Topologies to geometries in protein folding: Hierarchical and nonhierarchical scenarios

Ariel Fernández
James Franck Institute and Department of Chemistry, The University of Chicago, Chicago, Illinois 60637 and Instituto de Matemática, Universidad Nacional del Sur, Consejo Nacional de Investigaciones Científicas y Técnicas, Avenida Alem 1253, Bahía Blanca 8000, Argentina

Andrés Colubri
Instituto de Matemática, Universidad Nacional del Sur, Consejo Nacional de Investigaciones Científicas y Técnicas, Avenida Alem 1253, Bahía Blanca 8000, Argentina

R. Stephen Berry[a)]
James Franck Institute and Department of Chemistry, The University of Chicago, Chicago, Illinois 60637

(Received 25 August 2000; accepted 3 January 2001)

This work presents a method to portray protein folding dynamics at a coarse resolution, based on a pattern-recognition-and-feedback description of the evolution of torsional motions of the backbone chain in the hydrophobic collapse of the protein. The approach permits theory and computation to treat the search of conformation space from picoseconds to the millisecond time scale or longer, the time scales of adiabatic evolution of soft-mode dynamics. The procedure tracks the backbone torsional coordinates modulo the basins of attraction to which they belong in the Ramachandran maps. The state and history of the backbone are represented in a map of local torsional states and hydrophobicity/hydrophilicity matching of the residues comprising the chain, the local topology matrix (LTM). From this map, we infer allowable structural features by recognizing patterns in the LTM as topologically compatible with particular structural forms within a level of frustration tolerance. Each such 3D realization of an LTM leads to a contact map, from which one can infer one or more structures. Introduction of energetic and entropic terms allow elimination of all but the most favored of these structures at each new juncture. The method's predictive power is first established by comparing ''final,'' stable LTMs for natural sequences of intermediate length ($N \leqslant 120$) with PDB data. The method is extended further to β-lactoglobulin (β-LG, $N=162$), the quintessential nonhierarchical folder.
[DOI: 10.1063/1.1350660]

I. INTRODUCTION

The *in vitro* folding of a natural protein is characterized by two properties that any theoretical model needs to elucidate: expediency and robustness.[1–8] ''Expediency'' is the ability of a system to correct misfolds and reach a ''native'' or physiologically active structure within milliseconds or seconds. ''Robustness'' is the insensitivity of the folding process to fluctuations and variations in the pathway. Such properties suggest using a coarse description of the flow governing the soft-mode dynamics[1–5] in terms of its key coordinates. Such a description, completely determined by the primary sequence, allows us to disregard considerable geometric detail. This approach has found theoretical[1–3] as well as experimental support.[4,5] Sosnick *et al.*[4] postulated such a rough exploration for a small globular protein whose structural nucleus has a large entropic barrier along the pathway to the hydrophobic collapse of the chain.

The notion of ''topological search'' in the folding context has been introduced in recent times:[3] the backbone torsional dynamics is represented by a rough statistical version of itself based only on local torsional states, which are characterized simply by the basins they occupy in their Ramachandran (R-) maps. Thus two local (Φ,Ψ)-torsional configurations are regarded as equivalent if they belong to the same basin. The present work builds on previous topological treatments of torsional dynamics, now elaborated by incorporation of both backbone entropies based on the ''areas of the lakes,'' i.e., the microcanonical entropies of the R-map basins, and energetic considerations of side-chain interactions. We begin by summarizing our topological approach,[1–3] which we shall abbreviate as ''RaTS,'' for ''Random Topological Search.'' Next, we show a summary of the structures the method finds for a collection of small proteins containing up to 120 amino acid residues. Then, we go on to describe a method to infer geometry from the topological analysis and apply this to the dynamics of folding of β-lactoglobulin. A brief, preliminary version of these results, especially regarding β-lactoglobulin, has been published recently.[1(d)]

There are other treatments of backbone dynamics that approach computations from a simplified, coarsened viewpoint. The LINUS algorithm[3] deals with such representations. However, our RaTS approach does not attempt to describe real dynamics; rather, it describes the evolution of the

a)Electronic mail: berry@uchicago.edu

constraints on the dynamics imposed by appearances of motifs of secondary and tertiary structure. These appearances emerge when they are recognized among randomly changing patterns of occupancies of R-basins as special patterns compatible with those structural motifs. The LINUS approach freezes coordinates as folding evolves; the RaTS approach merely slows the searching process among R-basins as folding evolves. Another useful comparison of the two approaches is the inherently hierarchical approach of LINUS, in which local structural features must form first, in contrast to that of RaTS, in which long-range scaffolding is often necessary to stabilize local secondary structures. The importance of this point will become apparent in the discussion of β-lactoglobulin later in this paper.

The description of torsional dynamics by topological dynamics characterizes isomers by a time-dependent sequence of local isomers described coarsely in terms of the evolving populations of all the R-basins of the backbone. This yields a discretized mechanistic picture in which the relevant torsional information is encoded in a matrix of N columns (N =the number of residues in the chain) called the local topological matrix (LTM). All the other motions, the fast stretching and bending vibrations, are tacitly averaged. Furthermore, in our previous work, information regarding specific torsion angles was dropped, and only the Ramachandran basin occupancies were considered.

For the sake of illustration, let 1, 2, and 3 denote, respectively, the R-basins compatible with the extended β-sheet conformation, with the compact right-handed (R) α-helix conformation, and with the compact left-handed α-helix conformation; ''4'' denotes the extra basin only present in glycine. In terms of the Φ and Ψ angles at the minima of the R-basins, the three basins accessible for alanine-like residues, the vast majority of amino acids, correspond very approximately to $\{-120°,130°\}$, $\{-80°,-40°\}$, and $\{75°,50°\}$. Thus, the two typical topological patterns or consensus windows compatible with β-hairpin two-residue turns are ...1111(33)1111..., and ...1111(42)1111..., where hairpin turn windows are given in parentheses. Similarly, the topological patterns for common reverse turns are ...1111(13)1111..., and ...1111(22)1111..., while the pattern for an R-α-helix turn is ...2222... . The structures described by such patterns are stabilized enthalpically by forming contacts whose binding energies compensate for the loss of side-chain and backbone torsional entropy.

The patterns corresponding to those structures emerge at rates compatible with real folding rates only if the pattern recognition procedure has some tolerance both to mismatchings of hydrophobic residues and to torsional incongruities.[1(d)] Thus, a consensus window such as ...1111(2223)1111... must be recognized as a β-sheet hairpin with four-residue turns,[1,3] just like the perfect pattern ...1111(2222)1111... .

The evolution of the LTM is determined by the interbasin transitions whose rates decrease as patterns compatible with structural motifs appear in the LTM, an ''if-you-see-it-freeze-it'' strategy. On the other hand, the hopping rates increase if existing topological patterns dismantle, a process associated with the formation of a 33% out-of-consensus critical bubble in the LTM.[9] Thus, the mean hopping rate for free residues is $10^{11}\,s^{-1}$: this drops to $10^{7}\,s^{-1}$ or $10^{3}\,s^{-1}$, respectively, as soon as the residues are recognized to be part of a secondary or tertiary pattern in the LTM.[1,3] In previous treatments,[1,3] we allocated equal probabilities to all the accessible R-basins. Now, we refine this by setting their probabilities proportional to their constant-energy areas, i.e., to the exponentials of their microcanonical entropies. This makes the rates of interbasin transitions proportional not only to the hopping frequency but also to the ratio of destination area to the initial R-basin area.

A structural motif is recognized and recorded as a pattern in the contact matrix (CM) when a pattern of occupancies of R-basins corresponds to the topology of a specific contact map.[1,9,10] The CM is determined from an operational definition of contact: two residues are in contact when their barycenter distance d is less than the maximum distance at which the attractive interaction of the residues is at least 1/2 kT in which, in effect, means $d<8$ Å.[1,9,10]

The iteration of two successive operations determines the LTM-CM dynamics: first, the pattern recognition operation π: LTM→CM, and then the renormalization feedback operation ρ: CM→LTM, prescribing how the next pattern recognition on an LTM is to be performed, according to the long-range interactions encoded in the latest CM.

Structural ambiguity may arise in the search for topological patterns in the LTM, because each R-basin contains a vast geometric latitude associated with a range of (Φ,Ψ) angles.[1] For example, the same basin that contains the local (Φ,Ψ) conformation ascribable to an α-helix turn with nonzero pitch also contains the local conformations of a two-residue β-turn with zero pitch. In principle, this structural ambiguity should be resolved *a posteriori*, as structural development causes one pattern to outgrow its competitors: that is, those sharing common consensus windows in the LTM. Initially, both structural motifs get recorded as alternatives; then, a bifurcation of folding pathways occurs when one pattern, typically misfolded, disappears, while the other, which offers a better possibility for structural growth or hierarchical development, grows and stabilizes.[1,9,10]

This structural ambiguity could, one imagines, become computationally unmanageable for proteins that do not conform to a hierarchical model of structural development, but instead, pass through some significant but non-native ''misfolded'' intermediates along their dominant pathways. This situation is illustrated by the experimentally probed folding of β-lactoglobulin (β-LG) with chain length $N=162$.[11–14] The native structure for this protein has a predominant β-sheet motif with a predominantly α-helical motif in its first on-path intermediates.[11–14] Furthermore, mounting evidence rules out the ''molten globule scenario''[15] as a plausible hierarchical model for the folding of this protein.[13,14]

For these reasons, β-LG becomes an ideal study case to test a refined, more rigorous means of defining the π operation. Instead of characterizing each structural motif only topologically in order to define each recognizable pattern, we now impose on the π map a geometric interpretation via energetically optimized 3D realizations of each LTM (Fig. 1). In other words, systematically generated 3D geometries

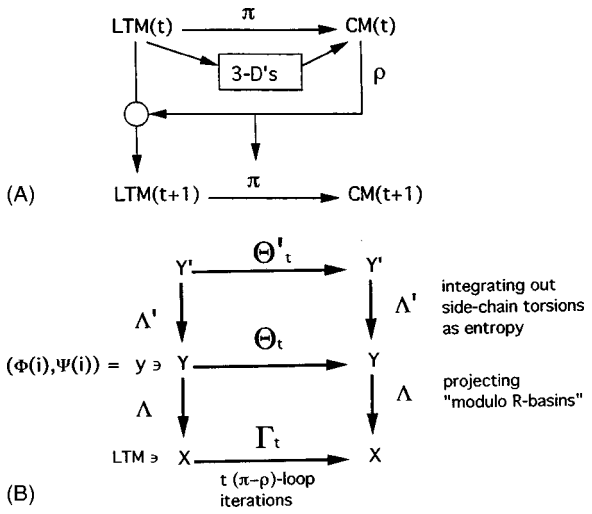

FIG. 1. (A) Formal scheme of a single pattern-recognition-and-feedback iteration at the coarse-grained level of simulation of the torsional dynamics in which torsional states are specified only with respect to the Ramachandran basin to which they belong. The elements of the iteration are the local topological matrix (LTM) and the contact matrix (CM), realized by two operations, the pattern recognition (π) and the renormalization or rate-rescaling operation (ρ). The LTM is also read to reveal 3D realizations of its topology, as shown in the next figure, as well as specific structural realizations. (The latter step may require the entropic and energetic considerations introduced through the PROCHECK program.) The operations required to determine the LTM at two consecutive instants t and $t+l$ [in real time, $t \times 64$ ps and $(t+l)\times 64$ ps] are represented. (B) Diagram relating the rough flow Γ_t generated by π–ρ loop iteration, the backbone dynamics flow Θ_t, and the full torsional dynamics flow Θ'_t, derived from geometric interpretation of the LTM. The interpretive steps are commutative, implying that the steps are self-consistent for any backbone torsional state y and any detailed conformation y' $\Lambda'\Theta'_t(y')=\Theta_t\Lambda'(y')$; $\Lambda\Theta_t(y)=\Gamma_t\Lambda(y)$; thus, all three descriptions of the dynamics are compatible with each other.

realizing the LTM enable the energy-based (and free-energy-based) ranking of the geometries corresponding to each topological pattern. Such 3D realizations will be denoted R-walks, as they are inscribed in a flexible lattice we call the Ramachandran (R-) lattice (Sec. III). Thus, for each N-sequence of R-basins, one for each residue, we assign a set of geometric realizations whose (Φ,Ψ) coordinates lie in the chosen R-basins according to an empirical probability distribution obtained from the program PROCHECK.[16] This is sufficient to define the CM matrices uniquely and associate them with specific geometries (Fig. 1), so the structural ambiguity of the topological (LTM) representation is removed. The ambiguity is now replaced by a multiplicity of detailed folding pathways consistent with the ''funnel paradigm'' which attempts to account for the expediency of the folding process.[17] The next simplification, done only tentatively until a multiple-path-following algorithm is in hand, is to choose the structure of lowest free energy and identify the pattern with that single structure.

