Folding a protein by discretizing its backbone torsional dynamics

Ariel Fernández
Instituto de Matemática (INMABB), Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Sur,
Avenida Alem 1253, Bahía Blanca 8000, Argentina

(Received 26 June 1998; revised manuscript received 12 November 1998)

The aim of this work is to provide a coarse codification of local conformational constraints associated with each folding motif of a peptide chain in order to obtain a rough solution to the protein folding problem. This is accomplished by implementing a discretized version of the soft-mode dynamics on a personal computer (PC). Our algorithm mimics a parallel process as it evaluates concurrent folding possibilities by pattern recognition. It may be implemented in a PC as a sequence of perturbation-translation-renormalization $(p\text{-}t\text{-}r)$ cycles performed on a matrix of local topological constraints (LTM). This requires suitable representational tools and a periodic quenching of the dynamics required for renormalization. We introduce a description of the peptide chain based on a local discrete variable the values of which label the basins of attraction of the Ramachandran map for each residue. Thus, the local variable indicates the basin in which the torsional coordinates of each residue lie at a given time. In addition, a coding of local topological constraints associated with each secondary and tertiary structural motif is introduced. Our treatment enables us to adopt a computation time step of 81 ps, a value far larger than hydrodynamic drag time scales. Folding pathways are resolved as transitions between patterns of locally encoded structural signals that change within the 10 μs–100 ms time scale range. These coarse folding pathways are generated by the periodic search for structural patterns in the time-evolving LTM. Each pattern is recorded as a contact matrix, an operation subject to a renormalization feedback loop. The validity of our approach is tested *vis-a-vis* experimentally-probed folding pathways eventually generating tertiary interactions in proteins which recover their active structure under *in vitro* renaturation conditions. As an illustration, we focus on determining significant folding intermediates and late kinetic bottlenecks that occur within the first 10 ms of the bovine pancreatic trypsin inhibitor renaturation process. The probed cooperativity and nucleation effects, as well as diffusion-collision stabilization of secondary structure are shown to result from the persistence of relatively stable patterns through successive $(p\text{-}t\text{-}r)$ cycles, thus acting as seeding patterns for further growth or hierarchical development. [S1063-651X(99)10504-X]

PACS number(s): 87.15.He, 87.10.+e, 87.15.Nn

I. INTRODUCTION AND OUTLINE OF THE WORK

This work deals with a coarse computation of the long-time limit in the dihedral torsional dynamics of a peptide chain and its bearing on the folding process of a natural protein. This limit is studied by introducing a discrete codification of the set of local conformational constraints required to form each folding motif. We aim for a rough theoretical understanding of the microscopic expedient by which natural proteins fold into their active conformation under *in vitro* renaturation conditions. Our approach is based on a discretization of the torsional states of the chain taking into account the dominant rotameric forms (torsional isomers) for each residue or unit. We focus on proteins that recover their activity as a result of the actual dominance of folding pathways [1–8] under appropriate renaturation conditions. The discretized torsional dynamics makes it necessary to cast the folding process as the evolution of patterns of locally codified signals, the recognition of which is an inherently parallel operation in which patterns are translated and registered on a contact matrix (CM).

The vast gap between the time scales accessible to molecular dynamics computations, typically in the range 10^{-12}–10^{-9} s, and those inherent to transitions between contact patterns (CP's) formed by the chain, typically in the range $10\,\mu$s–10^{2} s, suggests the need for a *semiempirical* treatment to elucidate the microscopic origin of the *expediency* of the folding process [9,10]. Recently, a considerably large computation time step of the order of 10–15 ps has been reached by taking into account the adiabatic ansatz [11] leading to the elimination of hard modes (stretching vibrations and angular planar vibrations) and restricting the dynamics to the soft-mode manifold of dihedral torsional motions [12]. Thus, computations taking into account the geometry of the soft-mode manifold have been actually carried out [13]. Such computations have made the microsecond timescale accessible, thus elucidating the nature of early folding events. We must go one step further in our simplification of the chain dynamics: The 1–10 ms time scale must be made accessible in order to encompass folding events relevant to the global hydrophobic collapse, formation of the molten globule [2], and other meaningful kinetic bottlenecks [4,7,8]. This is precisely the aim of this work.

In a preliminary step towards implementing the required semiempirical treatment on a computer, a basic representational tool, the local topological constraint matrix (LTM), is introduced. This matrix coarsely describes the torsional state of the each residue or unit of the chain in a particular conformation. This is done by indicating the basin of attraction where the local torsional state of the chain belongs at a given time. Each basin represents a distinctive rotamer or torsional isomeric state for a single unit in the chain, and there are only 2–4 such basins per residue. This fact considerably simplifies the dynamical analysis of the folding problem. In other words, conformations are locally viewed modulo basins of attraction within which equilibration is incommensurably fast compared with folding time scales [3].

The previous argument suggests that a rough computation of the folding process requires three basic operations: (a) an operation prescribing time-dependent rules for the evolution of the LTM; (b) the periodic evaluation of the LTM in search of concurrent folding possibilities; and (c) an operation linking the latest LTM evaluation with the generation of new LTM's according to (a). Thus, an actual computation would require a parallel algorithm. Accordingly, this algorithm is described to within an implementation level in Secs. II–IX. Section II introduces the representational tools required to define the LTM. Sections III and IV describe the entire outline of the basic operations (a)–(c) given above. Sections V–VII give the rules that determine the time-evolution of the LTM, while Secs. VIII and IX describe how the algorithm actually works within a parallel architecture, reproducing basic features of the folding process.

The rest of the work is devoted to an actual implementation of the inherently parallel algorithm described in Secs. II–IX on a personal computer (PC), that is, on a sequential machine within a conventional architecture. The PC implementation makes our results reproducible. It requires, in turn, three basic consecutive operations that represent the sequential or conventional counterpart of the parallel operations (a)–(c) given above: *Perturbing* the LTM, *reading or translating* it into a CM, and *renormalizing* the chain in a way that affects the way in which new perturbations of the LTM are applied. The LTM must be occasionally frozen in the sequential PC in order to perform the last two operations. All three operations naturally form a feedback loop or cycle. In turn, the loop is iterated to generate the coarse torsional dynamics. Formally speaking, we identify a favored folding pathway by iterating a perturbation-translation-renormalization $(p\text{-}t\text{-}r)$ cycle.

The perturbation operation itself is implemented so that it coarsely determines the dynamic flow of the system by its iterative application on the LTM. The translation operation is inherently a pattern recognition step and as such searches for folding patterns in the LTM. Each folding pattern is recorded in a CM if and only if all local torsional constraints required to form the pattern are fulfilled and the putative interacting contour regions are matched in terms of their hydrophobicity. Thus, *the fulfillment of torsional constraints emerges as a consensus window in the LTM*. Finally, the renormalization operation redefines the chain contour distances and torsional time scales *vis-a-vis* the latest CM translated. This is done in such a way that the torsional degrees of freedom engaged in the latest CM recorded have a longer evolving time scale than free residues. Thus, preexisting structure is rightly regarded as a (metastable) seeding pattern or kernel upon which structure-growth steps may take place in accord with the nucleation scenarios for protein structure formation [4,5].

The sequential implementation on the PC of the inherently parallel folding algorithm following the tenets given above is expounded in detail in Secs. X and XI of this work. In turn, Sec. XII is devoted to an illustration of our approach by elucidating the favored folding pathway for the protein bovine pancreatic trypsin inhibitor (BPTI), from the coil conformation to its biologically competent folding. This illustration reveals the microscopic features of torsional dynamics that are responsible for the expediency and robustness of the protein renaturation process under realistic physiological conditions.

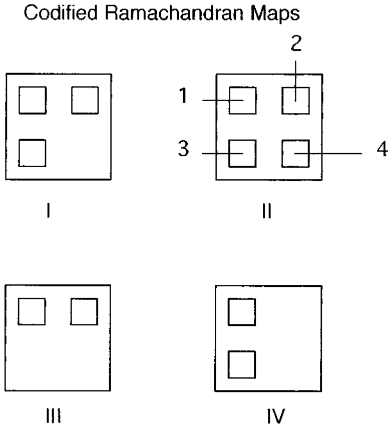

FIG. 1. Discrete codification of local torsional states of aminoacids (residues) by indicating the basin (1, 2, 3, or 4) in the Ramachandran map where the torsional coordinates Φ,Ψ lie. There are four types of maps I–IV, depending on whether the residue is alanyl-like (I), glycine (II), precedes a proline (III), or proline (IV). Thus, a Ramachandran discrete variable $R(\mathbf{y},n)=1,2,3,4$, indicates the basin for the nth residue in the conformation roughly defined by the LTM $\mathbf{y}$.

