A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed helical proteins

Andrzej Kolinski,[a] Adam Godzik, and Jeffrey Skolnick
Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, California 92037

(Received 6 November 1992; accepted 21 January 1993)

Starting from amino acid sequence alone, a general approach for simulating folding into the molten globule or rigid, native state depending on sequence is described. In particular, the 3D folds of two simple designed proteins have been predicted using a Monte Carlo folding algorithm. The model employs a very flexible hybrid lattice representation of the protein conformation, and fast lattice dynamics. A full rotamer library for side group conformations, and potentials of mean force of short and long range interactions have been extracted from the statistics of a high resolution set of nonhomologous, 3D structures of globular proteins. The simulated folding process starts from an arbitrary random conformation and relatively rapidly assembles a well defined four helix bundle. The very cooperative folding of the model systems is facilitated by the proper definition of the model protein hydrogen bond network, and multibody interactions of the side groups. The two sequences studied exhibit very different behavior. The first one, in excellent agreement with experiment, folds to a thermodynamically very stable four helix bundle that has all the properties postulated for the molten globule state. The second protein, having a more heterogeneous sequence, at lower temperature undergoes a transition from the molten globule state to the unique native state exhibiting a fixed pattern of side group packing. This marks the first time that the ability to predict a molten globule or a unique native state from sequence alone has been achieved. The implications for the general solution of the protein folding problem are briefly discussed.

I. INTRODUCTION

In the previous paper in this series,[1] we described a reduced, however, quite detailed representation of protein conformation, and employed knowledge based potentials of mean force derived from the statistics of a database of high resolution, 3D structures of globular proteins. Two highly idealized sequences, poly-*l*-alanine and poly-*l*-valine, were studied. It is known that alanine has a strong propensity to be in the α helical conformation, while valine prefers expanded states and is frequently seen in β sheets of globular proteins.[2] The Monte Carlo algorithm correctly predicted the expected class of 3D structures for both sequences. The simulated helix–coil transition of polyalanine has been found to be more cooperative than the coil–(β-type) globule transition of polyvaline. In the last case, it has been shown that compactness of the globular state itself is not sufficient to induce any appreciable amount of secondary structure. Rather the short range interactions, together with cooperativity of the hydrogen bond network, are responsible for the large increase in the secondary structure associated with the collapse transition. These test simulations demonstrated that the model reproduces the general features of long time protein dynamics and the folding–unfolding transition. However, the obtained 3D structures were not unique because homopolypeptide sequences were used. In the present work, the sequences of two small designed helical proteins[3,4] were used in the Monte Carlo simulations of the folding process. The only information specific for the folded proteins, employed as input for the Monte Carlo folding algorithm, was their amino acid sequence. Therefore the obtained 3D structures, and the folding pathways, are *de novo* predictions.

Most features of the present realization of the protein model are identical to those presented previously.[1] The similar geometrical representation of the protein conformation, a very similar model of dynamics (albeit, two new elemental moves have been added to the previous set), and a similar interaction scheme (with one, very important, extension) are used in this work. The major difference between the present formulation and our previous realization lies in the introduction of the cooperativity of the side group packing. This update is essential for modeling the late intermediates[5–7] of the protein folding process. It facilitates the transition from liquidlike side chain packing in the so-called molten globule state to the nativelike state[7–10] which possesses a well-defined pattern of the side group contacts within the unique fold of the main chain. This cooperativity[9] in the protein folding process will be discussed below in detail. Let us note here that for homopolypeptides, such as those studied in the previous work, due to lack of specificity in the long range interactions, inclusion of cooperativity of the side chain packing is not crucial. As we will see, even some heterosequences, possessing an ap-

[a]Also Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland.

 0021-9606/93/097420-14$06.00

parently too homogeneous hydrophobic interior, cannot undergo a cooperative transition from the molten globule to the nativelike state. On the other hand, it will also be shown that if a reduced representation of the protein conformation is used, the direct account for cooperativity in side chain fixation[9] is a necessary requirement to obtain unique folds exhibiting nativelike packing of the model protein.

The paper is organized as follows. In the next section (II), we discuss the protein model and the simulation method. Since a complete description appeared in the previous paper,[1] the geometrical representation and the Monte Carlo dynamics scheme are very briefly reviewed. However, the interaction scheme is discussed in detail due to improvements in the model. The potentials have been improved in a number of minor ways and a major extension accounts for the cooperativity of side chain packing. In Sec. III, the results of the simulations are presented. In particular, the folding pathways, including a discussion of early intermediates, are described. Most important to the process are the late intermediates, including the molten globule state. Questions concerning the nature of this intermediate and the character of the transition to the well-defined nativelike state are addressed. In Sec. IV, we discuss the properties of our model systems in the context of other theoretical work and recent experimental findings. Section V contains the summary of the present work. Possible extensions of these *de novo* predictions (starting from the sequence alone) of protein 3D structures are also briefly discussed.

II. METHOD

A. Representation of the protein conformation

A reduced representation[11] of the main chain backbone as well as the side groups is used in this study. The main chain backbone employs the α-carbon lattice representation. The α-carbon–α-carbon virtual bonds belong to the following set of 56 vectors: {[2,1,1],.[2,1,0],.[1,1,1]} in the metric of the underlying simple cubic (sc) lattice. The cubic lattice unit is assumed to be equivalent to 1.7 Å (Angstroms). Consequently, the Cα–Cα distances are allowed to fluctuate between 2.94 and 4.16 Å, with the average value of the virtual bond very close to 3.80 Å. Some sequences of two backbone vectors are excluded in order to keep the local geometry very close to that of real proteins. Too narrow valence angles for virtual bonds (below 78.5°) as well as too open ones (above 141.1°) are prohibited. Within these limitations, the lattice Cα trace can fit the geometry of globular protein x-ray structures with an average accuracy in the range of 1.0 Å root-mean-square (rms) deviation.[12]

The excluded volume of the main chain backbone is implemented so that no Cα can approach another (with the exception of the nearest neighbors along the chain) at a distance smaller than $8^{1/2}$ lattice units, which is equivalent to 4.81 Å. This distance is somewhat larger than the minimum distance seen in real proteins. On the other hand, taking into consideration the intrinsic resolution of the lattice and the large number of possible realizations of this minimal distance, the binary radial distribution of model Cαs quite closely mimics that seen in real proteins. One more reason for this similarity is that the model side chains have on average a somewhat too small repulsive core. What is more important, is that for several tested proteins, the volume of the model system at low temperature is very close to the corresponding real structures.

The side groups are modeled as single spheres. The pairwise contact distances have been derived from the statistics of a high resolution, nonhomologous, set of the 3D structures extracted from the Brookhaven Protein Data Bank (PDB).[13,14] A pair of side groups are counted as being in contact when any pair of their heavy atoms are at a mutual separation smaller than 4.2 Å. This criterion leads to a relatively narrow distribution of the center of mass distances between contacting side group pairs. When the two side groups are at a distance smaller than the average value plus one standard deviation from the above distribution, they are considered to be in contact. The numerical values of the contact distances, R_{ij}, and their standard deviations, ΔR_{ij}, for all pairs of side groups are given in Table I. The average distance minus one standard deviation was taken as the diameter for the soft repulsive cores of the spherical side groups. A schematic representation of the interaction regions of the model protein are shown in Fig. 1. This treatment of the side groups refines the previously used representation.

The location of the side chain center of mass with respect to the main chain backbone depends on the backbone conformation and on the rotational isomeric state of the side group (where applicable). For each set of two consecutive lattice backbone vectors, the discrete set of rotamers in the reduced representation has been derived based on the statistics of PDB structures. Subsets of the real rotamers, whose centers of mass were closer than 1.7 Å, were treated as a single (averaged) model rotamer. All the rotamers have their single sphere representations with off-lattice positions of their centers of mass. With these criteria, some amino acids always have just one rotamer for a given conformation of the main chain. For example, this is true for alanine, valine, serine, and of course glycine. In the case of glycine, the center of interaction is assumed to be at the same point as the model Cα position. For the other amino acids, the number of rotamers is larger and depends on the main chain conformation. For arginine, the number of rotamers varies from 11 to 13, as the main chain backbone fragment assumes different valence angles between the virtual bonds connecting the Cαs of interest.

B. Model of Monte Carlo dynamics

The set of elemental moves employed in the Monte Carlo dynamics contains the following micro modifications of the chain conformation. The move set includes the previously described two (virtual) bond spike moves, end moves, four bond moves (where a randomly selected four bond fragment is removed, and another one chosen from a large prefabricated library is inserted instead), and eight bond moves (combinations of the single Cα jump to the 26

TABLE I. Numerical values of the pairwise contact distances R and their standard deviations ΔR (even lines) in Å · 100.