Relevant experimental probes[4] are fluorescence quenching, circular dichroism (CD), and proton exchange (HX) labeling; these complement each other in laying a basis for a scenario for the folding of small globular proteins. Three distinguishable stages have been inferred from the stopped-flow experiments: (a) An ultrafast burst phase in the submillisecond range is thought to be the simple initial contraction of the chain as it avoids the poor solvent resulting from the sudden denaturant dilution. This good-to-poor solvent contraction is highly nonspecific and generic to any polymer. (b) A hydrophobic nucleus is inferred to form in milliseconds, with the native-like topology, which nucleates collapse without yet bringing a large fraction of the protein to its native structure. (c) Subsequent folding steps take the system downhill in free energy as it undergoes massive chain collapse with concurrent formation of extensive secondary and tertiary structure. These final steps become rate limiting only if reorganization is required to correct misfoldings.

This scenario is compatible with the RaTS model in four aspects: (a) The time scale separation between rapid intrabasin thermalization and slow interbasin hopping validates our operational procedure for the random topological search.[18] (b) Although local secondary structure might form first,[6] it alone does not necessarily lead to further structural development since, by itself, secondary structure may be unstable and require tertiary buttressing that forms on diffusional time scales.[4] An example of such behavior, β-lactoglobulin, which displays an early tendency toward taking on an organized structure but, during the first 100 ns of its folding process displays a highly mobile, dynamic, ''flickering'' equilibrium between random coil and α-helix: this system is discussed in detail later. (c) A misfold stabilized by a hierarchical folding of ever-increasing complexity becomes far more difficult to rectify than a misfold during formation of an embryonic collapse-inducing nucleus.[4,5] (d) In structured regions, changes of basin occupancies rarely appear in the R-maps as folding progresses, and the relative locations of those present remain unaltered.[16] If enough of those rare appearances of ''erroneous'' occupancies occur—in 30% of the dihedral pairs in a structured region—then that region dismantles. (It is possible that some R-basins become inaccessible as the chain becomes tightly packed, but the method presented here would not reveal this.)

II. TOPOLOGICAL PATTERN RECOGNITION

As indicated in the Introduction, random interbasin ''flipping'' generates the evolution of the LTM,[1,10] according to structure- and time-dependent rules dependent on the patterns recognized. The pattern recognition is based on a topological identification resting on the following premises:[1,9,10]

(a) Each R-basin defines the class of local curvature of the backbone chain according to the following correspondences: basin 1→extended state; basin 2→''compact'' convex state; basin 3→''compact'' concave state; basin 4 (glycine only)→maximally extended state.
(b) The pitch of a turn is defined by an R-basin combination (see below).

Thus, the symbolic dynamics, which we call ''π–ρ loop dynamics'',[1,10] results from the iteration of the two alternating operations, π and ρ. Results and predictions from such computations for natural proteins of intermediate length ($N \leqslant 120$) are shown in Table I. Whenever the active structure has been identified and entered as such in the Protein Data Bank (PDB), be it by x-ray diffraction or NMR, its associ-

TABLE I. Predicted topological resolution of protein active structure: stable LTM matrices—that is, matrices invariant under further (π-ρ)-iteration—were obtained by (π–ρ) loop dynamics for several proteins with length $N \leqslant 120$, identified by their respective PDB accession code. In a few instances, a dash appears in the LTM; this indicates that the corresponding residue within the otherwise stable topology fluctuates between Ramachandran (R-) basins, although the Protein Data Bank (PDB) assigns a specific structure to each of these. The fixed R-basin for each nonfluctuating residue in the LTM is the same as that which would result by viewing the (Φ,Ψ)-coordinates of the Protein Data Bank structure modulo Ramachandran basins. The residues at the extremities of the backbone chains have specific assignments in the PDB which may be specific to the crystal structure; the protein in solution or *in vivo* may well have nonrigid extremities, as captured by the LTM picture.

File name: 1pit.pdb-Sequence: No name-Model: 1-Number of units: 58

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ARG	PRO	ASP	PHE	CYS	LEU	GLU	PRO	PRO	TYR	THR	GLY	PRO	CYS	LYS	ALA	ARG	ILE	ILE	ARG
⋯	1	2	2	2	2	1	1	1	1	2	⋯	2	1	2	1	1	1	1	1
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
TYR	PHE	TYR	ASN	ALA	LYS	ALA	GLY	LEU	CYS	GLN	THR	PHE	VAL	TYR	GLY	GLY	CYS	ARG	ALA
1	1	1	1	2	2	2	3	⋯	1	1	1	1	1	1	2	⋯	1	3	1
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58		
LYS	ARG	ASN	ASN	PHE	LYS	SER	ALA	GLU	ASP	CYS	MET	ARG	THR	CYS	GLY	GLY	ALA		
2	2	1	1	1	2	1	2	2	2	2	2	2	2	1	1	4	⋯		

File name: pdb1aar.ent-Sequence: A-Model: 1-Number of units: 76

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
MET	GLN	ILE	PHE	VAL	LYS	THR	LEU	THR	GLY	LYS	THR	ILE	THR	LEU	GLU	VAL	GLU	PRO	SER
3	1	1	1	1	1	1	2	1	4	1	1	1	1	1	1	1	1	2	1
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
ASP	THR	ILE	GLU	ASN	VAL	LYS	ALA	LYS	ILE	GLN	ASP	LYS	GLU	GLY	ILE	PRO	PRO	ASP	GLN
1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	1	1	2	2	2
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
GLN	ARG	LEU	ILE	PHE	ALA	GLY	LYS	GLN	LEU	GLU	ASP	GLY	ARG	THR	LEU	SER	ASP	TYR	ASN
1	1	1	1	1	3	4	1	1	1	1	2	2	1	1	2	2	2	1	3
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76				
ILE	GLN	LYS	GLU	SER	THR	LEU	HIS	LEU	VAL	LEU	ARG	LEU	ARG	GLY	GLY				
1	2	1	3	1	1	1	1	1	1	1	1	1	1	3	3				

File name: pdb1c3t.ent-Sequence: A-Model: 1-Number of units: 76

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
MET	GLN	LEU	PHE	VAL	LYS	THR	LEU	THR	GLY	LYS	THR	LEU	THR	VAL	GLU	LEU	GLU	PRO	SER
3	1	1	1	1	1	2	1	2	3	1	1	1	1	1	1	1	1	2	2
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
ASP	THR	VAL	GLU	ASN	LEU	LYS	ALA	LYS	ILE	GLN	ASP	LYS	GLU	GLY	ILE	PRO	PRO	ASP	GLN
1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	1	1	2	2	1
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
GLN	ARG	LEU	ILE	PHE	ALA	GLY	LYS	GLN	LEU	GLU	ASP	GLY	ARG	THR	LEU	SER	ASP	TYR	ASN
1	1	1	1	1	3	4	1	1	1	1	1	4	1	1	2	2	2	2	3
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76				
LEU	GLN	LYS	GLU	SER	THR	ILE	HIS	LEU	VAL	LEU	ARG	LEU	ARG	GLY	GLY				
1	1	1	3	2	1	1	1	1	1	1	1	1	1	4	3				

File name: pdb1a6v.ent-Sequence: L-Model: 1-Number of units: 110

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
GLN	ALA	VAL	VAL	THR	GLN	GLU	SER	ALA	LEU	THR	THR	SER	PRO	GLY	GLU	THR	VAL	THR	LEU
3	1	1	1	1	1	1	2	1	1	1	2	1	1	4	2	1	1	1	1
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
THR	CYS	ARG	SER	SER	THR	GLY	ALA	VAL	THR	THR	SER	ASN	TYR	ALA	ASN	TRP	VAL	GLN	GLU
1	1	1	1	2	2	2	1	1	1	2	2	2	3	1	1	1	1	1	1
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
LYS	PRO	ASP	HIS	LEU	PHE	THR	GLY	LEU	ILE	GLY	GLY	THR	ASN	ASN	ARG	ALA	PRO	GLY	VAL
1	1	3	3	1	1	1	1	2	1	1	3	⋯	1	1	1	1	1	4	1
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
PRO	ALA	ARG	PHE	SER	GLY	SER	LEU	ILE	GLY	ASN	LYS	ALA	ALA	LEU	THR	ILE	THR	GLY	ALA
1	2	2	1	1	1	⋯	1	1	4	1	1	1	1	1	1	1	1	3	1
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
GLN	THR	GLU	ASP	GLU	ALA	ILE	TYR	PHE	CYS	ALA	LEU	TRP	TYR	SER	ASN	HIS	TRP	VAL	PHE
1	2	1	1	1	3	1	1	1	1	1	1	1	1	⋯	1	1	1	1	1
101	102	103	104	105	106	107	108	109	110										
GLY	GLY	GLY	THR	LYS	LEU	THR	VAL	LEU	GLU										
1	2	3	1	1	1	1	1	3	3										

ated topology coincides with our prediction, except for a few artifacts and uncertainties, as shown in Table I.

The pattern recognition operation is based on the progressive maximization of the number of hydrophobic contacts within the evolving constraints determined by the LTMs, as new topologies are developed. We shall now focus on the precise determination of the LTM→CM map. The information inferred from the LTM requires making a topological distinction between zero-pitch turns or bends, realized by R-basin windows (22) and (33) for two-residue reverse and β-hairpin turns, respectively; and nonzero pitch turns, realized for instance as (13), or (42). Obviously the

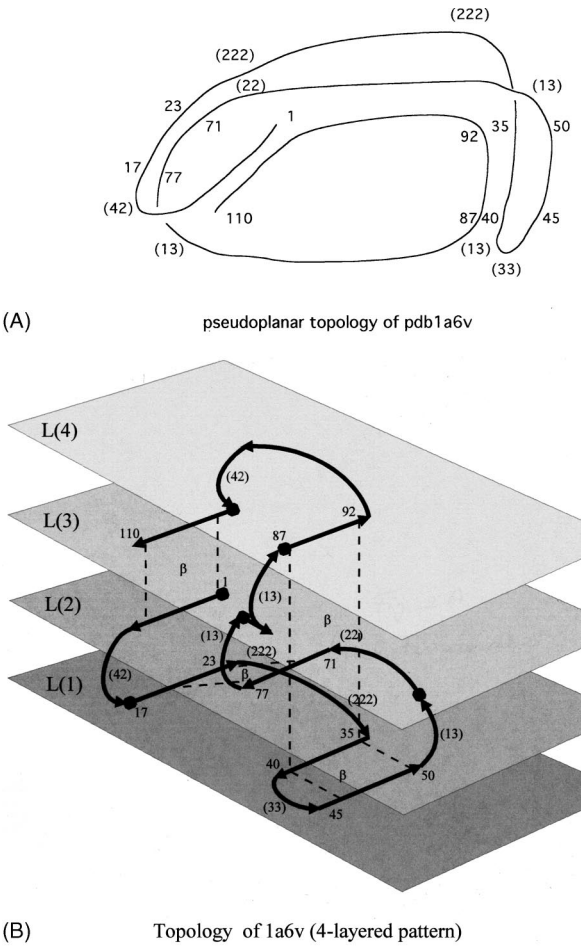

FIG. 2. (A) Predicted pseudoplanar topology for the protein with Protein Data Base (PDB) accession code $1a6v$ (sequence L), as inferred from its stable LTM. (B) Four-layer topological representation of the folded $1a6v$. The layer connections are determined by the zero and nonzero pitch turns of the chain. The contour position at which a new layer is accessed is marked by a dark circle, and β-sheets and tertiary interactions involving β-strands are represented by dashed lines.

latter case is only realized provided the first of the two residues is glycine (Gly). Thus, once the peptide chain has oriented properly, its topology may be envisioned in terms of a layered manifold of planes, in which passage to an adjacent layer is achieved by a nonzero pitch turn or loop, while zero-pitch turns or loops generate in-layer foldings. This representation makes precise the concept of ''β-sheet tertiary topologies'' because it is not possible to represent such complex β-sheet motifs in a single plane.[19] It allows us to distinguish between coplanar multi-β-strand motifs and multilayered β-strand topologies, with different strands in different layers (cf. Fig. 2).

