Semiempirical prediction of protein folds

Ariel Fernández,[1,2,*] Andrés Colubri,[1] and Gustavo Appignanesi[1,3]
[1]*Instituto de Matemática, Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Sur, Avenida Alem 1253, Bahía Blanca 8000, Argentina*
[2]*Max-Planck Institut für Biochemie, Abteilung Strukturforschung, Am Klopferspitz, Martinsried bei München, D-82152 Germany*
[3]*Departamento de Química e Ingeniería Química, Universidad Nacional del Sur, Bahía Blanca 8000, Argentina*
(Received 13 July 2000; revised manuscript received 14 November 2000; published 10 July 2001)

We introduce a semiempirical approach to predict *ab initio* expeditious pathways and native backbone geometries of proteins that fold under *in vitro* renaturation conditions. The algorithm is engineered to incorporate a discrete codification of local steric hindrances that constrain the movements of the peptide backbone throughout the folding process. Thus, the torsional state of the chain is assumed to be conditioned by the fact that hopping from one basin of attraction to another in the Ramachandran map (local potential energy surface) of each residue is energetically more costly than the search for a specific (Φ, Ψ) torsional state within a single basin. A combinatorial procedure is introduced to evaluate coarsely defined torsional states of the chain defined ''modulo basins'' and translate them into meaningful patterns of long range interactions. Thus, an algorithm for structure prediction is designed based on the fact that local contributions to the potential energy may be subsumed into time-evolving conformational constraints defining sets of restricted backbone geometries whereupon the patterns of nonbonded interactions are constructed. The predictive power of the algorithm is assessed by (a) computing *ab initio* folding pathways for *mammalian ubiquitin* that ultimately yield a stable structural pattern reproducing all of its native features, (b) determining the nucleating event that triggers the hydrophobic collapse of the chain, and (c) comparing coarse predictions of the stable folds of moderately large proteins ($N \sim 100$) with structural information extracted from the protein data bank.

DOI: 10.1103/PhysRevE.64.021901 PACS number(s): 87.10.+e, 87.15.He, 87.15.Cc, 87.14.Gg

I. INTRODUCTION

One of the core long-standing problems in molecular biophysics is the *ab initio* prediction of the native three-dimensional (3D) structure and expeditious pathways of a protein folding under *in vitro* renaturation conditions [1–6]. Recent research [7–10] reveals that a meaningful approach to this problem must reconcile the local conformational constraints imposed by steric hindrances on (Φ, Ψ) torsions of individual residues with the nonbonded potential energy terms responsible for the large-scale organization of the chain.

The local torsional restrictions are determined by the so-called Ramachandran maps [11]. Such plots represent local potential energy surfaces associated with each of the 20 types of residues, mapping the local (Φ, Ψ)-torsional coordinates of each residue onto the energy axis and subsuming all bonded interactions. Thus, this surface is built exclusively taking into account intraunit elastic torsional, dipole-dipole, and Lennard-Jones terms determining the local steric hindrances and propensities that constrain torsional motion.

The basic topographic features of the Ramachandran plot remain invariant throughout the whole series of conformational changes that take place during the folding process [7–10,12]. This fact clearly suggests that an efficient exploration of conformation space may be achieved by separating local terms from nonbonded potential energy contributions, and incorporating the former as determinants of a discretized coarse representation of the torsional state of the chain. Within this discretized framework, local torsional states of individual residues may be viewed modulo the basins of attraction in the Ramachandran plot (R basins): Two local isomers are coarsely regarded as ''the same'' provided they belong to the same R basin. Thus, the basic assumption of our model may be formulated as follows: Since interbasin hopping is slower than intrabasin exploration, the torsional dynamics of the chain are enslaved or subordinated to the sequence of interbasin transitions. Then, since the number of R basins per aminoacid is discrete and small (1-4), the folding problem may be essentially digitalized within a context in which an N sequence of R basins ($N =$ length of the chain) represents a distinctive set of torsional constraints.

This basic assumption leads to a discretized mechanistic picture in which the relevant torsional information is encoded in a time-dependent matrix of N columns ($N =$ length of the chain) called local topological constraints matrix (LTM) [7–10]. To fix notation, let the digits 1, 2, 3, 4 denote, respectively, the R basins containing the extended β-sheet conformation, the compact right-handed (R) α-helix conformation, the compact left-handed α-helix conformation, and the extra basin only present in Gly. In turn, an N-vector of such digits will denote a coarse torsional state (topology) of the entire chain. Within this framework, not only basic secondary motifs may be codified, but also the conformations of hairpin turns and bends [13] may be discretely expressed modulo R basins [10]: Typical topological patterns or consensus windows compatible with β-hairpin two-residue turns [13] are ...1111(33)1111..., ...1111(42)1111..., where hairpin turn windows are given in brackets; similarly, the topological patterns for common reverse turns are: ...1111(13)1111...,

*Author to whom correspondence should be addressed. Email address: arifer@criba.edu.ar