	GLY	ALA	SER	CYS	VAL	THR	ILE	PRO	MET	ASP	ASN	LEU	LYS	GLU	GLN	ARG	HIS	PHE	TYR	TRP	CYX
GLY	392	394	393	424	466	433	462	429	472	410	425	471	457	450	435	467	438	448	435	455	452
GLY	22	18	37	17	44	51	51	34	47	41	58	41	65	44	55	85	51	66	75	80	24
ALA	394	389	403	438	453	442	465	437	455	422	422	465	471	443	442	470	436	449	443	460	407
ALA	18	19	37	15	47	45	56	41	53	43	49	40	60	54	41	64	48	64	68	91	50
SER	393	403	406	424	460	440	488	451	457	398	433	469	470	433	457	452	432	492	498	505	473
SER	37	37	56	60	50	56	47	60	84	43	41	45	69	52	56	69	54	63	67	71	44
CYS	424	438	424	322	477	449	480	445	445	491	500	478	590	482	465	486	464	479	467	612	0
CYS	17	15	60	99	55	44	50	11	66	44	39	48	47	65	31	82	48	86	70	52	0
VAL	466	453	460	477	508	504	524	503	521	492	485	529	517	470	494	516	497	506	517	517	498
VAL	44	47	50	55	63	62	64	60	65	48	52	52	61	77	50	79	63	78	79	95	46
THR	433	442	440	449	504	463	513	465	483	438	455	516	509	439	474	509	484	527	511	527	447
THR	51	45	56	44	62	79	66	50	77	55	52	50	69	57	66	82	54	71	75	80	60
ILE	462	465	488	480	524	513	547	535	524	504	510	544	545	501	527	526	483	540	530	527	512
ILE	51	56	47	50	64	66	69	52	63	73	62	57	77	68	61	74	83	76	78	101	57
PRO	429	437	451	445	503	465	535	489	493	475	489	505	485	490	480	521	454	469	467	483	464
PRO	34	41	60	11	60	50	52	49	41	56	43	56	89	48	58	69	46	64	71	92	28
MET	472	455	457	445	521	483	524	493	522	490	485	537	529	466	497	527	490	522	503	538	547
MET	47	53	84	66	65	77	63	41	60	75	72	63	68	70	67	81	58	68	79	77	63
ASP	410	422	398	491	492	438	504	475	490	472	451	496	462	455	465	463	482	514	524	505	442
ASP	41	43	43	44	48	55	73	56	75	66	53	61	60	66	59	74	50	67	72	77	53
ASN	425	422	433	500	485	455	510	489	485	451	476	496	465	465	464	494	477	500	502	514	467
ASN	58	49	41	39	52	52	62	43	72	53	47	58	73	57	59	96	59	83	80	86	67
LEU	471	465	469	478	529	516	544	505	537	496	496	536	522	501	509	510	488	528	523	513	501
LEU	41	40	45	48	52	50	57	56	63	61	58	53	53	52	63	81	69	73	80	91	44
LYS	457	471	470	590	517	509	545	485	529	462	465	522	522	583	488	519	476	477	485	479	538
LYS	65	60	69	47	61	69	77	89	68	60	73	53	86	68	71	98	78	75	91	93	76
GLU	450	443	433	482	470	439	501	490	466	455	465	501	583	501	501	487	501	500	518	522	444
GLU	44	54	52	65	77	57	68	48	70	66	57	52	68	60	44	71	56	74	66	85	45
GLN	435	442	457	465	494	474	527	480	497	465	464	509	488	501	499	483	492	493	492	519	462
GLN	55	41	56	31	50	66	61	58	67	59	59	63	71	44	68	82	56	69	76	92	42
ARG	467	470	452	486	516	509	526	521	527	463	494	510	519	487	483	515	476	464	510	454	519
ARG	85	64	69	82	79	82	74	69	81	74	96	81	98	71	82	96	77	94	89	98	77
HIS	438	436	432	464	497	484	483	454	490	482	477	488	476	501	492	476	511	473	486	488	508
HIS	51	48	54	48	63	54	83	46	58	50	59	69	78	56	56	77	63	75	99	99	59
PHE	448	449	492	479	506	527	540	469	522	514	500	528	477	500	493	464	473	533	551	563	548
PHE	66	64	63	86	78	71	76	64	68	67	83	73	75	74	69	94	75	82	78	89	47
TYR	435	443	498	467	517	511	530	467	503	524	502	523	485	518	492	510	486	551	528	553	490
TYR	75	68	67	70	79	75	78	71	79	72	80	80	91	66	76	89	99	78	83	108	101
TRP	455	460	505	612	517	527	527	483	538	505	514	513	479	522	519	454	488	563	553	567	478
TRP	80	91	71	52	95	80	101	92	77	77	86	91	93	85	92	98	99	89	108	72	83
CYX	452	407	473	0	498	447	512	464	547	442	467	501	538	444	462	519	508	548	490	478	291
CYX	24	50	44	0	46	60	57	28	63	53	67	44	76	45	42	77	59	47	101	83	48

nearest sc-lattice vertices, and two four-bond replacements of the adjacent segments).[1] In every case, there are the appropriate random modifications of the affected side chains. There are also separate rotamer modifications, which are the most frequently attempted moves.

Two new moves were implemented in the present version of the algorithm. The first is the ten bond move, where randomly selected bonds i and $i+9$ undergo correlated fluctuations (and/or rotations), and the intervening bonds translate by the corresponding small distance. Starting from a randomly selected point up to a terminal segment, a large fragment of the chain can move by a lattice unit. This kind of rigid body translation of the subchain, while rarely successful, may be helpful in surmounting some local energy maxima.

An arbitrary time unit is defined as that required for the range of N (where N is the number of residues in the model chain) rotamer modifications, N-2 spike moves, 2

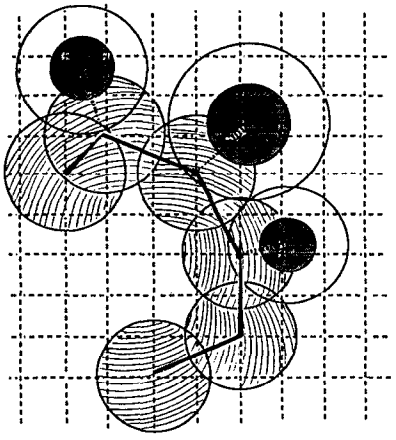

FIG. 1. Schematic illustration of the reduced representation of the protein used in this work. The shaded areas correspond to the excluded volume of the main chain and the strongly repulsive radii of the side groups. The open circles correspond to the contact radii of the side groups. The underlying cubic lattice is schematically shown by the dashed line grid.

end moves, N-4 four-bond moves, N-8 eight-bond moves, N-10 ten-bond moves, and one attempt at subchain, rigid body, translation.

The model of dynamics allows for the diffusive motion of assembled elements of secondary and supersecondary structure. This motion can occur by segmental diffusion, (via a succession of a large number of small, local rearrangements) or by rigid-body diffusion (via a succession of small distance jumps of large pieces of the model chain). Thus, presumably the model dynamics has no built-in bias towards a specific mechanism of protein assembly. As a matter of fact, the folding trajectories, while mostly following the "on site" mode[15] of structure assembly, also exhibit some elements of the collision-diffusion[16,17] mechanism, especially when very early folding events are observed.

C. Interaction scheme

All the energy values (dimensionless) are given in kT units, where k is Boltzmann's constant, and T is the temperature. The potential describing the short range interactions (i.e., those which are local down the chain) used in this work is exactly the same as described previously. Of course, the angular correlations between the pair of side-chain vectors are specific to the particular pair of amino acids. Consequently, the local side chain orientational coupling energy depends on the sequence, and reflects local propensities for a particular type of secondary structure.

$$E_{\text{local}}=\Sigma E_2(\phi_{i,i+2},a_i,a_{i+2})+\Sigma E_3(\phi_{i,i+3},a_i,a_{i+3}) + \Sigma E_4(\phi_{i,i+4},a_i,a_{i+4}), \quad (1)$$

where a_i denotes the amino acid type at position i down the sequence, and $\phi_{i,k}$ is the angle between the reduced representation, side group vectors (the vectors from Cα to the center of the side group) i and k. The potentials E_2, E_3, and E_4 have been derived from the statistics in the structure database (see Ref. 1 for the list of proteins). The numerical values are determined as the negative logarithm of the frequency of occurrence of an angle in a particular range (ten intervals were used for the discretization) with respect to the uniform, unbiased distribution. An example of the numerical values for two pairs (alanine–alanine, and valine–valine) has been given in Ref. 1. Since it is not practical to present here a table or histogram for each pair of amino acids, these supplementary materials are available by anonymous ftp [ftp.scripps.edu].

Additionally, there is an amino acid nonspecific, lattice-correction term of the short range interactions which biases the distribution of the distances between the ith and $i+3$th Cαs towards the bimodal distribution seen in real proteins. The system experiences a small energy decrease for all the local conformations of the backbone which are compact (turnlike) or very expanded (presumably β-strand type), regardless of the sequence.