To illustrate the multilayered representation of the LTM, we use the predicted stable LTM for the chain with accession code pdb1a6v ($N=110$), given in Table I. A multisurface representation of the LTM topology is given in Fig. 2. This protein has a nonzero pitch (42)-turn at contour positions 15 and 16. This marks a descending move from an upper layer, here denoted L(2), to the ground layer, L(1). The ground layer contains the (17–23) β-strand, followed by a loop defined by two three-residue (222)-turns at contour positions 25–27 and 31–33, and another β-strand in the region 35–40. Still within L(1), this big looped region is followed by an in-plane (33) β-hairpin turn, with an orientation opposite to the previous (222)-turns, followed by the (45–50) β-strand. Thus, the entire antiparallel (35–40)–(45–50) β-sheet is a hairpin within L(1).

At this point, the nonzero pitch (13)-turn at contour positions 51,52 brings the chain back to the L(2) layer, which also contains an in-plane (22)-turn at positions 62.63, followed by the (71–77) β-strand. In contrast to the (35–40)–(45–50) β-sheet, the antiparallel tertiary β-sheet (17–23)–(71–77) is made up of strands in different layers. This is an L(1)–L(2) tertiary interaction between β-strands. A new (13)-turn at contour positions 78,79 takes us to a third layer L(3), and to the ''unstructured'' 81–84 coil region, from which a fourth layer, L(4), is accessed through the (13)-turn at positions 85, 86. This layer contains the (87–92) β-strand. Thus, the three-strand complex β-sheet motif (87–92)–(35–40)–(45–50) is a multilayered L(4)–L(1)–L(1) β-sheet. Residues in different layers may be close in space; thus the L(4)–L(1)–L(1) tertiary β-sheet can be as geometrically plausible as an L(1)–L(1)–L(1) planar 3-strand β-sheet. A topologically descending passage takes place at contour positions 101–103, due to a (123)-turn. This turn brings the (104–108) β-strand back to layer L(3), and generates the parallel β-sheet from the sequences (9–13) and (104–108) at the extremities of the molecule, in an L(2)–L(3) topology.

As pointed out previously, in complex proteins, structural ambiguities cannot be resolved merely by a topological characterization of the emerging patterns. It is necessary to specify the significant 3D realizations of each topological matrix (LTM). This situation will be illustrated and elaborated in Secs. III and IV. We now describe the construction of the R-lattice, then analyze the various contributions to the long-range potential which govern the energetics of the R-lattice exploration, and finally show how to implement this exploration computationally.

III. FROM THE LTM TO THE CM

A. Geometry-mediated pattern recognition in the LTM

To infer a geometric structure from the pattern recognized on a given LTM, we first generate the 3D realizations of the LTM by confining the local (Φ,Ψ)-coordinates to the R-basins designated by the LTM according to the statistical weights determined from a structural database.[16] To achieve this, we must first define the geometric framework on which 3D realizations of each LTM may be determined.

As pointed out in Sec. I, to go beyond our previous treatments, we now introduce energetic and entropic information into the model. We begin the move from topology to geometry by ascribing values to dihedral angles, not just basin assignments, by including inter-residue interaction energies, and by ascribing relative entropies to each R-basin explicitly, based on the relative areas of the R-basins at the height of the lowest saddle linking them.

Up to this point, we could interpret the LTM as defining an abstract ''stick'' lattice with the α-carbon of a residue at each node, and 2, 3, or 4 ''sticks'' from that node. The sticks

point to (or correspond to) the 2, 3, or 4 R-basins that can be chosen as we go to the next residue in the LTM. Now, we extend this by including interaction energies and allowing variability of the dihedral angles within R-basins. We account for energies of inter-residue interactions by evaluating their Lennard-Jones, electrostatic, hydrophobic, and dipole–dipole interactions, in order to infer explicit geometry from the LTM. Allowing geometric distributions within each R-basin makes each stick into a fuzzy cone; inclusion of inter-residue interaction energies allows us to eliminate structures, specifically through the repulsive term in the Lennard-Jones potential, corresponding to walks on the lattice that would correspond to self-intersecting paths. At the topological level at which we construct the LTM, there is no way to incorporate self-avoidance; obviously, to apply the LTM analysis in a robust way to protein folding, we need to maintain self-avoidance. The previous analyses based on purely topological considerations and small systems happened to yield patterns *consistent with* self-avoiding geometries; the problem of excluded volume simply did not arise. Now, as we extend the method to deal with larger systems and their structural ambiguities, we must be prepared to handle this issue.

In order to introduce a locally optimized lattice, we define nodes representing allowed spatial positions for the α-carbons obtained by taking into account both the local steric hindrances inherent to the backbone and those applied on the backbone by the side chains.[1,2,6] These geometric constraints are subsumed in the new R-maps which govern the local (Φ,Ψ)-dynamics of each residue. Such R-maps encompass the local (Lennard-Jones, torsional, and local dipole–dipole) contributions to the soft intrachain, inter-residue potential terms. Thus, the new, locally optimized (rigid) lattice is constructed taking into account the minima in the Ramachandran plots.

In contrast with the approach presented in this work, molecular dynamics calculations describe folding in terms of mechanics by solving equations of motion. The necessary complexity of the requisite algorithms makes it difficult to develop descriptions on the long time scales relevant for folding,[5–8,10,16,19–22] typically of the order of milliseconds or longer. Lattice models, typically cubic,[8,23] have, by contrast, been introduced to mimic the slow conformational search performed by a foldable peptide chain. These models lack capacity to incorporate configurational entropies or to account for the local steric hindrances and backbone geometric constraints determined by the R-maps. The method used here, based on random ''flips'' of dihedral angles, begins to approach a Metropolis Monte Carlo method when we introduce energetic and entropic considerations to eliminate ambiguities in structural assignments.

The invariance of the R-map topography throughout the folding process has been confirmed, using recent programs, notably PROCHECK,[16] which determines the statistical distribution of (Φ,Ψ)-points for each structural motif. We use the PROCHECK database of 163 proteins with a resolution of 2 Å or better (cf. Refs. 16, 10) to plot the observed local torsional coordinates of native foldings. With at least 94% probability, each point plotted indeed lies in the R-basin of attraction in the R-map compatible with the structural element in which the residue is engaged. The ''error rate'' for dihedral angles in the proteins in the PROCHECK database is at most 6%. This kind of reliability of topographical features of the R-map is essential to allow us to define a flexible geometric version of the locally optimized lattice, the R-lattice, which expresses chain conformations as self-avoiding walks (R-walks). As in our previous work, each walk is determined by a random choice of R-basins, one for each residue; now, we impose a nonlocal coordinate optimization to minimize the free energy, including entropic and long-range potential energy contributions of the backbone chain, given that the residues occupy the given sequence of R-basins. Thus, the R-walks evolve according to the backbone entropic and enthalpic ''forces,'' the first of which are represented by the relative ''lake'' areas of the R-basins, as given in Table II, and the second, the long-range potential energy contributions.

As stated above, a rigid, locally optimized lattice determined by the Ramachandran energy minima for each residue would not be suitable to deal with the folding problem, since long-range potential terms may considerably distort the locally optimized torsional coordinates even as they remain confined to R-basins. As a compensation, an elastic contribution to the potential must be introduced to penalize lattice distortions correlated with torsional displacements away from the minima in the R-basins. The geometric framework thus constructed encompasses the local contributions to the

TABLE II. Normalized lake areas for basins of attraction in the Ramachandran maps for all amino acid residues, expressed as percentage probability or fraction of the total lake area. The total lake area is itself a fraction of the (Φ,Ψ)-torus area, $2\pi\times2\pi$. Basin 1 contains, among others, the extended local β-sheet conformation; basin 2 contains the right-handed α-helix local conformation and basin 3, the left-handed helix local conformation. Conventional three-letter abbreviations for the names of all amino acid residues have been adopted (Ref. 17). The row labeled ''Prec. Pro'' is associated with any residue preceding proline (Pro) other than glycine (Gly), which remains unaffected, or proline itself, which would be thus restricted to basin 1 with 100% probability.

Residue	Basin 1	Basin 2	Basin 3	Basin 4
Ala	55	41	4	0
Arg	52	42	6	0
Asn	46	36	18	0
Asp	50	42	8	0
Cys	50	45	5	0
Gln	48	43	9	0
Glu	53	41	6	0
Gly	26	24	30	20
His	46	42	12	0
Ile	56	42	2	0
Leu	54	42	4	0
Lys	51	42	7	0
Met	55	40	5	0
Phe	55	41	4	0
Pro	51	49	0	0
Prec. Pro	78	0	22	0
Ser	56	38	6	0
Thr	51	46	3	0
Trp	52	44	4	0
Tyr	54	40	6	0
Val	56	42	2	0

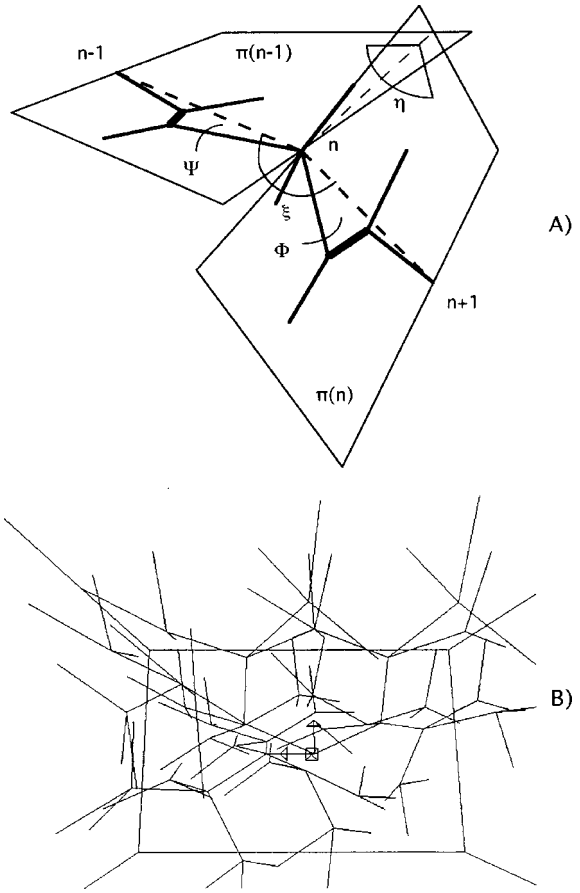

FIG. 3. (A) Representation of the local conformation of the peptide chain backbone at contour position n. The torsionally rigid $(n-1)-n$ and $n-(n+1)$ peptide bonds are represented by the thickest solid lines, α-carbons are numbered consecutively by index n, and virtual bonds between α-carbons are shown by dashed lines. For any two consecutive α-carbons $n-1$ and n, the entire backbone bond pattern joining them lies in the plane denoted $\pi(n-1)$. The local conformation of the backbone may be specified by the (η,ξ) variables representing the planar angles between consecutive planes and consecutive virtual bonds, respectively. (B) The rigid or locally optimized 3D lattice for a polyalanine chain with $N=5$. The segments are scaled to represent the virtual bond length 3.7 Å. The R-values adopted correspond to the minima in the R-map for alanine. The coordinate system is shown centered at the α-carbon at contour position $n=1$. Only self-avoiding walks are shown ($r_{ij}>3.7$ Å for any pair of units (i,j) with $j \geqslant i+3$).

potential energy by codifying the information contained in the R-maps.