The layout of this work, although intricate, is a natural one. Many concurrent folding possibilities may occur in a flexible peptide chain at a given time and thus the analysis must be, in principle, parallel, as described in Secs. II–IX. However, to make our results reproducible, we must present them on a conventional computer architecture suitable for a PC. This is done in Secs. X–XI and illustrated in Sec. XII.

II. DISCRETIZED CODIFICATION OF LOCAL TORSIONAL STATES OF THE PEPTIDE CHAIN

To specify the context referred to above, we shall coarsely define the conformation manifold to determine a discrete version of the soft-mode dynamics of the peptide chain. To reach this goal, we shall codify the local torsional state of a residue according to the basin of attraction where its two torsional dihedral coordinates Φ,Ψ lie within a local potential energy surface, the so-called Ramachandran map [14], as indicated in Fig. 1. In essence, the Ramachandran map governs the local torsional dynamics of a single aminoacid residue of the peptide chain. Such dynamics is not correlated to those of nearest-neighbor residues due to the torsional rigidity of the backbone bond in between residues [14]. These considerations lead us to naturally define a ''Ramachandran variable,'' denoted $R(\mathbf{y},n)$, indicating the basin of attraction of residue n in the conformation coarsely defined by the LTM $\mathbf{y}$.

In our codification of the local torsional dynamics we classify residues or aminoacids as follows: L-alanyl-like, glycine, proline, and any residue preceding proline (Fig. 1). Thus, since an alanyl-like residue with contour position n has three basins of attraction [14], we would get three possible

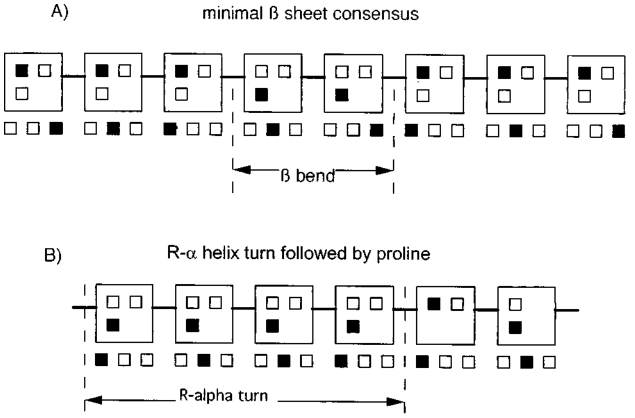

FIG. 2. (a) LTM consensus window for a minimal β-sheet structural motif. (b) LTM consensus window for a right-handed α-helix turn interrupted by a proline (and a residue preceding proline).

values depending on $\mathbf{y}$: $R(\mathbf{y},n)=1,2,3$, while if glycine is at the nth position, we would get $R(\mathbf{y},n)=1,2,3,4$, again depending on $\mathbf{y}$. On the other hand, if proline happens to be the nth residue, we would get $R(\mathbf{y},n)=1,3$, while if the nth residue precedes proline, we would obtain $R(\mathbf{y},n)=1,2$. This codification is consistent with the existence of local torsional isomers coarsely represented as basins of attraction in the Ramachandran plots [14]. Generically speaking, we intend to solve the folding problem by translating and renormalizing evolving patterns of locally encoded structural signals. Our goal is to provide a consistent dynamical picture in which long-range intrachain interactions are determined by a set of local torsional constraints.

As shown in the following sections, the shortest mean time for a change in the patterns is of the order of 81 ps, a value far larger than the hydrodynamic drag time scale of 15 ps adopted in the continuum soft mode analysis [13,15]. This considerably longer time step for the evaluation of the LTM by pattern recognition techniques [16], together with the coarser level of structural resolution makes time scales relevant to folding computationally accessible within the predetermined level of accuracy.

III. OUTLINE OF A PARALLEL COMPUTATION OF COARSE SOFT-MODE DYNAMICS

Following the tenets expounded in the previous section, the discretized torsional dynamics may be computed according to the following formal scheme: (a) We introduce a ternary variable $G(n)=1,2,3$ indicating, respectively, whether the nth residue along the chain is hydrophobic, neutral or hydrophilic (polar). (b) We determine the type of Ramachandran plot (I–IV), as indicated in Fig. 1, for each residue $n=1,\ldots,N$. (c) We define an LTM $\mathbf{y}$ by two rows $\{R(\mathbf{y},n),G(n)\}_{n=1,\ldots,N}$, as illustrated in Figs. 2 and 3. Thus, $R(\mathbf{y},n)=1$ indicates that the nth residue has adopted the extended conformation compatible with a β sheet; $R(\mathbf{y},n)=2$ indicates that either the nth residue has adopted a locally compact conformation compatible with a β bend (zero pitch), or with a left-handed α helix; finally, $R(\mathbf{y},n)=3$ indicates that the conformation of the nth residue is compatible with the formation of a β bend or with a right-handed α helix [14]. (d) We perturb the LTM by simulating interbasin transitions according to fixed transition probabilities to coarsely simulate the torsional isomerizations. (e) At fixed time intervals we search for *consensus* regions (patterns) of torsional isomers along the chain (Figs. 2 and 3). By consensus we simply mean regions of the chain where the local topological constraints associated to the formation of a particular folding or structural element are satisfied. In this way, a consensus window emerges as a pattern of structural signals encoded locally along the aminoacid sequence. The broad latitude (from 30° up to 60°, [14]) in local torsional coordinates within the Ramachandran basin [14], and the vast structural distorsion it leads to, implies that the discrete codification cannot be implemented at the geometric level. Rather, the interbasin transitions are meant to mimic changes in the local topological constraints to which the flexible chain is subject in order to reach specific structural patterns. (f) We evaluate and translate such patterns into a CM evolving within the time scale range 10 μs–10 s. Thus, translation becomes a pattern recognition and therefore, a parallel operation taking place at regular intervals. Each pattern within an LTM emerges with a certain probability that is effectively computed as the number of evaluations of the LTM that yield the particular structural motif associated to the pattern divided by the total number of evaluations of the same LTM. (g) The translation operation is subject to a feedback loop, whereby a renormalization operation readjusts the basin mean transition frequencies according to the latest CP translated, and 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* by slowing down or speeding up basin transitions, depending on whether new interactions are formed or dismantled. The rationale behind this is that, as a

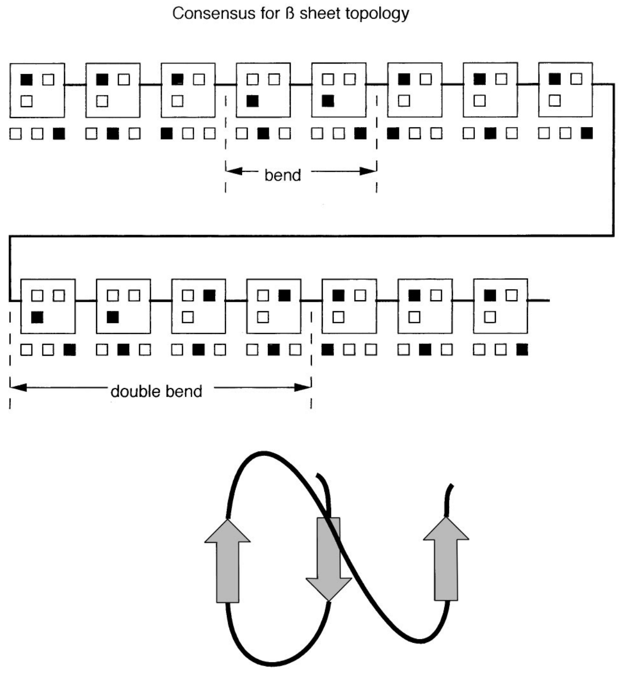

FIG. 3. A LTM consensus window for the complex three-strand antiparallel β-sheet motif shown.

new CP is formed engaging residues that were previously free, the mean energies of their Ramachandran basins decrease by ΔH, the change in enthalpy due to the formation of long-range intrachain contacts. Thus, the local activation barriers for basin transitions within a Ramachandran map become larger than in the case of free residues (the renormalization operation ''slows down'' the residues engaged in a recognized pattern). (h) Nucleation steps and the cooperativity in the formation of secondary structure are accounted for by means of the renormalization operation: Suppose the LTM is evaluated at a given time and a short consensus region is detected. Then, the residues which generated this initial consensus window become endowed with basin transition frequencies which are lower than those of the neighboring residues and, consequently, the consensus region initially formed has a chance to grow upon successive evaluations of the LTM.