The energy of hydrogen bonds is calculated as follows:

$$E_{h\text{-bond}}=\sum\sum E^{\text{H}}\delta(i,j)+\sum\sum E^{\text{HH}}\delta(i,j,i\pm1,j\pm1) \quad (2)$$

where E^{H} is the strength of the model hydrogen bond and $\delta(i,j)=1$, when the following criteria (3) and (4) are satisfied, and otherwise it is zero:

$$|(b_i-b_{i+1})\cdot r_{ij}|<6; \quad r_{ij}\leqslant\sqrt{17} \quad \text{and} \quad |i-j|\geqslant4, \quad (3)$$

$$|(b_j-b_{j+1})\cdot r_{ij}|<6; \quad r_{ij}\leqslant\sqrt{17} \quad \text{and} \quad |i-j|\geqslant4. \quad (4)$$

No account is taken of the donor–acceptor directionality of the H bonds. However, there can only be up to two H bonds connecting every amino acid (except proline, which can participate in only one hydrogen bond). There are no explicit hydrogen bonds with side groups. E^{HH} is the cooperative contribution to the energy of the hydrogen bond network. $\delta(i,j,k,l)$ is equal to 1 when amino acids i and j are hydrogen bonded to each other when amino acids k and l are also hydrogen bonded. Otherwise, it assumes the value 0. The summation in the Eq. (2) is performed so that each hydrogen bond and each cooperativity contribution (pair of hydrogen bonds) is counted once. The above definition of the model hydrogen bond reproduces about 90% of the backbone hydrogen bonds deduced by the Kabsch–Sander method[18] for the 3D structures of globular proteins. The estimation is done by first fitting the lattice Cα trace to the PDB structures, then running the simulation program with a fixed backbone, and finally, comparing the resulting model network of hydrogen bonds with that calculated by the DSSP program.[18]

Several terms contribute to the energy of the long range interactions of the model system. First, there is the one body term, which is amino acid specific and is only a function of the distance of the center of mass of the side group from the center of mass of the protein under consideration. The approximately spherical shape of globular proteins, and the very small contribution of this term to the total energy of the model system permits the spherical approximation to be used. This term reflects the tendency of some amino acids to be buried inside the globule, while others prefer to be exposed on the surface. The numerical values of this potential are given in Table II. The implementation of the one body term in the simulation algorithm requires the *a priori* estimation of the radius of gyration of the globular state. This is relatively easy to do, and a quite accurate value for most proteins can be obtained from

$$s=2.2\exp(0.38\ln N) \text{ in Å}. \quad (5)$$

The two body interactions of the side groups are also derived from the statistics of the PDB database. The values of the contact energy, ε_{ij} (the negative logarithm of the observed frequency of the particular pair relative to the random frequency of the corresponding pair calculated employing the Bragg–Williams approximation[19]), are given in Table III. The pairwise energy of the model system is then counted as follows:

$$E_{ij}=\begin{cases}E_{\text{rep}}; & \text{for} \quad d_{ij}<R_{ij}-\Delta R_{ij}\\ \varepsilon_{ij}; & \text{for} \quad R_{ij}-\Delta R_{ij}<d_{ij}<R_{ij} \text{ and } \varepsilon_{ij}>0,\\ \varepsilon_{ij}\cdot f/\{(R_{ij}-d_{ij})^6+1\} & \text{for} \quad \varepsilon_{ij}<0\end{cases} \quad (6)$$

TABLE II. The numerical values of statistical centrosymmetric one-body potential for various side groups.

	$k=2^a$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
GLY	−0.86	−0.18	−0.27	−0.07	−0.08	−0.26	−0.08	0.17	0.16	0.19	0.05	0.00	0.06	−0.18	−0.30	−0.11	0.00
ALA	0.00	−0.08	0.15	−0.16	−0.10	0.04	−0.01	0.05	0.00	0.03	−0.05	0.01	0.04	0.12	−0.14	0.06	−0.23
SER	0.00	0.28	−0.02	−0.03	0.12	0.18	0.32	0.12	0.03	−0.02	−0.11	−0.08	−0.33	−0.21	−0.13	−0.41	0.00
CYS	0.00	0.00	−0.71	−0.50	−0.08	−0.59	−0.17	−0.24	0.20	0.52	0.06	0.07	0.81	0.00	0.00	0.00	0.00
VAL	−0.35	−0.34	−0.50	−0.40	−0.48	−0.33	−0.36	−0.14	0.00	0.19	0.32	0.54	0.81	0.52	1.02	0.90	0.00
THR	−0.87	0.19	−0.06	0.00	0.05	0.39	−0.12	0.04	0.06	0.01	−0.02	−0.18	−0.08	0.08	0.04	−0.13	0.00
ILE	0.00	−0.72	−0.48	−0.52	−0.54	−0.54	−0.39	−0.23	−0.05	0.41	0.56	0.42	1.08	2.20	1.40	1.49	0.00
PRO	0.00	0.49	0.89	0.61	0.51	0.32	0.02	0.10	0.01	−0.17	−0.02	−0.24	−0.28	−0.28	0.13	−0.55	−0.84
MET	0.00	0.00	−0.28	−0.47	−0.31	−0.39	−0.48	0.00	0.06	0.11	0.28	0.45	0.43	0.30	0.00	0.00	0.00
ASP	0.00	0.19	0.86	0.39	0.72	0.58	0.50	0.35	0.13	−0.17	−0.22	−0.26	−0.37	−0.68	−0.79	−0.43	−0.72
ASN	0.00	−0.13	0.42	0.41	0.19	0.58	0.60	−0.08	−0.11	−0.02	−0.32	−0.34	0.15	0.12	−0.15	−0.53	0.00
LEU	0.00	−0.18	0.00	−0.32	−0.57	−0.45	−0.34	−0.11	−0.08	0.01	0.43	0.58	1.02	1.06	0.89	0.98	0.00
LYS	0.00	2.24	1.65	1.93	1.92	1.18	1.09	0.31	−0.05	−0.25	−0.24	−0.57	−0.73	−0.49	−0.46	−0.80	−0.33
GLU	0.00	0.50	1.30	1.33	1.11	1.01	0.52	0.29	0.12	−0.35	−0.31	−0.16	−0.57	−0.65	−0.71	−0.31	−0.61
GLN	0.00	0.00	0.74	1.05	0.61	0.44	0.56	−0.05	0.07	−0.23	−0.10	−0.32	−0.45	−0.30	−0.61	0.00	0.00
ARG	0.00	0.00	−0.12	0.62	0.95	0.44	0.22	−0.01	−0.23	−0.09	−0.25	−0.07	−0.26	0.25	0.06	−0.23	0.00
HIS	0.00	−0.75	−0.20	−0.28	−0.35	0.14	0.12	0.00	−0.01	0.07	0.02	0.15	0.04	−0.33	0.00	0.00	0.00
PHE	0.00	−0.73	−0.88	−0.73	−0.46	−0.61	−0.26	−0.27	−0.04	0.44	0.52	1.20	1.29	1.61	1.10	0.00	0.00
TYR	0.00	0.00	0.16	0.08	0.48	−0.12	−0.27	−0.44	−0.15	−0.02	−0.08	0.35	0.73	0.16	2.29	0.00	0.00
TRP	0.00	0.00	−0.29	0.39	−0.44	−0.30	−0.54	−0.36	−0.14	0.27	0.25	0.39	1.75	0.34	0.00	0.00	0.00

[a] $k=i$ indicates $0.1 \cdot i \cdot S < r < 0.1 \cdot (i+1) \cdot S$, with S the expected radius of gyration of the native state, and r is the distance of the center of the side group from the protein center of mass.

where E_{rep} is the repulsive force, d_{ij} is the distance between the two side groups, and f is the angular factor, equal to the square of the cosine between the backbone vectors, connecting α-carbons i-2 with $i+2$, and j-2 with $j+2$, respectively. This factor reflects the tendency of the elements of secondary structures seen in real proteins to be close to parallel. Thus there is a small change in comparison with the previous version of the model; now, the side chain interactions are now coupled to the orientation of the corresponding fragment of the backbone instead of the relative orientation of the side groups.

The major improvement in the tertiary interaction potential comes from the four-body interactions, which introduces cooperative coupling of the pairwise interactions. It is known from experiment that the transition from the molten globule to the native state is very cooperative and involves the fixation of side chain–side chain contacts. Therefore, to reproduce this essential feature of protein folding, we have introduced a cooperative term into the side chain–side chain interactions within the framework of the reduced representation of the protein conformation. On a very phenomenological level, the inclusion of such terms

TABLE III. Potential of pairwise interactions for side groups (including backbone, and glycine for which the interactions are centered on the Cα) in kT units.