B. Constructing the Ramachandran lattice

The local conformation of the chain at contour position n may be represented by the position of the α-carbon n relative to its closest neighbors $n-1$ and $n+1$ (Fig. 3). Specifying this local conformation requires two soft planar angular degrees of freedom, here denoted η, ξ, which are functions of the torsional variables Φ, Ψ: η is the angle between the planes $\pi(n)$ and $\pi(n+1)$ containing the $(n-1)-n$ and the $n-(n+1)$ peptide bonds, respectively, and ξ is the angle between the $(n-1)-n$ and $n-(n+1)$ virtual bonds.

Depending on the type of residue, locally optimal η, ξ-conformations are obtained by locating the minima in the R-map which displays the energy dependence of the Φ, Ψ-torus.[1,19] Thus, L-alanine-like residues, the most common type, possess three basins of attraction corresponding to three distinctive torsional isomers, one containing the extended β-sheet local geometries (basin 1), one compatible with right-handed compact geometry of the α-helix motif (basin 2), and a third basin compatible with the local left-handed α-helix turn conformation (basin 3). In addition, each such basin is compatible with the geometries of reverse or β-hairpin turns. Thus, in our topological notation, the two-residue turns read: (13), (22), (33), and (42). On the other hand, glycine has four basins (1 to 4), proline has only two (1 or 2), and a residue preceding proline, other than proline itself or glycine, has also two basins (1 or 3).[1,10,16,19]

The locally optimized geometries or minima in the R-map of each residue yield the corresponding locally-optimized (η,ξ) values; we call these coordinates of the minima the ''R-values.'' The rigid spatial lattice is then the set of all possible sequences of such locally optimized conformations and may be inductively constructed as follows: (a) Store all possible R-values for residue n, with $n=1,...,N$. (b) Spatially fix α-carbon 1 at a point in three-dimensional space and determine the spatial coordinates of α-carbon 2 within a sphere of radius 3.7 Å centered at α-carbon 1 by using the R-values of residue 1, thus building the first step of the backbone. (c) Given an allowable position for α-carbon n, use the set of R-values of residue n to determine the possible locations of α-carbon $n+1$ within a sphere of radius 3.7 Å centered at α-carbon n. The locally optimized lattice for a poly L-alanyl-like chain of length $N=5$ is shown in Fig. 3(B).

Following the scheme of Fig. 1, we now generate full R-walks using a prescription to choose the R-basin for each residue. This process eliminates all but the most favorable structure. Each choice of R-basin and its associated R-value is determined according to a probability p which is dictated by the basin entropy: $p=A/B$, where A is the ''lake'' area of the R-basin, filled up to the height of the lowest-lying saddle, and B is the sum of all lake areas in the R-map, a fraction of the torus area $2\pi\times 2\pi$. These areas may be computed using the PROCHECK program, which constructs a normalized statistical-sample plot of Φ, Ψ-torsional coordinates from native foldings of an ensemble (database) of natural proteins with a resolution of 2 Å or better.[10,16] The probability of a local chain conformation lying in a particular R-basin is obtained by counting the number of Φ, Ψ-points falling within the basin and dividing this quantity by the total number of plotted Φ, Ψ-points (cf. Table II). Table II demonstrates that in the random coil, the extended conformation is a dominant local structural motif because it is entropically favored and long-range intramolecular potential energy terms exert little influence.

The backbone conformational entropy change ΔS_b to go from a random coil to a contact pattern b, in which a family F of residues in the chain lie in specific R-basins, may be readily computed as

$$\Delta S_b = R \sum_{j \text{ in } F} \ln[A(j)/B(j)], \tag{1}$$

where $A(j)$ is the lake area for the specified R-basin for

residue j, and $B(j)$ is the sum of all lake areas in the R-map for residue j. Thus, the backbone conformational entropy change associated to a $v \rightarrow w$ R-walk transition is

$$\Delta S_b(v,w) = R \sum_{j=1,2,\ldots,N} \ln[A(j,w)/A(j,v)], \tag{2}$$

where $A(j,w)$, $A(j,v)$ are the R-basin lake areas for residue j in R-walks w and v, respectively.

C. Energetics governing R-lattice exploration

Canonical ensemble simulations in the R-lattice require that, in addition to long-range potential energy contributions represented as Coulombic, U_{el}, and effective hydrophobic, U_{solv}, terms, we introduce an elastic-force penalty, U_{lat}, for deformations of the locally optimized lattice, as follows:

$$U_{\mathrm{lat}} = \sum_{j=1,\ldots,N} U_{\mathrm{lat},j} = \sum_{j=1,\ldots N} k_j(\cos\eta_j - \cos\eta_{j,0})^2 + g_j(\cos\xi_j - \cos\xi_{j,0})^2, \tag{3}$$

where $U_{\mathrm{lat},j}$ is the contribution to lattice distortion energy at residue j, k_j, and g_j are local elastic moduli,[21,22] and $\eta_{j,0}$, $\xi_{j,0}$ are the R-values for residue j.

While semiempirical Coulomb parameters for the two-body terms have been compiled[21,22] and rationalized, the solvophobic potential is an effective contribution[24–26] and as such it is sometimes not included in MD simulations. The effective solvophobic potential arises as a result of the entropy-driven solvophobic effect:[24,25] the solvophobic association of nonpolar groups arises from entropy minimization associated with the ordering of solvent around nonpolar moieties, an effect not compensated by enthalpy-lowering interactions between those moieties and the solvent.[25] It has been shown[9,24–26] that the net free energy decrease due to the formation of a hydrophobic $(h-h)$ contact is proportional to the change in solvent-exposed area; the proportionality constant was estimated at 319 J/Å^2 mol, a figure derived in these works directly from the surface tension of water.

Thus, while a representation that includes explicit protein–solvent interactions exhibits the $h-h$ association as entropically driven, a description without explicit consideration of solvent molecules needs an effective solvophobic force to account for the tendency to minimize the solvent-exposed area. Using accepted average dimension parameters for the side chains of the amphiphilic or hydrophobic residues,[26] we fix the exposed area change at 40 Å^2 for contacts between relatively small residues (Ala, Gly, Pro), or between relatively small and relatively large residues (Val, Ile, Leu, Phe, Tyr, Trp, Met, His), and 80 Å^2 if both residues have relatively large side groups. This gives an average energy decrease per $h-h$ contact of 12.4 kJ/mol and 24.8 kJ/mol, respectively (cf. Ref. 9). This semiempirical parametrization enables our model to distinguish between ''nuclear residues,'' that is, those instrumental in creating a relatively stable topological nucleus triggering the hydrophobic collapse (Val, Ile, Leu, Phe, Tyr, Trp, Met, His), and the remaining hydrophobic residues. On the other hand, intercalation of a single water molecule between two h beads (or two amphiphilic beads) drives them apart with the thermal fluctuations of ordinary temperatures.[9,23,24]

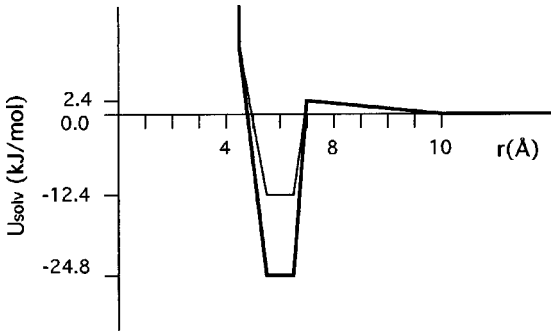

FIG. 4. The effective solvophobic two-body potentials for free residues with added repulsive Lennard-Jones term. Two families of residues are distinguished: the large hydrophobic residues (Leu, Ile, Val, His, Tyr, Trp, Phe, Met) with a potential well of-24.8kJ/mol (thick line plot), and the remaining hydrophobic residues, with a well of-12.4kJ/mol (thin line plot). The insensitivity of the results with respect to the different shapes of the effective potential holds valid provided the most favored proximity range 5.7–6.4 Å and the negligible-force range $r > 8$ Å remain invariant.

To define the solvophobic potential empirically, we first notice that the solvent-exposed area of a hydrophobic or amphiphilic residue is reduced by an amount depending on the contact hierarchy to which the residue belongs,[26] while the free energy of contact formation is linearly dependent on the concurrent reduction of the exposed area. For a generic R-walk w, $U_{\mathrm{solv}}(w)$ becomes the sum of two-body terms which, for pairs of free residues, adopts the form given in Fig. 4. The results are insensitive to the detailed shape of this effective potential, provided the most favored proximity range 5.7–6.4 Å and the range at which forces become negligible, $r > 8$ Å, remain invariant. On the other hand, the description requires two-body scaling factors λ_s, λ_t $(0 < \lambda_t < \lambda_s < 1)$ to account, respectively, for residues already engaged in secondary or tertiary structure which therefore have already partially reduced their solvent-exposed areas. Furthermore, because the strength of the contact depends on the extent of reduction of the exposed area associated with hydrophobic pairing, the residue starting with the most solvent contact is the one which determines the strength of the putative contact. Thus, the solvophobic potential for a generic walk w becomes:

$$U_{\mathrm{solv}}(w) = \sum_{(i,j)\ \mathrm{in}\ W} U_{\mathrm{solv},ij}(r_{ij}(w)) + \lambda_s \sum_{(i',j')\ \mathrm{in}\ W'} U_{\mathrm{solv},i'j'}(r_{i'j'}(w)) + \lambda_t \sum_{(i'',j'')\ \mathrm{in}\ W''} U_{\mathrm{solv},i''j''}(r_{i''j''}(w)), \tag{4}$$

where W is the family of pairs (i,j) of free hydrophobic or amphiphilic residues along the chain with $j \geqslant i+3$, W' is the family of pairs with at least one residue in the pair engaged in secondary structure, and W'' is the family of pairs with at least one residue in the pair engaged in tertiary structure; and $r_{ij}(w)$, $r_{i'j'}(w)$, $r_{i''j''}(w)$ are, respectively, the distances

between residues i and j, i' and j', and i'' and j'' in the R-walk w, obtained by optimizing local torsional coordinates within the R-basins that define the R-walk w.

The scaling factors λ_s and λ_t have been determined empirically by calibrating the simulations to reproduce the earliest intermediate whose stability already requires tertiary contact buttressing, along the dominant experimentally probed pathways, as illustrated in Sec. IV, below. Thus, in this work we have adopted the values $\lambda_s=0.55$ and $\lambda_t=0.27$. This scaling implies that any residue engaged in a contact hierarchy higher than tertiary should be considered buried, and any further potential energy decrease associated with further hydrophobic contact can only be of the order of thermal fluctuations.

D. The R-lattice dynamics

As previously discussed, the R-lattice dynamics may be subsumed into the π–ρ loop dynamics since the pattern recognition becomes unambiguous—although computationally costly—when R-walks are used to interpret the LTMs (Fig. 1). The R-lattice dynamics may be generated according to the following rules:

(a) An R-walk may be represented by an N-component vector w, where $w_j=1,2,3$ (or 4 in case the jth residue is Gly) indicates the chosen R-basin for residue j.

