goal is to design a sequence which has very low energy in a given (arbitrary) conformation. Folding simulation with the same potential function as was used at the design stage will then reveal whether this conformation can be reached in a reasonable time. Combination of design and folding "in one pair of hands" makes it possible to address the basic questions of protein folding and evolution separately from the problem of finding the correct potential functions for protein simulations.

In this study we model proteins as positioned on a cubic lattice. The Hamiltonian of a model protein is determined by the set coordinates of its monomers $\{r_i\}$ and (quenched) sequence of monomers $\{\sigma_i\}$ which denotes the identity of each monomer. Contact approximation is taken for the Hamiltonian,

$$E(\{\sigma_i\},\{r_i\}) = \frac{1}{2}\sum_{i,j}^{N} U(\sigma_i\sigma_j)\Delta(r_i - r_j), \qquad (1)$$

where $N$ is the total number of monomers and $\Delta$ defines the contact potential between them: $\Delta(r) = 1$ if monomers are lattice neighbors and 0, otherwise. We consider our model proteins positioned on a cubic lattice with unit bond length.

The set of potentials $U(\alpha,\beta)$ characterizes energies with which a monomer of type $\alpha$ interacts with a monomer of type $\beta$. First we tried two-letter sequences (hydrophobic-hydrophilic) like ones used in two-dimensional lattice models of proteins [4]. However, two-letter sequences appeared to be inappropriate for studying protein folding in three-dimensional models (see below). Therefore in what follows twenty-letter representation of protein sequences was used. In this case $U(\alpha,\beta)$ is a $20 \times 20$ matrix; as an example we used the one derived by Miyazawa and Jernigan [3] from the statistical distribution of contacts in native proteins.

In this work we studied folding of 80-monomer chains. Following the idea to combine folding and design we choose (arbitrarily) a target structure which is in our case a compact conformation of a chain on the cubic lattice. An example of the target structure is shown in Fig. 1.

After the target structure is picked, sequence design should be made to find a sequence which fits the target structure with low energy as determined according to Eq. (1) where coordinates $\{r_i\}$ correspond to target conformation. To this end the sequence-space Monte Carlo (MC) procedure of design was used [5,6]. The idea of sequence design is very simple: For the design purposes just view Eq. (1) as one where coordinates of the target structure $\{r_i\}$ are quenched but sequence variables $\{\sigma_i\}$ are annealed and Eq. (1) should be optimized with respect to them. This leads naturally to the idea of simulated annealing in sequence space; the procedure is straightforward and the details are published in [5,6].

The following argument based on the theory of heteropolymers allows us to estimate whether the native (target) structure corresponds to the global energy minimum for the designed sequence.

We divide the set of all conformations into two groups: the ones which have significant similarity with the tar-

0031-9007/94/72(24)/3907(4)$06.00



38

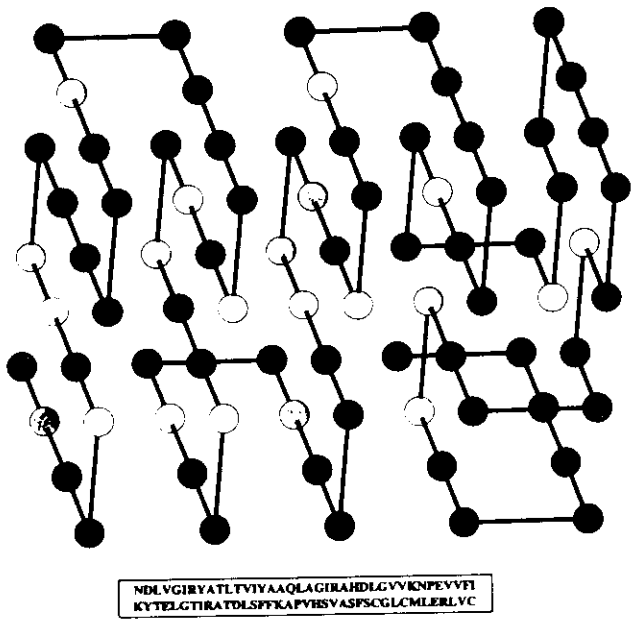



FIG. 1. An example of a compact conformation of an 80-monomer on a cubic lattice and the optimized sequence. Amino acids of different types are shown by different gray scale for illustrative purposes. This conformation as well as several other conformations with their sequences (not shown) were used as native structures in our studies. The shown sequence was designed to have low energy in the shown conformation.

get structure and the remaining vast majority of conformations which have marginal or no similarity with the target structure, just like two randomly superimposed conformations.

For conformations which are not similar to the target structure the designed sequence is effectively random and therefore the statistics of their energies are equivalent to those of a random heteropolymer. (A similar argument was first given by Bryngelson and Wolynes in their discussion of the "minimal frustration" model of protein folding [7].)

The important feature of random heteropolymers is that there exists a threshold energy $E_c$ such that the probability to find conformations with energy well below $E_c$ is extremely small [8–11]. Therefore the successful design should create sequences whose energy $E_N$ in the native conformation is well below $E_c$: In this case random conformations (structurally nonsimilar to the native state) will not have energies close to that of the native conformation and therefore will not serve as deep energetic traps for folding.

$E_N$ is known directly from Eq. (1) for the designed sequence. To estimate $E_c$ we use the replica mean-field theory of heteropolymers [8–11]. $E_c = E_0 - JN(2\ln\gamma)^{1/2}$, where $\gamma$ is the number of conformations per monomer. The important parameters $E_0$ and $J$ are the mean and the standard deviation of interaction energies. Since we are using parameters which are obtained from protein statistics, we have only relative energies and do not know the absolute energy scale for those parameters. So we use the energy unit at which $J = 1$. This requires multiplication of all parameters by a scaling factor. To determine this scaling factor we generated a set of 1000 random sequences (all having the same amino-acid composition) and fitted them into the target structure adjusting the scaling factor so that $J^2 = (\langle E^2\rangle - \langle E\rangle^2)/N = 1$. $\langle E\rangle = E_0$ and $\langle\,\rangle$ denotes averaging over the set of random sequences. We took $\gamma = 3.5$ which takes into account excluded volume and certain degree of compactness of unfolded conformations for which variance of interactions $J$ is estimated. The estimates were done for two sets of parameters: "two-letter" code with monomers of two types ("$H$" and "$P$") so that $U(H,H) = -1$; $U(H,P) = U(P,P) = 0$ and the twenty-letter set of Miyazawa and Jernigan. The amino-acid composition was set to be 50% $H$ and 50% $P$ monomers for two-letter chains and corresponding to averaged composition in proteins [12] for the twenty-letter set. The results for 80-monomer chains are given below.

(1) Two-letter heteropolymers: $E_c = -72.3$, $E_N = -61$. The model is not specific enough to have unique structure: All possible energy levels are multiple degenerate. No folding to unique structure is possible in that case.

(2) Twenty-letter parameters: $E_c = -123.6$, $E_N = -156.5$. The estimated gap is pronounced, $\approx 23T$ at the temperature at which most of the simulations have been done. In all that follows twenty types of monomers are used and the results are reported for that model.

Now we simulate folding of the designed sequence using the simple lattice Monte Carlo folding algorithm [13–17] and energy function given by Eq. (1). The move set which we used allows corner flips and crankshaft motions but excludes multiple occupancies of lattice sites. It was argued in [17] that such a move set makes cubic lattice simulation ergodic.

Simulations started from random coil conformations. There was made a total of 1000 runs starting from different randomly chosen coil conformations. The main result of this work is that in each run chain folded into the unique target conformation with mean first passage folding time close to $10^6$ MC steps.

A typical folding trajectory recorded at temperature $T = 1$ is shown in Fig. 2. Analysis of energy changes with Monte Carlo time shows that structures with energy lower than the energy of the native state have never been encountered. In order to estimate whether this conclusion is sensitive to the move set we repeated simulation with enhanced move set which allowed also for 3,4,5-monomer crankshaft moves. The results are similar: Again the target state was the lowest energy one, and the trajectory was similar to the one shown in Fig. 2.

Pronounced fluctuations around the minimum energy structure make this model close to the Molten Globule (MG) [18,19]. This is due to the fact that there are

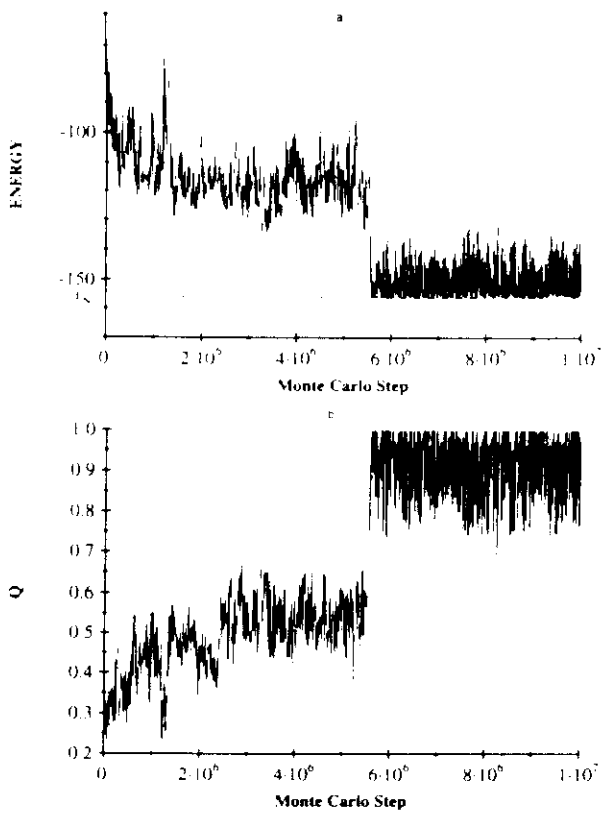



FIG. 2. A typical MC trajectory of folding simulations for 80-monomer chain. (a) The dependence of energy on MC step. The energy of the native conformation is shown as $E_N$. (b) The dependence of normalized number of native contacts on MC step. The maximal number of contacts $N_{\text{total}} = 105$ for a compact 80-monomer. For each conformation we normalize the number of native contacts, $Q$, by $N_{\text{total}}$ so that $Q = 1$ corresponds to the native conformation.

no side chains in the model, which tight packing distinguishes the MG from the native state and makes the native conformation more rigid [19].

The ability to fold appears to be a virtue of designed sequences and is temperature dependent, as expected. Steepness of the curves in Fig. 3 is consistent with the assertion [5] that designed sequences have a first-order folding transition. Applied to proteins this suggests that the coil-MG transition may be also first order, like the native-MG one. The first-order character of the native-MG transition, however, may be due to a different reason, side-chain freezing [19], which is not considered in the present model (see [20] for the discussion of first-order transitions in macromolecules).