	GLY	ALA	SER	CYS	VAL	THR	ILE	PRO	MET	ASP	ASN	LEU	LYS	GLU	GLN	ARG	HIS	PHE	TYR	TRP	CYX
GLY	0.3	0.1	0.0	0.1	0.2	−0.0	0.3	0.4	0.4	−0.4	−0.1	0.5	1.2	0.6	−0.1	0.5	0.3	0.1	−0.4	−0.5	−0.3
ALA	0.1	−0.1	−0.1	0.1	−0.6	−0.3	−0.6	0.3	0.1	0.1	0.1	−0.4	1.0	0.5	−0.0	0.5	−0.1	−0.6	−0.5	−0.7	−0.3
SER	0.0	−0.1	−0.6	−0.4	0.3	−0.5	0.4	0.4	−0.0	−0.9	−0.3	0.4	0.5	−0.2	0.0	0.0	−0.6	−0.0	−0.1	−0.1	−0.4
CYS	0.1	0.1	−0.4	−3.3	−0.8	0.0	−1.1	−0.0	−1.8	0.3	0.3	−0.5	1.6	0.8	0.1	0.4	−1.2	−1.5	−0.1	−0.5	2.0
VAL	0.2	−0.6	0.3	−0.8	−0.9	0.2	−0.7	0.1	−0.5	0.7	0.2	−0.6	0.9	0.4	0.2	0.3	−0.2	−0.8	−0.3	−0.8	−0.1
THR	−0.0	−0.3	−0.5	0.0	0.2	−0.3	0.0	0.3	0.2	−0.6	−0.3	0.3	0.6	−0.2	−0.2	0.1	−0.6	−0.0	−0.2	0.0	−0.6
ILE	0.3	−0.6	0.4	−1.1	−0.7	0.0	−0.8	0.3	−0.7	0.6	0.7	−0.6	0.8	0.4	0.3	0.4	−0.2	−0.8	−0.5	−0.9	−0.4
PRO	0.4	0.3	0.4	−0.0	0.1	0.3	0.3	0.1	−0.2	0.6	0.4	0.5	1.1	0.6	−0.1	0.5	−0.5	−0.2	−0.9	−0.7	0.1
MET	0.4	0.1	−0.0	−1.8	−0.5	0.2	−0.7	−0.2	−1.1	1.0	0.2	−0.6	0.5	0.0	0.2	0.1	−0.9	−1.1	−0.6	−1.3	−0.3
ASP	−0.4	0.1	−0.9	0.3	0.7	−0.6	0.6	0.6	1.0	0.3	−0.6	0.8	−0.6	0.3	0.0	−1.1	−1.0	0.4	−0.3	−0.2	−0.2
ASN	−0.1	0.1	−0.3	0.3	0.2	−0.3	0.7	0.4	0.2	−0.6	−0.4	0.5	0.1	−0.4	−0.4	0.0	−0.5	0.2	−0.5	−0.1	0.1
LEU	0.5	−0.4	0.4	−0.5	−0.6	0.3	−0.6	0.5	−0.6	0.8	0.5	−0.6	1.1	0.7	0.4	0.2	−0.3	−0.9	−0.1	−0.8	−0.4
LYS	1.2	1.0	0.5	1.6	0.9	0.6	0.8	1.1	0.5	−0.6	0.1	1.1	1.9	−0.7	0.4	1.7	0.3	0.3	−0.2	0.1	1.3
GLU	0.6	0.5	−0.2	0.8	0.4	−0.2	0.4	0.6	0.0	0.3	−0.4	0.7	−0.7	0.6	0.2	−0.9	−0.9	0.1	−0.0	0.1	0.6
GLM	−0.1	−0.0	0.0	0.1	0.2	−0.2	0.3	−0.1	0.2	0.0	−0.4	0.4	0.4	0.2	0.0	−0.5	0.9	0.0	−0.5	−0.2	−0.6
ARG	0.5	0.5	0.0	0.4	0.3	0.1	0.4	0.5	0.1	−1.1	0.0	0.2	1.7	−0.9	−0.5	0.2	−0.4	−0.4	−0.4	−0.9	0.5
HIS	0.3	−0.1	−0.6	−1.2	−0.2	−0.6	−0.2	−0.5	−0.9	−1.0	−0.5	−0.3	0.3	−0.9	0.9	−0.4	−1.4	−1.0	−0.8	−1.2	0.3
PHE	0.1	−0.6	−0.0	−1.5	−0.8	−0.0	−0.8	−0.2	−1.1	0.4	0.2	−0.9	0.3	0.1	0.0	−0.4	−1.0	−1.5	−0.5	−1.3	−0.1
TYR	−0.4	−0.5	−0.1	−0.1	−0.3	−0.2	−0.5	−0.9	−0.6	−0.3	−0.5	−0.1	−0.2	−0.0	−0.5	−0.4	−0.8	−0.5	−0.8	−0.5	−0.9
TRP	−0.5	−0.7	−0.1	−0.5	−0.8	0.0	−0.9	−0.7	−1.3	−0.2	−0.1	−0.8	0.1	0.1	−0.2	−0.9	−1.2	−1.3	−0.5	−0.8	−1.2
CYX	−0.3	−0.3	−0.4	2.0	−0.1	−0.6	−0.4	0.1	−0.3	−0.2	0.1	−0.4	1.3	0.6	−0.6	0.5	0.3	−0.1	−0.9	−1.2	−5.3

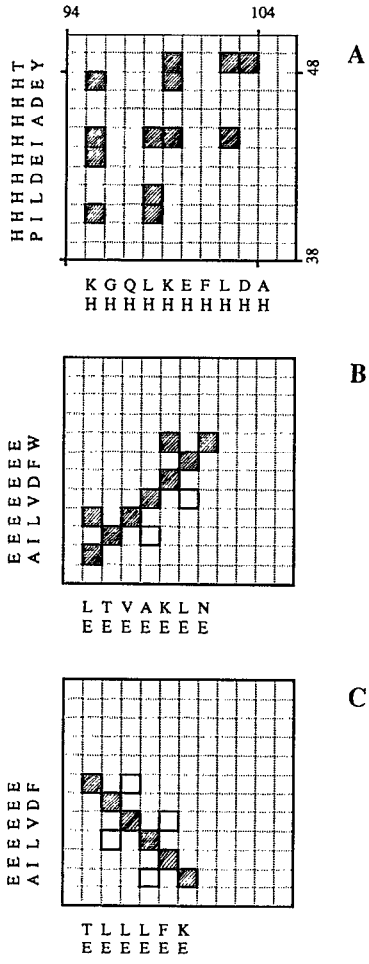

FIG. 2. Representative side chain–side chain contact patterns extracted from globular proteins for (A) helices, and (B) parallel, and (C) antiparallel β sheets are indicated by open squares. The shaded squares indicate that portion of the contacts experiencing cooperative interactions according to Eq. (7).

can be rationalized as follows. On analyzing side group contact maps of real native proteins, there are many regularities evident. These regularities in the contact maps reflect the more or less regular, solidlike packing of protein structures. Indeed, the various secondary structure motifs exhibit characteristic contact map patterns. Figure 2 shows representative side group contact patterns extracted from globular proteins for parallel helices, and parallel and anti parallel β sheets. Furthermore, these patterns are rather insensitive to the exact details of how the contacts are defined. For example, when two parallel helices are in contact, there is a very typical repeating pattern of side group contacts [see Fig. 2(A)]. Provided that residues i and j are in contact, it is very likely that residues $i+4$ and $j+4$ are also in contact. In addition, the contact between residues i-3 and $j+3$ is very frequently observed, as well as contacts between residues $i+3$ and $j+4$. For two parallel β strands [Fig. 2(B)], the contact between residues i and j makes the contact between residues $i+k$ and $j+k$ (with $k=1,2,3,4$) more probable. For antiparallel motifs, [Fig. 2(C)] corresponding correlations exist. Several templates representing a cooperative contribution to the pairwise interactions of the side groups have been examined in order to find the ones which have no bias towards a particular type of secondary structure. It appears that coupling the contacts i,j with $i\pm3$, $j\pm3$ contacts, and with contacts of the type $i\pm4$, $j\pm4$ generates proteinlike packing patterns for helical proteins as well as for β-type proteins, within the framework of the reduced representation of the side groups used here. The stability of α/β barrels also seems to be quite enhanced when the above templates are used. The shaded squares in Fig. 2 are the portions of the patterns experiencing enhanced cooperativity resulting from i,j and $i\pm3$, $j\pm3$, and $i\pm4$, $j\pm4$ coupling. Clearly, the contact templates shown by the shaded squares in Fig. 2, permit both helical and β-sheet nativelike packing to be enhanced. Thus the tertiary interaction energy of the model systems, including the cooperative contribution is as follows:

$$E_{\text{ter}}=\sum_i\sum_j\left\{E_{ij}+\sum_k\sum_n(\varepsilon_{ij}+\varepsilon_{i+k,j+n})\cdot C_{ij}C_{i+k,j+n}\right\};\quad |k|=|n| \tag{7}$$

with $C_{i,j}=1$ if the side groups of residues i and j are in contact; otherwise, $C_{i,j}=0$. k and $n=\pm3$, and k and $n=\pm4$. E_{ij} is defined in Eq. (6).

The last term [Eq. (7)] of the interaction scheme of the present model effectively reduces the number of possibilities of the side chain packing, thereby enhancing proteinlike patterns. Presumably, with a proper sequence, the model system should be able to assume a well defined, and unique, packing in the globular state which is very similar to that in real proteins. Moreover, it should be possible to achieve this transition in a relatively short simulation time. Indeed, in our simulations, the transition from the molten globule to the nativelike state is probably speeded-up several times in comparison with earlier folding events. Nevertheless, the side group fixation process remains the most time consuming part of the simulations, reflecting the very long time required for this late stage process seen in the folding of real proteins.[9]

The problem with all statistically derived potentials of mean force[11] is that some degrees of freedom are usually missed when computing the partition function of the reference state. The present model uses several potential functions of the mean force derived this way, and some built "by hand" potentials (cooperative terms). Consequently, the simulation algorithm scales the various energy contributions by constant factors. These factors were tuned in order to obtain a "reasonable" level of secondary structure in denatured state, and to produce stable, nativelike structures of a few, small test proteins. In order to avoid all the irrelevant details, the thermodynamics of the folding transition will be characterized by the relative contributions of the various energy terms (in kT units) of the model system.

III. RESULTS

Two designed proteins were studied. Their sequences[3,20] are the following:

SEQI = (GELEELLKKLELLKG PRR)$_3$ GELEELLKKLELLKG.

SEQII = GEVEELLKKFKELWKGPRRGEIEELFKKFKELIKGPRRGEVEELLKKFKELWKGPRRGEIEELF KKFKELIKG.

The first one has been designed to adopt the four helix bundle topology.[3] The second is an improved version, with a more heterogeneous amino acid composition.[4] For the first sequence, we can compare our simulation results with substantial experimental studies.[3] The simulations of the second protein are mostly in the character of theoretical predictions, which have yet to be confirmed by real experiments.