...1111(22)1111..., while the pattern for an R-α-helix turn is ...2222... The recognition of such patterns depends on frustration-tolerant matchings of hydrophobic residues and ion pairs and on the tolerance to torsional incongruities [8], a plasticity that is senseless at the geometric level, where mismatches are energetically penalized and matchings are energetically favored by the nonbonded intramolecular potential. Thus at the topological level, consensus windows such as ...1111(2223)1111... are ''recognized'' as β-sheet hairpins with four-residue turns [8] just like the perfect pattern ...1111(2222)1111... . The stability of such patterns is contingent upon the compensatory role of the enthalpic loss due to contact formation with respect to the actual loss in side-chain torsional entropy and thus, the thermodynamic potential will tend to ''correct'' torsional incongruities.

At this level of description, the LTM evolution is determined by the interbasin transitions whose rates decrease as patterns compatible with structural motifs are recognized in the LTM, an ''if-you-see-it-freeze-it'' computation strategy. On the other hand, the hopping rates increase if existing topological patterns are dismantled due to the formation of a 33% out-of-consensus critical bubble in the LTM [14]. Thus, the mean hopping rate for free residues is $10^{11}\,s^{-1}$, and a deceleration is applied reducing it to 10^{8} or $10^{3}\,s^{-1}$, respectively, depending on whether the residue is detected as part of a secondary or tertiary pattern at the time when the LTM is evaluated [7–10].

The outcome of the pattern recognition is recorded periodically, say every 100 ps—the minimal time for a pattern change (cf. [7–10,14])—as a contact matrix (CM). This recording is based on the topological compatibility of the pattern ''read'' in the LTM *vis-à-vis* a specific structural motif [7,8]. In turn, the CM is determined based on an operational definition of contact: Two residues are in contact when their α-carbon distance is shorter than the maximum distance associated with an energetic decrease of at least $RT/2$ in the longest-range contribution to the intramolecular potential.

A feedback or renormalization mechanism ensures that the CM dynamics will entrain the LTM evolution in the long-time limit as patterns are hierarchically developed [7–10]. Thus, the iteration of two generic operations determines the LTM-CM dynamics: The pattern recognition operation π:LTM$\rightarrow$CM and the renormalization feedback operation ρ:CM$\rightarrow$LTM, prescribing how the next pattern recognition on an LTM is to be performed according to the long-range interactions encoded in the latest CM.

As N gets larger, a clearcut assignment of topological patterns in the LTM is marred by structural ambiguity: Each R basin contains a vast geometric latitude that may encompass different rotameric isomers [11] that will be accomodated in mutually excluding geometries of the whole chain as folding possibilities grow exponentially with N. For instance, the same basin that contains the local (Φ,Ψ) conformation associated with an α-helix turn (nonzero pitch) also contains the local conformations of a two-residue β turn (zero pitch). The structural ambiguity may only get resolved *a posteriori* as structural development causes one pattern to outgrow its competitors, that is, those sharing common consensus windows in the LTM. Initially, both structural motifs might get recorded and then a bifurcation of folding pathways occurs, except that the misfolded pattern will be ephemeral when compared with the one that offers a better possibility for structural development [7–9]. As can be surmised, the level of pathway bifurcation also grows exponentially with N, eventually rendering the topological treatment unmanageable.

While in small proteins the π-ρ loop algorithm is highly efficient in the prediction of nativelike topologies, the structural ambiguity in pattern recognition may become computationally insurmountable for larger proteins. Thus, we need a new and rigorous means to define the π operation by making use of information encoded in the nonbonded potential.

Instead of topologically characterizing each structural motif in order to recognize patterns, as done in previous work, we shall factorize the π map through energetically optimized 3D realizations of each LTM (Fig. 1). In other words, an optimized 3D geometry realizing the LTM will act as a mediator in the recognition of topological patterns. Thus, CM matrices will become directly accessible from the resulting geometries [15].

Based on this conceptual framework, an algorithm for structure prediction is introduced in this work. The algorithm reveals the time evolution of backbone geometries determined by nonbonded potential energy optimization [16] under the constraints imposed by the evolving LTM. This requires engineering a semiempirical potential designed to evaluate the contribution of each type of nonbonded interaction. The LTM's are generated using the geometry-mediated (π-ρ) loop routine. Each LTM serves as the set of constraints for geometric optimization, and reciprocally, each optimal geometry determines how to generate the next LTM.

The fact that the model focuses on the backbone conformation might lead one to the belief that other degrees of freedom such as side-chain dihedral torsions have been disregarded. This is not so. Our model incorporates implicitly a level of conformational detail that enables a self-consistent means of coarsely generating torsional dynamics of the peptide backbone. This does not entail truncating the number of torsional degrees of freedom and retaining only the (Φ, Ψ) backbone description: While torsional degrees of freedom of the side chains are not included explicitly as dynamic variables, they surface in the model as conformational entropy and define the strength of two-body contributions, both determinants of the coarse torsional dynamics of the backbone. On the other hand, in order to obtain a self-consistent (Φ, Ψ) description of backbone dynamics, as provided by the geometry-mediated (π-ρ) algorithm, we must go into further detail *vis-à-vis* the backbone itself, incorporating the dipoles localized along the chain. This is necessary to construct two nonbonded contributions: the dipole-dipole and hydrogen-bond terms, in turn determinants of the geometry-mediated dynamics.