The temperature dependence of entropy can be obtained from temperature dependence of energy $E(T)$ using the thermodynamic relation

$$s(T) = s(\infty) + \frac{1}{N}\left(\frac{E(T)}{T} - \int_T^\infty \frac{E(t)}{t^2}dt\right). \qquad (2)$$

Here $s(\infty)$ is a high-temperature (athermal) limit of entropy. The value $s(\infty) = \ln(4.68) + \frac{1}{6}\ln(79)/79$ is known since at high $T$ it coincides with that for an athermal polymer on a cubic lattice [21]. Our simulations were

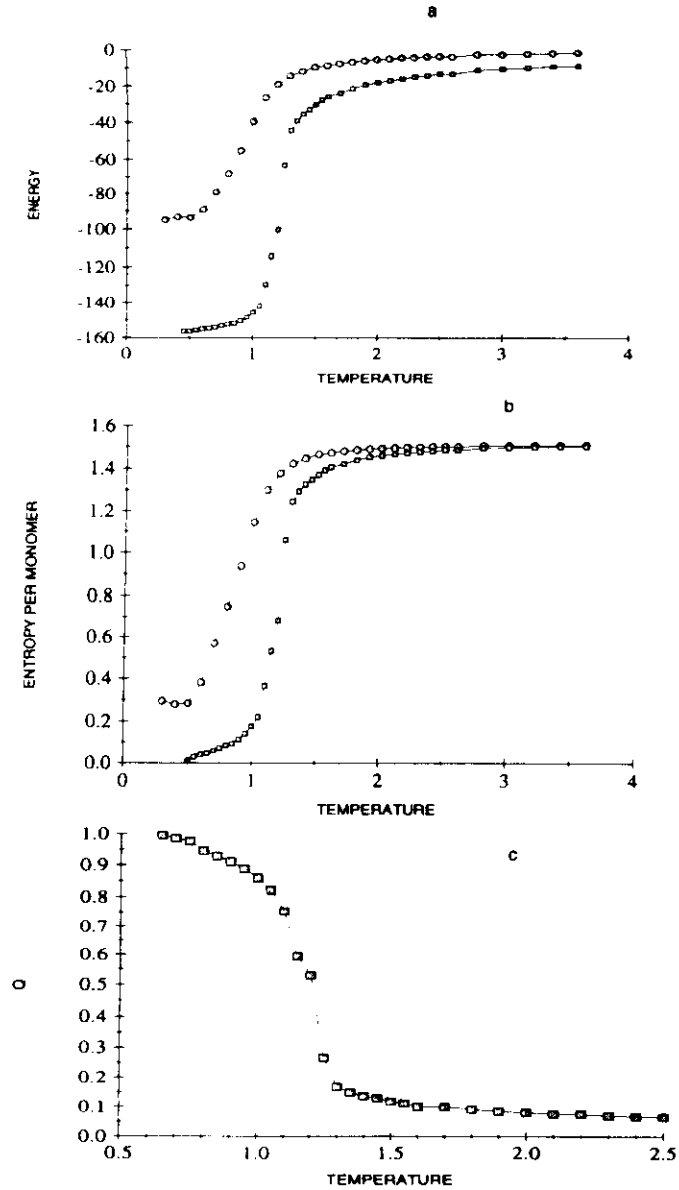



FIG. 3. Temperature dependence of energy $E$ (a), configurational entropy per monomer (b), and structural similarity with the native state (c) for the designed sequence (squares) and for a random sequence with the same amino-acid composition as the designed one (circles). At each temperature $10^8$ MC steps were made, and average energy $E$ and structural similarity with the target conformation $Q$ were determined as an average over the whole run at a given temperature. The calculation of the entropy curve is explained in the text.

performed in the temperature range $0.5 < T < 10.0$. We took $s(T = 10) = s(\infty)$. Only part of the temperature dependence corresponding to the temperature range $0.5 < T < 3.6$ is shown to provide a reasonable scale to show the transition. The truncated part at $T > 3.6$ is a trivial base line. In the low-temperature limit $s(T = 0.5) = 0.007$. The smallness of this number is consistent with the main result of this work—that designed sequences repetitively return to the target (native) conformation.

The same procedure was used then to calculate confor-

mational entropy of the random sequence. The number of conformations is determined as usually $M = \exp(Ns)$. In this case the same rate of annealing leads to freezing without development of unique structure: Different runs end up in different, unrelated conformations. The number of such frozen low-temperature conformations is estimated from the calculation of entropy (Fig. 3) to be $\sim 10^9$. Note also that even in the denatured state energy of designed sequence is noticeably lower than that of the random sequence.

Low-temperature freezing for a random sequence is a kinetic phenomenon: It was shown in [22] that in this case the global minimum cannot be reached by *any* algorithm in a reasonable [less than "Levinthal" $\exp(\alpha N)$] time. This does not contradict the assertion [9] that random sequences can have a thermodynamically stable unique structure in a certain temperature range. The reason is that the unique structure of random sequences becomes thermodynamically stable only at temperatures lower than $T_c$, the glass transition temperature [7–11,22,23]. However, as was shown in [22] (see also the excellent discussion in [23]), at $T < T_c$ the kinetics become extremely slow because the ruggedness of energy landscape of random sequences turns out to be crucial at temperatures lower than $T_c$. Sequences with large gaps have native structures which are stable at $T > T_c$, resolving therefore the contradiction between the requirement of thermodynamic stability and kinetic accessibility which is characteristic of random sequences.

Analysis of the curve $Q(T)$ in Fig. 3(c) suggests that the native state is sufficiently stable at temperatures at which simulations were done. For example, at $T = 0.8$, $Q \approx 0.95$ which means that 95% of native contacts persist throughout the simulations. Conformations which have 95% of native contacts differ from the native one (shown, e.g., in Fig. 1) by "tails" of 3–4 monomers long stretching out of the native structure. The alternative interpretation of this result would be that the chain spends 95% of the time in the native state and 5% of the time in unfolded conformation. The analysis of simulation data at $T = 0.8$ suggests that the chains spend practically all the time in or near native conformation, so that short-tail fluctuations account for the fact that $Q < 1$. This can be also illustrated from the estimate of entropy at $T = 0.8$, $S = 0.05$ per monomer, which suggests that fluctuations cover $\sim 100$ conformations, each only slightly (by 3–4 monomers) different from the native state. This is consistent with the "short-tail stretching" picture.

The same experiments were repeated with several other sequences and several other randomly chosen target structures for proteins of different lengths (36–100 monomers). One target conformation even had a quasiknot (Abkevich, Grosberg, and Shakhnovich, unpublished results). In all cases the results of simulations are qualitatively the same and are quantitatively close to the ones presented in this work.

Our design procedure generated sequences for which the target structure is likely to be the global (or at least accessible stable local) energy minimum separated by a pronounced energy gap from the set of non-native conformations. It is remarkable to note that such thermodynamically oriented design solved at the same time the *kinetic* problem making the native structure also kinetically accessible. This may represent a simple and universal principle of evolutionary selection of one-domain proteins with stable and kinetically accessible native conformation.

I am grateful to Victor Abkevich, Alexander Gutin, Martin Karplus, Peter Leopold, Oleg Ptitsyn, and Andrej Sali for interesting discussions. Graphic program ASGL by Andrej Sali was used to generate some of the plots. This work was supported by the Packard Foundation.

[1] C. Levinthal, J. Chem. Phys. **65**, 44 (1968).
[2] C. Wilson and S. Doniach, Proteins: Struct. Funct. Genetics **6**, 193 (1989).
[3] S. Myazawa and R. Jernigan, Macromolecules **18**, 534 (1985).
[4] K.F. Lau and K.A. Dill, Macromolecules **22**, 3986–3997 (1990).
[5] E.I. Shakhnovich and A.M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).
[6] E.I. Shakhnovich and A.M. Gutin, Protein Eng. **6**, 793 (1993).
[7] J.D. Bryngelson and P.G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[8] E.I. Shakhnovich and A.M. Gutin, Biophys. Chem. **34**, 187 (1989).
[9] E.I. Shakhnovich and A.M. Gutin, Nature (London), **346**, 773 (1990).
[10] C. Sfatos, A. Gutin, and E. Shakhnovich, Phys. Rev. E **48**, 465 (1993).
[11] A. Gutin and E. Shakhnovich, J. Chem. Phys. **98**, 8174 (1993).
[12] T. Creighton, *Proteins Structure and Molecular Properties* (Freeman, San Francisco, 1992).
[13] A. Sali, E.I. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[14] E.I. Shakhnovich, G.M. Farztdinov, A.M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[15] R. Miller, C. Danko, M.J. Fasolka, A.C. Balazs, H.S. Chan, and K.A. Dill, J. Chem. Phys. **96**, 768 (1992).
[16] C. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369–6372 (1993).
[17] H.J. Hilhorst and J.M. Deutch, J. Chem. Phys. **63**, 5153 (1975).
[18] O.B. Ptitsyn, in *Protein Folding* (Freeman, New York, 1992), Chap. 6, pp. 243–300.
[19] E.I. Shakhnovich and A.V. Finkelstein, Biopolymers **28**, 1667 (1989).
[20] M. Karplus and E. Shakhnovich, in *Protein Folding* (Ref. [18]), Chap. 4, p. 127.
[21] P.G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca, NY, 1970).
[22] J.D. Bryngelson and P.G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[23] H. Fraunefelder and P.G. Wolynes, Phys. Today **47**, 58 (1994).

Reprinted from Biochemistry, 1994, 33.


# Specific Nucleus as the Transition State for Protein Folding: Evidence from the Lattice Model†

V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich*

*Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge Massachusetts 02138*

*Received February 1, 1994; Revised Manuscript Received June 10, 1994*•

ABSTRACT: We have studied the folding mechanism of lattice model 36-mer proteins. Using a simulated annealing procedure in sequence space, we have designed sequences to have sufficiently low energy in a given target conformation, which plays the role of the native structure in our study. The sequence design algorithm generated sequences for which the native structures is a pronounced global energy minimum. Then, designed sequences were subjected to lattice Monte Carlo simulations of folding. In each run, starting from a random coil conformation, the chain reached its native structure, which is indicative that the model proteins solve the Levinthal paradox. The folding mechanism involved nucleation growth. Formation of a specific nucleus, which is a particular pattern of contacts, is shown to be a necessary and sufficient condition for subsequent rapid folding to the native state. The nucleus represents a transition state of folding to the molten globule conformation. The search for the nucleus is a rate-limiting step of folding and corresponds to overcoming the major free energy barrier. We also observed a folding pathway that is the approach to the native state after nucleus formation; this stage takes about 1% of the simulation time. The nucleus is a spatially localized substructure of the native state having 8 out of 40 native contacts. However, monomers belonging to the nucleus are scattered along the sequence, so that several nucleus contacts are long-range while other are short-range. A folding nucleus was also found in a longer chain 80-mer, where it also constituted 20% of the native structure. The possible mechanism of folding of designed proteins, as well as the experimental implications of this study is discussed.