Essentially the tenets (a)–(h) enable us to define a p-t-r iterative algorithm endowed with predictive potential and most suitable to mimic the parallel nature of the exploration of conformation space.

IV. 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 α 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 (Fig. 2). Because of the highly helix-disruptive tendencies of glycine [14,19,20], if its local diagram appears within a consensus window, the entire helix turn containing glycine is obliterated from the CM. The disrupting tendencies of the proline, on the other hand, do not require special instructions as the $R(\mathbf{y},n)=3$ value cannot occur for a residue preceding proline, as shown in Figs. 1 and 2. A consensus window translating into the (R) α helix is indicated in Fig. 2. Similarly, for a left-handed α helix, we must demand permanence of $R(\mathbf{y},n)=2$ value, while retaining all other conditions regarding hydrophobic periodicity along the chain.

Likewise, being pleated structures, β sheets are characterized by the persistence of the extended local conformation basin marked by $R(\mathbf{y},n)=1$. In order to fulfill hydrophobic/polar compatibilities, the G values must be preserved in a parallel or antiparallel fashion, depending on the relative orientation of the strands in the β sheet (Figs. 2 and 3). For the sake of illustration, a three-strand β-sheet topology is displayed in Fig. 3 together with its LTM consensus pattern. A structural pattern in the same topology class [17] will be generated in our (p-t-r) simulation of the folding of [18–21].

Turns and bends may be a determinant of the β sheet or simply required to form hydrophobic contacts, thus they will be treated generically, regardless of whether or not they realize β-sheet topologies. Should such motifs require closure of chain loops, they would require a $R(\mathbf{y},n)=2$ or $R(\mathbf{y},n)=3$, consensus window in the LTM at the time of its evaluation.

V. LOCAL CHOICE OF TRANSITIONAL DIRECTION FOR EACH RESIDUE AT A GIVEN TIME

Prior to determining the transitional timescale for the local changes in the value of $R(\mathbf{y},n)$, we need to operationally determine each local direction for a the basin transition taking place in the local Ramachandran landscape. Thus, the basin transition marked by a change $R(\mathbf{y},n)\rightarrow R(\mathbf{y}',n)$ will be clockwise ($1\rightarrow 2$, $3\rightarrow 1$ in an alanyl-like residue) or counterclockwise ($2\rightarrow 1$, $1\rightarrow 3$) depending on a lottery based on the fixed direction probabilities. Thus, if $p(+,n,j)$, $p(-,n,j)$ denote, respectively, the probability of a clockwise or counterclockwise transition when the nth residue is within basin $j[R(\mathbf{y},n)=j]$, we obtain for an alanyl-like residue

$$\begin{aligned} p(+,n,1) &= \exp[-(E^{\neq}_{12}-E_1)/RT]/\{\exp[-(E^{\neq}_{12}-E_1)/RT]+\exp[-(E^{\neq}_{13}-E_1)/RT]\} \\ &= 1/\{1+\exp[-\Delta E^{\neq}/RT]\}, \end{aligned} \tag{1}$$

$$p(-,n,1)=1-p(+,n,1), \tag{2}$$

where $E^{\neq}_{12},E^{\neq}_{13}$ represent, respectively, the energies of the transition states (saddle points in the Ramachandran landscape) corresponding to the transitions $1\rightarrow 2$, $1\rightarrow 3$ in an alanyl-like residue; $(E^{\neq}_{12}-E_1)$ and $(E^{\neq}_{13}-E_1)$ are, respectively, the kinetic barrier for the $1\rightarrow 2$ and $1\rightarrow 3$ transitions; and $\Delta E^{\neq}=E^{\neq}_{13}-E^{\neq}_{12}$. In essence, $[\Delta E^{\neq}/RT]$ must be treated as a parameter to be independently determined, as the transition state energies within Ramachandran landscapes are not known precisely. This is so since MD trajectories spend virtually no time in saddle points and thus their energies are not directly accessible through such computations. Nevertheless, the parameter $\Delta E^{\neq}/RT$ may be estimated indirectly and independently using data which is not directly generated by our computations in order to determine interbasin transitional probabilities for all alanyl-like residues. For that purpose, a p-t-r simulation is carried out (see below) for a test oligopeptide chain made up of four alanyl-like residues capable only of forming a single right-handed α-helix turn [cf. Fig. 2(b)] and compared with previous independent molecular dynamics computations on the same test system [19]. The comparison yields the estimate $\Delta E^{\neq}/RT=-\ln 1.88$. Likewise, for the nth alanyl-like residue we also get $p(+,n,2)=0$, $p(-,n,2)=1$; $p(+,n,3)=1$, $p(-,n,3)=0$.

These equalities result from the fact that there is no transition state or saddle point between basins 2 and 3 in the Ramachandran landscape [14]: A right-handed helix (basin 3) requires dismantling, or turning into the ''extended conformation'' (basin 1), before it may be turned into a left-handed helix (basin 2). Thus, the only possible pathway for such a local transformation is $3 \rightarrow 1 \rightarrow 2$.

We know from our computer experiments that the time evolution of the LTM is crucially dependent on the parameter $\Delta E^{\neq}/RT$. This is expected since the extended conformation of a residue must be favored entropically (it has a larger Ramachandran basin), as well as enthalpically (hydrogen bonds are 1.6 times stronger in β sheets than in α helices) [13,14]. Thus, although meaningful quantitative information on the sensitivity of the results to parameter changes is difficult to assess, we have noticed that a 10% decrease in the $\Delta E^{\neq}/RT$ value is responsible for the formation of critical bubbles or β-sheet-destruction kernels that preclude the persistence of β sheets within folding time scales. On the other hand, the value given above and obtained from independent sources [19] is adequate to reproduce the chronology of events and elucidate the significant kinetic bottlenecks in the dominant folding pathway of BPTI, as shown in Sec. XII.

The other types of residues are treated similarly. Thus, if the nth residue is glycine all probabilities are $p(\pm,n,j)=1/2$, for $j=1,2,3,4$. This is so since glycine has such a small side chain (a hydrogen) that the negligible steric hindrance makes all torsional transition states virtually identical in energy. On the other hand, if the nth residue precedes a proline, we get $p(+,n,1)=1$, $p(-,n,1)=0$; $p(-,n,2)=1$, $p(+,n,2)=0$. Finally, for a proline at contour value n, we obtain $p(+,n,1)=0$, $p(-,n,1)=1$; $p(+,n,3)=1$, and $p(-,n,3)=0$.

VI. DISTRIBUTION OF BASIN TRANSITION FREQUENCIES IN THE GENERATION OF LTM's

Our discretized model of topological dynamics covers the spectrum of activated molecular motions evolving within the 1 ps–1 ms time range. Thus, faster diffusional-like unhindered torsions [19,20] in a free residue have been integrated out as conformational entropy of the state defined by the coarse LTM representation (cf. [16]). Such motions well into the ps range, as well as the faster hard-mode dynamics, determine the rodlike shape of the protein molecule and the all-or-none nature of the loop closure steps when viewed within the timescale window between two LTM states. Thus, the microscopic mean time range relevant to LTM transitions is 10^{-11}–10^{-3} s, covering the time scale for internal motions (10^{-11} s) of the order of the calculated diffusional displacements of flexible hinged domains [19] and, at the other end of the spectrum, the limiting values (10^{-4}–10^{-3} s), typical mean time scales for the fast exchange between folded and unfolded states with respect to tertiary interactions engaging two secondary elements [18]. Within this range, we encompass the mean time frame of 10^{-7} s, typical for a localized helix-unwinding event leading to a bubble ([19], cf. also [16]).

These considerations lead us to define a temperature-dependent normalized distribution of transition periods, $w=w(\tau)$ for the N independent basin transitions within the Ramachandran local landscapes. This distribution features three Gaussian peaks centered at characteristic periods 10^{-11}, 10^{-7}, and 10^{-3} s. The transition times are assigned from this distribution in such a way that the effect of thermal fluctuations on the formation of consensus and thus, on structural transitions is incorporated. Each Gaussian peak has a dispersion $\sigma^2=gT$, where the constant g depends on the actual denaturation temperature $T(\text{denat.})$ and on the consensus-based interpretation of denaturation, as shown below.