Furthermore, within the timescales relevant to an adiabatic separation of side-chain and backbone motions, the residues are regarded as solid ellipsoids centered at the α-carbons with three effective Lennard-Jones radii. Thus, besides the standard value of 1.85 Å adopted as the radius along the virtual bond joining adjacent aminoacids—the

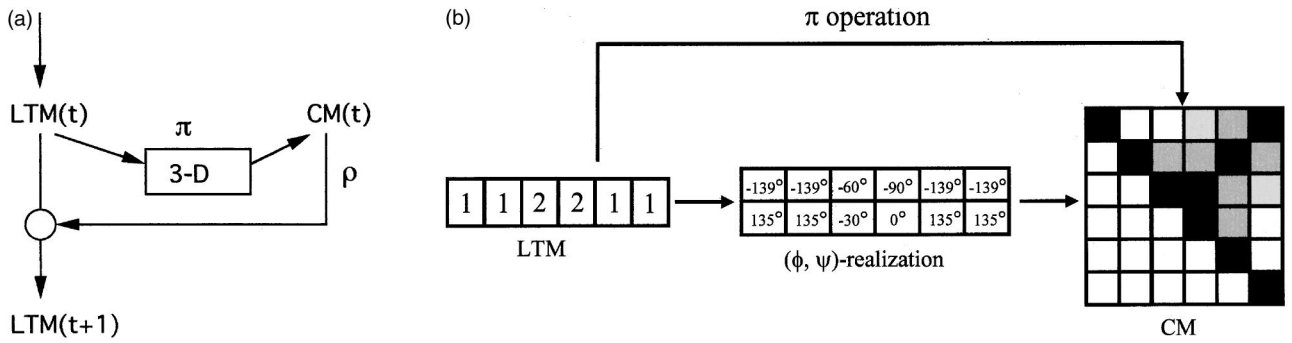


FIG. 1. (a) Scheme of a single pattern-recognition-and-feedback (π-ρ) iteration in the simulation of the torsional dynamics as entrained by its coarse version where torsional states are viewed with regards to the Ramachandran basin to which they belong. The topological or coarse dynamics are monitored by the time evolution of the local topological constraints matrix (LTM). Two basic operations determine the enslavement of the LTM dynamics to the long-range organization encoded in the contact matrix (CM): The pattern recognition (π) and the renormalization operation (ρ). To remove structural ambiguity, the former may be mediated through an optimal 3D realization of the LTM, a preliminary operation that enables the determination of contact patterns not systematically identifiable in the LTM. The operations required to determine the LTM at two consecutive discretized instants t and $t+1$ are represented [7–10]. The feedback renormalization affects the way the LTM $(t+1)$ is generated from LTM (t), and also the way the LTM $(t+1)$ matrix is evaluated to detect patterns recorded in CM $(t+1)$. (b) Illustration of a geometry-mediated detection of a simple pattern: A β-sheet motif has been recognized in the LTM by optimizing torsional values within the R basins given in the LTM.

''along-backbone dimension''—the residue is endowed with two other dimensions, surfacing depending on whether side chains interact laterally—as in secondary structure—or they are engaged in a headon collision leading to a tertiary interaction.

To summarize, there is an essential premise upon which both the previous topological (π-ρ) loop algorithm and the new unambiguous geometry-mediated algorithm are based: In order to explore a region in the torsional space of a residue, it is first necessary that the residue finds its correct basin (R basin) in the local potential energy surface or Ramachandran map [7]. By ''correct basin'' we simply mean the basin that contains the targeted region in (Φ, Ψ) space. Since interbasin hopping is far slower than intrabasin equilibration, it is natural to assume that interbasin hopping enslaves or subordinates the folding process when the latter is viewed as the long-time limit of torsional dynamics. Thus we can claim that the folding problem may be reduced to the problem of determining the time evolution of torsional constraints, where such constraints may be regarded as the R basins to which residues are confined at a given time. This idea stands in sharp contrast with significant efforts aimed directly at computing the torsional dynamics, or simplified versions, or caricatures of them [2–4].

In view of this, the entire problem of *ab initio* prediction of folding pathways boils down to devising a means of computing the interbasin hopping dynamics. A first attempt has been given in [7—10]. In this earlier version, the evolution of torsional constraints is encoded in a time-evolving matrix (the LTM), which assigns an R basin to each residue and chooses interbasin hopping rates (transition rates) from Gaussian distributions picked according to the topological patterns detected in the LTM. Thus, if a residue is free at a given time, that is, it is not recognized as part of any topological pattern, its mean interbasin hopping time is 10 ps; likewise, if it is recognized as part of secondary or tertiary structure, its mean transition times are 10 ns and 1 ms, respectively [7–10]. This simulated dynamics of torsional constraints has some inherent limitations. (a) It fails to give an accurate description when the topological (modulo R basin) resolution is structurally ambiguous, or different structural patterns are possible for a given combination of R basins [7]. (b) The resolution of hopping rates is not fine enough to truly take into account the extent of structural engagement of a residue (free, secondary, or tertiary are simply not fine enough categories as indicated below).