Theoretical studies of the thermodynamics and dynamics of protein folding have been reviewed recently in Karplus and Shakhnovich (1992). The authors pointed out that different approaches should be taken to study different parts of configurational space. The neighborhood of the native state and the dynamics of thermal fluctuations around this state can be studied in detail using an all-atom representation of a protein and applying molecular dynamics to simulate the system. The simplistic point of view would be to extend these calculations further to explore more of the configurational space and to also address the folding problem. However, this is impossible due to the obvious time limitations of such calculations. This implies that simplified models should be used to study folding. These models should be adequate to the problem, but free of details that are relevant on time and length scales much smaller than the ones at which interesting folding events occur. The adequacy of a model for the folding problem requires that model proteins possess a unique structure that is thermodynamically stable at physiological temperatures. The model should have the Levinthal paradox, i.e., an astronomically large number of conformations that cannot be scanned exhaustively in a folding simulation.

The idea of "preaveraging" irrelevant fast degrees of freedom leads to low-resolution models such as "beads on a string" (Lifshitz et al., 1978) or closely related lattice models (Ueda et al., 1978; Shakhnovich & Gutin, 1990a; Covell & Jernigan, 1990; Lau & Dill, 1989; Skolnick & Kolinski, 1990a,b). In such models, a group of atoms of a protein is represented by one effective monomer; one could visualize this as a $C_\alpha$ representation of protein folds. These models capture important aspects of the protein folding problem: an astronomically large number of conformations, the polymeric structure of the chain, and the chain heterogeneity [monomers (although represented by structureless "beads") may be of different types manifested by interresidue interactions of different strengths and signs]. The identity of a model protein is determined by the sequence of monomers. How can one study the folding of such model proteins? The key requirement is that simulations be unbiased to the native state and converge repetitively to one conformation independent of initial conditions—just as real proteins do.

The straightforward and desirable approach to folding proteins even within simplified models is to use the natural amino acid sequences of some moderately sized proteins and simulate their folding by Monte Carlo or molecular dynamics (Wilson & Doniach, 1989; Skolnick & Kolinski, 1990b). However, the major problem with this approach is that protein sequences may have been evolutionarily designed to satisfy folding requirements with a certain "exact" force field. Simulations necessarily use some approximate force field for which the native structure may be neither a global nor a pronounced kinetically accessible local minimum. When the force field is not completely adequate, the natural sequence is effectively random. Therefore, in order to explore this avenue, knowledge of the precise force field is necessary. The attempts to overcome this difficulty were based on the introduction of certain biases [e.g., making only the native contacts favorable (Ueda et al., 1978) or forcing the chain to acquire native secondary structure (Skolnick & Kolinski, 1990b)]. However, model Hamiltonians where such biases are introduced are somewhat unphysical.

A possible approach to unbiased simulations is to study short chains for which some subset of conformations can be enumerated. Then a nonspecific parameter, such as the average attraction between monomers, can be chosen in such a way that the global minimum would belong to this enumerated subset and therefore is known. Folding simula-

† E.I.S. was supported by the Packard Foundation.
• Abstract published in *Advance ACS Abstracts*, July 15, 1994.

0006-2960/94/0433-10026$04.50/0 

tions would reveal whether this conformation of the global minimum is accessible or not.

This approach was taken by Shakhnovich et al., (1991) to study 27-mer chains on a simple cubic lattice and later by Miller et al. (1992) and Camacho and Thirumalai (1993) for 2-dimensional lattices. The folding of model proteins in these works was studied by lattice Monte Carlo simulations (Verdier, 1973; Hilhorst & Deutch, 1975). The important result obtained by Shakhnovich et al. (1991) is that folding and nonfolding sequences exist. The detailed analysis based on folding simulations for 200 random sequences (Sali et al., 1994) showed that the difference between folding and nonfolding sequences is that folding sequences had as their native structure a pronounced global energy minimum (Sali et al., 1994). Unfortunately, such an approach cannot be extended beyond 27-mer chains in three dimensions because the computational complexity of enumeration grows dramatically as chain length increases. However, it is clear that one should not necessarily enumerate all conformations; what is really very important is to know the global minimum conformation and its relation (in energy scale) to the multitude of remaining conformations.

This leads to the idea of combining design and folding to study the folding of longer protein size chains. The idea is simple: to design a sequence that will deliver sufficiently low energy to a given structure, so that one can be certain that this "target" structure represents a pronounced global minimum for this sequence. The specific choice of force field is not essential at this stage, provided that the design of a sequence satisfying the conditions mentioned above is possible with this force field. This sequence then can be subjected to a folding simulation with the same force field that was used at the design stage. At this point, one can hope that the simulation will converge to the target conformation for which the sequence was designed. The key idea here is to use the same force field for the folding simulation and for sequence design. This allows us to address the fundamental questions of protein folding separately from the practically very important but difficult question of which force fields are the most appropriate to study real proteins.

A step in this direction was made in a recent work by O'Toole and Panagiotoupoulos (1992) in which symmetric native structures and a simplified 2-letter, HP (hydrophobic, polar), representation of protein sequences were used. The design was based on the requirement to place more hydrophobic groups inside and hydrophilic groups outside. However, this attempt was not successful for longer chains since the designed sequences did not fold to their target structures. This is likely due to the deficiency of the 3-dimensional, 2-letter HP model, which does not have a stable unique conformation of the global energy minimum (Shakhnovich, 1994).

The idea of combining design and folding was realized successfully recently when an effective sequence design algorithm based on a Monte Carlo (MC) optimization procedure in sequence space became available (Shakhnovich & Gutin, 1993a,b). This made it possible to use a more realistic sequence representation of monomers of 20 types and allowed lattice model folding of proteins of different lengths (36–100) (Shakhnovich, 1994). This approach provides a unique opportunity to study the mechanism by which model proteins solve their folding problems, which is by no means simpler than that of real proteins. Indeed, the shortest of the model proteins we worked with is a 36-mer, which has $4.68^{35} \sim 10^{25}$ conformations (Sykes, 1963), too large a number to be scanned exhaustively. (For 100-mers, which also fold in our simulations, this number is $10^{75}$!) Since we eventually can trace any intermediate conformation in the simulation, a very detailed study of the mechanism of folding can be done for the model proteins.

The statistical mechanics of proteins with designed sequences was discussed by Shakhnovich and Gutin (1993a), who showed that the sequences undergo a first-order folding transition to the native state. [For the qualitative explanation of the nature of first-order transitions in biomolecules, see Karplus and Shakhnovich (1992).] The phenomenological model of Bryngelson and Wolynes (1987), where the idea of design was encapsulated in the "principle of minimal frustration", also implied that transition to the native state may have first-order character. But the mechanism of first-order transitions is known to involve nucleation and growth (Lifshitz & Pitaevskii, 1981). Therefore, it is natural to expect the nucleation growth mechanism of protein folding.

The idea of a nucleation growth mechanism in protein folding was suggested by Levinthal in a largely unavailable publication (Levinthal, 1969) and was pursued in the subsequent work of Tsong et al. (1972) on the basis of kinetic analysis of experimental data and by Wetlaufer (1973) on the basis of observation of existing protein structures. The nucleation mechanism was also discussed in a recent work (Mault & Unger, 1992). In these works, the nucleation growth mechanism was based on phenomenological models, and detailed microscopic study to support or reject this hypothesis was missing. In this study, we suggest a detailed microscopic analysis on the basis of the lattice model of protein folding.

## METHODS

We use a 36-mer chain on a cubic lattice as a basic model (some results for longer chains will be sketched in the Discussion). We tried two different arbitrarily chosen compact native structures in order to determine which conclusions depend on the structural features of the native state and which are independent of it (Figure 1). The next step was to design a sequence that fits the native structure with a low energy. To this end, we used a MC optimization algorithm in sequence space, documented in detail by Shakhnovich and Gutin (1993a,b).

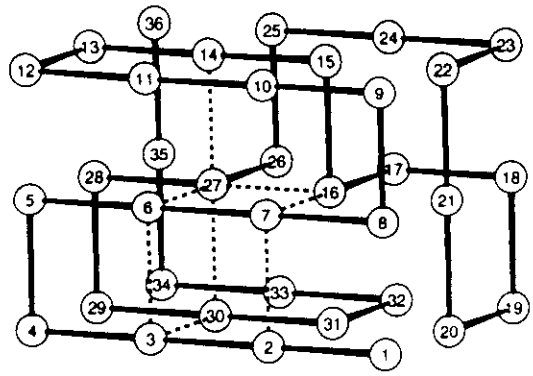
The energy function that we used throughout this work is taken in the nearest-neighbor approximation (Miyazawa & Jernigan, 1985):

$$E_0(\{\sigma_i\}) = \frac{1}{2}\sum_{i,j}^{N} U(\sigma_i\sigma_j)\Delta(r_i^{\text{native}} - r_j^{\text{native}}) \qquad (1)$$

where $N = 36$ is the total number of monomers and $\Delta$ defines the contact potential between them: $\Delta(r) = 1$ if $r_{\text{low}} < r < r_{\text{high}}$ and $\Delta(r) = 0$ otherwise. Our model protein is positioned on a simple cubic lattice with bond length of 3.8 Å. The target native conformation is set through the coordinates of its monomers $\{r^{\text{native}}\}$. Any two monomers that are 3.8 Å apart (so that, say, $r_{\text{low}} = 3.7$ Å and $r_{\text{high}} = 3.9$ Å) are considered to be in contact. For the set of potentials $U(\sigma_i\sigma_j)$, we used parameters determined by Miyazawa and Jernigan (1985) (MJ) from the statistical distribution of contacts in native proteins. The sequence design algorithm was run at low selective temperature [see Shakhnovich and Gutin (1993a)], $T_{\text{sel}} = 0.2$, to provide sequences that fit the native structure with sufficiently low energy.

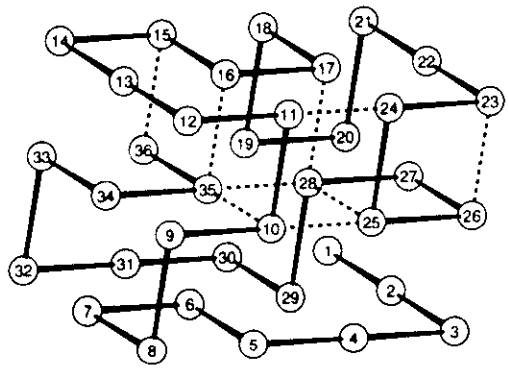
The MC procedure in sequence space requires the initial setting of amino acid composition. We tried several choices. First we designed proteins with an "average" amino acid

a



SQKWLERGATRIADGDLPVNGTYFSCKIMENVHPLA
GTDLYRDGLNENYRQRYAVSTFVPQPDPIHDVYLQP
PDHLLDRYSTRVVDGESYFYHTNTTSRILERCSLAA

b



GPVQNDINIEHLSAGDPGRFCWLYKVMERTSLAAKT

FIGURE 1: Target conformations used in this study. Sequences are shown that fit the corresponding conformations with sufficiently low energy to make sure that these conformations are global energy minima for the designed sequences. We worked with three sequences designed for structure a and one sequence designed for structure b. Most calculations were done with structure a; however, for the sake of control the nucleus was also determined for structure b. Dashed lines denote contacts belonging to the nucleus.

composition taken from Table 1.1 of Creighton (1992). Another choice was to take a composition like that of the small 36-residue protein pancreatic bird polypeptide (1 ppt).