A. Monte Carlo folding experiments

First, a set of random conformations of the model polypeptides was generated. These random coils correspond to the denatured state of the model protein. The random coil state of the model polypeptides has a mean radius of gyration 1.5–3 times larger than that of the low temperature states, and the helix content of the random coil is in the range of 10%–15%. Starting from these random states, the Monte Carlo algorithm simulates the temperature driven renaturation. In most runs, the system temperature changed by about 5%–10% during annealing. 24 independent folding experiments were performed for SEQI, and 14 for SEQII. In all cases, the system adopted the four helix bundle topology. Representative snapshots of the folding pathway for SEQII are shown in Fig. 3. In almost all cases, folding started from assembly of an α-helical hairpin, with a somewhat larger probability for the central hairpin relative to the two other possibilities. This hairpin, being a very early intermediate of the folding process, usually dissolves many times before a more pronounced intermediate form. This, still rather early, intermediate consists of three helices which are usually in almost correct registration. "On site" assembly of the fourth helix completes the bundle.

There are two possible topologies of four helix bundle with tight loops, the right turning and the left turning one (see schematic drawing in Fig. 4). The model systems adopt both topologies. The result of 11 of 24 runs for SEQI was the right turning bundle, while for SEQII, both topologies were obtained 7 times. For each sequence, the left and right turning bundles can be separated and used to extract a set of distance constraints, which subsequently can be used to generate a "consensus" fold for both topologies. Table IV presents the statistics of the obtained folds, including the root-mean-square (rms) separation (measured between Cα traces alone) between the particular folds belonging to the same class.

Of course, the fact that the two different topologies for the same sequence were obtained necessitated further investigation. Three possible interpretations of this result have to be considered. First, perhaps these sequences (or perhaps one of them) can really adopt two isoenergetic topologies. Second, the annealing procedure "quenches" the system too fast, and consequently the model system falls into the global free energy minimum as well as into a

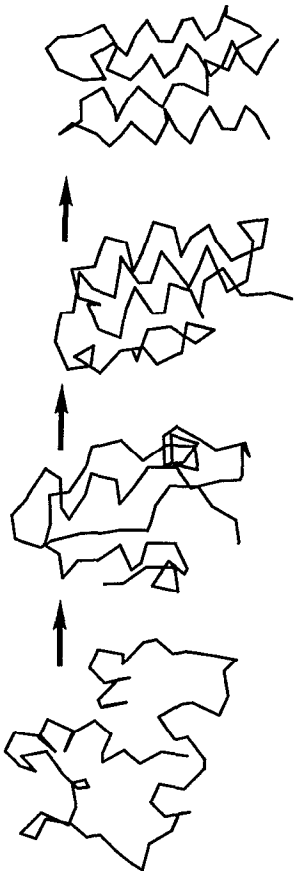

FIG. 3. Representative folding pathway of the SEQII. Only the α-carbon traces are shown for clarity.

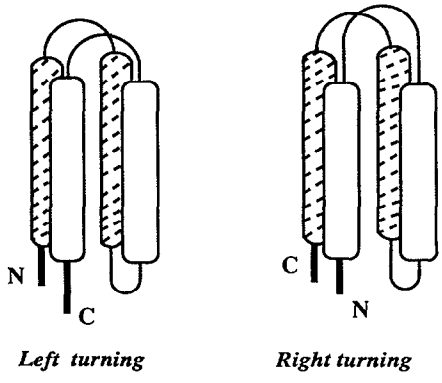

FIG. 4. Two different topologies of four helix bundles.

TABLE IV. Statistics of the Monte Carlo folding experiments.

Sequence	Type of fold	Number of folds	Average rms from consensus structure	Average radius of gyration	Average energy[b] (kT)
SEQI	Right turning	11	4.1 Å (3.3)	11.6 Å	−320.1
SEQI	Left turning	13	4.2 Å (3.6)	11.6 Å	−321.7
SEQII	Right turning	7	3.6 Å (2.9)[a]	10.6 Å	−422.3
SEQII	Left turning	7	3.7 Å (3.5)	10.7 Å	−407.6

[a]The two most distant structures converge after 120 000 MC steps to 2.3 Å rms. The numbers in parenthesis correspond to rms after the long isothermal runs.
[b]Energy after long isothermal runs (40 000–80 000 MC steps).

very deep local minimum, corresponding to the alternative topology. Finally, one has to consider if the model is at all able to differentiate between the various topologies due to the reduced representation of the polypeptides and other simplifications. It will be shown that the last possibility can be quite safely eliminated, using several methods to assess a proper fold in each case (more precisely, it has been proven to be possible for simple folds; the problem of more complicated topologies needs further investigation).

B. Analysis of the SEQI folds

Having consensus folds, long, isothermal runs have been performed at a temperature just below the folding temperature. Every run consisted of 40 000 to 80 000 Monte Carlo steps, every step involves on the range of N (the number of residues in the protein chain which equals 73), moves of every type, described in the previous section. The folded structure remains topologically stable over the entire run. However, there are substantial fluctuations of the configurational energy and the radius of gyration of the structure. Both the left turning and right turning bundles behave this way. Moreover, the average energy of both folds is essentially the same. Numerical values for SEQI, and SEQII are compared in Table V. Representative flow charts for the SEQI radius of gyration, configurational energy, and helix content are given in Fig. 5. The range of energy and conformational fluctuations[8,9] are very typical of the so-called "molten globule state" of globular proteins. The molten globule, being a putative late intermediate of the folding process, is often characterized as a structure having the general topology of the final fold, essentially the same (or very similar) secondary structure, whose size is swollen by 10%–20% with respect to the native state, and poorly defined tertiary interactions. The last observation means that the hydrophobic core of the globule is formed; however, the side chains are not fixed,[8] and the typical native pattern of the interresidue contacts is not yet developed. In other words, the protein interior behaves like a liquid. This liquidlike, mobile structure is probably the most pronounced property of the molten globule state, with several experimental consequences. In order to address this issue more directly, we analyzed the lifetime of the side group–side group contacts. In Fig. 6, the average contact map for the right turning bundle is shown in the upper part of the plot, where those contacts which are observed in at least 50% of the trajectory are marked by black squares. The lower part of the diagram shows the first dissolution time of these contacts. It is easy to see that none of the contacts are completely fixed, and very few of them survive for even 10 000 MC steps. These longer living contacts do not form any well defined pattern or network across the bundle. There is one more computational experiment which proves that this model protein behaves as molten globule. Suppose one fixes the main chain backbone, randomizes the side group rotamers, and then allows the system to relax. This relaxation process for SEQI is extremely fast (a few MC steps) and does not involve any specific rearrangement of the entire structure. The relaxation process is rather similar to the fluctuations seen in the long isothermal runs.

TABLE V. Energy distribution of various conformations.[a]

Sequence	Conformation	Average total energy	Local side chain coupling energy	Hydrogen bond energy	Side chain rotamer energy	Lattice correction energy	Long range tertiary energy	Side chain contact template energy
SEQI	Right turning bundle	−320.1	−73.2 (22.9%)	−59.5 (18.6%)	−25.7 (8.0%)	−33.1 (10.3%)	−82.4 (25.7%)	−46.3 (14.4%)
SEQI	Left turning bundle	−321.7	−63.6 (19.8%)	−57.1 (17.7%)	−26.5 (8.2%)	−31.2 (9.7%)	−87.5 (27.2%)	−55.9 (17.4%)
SEQI	Denatured state	−125.7	−33.6 (26.7%)	−22.5 (17.9%)	−26.4 (21.0%)	−18.7 (14.9%)	−22.4 (17.8%)	−2.1 (1.7%)
SEQII	Right turning	−422.3	−55.8 (13.2%)	−66.0 (15.6%)	−17.0 (4.0%)	−31.5 (7.5%)	−138.2 (32.7%)	−113.8 (27.0%)
SEQII	Left turning	−407.6	−54.5 (13.4%)	−57.0 (14.0%)	−20.5 (5.0%)	−30.9 (7.6%)	−135.9 (33.4%)	−108.9 (26.7%)
SEQII	Denatured state	−151.5	−23.8 (15.7%)	−17.0 (11.2%)	−16.7 (11.0%)	−19.1 (12.6%)	−47.7 (31.5%)	−27.3 (18.0%)

[a]Energy is in units of kT. The numbers in parentheses indicate the relative contribution to the total energy. See the text for additional details.

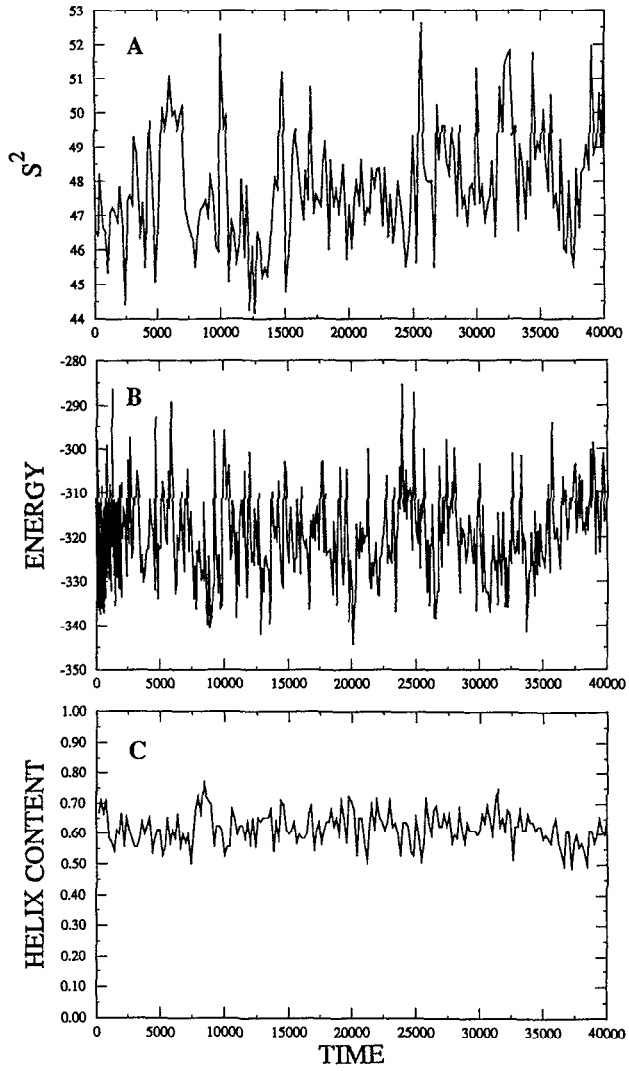

FIG. 5. Flow charts of the radius of gyration (A), the configurational energy (B), and the helix content (C), for long isothermal simulations of the right turning bundle of SEQI.