Thus, the new algorithm given in this work serves a purpose: In its implementation we have taken care of the caveats mentioned above. The time evolution of torsional constraints is now based on an optimized geometric realization of the LTM, while the interbasin hopping probability at a given time depends directly on the extent of structural involvement of the residue at that time. The latter quantity is computed by evaluating the energetic and entropic changes that would result in the entire structure if the given residue would change its R basin. In this way the structural multiplicity resulting from working directly at the topological level is removed, while a finer resolution of basin hopping rates reflects the fact that the more structurally engaged a residue is, the least likely it will be prone to change its R basin.

II. THE ALGORITHM

The algorithm introduced in this work hinges on the fact that local contributions to the potential energy may be subsumed into time-evolving conformational constraints defining sets of restricted backbone geometries framing each pattern of nonbonded interactions [7–10,17]. In turn, each such pattern dictates how the new set of constraints will emerge and thus conditions the way in which the new pattern will be generated. This approach may be contrasted with existing algorithms where all terms—local and nonbonded—are treated together as a whole [18] without introducing hierarchical considerations.

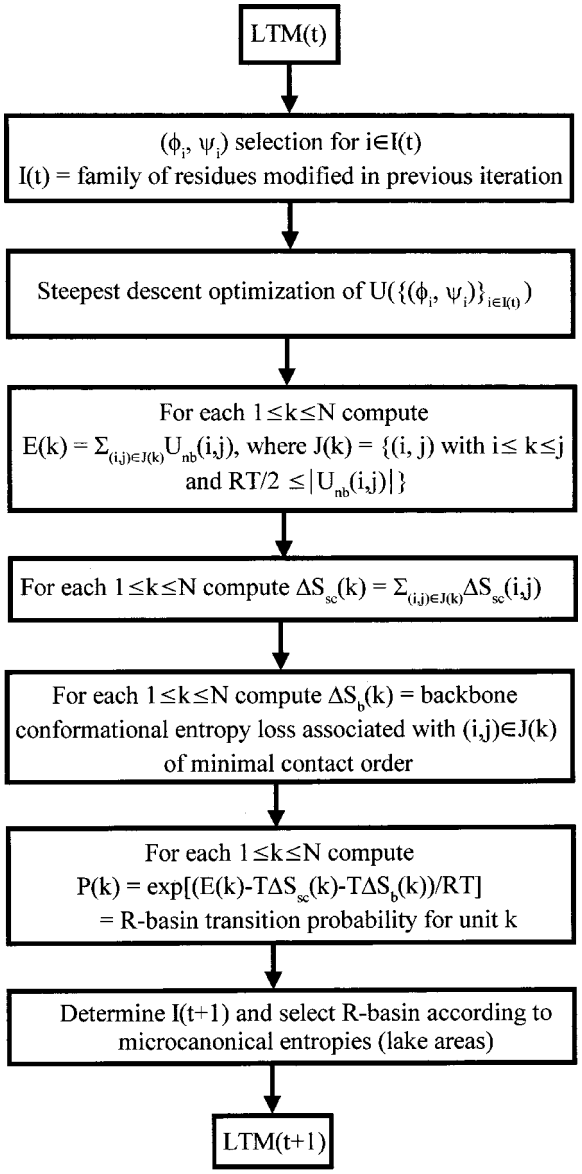


FIG. 2. Scheme of a single geometry-mediated π-ρ iteration carrying the system from the chain topology LTM(t) to LTM($t+1$).

A single complete iteration of our algorithm is schematically given in Fig. 2, which describes a single 3-D-mediated π-ρ iteration. Each such iteration maps one topological state of the chain, LTM(t), onto the next, LTM($t+1$), a transition entailing 10 ps in real time, the minimal time for a significant change in contact patterns involving the fastest interbasin hopping for free residues (cf. [7–10,14,17]). Thus, each iteration involves the following steps.

(1) A (Φ, Ψ)-coordinate assignment within the chosen R basins for LTM(t) is performed. The assignment is restricted to those residues belonging to the family $I(t)$ prescribed to have changed their R basins in the previous iteration. This assignment is based on an intra-R-basin probability distribution of torsional coordinates. In this work we have adopted a (Φ, Ψ) plotting for each residue obtained from a structural database of 163 proteins with 1-Å resolution or better, generated by the PROCHECK program [15]. Our program is equipped to accommodate alternative database sources for the torsional realization of an R basin.