The 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 [19], with typical time scales for localized helix disruptions, and with diffusion-collision models [20] in which secondary structure is stabilized further by forming tertiary contacts [18]. Furthermore, the decrease in heat content (ΔH) due to the formation of secondary or tertiary intrachain interactions is responsible for the lowering of the average energies of the Ramachandran basins with concurrent increase in the local transition times (cf. Sec. XI).

Accordingly, the LTM evolves through 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 respectively of type I, II, or III. Thereafter, a new direction and a new escape time is assigned from the lottery for each new basin transition after the previous transition has been completed. This procedure warrants the maximum permanence time in the extended local conformation 1 for a free alanyl-like residue, in accord with observed facts [19]. 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

$$[f^{-1}-10^{-11}\ \text{s}|\leqslant|\tau'-10^{-11}\ \text{s}|, \quad \text{with } \tau' \text{ satisfying}$$
$$w(\tau')=\text{Infimum}_{\tau}\{w(\tau)\geqslant 1/[2N]\}. \tag{3}$$

The condition yielding the shortest escape time τ' arises from the fact that there are at most $2N$ possible local transitions in the peptide chain (a maximum of two transitional directions for each residue in each given basin). Such considerations yield the value $\tau'\approx 1$ ps at $T=298$ K. Thus, the real time interval between two consecutive readings or evaluations of the LTM must be taken to be $3^4=81$ ps, that is, the minimum time to get a CP transition (formation of a β bend or helix turn engaging four residues) with the fastest basin-transition times.

The other two peaks in the distribution $w=w(\tau)$ correspond, respectively, to mean escape times for residues in secondary and tertiary structural elements. Again, the same considerations apply in regards to the escape time assignment to basin transitions for class II and class III residues. These rules imply that the residues in loops, bends, or turns

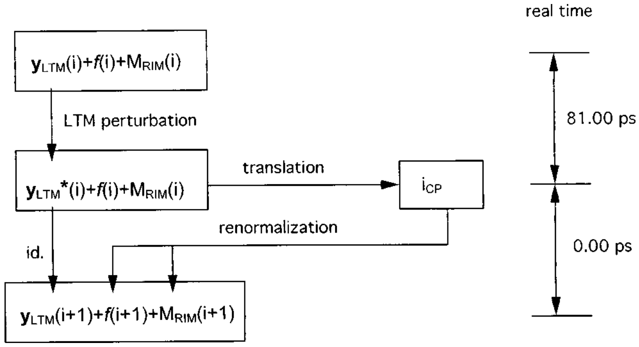

FIG. 4. The generic chart of operations for the personal computer realization of the ith perturbation-translation-renormalization p-t-r cycle iteration. The pattern recognition or translation operation requires storage of the generic patterns, such as those illustrated in Figs. 2 and 3. The LTM perturbation requires one sequential computation for each LTM entry following working equations (4) and (5). Each renormalization operation takes place while the LTM is quenched in the same state, LTM*, in which it was translated. This operation defines the vector $f(i+1)$ of mean frequencies, and the reading instructions manual, $M_{\mathrm{RIM}}(i+1)$, to be used in the $(i+1)$th p-t-r cycle.

concurrently formed with any secondary or tertiary structural element adopt the cadence of the structural element itself.

The mechanistic aspects of period distribution imply that this operation is subject to renormalization 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 an intrachain interaction, or, equivalently, they belong to class II or III. Thus, since a CP transition reclassifies the residues, it also dictates the range from which new transition time assignments are drawn (cf. Fig. 4). The time range for a specific residue remains the same as before the CP transition if the transition does not alter the class of the residue, and changes if the CP transition transfers the residue to a different class. Thus, the renormalization operation is essentially the means by which long-range correlations in the LTM are introduced as a consequence of the latest large-scale motions.

VII. TEMPERATURE T IN THE DISCRETIZED LTM DYNAMICS

Secondary structure dismantling materializes and is recorded as such by deletion in the CM whenever a consensus bubble forms amongst class II residues engaged in the structure. 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 consensus window length [14,18] must be out of phase with the consensus required values at the time when the reading of the LTM takes place. Because of the renormalization loop, this transition at the CM level immediately transfers a new set of constraints for the generation of new LTM's. The residues previously engaged in the structure and in its concurrent loops are reclassified, being transferred from class II to the higher frequency class I.

Cooperative effects reflect themselves mechanistically in the formation of the consensus bubble: For instance, in the α-helix motif, the larger the helix, the more improbable it is to find a 30% out-of-phase subsequence of class II residues. Furthermore, these considerations enable us to estimate the constant g which determines the effect of thermal fluctuations on the period distribution: At the denaturation or melting temperature $T(\text{denat.})$, every helix formed must develop a consensus bubble evaluated and recorded with the next reading of the LTM. Thus, if σ is ''large enough,'' the period distribution in the helix is broad enough so that consensus cannot be preserved: The period range, of the order of σ, is such that a helix consensus cannot survive two consecutive readings. From these considerations, and taking into account our empirical estimate of the denaturation dispersion fixed at $\sigma = 10^{-8}\,\mathrm{s}$, and the typical experimental $T(\text{denat.}) = 313\,\mathrm{K}$ for proteins such as the ones studied in this work [17]), we get $g \approx 3.2\times 10^{-19}\ \mathrm{s}^2\ \mathrm{K}$.

VIII. NUCLEATION EFFECTS AND PROTEIN FOLDING COOPERATIVITY

Complex structural patterns such as the one presented in Fig. 3 do not result from all-or-none processes. Rather, a nucleating event involving the largest cost in conformational entropy takes place first and is subsequently followed by a sequence of folding events that seek to minimize additional losses in conformational entropy. Two theoretical approaches to biopolymer folding, the variational treatment rooted in the sequential minimization of entropy loss (SMEL) [9,16], and the theory of hydrophobic zippers [10] take into account the cooperative nature of formation of structural motifs triggered by nucleating events in their attempt to account for pathways that minimize the entropic cost of each step.

In our model, cooperativity and nucleation work through successive renormalization loops as expedients for the folding of peptide chains: Suppose a nucleating pattern has been identified in a reading of the LTM, be it a helix turn such as the one depicted in Fig. 2(b), or a turn involving a loop closure (Figs. 2(a) and 3). Then, the residues engaged in the respective nucleating consensus window are automatically locked into lower frequencies of the order of 10^7 Hz via the renormalization operation that immediately follows the translation of the nucleation pattern into a CP registered in the CM. Thus, the nucleation consensus will ''survive'' or be preserved for approximately 1000 evaluations of the LTM, since the evaluation frequence is of the order of 81 ps at 298 K [cf. Eq. (3)], the typical working temperature for *in vitro* folding of BPTI. This time span allows for a chance to enlarge the original consensus window by progressive torsional isomerizations of residues adjacent to those engaged in the nucleating step. Once adjacent residues form consensus compatible with the original one in relation to a specific structural motif, they are also locked into the 10^7 Hz frequency domain and thus they contribute to the seeding process.

Summarizing, the key feature of renormalization that enables us to deal with cooperativity in the formation of secondary structure is the survival of a nucleating pattern warranted by the reclassification of class I residues as class II residues once the nucleating pattern occurs. The argument holds *mutatis mutandis* for tertiary interactions formed cooperatively 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 Hz, thus warranting the survival of the respective nucleation pattern.

IX. THE RENORMALIZATION OPERATION AND RENORMALIZED CONTOUR DISTANCES

As stated above, the renormalization operation is the expedient by which long-range correlations are introduced. It involves two concurrent roles: Redefining contour distances between different residues relative to the latest CP translated, thus determining how to read and evaluate the LTM, and placing a new set of constraints on the generation of new LTM's after a CP transition has been recorded as a change in the CM. Both roles are specified at a given time according to the latest CP generated. The renormalization of constraints for the basin transitions is articulated through the operational feedback loop (Fig. 4), introducing the latest CP generated as input for the reclassification of residues. On the other hand, the renormalization of contour distances points to the root of cooperativity, since unfavorable long-range interactions might eventually materialize with lower entropic costs, as illustrated in Sec. XI.