(b) An initial random coil R-walk is created by assigning R-basins to residues according to realizations of a discrete random variable whose values 1,2,3,4 label the R-basins in the R-map. Its probability distribution is determined by the local backbone conformation entropies, as dictated by the lake areas of the R-basins given in Table II.

(c) A transition between LTMs is realized by the acceptance of new choices of R-basins for a randomly chosen family of residues along the sequence. The probability of an R-basin transition in a chosen residue is now the quotient of the respective lake areas of the final state divided by that of the initial state, as obtained from Table II. In previous work, all basins were assumed to have equal probability.

(d) Let Z denote an (η,ξ)-coordinate vector and $Z^*(v)$, $Z^*(w)$ be chosen Z vectors for the R-walks v and w, respectively, obtained from the probability distribution of (Φ,Ψ)-points given by PROCHECK plotting[16] within each R-basin. Starting with R-walk v, a new chosen N sequence of R-basins, w, obtained by applying rule (c) and involving R-basin changes within a family of residues F, becomes an acceptable transition $v\rightarrow w$ if and only if one of the following two Metropolis-type conditions, d(1) or d(2), is fulfilled: d(1), the free energy drops in the new geometry, or d(2), the free energy increases and a temperature-based, random-variable choice meets a selection criterion. Explicitly, these conditions are the following:
d(1) $\Delta U-T\Delta S_{sc}=\Delta U(v,w)-T\Delta S_{sc}(v,w)\leqslant 0$, where $U=U_{\rm lat}+U_{\rm solv}+U_{\rm el}$; $\Delta U(v,w)=[U(Z^*(w))-U(Z^*(v))]$, and $\Delta S_{sc}(v,w)$ denotes the change in side-chain conformational entropy, given by

TABLE III. Side-chain torsional entropy parameter (exponent ζ), indicating the number of torsional degrees of freedom of the side chain.

Ala	1
Val	3
Leu	4
Ile	4
Gly	0
Pro	0
Cys	2
Met	4
His	2
Phe	2
Tyr	3
Trp	2
Asn	2
Gln	3
Ser	2
Thr	3
Lys	5
Arg	6
Asp	2
Glu	3

$$\Delta S_{sc}(v,w)=R\sum_{j\ \text{in}\ F}\ \sum_{m=1,\ldots,\zeta(j)}\ln[Q_{j,m}(w)/Q_{j,m}(v)]. \quad (5)$$

Here, $m=1,\ldots,\zeta(j)$ labels the various side-chain torsional degrees of freedom for residue j, and $Q_{j,m}(v)$ represents the perimeter measure of the portion of unit circle available to the mth side-chain torsional variable for residue j in CM(v). The ζ value for each kind of residue (Table III) is obtained by counting the number of side-chain unconstrained torsional dihedrals for each residue. The following expression, valid only for the engaged hydrophobic residues, has been adopted to simplify the computations:

$$\prod_{m=1,\ldots,\zeta(j)}[Q_{j,m}(w)/Q_{j,m}(v)]\approx q^{-\zeta(j)}, \quad (6)$$

where $q\approx 2$ (cf. Refs. 10, 25) is the side-chain torsional restriction factor for a free residue in CM(v), which has become engaged in a hydrophobic contact in CM(w). The second condition allows thermal fluctuations upward in free energy: $\Delta U-T\Delta S_{sc}>0$, **and**

$$0<r^*<\{\Pi_{j\ \text{in}\ F}\Pi_{m=1,\ldots,\zeta(j)}[Q_{j,m}(w)/Q_{j,m}(v)]\}\times\exp[-\Delta U(v,w)/RT],$$

where r^* represents a realization of a uniformly distributed random variable r in the real line interval [0,1]. Actually, conditions d(1) and (2) incorporate the side-chain entropic contribution to the free energy since the side-chain torsional coordinates have not been explicitly included in the choice of R-walks.

(e) Consistency with tenets (a) and (c) requires that the N-component vector $Z^*(v)$ with residue j in R-basin v_j must be determined (after the proper coordinate conversion) by sampling the normalized distribution of plotted (Φ,Ψ)-points in R-basin v_j, $j=1,\ldots,N$, obtained from the PROCHECK database.[16]

The underlying LTM→CM→LTM feedback dynamics,

subsumed in the LTM→R-walk→CM→LTM dynamics described in this section, has been described elsewhere[1,9,10] and needs not to be reviewed here. This section has dealt exclusively with how to resolve geometric ambiguities in assignments in the LTM→CM step by first systematically determining the optimal 3D realizations (R-walks) of an LTM, and then indicating when to accept a change in the LTM according to the specific conformational changes to which it leads.

This procedure is realized in a computer program that will be made available upon request to the authors. The authors request that users return comments, corrections, and improvements, with the understanding that these will be incorporated, with full acknowledgments, in subsequent distributions.

IV. EXPLORING THE NONHIERARCHICAL FOLDING SCENARIO FOR β-LACTOGLOBULIN

The coarse dynamics of ''modulo R-basins'' entrains the torsional dynamics of a peptide chain for time scales vastly longer than the thermalization (or equilibration) time scales within R-basins.[1,10,9] However, if the protein is conformationally plastic, as is unfolded and even partially folded β-lactoglobulin or β-LG, the topological information encoded in the LTM leads to a structural multiplicity whereby a patterned window of the LTM might correspond to consensus regions of different structural patterns with comparable statistical weight. Furthermore, such multivalued LTM→R-walk→CM assignments lead to bifurcating folding pathways. In this situation, the competing structural pattern with the highest possibility of structural growth and hierarchical development will eventually outgrow the others sharing the same LTM window at a particular time (cf. Refs. 1, 10, 27). It is a sort of Darwinian selection at the molecular level.

To illustrate these ideas we shall focus on the folding of β-LG under typical naturation conditions[11–14] and in the absence of trifluoroethanol (TFE) or other organic solvents which might add extra stabilization to α-helical motifs. The conformational plasticity of β-LG seems to favor the folding scenario called ''nonhierarchical'' as postulated by Lim.[28] The early stages of folding in such scenarios tend to produce secondary-structure motifs that do not appear in the native structure, in preference to the molten globule[15] or motifs appearing in the native structure.[11–14] (The term ''misfolded'' is sometimes used here, but because those early-stage motifs may well be on pathway, it seems inappropriate to suggest that they are erroneous.) Thus, it is believed that the predominantly β-sheet native folding of β-LG might not be a mere structural refinement of a predominantly β-sheet intermediate whose formation is in turn the kinetic bottleneck of the folding pathways. Rather, the experimental evidence implies that some kind of intermediate with extensive α-helix regions is invariably involved and, due to the larger lake areas of the R-basins for β-sheets, cf. Table II, their conversion to the latter motif appears to be an entropically driven transition. However, most experimental probes of this scenario have been either highly context dependent (e.g., halogenated organics have been used to stabilize the α-helical motifs) or are based on dissection experiments[12] that examine the folding of small fragments of β-LG, which preclude the elucidation of the role of tertiary long-range interactions on the stabilization of the transient α-helical motifs.

Given the conformational plasticity of β-LG and the occurrence of non-native, predominantly α-helical intermediates along its folding pathways, this protein becomes a crucial study case to validate the folding model presented here. Specifically, we ask whether this method can successfully trace paths down the folding path and, along the way, resolve pathway-bifurcating structural ambiguities that arise during the earliest stages of folding, but are only now becoming experimentally accessible.[14] We carried out 62 runs in the R-walk stage, each comprised of 2×10^7 iterations (1 iteration=64 ps in real time[1,10]) of the LTM→R-walk→CM→LTM cycle at $T=318$ K. These were done by simulating a reducing environment to enhance the strength of intramolecular disulfide Cys–Cys (Cys=cysteine residue) covalent bonds. Specifically, we set the depths of their potential wells three times that for a single average solvophobic contact (cf. Fig. 4). This ensures relatively fast disulfide bond reshuffling along the folding pathways, relative to the rate of *cis–trans* proline isomerization[19] or other subordinating slow processes concurrent with the folding dynamics. Thus, the model assumes no changes from the more probable ''*trans*'' Ramachandran map for the proline throughout the simulations.

Seven snapshots of metastable CMs, each with minimum lifetime of 10^2 iterations (≈6.40 ns.[3,10]), along the most probable pathway are shown in Figs. 5(A)–(G); Fig. 5(H) is the CM of the native β-LG taken from the protein database. Table IV displays the final, stable LTM that maps onto the CM shown in Fig. 5(G). Pathways very similar on the microsecond scale and longer (but not necessarily on the scale of 100 ns or less) appeared in 28 runs with a time dispersion for each snapshot of less than 10^{-7} s. The snapshots were obtained respectively at 6.4×10^{-8}, 3.0×10^{-7}, 7.9×10^{-7}, 1.8×10^{-5}, 5.2×10^{-4}, 8.8×10^{-4}, and 10^{-3} seconds. The information content in these figures must be complemented by comparison with the time evolution of the Shannon entropy,[10,27] shown in Fig. 6. This quantity reveals the time-dependent dispersion of the probability over the CM patterns. Thus, since the Shannon entropy is computed with respect to the *geometric* CM partition of the (Φ,Ψ)-conformation space into mutually disjoint classes, its high value in the submicrosecond regime shows that the structural ambiguity inherent in the *topological* (LTM) representation of the torsional dynamics is greatest in the earliest stages of folding (10 to 100ns). Each R-basin is capable of supporting several local structural motifs that are equally probable until we introduce nonbonded potential energy terms (i.e., the $U_{\rm lat}$, $U_{\rm solv}$, $U_{\rm el}$ contributions neglected in the construction of the R-map) to discriminate among them.[19,27] This fact is marked by the relatively high values of the Shannon entropy during the early stages (10–100 ns) of folding, as the system begins to nucleate and form local structures as it emerges from its random-coil state (for which the Shannon entropy is zero or very small). Later, as the CM dynamics entrains the torsional dynamics ($t>1$ μs,[3,10,27]), the fall in Shannon entropy shows the pathways converging, as the process elimi-