The nature of this molten globular structure of the SEQI folds and the rather remarkable agreement with recent experiments will be discussed in more detail below. Now, let us analyze the strikingly different behavior of the SEQII folds.

C. Analysis of the SEQII folds

This sequence has been designed by introducing a number of amino acid substitutions in SEQI. The major objective was to replace the very uniform leucine core of the expected bundle by a more heterogeneous set of hydrophobic residues. As one may see from inspection of the data from Tables IV and V, there are several differences between the models of both proteins. First, based on the numerical value of the radius of gyration, the packing density of the SEQII folds seems to be larger. The radius of gyration for these folds is smaller, in spite of the fact that there are several large side chains (PHE,TRP) incorporated in the new sequence. Then, the configurational energy (at the same temperature) of the folded structures is considerably smaller in the case of SEQII. What is worth noticing is that there is a small but highly reproducible difference between the energy of the right handed and the left handed bundle. The right handed fold is about 15 kT lower in energy.

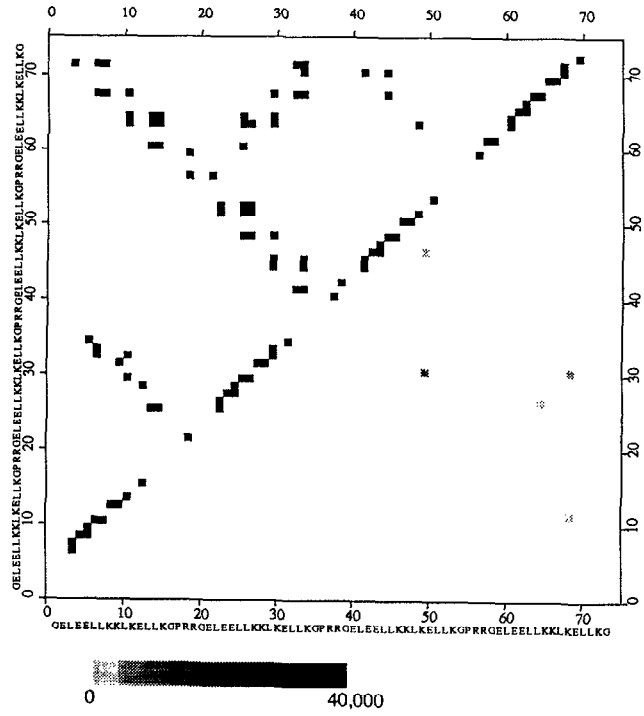

FIG. 6. The contact map for the SEQI right turning bundle (in the upper left corner of the diagram), and the illustration of the first dissolution time of the contacts (in the lower right corner of the diagram).

The consensus structures for the right vs the left turning bundle exhibit a smaller rms deviation, suggesting that the right handed fold is better defined. Moreover, when the two most distant right turning folds (3.7 Å rms) obtained from the annealing procedure are then subjected to isothermal runs, they converge to structures with an rms deviation equal to 2.3 Å, with *almost the same* pattern of the side group contacts. Namely, 71 contacts are exactly the same out of 93 long distance contacts ($|i-j|>5$) seen in one of the folds and 95 in the other. This is the level of reproducibility of a contact map of a protein which has been crystallized in different conditions, or for contact maps of very close homologues. The above strongly suggests that the right turning fold of SEQII is unique and resembles the folds of natural globular proteins.

Now let us consider long isothermal simulation runs for the folded structures. These runs start from randomized (liquidlike) configurations of the side chains. First, let us note that the helix content is approximately the same for both sequences, and beside small fluctuations, it does not change during these simulations [See Fig. 5(C)]. There is clearly a fast side chain reorientation process, where they roughly adjust to the local environment. This is then followed by a much longer relaxation involving the entire structure, which adopts a denser, lower energy state [see

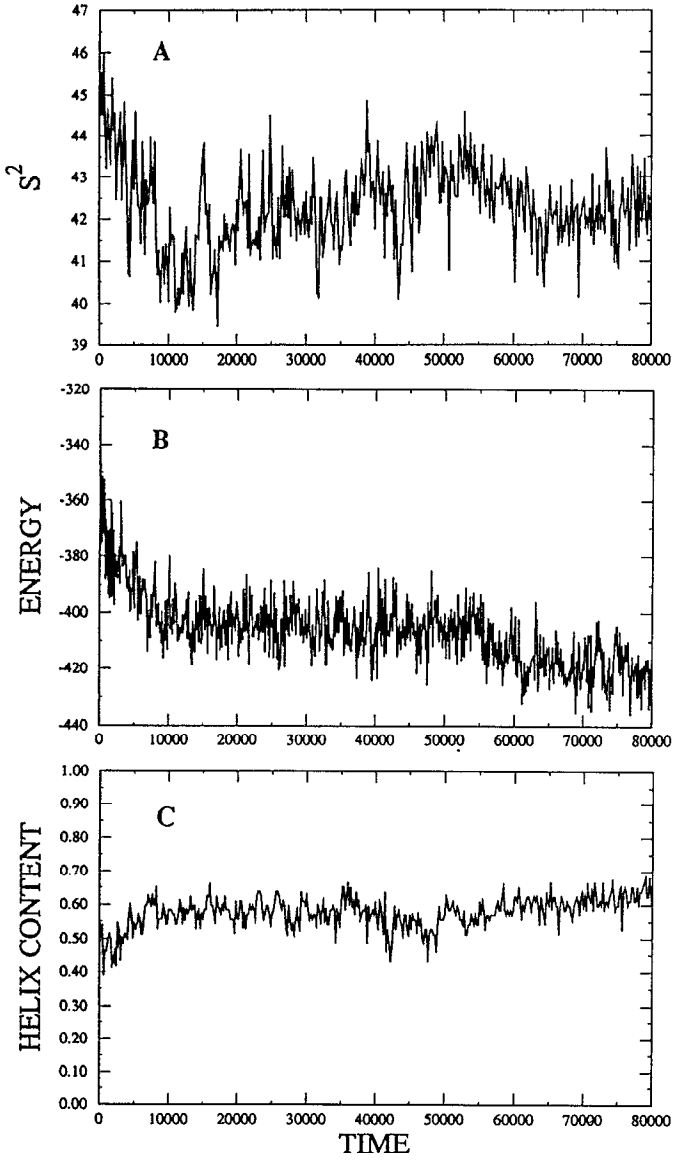

FIG. 7. Flow charts of the radius of gyration (A), the configurational energy (B), and the helix content (C), for long isothermal simulations of the right turning bundle of SEQII.

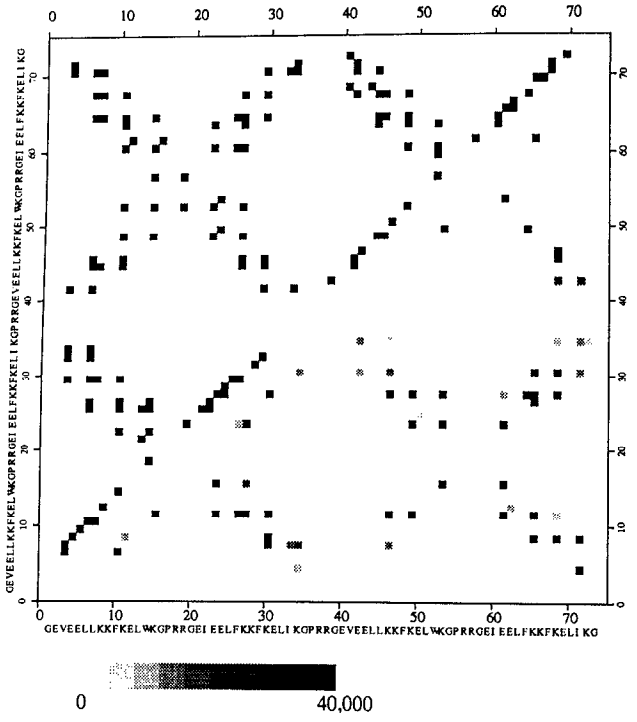

FIG. 8. The contact map for the SEQII right turning bundle (in the upper left corner of the diagram), and the illustration of the first dissolution time of the contacts (in the lower right corner of the diagram).