X. ROUGH PROTEIN FOLDING ON THE PERSONAL COMPUTER

The aim of this section is to actually prescribe the implementation of the perturbation-translation-renormalization (p-t-r) cycle operations on a PC. A suitable PC is one roughly matching the following specifications: RAM memory, 128 MB; processor speed, 300 MHz; hard disc memory, 6 GB.

As previously stated, the p-t-r computation mimics a parallel synchronous algorithm designed to predict the active structure of biomolecules reached within times incommensurably shorter than those required for thermodynamic equilibration. Such computations require a parallel evaluation of concurrent folding possibilities at regular intervals. Each reading, in turn, defines the set of constraints to which the system is subject when undertaking the next folding stage.

A sequential computation on a PC implies that we would need to quench the LTM every 81 ps, as indicated in Fig. 4, to sequentially perform a translation or pattern recognition. In turn, the recorded pattern would determine via renormalization the generation of the new sequence of LTM's to be read after another 81 ps have elapsed. As we recall, the LTM evolves in time by basin transitions requiring a choice of transition direction in the Ramachandran landscape [cf. Eqs. (1,2)] and a frequency (or transition time) chosen from a Gaussian distribution determined by renormalization at every reading timestep from a set of three distributions: one for free residues (class I, mean frequency $f=10^{11}$ Hz), one for residues engaged in secondary structure (class II, mean frequency $f=10^7$ Hz) and one for residues engaged in tertiary interactions (class III, mean frequency $f=10^3$ Hz). Renormalization assigns a distribution to each residue for a given codified Ramachandran map in the LTM, and the sequential computer mimics the dynamics resulting from the fact that a new frequency is chosen from the same distribution after each transition has been completed.

We may realize the p-t-r computation iteratively, by quenching the LTM dynamics during the t-r operations, as shown in Fig. 4. In this figure, the ith iteration is displayed as preparing the LTM for the $(i+1)$th iteration. The LTM has been added two modules: a mean frequency N entry vector (f) and a reading instruction manual (RIM). Each entry in the frequency row may take one of three values 1, 2, or 3, according to whether the mean frequency assigned by renormalization to the particular unit or residue is 10^{11}, 10^7, or 10^3 Hz. The RIM determines the renormalized distances to be considered when searching for consensus in regions containing already windows of secondary of tertiary structure. Following the RIM, if a 2 or two–three–mean frequency window already translated into a CM is found within a window where consensus is being searched, it must excluded from the new window being examined.

In order to inductively define the ith p-t-r iteration, as shown in Fig. 4, we shall consider given all three modules, $\mathbf{y}_{\mathrm{LTM}}(i)$, $f(i)$ and $M_{\mathrm{RIM}}(i)$, where i refers to the ith iteration. The $\mathbf{y}_{\mathrm{LTM}}(i)$ must be perturbed during 81 ps to obtain the $\mathbf{y}^{*}_{\mathrm{LTM}}(i)$ at the end of the time step. This corresponds in real time to a 81 ps progress of the torsional dynamics since the $\mathbf{y}^{*}_{\mathrm{LTM}}(i-1)=\mathbf{y}_{\mathrm{LTM}}(i)$ has been translated. The $\mathbf{y}^{*}_{\mathrm{LTM}}(i)$ is then ready to be translated into the i_{CM}. The perturbation $\mathbf{y}^{*}_{\mathrm{LTM}}(i)$ is determined exclusively by the $\mathbf{y}_{\mathrm{LTM}}(i)$ together with the mean frequencies encoded in the vector $f(i)$. Thus, the perturbation actually corresponds to the running of the LTM dynamics for 81 ps, until the LTM is quenched again for pattern recognition. The pattern recognition or translation, in turn, defines the next vector $f(i+1)$ of mean frequencies and the next reading instructions manual $M_{\mathrm{RIM}}(i+1)$, both entities being produced by the ith renormalization operation. This information, together with the previously perturbed $\mathbf{y}^{*}_{\mathrm{LTM}}(i)=\mathbf{y}_{\mathrm{LTM}}(i+1)$ is all that is needed to now run the $(i+1)$th t-r interation, which would involve first perturbing the $\mathbf{y}_{\mathrm{LTM}}(i+1)$, then translating the $\mathbf{y}^{*}_{\mathrm{LTM}}(i+1)$ and finally reassigning according to $\mathbf{y}_{\mathrm{CM}}(i+1)$, mean frequencies and the new RIM. The inductive definition of the sequential algorithm is now complete.

In practice, an actual bottleneck in the sequential computation is the pattern recognition or translation on the LTM, an inherently parallel operation. If N is not too long ($N\approx 50$–100) this operation may be accessible to a sequential machine engaged in column-by-column reading with concurrent memory storage. A state-of-the-art PC as specified above takes approximately 0.71 ms of real time to sequentially translate each LTM into a CM with $N=100$, storing nucleation folding events and pattern prototypes, such as those specified in Figs. 2 and 3.

A. The perturbation operation on the PC

The perturbation $\mathbf{y}_{\mathrm{LTM}}\rightarrow\mathbf{y}^{*}_{\mathrm{LTM}}$ is the first aspect to be considered in a sequential computation, as it determines the dynamic flow in our coarse description. In order to determine the state of each entry in the LTM once a time step of 81 ps has been completed, one must consider the following scheme: (i) Classify residues according to whether their mean frequency value assigned in the ith p-t-r cycle is 1, 2, or 3. If residue n has been previously assigned the 2 or 3 frequency value, store its actual frequency $F(n)$ chosen from the respective Gaussian distribution and record the number

$Q_i(n)$ of p-t-r cycles that have taken place since the 2 or 3 values have been originally assigned until the ith cycle is performed. Such units are given a different treatment, following instruction item 4. For the free units at the time of the ith iteration, instructions 1–3 are to be followed. (ii) Record the value 1, 2, or 3 of the variable $R(\mathbf{y}^*_{\mathrm{LTM}}(i-1),n)=R(\mathbf{y}_{\mathrm{LTM}}(i),n)$ for each entry n with mean frequency value 1, quenched throughout the $(i-1)$ translation or pattern recognition. (iii) For each residue or column n during the ith iteration, consider the Gaussian distribution peaked at $f_n(i)=10^{11}$ Hz. Using this distribution, generate by means of a Monte Carlo simulation as many frequencies f_k, $k=1,\ldots,n^*$ as necessary, so that the following inequalities hold:

$$\sum_{k=1,\ldots,n^*-1}(2\pi/f_k)<81\ \mathrm{ps}\leqslant\sum_{k=1,\ldots,n^*}(2\pi/f_k). \quad (4)$$

(iv) Determine the state of the R entry at the time of the next reading as follows: It changes provided

$$3/2\pi\geqslant\left(81\ \mathrm{ps}-\sum_{k=1,\ldots,n^*-1}(2\pi/f_k)\right)f_{n^*}\geqslant\pi/2. \quad (5a)$$

It remains the same as in the $(i-1)$th translation if

$$\left(81\ \mathrm{ps}-\sum_{k=1,\ldots,n^*-1}(2\pi/f_k)\right)f_{n^*}<\pi/2 \quad \text{or} \quad >3/2\pi. \quad (5b)$$

In the event that Eq. (5a) is fulfilled, to determine the new value of the R entry, one must choose n^* directions from the lottery specified by the fixed probabilities given in Secs. V and VI. (v) Suppose entry n has been assigned a frequency value of 2 or 3 a number $Q_i(n)$ of p-t-r cycles before the ith cycle. Then, since $2\pi/F(n)\geqslant 81$ ps, its state to be read in the ith pattern recognition remains the same if $[2\pi/F(n)-Q_i(n)81\ \mathrm{ps}]F(n)<\pi/2$ or $>3/2\pi$, and changes otherwise. If a consensus bubble in a secondary or tertiary structure is formed by having 30% of out-of-phase R states at the time of the ith translation, then the entire consensus units are reassigned the 1 frequency value by the ith renormalization operation, thus indicating the dismantling of a structural pattern.

In this way, following instructions (i)–(v), we define the perturbed matrix $\mathbf{y}^*_{\mathrm{LTM}}(i)$, which is quenched during the ith translation-renormalization, as indicated in Fig. 4. This is precisely the matrix fed into the $(i+1)$ iteration, that is $\mathbf{y}^*_{\mathrm{LTM}}(i)=\mathbf{y}_{\mathrm{LTM}}(i+1)$, which, in turn, must be perturbed according to the reassignment $f(i+1)$ of frequencies which took place in the ith iteration.