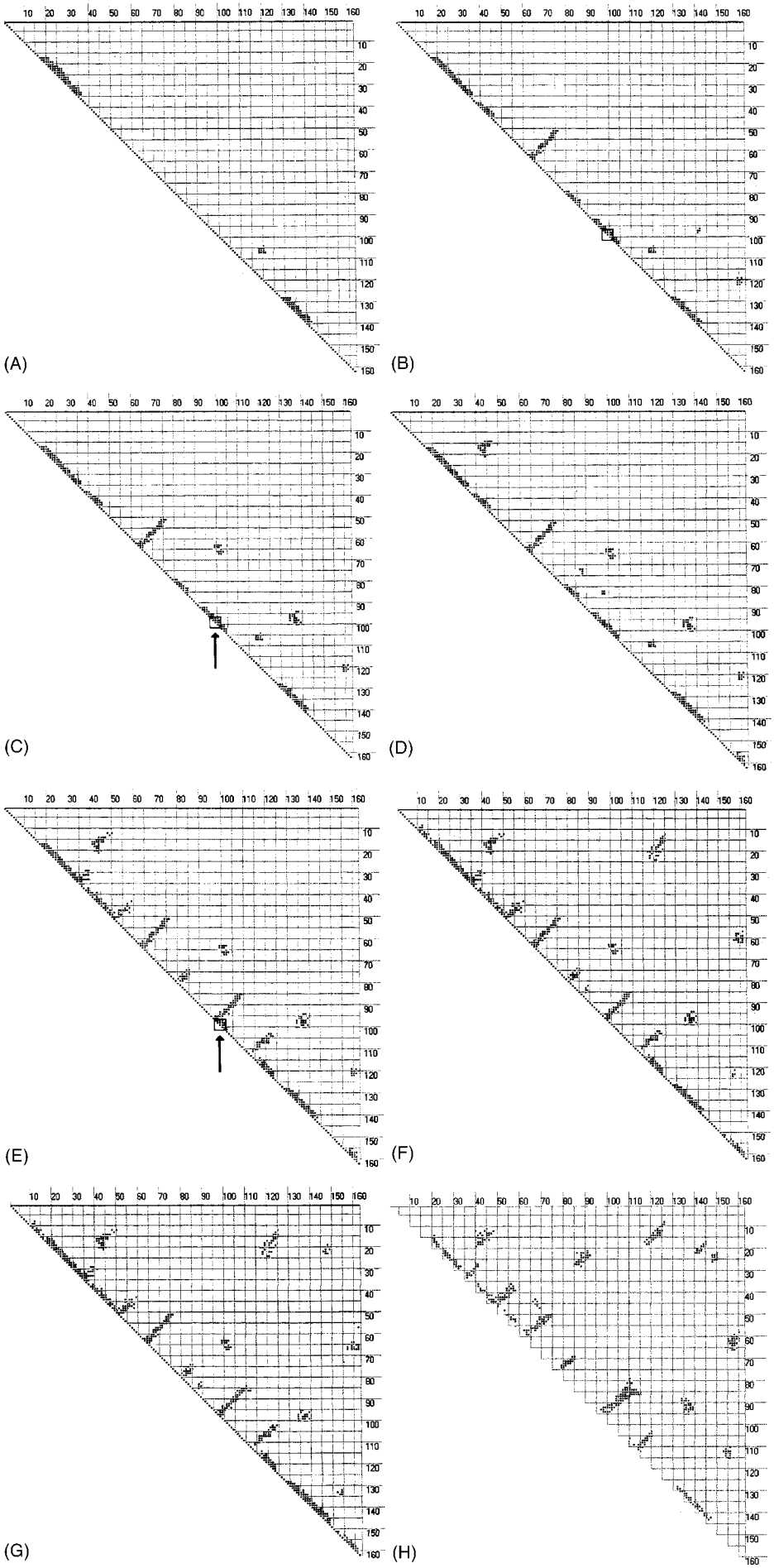

FIG. 5. Eight snapshots in the time evolution of the contact matrix for β-lactoglobulin along the dominant folding pathway simulated in the LTM→R-walk→CM→LTM dynamics. The first seven images were obtained respectively at: (A) 6.4×10^{-8}: (B). 3.0×10^{-7}; (C). 7.9×10^{-7}; (D). 1.8×10^{-5}; (E). 5.2×10^{-4}; (F). 8.8×10^{-4}; and (G). 10^{-3} seconds: (H) is the CM taken from the Protein Data Base structure for this protein. (1 iteration=64ps in real time.) The arrows in Figs. 5(C) and 5(E) indicate a region of local topological invariance and structural multiplicity (see the main text).

TABLE IV. Predicted stable LTM for β-lactoglobulin (PDB accession code *1beb*). Its projection onto a CM is shown in Fig. 5(C).

PDBCode: 1BEB-Sequence: A-Model: 1-Number of units: 162

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
LEU	ILE	VAL	THR	GLN	THR	MET	LYS	GLY	LEU	ASP	ILE	GLN	LYS	VAL	ALA	GLY	THR	TRP	TYR
⋯	⋯	⋯	⋯	⋯	1	1	1	3	1	1	2	2	2	2	1	⋯	1	1	1
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
SER	LEU	ALA	MET	ALA	ALA	SER	ASP	ILE	SER	LEU	LEU	ASP	ALA	GLN	SER	ALA	PRO	LEU	ARG
1	2	1	1	1	1	2	1	2	2	2	2	1	⋯	2	2	1	1	3	1
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
VAL	TYR	VAL	GLU	GLU	LEU	LYS	PRO	THR	PRO	GLU	GLY	ASP	LEU	GLU	ILE	LEU	LEU	GLN	LYS
1	1	1	2	1	1	1	1	1	2	2	4	1	1	1	1	1	1	1	1
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
TRP	GLU	ASN	GLY	GLU	CYS	ALA	GLN	LYS	LYS	ILE	ILE	ALA	GLU	LYS	THP	LYS	ILE	PRO	ALA
1	2	1	⋯	1	1	1	1	1	1	1	1	1	1	1	⋯	2	1	2	2
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
VAL	PHE	LYS	ILE	ASP	ALA	LEU	ASN	GLU	ASN	LYS	VAL	LEU	VAL	LEU	ASP	THR	ASP	TYR	LYS
1	1	1	1	1	1	3	3	1	2	1	1	1	1	2	1	1	1	⋯	2
101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
LYS	TYR	LEU	LEU	PHE	CYS	MET	GLU	ASN	SER	ALA	GLU	PRO	GLU	GLN	SER	LEU	VAL	CYS	GLN
2	1	1	1	1	1	1	1	1	2	2	1	2	2	1	2	1	1	1	1
121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
CYS	LEU	VAL	ARG	THR	PRO	GLU	VAL	ASP	ASP	GLU	ALA	LEU	GLU	LYS	PHE	ASP	LYS	ALA	LEU
1	1	1	2	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2
141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
LYS	ALA	LEU	PRO	MET	HIS	ILE	ARG	LEU	SER	PHE	ASN	PRO	THR	GLN	LEU	GLU	GLU	GLN	CYS
2	2	1	1	1	2	1	1	1	1	1	1	2	2	2	2	2	1	1	⋯
161	162																		
HIS	ILE																		
⋯	⋯																		

nates initially competing structural kernels which lack further possibilities for development or growth.

To illustrate this initial burst of folding possibilities in the early stages of folding, consider, for instance, the α-helical turn in the 97–101 region, as shown in Figs. 5(A)–(C). One cannot identify this motif simply by examining the 97–101 window in the LTM. A window compatible with that α-helical turn is also compatible with a β-turn, since both local structural motifs lie in the same R-basin.[3,10] Thus, we must use an optimization that includes the full intramolecular potential and not just the local terms used to construct the R-map, in order to discriminate between these two local structural motifs. This inclusion significantly alters the statistical weight of the two local torsional isomers, eventually favoring the β-turn. However, the long-range interactions are not called on until the CM dynamics has entrained the torsional dynamics, i.e., until patterns reveal the formation of secondary structures. Indeed, as shown in Figs. 5(C) through (E), the 97–101 pattern in the LTM window remains invariant during this stage. The structural transitions take place within the same R-basins for the residues in the window 97–101, while the structural assignment displayed in the CM transforms from α-helix turn to β-turn, as the dynamics of tertiary nonbonded contacts dominates the torsional dynamics beyond the 1 μs threshold.

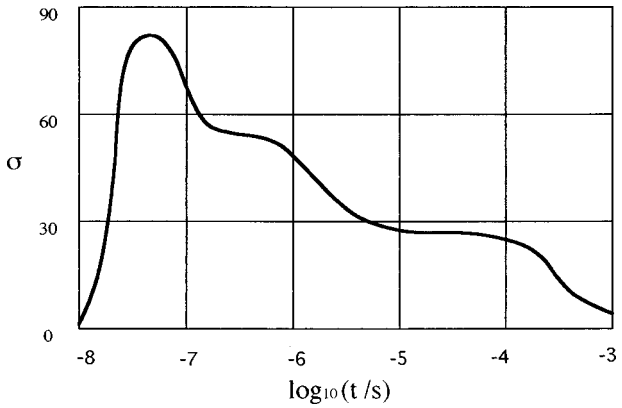

FIG. 6. The shape of the folding basin or funnel for β-lactoglobulin as determined by the time-dependent Shannon information entropy with respect to a partition of conformation space in mutually disjoint regions associated with contact patterns.

Examination of Figs. 5(A)–(C) reveals an early development of transient α-helical structure concurrently with the formation of a β-sheet in the 62–77 region within the 10–800 ns scale. Concurrently, the native (106,109) and then the non-native (121,160) disulfide bonds form since they entail a minimal loss in conformational entropy. The pattern shown in Fig. 5(C) is then stabilized by scaffolding α-helix–α-helix and α-helix–β-sheet tertiary interactions that develop within the 10–20 μs time scale [Fig. 5(D)]. Such interactions are not observable in the dissection experiments[12] because they strip away the stabilizing tertiary scaffolding. Our computations indicate that their role is crucial in stabilizing the nonnative secondary structure, and hence that experiments should be designed to probe the formation of that tertiary structure.

During the early stages, to less than about 10 μs and prior to the formation of the tertiary contacts, sequential images of the CM reveal the transient role of ''flickering'' intermediate secondary structures, predominantly α-helices. In

roughly the first 100 ns, as many as 30% of the residues change R-basin from one CM to the next. Then, when enough tertiary structure forms to stabilize the secondary helical structures by blocking the mobile helix–coil equilibration, the relevant regions undergo a progressive refolding into entropically more favorable β-sheets (cf. Table II), as revealed by contrasting Figs. 5(D) and 5(E). *This process is a kind of internal catalysis; the formation of tertiary structure effectively lowers the free-energy barrier between the α-helix and the β-sheet, while raising the barrier between helix and coil.*

The free-energy decrease that drives these transitions is reflected in the profusion of new tertiary structures and in the transformation of the secondary structures from transient α-helices into stable β-sheets and loops. However, the topological classes of many local structural motifs (e.g., reverse turns competing with α-helix turns) remain the same, as the examination of the 97–101 window illustrates. The folding transition between the intermediates displayed in Figs. 5(E) and 5(F) represents a kinetic bottleneck since it entails reshuffling a disufide bond (121,160)→(66,160) induced by the earlier formation of (10–24)–(115–125) long-range tertiary structure. The CM displayed in Fig. 5(G) has all the significant structural elements of the native fold of β-LG, as reproduced in the CM representation of the species with Protein Data Bank accession code: 1BEB, Fig. 5(H). This result is an indication that there is predictive power in a coarse-grained description based on occupancy of Ramachandran basins, followed by a finer-grained description that addresses specific geometry; that is, it appears useful to picture the folding process as one of first finding the appropriate R-basins and then, the right structure.

V. THERMODYNAMIC SIGNATURE OF THE TOPOLOGICAL COLLAPSE

The protein's topological attainment of a collapse-competent nucleus with a structure that is not yet highly specific is illustrated by the plateau in the Shannon entropy, Fig. 6. This form of representation stimulates questions that need to be addressed to complete the analysis of our data. Given that a topology specified by an LTM has many geometric realizations, and thus, an inherent conformational entropy, are there topologies compatible with structures—or groups of structures—that can survive long enough to be characterized as thermodynamic intermediates along the folding process? For example, do any of the patterns in Figs. 5(A)–(G) represent an observable, thermalized intermediate? What are the time-dependent thermodynamics of the folding process if we coarse-grain our classification to a topological level? Is there a thermodynamic signature for the hydrophobic collapse? If so, is there an identifiable collapse-inducing nucleus (cf. Ref. 4)?

These questions may be answered by determining the time dependence of the free-energy change ΔG, the energy change ΔU, and the exposed surface area work $\Delta W=\Delta H-\Delta U$ performed on the solvent as the chain folds from random coil along its dominant pathway. Strictly, we examine how the model system moves along a mean dominant pathway. At the level of mean pathways, these quantities are computationally accessible via detailed balance even at the topological level of resolution.[1d] To implement this premise, let us consider a constructive transition $v\rightarrow w$ between two patterns v, w with respective CMs denoted CM(v) and CM(w), with w the more ordered. The detailed balance principle implies the following relation:

$$D_v \exp(-U(v)/RT) r_{vw} = D_w \exp(-U(w)/RT) r_{wv}, \tag{7}$$

where D_v, D_w denote the respective degeneracies of CM(v) and CM(w), $U(v)$ and $U(w)$, the respective energies averaged over all geometric realizations of v and w, and r_{vw} and r_{wv}, the forward and backward transition rates at equilibrium, from CM(v) to CM(w), respectively. Thus, within our resolution level, we get

$$D_v = \prod_{n=1,\ldots,N} A(n,v) \prod_{m=1,\ldots,\zeta(n)} Q_{n,m}(v), \tag{8}$$

and a similar expression for D_w.