(2) A steepest-descent optimization [16] is performed adopting an effective semiempirical potential $U=U_{\rm nb}+U_R$ made up of a nonbonded potential $U_{\rm nb}$, and a local elastic Ramachandran term denoted U_R, penalizing departures from the local (Φ, Ψ) minimum in each R basin [17]. The nonbonded potential has been especially engineered to incorporate all meaningful pairwise additive contributions of the form $U_{\rm nb}(i,j)$ representing interactions between pairs (i,j)'s of nonadjacent aminoacids, while U_R is made up of single-unit $[U_R(j)]$ contributions. The optimization is confined to the set of residues $I=I(t)$ that have been required to change their respective R basins in the previous interation LTM($t-1$)$\rightarrow$LTM(t). Thus, the optimization process is confined to the set of variables $\{(\Phi_i,\Psi_i)\}$, $i\in I(t)$, where (Φ_i,Ψ_i) denotes the torsional state of residue i. The detailed backbone conformation for the remaining aminoacids obtained in the LTM($t-1$)$\rightarrow$LTM(t) iteration is retained.

(3) The CM(t) contact matrix is computed directly from the optimized geometry $\{(\Phi_i^*,\Psi_i^*)\}$, $i=1,...,N$, obtained in step (2), by determining the distances between nonadjacent α carbons. The aminoacid pair (i,j) is considered to be in contact if $U_{\rm nb}(i,j)\leqslant -RT/2$.

(4) For each detected contact between residues i and j, we compute the effective loss in side-chain entropy $\Delta S_{\rm sc}(i,j)=\Delta S_{\rm sc}(i)+\Delta S_{\rm sc}(j)$ associated with the partial torsional constraints imposed on the side chains of the aminoacids involved in the (i,j) contact.

(5) The family $J(k)$ of pairs (i,j) of residues such that $i\leqslant k\leqslant j$ and $|U_{\rm nb}(i,j)|\geqslant RT/2$ is determined. This is the family of attractive (contact) or repulsive (anticontact) interactive pairs whose survival depends on the torsional conformation adopted by residue k.

(6) We compute the minimal loss in backbone conformational entropy $\Delta S_b(k)$ associated with the closure of a loop of size $L(k)$, where $L(k)$ is the minimal contact order (chain contour distance in number of aminoacids) for all pairs in $J(k)$. Thus, if $L(k)=j^*-i^*$ with $i^*,j^*\in J(k)$, we get $0\geqslant\Delta S_b(k)=R\ln[\Pi_{i^*<j'<j^*}A_{j'}/4\pi^2]$, where $A_{j'}$ is the microcanonical lake area of the R basin [17] required for loop unit j' to be in, and $4\pi^2$ is the total area of the product of two unit circles (one for each local torsional variable). The lake areas of residues are computed by considering the area enclosed by the contour equipotential line in the Ramachandran map that includes the lowest-lying saddle point. Such a contour line is identified with the basin rim. These microcanonical entropies are given in Table I and have been estimated from a protein structure database obtained from the PROCHECK program [15]. Thus, the probability of falling in a specific R basin for a particular aminoacid is empirically determined as the number of plotted (Φ, Ψ) points belonging to that R basin and taken from different native folds for the specified aminoacid, divided by the total number of plotted points for that aminoacid [17].

(7) For each residue k with $1\leqslant k\leqslant N$, we compute the inter-R-basin hopping probability within timespan

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

Ramachandran typology	Basin 1	Basin 2	Basin 3	Basin 4
Ala-like	0.52	0.40	0.08	0.00
Gly	0.26	0.24	0.30	0.20
Prec. pro	0.78	0.00	0.22	0.00
Pro	0.51	0.49	0.00	0.00

$(t,t+1)$, $p(k)=\exp\{[\Sigma_{(i,j)\in J(k)}\{U_{\rm nb}(i,j)-T\Delta S_{\rm sc}(i,j)\}-T\Delta S_b(k)]/RT\}$ if the exponent is smaller than zero and $p(k)=1$ otherwise. In other words, the torsional mobility of a residue depends on the stability of the pattern that may be disrupted by its inter-R-basin hopping. This prescription reflects the essential operating tenet of the π-ρ loop algorithm: The R basin hopping rate of a residue depends on the level of structural engagement of the residue. In turn, the level of engagement of residue k is quantified by the overall free energy term: $\Sigma_{(i,j)\in J(k)}[U_{\rm nb}(i,j)-T\Delta S_{\rm sc}(i,j)]-T\Delta S_b(k)$. The backbone entropic contribution, $-T\Delta S_b(k)$, is necessary to ensure the cooperativity in the emergence of stable long-range organization: If a long-range contact is formed, its fragility is due to its high entropic cost. Thus, there will be a residue k in the loop region for which the term $-T\Delta S_b(k)$ will be large and positive thus favoring its inter-basin hopping [high probability $p(k)$]. This hopping will obviously lead to the concurrent destruction of the long-range contact. On the other hand, if the long-range contact emerges only after several contacts have formed within its putative loop, the backbone entropy contribution $-T\Delta S_b(k)$ for any residue k belonging to the remaining portions of the loop will be comparatively small (and positive), thus the residue k will be less prone to change its R basin. This implies that the long-range contact formed in a cooperative ''hierarchical'' fashion is more stable than one formed at a large entropic expense, in accord with known observations [6].