In real time, a single LTM perturbation following instructions (i)–(v) takes 0.60 ms maximum. This upper bound is found in the case where all residues are free, and therefore, their frequences are to be chosen from the distribution with the highest mean frequency. Thus, the most conservative estimation of the total PC computation time involved in the estimated 10^7 p-t-r loop iterations, as required to satisfactorily penetrate relevant folding times, is 1.31×10^4 s.

B. Translation operation in the PC

In each iterative feedback loop, the translation operation follows after the perturbation operation and requires the quenching of the dynamic flow (coarse torsional motion), which is realized by freezing the LTM* while the translation and renormalization operations are performed (cf. Fig. 4). Due to its inherently parallel nature, the translation operation is the most time consuming in a sequential machine. It is applied to the perturbed LTM, the LTM*, and maps it onto a CM after having identified all structural patterns. This map entails three possible steps: (1) It records newly formed structural patterns as contacts in the CM; (2) It deletes CP's if meaningful consensus bubbles arise in the LTM* at the time of its evaluation; (3) It records the eventual growth of preexisting patterns.

Since a pattern involves the formation of intrachain contacts, the first thing that a pattern recognition on the LTM* must take into account is to search for consensus windows of local compact conformations, that is, uniform windows with either Ramachandran value $R=2$ or 3. Such windows represent bendings of the chain that *may* induce contact formation. That is, they are *necessary* (but not sufficient) conditions for the occurrence of loops, β bends, or α-helix turns, the seeding elements for structure formation leading to antiparallel contacts (Figs. 2 and 3). On the other hand, potential regions of parallel contact require windows made up of two adjacent compact subwindows, each with different R (one subwindow with $R=2$ followed by one with $R=3$, or vice versa) (cf. Fig. 3). Thus, once such necessary conditions are taken into account, the search for possible foldings is considerably simplified: The ''compact windows'' must be identified first in a systematic search for structural patterns.

The next step in translation is to check whether putative interacting regions flanking the compact windows are in the proper local torsional conformation (extended $R=1$ conformation for β-sheet interactions, $R=2$,3 for helices). Finally, a matching of the hydrophobicity/polarity/neutrality map should hold for the putative interacting regions defined by the compact window, unless tertiary contacts form first, and they later caffold the formation of secondary structure, in which case the tertiary contacts serve as templates in the matching (see Sec. XII).

On the other hand, the detection of consensus bubbles in a preexisting structure is actually straightforward: The program must simply check whether the Ramachandran values of a previously identified consensus window remain unaltered or whether 30% of consensus has been lost at the time of the reading by formation of an out-of-phase subwindow, in which case the corresponding CP is deleted by the translation operation.

Finally, with regard to the detection of patterns of structural growth, the same considerations arise as for the identification of new patterns: The preexisting structure must possess a ''compact subwindow'' within its consensus window, otherwise it wouldn't have arised in the first place. Thus, the consensus demands for a growth folding event to materialize are the same as those needed to form a recognizable pattern once a compact window has been detected.

C. The renormalization operation in the PC

Since, as indicated in Fig. 4, the renormalization operation follows after translation and is determined by it, its ac-

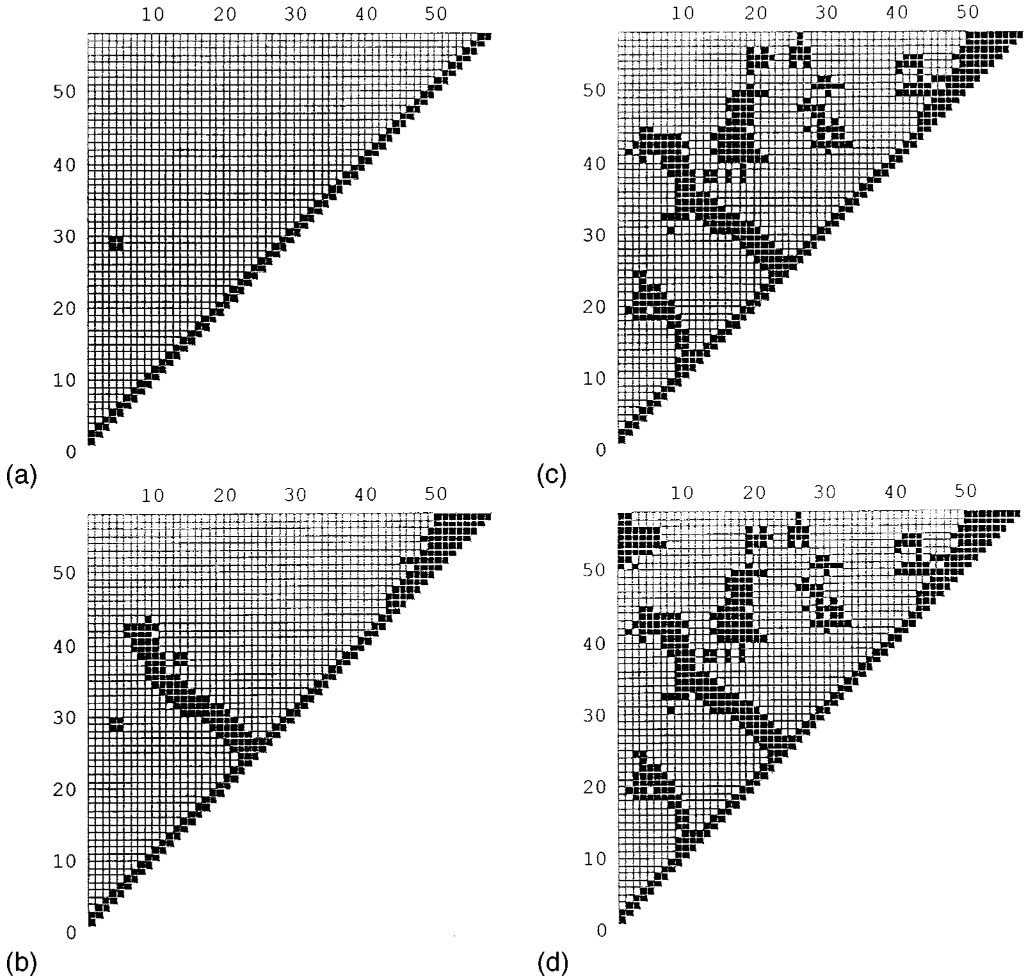

FIG. 5. (a)–(d) Four snapshots of the CM taken, respectively, at 3.2×10^{-4}, 1.3×10^{-3}, $1.3\times10^{-3}+3.2\times10^{-7}$ s, and $1.3\times10^{-3}+3.2\times10^{-7}+0.5\times10^{-2}$ s in the course of 10^7 p-t-r cycles for the BPTI.

tual implementation becomes relatively simple: Suppose a pattern has been detected and recorded as such in the CM (an α helix as a segment parallel to the main diagonal, a β sheet as a segment perpendicular to it, and looped, native and tertiary contacts as connected and simply connected shaded regions [cf. Figs. 5(a)–(d)]. Then, if residues m and n represent the contour extremities of the preexisting intrachain interaction, a new ''renormalized'' chain is defined by regarding residues m and n as consecutive along the new chain, and new perturbation rules apply to its possible LTM's. In this way, all residues in between m and n, being already structurally engaged, have been virtually removed from further secondary structure search: All new consensus windows are searched in the renormalized chain by assuming that the region of the chain deleted by renormalization has a compact window within, and as such, is only susceptible of structure growth, should the right consensus windows flanking residues m and n form.

Its other role, the reassignment of mean frequencies to all residues according to the latest CM generated is easily introduced: Residues engaged in secondary structure are slowed down with respect to free residues and those engaged in tertiary contacts are slowed down even further. Thus, the hierarchical nature of structure formation translates dynamically into a hierarchy of time scales. In this way, renormalization modifies the RIM, defining the way in which the next LTM* is to be generated and how new consensus windows are to be searched.

According to the mechanism of the perturbation-translation-renormalization cycle, the fragility of secondary structure depends on the number of cycles it survives before a consensus bubble is registered upon translation. However, in agreement with the reaction-diffusion model [20], we predict that this structure may be further stabilized by forming tertiary interactions with the net effect of slowing down all oscillators involved by four orders of magnitude: It would take on the average $Q=10^9$ cycles to detect a consensus bubble in this upper level of structure hierarchy. This stabilization becomes operational in the shaping of the active folded structure, as shown in Sec. XII.

