# Self-organization and mismatch tolerance in protein folding: General theory and an application

Ariel Fernández[a] and R. Stephen Berry[b]
*Department of Chemistry and the James Franck Institute, The University of Chicago, Chicago, Illinois 60637*

(Received 13 September 1999; accepted 21 December 1999)

The folding of a protein is a process both expeditious and robust. The analysis of this process presented here uses a coarse, discretized representation of the evolving form of the backbone chain, based on its torsional states. This coarse description consists of discretizing the torsional coordinates modulo the Ramachandran basins in the local softmode dynamics. Whenever the representation exhibits ''contact patterns'' that correspond to topological compatibilities with particular structural forms, secondary and then tertiary, the elements constituting the pattern are effectively entrained by a reduction of their rates of exploration of their discretized configuration space. The properties ''expeditious and robust'' imply that the folding protein must have some tolerance to both torsional ''frustrated'' and side-chain contact mismatches which may occur during the folding process. The energy-entropy consequences of the staircase or funnel topography of the potential surface should allow the folding protein to correct these mismatches, eventually. This tolerance lends itself to an iterative pattern-recognition-and-feedback description of the folding process that reflects mismatched local torsional states and hydrophobic/polar contacts. The predictive potential of our algorithm is tested by application to the folding of bovine pancreatic trypsin inhibitor (BPTI), a protein whose ability to form its active structure is contingent upon its frustration tolerance.  [S0021-9606(00)50611-4]

## I. INTRODUCTION: MOTIVATIONS AND OUTLINE OF THE WORK

Two dominant properties characterize the process of folding of a natural protein under physiological solvent conditions: Expediency and robustness.[1–9] Both properties have only been implicit in theoretical approaches to the so-called protein folding problem, mainly because the microscopic models have no explicit way to accommodate them.[10–15] In turn, a major impediment has precluded the implementation of suitable simulation in any sort of biopolymer:[16–19] The vast gap between the very short time scales of soft-mode torsional degrees of freedom of the chain and the much slower folding events associated with formation of secondary and higher-order structures. Thus, while the subpicosecond-to-nanosecond range in the time scale spectrum has been effectively described in molecular-dynamics simulations,[15] the $10\,\mathrm{ns}-10^2$ s range relevant to the emergence of long-range structural organization has been accessible only with kinetic models based on experimentally fitted Arrhenius-type rate coefficients and other empirical assumptions.[3,9] (It is now entirely possible to construct master equations for such organization from statistical samples of sequences of stationary states on the potential surface, but this has yet to be carried out for even a simple protein model.[20–23]) Furthermore, the dearth of thermodynamic data on isolated secondary structure motifs, an impossibility in the absence of stabilizing tertiary interactions, has inhibited the implementation of predictive free-energy minimization algorithms for proteins, such as the ones used with moderate success for RNA (ribonucleic acid) folding.[24,25]

To bring this problem into addressable form that exhibits the microscopic origins of the efficiency and robustness of folding, we introduce an approach that simulates, in a somewhat abstracted or simplified fashion, the way long-range contacts establish themselves and lead to a succession of folding steps. To achieve a workable level of simplification, we first assume we can neglect all the coordinates of the protein except the torsional angles of the backbone, the $\Phi$, $\Psi$-angles in conventional nomenclature. Next, we discretize the microscopic dynamics, simplifying the torsional state description for each residue. This is done by mapping the $\Phi$, $\Psi$—local torsional coordinates onto a discrete set of rotamers or torsional isomers. Specifically, each basin in the so-called Ramachandran potential energy map governing the local torsional dynamics,[17,18,24,26] (pp. 175 and 176 of the last of these), is identified with a local rotamer for each residue along the peptide chain. Thus, the sequence of rotameric states, one for each residue or contour unit, specifies a coarse description of the global torsional state of the backbone chain. (The torsion angle $\omega$, between residues, has a virtually unique value for most residue pairs and is therefore omitted from this description.[26])

We denote this sequence, given in matrix form, as the LTM or local topology matrix. The $n$th column in the LTM indicates the type of Ramachandran map, the torsional state

[a]Permanent address: Instituto de Matemática, Universidad Nacional del Sur, Consejo Nacional de Investigaciones Científicas y Técnicas, Bahía Blanca 8000, Argentina.
[b]Electronic mail: berry@rainbow.uchicago.edu

 

and the $h/p/n$ (hydrophobic/polar/neutral) class of the $n$th residue of the sequence, beginning in the usual way at the *N*-terminus. Thus, the precise torsional geometry is immaterial; rather, the mapping specifies the topological compatibility of the sequence of torsion angles with specific structural motifs, not necessarily with unique local structures. If, for example, $L$ sequential residues are, at some instant, in the right-handed local bending state of the chain corresponding to a specific Ramachandran basin, the topology of that sequence could be equally compatible with a zero-pitch ß-bend or loop, or with an $\alpha$-helix turn with nonzero pitch. This use of topologically based classes independent of geometric specifics is amply justified: If we were to specify geometries of local torsional states, the 30°–40° variation within each Ramachandran basin[24] would allow such a vast range of distortions as to render the geometry of any structural motif unidentifiable. Furthermore the method makes it unnecessary to take explicit account of the forces responsible for the nonbonded interactions; instead, the identification of structure-adapted patterns recognizes implicitly that those interactions are present and are adequate to establish and maintain the recognized structures.

This coarsening of conformation space would fail to render a useful structural picture if the topology of the Ramachandran maps and the relative location of the basins were sensitive to the conformations of nearby residues. Fortunately, that is not the case.[24] The local environment of each residue is not significantly affected by the interaction of its side chain with the other molecular groups within the amino acid and its nearest neighbors. This is demonstrated by the fact that the torsional coordinate values for each residue in the crystal structure of a protein lie invariably within the Ramachandran basin corresponding to the structural feature to which the residue belongs (cf. Ref. 24).

This coarse construction of conformation space enables us to codify the soft-mode dynamics determined by the short-range Lennard-Jones and local torsional contributions to the intramolecular potential. However, a consistent picture must also retain the compatibility of the local torsional states with long-range electrostatic and solvophobic interactions.[16,27] Thus, the energetic contributions of different ranges should be treated hierarchically (Fig. 1) as a set of mutually interactive constraints in which local torsional states either frustrate or allow long-range interactions, and, in turn, long-range contacts entrain or inhibit the local torsional dynamics.

A concrete, semiempirical realization that enables us to simulate folding processes begins with rapid, random changes among representations of the backbone chains as sequences of local, intrachain conformations: We assign each residue to one of four classes, which specifies the allowed, discretized assignments of each pair of $\Phi$, $\Psi$—angles in the successive residues. The four classes of residues have up to 4 possible combinations of the $\Phi$, $\Psi$—angles, corresponding essentially to the four combinations of *cis* and *trans* conformations of each angle; the allowed combinations for each kind of residue correspond to the minima of the corresponding Ramachandran plot. A recognition procedure, based on scrutiny of successions of classes and conformations of successive residues, reveals the appearance of patterns of those sequences that we can identify as characteristic of organized secondary or tertiary structures. As soon as we identify such a pattern, we severely reduce the rate at which it changes conformation, thereby ''freezing'' segments of the chain and removing them from the ''free'' random motions of the denatured protein.