Since we are using MJ parameters that were obtained from protein statistics, we have only relative energy and do not know the absolute energy scale for this set of parameters (Finkelstein et al., 1993). So we use the energy unit at which DU = 1, where DU = $((\langle U^2\rangle - \langle U\rangle^2))^{1/2}$ is the standard deviation of the energy of different interactions; this is the measure of their heterogeneity. $\langle\ \rangle$ denotes averaging over all possible pairwise interactions in the given sequence:

$$\langle U^{\mathrm{P}}\rangle = \frac{2}{N(N-1)}\sum_{i<j}^{N} U^{\mathrm{P}}(\sigma_i\sigma_j) \qquad (2)$$

The design procedure generated a number of sequences; we intensively studied the ones shown in Figure 1.

The lattice Monte Carlo simulations of the folding of designed sequences are done with a standard algorithm well documented in earlier works (Verdier, 1973; Hilhorst & Deutch, 1975; Sali et al., 1994). The standard move set was taken to include corner flips and crankshaft motions (Hilhorst & Deutch, 1975). The Metropolis criterion with the energy function (eq 1) was used (Metropolis et al., 1953) to accept or reject moves.

To measure the structural similarity between a current conformation and the native state, we used the similarity parameter $Q$ (Shakhnovich & Gutin, 1989a,b, 1990a), which is the normalized number of native contacts in a conformation:

$$Q = \frac{N_{\mathrm{native}}}{N_{\mathrm{total}}}$$

where $N_{\mathrm{total}}$ is the number of contacts in the compact conformation; $N_{\mathrm{total}} = 40$ for the 36-mer. It follows from this definition that $Q = 1$ in the native state.

Simulations started from the random coil (see an example in Figure 2) and ended when the native target structure was reached (Figure 3). The mean first passage time for reaching the native state was $\sim 10^6$ Monte Carlo steps at $T = 0.90$, at which all simulations reported in this work were performed. The native conformation (shown in Figure 1a,b for corresponding sequences) had the lowest energy among all conformations found in the simulations. To test this, a long simulation of $10^9$ Monte Carlo steps was run to make sure that no other structures with energy equal to or lower than the energy of the native structure were encountered. This was indeed the case, which made us sufficiently confident that the native is the global minimum of energy.

## SEARCH FOR THE NUCLEUS

Exploring implications of the first-order transition kinetics of folding we expect that the chain overcomes the main free energy barrier via a nucleation growth mechanism. There are two slightly different definitions of nuclei in the kinetics of the first-order transitions (Lifshitz & Pitaevskii, 1981). The critical nuclei correspond to transition states (free energy barriers). There is a probability of roughly 1/2 that the new phase will grow further after the critical nucleus is formed and a probability of 1/2 that it will dissolve. One can also define a postcritical nucleus, i.e., the minimal sized fragment of the new phase that inevitably grows further to the new phase. Certainly there is no great difference between the two ways of defining the nucleus because the postcritical nucleus simply should have energy a few $k_{\mathrm{B}}T$ lower than the critical one, the barrier state, in order to make the subsequent growth unidirectional and irreversible. In our study, we will be interested in postcritical nuclei, i.e., ones that subsequently grow into the folded state.

The main difficulty in finding a nucleus comes from the fact that they are very short-lived before they grow further into the native or near-native conformation. By no means should they be confused with intermediates that are long-lived and detectable because they are sufficiently deep local minima. We define a nucleus as a set of contacts that satisfies the following two conditions: (i) Formation of a nucleus is a sufficient condition for folding; i.e., after a set of contacts that constitutes the nucleus is formed, the subsequent folding is guaranteed and is very fast (in our search for a nucleus we required that folding should take place in less than 50 000 MC steps after the nucleus is formed). We are therefore looking for postcritical nuclei. (ii) Formation of a nucleus is a necessary condition for folding; i.e., the pattern of contacts corresponding to the nucleus is *always* present in "prefolding conformations" when the number of native contacts is relatively small, but subsequent folding is very fast.

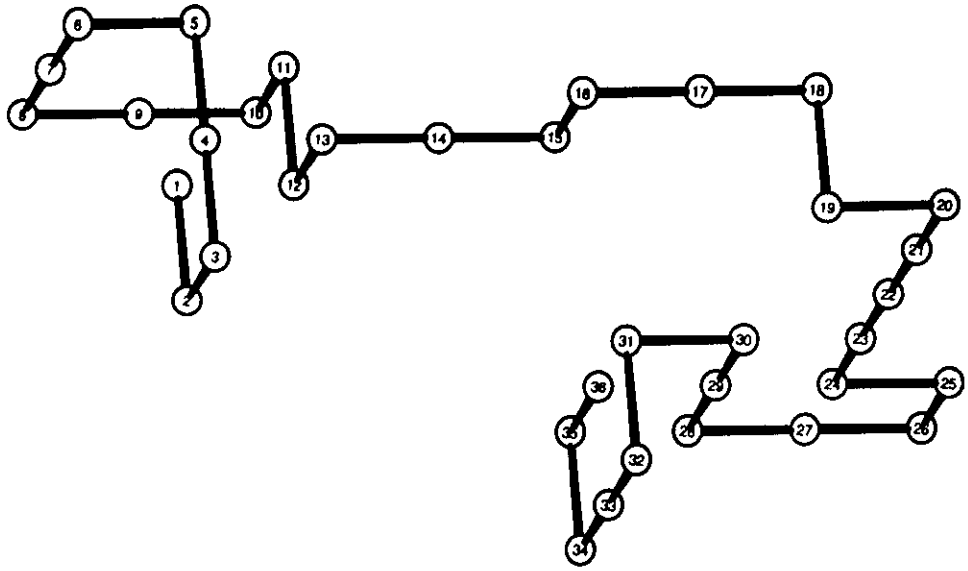


FIGURE 2: Example of a starting random coil conformation.

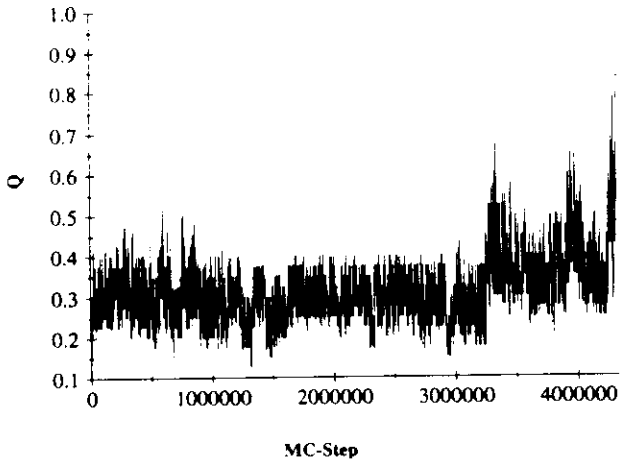



FIGURE 3: Example of a folding trajectory starting in the random coil conformation and ending in the native state. Such types of trajectories were used throughout this study.

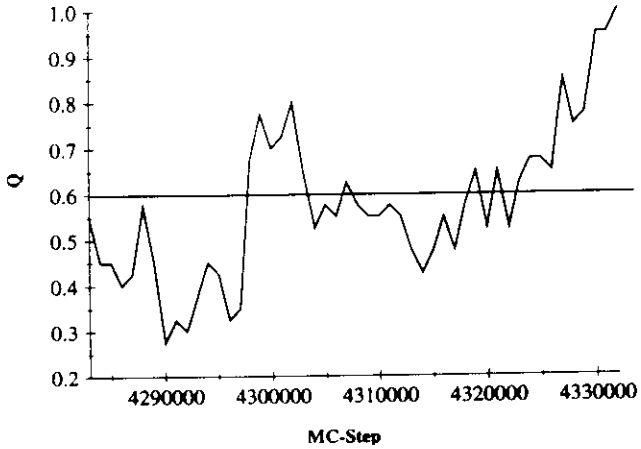



FIGURE 4: Part of the folding trajectory from Figure 3 used to search for the nucleus. The horizontal line illistrates the criterion $Q < 0.6$ for the choice of conformations relevant to the search for the nucleus.

The last condition requires some explanation. It is trivial that in the vicinity of the native state where $Q \approx 1$, some contacts will consistently appear just before the native state is reached. What we are interested in is the *minimal* set of contacts that *must* be formed before folding proceeds to the native state. To this end, we should analyze conformations that are not too close to the native state. As inspection of Figure 3 suggests, in the largest part of the trajectory the chain is fluctuating in conformations with $Q$ not exceeding 0.6. This means that we should search for nuclei by analyzing sets of contacts that are present in conformations belonging to steep parts of the trajectory (Figure 4) but that are structurally different from the native state. To this end, we analyzed all conformations with $Q < 0.6$ (see Figure 4) that are separated by less than 50 000 Monte Carlo steps from the final step of the simulation when the native state was reached. The data were collected over 10 runs, each starting from a random coil and ending in the native conformation. Our analysis was aimed at revealing the set of contacts common to all 10 runs.

We discovered that rapid folding always takes place after the formation of a distinct set of eight contacts (shown by dashed lines in Figure 1a) for the first target structure and nine contacts for the second target structure (Figure 1b). We can see that contacts forming the nucleus are located in the native structure in the vicinity of each other, not in random positions. These contacts form a spatially localized substructure, which serves as a nucleus for folding. The formation of this set is indeed both necessary and sufficient for fast folding. It is the necessary condition because this set of contacts is formed for the first time only several thousand Monte Carlo steps before the native state is reached and in conformations for which the number of native contacts is relatively small (less than 25 out of 40). It was also a sufficient condition because after the nucleus had been formed the native state was always reached in less than 50 000 Monte Carlo steps, or about 1% of the total Monte Carlo time of folding from a random conformation.

Another important finding was that the position of the nucleus was nonspecific to the sequence chosen: for all three sequences shown in Figure 1a, the position of the nucleus was the same. We analyzed folding trajectories for 30 more sequences designed to have the native structure, as shown in Figure 1a, and found that in all these trajectories formation of the nucleus shown in Figure 1a preceded subsequent fast folding to the native state. To avoid confusion here, we should stress that although these sequences are nonhomologous, they are not independent either: they were all designed to have the structure shown in Figure 1a as the global minimum conformation.