Fig. 7(A)]. The length of the last relaxation process indicates a very collective rearrangement suggestive of the molten globule to nativelike state transition. The fluctuations of all the measured properties at equilibrium (compare Fig. 7 with Fig. 5) are much smaller than those seen for the SEQI sequence. In the upper part of the diagram shown in Fig. 8, the contact map for the right handed fold of SEQII is shown using exactly the same convention as the one presented in Fig. 6 for SEQI. The lower part of the plot shows the first dissolution time of the contacts. In contrast to the situation seen for SEQI, a substantial fraction of the contacts of the right handed fold of SEQII are present during the entire long simulation run (both experiments were performed under identical conditions). Moreover, the pattern of these "strong" contacts forms a rather dense network involving the entire 3D structure. In other words, for the case of SEQII, the model system undergoes a transition from the molten globule state to the native like state with a fixed (solidlike) pattern of tertiary interactions.

D. Cluster analysis of the side group packing

The Monte Carlo trajectories of long, equilibrium runs of both SEQI and SEQII have been further analyzed in the following ways:

Instantaneous contact (between side groups) maps were stored at regular intervals along the trajectory, thereby providing 600 snapshots. Next, for every possible pair of snapshots, the similarity of the contact maps was calculated as the number of identical contacts in both structures. The resulting 600×600 matrix describes the simulation history, with the highest similarity score denoting that the molecule returns twice to exactly the same conformation. In the original matrix, high scoring pairs appear randomly in different positions in the matrix, indicating that both molecules were indeed in the equilibrium and that the number of conformations sampled in both simulations was much smaller then the length of each simulation. There was a marked difference between the two structures. For SEQI, the two most distant structures shared only 40% of their contacts, while for SEQII, this percentage was equal to 75%. At this point, however, it was not clear, whether both structures are moving around a single minimum, and only the width of the energy well is wider for SEQI, or whether the two energy surfaces are qualitatively different.

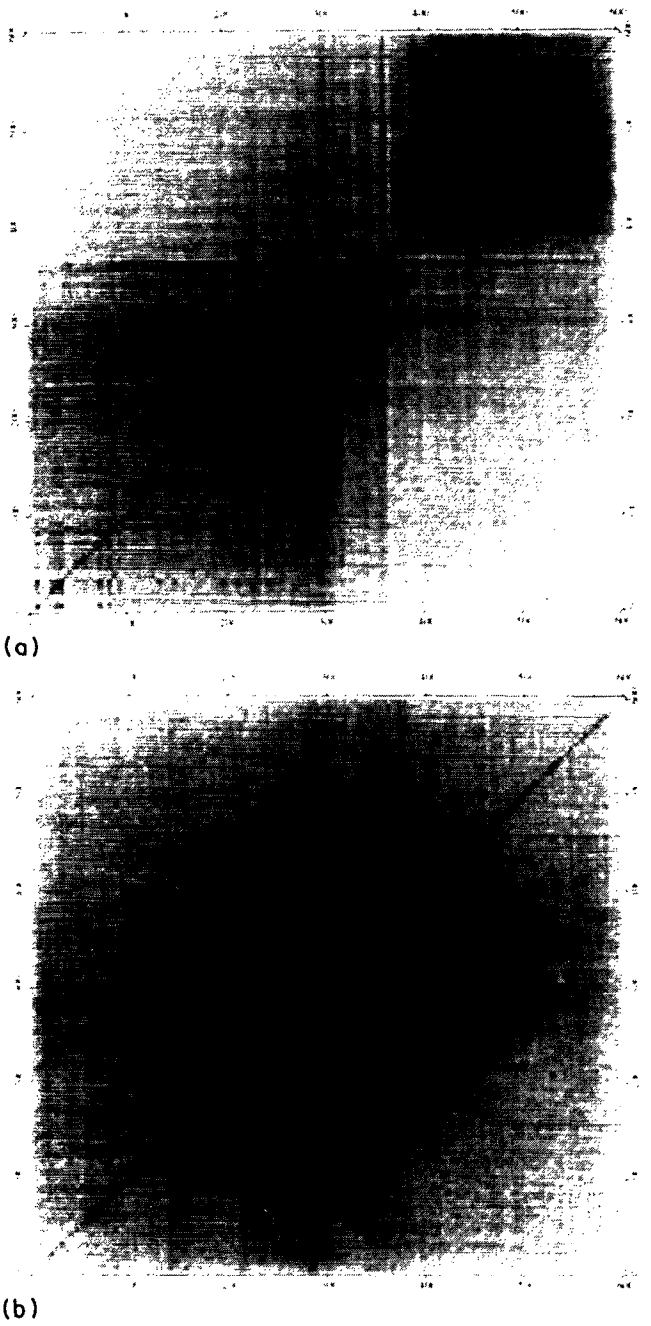

FIG. 9. Clustering of the conformational states for long isothermal runs of the right turning bundle of SEQI, (a), and the corresponding diagram for SEQII, (b). For SEQI, the black spots correspond to 60% overlap of the contact maps. For SEQII, the corresponding threshold is equal to 90%. See the text for additional details.

To answer this question, the original order of trajectory points was then changed in such a way as to bring together similar structures and to maximally separate dissimilar ones. This was done by rearranging the "history matrix" to minimize its moment of inertia, i.e., by moving larger elements closer to the diagonal. This resulted in clustering of similar configurations. As shown in Figs. 9(A) and 9(B), the resulting history matrix for SEQI and SEQII are quantitatively different. For SEQII [Fig. 9(B)], there is only one cluster of configurations, having over 90% contact overlap centered around a large plateau of the lowest energy configuration, which is repeated several times in the course of simulation. Therefore, it can be concluded, that SEQII oscillates around one, unique structure. In contrast, for SEQI there are two well defined and one dispersed clusters of configurations, with contact overlap between groups on the level of 50%–60%. This value is typical for the contact map similarity between structural homologues with weak sequence similarity. Accordingly the behavior of SEQI can be described as continuous motion among a few distinct structures, which share the same topology, but where a substantial portion of the interactions stabilizing the structures in each cluster is different. These results shed an interesting light on the differences at the molecular level between the native conformation and the molten globule states of proteins.

IV. DISCUSSION

The designed protein with the first sequence (SEQI) has been extensively studied by various experimental techniques. It has been shown (via NMR and CD experiments) that the folded state is characterized by well defined secondary structure. The thermal denaturation studies indicate that the low temperature state is very stable; however, unfolding is less cooperative than in natural helical proteins. H/D exchange and binding of hydrophobic dyes clearly show that the protein has a very mobile interior, and the packing pattern of the hydrophobic leucine core is poorly defined. The authors of these experimental studies[3] conclude that this designed protein exhibits both the properties of the molten globule (liquid like hydrophobic interior) and native state (very stable fold). Experiments on this protein and other related four-helix bundles, formed by association of shorter polypeptides, suggests that the left handed and the right handed folds may be isoenergetic.

The present simulations of the SEQI folding, and the subsequent analysis of the folded state show remarkable agreement with the available experimental data. Clearly, the model system exhibits all the features of the molten globule state. While there is no clear experimental evidence that the protein adopts both the left handed and the right handed folds, the simulation results suggest that this may be the case (see *Note added in proof*). The results of our cluster analysis of the contact maps of both proteins hopefully give new insight into the nature of the molten globule state of proteins.

Using the same interaction scheme, the simulations of the second protein (SEQII) lead to well defined folds with a rather strong indication that the lowest free energy state is the right turning bundle. Let us note here that our recently developed inverse folding algorithm[21] selects the right turning four helix bundle generated by the Monte Carlo as the most specific fold for the SEQII sequence, when compared with the left handed fold and with all other high resolution 3D structures (or their fragments) from the PDB database. This is again in agreement with the experimental data.[4] The resolution of the predicted structure is on the level of 3 Å rms for the α-carbon trace,

with well defined side group packing at the same level of accuracy. This provides a good starting point for modeling of the full heavy atom structure with reasonable accuracy.[22] The computational tools necessary for the reconstruction procedure are already available; however, they need to be improved.

What is worth noting is that the simulated folding pathways of SEQII have all the general features of the folding pathways typically seen for single domain globular proteins, with the major intermediate being the molten globule state. This is to our knowledge the first simulation of a real sequence, which starting from a random expanded conformation, forms some early intermediates, then assembles into the correct topology, and finally, after a long time associated with the collective process of side group fixation, crosses from the molten globule to the native like state.

What differentiates the present approach from other reduced representation simulations[15] of protein folding? First, there is a very flexible, and quite accurate lattice representation of the protein conformation. The dynamic Monte Carlo scheme, due to various advantages of the discretized algorithms, is probably about 2 orders of magnitude faster than an equivalent off-lattice, MC version. The speed-up factor in comparison with Brownian dynamics is even larger. Consequently, the lattice simulations correspond to longer physical times. Probably a quite substantial contribution to the speed-up arises from the smoothing of the energy surface by the lattice approximation, and by the proper account of the cooperativity of the folding process.