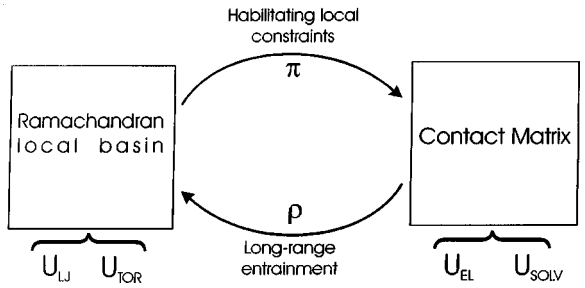


FIG. 1. Basic scheme of a single iteration loop for the $\pi$-$\rho$ parallel computation ($\pi$=projection, $\rho$=renormalization), revealing the hierarchical interplay between short and long-range terms in the intramolecular potential in a pattern-recognition representation with feedback. The Lennard-Jones ($U_{LJ}$) and local torsional terms including dipole–dipole interactions ($U_{TOR}$), define the Ramachandran PES for each residue, while the long-range electrostatic ($U_{EL}$) and solvophobic ($U_{SOLV}$) terms determine the contact patterns (CP's) recorded in the contact matrix (CM). Local torsional states are discretely mapped modulo the Ramachandran basin to which they belong. In this way the global torsional state of the chain is coarsely defined by a discretely codified matrix, the LTM, indicating in which local basin the torsional coordinates of each residue lie at a given time.

We designate the successions of types and conformations of residues that tend to self-organize readily as contact patterns (CPs). Contact patterns may be long or short, may be appropriate for forming loops, helices, or sheets (or any other organized structure that might appear). Empirical data from detailed kinetic experiments enable us to associate specific rates with structures and contact patterns, and thereby to model the folding kinetics. After some further explication, we describe the method of relating the contact pattern to a local topological matrix (LTM) and to the kinetics of folding. Then we describe how mismatch tolerance is introduced into the model, and how the model indicates quantitatively that real proteins probably need such tolerance as they fold. In the final section, we apply the method to describe the folding of bovine pancreatic trypsin inhibitor (BPTI). Then, in the following paper, we relate this method, essentially a pattern-recognition or topological approach but with implicit implications regarding topography, to a formulation much more explicit regarding topography.

A contact pattern arising from nonbonded interactions is stochastically generated in our model when the putative contacting units come sufficiently near. ''Near'' is a designation specified by the degree of tolerance or, from the opposite perspective, ''torsional frustration'' we allow in the model. Quantitatively, the specification is the range of values of the torsion angles that designate a residue as lying within a specific Ramachandran basin. A major goal of these papers is assessing what that degree of tolerance should be, for the model to represent the folding process well enough to be useful. The term ''torsional frustration'' is used here to indicate that a set of successive rotameric states is incompatible

with—i.e., outside the allowed tolerance limits for—the regional restrictions for establishing a purported contact pattern (CP). We are not using frustration in the conventional sense of mismatch in the complementarities of residues engaged in a putative contact.[27–29] Thus, the probability that a particular CP be established will be computationally determined by the value of the input variable specifying the allowable level of torsional tolerance when attempting to form the CP at a particular time. Outside this range, the system exhibits torsional frustration, in the terms we use here.

Whenever a specific CP has been established, the dynamics of the chain, specifically the dynamics that transform the LTM, change in the region of the new CP. This is implemented as a reduction of the interbasin transition frequency for those residues engaged in the CP, including those in its interresidue loops and other structural motifs. Thus, our examination of the topography and dynamics of the potential surface begins by exploring the potential energy surface (PES) stochastically and applying an iterative pattern-recognition process with a feedback loop, within which each CP is activated according to the frustration-dependent tolerance level.

This treatment encompasses three fundamental aspects: (a) The minimal torsional frustration tenet, reflecting the protein's tendency to maximize the topological compatibility of short- and long-range interactions;[24] (b) the hierarchical entrainment of the different modes into secondary and then tertiary structures, reflecting the locking-in of local torsional transitions within Ramachandran basins as the corresponding CPs are established; and (c) the possibility of entering any of several competing CP basins with different probabilities according to the level of contact frustration involved.[28–30]

According to (c), and given the tolerance to classical frustration, different competing CPs could lead to different folding pathways, generating either a common structure or different, competing target structures. The former is probably the most widely held picture today of the folding process; the latter is natural for any random-sequence heteropolymer and would be very difficult to disprove, even in sequences reaching the ''native'' targets of natural selection. The possibility of multiple native structures having roughly equivalent physiological activity needs careful scrutiny, particularly if they differ in their scaffolding but not in the configurations of their active sites. The uniqueness inferred from crystal structures of natural proteins may be a result of the strict stereochemical demands involved in crystal formation, rather than uniqueness of the folded structure in solution or *in vivo*, as suggested by experiments done by the Frauenfelder group.[30]

We denote by $i \rightarrow j$ the transition to establish the specific CP $j$ from a prior pattern $i$. The distribution of barriers to CP transitions depends on the level of torsional tolerance or frustration of the CP $j$. In our procedure this distribution is selected, either by trial and error or from empirical data; it is meant of course to mirror a real tolerance. A zero-frustration barrier for the $i \rightarrow j$ transition is the most improbable because its entropic cost is the highest but, on the other hand, if chosen, it yields a 100% probability for the transition. The plasticity arising from the tolerance, paramount to biological activity, makes it feasible to form a motif with high orientational demands, such as the closure of a large loop. Barriers involving relatively small conformational entropy losses are relatively improbable but are easy to surmount, and therefore, are kinetically important. Furthermore, once a barrier is surmounted, the long-range potential between nonbonded residues exerts a drag which assists the local torsional changes that guide the system toward a basin bottom that is a native, active state or better, a set of interconnecting states.[29]

Finally, yet another kind of tolerance is incorporated in our treatment: The tolerance to classical frustration resulting from $h/p/n$ complementarily mismatches in the side chains.[28–30] This form of frustration, unlike the torsional frustration which may be ''corrected'' (in pattern recognition terminology[31]) or ''funneled out'' (in dynamicist's terminology[29]), has no funnel or staircase topography to remove such frustration and ultimately induce the relaxation of the frustrated LTM into a state of perfect contact matching. Only fluctuations can remove this kind of mismatch. Nevertheless, tolerance is needed for this kind of frustration, as a means to stimulate the formation of tertiary contacts which provide their eventual thermodynamic compensation, and sometimes even induce a cooperative structural growth, as illustrated in the following paper [J. Chem. Phys. **112**, 5223 (2000)].

The paper proceeds as follows. In Sec. II we define the LTM, the rules governing the LTM dynamics, and their represcription every time a new CP has been recognized or read in the LTM. In essence, we implement the ''stiff'' version of the parallel folding algorithm. The technical details of the actual implementation of the inherently parallel algorithm on a conventional personal computer (PC) are presented as supplementary material in the electronic archive. Section III gives an analysis of the pattern recognition operation now viewed as a stochastically activated map, LTM→CP. Special emphasis is placed on the relation between this map and our theoretical underpinnings of the frustration tolerance of the folding process. Section IV presents the application of the method to BPTI, and Sec. V consists of concluding remarks.

## II. DISCRETIZED TORSIONAL DYNAMICS OF THE PEPTIDE CHAIN: A PARALLEL COMPUTATION

### A. Representational tools

The Ramachandran map or local PES[3,24,26,27] governs the local $\Phi$, $\Psi$—torsional dynamics of each residue or amino acid along the peptide chain. Such maps present a discrete and small number of basins of attraction enabling us to codify the torsional state of the residue discretely by defining the $\Phi$, $\Psi$—coordinate vector modulo the basin of attraction to which it belongs. This coarse simplification of local torsional states is feasible because intrabasin stretching and bending vibrations are much faster and equilibrate much faster than the torsions that determine the conformation at each residue of the polypeptide chain.[32] We may thereby classify residues according to their particular type of Ram-

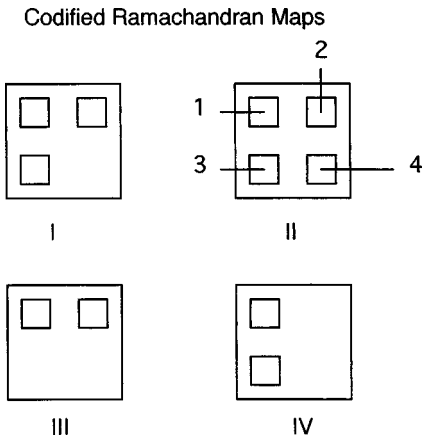


FIG. 2. Discrete codification of local torsional states of amino acid residues by designation of the basin (1, 2, 3, or 4) in the Ramachandran map where the torsional coordinates $\Phi$, $\Psi$ lie. Each basin is labeled according to its relative position in the two-torus or local conformation space. There are four types of maps, I–IV, according to whether the residue is *L*-alanyl-like (I), glycine (II), precedes a proline (III), or is proline (IV). Thus, a Ramachandran discrete variable $R(\mathbf{y},n)=1,2,3,4$, indicates the basin for the $n$th residue in the conformation roughly defined by the LTM $\mathbf{y}$.

achandran map, and label the basins or torsional isomers according to their relative position in the $\Phi$, $\Psi$-torus, using the scheme indicated in Fig. 2.