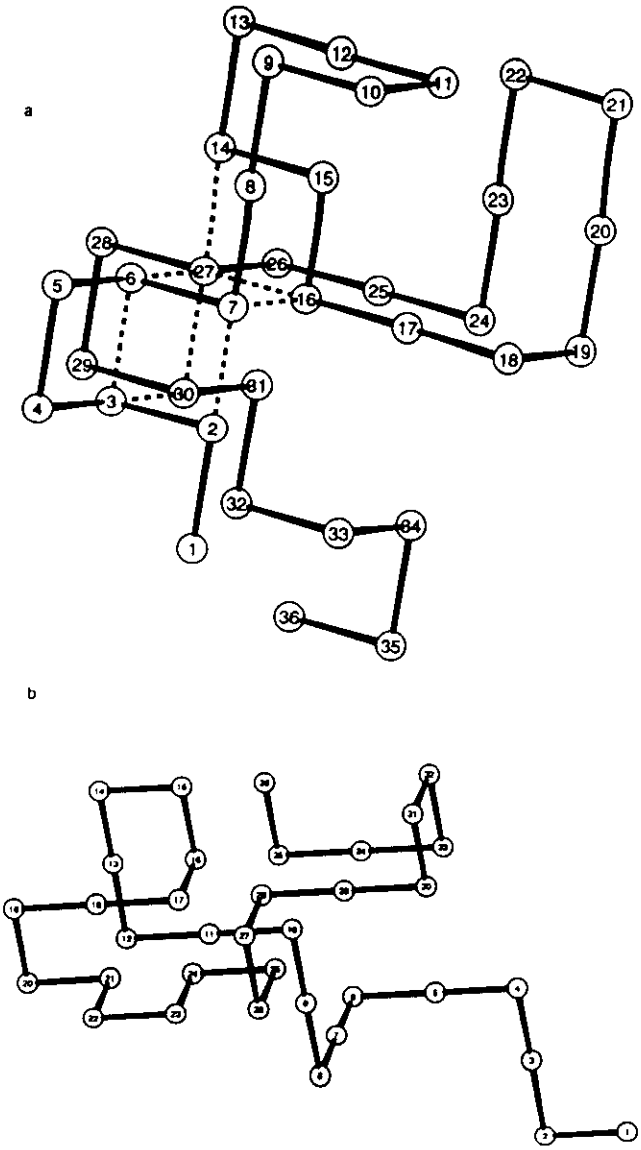


FIGURE 5: (a) Example of a starting conformation containing nucleus contacts. Otherwise the conformation was random. (b) Control: Starting conformation containing the same number of native-like contacts as in a, but without nucleus contacts.

## EXPLORING THE NUCLEATION MECHANISM

As the first test of the proposed nucleation mechanism, we studied folding trajectories that started from a conformation with a preformed nucleus; otherwise this conformation was completely random and noncompact (see Figure 5a). It contained about ten native contacts (eight belonging to the nucleus and two randomly formed). Therefore, $Q \approx 0.25$ in a starting conformation. When simulations were started from conformations with the preformed nucleus, as shown in Figure 5a, the native state was reached quickly (on average in less than in 30 000 MC steps, and in many runs in less than in 1000 MC steps). The time course of a typical simulation, which started from a conformation with a preformed nucleus is shown in Figure 6.

However, the question may arise whether fast folding from a conformation with a preformed nucleus is due to formation of the nucleus or whether any eight native contacts in the starting conformations provide such fast folding. In order to address this issue, we ran a control experiment starting from several conformations that contained at least eight native contacts but that were different from the nucleus ones (see

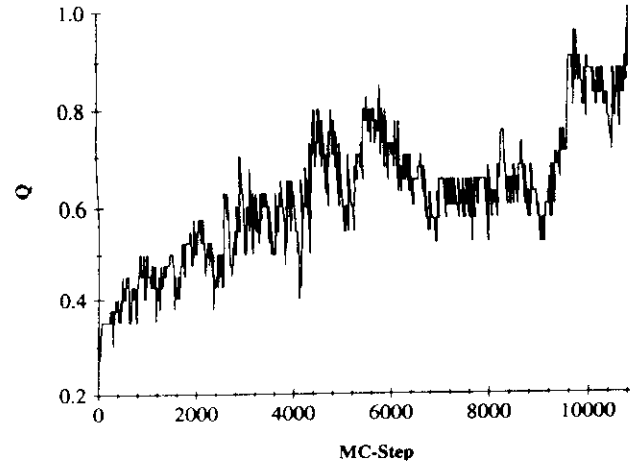



FIGURE 6: Typical folding trajectory for runs that start from a conformation with a preformed nucleus, as shown in Figure 5a.

an example of such starting conformation in Figure 5b). In all of these control experiments, the folding trajectories were practically indistinguishable from the ones that started from completely randomized conformations (Figure 3). The folding time distribution was unaffected by the choice of initial conformation in this case and yielded the same mean first passage folding time as before of close to 1 million MC steps. This can be rationalized if we look at any arbitrary trajectory that starts from a random coil (Figure 3). In fact, 8–10 native-like contacts (i.e., conformations with $Q \approx 0.2$–$0.25$) are formed at the very beginning of the simulation (in less than 100 000 MC steps). However, this does not lead to rapid folding: a few million more steps are required to reach the native structure. Only formation of the *specific* subset of contacts, the nucleus, results in rapid folding.

As was mentioned before, formation of the postcritical nucleus corresponds to the transition over the main free energy barrier. This implies that there must be a significant difference in folding mechanism when the simulations start from completely randomized conformations and when they start from a conformation with a preformed nucleus, as shown in Figure 5a. In the first case, one should expect that the rate-limiting stage is overcoming the main barrier or formation of the nucleus, while in the latter case the motion to the native state would be downhill in free energy space, representing an effective pathway or funnel (Leopold et al., 1992).

To test this, we compared the statistical characteristics of the folding process in both cases. In the case where folding started from a random conformation, we evaluated at each trajectory, after each 1000 MC steps, the number of all current contacts ($y$) as well as the number of the native contacts ($x$). The frequency with which specific pairs $(x,y)$ were found in 10 folding trajectories was evaluated to calculate the probability $P(x,y)$ of finding a conformation with $y$ contacts, $x$ of which are the native ones. These results are illustrated in Figure 7a. Both $x$ and $y$ can take values from 0 to 40, and $41^2 = 1681$ dots correspond to 1681 possible pairs of $x$ and $y$. The higher $P(x,y)$, the lighter the corresponding dot on Figure 7.

We should note here that our experiments were aimed at the estimation of the mean first passage time, and therefore simulations ended when the native conformation had been reached. This explains the apparent low population of the native state in Figure 7. In fact, the native state was rather stable at that temperature, having $\langle Q \rangle \approx 0.8$ where $\langle\ \rangle$ denotes thermal averaging over long (equilibrium) trajectories.

One can see that conformations with approximately 25 total and 15 native contacts are most frequent. This is certainly

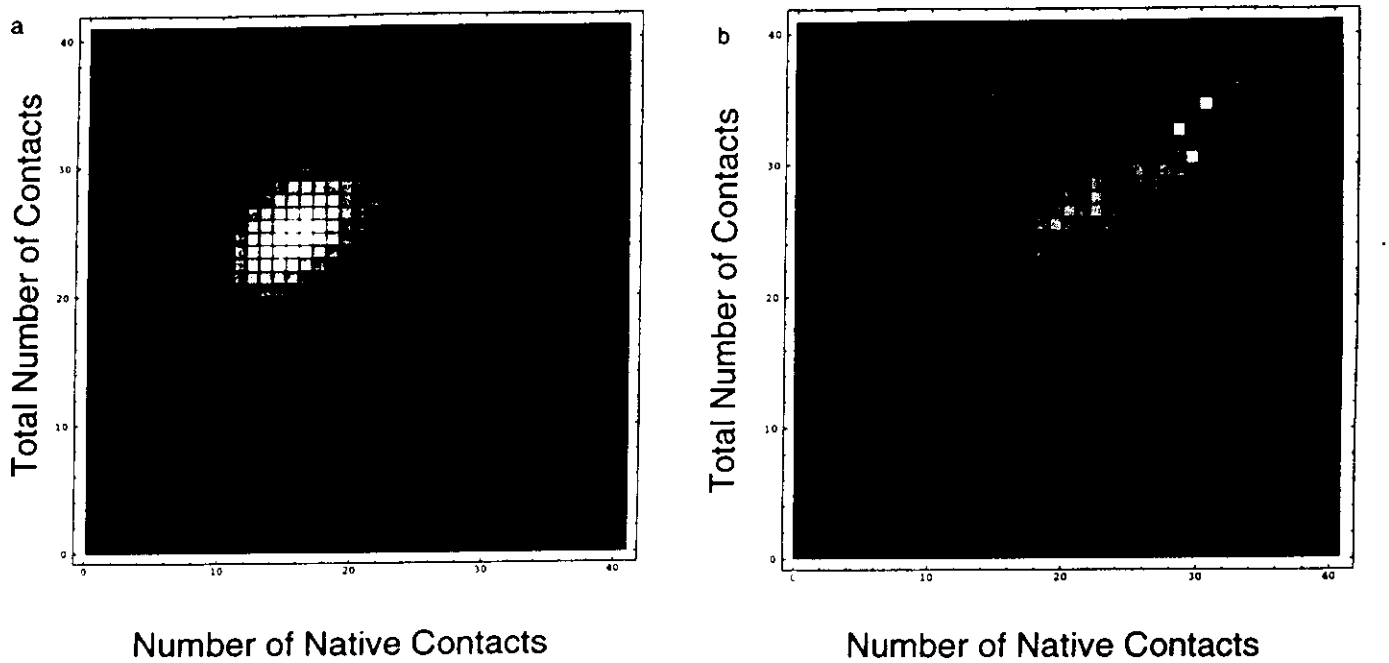



FIGURE 7: Density plots illustrating the frequencies with which conformations having a specified number of native contacts (abscissa) and total number of contacts (ordinate) are found in simulations. The brighter the dot with coordinates $(x,y)$, the more frequently conformations with $x$ total and $y$ native contacts were found. (a) Simulations starting from completely random conformations; (b) simulations starting from conformations with a preformed nucleus, as shown in Figure 5a.

a prebarrier minimum of free energy or a folding intermediate. Conformations with more than 30 native contacts are rare. This means that the chain spends most of its folding time fluctuating around the intermediate state until it reaches a conformation(s) corresponding to the free energy barrier, after which folding is fast. The computer experiments described earlier show that this is the set of conformations containing the nucleus.

The same calculations were performed when folding started from conformations that contained a preformed nucleus, as shown in Figure 5a. The only difference was that the numbers of native and all contacts, $x$ and $y$, were evaluated at each tenth Monte Carlo step because the folding time was substantially smaller. The results are illustrated in Figure 7b. There is a clear difference between the plots shown in Figure 7a,b. In the case of folding from a preformed nucleus, the number of native contacts is very strongly correlated with the number of all contacts, as the light area on Figure 7b is stretched along the main diagonal $x = y$. It is also important to note that there is no maximum of $P(x,y)$ on Figure 7b, and we can see that $P(x,y)$ is approximately constant in the area close to the diagonal $x = y$ and vanishes everywhere else, which implies that in this case the chain is not wandering randomly through conformational space but folds quickly, increasing the number of native contacts at an approximately constant rate (a clear indication of the propagation mechanism). Of course the polypeptide chain still has a tremendous number of conformations, but the constant value of $P(x,y)$ suggests that a directed assembly takes place after the nucleus is formed. Thus, the addition of any native contact decreases free energy, and this driving force directs the process. No significant free energy barriers are found in this part of the configurational space.