In turn, $r_{vw}=f_{vw}\exp(\Delta S(v,w)/R)$ with f_{vw} being the frequency of the $v\rightarrow w$ transition,[1,10,27] and

$$\Delta S(v,w) = \Delta S_b(v,w) + \Delta S_{sc}(v,w)$$

$$= R \sum_{n=1,\ldots,N} \sum_{m=1,\ldots,\zeta(n)} \ln\{[A(n,w)Q_{n,m}(w)]/[A(n,v)Q_{n,m}(v)]\}, \tag{9}$$

the total entropic change associated with the transition. Thus, the activation barrier for the constructive transition is entropic, since a constructive transition entails a loss in conformational freedom.[1,10,27] On the other hand, the reverse rate is $r_{wv}=f_{wv}\exp(-\Delta H(w,v)/RT)=f_{wv}\exp(\Delta H(v,w)/RT)$, where $\Delta H(w,v)$ is the enthalpy change associated with the dismantling transition.[1,10,26] Thus, combining Eqs. (7)–(9), we get

$$\Delta H(v,w) - \Delta U(v,w) = RT \ln(f_{vw}/f_{wv}) = \Delta W(v,w). \tag{10}$$

The transition frequencies f_{vw}, f_{wv} are given by[9,26]

$$f_{vw} = f \prod_{n=1,\ldots,N} [A(n,w)/A(n,v)], \tag{11}$$

$$f_{wv} = f' \sum_{n \text{ in } I(v,w)} [M(n)(B(n)-A(n,w))/B(n)], \tag{12}$$

where $f=10^{11}\,\mathrm{s}^{-1}$ and $f'=10^{8}\,\mathrm{s}^{-1}$ are, respectively, the mean rates of interbasin hopping for residues in v and w involved in the $v\rightarrow w$ transition,[9] $I(v,w)$ denotes the set of residues affected by the $v\rightarrow w$ transition, and $M(n)$ is the total number of R-basins accessible to residue n. The second factors in the right-hand side of Eqs. (11) and (12) arise from the difference in multiplicities or lake areas accessible to the residues in the R-walks v and w. The forward frequency f_{vw} given in Eq. (11) results from the probability $\Pi_{n=1,\ldots,N}[A(n,w)/A(n,v)]$ of confining residues to particular basins to yield the $v\rightarrow w$ transition. On the other hand, the expression for the backward frequency [Eq. (8)] is justified since the probability for residue n to lie outside the topological region required to form w is $(B(n)-A(n,w))/B(n)$, so the total number of effective realiza-

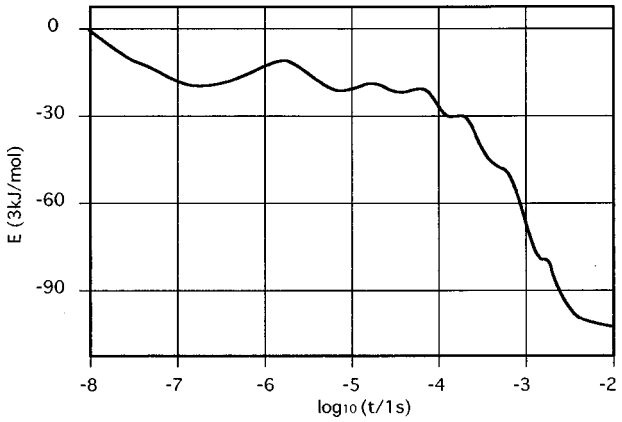

FIG. 7. Time evolution of the effective exposed-area work performed on the solvent during the folding of β-lactoglobulin. The curve actually consists of a set of discrete but very closely spaced points; each point represents an average taken over 128ps [two consecutive LTM readings (Refs. 1, 9, 10, 25)].

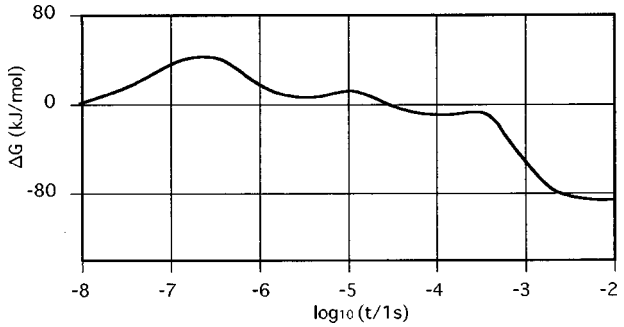

FIG. 8. Time evolution of the free-energy change during the folding of β-lactoglobulin. Each point in the plot is an average over 128 ps, as in Fig. 7.

tions of the dismantling transition $w \rightarrow v$ (analogous to ''effective collisions'' in transition state theory) is $\Sigma_{n \text{ in } I(v,w)} \times \{M(n)[B(n)-A(n,w)]/B(n)\}$.

Let us now obtain the free-energy change $\Delta G(v,w)$ associated with the $v \rightarrow w$ transition. A critical bubble nucleating the dismantling of w consists of a subwindow of 33% of the residues in CM(w) being in the ''wrong'' R-basins.[9] Thus, we may write[1,9]

$$f_{wv} \exp(\Delta H(v,w)/RT) = f' 2^{-1/3L_{vw}} 2/3(L_{vw}), \tag{13}$$

where the right-hand side of Eq. (13) is the inverse of the Zwanzig time to create a critical consensus bubble of size $(1/3)L_{vw}$.[9,19] Equation 13 yields

$$\Delta H(v,w) = RT\{(-1/3)L_{vw} \ln 2 + \ln[(2/3)L_{vw} f'/f_{wv}]\}, \tag{14}$$

where L_{vw} is the number of residues that must change basins in order to take the system from CM(w) to CM(v). The factor $2/3(L_{vw})$ counts the possible positions in the critical size subwindow along the L_{vw} window. Now, combining Eqs. (9) and (14) we compute the desired free-energy increment $\Delta G(v,w) = \Delta H(v,w) - T\Delta S(v,w)$.

Equations (9)–(14) are the working expressions to obtain the time-dependent effective work and free energy from the kinetic data associated with the folding process generated by our coarse computation. We now specialize our results for β-LG. Figure 7 displays the time evolution of the effective work along the dominant pathway. We focus now on the part of the total effective work that is performed on the solvent. This is the quantity that signals a sharp contraction of solvent-exposed volume within the submillisecond time scale, in accord with experimental data on the kinetics of hydrophobic collapse.[4,5,11–14] This means that within the submillisecond range, the collapse-inducing nucleus is formed [identified in Fig. 5(E)], and immediately the hydrophobic residues begin a drastic reduction in their solvent-exposed area, a signature of the hydrophobic collapse. Figure 7 also reveals that, after an initial reduction of exposed area in the 10^{-7} s time scale due to the formation of the non-native α-helices, the molecule spends considerable time tilting residues away from that compact, low-entropy, non-native conformation and into β-strand conformation. Because the extended β-strand conformation has the greater basin area, this is at least partially an entropy-driven transformation (cf. Table II) requiring the system to explore suboptimal topologies until a collapse-competent nucleus is established within the 10^{-4} s time scale. By that time, the prevailing motif is the β-sheet [cf. Figs. 5(A)–(G)].

The coarse-grained ''flipping'' search during the first stages of the folding process ($t < 10^{-4}$ s) is a computation vastly faster than an explicit molecular dynamics or molecular mechanics search to explore the early stage of folding. The application of flipping and structure-induced slowing at this stage, when the ''search volume'' is relatively large, is an abstract representation of the expediency of the folding process. When the protein reaches its collapsed state, a mechanical search strategy would have to deal with a drastically reduced and rugged conformation space with high kinetic barriers and consequent costly structural corrections. By contrast, in the initial stages, corrections of non-native patterns in the topological search do not require surmounting high barriers. Those early stages can be thought of as the system exploring the high rolling plain, before it starts its way down the rough walls of a basin or funnel. The most important of such corrections, the tilting away from R-basin 2 to R-basin 1 of most misfolded helical residues [Figs. 5(A)–(D)], happens within the 10^{-7} s to 10^{-6} s range. These movements involve relatively small changes in exposed area compared with those of the hydrophobic collapse, as indicated in Fig. 7.

To complete our analysis, in Fig. 8 we show the time evolution of the free-energy change during the folding process. After an initial incremental drop and rise (due to reduction of the random-coil conformational entropy with little compensating enthalpic gains), the free energy falls to a marginally stable plateau region after about 10^{-4} s, before undergoing a drastic drop in the millisecond time scale. This drop is our indication of the hydrophobic collapse. Finally, the system reaches the free energy of the native state. The picture makes apparent that no thermodynamic intermediate with a lifetime of milliseconds or even of tenths of milliseconds occurs, consistent with kinetic experiments that fail to detect significant populations of folding intermediates.[4,5,11–14] Thus, on the millisecond scale, the collapse-

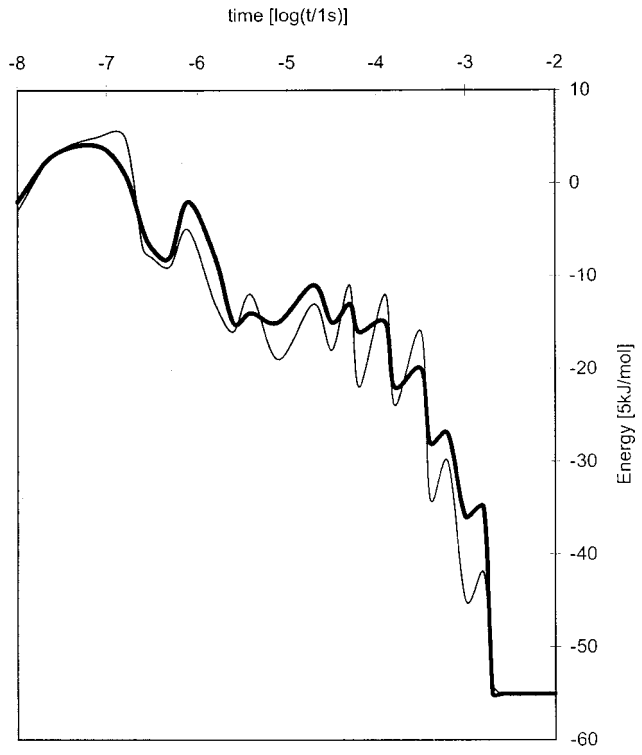

FIG. 9. Time evolution of the R-basin-thermalized energy along the most frequent (thick line) and least reproducible (thin line) folding pathway for β-lactoglobulin. The chain conformation is topologically resolved and time is coarse-grained with a 94 ps time step.

competent nucleus, while possessing some native-like topological features [cf. Figs. 5(C)–(F)], is not significantly more stable than the random coil. However, during the plateau interval up to ca. 10^{-4} s, significant local structure appears [Figs. 5(A)–(D)], notably α-helical structure, even while the solvent-exposed volume suffers no significant or progressive contraction. This picture is not consistent with a hierarchical ''simple-to-complex'' scenario in which each new local structure is part of the final, native configuration.[6] Rather, the results suggest that the system first searches for the correct *topology*, then searches within that topology for a path toward its native structure.[4,5] That is, the system finds the native-like occupancies of the Ramachandran basins, but, because of kinetic considerations, goes first to the most accessible but not the most stable structures accessible within that set of occupancies. Only when formation of enough tertiary structure forms to quench the fast local kinetics carrying the system between random and secondary structures can the chain go over to the thermodynamically optimal native form.