The coarse-grained version of the torsional dynamics of the chain is then given by the time evolution of the LTM. The $n$th column in this matrix is made up of two objects (Figs. 3 and 4). One indicates the type of Ramachandran map for the $n$th residue and the particular basin within this map where the $n$th residue is at the chosen time; the other object is fixed and indicates the $h/p/n$ class of the residue. Thus, the topological compatibility of successive local rotamers with a specific structural pattern may be diagnosed from the pattern in a window in the LTM, as illustrated in Figs. 3 and 4. We shall call such an array a ''consensus pattern.'' The emergence of a consensus pattern in a window indicates the fulfillment of local topological constraints which, when met, engender the formation of a structural motif involving long-

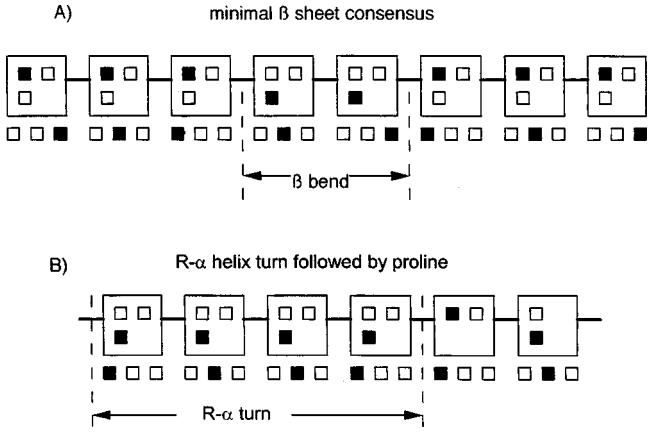


FIG. 3. Typical consensus windows in the local topography map (LTM): (a) LTM consensus window for a minimal ß-sheet structural motif; (b) LTM consensus window for a right-handed $\alpha$-helix turn interrupted by a proline (and a residue preceding proline).

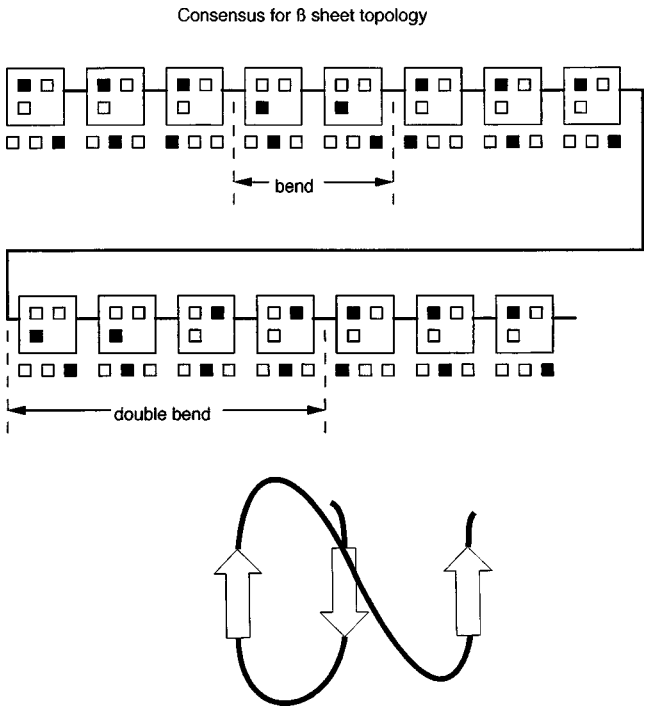


FIG. 4. An LTM consensus window for the complex three-strand antiparallel $\beta$-sheet motif shown.

range, nonbonded interactions for the residues corresponding to that window.

These considerations lead us to define a ''Ramachandran variable,'' $R(\mathbf{y},n)$, indicating the basin of attraction of residue $n$ in the conformation coarsely defined by the LTM $\mathbf{y}$. We classify residues or amino acids into four groups: *L*-alanyl-like, glycine, proline, and any residue preceding proline (Fig. 2). Thus, because an alanyl-like residue with contour position $n$ has three basins of attraction,[24] there are three possible values for $R$, depending on $\mathbf{y}$: $R(\mathbf{y},n)=1,2,3$, while if glycine is at the $n$th position, we would get $R(\mathbf{y},n)=1,2,3,4$, again depending on $\mathbf{y}$. A proline residue may have only $R(\mathbf{y},n)=1,3$, while the $n$th residue preceding proline may have only $R(\mathbf{y},n)=1,2$.

A seven-step procedure allows us to compute the discretized torsional dynamics thus:

(i) We introduce a ternary variable $G(n)=1,2,3$ indicating respectively whether the $n$th residue along the chain is hydrophobic, neutral or hydrophilic (polar).

(ii) We determine the type of Ramachandran plot (I–IV), as indicated in Fig. 2, for each residue $n=1,...,N$.

(iii) We define an LTM $\mathbf{y}$ by two rows $\{R(\mathbf{y},n),G(n)\}_{n=1,...,N}$, as illustrated in Figs. 3 and 4. Thus, $R(\mathbf{y},n)=1$ indicates that the $n$th residue has adopted the extended conformation *compatible* with a $\beta$-sheet; $R(\mathbf{y},n)=2$ indicates that either the $n$th residue has adopted a locally compact conformation *compatible* with a $\beta$-bend (zero pitch), or with a left-handed $\alpha$ helix; finally, $R(\mathbf{y},n)=3$ indicates that the conformation of the $n$th residue is *compatible* with the formation of a $\beta$-bend or with a right-handed $\alpha$ helix.[24]

(iv) We perturb the LTM by simulating interbasin torsional transitions according to fixed transition probabilities.

(v) At fixed time intervals we search for windows exhibiting consensus patterns of torsional isomers along the chain (Figs. 3 and 4).

(vi) We evaluate and translate such patterns into a contact matrix (CM) whose changes we monitor throughout a range of times from 10 $\mu$s to $10^2$ s, depending on the system and the conditions. Thus, diagnosing the evolving structure becomes a periodic pattern recognition and therefore, a parallel operation taking place at regular intervals.

(vii) The recognition operation $\pi$, which is actually a projection, is subject to a feedback loop, whereby a renormalization operation, $\rho$, readjusts the inter-basin mean transition frequencies according to the latest CP identified. The contour ranges of intrachain interactions and contour distances are renormalized relative to the latest CP formed. In other words, the renormalization operation introduces long-range correlations on the LTM according to the scheme given in Fig. 1.

### B. Folding as a pattern recognition operation

The dominant secondary structure motifs can be identified as recognizable patterns emerging in the time-dependent LTM $\mathbf{y}=\mathbf{y}(t)$. Thus, the right-handed $\alpha$-helix requires a window of residues with $R(\mathbf{y},n)=3$. Without loss of generality and for the sake of notation, we shall identify this motif by a window of the LTM $\mathbf{y}$ with $R(\mathbf{y},n)=3$ and a periodic $G(n)=1=G(n+3)$ or $G(n)=1=G(n+4)$ hydrophobicity (Figs. 3 and 4). Because glycine is highly disruptive of an $\alpha$-helix,[24,33] if its local diagram appears in what would otherwise be a helix-forming consensus pattern in a 4- or 5-residue window, the entire helix turn containing glycine is obliterated from the CM. The disrupting tendencies of proline, on the other hand, do not require special instructions because $R(\mathbf{y},n)$ cannot take the value 3 for a residue preceding proline, as shown in Figs. 2–4. Figure 3 shows a minimal $\beta$-sheet and a right-handed $\alpha$-helix turn interrupted by a proline (and a residue preceding proline). Similarly, for a left-handed $\alpha$-helix, we must demand persistent values of 2 for $R(\mathbf{y},n)$, while retaining all other conditions regarding hydrophobic periodicity along the chain. Figure 4 shows tertiary structure, as the complex 3-strand antiparallel $\beta$-sheet.