We studied the role of the nucleus in the initiation of folding in our model. However, for conformations with a nucleus, proximity to the transition state may also play an important role in unfolding. To this end, we studied longer trajectories, during which several folding–unfolding events occurred (Figure 8). Inspection of these trajectories reveals two possible

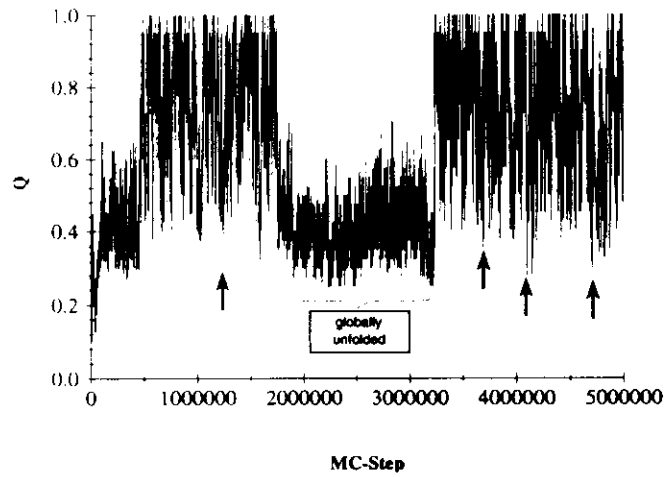



FIGURE 8: Part of a longer trajectory containing local unfolding events shown by arrows and global unfolding shown by the bracket. Local unfolding is as deep as the global one; however, locally unfolded conformations usually refold in less than 20 000 MC steps.

scenarios of transient unfolding. The first type of behavior corresponds to significant unfolding (up to $Q \approx 0.2$), but after 10000–20000 MC steps the chain refolded back. Such unfolded conformations after which the chain refolds quickly will be called locally unfolded. However, sometimes the same degree of unfolding to $Q \approx 0.2$ led to more dramatic consequences: a few million MC steps were required for the chain to refold (see Figure 8). Conformations that required such a long time to refold will be called globally unfolded.

The question then is what is the difference between globally and locally unfolded conformations? We studied 10 different long (up to 100 million MC steps) folding trajectories (part of one is shown in Figure 8) and examined all locally unfolded conformations with less than 16 native contacts. We found that all of these conformations contained the intact nucleus, while globally unfolded conformations missed contacts from the nucleus. An implication of this observation is that although fluctuations in the folded state are significant, some contacts

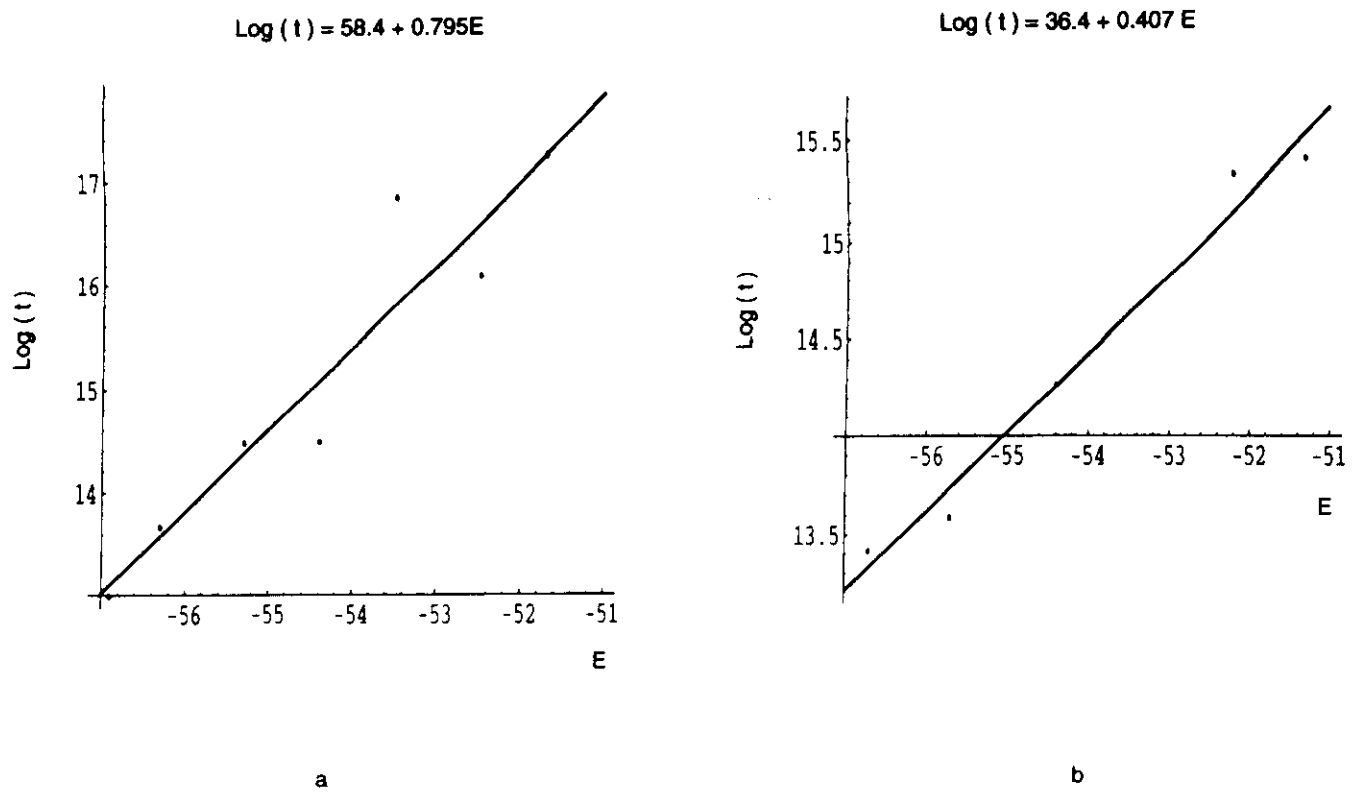



FIGURE 9: (a) Dependence of the log of MFPT of folding on the energy of the native state for a number of sequences having different energy of nucleus contacts and the same total energy of the other contacts. (b) Same dependence but for the set of sequences having the same energy of nucleus but different total energy of remaining contacts.

are more stable than others—a clear indication of the heterogeneity of the folded conformation in our model, which we relate to the molten globule state in proteins (see the Discussion section). This result shows also that there is no other nucleation site in our chains. If there were, we would see either of the nucleation sites preserved in locally unfolded conformations, but we definitely see (repetitively) only one subset of contacts common to all locally unfolded conformations. We note, however, that this conclusion is drawn for 36-mer chains, and it is certainly possible that longer chains may have multiple nucleation sites. A very interesting question, then, is at what size of the chain (if any) the nucleation regime changes from one nucleus to multiple nuclei.

The next issue we addressed was the dependence of the folding time on the interaction energy of the contacts constituting the nucleation site. To this end, using the same design procedure (Shakhnovich & Gutin, 1993a,b), we selected a set of sequences having a different total energy for the 8 nucleus contacts but a similar total energy for the remaining 32 native contacts. The objective was to study how the stability of the nucleus affects the rate of folding. The result is presented in Figure 9a, where dependence of the logarithm of the folding time on the total energy of the native conformation, normalized by $kT$, is shown. We would like to emphasize that although we plot the mean folding time vs the *total* energy of the native conformation, the sequences corresponding to the different data points in Figure 9a differ by the energy of nucleus contacts only, having similar energy for the remaining contacts in the native conformation. The dependence presented in Figure 9a is close to linear with a slope of 0.8. This should be contrasted with the results of a control experiment in which sequences were chosen to have similar energy for nucleus contacts and differ in energy for the remaining contacts (Figure 9b). In this case, the dependence of log(time) on the energy of the structure is also close to linear, but the slope is half as great (0.4). This indicates that stabilization of the nucleus is more important for rapid folding than the stabilization of other contacts, although the latter may indirectly stabilize the nucleus, decreasing the entropic cost of its formation. This gives rise to the acceleration of folding in that case.

## DISCUSSION

In this section, we will discuss two aspects of the present study. First, we discuss the lattice model results and their implications. In the second part of the Discussion, we will discuss the applicability of simplified lattice models to the study of the folding of real proteins: features that lattice models catch and features that they miss.

In this paper, we have provided a body of evidence that the folding mechanism of lattice proteins involves the formation of a *specific* nucleus as a *transition state*, with its subsequent growth. This is not at all unexpected because nucleation growth is a standard mechanism of cooperative (first-order) transitions; for instance, the vapor–liquid transition is well known to involve a nucleation growth stage (Lifshitz & Pitaevskii, 1981). There is, however, an essential difference between the nucleation growth mechanism in simple liquids and that in model proteins. In liquids a nucleus is nonspecific and is fully characterized by its size. In model proteins the nucleus is *specific*, which means that a particular set of contacts, constituting a *transition state*, should be formed to cause subsequent fast folding to the native state.

The folding process in each molecule involves two stages, which we can characterize as stochastic and deterministic. The stochastic stage is rate-limiting (the stage at which the nucleus is formed via random search). Of course this does not imply that the protein should "wait" for a multiparticle collision to form the nucleus. Since the nucleus is a substructure of the native state, its contacts are attractive and therefore the partly formed nucleus does not disappear. The possibility of a stochastic search to form nuclei was pointed

out by Wetlaufer (1973). The stochastic search for the nucleus takes place in the intermediate that is formed at the burst stage of the folding process (in less than 30 000 MC steps). This burst intermediate can be seen as a light area on Figure 7a as a partly compact state (having 20 out of 40 contacts, 10–12 of which are the native ones). This intermediate represents a multitude of rapidly interconverting conformations, corresponding to a prebarrier free energy minimum. Formation of a burst semicompact intermediate precedes the formation of a nucleus, which is formed later when native contacts in the intermediate include, for the first time, the nucleus ones.

The subsequent folding, after the nucleus is formed, is fast and practically unidirectional. This is not surprising because formation of a nucleus is equivalent to overcoming the main free energy barrier. We observe *a folding pathway* that is the postnucleus assembly of the protein associated with the directed motion downhill in free energy. Indeed, as inspection of Figure 7b suggests, the number of native contacts grows steadily with the increase of the total number of contacts, i.e., roughly speaking, in this regime every added contact is a native one. It can also be seen that the chain does not encounter, at this temperature, significant barriers as it progresses through the pathway; the motion in configurational space is rather diffusion-like. The evidence for this is the approximate constant density in the light region of the diagram of Figure 7b.

It was suggested in previous works, implicitly (Wetlaufer, 1973) or explicitly (Rooman et al., 1992a,b), that at least a considerable part of the nucleus should be formed by contacts between residues that are close to each other in sequences (local contacts). Our analysis is consistent with these assertions. Inspection of Figure 1 shows that the nucleus is formed by both long-range as well as short-range contacts, with some predominance of the long-range contacts. However, the relation between the numbers of short-range and long-range contacts in the nucleus may depend on the potential chosen since the local component of the potential may increase the number of local contacts in the nucleus. This question requires further study. We believe, however, that some long-range contacts must always be present in the nucleus since such contacts are most effective in decreasing entropy of the transition state and thus creating an "entrance" to the pathway.