(8) To generate LTM$(t+1)$, we take into account two facts: (a) Unit k changes R basin within timespan $(t,t+1)$ with probability $p(k)$, and thus the family $I(t)$ of residues that change their R basin in the time interval $(t,t+1)$ is determined: (b) The probability of hopping in residue k from R basin n to R basin m ($n,m=1,2,3,4$ as applicable, cf. Sec. I) is given by the quotient $A(m)/|B(k)-A(n)|$, where $A(n),A(m)$ are, respectively, the lake areas [17] or microcanonical entropy areas of R basins n and m, and $B(k)$ is the sum of all such lake areas for residue k, a fraction of the total (Φ, Ψ)-area $2\pi \times 2\pi = 4\pi^2$.

The structural fluctuations may be investigated by examining the behavior of $I(t)$. The family $I(t)$ becomes progressively smaller as large-scale organization arises and becomes consolidated. This trend, in principle easing the task of the optimization mechanism, is counteracted by the fact that the region of the potential energy surface explored in these later stages becomes progressively more rugged: After hydrophobic collapse, further hydrophobic interactions entail more and more excluded volume effects (anticontacts) because of chain compactification. This implies that the short-range repulsive forces of Lennard-Jones type become ubiquitous, defining an increasing ruggedness in the energy landscape and rendering the optimization method ineffective due to the profuseness of energy traps.

Step (4) assumes that the working model for the peptide chain is essentially an α-carbon backbone model that subsumes the geometry of the side chain as an entropic contribution, $\Delta S_{\rm sc}$. This side-chain entropy is conformation dependent since a free residue possesses a larger ''effective torsional volume'' than a residue engaged in a contact pattern. In turn, the effective torsional volume may be regarded as the product of available unit-circle regions for each torsional degree of freedom of the side chain, as shown in Sec. IV.

In consistence with this view, the repulsive Lennard-Jones contribution to the nonbonding potential is tailored by regarding each residue as an ellipsoid: the van der Waals's radius of a residue along the virtual fixed-length bond joining adjacent α carbons [11] is fixed at 1.85 Å, that is, half of the contour distance between residues. A slightly smaller value, $r_{\rm vW}=1.75$ Å, has been adopted along the direction of secondary interactions. That is so since side chains in α helices or β sheets interact laterally and thus their actual dimensions do not alter the backbone geometry. However, in tertiary interactions, side chains contribute to increase the effective dimension $r_{\rm vW}^{\perp}$ of the residue along the direction of interaction. Thus, the ''orthogonal'' van der Waals's radius $r_{\rm vW}^{\perp}$ for the residue will be fixed at $r_{\rm vW}^{\perp}=2.5$ Å, the average dimension of the side chain felt in a ''headon'' collision.

III. ENGINEERING AN INTRAMOLECULAR POTENTIAL

A nonbonded intramolecular potential is engineered to search—via steepest descent or equivalent optimization methods [16]—for suitable backbone geometries under the constraints dictated by the LTM. This enables us to detect the optimal contact patterns CM's compatible with a given LTM, an operation that prescribes how the next LTM is to be generated according to the steps (1)–(8) defined in Sec. II.

In order to treat each LTM or N vector of R basins as a set of constraints within which optimization is performed, we need to penalize energetically departures from the local optimization in the Ramachandran map. Such deviations are determined by the influence of nonbonded terms responsible for the onset of large-scale organization. Thus, the appropriate engineering of a suitable potential is paramount to succeed in the pattern identification.

The effective intramolecular potential U must be taken to be $U=U_{\rm nb}+U_R$, where $U_{\rm nb}$ is the nonbonded contribution and U_R denotes an empirical local contribution introduced

here to penalize deviations from the energy minimum in the *R* basin dictated by the LTM. This last contribution is then given as

$$U_R = \sum_{j=1,\ldots,N} U_{R,j}$$

$$= \sum_{j=1,\ldots,N} C_j[1-\cos(\Phi_j-\Phi_{j,0})]^2 + G_j[1-\cos(\Psi_j-\Psi_{j,0})]^2$$

$$\approx \sum_{j=1,\ldots,N} C_j(\Phi_j-\Phi_{j,0})^4 + G_j(\Psi_j-\Psi_{j,0})^4, \qquad (1)$$

where $U_{R,j}$ is the effective local distortion energy at residue j, C_j and G_j are local elastic moduli [18,19], and $\Phi_{j,0}, \Psi_{j,0}$ are the minima in a given *R* basin for residue j. The moduli are chosen so that the 5-kcal/mol equipotential contour in the Ramachandran plot serves as a rim for the *R* basin (cf. [11], Vol. 1, p. 268). Thus, for a 1-alanine-like residue (with three possible *R* basins), the moduli for the *R* basin 1, containing the extended β-sheet conformation, must be chosen so that an $\sim 80°$ increment in $\Phi_j-\Phi_{j,0}$ or $\Psi_j-\Psi_{j,0}$ represents an energy increase of 5 kcal/mol.