Being pleated structures, $\beta$-sheets are characterized by the persistence of a sequence of local conformations with $R(\mathbf{y},n)=1$. In order to fulfill hydrophobic/polar compatibilities, the $G$-values must match in a parallel or antiparallel fashion, $h$-to-$h$ and $p$-to-$p$, depending on the relative orientation of the strands in the $\beta$-sheet (Figs. 3 and 4). For illustration, Fig. 4 shows a 3-strand $\beta$-sheet topology together with its LTM consensus pattern. A structural pattern in the same topology class[33] will be generated in our simulation of the folding of BPTI.[32,34–36]

Turns and bends may contribute to formation of the $\beta$-sheet, or may simply allow the chain to form hydrophobic contacts. Since actual geometry is immaterial to the LTM description, we can treat turns and bends as generic structural motifs, regardless of whether or not they realize $\beta$-sheet topologies. To close a loop, such a motif requires a consensus window with $R(\mathbf{y},n)=2$ or $R(\mathbf{y},n)=3$ in the LTM at the time of its evaluation.

### C. The distribution of interbasin transition frequencies in the evolution of LTMs

Our discretized model of topological dynamics covers the spectrum of activated molecular motions occurring in the time range from 1 ps to 1 ms. Faster internal motions, including rapid, diffusionlike unhindered torsions[32,35] in a free residue have been averaged implicitly and appear as conformational entropy of the state defined by the coarse LTM representation (cf. Ref. 19). Thus, the time range relevant to LTM transitions runs from $\sim 10^{-11}$ s for the calculated diffusional displacements of flexible hinged domains[27] to $\sim 10^{-4}-10^{-3}$ s, for the fast exchange between folded and unfolded states in which two secondary elements engage in a tertiary interaction.[34] Within this range is the typical mean time of $10^{-7}$ s for a localized helix-unwinding event leading to a bubble.[16,19,32,33]

These considerations lead us to define a temperature-dependent normalized distribution of transition periods, $\omega=\omega(\tau)$, for the $N$ independent basin-to-basin transitions corresponding to the local Ramachandran landscapes. Whether a local transition takes the system from a correct basin to an incorrect basin depends on what basin is regarded as correct on the basis of a putative structural element being identified in the consensus map. We assume that the distribution $\omega=\omega(\tau)$ has three Gaussian peaks centered at characteristic periods $10^{-11}$, $10^{-7}$, and $10^{-3}$ s, which we shall denote as I, II, and III. These transition times are assigned within this distribution to incorporate the effect of thermal fluctuations on the formation of consensus and thus, on structural transitions. Each Gaussian peak has a dispersion $\sigma^2=g_J T$, ($J=$I, II, or III) where the constant $g_J$ depends on the actual denaturation temperature $T_{\mathrm{dena}}$ and on the consensus-based interpretation of denaturation, as shown below.

The trimodal transition-time distribution allows us to classify residues in three classes: Class I contains all free residues, that is, residues not engaged in any structural motif, with mean basin transition period $10^{-11}$ s; class II contains all residues with mean transition time $10^{-7}$ s engaged in secondary structure but not in tertiary interactions; and class III contains all residues engaged in tertiary structure, whose mean transition time is $10^{-3}$ s. This classification of residues according to their inherent mean basin transition times is compatible with fluorescence depolarization probes for unhindered torsional motions,[32] with typical timescales for localized helix disruptions, and with diffusion-collision models[35] in which secondary structure is stabilized further by forming tertiary contacts.[34] The increase in these local transition times is directly associated with transitions from less-strongly to more-strongly bound structures, with corresponding decreases in enthalpy.

Accordingly, a single step in the evolution of the LTM is fixed by a lottery from which first the direction of the basin transition is chosen and then the transition times are assigned from within Gaussian distributions centered at 10 ps, 100 ns, and 1 ms, depending on whether the residue is of type I, II,

or III—which, in turn, is a function of the contact pattern of its immediate environment. Thereafter, a new direction and a new escape time is assigned from the lottery for each new transition. This procedure yields the maximum permanence to the extended local conformation 1 for a free alanyl-like residue, in accord with observations.[32] The transitional frequency $f=2\pi/\tau$, corresponding to a residue not engaged in an intrachain interaction or loop (a class I residue) satisfies the inequality

$$|f^{-1}-10^{-11}\,\mathrm{s}|<|\tau'-10^{-11}\,\mathrm{s}|,$$

with $\tau'$ satisfying

$$\omega(\tau')=\mathrm{Infimum}_{\tau}\ \{\omega(\tau)\geqslant 1/[2N]\}. \tag{1}$$

The condition yielding the shortest escape time $\tau'$ arises from the fact that there are at most $2N$ possible local transitions in the peptide chain, with a maximum of two transitional directions for each residue in each given basin. Such considerations yield the value $\tau'=6$ ps at $T=298\,^{\circ}$K. Thus, the time step between two pattern searches of the LTM must be in the range $(2^6/6)\times 6\ \mathrm{ps}=64\ \mathrm{ps}$, that is the minimum time to get a CP transition corresponding to the formation of a stable $\beta$-bend or helix turn engaging six residues, based on the shortest basin-transition times. This mean first-passage time has been computed in Ref. 3 (see also Sec. III).

The other two peaks in the distribution $\omega=\omega(\tau)$ correspond, respectively, to mean escape times for residues in secondary and tertiary structural elements. Again, the same considerations apply in relating the transition times to the search frequencies for class II and class III residues. These rules imply that the residues in loops, bends, or turns concurrently formed with any secondary or tertiary structural element adopt the cadence of the structural element itself.

As a consequence, the rate of pattern search is subject to redetermination with each CP transition: A CP determines which columns in the LTM correspond to free or class I residues and which correspond to residues engaged in class II (secondary) or class III (tertiary) interaction. When the class of a residue or set of residues changes, so does the time scale of its dynamics.

### D. The temperature ($T$) in the discretized LTM dynamics

Secondary-structure dismantling or local denaturation materializes and is recorded as such by deletion in the CM whenever a pattern we call a ''consensus bubble'' forms amongst a set of contiguous class II residues. By ''consensus bubble'' we mean that in the $R$-row of the LTM, a consecutive sequence of Ramachandran variables of length 30% of the total length of the consensus window[16,32] must fail to match the values required for the consensus pattern at the time the LTM is read. The residues previously engaged in the structure and in its concurrent loops are reclassified from class II to the higher-frequency class I.

Cooperative effects reflect themselves mechanistically in the formation of the consensus bubble: For example in the $\alpha$-helix motif, the larger the helix, the more improbable is a 30% mismatch within a class II sequence of residues. Furthermore, these considerations enable us to estimate the constant $g_J$ which determines the scales of thermal fluctuations of the Gaussian period distributions. At the denaturation or melting temperature $T_{\mathrm{denat}}$, virtually every helix in the system must develop a consensus bubble between two successive evaluations of the LTM. If the deviation $\sigma$ is ''too large,'' the period distribution in the helix is so broad that consensus cannot be preserved: the range of fluctuation times, of the order of $\sigma$, is so broad that a helix consensus is unlikely to survive two consecutive readings. From our empirical estimate of the denaturation dispersion fixed at $\sigma=10^{-8}$ s, and the typical experimental $T_{\mathrm{denat}}=313\,^{\circ}$K for proteins such as the ones studied in this work,[33] we get a dispersion for class II of $g_{\mathrm{II}}=3.2\times 10^{-19}\,\mathrm{s}^2/{}^{\circ}\mathrm{K}$.