XI. ESTIMATION OF THE KINETIC BARRIER OF AN ELEMENTARY FOLDING STEP

In order to elucidate the kinetic bottlenecks and significant events along a folding pathway, it is of paramount importance to estimate the kinetic barriers for elementary transitions between CP's. This requires that we renormalize the potential energy surface (PES) to that coarse level of description and adopt an adiabatic ansatz (cf. [16]) whereby (a) each

CP is associated with a superbasin of attraction in the PES, containing as many substrates as LTM's may realize for that same CP and (b) equilibration within a superbasin is incommensurably faster than intersuperbasin or CP transitions [16]. We shall start by estimating the barrier B for a constructive folding step where the number of contacts is increased as a result of the transition. Thus, if $a_{\mathrm{CP}} \rightarrow b_{\mathrm{CP}}$ generically represents such a process, our aim is to determine the actual entropic cost $-T\Delta S_{ab} = B_{ab} = E^{\neq}_{ab} - E_a$ entailed in the formation of the new contacts [9,16]. Here $E^{\neq}_{ab}$ represents the effective transition state energy and E_a, the thermal average or expected energy of the super basin a_{CP}. This entropic cost quantifies the loss in conformational freedom involved in the formation of the more constrained structural pattern b.

We shall denote by $\Omega(a)$ and $\Omega(b)$ the LTM multiplicities of a_{CP} and b_{CP} that is, the number of possible LTM's translatable into a_{CP} and b_{CP} respectively. Thus, we obtain

$$\Omega(a) = \prod_{n=1,\dots N} q(a,n); \quad \Omega(b) = \prod_{n=1,\dots N} q(b,n), \tag{6}$$

where $q(a,n)$ indicates the number of possible values of the nth Ramachandran variable $R(\mathbf{y},n)$ for all LTM's $\mathbf{y}$'s translatable into a_{CP}. The same statement holds for $q(b,n)$, b_{CP}. Thus, we obtain

$$q(a,n) = 1, \tag{7}$$

if the nth residue is engaged in a structural element recorded in a_{CP}; or $q(a,n) = 2$, 3, or 4, if the nth residue is *not* engaged in any structural element of a_{CP} *and* the nth residue is proline or a residue preceding proline $[q(a,n)=2]$, the nth residue is alanyl-like $[q(a,n)=3]$, or the nth residue is glycine

$$q(a,n) = 4. \tag{7'}$$

Then using Boltzmann relation, the adiabatic kinetic barrier B_{ab} may be estimated as

$$B_{ab} = -RT \ln[\Omega(b)/\Omega(a)] = RT \sum_{n_1 \leqslant n \leqslant n_2} \ln q(a,n), \tag{8}$$

where $n_1 \leqslant n \leqslant n_2$ indicates the set of residues engaged in the formation of the new contacts absent in a_{CP} but present in b_{CP}. Since no more than two consecutive such residues may be proline, and given the type of LTM patterns recorded in the CM (cf. Figs. 2 and 3), it is certain that $B_{ab} \geqslant 0$.

On the other hand, the expected energy E_a of a_{CP} is

$$E_a = \Omega(a)^{-1} \sum_{\mathbf{y} \text{ in } a_{\mathrm{CP}}} E(\mathbf{y}), \tag{9}$$

$$E(\mathbf{y}) = H(a_{\mathrm{CP}}) + \sum_{n=1,\dots,N} E(R(\mathbf{y},n)), \tag{10}$$

where $H(a_{\mathrm{CP}})$ indicates the heat content or enthalpy of a_{CP} $[H \leqslant 0$, with $H(\text{random coil} = 0)]$, a quantity exclusively dependent on the pattern of long-range intrachain contacts, and $E(R(\mathbf{y},n)) = 0$, 1, or 2 kcal/mol depending on whether $R(\mathbf{y},n) = 1$, 3, or 2, respectively (see Fig. 1 and [14]).

Then, the transition state energy $E^{\neq}_{ab}$ may be adiabatically estimated as

$$E^{\neq}_{ab} = H(a_{\mathrm{CP}}) + \Omega(a)^{-1}\left[\sum_{\mathbf{y} \text{ in } a_{\mathrm{CP}}} \sum_{n=1,\dots,N} E(R(\mathbf{y},n))\right] + RT \sum_{n_1 \leqslant n \leqslant n_2} \ln q(a,n). \tag{11}$$

Moreover, the reverse kinetic barrier $B_{ba} = E^{\neq}_{ab} - E_b$, which is equal to ΔH_{ba}; the heat that must be transferred to the system in order to break the contacts that appear in b_{CP} and not in a_{CP}, is

$$B_{ba} = E^{\neq}_{ab} - H(b_{\mathrm{CP}}) + \Omega(b)^{-1}\left[\sum_{\mathbf{y} \text{ in } b_{\mathrm{CP}}} \sum_{n=1,\dots,N} E(R(\mathbf{y},n))\right]. \tag{12}$$

Thus, working Eqs. (6)–(12) characterize completely the kinetics of protein folding defined at the level of elementary transitions between contact patterns.

XII. RESULTS: CM PATHWAY FOR BPTI

The aim of this section is to predict the dominant folding pathway and structural features of folded BPTI by a mechanistic analysis of its long-time dynamics resulting from the simplified model of torsional motion at 298 K [7,18–21]. The simulation is sequentially implemented on a PC following the basic tenets already expounded.

We shall focus on determining significant folding intermediates and the late kinetic bottlenecks which occur within the first 10^{-2} s of the renaturation process, a time span that requires 10^7 p-t-r cycles. Our aim is to show how the dominant sequence of CM transitions for the BPTI describes the dominant folding kinetics, reproducing the essential cooperative features of the experimentally probed folding pathways, including the late scenario in which tertiary interactions direct and stabilize the native Cys-Cys (5,55) contact (cf. [7,18–21], Cys=cysteine aminoacid). The actual refolding conditions involve special redox conditions, so that the kinetics of formation, dismantling and recombination of intrachain Cys-Cys disulfide bonds lies within the 10^{-7} s time scale [7,21], and does not interfere with the folding process. This is so since the fastest Cys5-Cys30 interaction takes 4×10^{-5} s to form, as shown below.

In order to analyze the kinetics along the dominant folding pathway obtained by the p-t-r algorithm, we need to make precise the following two points: (a) An operational definition of contact based on a proximity of 7 Å has been adopted in the CM development in consonance with previous treatments [17–21]. (b) According to the discrete coding of the soft-mode dynamics and the results of Sec. XI, the mean time of a constructive contact pattern transition $a_{\mathrm{CP}} \rightarrow b_{\mathrm{CP}}$ is estimated as $\tau \exp(B_{ab})$, where $\tau = 10^{-11}$ s, and the barrier B_{ab} may be estimated using working Eq. (8).

How does cooperativity operate in the folding of BPTI? To answer this question we examine the estimated mean times [Eqs. (6)–(12)] to form native Cys-Cys disulfide contacts starting from a random coil conformation:

$$\tau(5{,}55)\approx 10^{4}\ \text{s},\ \tau(30{,}51)\approx 1.6\times 10^{-1}\ \text{s}. \tag{13}$$

Throughout this section, the numbers in brackets, as in Eq. (13), will denote the contour values of residues along the chain.

Direct inspection of Eq. (13) reveals that such native contacts take a long time to form and can only be created cooperatively within the time scales under investigation. On the other hand, the third native Cys-Cys contact (14,38) may form directly within time scales commensurate with the occurrence of tertiary contacts:

$$\tau(14{,}38)\approx 1.3\times 10^{-3}\ \text{s}. \tag{14}$$

The sequence of CM's taken within the time span of 10^7 *p-t-r* cycles is consistent with the previous analysis. Four significant snapshots of the time evolution of the CM have been taken for 17 runs of the parallel computation and are displayed in Figs. 5(a)–(d). All runs yield 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, 3.2×10^{-4}, 1.3×10^{-3}, and $1.3\times 10^{-3}+3.2\times 10^{-7}$ s (the third snapshot is taken 311 LTM readings after the second), and $1.3\times 10^{-3}+3.2\times 10^{-7}+0.5\times 10^{-2}$ s. A direct analysis of the CM evolution indicates that the nonnative (5,30) disulfide bond forms first because a loop of size 25 poses no orientational constraints on its polar groups. Thus we get the estimate: $\tau(5{,}30)\approx 3.2\times 10^{-4}$ s. This is a good estimate, since the (5,30) contact indeed shows up on the CM snapshot taken precisely at that time, as featured in Fig. 5(a).