The results reported in the present paper were obtained for the 36-mer model proteins. A very important question is whether these results are valid for longer sequences. Our approach allows for folding longer sequences (at least up to 100-mers) (Shakhnovich, 1994a,b). To test the conclusions of a,b this work, we studied a nucleation mechanism of folding for a 80-mer chain. Using the same procedures as described in this work earlier, we found the nucleus for the 80-mer to have 22 out of 105 contacts. This included 16 monomers. Although we observed a single nucleus for the 80-mer chains, we cannot exclude a multiple-nuclei mechanism for longer chains. These multiple nuclei could be associated with folding domains [observed recently in hen lysozyme (Miranker et al., 1991)], which may or may not develop into the structural domains of native proteins.

The folding of long chains (36–100 residues) was possible only because these sequences were designed to have the native state as a pronounced energy minimum, and a special design procedure was necessary to generate such sequences (Shakhnovich & Gutin, 1993a,b). Long random sequences were not able to fold (Shakhnovich, 1994a). A complementary approach to study folding was taken in the recent paper by Sali et al. (1994b), where *short random* sequences were taken to study the "minimal requirements" for "one-shot" selection of folding sequences from the pool of random sequences. Analysis of the folding of short quasirandom folding sequences also revealed an activation mechanism, but it differs from the one found in the present study. The transition state for short random sequences turned out to contain 80–95% of native contacts (compare with the nucleus that has 8 particular contacts out of 40). This difference may be due to a number of reasons. First of all, a random interaction energy model was studied by Sali et al. (1994b), while here we studied a more realistic sequence model. This difference may be important since in the former model energies of contacts are totally uncorrelated, while in the latter energies of different contacts are correlated. Indeed, the identity of a protein is characterized in the letter model by $N$ "letters" (primary structure), which determines $\sim N^2$ interactions between every pair of monomers. This implies that these interaction energies cannot be independent. In the random interaction energy model, the identity of a protein is defined through setting all $\sim N^2$ interactions between any pair of contacts independently. Correlations may be important for nucleus formation, which is a contiguous subset of native, stable contacts. A second reason, which is more likely to explain the difference between the results of two models, is that present sequences were designed to enable the folding of long chains. It is likely that design in the sequence model generated a contiguous subset of strong contacts, which turned out to be a nucleus. It was pointed out by Sali et al. (1994b) that the model used there is likely to describe the folding of prebiological, short, and poorly optimized sequences. As longer proteins evolved, their folding may have required sequence design that developed a more effective nucleation growth mechanism. Indeed, the characteristic folding "time" of random 27-mers in Sali et al. (1994b) was 20–50 million steps, while in the present study 36-mers fold in 1–5 million steps and designed 80–100 mers fold in 5–10 million MC steps (Shakhnovich, 1994a,b).

The results reported in this paper were obtained using Monte Carlo simulation in the lattice model. Two questions are in order now: how representative is Monte Carlo for the kinetics of folding, and what is the relationship between lattice models and real proteins?

A comprehensive study of the role of lattice and move sets in the apparent dynamics of a polymer was performed by Skolnick and Kolinski (1990a, 1991), who showed that there is no significant dependence of observed dynamics on the choice of lattice (diamond or 210) or move set. Moreover, Rey and Skolnick (1991) compared the simulation results obtained by Monte Carlo on the simplest (diamond) lattice and by off-lattice Brownian dynamics. Their conclusion is that the main dynamic features observed are independent of the simulation technique chosen. It was shown also by Skolnick and Kolinski (1990) that the choice of local moves only, being most natural, provides the most realistic time scale picture, as judged in comparison with the master equation calculation.

Thus, in our view, the Monte Carlo approach (taking into account its computational effectiveness) may be plausible for depicting key features of kinetic processes associated with protein folding. However, it is unlikely that MC simulations can provide a description of all of the microscopic details of the process. Rather, general features, which are observed over thousands of steps, are of interest. This is the case in our study in which we focus on nucleus formation that takes place in $10^6$ steps.

The most important question concerning the approach taken in this study concerns the relationship between lattice model

proteins and real proteins. There is one obvious feature of real proteins that our model misses. This is the presence of side chains with their degrees of freedom and tight packing in the native state. Therefore, our model is aimed toward describing stages (if any) of the protein folding process that do not include the tight packing of side chains. It is widely believed now that packing of side chains occurs at the transition from molten globule (MG) to native (N) conformation. Experimental evidence has been accumulating (Williams et al., 1991; Hughson et al., 1990; Peng & Kim, 1994) that the molten globule, when at equilibrium, retains a significant part of the native-like backbone fold, in accord with theoretical predictions (Shakhnovich & Finkelstein, 1982, 1989; Finkelstein & Shakhnovich, 1989). It was suggested (Ptitsyn, 1973, 1987) that a "native-like" molten globule may be a universal intermediate on the protein folding pathway. Subsequent experimental findings (Ptitsyn et al., 1990; Matouschek et al., 1990, 1992; Jennings & Wright, 1993) strongly buttressed this point, providing evidence [especially in Jennins and Wright (1993)] that the transient long-lived intermediate is structurally close to the equilibrium "native-like" molten globule.

The folded state in our model should be related to the "native-like" molten globule. It is interesting to note that a chain in this state fluctuates around the native fold, but these fluctuations are inhomogeneous (nucleus contacts are fluctuating less than other contacts) (see Figure 8). This is in accord with experimental information about the molten globule (Hughson et al., 1990; Baum et al., 1989).

The nucleus transition state that we observe in this work is the transition state between a coil, or a structureless compact intermediate without unique structure (Elove et al., 1992; Radford et al., 1992), and the molten globule with elements of native-like fold. By no means should it be confused with the transition state between the native state (N) and the molten globule (MG), which is usually associated with the transition state for folding because the MG–N transition is the rate-limiting step for the whole process. This transition N–MG state is known, both from theory and experiment (Segawa & Sugihara, 1984; Shakhnovich & Finkelstein, 1989; Bycroft, 1990), to be close to the native state, differing from it by some small expansion [so small that protein core is mainly inaccessible to the solvent: see Segawa and Sugikhara (1984) and Matouschek et al. (1992)].

A significant simplification of the model is that it did not include explicitly secondary structure segments, which are stabilized by H-bonds and are able to move as a whole. This question is related to the secondary structure framework and related diffusion–collision hypotheses of folding (Kim & Baldwin, 1982; Karplus & Weaver, 1976). The physical mechanism assumed in these hypotheses is that native-like secondary structure is formed at early stages so that subsequent folding includes movements of segments as a whole, without their restructuring due to long-range interactions. This may give a kinetic advantage because the degrees of freedom associated with secondary structure become frozen, and the remaining search is feasible because it includes far fewer conformations. Therefore, in order to facilitate kinetics, secondary structure elements, after having been formed at the ultrafast stage of folding, should be so stable that their characteristic folding–unfolding interconversion time in the absence of long-range interactions is longer than the time of formation of long-range contacts [in the millisecond time range (Radford et al. 1992; Bycroft et al., 1990; Jennings & Wright, 1993)]. The only way to increase the interconversion time from basic nanoseconds to milliseconds, which is consistent with the second law, is to increase the stability of the helix. This requires ~10 kcal/m/helix of stabilization, which implies that the Boltzmann probability of such a stable isolated helix will be very close to 1. Recent studies of isolated fragments of myoglobin corresponding to helical segments in its native secondary structure did not lend evidence supporting the suggestion that isolated helixes are stable in the absence of long-range interactions (Waltho et al., 1993; Shin et al., 1993). Certainly, some fluctuating elements of native-like and nonnative secondary structure may form quickly. However, it is unclear (at least to us) how the formation of marginally stable fluctuating $\alpha$-helixes and $\beta$-strands, with their degrees of freedom in equilibrium with all other degrees of freedom, can provide any *kinetic* advantage leading to the resolution of Levinthal's paradox.

Of course our calculations cannot rule out the framework-type mechanism because movements of helixes or $\beta$-strands as a whole are not included in the move set. However, what they show is that this mechanism, even if valid, is not the only, or necessary, way to solve the Levinthal paradox. Our calculations give an *example* that the protein folding problem, at a model level, can be solved without a framework-type mechanism.

The sequences we worked with in this model were designed to have the native conformation as a pronounced global *energy* minimum. The question is how can this optimization be related to real proteins. First of all, we note that a pronounced energy gap between the native state and the set of nonnative conformations is a *necessary* thermodynamic condition of the uniqueness of the native structure; this is independent of the model or the potential function chosen. The native structure must be thermodynamically stable at physiological temperature. This can be guaranteed only if the gap between the native structure and nonnative conformations is sufficiently large, i.e., many $kT$ (Shakhnovich & Gutin, 1990). In other words, a large energy gap protects a unique structure from destruction by thermal fluctuations. However, our results go further and suggest that a pronounced energy gap is also a *sufficient* condition for sequences to fold rapidly to the native conformation.

These considerations do not contradict the fact that proteins are not highly stable. Experimental results (e.g., Privalov, 1979) suggest that the temperature of denaturation for most proteins is not too high, and therefore the difference in free energy between the native conformation and denatured states is moderate: 10–12 kcal/mol for a 100-residue protein at physiological temperature (Privalov, 1992). In order to give a correct interpretation of the thermodynamic data on protein stability, one should note that what is known to be small is the difference in *free energy* between the native and denatured states; this includes the entropic contribution. Energy differences between the native and denatured states are much more pronounced, as measured by the latent heat of denaturational transition and its cooperativity. The entropic factor is also essential for lattice proteins, making the unfolding temperatures not too high ($\approx$1.1 in our energy units) and the lattice proteins marginally stable, like real ones.

## CONCLUSION: IMPLICATIONS FOR EXPERIMENT

In this study, we have presented a minimal theoretical model of protein folding. The model is free of internal inconsistencies and unphysical assumptions. Indeed, the simulations are not artificially biased toward the native state: all the chain "knows" when the simulation starts from a random coil conformation is the amino acid sequence. The Hamiltonian is physical: the interaction of, say, glycine with another glycine depends only

on the spatial distance between the residues and does not depend on their positions in the chain or in the native conformation. Model proteins resolve the Levinthal paradox, exhibiting fast folding to the unique global minimum conformation without scanning the astronomically large number of possible conformations. We presented the possible mechanism.

Like any theoretical work, this one deals with a simplified representation of proteins, and the adequacy of the model for the system it studies is at issue. The only nontautological way to estimate the adequacy of a model is to formulate its predictions and compare them with experiment.

It should be noted here that the model studied in this paper represents a generic protein and is aimed toward the study of *universal* features of protein folding unrelated to the specific structural features of a protein molecule.

The theoretical analysis presented in this work has several implications directly related to experiment, as follows.