### E. The time scale renormalization operation

As stated above, the renormalization or rescaling of the time scale is the procedure that signifies establishment of long-range correlations. It plays two concurrent roles: (1) by changing the rates of change of spatial relations, with the identification of each new CP, it redefines, implicitly, the topography of the potential surface in the way it characterizes interactions between residues, and (2) it puts a fresh set of constraints on the generation of new LTMs as soon as a new CP transition has been recorded. The renormalization is accomplished via the feedback loop (Fig. 1), establishing the latest CP as input for the reclassification of residues.

Complex structural patterns such as those presented in Fig. 3 do not result from single-step, all-or-none processes, although several residues may be involved in passage of the system over a single saddle, if the effective range of interactions is long.[37] More probably, at least one nucleating event corresponding to a sizeable downward step on the potential energy surface and often involving a large reduction in conformational entropy takes place early in the folding history; in effect, these changes in this ''big step'' define it as ''first'' in a sequence of structure-seeking steps. It has been suggested that the succeeding steps in that folding series are likely to follow a path that minimizes additional reductions in conformational entropy[9,10,16,19] until, for example, a new structural motif such as a new loop is required for further folding. Then another big step is required.

Once formed, a nucleated consensus pattern can be expected to persist for ~1000 evaluations of the LTM, since the evaluation frequency is of the order of 64 ps at 298 °K [cf. Eq. (3)] the typical working temperature for *in vitro* folding of BPTI. This timespan allows the original consensus window to act as a seed and grow by progressive torsional isomerizations of residues adjacent to those in the nucleating pattern. When adjacent residues achieve consensus compatible with the structural motif of the original seed, they also lock into the $10^7\ \mathrm{s}^{-1}$ frequency domain and thus contribute to the propagation of the growth process.

In summary, the key feature of renormalization which enables us to deal with cooperativity in the formation of secondary structure is the survival of a nucleating pattern achieved by the reclassification of class I residues into class II residues once the nucleating pattern occurs. The argument holds *mutatis–mutandis* for tertiary interactions formed co-

operatively from secondary structure. In this case, the reclassification of class II residues as class III residues translates into a drastic drift in the basin transition frequency, from $10^7$ to $10^3\ s^{-1}$, thus assuring the survival of the respective nucleation pattern by further stabilization.

## III. FRUSTRATION TOLERANCE IN THE FOLDING PROCESS

We denote by $L=L(i,j)$ the number of residues that must be placed in the correct Ramachandran basin so that the appropriate CP forms to bring about the transition $i\rightarrow j$. According to our model, $L$ is the minimum length of the consensus window that must be identified by the $\pi$ operation in order to yield the CP transition. Zwanzig, Szabo, and Bagchi estimated that the mean first passage time $t(i,j)$ to reach this consensus is[3]

$$t(i,j)=f^{-1}\times 2^L/L. \tag{2}$$

Equation (2) assumes an equal frequency of interconversion into and from the correct local basin. The mean interbasin transition frequency depends on whether the residues engaged in the CP transition are free in CP $i$ (where $f=10^{11}\ s^{-1}$), engaged in secondary structure which is part of CP $i$ ($f=10^7\ s^{-1}$), or engaged in tertiary structure ($f=10^3\ s^{-1}$).

The result given in Eq. (2) is independent of the initial number of correctly conformed units.[3] However, a shorter first-passage time than that of Eq. (2) for the CP transition might be expected because the long-range potential may drag the participating residues into the correct basin, much in the same way the forces envisioned for a frustration funnel can help the self-organization to occur faster than by random searching.[29] Ultimately, since it is generally supposed that the terminus of the folding process is at least a deep local minimum in free energy, the entropic cost attached to reaching that state must be more than compensated by the enthalpy lost by contact formation. This prompts us to introduce a torsional frustration tolerance, to simulate and allow steps along the folding pathway in which the drop of either the enthalpy or the entropy need not be a local maximum. In other words, a partial lack of consensus in the LTM indicating frustrated or imperfect conformity with local constraints may be overlooked with a certain probability by the pattern recognizer. To carry this out, we define a frustration-dependent input parameter, $z(i,j,F)$, for the CP transition, thus

$$z(i,j,F)=\mathrm{Arctanh}\{2[L(i,j)-F]/L(i,j)-1\}, \tag{3}$$

where $F=0,1,...,L(i,j)$ is the number of residues not in the target basin. Thus, $z(i,j)=\infty$ in the perfectly structured case, so $F=0$, while $z(i,j)=-\infty$ if there are no structural elements in place whatever, so $F=L(i,j)$. Then the probability $p(i,j,F)$, of activating the CP $j$ after having read the LTM in CP $i$, which is just the $i\rightarrow j$ transition probability, becomes

$$p(i,j,F)=[1+\exp(-z(i,j,F))]^{-1}. \tag{4}$$

Equation (4) is the canonical logistic activation function of parallel processing,[31] satisfying the limiting conditions

$$p(i,j,F)=1 \ \text{ if } \ F=0, \quad p(i,j,F)=0, \quad \text{if } \ F=L(i,j). \tag{5}$$

The value of the input variable, a measure of the degree of local correctness, is chosen to fulfill these conditions. Thus, we may accept the CP transition with a lower barrier $B(F)=RT\ln 2^{L(i,j)-F}$ than the zero-frustration activation barrier $B(0)=RT\ln 2^{L(i,j)}$. However, the former barrier is less probable according to Eqs. (3) and (4). This plasticity in the pattern recognition is meant to reflect the probable physics of the folding process on the PES, which presumably first tolerates and then corrects some degree of torsional incongruities. Nonzero frustration barriers become more probable as the sequence $L(i,j)$ gets longer. This is intuitively obvious, since larger structures can better accomodate torsional distortions, with their greater latitude of torsional freedom to correct any deviations from their correct Ramachandran basins.

The model needs yet another form of plasticity reflecting tolerance to side chain mismatches in putative contacts. This is the tolerance to classical frustration which only occurs when the thermodynamic cost of the $h/p/n$ mismatch is small enough to make the contact occur. Under such conditions, the probability of the CP transition is also dependent on $X$, the number of contact mismatches

$$p(i,j,F,X)=p(i,j,F)\times g(i,j,X), \tag{6}$$

where $g(i,j,X)$ represents the tolerance to $h/p/n$ mistakes in the putative contacts to be formed if the $i\rightarrow j$ transition occurs. Thus, if $M(i,j)$ is the total number of contacts to be made in the transition, we get

$$g(i,j,X)=\{1+\exp[-\mathrm{arctanh}\{2[M(i,j)-X]/M(i,j)-1\}]\}^{-1}. \tag{7}$$

Thus, the statistical weight of recognizing $X$ h/p/n mismatches in the pattern ranges from zero if $X=M(i,j)$ to 1 if $X=0$. On the other hand, as Eqs. (3)–(7) reveal, the respective tolerances to torsional incongruities or mismatches gets larger with the numbers of correct basins and putative contacts required in the zero-frustration (or perfectly-matched) state that establishes the CP $j$.