Nucleation windows of the form shown in Fig. 2(b) seeding the formation of an α helix appear in the (43–58) extreme of the molecule within the range of time scales 8.8×10^{-4}–9.8×10^{-4} s. The timescale of formation of (14,38) given in Eq. (14) is also a good estimate. The snapshot taken at 1.3×10^{-3} s [Fig. 5(b)] displays this native contact, as well as fully and partially developed secondary structure elements such as the α helix and a two-strand portion involving the contour region (20–33) of the β sheet topology shown in Fig. 3. The nucleating events leading to the two-strand portion of the β-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.

Tertiary contacts between the α helix and the complex β sheet require the closure of the loop in the contour region (33–43) and start developing between the incomplete β sheet and the helix 311 LTM evaluations after the time when the last snapshot displayed in Fig. 5(b) was taken. This coincides precisely with the time estimation for closure of the 10 loop with five polar groups within the (33–43) region of the BPTI:

$$\tau(\text{tertiary interaction})\approx 3.2\times 10^{-7}\ \text{s}. \tag{15}$$

We emphasize that this is a renormalized calculation that assumes the previous formation of secondary structure.

At this point, the nonnative (5,30) Cys-Cys bond is completely dismantled to give rise to the native (30–51) contact induced and stabilized by the tertiary interaction, as a direct observation of the third snapshot [Fig. 5(c)] reveals. This pathway reveals how the formation of (30,51) is expedited by cooperative folding, in agreement with recent findings [7]. Furthermore, during the timespan of development of the first tertiary contact, the complex β-sheet motif continues to develop, fostering other tertiary interactions between the two dominant secondary structure elements. Since folding and unfolding of tertiary structure exchanges on the 10^{-3} s fast nuclear magnetic resonance time scale, the locking of oscillators at the mean 10^3 Hz frequence peak warrants the survival of the initial tertiary consensus while contact (30–51) forms and the β sheet is completed.

Finally, the (5,55) disulfide bond that would initially take a forbiddingly long time to form, now entails the closure of a complex 29 loop with no polar orientation demands: This loop is made up of the quasicoil (5–20) region, the β sheet (20–30) region, and the quasicoil strand (51–55). Notice that the formation of the native (30,51) contact has short circuited the loop closure for the (5,55) contact, so that the estimated time for this interaction is

$$\tau(5{,}55)\approx 0.5\times 10^{-2}\ \text{s}. \tag{16}$$

This estimation of the rate-determining step in BPTI folding is corroborated by examination of the fourth snapshot [Fig. 5(d)], and confirms previous estimations [7,18–21], 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 the tertiary structural elements of the BPTI but it also shows how cooperative effects can serve as an expedient to aid the formation of native interactions shaping the hydrophobic core. A good agreement with experimental kinetic probes [7,18–21] has been found.

Another proof of the predictive potential of our treatment is provided by the fact that the CM for the predicted active folding of BPTI [Fig. 5(a)–5(d)] contains all meaningful structural elements already identified in the CM obtained from X-ray crystallographic data [22]. Direct comparison reveals that both the predicted and experimental CM's reveal the very same functionally competent structural elements. However, in this regard we must warn the reader that the validity of a finer comparative analysis is not appropriate since such an analysis would demand a coincidence in the actual definition of ''contact.'' We have adopted an operational definition: A contact materializes if the distance between two residues is less than 12 Å, the largest distance for meaningful intrachain interaction with energy decrease of at least 1/2 kT. However, the literature adopts (cf. [22]) CM's with an *arbitrary* maximum distance, typically 10 Å. Not surprisingly, the CM using the more restrictive definition of contact is more sparse than the one adopted in this work. However, our CM identifies *all functionally competent structural elements and no other*, in perfect coincidence with the experimental CM.

XIII. DISCUSSION

The folding of a flexible peptide chain is inherently a parallel process in which different portions of the chain perform independent searches in conformation space and the search becomes expeditious and robust for sequences which are targets of natural selection [1–8]. An algorithm intended to predict the dominant folding pathway should capture this feature. In this work we present one such algorithm rooted in the search for patterns determined by a discrete codification of the local torsional states of the peptide chain. This representation enables us to penetrate into the time scale range of 1–10 ms, relevant to the protein folding process. In essence we have simulated a coarse version of the soft-mode dynamics [11–13] on a PC. To make it feasible to implement on a PC an inherently parallel simulation, we periodically quenched the dynamics, registered all structural patterns formed or dismantled, and renormalized the chain accordingly. Thus, the implementation on a PC required the iteration of a perturbation-translation-renormalization cycle performed on a matrix of local torsional topologies whose evolution roughly describes the folding pathway.

Our treatment hinges upon a binary codification of local torsional isomerizations subject to the constraints imposed by both local and long-range correlations. In this way, patterns of locally encoded structural signals represent actual patterns of fulfilled local topological constraints that are recognized and translated as foldings of the chain. In turn, a feedback or renormalization loop operates whereby a new folding generated imposes new constraints upon the formation of patterns. On the other hand, long-range intrachain contacts and secondary and tertiary interactions are induced by the drastic reduction in the available conformation space due to local correlations.

The validity of the results stemming from our treatment is tested by showing that they basically reproduce the chronology and structural features of experimentally probed intermediates and kinetic bottlenecks in a protein that renatures *in vitro* following a dominant pathway. The probed cooperativity and nucleation effects, as well as diffusion-collision stabilization of secondary structure [20] are shown to result from the persistence of relatively stable patterns through successive perturbation-translation-renormalization cycles, thus acting as seeding patterns or kernels for further structure growth or hierarchical development. This work might prove instrumental given the actual need to provide a detailed microscopic picture, eventually covering the entire time scale spectrum encompassing both the continuum soft-mode dynamics as well as significant folding events [22].

ACKNOWLEDGMENTS

Financial support from the J. S. Guggenheim Memorial Foundation (New York City), and from the Alexander-von-Humboldt Stiftung (Bonn) is gratefully acknowledged.

[1] R. Jaenicke, in *Is There a Code for Protein Folding*? Protein Structure and Protein Engineering, edited by E. L. Winnacker and R. Huber (Springer, Berlin, 1988), pp. 16–36.

[2] O. B. Ptitsyn, and G. V. Semisotnov, in *The Mechanism of Protein Folding*, Conformations and Forces in Protein Folding edited by B. Nall, B. Dill, and K. Dill (American Association for the Advancement of Science, Washington, 1991), pp. 155–168.

[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. Appignanesi, 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] (a) N. Go and H. A. Scheraga, J. Chem. Phys. **51**, 4751 (1969); (b) Macromolecules **9**, 535 (1976).

[12] H. Cendra, A. Fernández, and W. Reartes, J. Math. Chem. **19**, 331 (1996).

[13] (a) S. He and H. A. Scheraga, J. Chem. Phys. **108**, 271 (1998); (b) **108**, 287 (1998).

[14] C. Cantor and P. Schimmel, *Biophysical Chemistry* (Freeman, New York, 1980), Vols. I–III.

[15] (a) Z. Guo and D. Thirumalai, Biopolymers **36**, 83 (1995); (b) M. Guenza and K. F. Freed, J. Chem. Phys. **105**, 3823 (1996).

[16] A. Fernández and A. Colubri, Physica A **248**, 336 (1998).

[17] J. S. Richardson and D. C. Richardson, in *The Origami of Proteins*, Protein Folding, edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), pp. 5–18.

[18] T. G. Oas and P. S. Kim, in *A Protein Folding Intermediate Analog, in ''Protein Folding,''* edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), pp. 123–128.

[19] C. Brooks III, M. Karplus, and B. Montgomery Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics,* Advances in Chemical Physics, Vol. LXXI (Wiley & Sons, New York, 1988).

[20] D. Bashford, M. Karplus and D. Weaver, in *The Diffusion-Collision Model of Protein Folding, in ''Protein Folding,''* edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), pp. 283–290.

[21] T. E. Creighton, *Understanding Protein Folding Pathways and Mechanisms in ''Protein Folding,''* edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, 1990), pp. 157–170.

[22] T. E. Creighton, *Proteins* (Freeman, New York, 1984), pp. 231–234.