In contrast with any previous treatments of the problem [1–6,18], $U_{\rm nb}$ is engineered semiempirically to reproduce major large-scale nonbonded interactions as well as middle-range or local structural refiners also serving as structure buttresses,

$$U_{\rm nb} = U_{\rm LJ} + U_{\rm solv} + U_{\rm Coul} + U_{\rm dip} + U_H + U_{\rm SS}. \qquad (2)$$

The terms in the right-hand side of Eq. (2) denote, respectively, the Lennard-Jones repulsive term determining the excluded volume effect ($U_{\rm LJ}$), the effective solvophobic term for the sum of pairwise attractions between solvophobic residues [17] and pairwise repulsive interactions between polar and hydrophobic residues ($U_{\rm solv}$), the sum of effective Coulombic ion-pair interactions ($U_{\rm Coul}$), the sum of nonbonded dipole-dipole pairwise interactions ($U_{\rm dip}$), the amide hydrogen-bond N—H...O═C backbone interactions (U_H) and the disulfide bridging between Cys (cysteine) residues ($U_{\rm SS}$).

The determination of each type of nonbonded interaction is contingent on a suitable classification of the aminoacids, a seemingly controversial issue. In this work we have adopted the one given in Fig. 3, which has been obtained by chemical intuition applied in sorting out all types of side chains.

A. The Lennard-Jones repulsive contribution $U_{\rm LJ}$

The nonlocal steric hindrances determining the excluded volume effect are incorporated as an effective contribution $U_{\rm LJ}$, which penalizes energetically incursions of an α carbon into the neighborhood of another as $(r-2r_{\rm vW})^{-12}$ if the approach is sideways, as in a secondary structure motif [11,17], and as $(r-2r_{\rm vW}^{\perp})^{-12}$ if the approach is headon and leads to a favorable tertiary interaction.

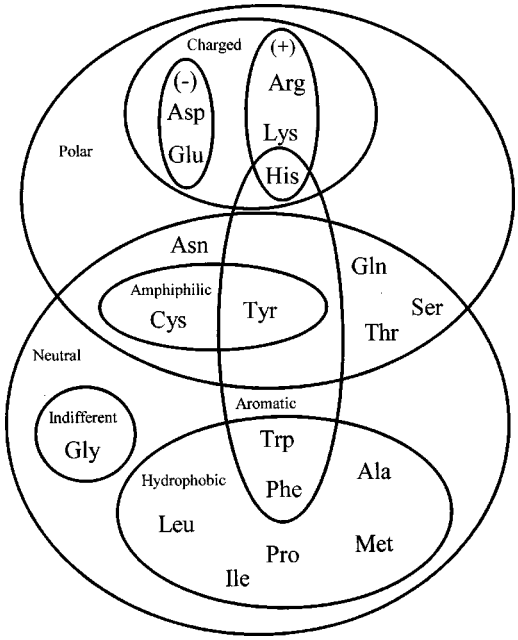


FIG. 3. Set-theoretic (Venn-diagram) classification of the 20 aminoacids.

B. The effective solvophobic contribution $U_{\rm solv}$

In contrast with Coulombic two-body terms [20,21], the solvophobic potential is an effective contribution [22–24] stemming from an entropic solvent effect, and as such it is typically excluded from molecular dynamics simulations (cf [18]). An effective solvophobic intramolecular potential arises as a result of the entropy-driven solvophobic effect [23]: The solvophobic association of nonpolar groups is known to be due to the need to minimize the entropy loss associated with the ordering of solvent around nonpolar moieties, an effect not compensated by enthalpy-lowering favorable interactions between the moieties and the solvent [24]. It has been shown [14,23–25] that the net free energy decrease due to the formation of a hydrophobic (*h*-*h*) contact may be rationalized as a surface-tension effect [23], and as such it is proportional to the change in solvent-exposed area with a proportionality constant estimated at ~ 78 cal/Å^2 mol (cf. [25]). Thus, while for the extended system (protein molecule+solvent), the *h*-*h* association is viewed as entropically driven, the restricted system (protein molecule) experiences a solvophobic force due to the tendency to minimize the solvent-exposed area. Adopting average dimension parameters for the side chains of the amphiphilic or hydrophobic residues [25], we fix the exposed area change at 48 Å^2 for contacts between relatively small hydrophobic residues (Ala, Pro), or between relatively small and relatively large residues (Val, Ile, Leu, Phe, Met, Trp), and ~ 70 Å^2 if such contacts involve exclusively the relatively large hydrophobic residues. This gives an average energy change per *h*-*h* contact of -3.8 kcal/mol and -5.47 kcal/mol, respectively (cf. [14,17]). This semiempirical parametrization enables our model to distinguish between ''nuclear residues,'' that is, those instrumental in creating a relatively stable nucleus triggering the hydrophobic collapse (Val, Ile, Leu, Phe, Met,