## IV. THE CM-PATHWAY FOR BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI): A TOLERANT PARALLEL COMPUTATION

The aim of this section is predicting at least one major folding pathway and the structural features of folded bovine pancreatic trypsin inhibitor (BPTI) via a coarse mechanistic analysis of its torsional long-time dynamics at 298 °K.[7,32,34–36] The folding process of this protein was first studied experimentally by Creighton[38,39] and theoretically by Camacho and Thirumalai,[40] who review later experimental work. Their theoretical approach was based on the successive formation of S–S bonds, to interpret the information obtained in the prior experimental studies. We focus on identifying significant structural patterns that correspond to formation of intermediates and on the late kinetic bottlenecks the system encounters in the first $10^{-2}$ s of the folding process, a timespan that requires $10^7(\pi-\rho)$ cycles. We show

how a dominant sequence of CM transitions for the BPTI in a frustration-tolerant simulation describes the dominant features of the folding kinetics, reproducing the essential cooperative aspects of the folding pathways thus far revealed by experiments. In particular, this includes the late stages in which tertiary interactions direct and stabilize the native Cys–Cys (5,55) contact.[7,25,27,29] The conditions in solutions where refolding occurs are reducing, making the Cys–Cys bonds labile enough that they do not lock the dynamics of formation of secondary structures. Consequently the kinetics of formation, dismantling and recombination of intrachain Cys–Cys disulfide bonds falls within the time scale[7] of $10^{-7}$ s,[29] and does not interfere with the folding process. This is incorporated in the model by taking into account that the fastest Cys5–Cys30 pairing takes $4\times10^{-5}$ s to form, as shown below. In other words, the sulfur–sulfur links form at the level of tertiary structure.

In order to analyze the kinetics along the dominant folding pathway, we adopt a precise operational definition of contact based on a proximity of 7 Å, the maximum distance for a significant decrease $(1/2kT)$ in the long-range potential, in consonance with previous treatments of contact pattern formation.[16,19,34]

How does cooperativity operate in the folding of BPTI? To answer this question we first examine the estimated mean times [Eqs. (8)–(10)] to form native Cys–Cys disulfide contacts if we assume a zero-frustration transition state:

$$\tau(5,55)=(2^{50}/50)\times10^{-11}\,\mathrm{s}\approx2\times10^{2}\,\mathrm{s},$$
$$\tau(30,51)\approx1.7\times10^{-3}\,\mathrm{s}. \tag{8}$$

Throughout this section, the numbers $i,j$ in parentheses, as in the $\tau(i,j)$ of Eq. (8), will denote the positions of residues along the sequence of the chain; if we allow nonzero-frustration, an $F$ value written within the same parentheses will specify the level of frustration tolerated in the formation of the contact.

The first relation of Eq. (8) implies that the (5, 55) native contact takes a long time to form even if some torsionally ''incorrect'' states are allowed to activate the Cys–Cys contact. For example, $\tau(5,55,F=5)=1.4\times10^{2}$ s. Obviously, as indicated in Sec. III, the larger the loop, the more tolerant its closure becomes to torsional incongruities. However, the (5, 55) contact requires closure of such a large loop that we can expect that forming it requires cooperativity within the timescales under investigation, as appears to be the case.[7,25,27,29] On the other hand, the second contact may form readily from less probable nonzero-frustration states: $\tau(30,51,F=3)\approx10^{-3}$ s. Also, the third native Cys–Cys contact (14,38) may form directly within time scales commensurate with the occurrence of tertiary contacts

$$\tau(14,38,F=5)\approx1.6\times10^{-4}\,\mathrm{s},$$
$$\tau(14,38,F=3)\approx1.3\times10^{-3}\,\mathrm{s}. \tag{9}$$

The sequence of CMs generated within the timespan of $10^{7}(\pi-\rho)$ cycles is consistent with the previous analysis. Four revealing images of the time evolution of the CM have been constructed by averaging 17 runs of the parallel computation; these are displayed in Figs. 5(a)–5(d). The results show steps along the (empirically) most probable type of pathway which has been reproduced in 14 of the 17 runs. All runs yielded virtually identical results, with a variance of occurrence of 1 ps for all significant kinetic bottlenecks of the folding process. The corresponding snapshot times averaged over all 17 runs are, respectively, $3.2\times10^{-4}$ s, $1.3\times10^{-3}$ s, $1.3\times10^{-3}\,\mathrm{s}+3.2\times10^{-7}$ s (where the third snapshot is taken only 311 LTM readings after the second), and $1.3\times10^{-3}\,\mathrm{s}+3.2\times10^{-7}\,\mathrm{s}+0.5\times10^{-2}$ s. The sudden transformation taking place between Figs. 5(b) and 5(c) is reminiscent of the ''proteinquakes'' described by Frauenfelder *et al.*,[41] although the process described here involves a more global change of structure than the proteinquakes of Ref. 41.

To understand the role of frustration tolerance we have also run 17 simulations of the stiff, totally intolerant version of the algorithm. In this case the structure development does not go beyond the CM displayed in snapshot of Fig. 5(b), which reveals the non-native disulfide bond (5,30), as well as part of a complex $\beta$-sheet motif of the type given in Fig. 4, and some helix structure. At this point, the intolerance to contact mismatches prevents the formation of further secondary structure and also prevents the formation of tertiary interactions which would stabilize the secondary structure already formed. The $\beta$-sheet complex motif is not completed, and even the incipient seeding motif is not stabilized, since that would require tolerance to mismatches in the tertiary contact as well as in the rest of the secondary structure motif and, as well, a torsional tolerance in the closure of the tertiary loops. The level of organization displayed in Fig. 5(b) is preserved through nearly $10^{3}$ further iterations after which kernels of structure destruction are formed, and the structure recedes back to the CM displayed in Fig. 5(a). Thus, the two CPs of Figs. 5(a) and 5(b) are kinetically related, presumably metastable, possibly analogous to the coexisting phaselike forms seen previously for protein models,[42] and frustration intolerance would force the system to oscillate between them without ever reaching the active folded form, as no further structural development is kinetically feasible.

Nucleation windows of the form shown in Fig. 3(b) that seed the formation of an $\alpha$-helix appear in the (43–58)-section of the molecule within the interval of $8.8\times10^{-4}$ to $9.8\times10^{-4}$ s. The timescale given in Eq. (9) for formation of (14,38) is also a good estimate. The image taken at $1.3\times10^{-3}$ s [Fig. 5(b)] displays this native contact, as well as fully and partially developed secondary structure elements such as the $\alpha$-helix and a two-strand portion involving the contour region (20–33) of the $\beta$-sheet topology represented in Fig. 4. The nucleating events leading to the two-strand portion of the $\beta$-sheet topology take place in the (20–33)-region of the chain within the same time interval as those triggering the formation of the helix. These motifs dismantle unless tertiary structure develops within their lifetime of $\sim10^{3}$ cycles, which can only happen in the tolerant version of our simulation model.

Tertiary contacts between the $\alpha$-helix and the complex $\beta$-sheet require the closure of the loop in contour region (33–43); this starts developing between the still-incomplete $\beta$-sheet and the helix just 311 LTM evaluations after the time the image in Fig. 5(b) was taken. This coincides pre-