Second, and probably much more important, is the inclusion of a cooperative, proteinlike interaction scheme. Proteinlike means that the level of secondary structure observed at high temperature mimics very well that seen in real denatured proteins. Moreover, non physical local conformations which are rarely seen in native as well as denatured proteins (as far as the data for the latter are available) are rarely seen in this model as well. The model network of hydrogen bonds, and the potential associated with the angular correlations between the side groups drive the model system to correct local conformations. On the level of the long range interactions, the cooperative hydrogen bond network keeps the folded structures within a large sea of proteinlike elements of secondary and supersecondary structures. This reflects the physical fact that while the energy difference between protein–water and protein–protein hydrogen bonds is rather marginal, the absence of hydrogen bonds within a protein costs a lot of energy. The cooperative contribution to the hydrogen bond energy is probably the most important factor which regularizes the secondary and the supersecondary structure. These interactions generate a proteinlike conformational space for the model chain backbone, and together with pairwise interactions of the side groups are able to generate reasonable looking elements of supersecondary structure. However, if the interaction scheme is truncated at this point, then liquidlike, poorly defined packing of the side groups occurs. To overcome this problem, cooperative coupling of the pairwise interactions of side chains has been introduced. Let us point out once more that the physical reason for this cooperative term comes from the well known cooperativity of the late folding events associated with the transition from the molten globule state to the native state.[9] In our model, the explicit cooperative contribution to the system's energy is extremely important, since it favors the protein like, well defined contact patterns between packed side groups. Parenthetically it has to be pointed out, that while these multibody interactions drive the model systems towards a solidlike fixation of the side chains in the native state, some sequences (e.g., SEQI) do not undergo the transition from the molten globule state to the nativelike state. Consequently, the regular, unique packing is obtained when appropriate and is not enforced.

In the present approach, great effort has been directed to designing an interaction scheme which reflects the most general regularities seen in the globular proteins. The knowledge based elements of the interaction scheme (hydrogen bond cooperativity, templates of the pairwise interactions, side chain–side chain angular correlations) seem to be necessary to generate reasonable models for protein dynamics and structure when a reduced representation is used. Otherwise, on using reduced representations of various types, and applying the straightforward translation of the detailed potentials to very approximate schemes, one loses all the fine details responsible for the very specific molecular properties of real proteins. Consequently, such model systems tend to adopt main chain backbone conformations and side chain packing which are qualitatively rather different from that seen in proteins. This is exactly what we tried to avoid in the present work.

Presumably, the proposed methodology will require future refinement such as more complex templates of multibody interactions. These will have to be derived from detailed analysis of the regularities of the contact maps. Further improvements will likely entail a more detailed definition of the side chain contacts (perhaps based on the partitioning the side group into smaller units), and probably some direct coupling of the particular templates of the side chain packing to the actual conformation of the main chain. Nevertheless, the present applications show that within the framework of a reduced representation, it is possible to obtain unique folds, starting from the random, denatured state, and to reproduce the transition from the molten globule to the nativelike state.

Preliminary studies of other proteins show that the present algorithm is capable of generating unique folds of other simple proteins. For example, the Rop monomer,[18] which is a 120 amino acid, left turning, four helix bundle protein, designed on the basis of naturally occurred Rop dimer, has been folded.[22,23] For more complicated topologies in β proteins and α/β proteins, no unique fold has been yet obtained. However, on running the present algorithm near the renaturation temperature, one may detect large fragments of correctly assembled elements of supersecondary structure using the same set of tertiary contact templates as in the studies described here. Consequently, a high fidelity method of secondary, and supersecondary structure prediction can probably be developed based on

the present model of protein structure and dynamics. This very interesting possibility will be explored in the near future.

V. SUMMARY AND CONCLUSION

The reduced representation of protein conformation used in this work employs a high coordination, lattice discretization of the Cα trace. Fluctuating virtual bonds between Cαs and a complex Monte Carlo dynamics scheme facilitate rather physical dynamics of the model system. The side groups, or rather sets of their rotamers, are modeled as single spheres, each with a repulsive core and an attractive envelope. Two terms contribute to the main chain backbone interactions. The first involves excluded volume, hard core interactions, and the second is a simplified model of the hydrogen bond network, which reproduces about 90% of the Kabsch–Sander estimate of the main chain hydrogen bonds in 3D structures of globular proteins. The explicitly introduced cooperativity of the hydrogen bonds drives the model system to various alternative conformations of the main chain trace, closely resembling various elements of secondary structure. This hydrogen bond scheme may be considered as a complicated (and highly decorated by the influence of other interactions) three-dimensional analog of the Zimm–Bragg model[24] of helix–coil equilibrium (but in our case, it is the folded state–coil equilibrium). It enforces proteinlike, main chain conformations and is independent of the specific protein sequence. What triggers the assembly of a particular type of structure is the "generalized hydrophobic moment," represented by the potential of the pairwise angular correlations between the side group directions along the main chain (only up to the fourth partner down the chain). This represents the short range, sequence specific contribution to the configurational energy.

The tertiary interactions consist of a centrosymmetric one body potential (a marginal term for compact states), pairwise and four body interactions. The latter represent an explicit cooperative term for side group packing, and consequently, they can facilitate the transition from the liquidlike (molten globule) to the solidlike (native) state of the interior of the model proteins.

Two examples of designed helical proteins have been studied. Starting from a random coil state, the Monte Carlo algorithm in each experiment was able to assemble the four helix bundle. All the known experimental facts have been predicted for the first sequence (SEQI), including the absence of final side chain fixation. For the second sequence (SEQII), the system undergoes a transition from the molten globule to the native state, in agreement with experimental evidence for closely related systems. For this sequence, the Monte Carlo simulations predict a very stable, right turning four helix bundle, with nativelike packing of the side groups. Consequently, it has been demonstrated, that at least for relatively simple cases, that *de novo* prediction (starting only from amino acid sequence) of protein folds is entirely possible. Preliminary results show that other simple helical proteins can be folded with good accuracy as well.

The question arises if the model has been explicitly, or in a convoluted way, implicitly tuned to always give helical solutions. This is not the case, since the Monte Carlo simulations correctly assemble elements of secondary and supersecondary structure in a few tested β and α/β proteins such as plastocyanin, γ-crystallin and flavodoxin. However, to date the resulting predicted topologies have been incorrect. There are several possible explanations of the failure of topology assembly for β and α/β proteins by the present version of the simulation algorithm. First, these proteins have a much more complicated topology, with a lot of possibilities for assembly of the secondary structure elements (with deep local free energy minima), and perhaps, the annealing procedure has to be performed much more slowly and for longer times. Second, the present tertiary interaction potentials do not account explicitly for interactions between side chains and various prosthetic groups, which may break the energetic degeneracy of the various generated topologies. This has to be explored in the future. Finally, the accuracy of the geometrical representation of the model protein, especially the side groups, and the cooperativity of the side chain interactions (more elaborate templates describing side chain packing are perhaps required) may need refinement in order to make formation of the secondary elements more selective, with better account for the turn and loop regions.

In conclusion, the present results demonstrate that *de novo* folding of globular proteins on computer is possible; however, more complex applications require further refinements of the proposed methodology.

Note added in proof. Subsequent to the acceptance of this paper, Handel, Wiliams, and DeGrado (Science, in press) demonstrated by incorporating two zinc binding sites at different locations into SEQI that the right and left turning bundles of SEQI are isoenergetic. Thus, the prediction of these simulations that both topologies of SEQI are isoenergetic has been independently confirmed by experiment.

ACKNOWLEDGMENTS

This research was supported in part by Grant No. GM37408 of the Division of General Medical Sciences, National Institutes of Health. Support of the Joseph Drown Foundation is also gratefully acknowledged. We thank Dr. William DeGrado for providing us with preprints and the sequence of SEQII, and Dr. Antonio Rey for useful discussions.

[1] A. Kolinski and J. Skolnick, J. Chem. Phys. **97**, 9412 (1992).
[2] M. Levitt, Biochemistry **17**, 4277 (1978).
[3] T. Handel and W. F. DeGrado, Biophysical J. **61**, A265 (1992).
[4] D. P. Raleigh and W. F. DeGrado, J. Am. Chem. Soc. **114**, 10079 (1992).
[5] D. A. Dolgikh, R. I. Gilmanshin, E. V. Brazhnikov, V. E. Bychkova, G.

V. Semisotnov, S. Y. Venyaminov, and O. B. Ptitsyn, FEBS Lett. **136**, 311 (1981).
[6] R. L. Baldwin and H. Roder, Curr. Biol. **1**, 219 (1991).
[7] H. Christensen and R. H. Pain, Eur. Biophys. J. **19**, 221 (1991).
[8] M. Ohgushi and A. Wada, FEBS Lett. **164**, 21 (1983).
[9] K. Kuwajima, Proteins **6**, 87 (1989).
[10] T. E. Creighton, Biochem. J. **270**, 1 (1990).
[11] R. L. Jernigan, Curr. Opinion Struct. Biol. **2**, 248 (1992).
[12] A. Godzik, A. Kolinski, J. Skolnick, J. Comp. Chem. (submitted, 1992).
[13] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Simanouchi, M. Tasumi, J. Mol. Biol. **112**, 535 (1977).
[14] PDB Newsletter, (1992).
[15] J. Skolnick and A. Kolinski, Annu. Rev. Phys. Chem. **40**, 207 (1989).
[16] M. Karplus and D. L. Weaver, Nature (London) **260**, 404 (1976).
[17] M. Karplus and D. L. Weaver, Biopolymers **18**, 1421-1437 (1979).
[18] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).
[19] T. L. Hill, *Statistical Mechanics* (McGraw-Hill, New York, 1956).
[20] W. F. DeGrado (private communication).
[21] A. Godzik, J. Skolnick, and A. Kolinski, J. Mol. Biol. **227**, 227 (1992).
[22] J. Skolnick, A. Kolinski, C. Brooks III, A. Godzik, and A. Rey, Science (submitted, 1992).
[23] A. Godzik, A. Kolinski, and J. Skolnick, J. Comp. Aided Mol. (in press, 1992).
[24] D. Poland and H. A. Scheraga, *Theory of Helix–Coil Transitions in Biopolymers* (Academic, New York, 1970).