These observations may be clearly mapped on the coarsely resolved potential energy surface. According to Eq. (7) and Ref. 1(d), the change in R-basin-thermalized energy $\Delta U(v,w)$, associated with the $v \rightarrow w$ transition is

$$\Delta U(v,w) = RT \ln[D_w r_{wv}/(D_v r_{vw})]. \tag{15}$$

Thus, the combination of Eqs. (7), (8), and the expressions for the Zwanzig's mean first passage times r_{vw}^{-1}, r_{wv}^{-1} [3,20] yield the coarse mean energy profile along the folding pathways resolved at the topological level. This cross section is displayed in Fig. 9. The most frequently found pathway is given by the thick solid line, while the least reproducible path that still leads to the native structure is shown in thin line. The plots reveal a submicrosecond energetic decrease associated with the formation of ''unproductive'' α-helical structure followed by a reorganization barrier encountered in the μs range, leading to the ultimate β-sheet motif. The steep staircase of the dominant folding pathway does not become apparent until the submillisecond range (10^{-4} s). This sudden change in the efficiency of landscape exploration reveals the formation of the collapse-competent nucleus which contains about 50% of the total β-sheet structure of the native state and possesses the correct topology to direct the folding process downhill, well into the ms range, as Fig. 9 shows. On the other hand, the least reproducible pathway does not show that steep staircase topography of a good structure seeker until late into the ms range. This is so since the unproductive helical structure formed initially is still outgrowing the β-sheet motif until the 100 μs range has been reached, and thus, it delays the formation of the collapse-inducing nucleus. Ultimately, however, the correct topology must have been reached judging from the staircase nature of the landscape cross section in the ms range. These explorations found no trajectories to the native structure that went directly to the native β-sheet.

VI. CONCLUDING REMARKS AND FUTURE DIRECTIONS

A theory of protein folding must reconcile the vast spectrum of structural detail of the peptide chain with the observed expediency and robustness of the folding dynamics. These two properties suggest a description that begins with a search for a ''correct topology,'' a means to describe the early stages of folding without the necessity of invoking specific structures. Such a method must be capable of generating and recognizing a loosely defined nucleus that can induce the hydrophobic collapse. This kind of scenario is clearly discernible in recent kinetic experiments[4,11–14] as well as in theoretical treatments.[1,29–31]

The approach we have presented previously[1] has these characteristics. It generates an evolving picture of the folding dynamics by reducing the torsional motions of the backbone to a coarse, digitized flow in a discrete conformation space, modulo the basins of attraction of the Ramachandran maps. This coarse picture of flow in conformation space was generated in earlier contributions[1] by a pattern-recognition-and-feedback iterative algorithm. Readings of the state of the backbone chain, specified by the LTM, pick out sequences of R-basins topologically compatible with established structural motifs of proteins. The regions in which these patterns are recognized then undergo flips of their torsion angles slower than those in unstructured regions.

The previous treatments based on this method contained ambiguities because some R-basin patterns are compatible with more than one structural motif. These ambiguities arise ultimately because each R-basin encompasses a significant range of values of the Φ and Ψ angles. For small proteins, this ambiguity can be managed because relatively few structural motifs appear. For longer chains, $N \geqslant 120$, the complexity of the configuration space is so great, and the ambiguities

are so numerous, that we have chosen to eliminate all but the most favored structural motif. We shall carry the method to a level one step more elaborate than that presented here, save several favored motifs, and order them according to their probability of occurrence. The contribution presented here is a method for identifying the geometries compatible with a given topology and LTM, and then to select the structure of lowest free energy. The present method says, in effect, that the folding pathway is dominated at each step by the most thermodynamically favored available move: the next advance in the method will allow for ''bursts'' of kinetically dominated moves.

To reach the goal of this paper, we have determined the entropic and energetic contribution to the free energy. These contributions to the free energy include:

(a) The side-chain-dependent lake areas of the basins in the Ramachandran map for each residue;[16]
(b) The effect on side-chain entropy of inter-residue contacts as they depend on residue types;
(c) The local steric constraints;
(d) The site-chain-burial consequences of the solvophobic force;
(e) The fine structure of each Ramachandran basin, due to the potential energy contributions that distinguish between local geometric motifs within the same topology; and
(f) The entropic contributions that create a bias towards β-strand conformation for residues perturbed away from helical or other ''compact'' local states, as revealed by the larger lake area of the Ramachandran basin associated with the β-strand conformation.

This procedure is only a little more demanding of computation than the previous, less specific method. However, it has been successful in portraying the evolution of β-lactoglobulin through its transient α-helical stage to its native β-sheet structure. The results suggest that sharp separation of proteins into disjoint classes of hierarchical and nonhierarchical folders may be an unfruitful direction. It may be more useful to inquire, instead, about two characteristics: (1) the time scales within which a system explores its various on-path motifs and whether the lifetimes of enduring motifs are due to entropic or energetic factors, and (2) the range of variability within the set of successful folding paths.

The method presented here finds the correct structures for many of the small globular proteins in the Protein Data Base. It is not flawless; even with the geometry, the method fails for some systems, such as hen lysozyme and α-lactalbumin. However, we can look forward to a natural next step to which reference was made above, in which more than the one most favored geometric realization is retained at each juncture. The appropriate extension will allow for folding pathways that pass through transient intermediate states that are suboptimal according to the PROCHECK criteria. This will be important because PROCHECK criteria are based on fully folded structures, not on the specific stabilizing forces of the partially folded structures. These forces include those that determine the contours of the individual R-basins and also any interactions between nonbonded neighbors. This approach, which will involve bringing in the full PROCHECK criteria gradually as the folding progresses, will be especially suited to parallel computation, a step we hope to make soon.

A second refinement that lies immediately ahead is incorporation of specific effective interaction potentials for the various side chains attached to the backbone. This task, already in progress, focuses explicitly on the dimensions and hydrogen bonding capacity of the side chains.

At this point, we must ask a final question: Is there a useful distinction between systems for which the pattern-recognition-and-feedback topological algorithm is adequate to describe folding, and systems for which we must resort to geometric realizations of the topologies, as described in this paper? In extreme cases, the answer is certainly ''yes.'' Moreover, the number of allowable structures is not the only consideration that can make it necessary to select the geometries to follow. Another is the factor of error tolerance necessary for the system to achieve successful folding.[1] Certain proteins such as bovine pancreatic trypsin inhibitor require tolerance to such a large fraction ($\sim$22%,[1(e)]) of ''incorrect'' residues that their evolution generates an explosive growth of acceptable patterns, rendering the algorithm forbiddingly costly, even for intermediate-length chains ($N\sim100$). Potentially worse still, if a chain could only fold successfully with a frustration tolerance surpassing 33% of a consensus window, then the topological approach would fail altogether, since patterns with an out-of-consensus bubble of 33% dismantle![1,9] Thus, the algorithm would be forced to accept patterns that it is inherently constrained to reject. In such situations, error tolerance or plasticity must have reached a threshold. Future contributions by these authors will address this issue.

ACKNOWLEDGMENTS

The authors wish to thank Professor Tobin Sosnick and Dr. Konstantin Kostov for their insightful comments. A.F. thanks Professor Robert Huber and Professor Jerome Percus for discussions and helpful suggestions. He also thanks the National Research Council of Argentina (CONICET) for its support. The U.S. part of this research was supported by a grant from the National Science Foundation.

[1] (a) A. Fernández, Phys. Chem. Chem. Phys. **1**, 861 (1999); (b) A. Fernández and R. S. Berry, J. Chem. Phys. **112**, 5212 (2000); (c) A. Fernández, K. Kostov, and R. S. Berry, *ibid.* **112**, 5223 (2000); (d) Proc. Natl. Acad. Sci. U.S.A. **96**, 12991 (1999); (e) A. Fernández, A. Colubri, and R. S. Berry, *ibid.* **97**, 14062 (2000).
[2] O. Ptytsin and G. V. Semisotnov, in *Conformations and Forces in Protein Folding*, edited by B. Noll and K. A. Dill (Am. Assoc. Adv. Sci., Washington, 1991).
[3] R. Baldwin and G. Rose, Trends Biochem. Sci. **24**, 26 (1999); *ibid.* **24**, 77 (1999).
[4] T. R. Sosnick, L. Mayne, and S. W. Englander, Proteins: Struct., Funct., Genet. **24**, 413 (1996); S. W. Englander, T. R. Sosnick, L. C. Mayne, M. Shtilerman, P. X. Qi, and Y. Bai, Acc. Chem. Res. **31**, 737 (1998); B. Krantz, L. B. Moran, A. Kentsis, and T. R. Sosnick, Nat. Struct. Biol. **7**, 62 (2000); L. Moran, J. P. Schneider, A. Kentsis, G. A. Reddy, and T. R. Sosnick, Proc. Natl. Acad. Sci. U.S.A. **96**, 10699 (1999).
[5] K. Zdanowski and M. Dadlez, J. Mol. Biol. **287**, 433 (1999).
[6] R. L. Baldwin and G. D. Rose, Trends Biochem. Sci. **24**, 26 (1999).
[7] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[8] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[9] A. Fernández and A. Colubri, Phys. Rev. E **60**, 4645 (1999).

[10] A. Fernández and A. Colubri, J. Math. Phys. **39**, 3167 (1998).
[11] K. Shiraki, K. Nishikawa, and Y. Goto, J. Mol. Biol. **245**, 180 (1995).
[12] L. Ragona, L. Confalonieri, L. Zetta, K. G. De Kruif, S. Mammi, E. Peggion, R. Longhi, and H. Molinari, Biopolymers **49**, 441 (1999).
[13] D. Hamada, Y. Kuroda, T. Tanaka, and Y. Goto, J. Mol. Biol. **254**, 737 (1995).
[14] V. Forge, M. Hoshino, K. Kuwata, M. Arai, K. Kuwajima, C. A. Blatt, and Y. Goto, J. Mol. Biol. **296**, 1039 (2000).
[15] E. I. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[16] P. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, J. Appl. Crystallogr. **26**, 283 (1993).
[17] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997)
[18] B. Vekhter, K. D. Ball, J. Rose, and R. S. Berry, J. Chem. Phys. **106**, 4644 (1997); B. Vekhter and R. S. Berry, *ibid.* **110**, 2195 (1999).
[19] C. Cantor and P. Schimmel, *Biophysical Chemistry* (Freeman, New York, 1980).
[20] R. Zwanzig, Proc. Natl. Acad. Sci. U.S.A. **92**, 9801 (1995).
[21] S. Lifson, in *Methods in Structural Molecular Biology*, edited by D. B. Davies, W. Saenger, and S. Danyluk (Plenum, London, 1981), pp. 359–385.
[22] I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, J. Biomol. Struct. Dyn. **7**, 391; *ibid.* **7**, 421 (1989).
[23] P. A. Kollman and K. A. Dill, J. Biomol. Struct. Dyn. **8**, 1103 (1991).
[24] O. Sinanoglu and A. Fernández, Biophys. Chem. **21**, 157 (1985).
[25] P. L. Privalov, in *Protein Structure and Protein Engineering*, edited by E.-L. Winnacker and R. Huber (Springer, Berlin, 1988), pp. 6–15.
[26] F. M. Richards, Annu. Rev. Biophys. Bioeng. **6**, 151 (1977); G. D. Rose and J. E. Dworkin, in *Prediction of Protein Structure and Principles of Protein Conformation*, edited by G. D. Fasman (Plenum, New York, 1989), pp. 625–634.
[27] A. Fernández, Ann. Phys. (Leipzig) **4**, 600 (1995); J. Stat. Phys. **92**, 237 (1998).
[28] V. I. Lim, FEBS Lett. **89**, 10 (1978).
[29] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10026 (1994).
[30] Z. Y. Guo and D. Thirumalai, Biopolymers **36**, 83 (1995).
[31] A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **92**, 10869 (1995).