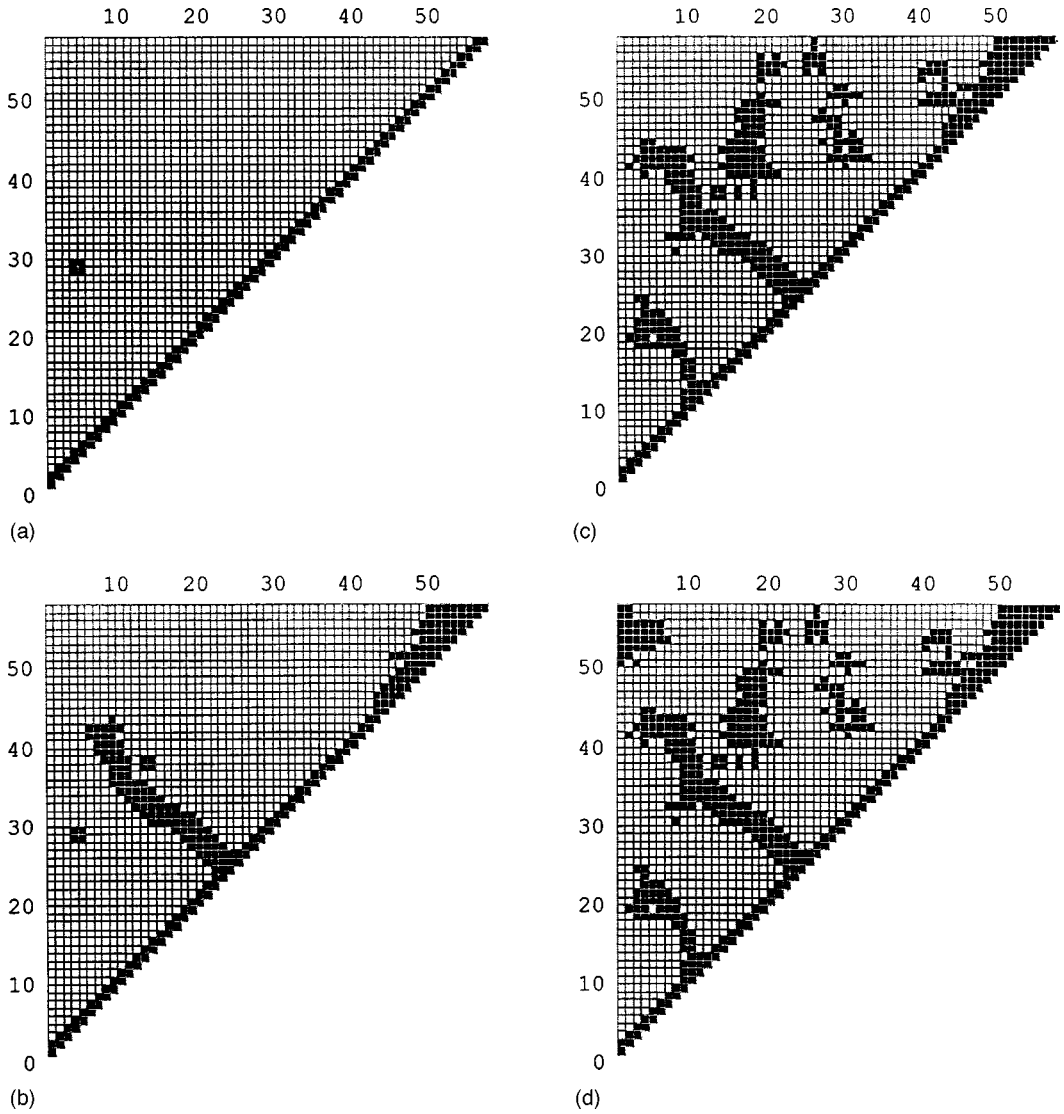


FIG. 5. Four successive contact maps of the CM evolution obtained by a frustration-tolerant simulation, with the axes denoting the amino acid residues in sequence and the filled, off-diagonal squares indicating the residues in contact at the time of the image; the images were taken, respectively, at (a) $3.2\times10^{-4}$ s, (b) $1.3\times10^{-3}$ s, (c) $1.3\times10^{-3}$ s$+3.2\times10^{-7}$ s, and (d) $1.3\times10^{-3}$ s$+3.2\times10^{-7}$ s$+0.5\times10^{-2}$ s in the course of $10^7$ parallel computation cycles for the BPTI. An $\alpha$-helical segment appears as a shaded region parallel to the main diagonal; a $\beta$-sheet, as a shaded segment perpendicular to it, and looped and tertiary contacts as connected and simply connected shaded regions.

cisely with the time estimation, $3.2\times10^{-7}$ s, for closure of the 10-loop within the (33–43) region. We emphasize that this is a renormalized calculation that assumes the previous formation of secondary structure. Nonetheless this result implies that formation of tertiary structure may begin as a very fast, ''proteinquake'' process, triggered by formation of a crucial secondary structure, in this instance the (33–43) loop.

At this time, the non-native (5,30) Cys–Cys bond is completely dismantled and is replaced by the native (30–51) contact induced and stabilized by the tertiary interaction, as the third image, Fig. 5(c), reveals. This pathway displays how the formation of (30,51) is expedited by cooperative folding, in agreement with recent findings.[7] Furthermore, during the development of the first tertiary contact, the complex $\beta$-sheet motif continues to grow, fostering other tertiary interactions between the two dominant secondary structure elements. Since folding and unfolding of tertiary structure occurs on the fast NMR time scale of $10^{-3}$ s, the locking of Class III structures to move below the $10^3\,\mathrm{s}^{-1}$ frequency peak enables the survival of the initial tertiary consensus while contact (30–51) forms and the $\beta$-sheet is completed. It could not have been completed unless there was frustration tolerance in the folding process enabling the initially weak $\alpha$-helix—$\beta$-sheet tertiary contact to form, thus allowing the completion and further engagement of the full three-strand $\beta$-sheet motif which ultimately interacts with the $\alpha$-helix.

Finally, the (5,55) disulfide bond that would initially take a prohibitively long time to form, now entails the closure of a complex 29-residue loop with no polar orientation requirements. This loop consists of the quasi-coil (5–20) region, the $\beta$-sheet (20–30) region, and the quasi-coil strand (51–55). Notice that the formation of the native (30–51)-contact short-circuits the loop closure for the (5,55) contact, so that the estimated time for the cooperative formation of this interaction is

$$\tau(5,55)\approx 1.1\times 10^{-2}\ \mathrm{s}, \quad \tau(5,55,F=3)\approx 5\times 10^{-3}\ \mathrm{s}. \tag{10}$$

This estimate of the rate-determining step in BPTI folding is corroborated by the fourth snapshot [Fig. 5(d)], and confirms previous estimates,[7,26,32,34–36] in the sense that contact (30,51) occurs $10^5$-fold more rapidly than (5,55).

Thus, the long-time dynamics obtained by means of our semiempirical microscopic model not only predicts with good accuracy what tertiary structural elements form in BPTI but it also models the kinetics of their formation and shows how cooperative effects expedite the formation of native interactions shaping the hydrophobic core. The kinetics stand in good agreement with experimental kinetic studies.[7,26,32,34–36]

Another demonstration of the predictive potential of the model is the way the CM for the predicted active folding of BPTI [Fig. 5(d)] contains all meaningful structural elements already identified in the CM obtained from x-ray crystallographic data.[26] The predicted and experimental CMs exhibit the same functionally significant structural elements. A finer-grained comparative analysis would not be appropriate in the present context since such an analysis would demand too much precision—and hence inflexibility—in the definition of contact. However, the literature adopts (cf. Ref. 30) CMs with an *arbitrary* maximum distance, typically 5–7 Å. Not surprisingly, the CM using a more restrictive definition of contact is sparser than the one adopted in this work. However, our CM identifies *all functionally significant structural elements and no other*, in agreement with the experimental CM.

This model, like earlier theoretical treatments,[43,44] finds optimal pathways that involve transient intermediate structures that form, assist the folding process and then dismantle in order to form other structures closer to the native.

Recent experimental evidence indicates that BPTI[45] and other systems[46] achieve expedient folding by a search for a topology that nucleates the center from which the hydrophobic collapse progresses. This scenario is not necessarily compatible with the hierarchical model put forth by Rose and co-workers.[47] Our model presented in this paper addresses the problem by giving precise, quantitative meaning to the kind of order associated with specific topologies, in the folding context. For the search in conformational space to be efficient, expedient and robust, there must be a suitable separation of time scales: Equilibration times within Ramachandran basins must be very much shorter than time scales for interbasin transitions.[48] This enables us to use a coarse-grained approach that one can consider analogous to transition state theory in chemical kinetics, where one treats fast vibrations as equilibrated enough to allow the use of regional partition functions. Here we treat the structural motifs as patterns of topologically compatible ''lumped'' torsional states classified simply by their Ramachandran basins. Within this coarse-grained picture, long-range organization appears as patterns are identified. As in Ref. 40, there is no need for explicit reference to forces between nonbonded residues; there, the authors looked on that as a limitation of their model, while we see it here as a strength because it allows us to avoid having to obtain information we would use only in intermediate stages of the analysis. The admittedly *ad hoc* renormalization operation in the method presented here ensures, by slowing their rates, that establishment of regions with long-range order stabilize with respect to torsional transitions. Thus, while the hierarchical model of Rose *et al.*[47] emphasizes a local bias in a global search, our model allows the folding protein to make use of both local and longer-range interactions.