(1) The cooperative character of the coil–molten globule transition in natural (i.e., evolutionarily optimized) protein sequences contrasted with the non-cooperative character and absence of unique structure in the randomized sequence. This explains the difference in experimental results for proteins [bovine carbonic anhydrase and staphylococcal $\beta$-lactamase (Uversky et al., 1992) and staphylococcal nuclease (Gittis et al., 1993)], where "all-or-none" transitions were reported, and for the quasirandom sequence of the F2 fragment of tryptophan synthase (Chafotte et al., 1991), where the transition is non-cooperative.

(2) Theory demonstrated the heterogeneity of the folded state (in the context of our model, molten globule), asserting that some contacts (in-nucleus) are less subject to fluctuations than other contacts (off-nucleus) (see Figure 8 and the discussion there). Corresponding, the nucleus contact interconversion rate is much slower, as is manifested in higher HD protection factors. Such heterogeneity in protection factors in the molten globule was indeed observed in a number of proteins (Hughson et al., 1990; Jeng et al., 1990). The explanation is simple: conformations with the nucleus correspond to the top of the barrier, the transition state. Therefore, fluctuations that go up to the barrier are most rare as they require higher energy. This makes the nucleus the most protected region in a molten globule.

(3) Our calculations predict a direct correspondence between the residues that are most protected from HD exchange in the equilibrium molten globule and the ones involved in folding the nucleus, i.e., *the first* stable set of contacts to be formed in the course of the folding process. This assertion is in accord with the experimental results for myoglobin (Jennings & Wright, 1993) and cytochrome C (Roder et al., 1988). The observation about the implications of mutations in nuclei on the folding rate makes this correspondence directly experimentally verifiable.

Our design–folding approach provided a possible conceptual framework to solve the protein folding problem. Within this approach, one can also address questions pertinent to the folding pathway of a specific protein, e.g., how to determine the folding nucleus in a given protein. To this end, it is necessary to take the native structure of this protein as a target conforamtion, design a sequence to fit the target conformation, and fold this sequence. This requires the incorporation of side chains into the lattice model, and the recent work by Skolnick and Kolinsky (1993) demonstrated the feasibility of such an endeavor. We are currently working along these lines.

## ACKNOWLEDGMENT

We thank Alexander Grosberg, Martin Karplus, Peter Leopold, Oleg Ptitsyn, and Andrej Sali for interesting and useful discussions. Graphic program ASGL by Andrej Sali was used to generate some of the plots.

## REFERENCES

Baum, J., Dobson, C. M., Evans, P. A., & Hanly, C. (1989) *Biochemistry 28*, 7–13.
Baumgartner, A. (1984) *Annu. Rev. Phys. Chem. 35*, 419–435.
Bryngelson, J. D., & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 7524–7528.
Bryngelson, J. D., & Wolynes, P. G. (1989) *J. Phys. Chem. 93*, 6902–6915.
Briggs, M., & Roder, H. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 2017–2021.
Bycroft, M., Matouschek, A., Kellis, A., Jr., Serrano, L., & Fersht, A. R. (1990) *Nature 346*, 488–490.
Camacho, C. J., & Thirumalai, D. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 6369–6372.
Chaffotte, A., Guillou, Y., Delepierre, M., Hinz, H.-J., & Goldberg M. (1991) *Biochemistry 30*, 8067.
Covell, D., & Jernigan, R. (1990) *Biochemistry 29*, 3287–3294.
Creighton, T. (1992) *Proteins. Structure and Molecular Properties*, W. H. Freeman & Co., New York.
Elove, G., Chaffotte, A., Roder, H., & Goldberg, M. (1992) *Biochemistry 31*, 6876–6883.
Finkelstein, A. V., & Shakhnovich, E. I. (1989) *Biopolymers 28*, 1668–1694.
Finkelstein, A. V., Gutin, A. M. & Badretdinov, A. Ya. (1993) *FEBS Lett. 325*, 23–28.
Gittis, A. G., Stites, W. E., & Lattman, E. E. (1993) *J. Mol. Biol. 232*, 718–724.
Goldstein, R., Luthey-Schulten, Z. A., & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 4918–4922.
Hilhorst, H. J., & Deutch, J. M. (1975) *J. Chem. Phys. 63*, 5153–5161.
Hughson, F., Wright, P., & Baldwin, R. (1990) *Science 249*, 1544–1548.
Jeng, M. F., Englander, W., Elove, G., Wand, A., & Roder, H. (1990) *Biochemistry 29*, 10433.
Jennings, P., & Wright, P. (1993) *Science 262*, 892–896.
Karplus, M., & Weaver, D. (1976) *Nature 160*, 404–406.
Karplus, M., & Shakhnovich, E. (1992) in *Protein Folding* (Creighton, T. E., Ed.) pp 127–195, W. H. Freeman and Company, New York.
Kim, P., & Baldwin, R. (1982) *Annu. Rev. Biochem. 51*, 459–489.
Kolinski, A., & Skolnick, J. (1993) *J. Chem. Phys. 98*, 7420–7433.
Lau, K. F., & Dill, K. A. (1990) *Macromolecules 22*, 3986–3997.
Leopold, P. E., Montal, M., & Onuchic, J. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 8721–8725.
Levinthal, C. (1969) in *Mossbauer Spectroscopy of Biological Systems. Proceedings of a Meeting Held at Allerton House, Monticello, IL* (Debrunner, P., Tsibris, J.-C., & Munck, E., Eds.) pp 22–24, University of Illinois Press, Urbana, IL.
Lifshitz, E. M., & Pitaevskii, L. P. (1981) *Physical Kinetics*, Pergamon, Oxford, U.K.
Lifshitz, I. M., Grosberg, A. Yu., & Khohlov, A. R. (1978) *Rev. Mod. Phys. 50*, 683–713.
Matouschek, A., Kellis, J., Jr., Serrano, L., Bycroft, M., & Fersht, A. R. (1990) *Nature 346*, 440–445.
Matouschek, A., Serrano, L., & Fersht, A. R. (1992) *J. Mol. Biol. 224*, 819–835.
Mault, J., & Unger, R. (1991) *Biochemistry 30*, 3816–3824.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., & Teller, E. (1953) *J. Chem. Phys. 21*, 1087–1092.

Miller, R., Danko, C., Fasolka, M. J., Balazs, A. C., Chan, H. S., & Dill, K. A. (1992) *J. Chem. Phys. 96*, 768–780.
Miranker, A., Radford, S., Karplus, M., & Dobson, C. (1991) *Nature 349*, 633–636.
Miyazawa, S., & Jernigan, R. (1985) *Macromolecules 18*, 534–552.
O'Toole, E. M., & Panagiotoupoulos, A. Z. (1992) *J. Chem. Phys. 97*, 8644–8645.
O'Toole, E. M., & Panagiotoupoulos, A. Z. (1993) *J. Chem. Phys. 93*, 3185–3190.
Peng, S., & Kim, P. (1993) *Biochemistry* (in press).
Privalov, P. L. (1979) *Adv. Protein Chem. 33*, 167–241.
Privalov, P. L. (1992) In *Protein Folding* (Creighton, T. E., Ed.) pp 127–195, W. H. Freeman and Company, New York.
Ptitsyn, O. B. (1987) *J. Protein Chem. 6*, 273–293.
Ptitsyn, O. B. (1992) in *Protein Folding* (Creighton, T. E., Ed.) Chapter 6, pp 243–300, W. H. Freeman and Company, New York.
Ptitsyn, O. B., Pain, R., Semisotnov, G., Zerovnik, E., & Razglyaev, O. (1990) *FEBS Lett. 262*, 20–24.
Radford, S., Dobson, C., & Evans, P. (1992) *Nature 358*, 302–307.
Rey, J., & Skolnick, J. (1991) *Chem. Phys. 158*, 199.
Rooman, M. J., & Wodak, S. J. (1992) *Biochemistry 31*, 10239–10249.
Rooman, M. J., Kocher, J.-P., & Wodak, S. J. (1992) *Biochemistry 31*, 10226–10238.
Sali, A., Shakhnovich, E. I., & Karplus, M. (1994a) *J. Mol. Biol. 3*, 1614–1636.
Sali, A., Shakhnovich, E. I. & Karplus, M. (1994b) *Nature 369*, 248–251.
Segawa, S., & Sugihara, M. (1984) *Biopolymers 23*, 2473–2488.
Shakhnovich, E. I. (1994a) *Phys. Rev. Lett. 72*, 3907–3910.
Shakhnovich, E. I. (1994b) in *Protein Structure by Distance Analysis* (Bohr, H., & Brunack, S., Eds.) IOS Press, Amsterdam.
Shakhnovich, E. I., & Finkelstein, A. V. (1982) *Dolk. Akad. Nauk SSSR 243*, 1247–1251.
Shakhnovich, E. I., & Finkelstein, A. V. (1989) *Biopolymers 28*, 1667–1681.
Shakhnovich, E. I., & Gutin, A. M. (1989a) *Biophys. Chem. 34*, 187–199.
Shakhnovich, E. I., & Gutin, A. M. (1989b) *J. Phys. A22*, 1647.
Shakhnovich, E. I., & Gutin, A. M. (1990a) *J. Chem. Phys. 93*, 5967–5971.
Shakhnovich, E. I., & Gutin, A. M. (1990b) *Nature 346*, 773–775.
Shakhnovich, E. I., & Gutin, A. M. (1993a) *Proc. Natl. Acad. Sci. U.S.A. 90*, 7195–7199.
Shakhnovich, E. I., & Gutin, A. M. (1993b) *Protein Eng. 6*, 793–800.
Shankhnovich, E. I., Farztdinov, G. M., Gutin, A. M., & Karplus, M. (1991) *Phys. Rev. Lett. 67*, 1665–1667.
Shin, H. C., Merutka, G., Waltho, J. P., Tennant, L., Dyson, H., & Wright, P. (1993) *Biochemistry 32*, 6356–6366.
Skolnick, J., & Kolinski, A. (1990a) *J. Mol. Biol. 212*, 787–817.
Skolnick, J., & Kolinski, A. (1990b) *Science 250*, 1121–1125.
Skolnick, J., & Kolinski, A. (1991) *J. Mol. Biol. 221*, 499–531.
Sykes, M. (1963) *J. Chem. Phys. 39*, 410.
Tsong, T. Y., Baldwin, R., & McPie, P. (1972) *J. Mol. Biol. 63*, 453.
Ueda, Y., Taketomi, H., & Go, N. (1978) *Biopolymers 17*, 1531–1548.
Uversky, V., Semisotnov, G., Pain, R., & Ptitsyn, O. (1992) *FEBS Lett. 314*, 89–92.
Verdier, P. H. (1973) *J. Chem. Phys. 59*, 6119–6126.
Waltho, J. P., Feher, V. A., Merutka, G., Dyson, H., & Wright, P. (1993) *Biochemistry 32*, 6337–6355.
Wetlaufer, D. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 697–701.
Williams, D., Harding, M., & Woolfson, D. (1991) *Biochemistry 30*, 3120–3128.
Wilson, C., & Doniach, S., (1989) *Proteins: Struct., Funct., Genet. 6*, 193–209.