Three generic facts support this approach:

(a) the Ramachandran maps maintain their topological invariance throughout the folding process; no basins appear, disappear or fuse.

(b) A hierarchical scenario would require that the local secondary structure survives long enough, approximately 1 $\mu$s, to bias further long-range assembly. This could well be inconsistent with results of recent experiments based on kinetic probes[46,47] that show local secondary structure is ephemeral unless scaffolded by tertiary interactions.

(c) Misfoldings cannot be corrected easily in a structure of increasing complexity that is carried out at the local level of geometry, whereas a rough search at a more coarse-grained topological level of approximate regional structural integrity, especially with a nonzero tolerance, is far more readily self-correcting.[47]

A brief description of this approach has been published recently.[49] An extension, to begin to link the pattern-recognition topological approach described here, to the topographical approach of Refs. 20 and 37, follows in the accompanying paper.[50]

## V. CONCLUSIONS

This work presents a model for pattern formation and self-organization of a chain, particularly but not exclusively a protein, with a means to incorporate in a consistent fashion some elementary features that intuition suggests ought to be demanded from moderately realistic caricatures of a folding protein, and a demonstration by application to the folding kinetics of BPTI of how the method provides a quantitative representation of the folding process. The elements of the presentation have been:

(1) A way to account for the essentially parallel, simultaneous nature of the protein's exploration of its conformation space, whereby uncorrelated regions of the peptide chain are able to search independently for concurrent folding possibilities—i.e., for formation of multiple nucleation sites.

(2) A means of bridging the timescale gap that separates the relatively fast torsional dynamics from slower folding events leading to secondary and tertiary structure. This is achieved by simplification via coarsening of torsional conformation space and of time scales.

(3) A means of incorporating the plasticity or error tolerance of the folding process with respect to torsional incompatibilities and contact mismatches during the formation of specific structural motifs.

(4) A means of rationalizing how this frustration tolerance may account for the inherent robustness and expediency in the folding of a natural protein.

(5) The application to bovine pancreatic trypsin inhibitor.

## ACKNOWLEDGMENTS

We would like to thank Dr. Konstantin Kostov for many stimulating discussions about this work. We would also like to thank Professor Tobin Sosnick for his helpful comments. A.F. wishes to thank the J. William Fulbright Foreign Scholarship Board for the Grant that made this collaboration possible. R.S.B. would like to acknowledge the support of the National Science Foundation for his part in the work.

[1] R. Jaenicke, in *Protein Structure and Protein Engineering*, edited by E. L. Winnacker and R. Huber (Springer-Verlag, Berlin, 1988), p. 16.
[2] O. B. Ptitsyn and B. V. Semisotnov, in *Conformations and Forces in Protein Folding*, edited by B. Nall, B. Dill, and K. Dill (American Association for the Advancement of Science, Washington, 1991), p. 155.
[3] R. Zwanzig, A. Szabo, and B. Bagchi, Proc. Natl. Acad. Sci. USA **89**, 20 (1992).
[4] R. L. Baldwin, Proc. Natl. Acad. Sci. USA **93**, 2627 (1996).
[5] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[6] J. Bohr, H. Bohr, and S. Brunak, Europhys. News **27**, 50 (1996).
[7] T. E. Creighton, N. J. Darby, and J. Kemmink, FASEB J. **10**, 110 (1996).
[8] M. Jamin and R. L. Baldwin, Nat. Struct. Biol. **3**, 613 (1996).
[9] A. Fernández and G. Appingnanesi, Phys. Rev. Lett. **78**, 2668 (1997).
[10] K. Dill, K. M. Fiebig, and H. S. Chan, Proc. Natl. Acad. Sci. USA **90**, 1942 (1993).
[11] N. Go and H. A. Scheraga, J. Chem. Phys. **51**, 4751 (1969).
[12] N. Go and H. A. Scheraga, Macromolecules **9**, 535 (1976).
[13] H. Cendra, A. Fernández, and W. Reartes, J. Math. Chem. **19**, 331 (1996).
[14] S. He and H. A. Scheraga, J. Chem. Phys. **108**, 287 (1998).
[15] S. He and H. A. Scheraga, J. Chem. Phys. **108**, 271 (1998).
[16] A. Fernández, J. Stat. Phys. **92**, 237 (1998).
[17] Z. Guo and D. Thirumalai, Biopolymers **36**, 83 (1995).
[18] M. Guenza and K. F. Freed, J. Chem. Phys. **105**, 3823 (1996).
[19] A. Fernández and A. Colubri, Physica A **248**, 336 (1998).
[20] R. E. Kunz and R. S. Berry, J. Chem. Phys. **103**, 1904 (1995).
[21] K. D. Ball and R. S. Berry, J. Chem. Phys. **109**, 8541 (1998).
[22] K. D. Ball and R. S. Berry, J. Chem. Phys. **109**, 8557 (1998).
[23] K. D. Ball and R. S. Berry, J. Chem. Phys. **111**, 2060 (1999).
[24] C. Cantor and P. Schimmel, *Biophysical Chemistry* (W. H. Freeman, New York, 1980).
[25] R. F. Gesteland and J. F. Atkins, *The RNA World* (Cold Spring Harbor Press, New York, 1993).
[26] T. E. Creighton, *Proteins* (Freeman and Co., New York, 1993).
[27] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).
[28] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[29] R. Zwanzig, Proc. Natl. Acad. Sci. USA **92**, 9801 (1995).
[30] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).
[31] D. E. Rumelhart, J. C. McClelland and t. P. R. Group, *Parallel Distributed Processing* (MIT Press, Cambridge, 1988).
[32] C. L. Brooks, M. Petitt, and M. Karplus, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics* (Wiley, New York, 1988).
[33] J. S. Richardson and D. C. Richardson, in *Protein Folding*, edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), p. 5.
[34] T. G. Oas and P. S. Kim, in *Protein Folding*, edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), p. 123.
[35] D. Bashford, M. Karplus, and D. Weaver, in *Protein Folding*, edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), p. 283.
[36] T. E. Creighton, in *Protein Folding*, edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), p. 157.
[37] K. D. Ball, R. S. Berry, A. Proykova, R. E. Kunz, and D. J. Wales, Science **271**, 963 (1996); see also R. S. Berry, N. Elmaci, J. P. Rose, and B. Vekhter, Proc. Natl. Acad. Sci. USA **94**, 9520 (1997), for application of this method to a protein model.
[38] T. E. Creighton, J. Mol. Biol. **113**, 275 (1977).
[39] T. E. Creighton, Prog. Biophys. Mol. Biol. **33**, 231 (1978).
[40] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA **92**, 1277 (1995).
[41] A. Ansari, J. Berendzen, S. F. Browne, H. Frauenfelder, I. E. T. Iben, T. B. Sauke, E. Shyamsunder, and R. D. Young, Proc. Natl. Acad. Sci. USA **82**, 5000 (1985).
[42] B. Vekhter and R. S. Berry, J. Chem. Phys. **110**, 2195 (1999).
[43] J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. USA **87**, 3526 (1990).
[44] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[45] P. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, J. Appl. Crystallogr. **26**, 283 (1993).
[46] M. Dadlez, Biochem. J. **36**, 2788 (1997).
[47] T. R. Sosnick, L. Mayne, and S. W. Englander, Proteins: Struct., Funct., Genet. **24**, 413 (1996).
[48] R. L. Baldwin and G. Rose, Trends Biol. Sci. **24**, 26 (1999).
[49] A. Fernández, K. Kostov, and R. S. Berry, Proc. Natl. Acad. Sci. USA **96**, 12991 (1999).
[50] A. Fernández, K. Kostov, and R. S. Berry, J. Chem. Phys. **112**, 5223 (2000) (following paper).