TABLE II. Multiplicative table indicating the empirically estimated enthalpic change (an adjustable parameter in our simulations) associated with a putative contact involving different types of residues. The repulsive interaction between a hydrophobic (or very hydrophobic) residue and a polar one is also incorporated. In the case of two polar residues, only ion pairs involving Glu^-, Asp^-, His^+, Lys^+, or Arg^+ become meaningful interactions ($U(i,j) \geq -RT/2$) under physiological acidic conditions with $4 \leq p\mathrm{H} \leq 6$ (see text).

	Very hydrophobic	Hydrophobic	Amphiphilic	Neutral	Polar
Very hydrophobic	−5.47	−4.56	−4.56	0	0.96
Hydrophobic	−4.56	−3.8	−3.8	0	0.8
Amphiphilic	−4.56	−3.8	−3.8	0	0
Neutral	0	0	0	0	0
Polar	0.96	0.8	0	0	

Trp) and the remaining hydrophobic residues. On the other hand, the intercalation of a single water molecule between two *h*-beads (or two amphiphilic beads) would drive them appart within the range of thermal fluctuations [17,24], fixed at $r \sim 8$ Å within the coarse topological description of the backbone inherent to our model.

To define empirically the solvophobic potential we first notice that the solvent-exposed area of a hydrophobic or amphiphilic (Cys, Tyr) residue is reduced depending on the contact hierarchy or level of hydrophobic burial to which the residue belongs [25], while the free energy of contact formation is linearly dependent on the concurrent reduction of the exposed area. Thus, the two-body terms for pairs of free hydrophobic or amphiphilic residues adopt the form of a flat well in the region $r = 5.7\text{–}7$ Å, according to typical extreme distances of secondary and tertiary structure interactions [11]. The insensitivity of the results with respect to the different shapes of this effective potential holds true provided the most favored proximity range 5.7–7 Å and the negligible-force range $r > 8$ Å remain invariant.

On the other hand, the burial dependence of the solvophobic force must be incorporated. Thus, two-body scaling factors λ_s, $\lambda_t (0 < \lambda_t < \lambda_s < 1)$ are introduced to account, respectively, for residues already engaged in secondary or tertiary structure and thus having undergone already partial reductions of their solvent-exposed areas. Because the strength of the contact depends on the reduction of the exposed area associated with hydrophobic pairing, the residue with the previous highest-order contact hierarchy is the one that determines the strength of the putative contact. Thus, the solvophobic potential for a specific backbone geometry reads

$$U_{\mathrm{solv}} = \sum_{(i,j)\ \mathrm{in}\ W} U_{\mathrm{solv},ij}(r_{ij}) + \lambda_s \sum_{(i'j')\ \mathrm{in}\ W'} U_{\mathrm{solv},i'j'}(r_{r'j'}) + \lambda_t \sum_{(i''j'')\ \mathrm{in}\ W''} U_{\mathrm{solv},i''j''}(r_{i''j''}), \quad (3)$$

where W is the family of pairs (ij) of free hydrophobic or amphiphilic residues along the chain with $j \geq i+3$, W' is the family of pairs with *at least one* residue in the pair engaged in secondary structure, and W'' is the family of pairs with *at least one* residue in the pair engaged in tertiary structure: and r_{ij}, $r_{i'j'}$, and $r_{i''j''}$ are, respectively, the distances between residues i and j, i' and j', and i'' and j'', obtained after optimization of local torsional coordinates within the R basins that define the LTM.

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

The solvophobic term U_{solv} also incorporates unfavorable (repulsive) interactions between polar or hydrophillic residues (Lys, Arg, His, Asp, Glu, Asn, Gln, Ser, Thr) and hydrophobic residues, and favorable interactions between amphiphilic and hydrophobic residues. Thus, the ''multiplicative'' Table II gives the potential energy contribution in kcal/mol associated with favorable and unfavorable pairwise hydrophobic interactions at the distance range 5.7–7 Å.

C. The effective Coulombic contribution U_{coul}

The effect of ion pairs such as Glu^-, Lys^+ on the stabilization of secondary structure under wide pH ranges (2–12) has been established by Marqusee and Baldwin [26]. Thus, using oligopeptide probes, it has been shown that an i, $i+3$, or i, $i+4$ spacing of ion pairs involving potentially charged residues (His^+, Lys^+ or Arg^+, Glu^- or Asp^-) becomes a stabilizing factor for an α helix commensurate with the required hydrophobic periodicity [26]. For this reason, our semiempirical treatment incorporates a potential well of ~ 1 kcal/mol trapping α carbons for ion pairs (i,j) with the appropriate contour spacing ($|j-i| \geq 3$) and lying within the typical α-carbon contact range 5.7–7 Å.

In contrast with the charged groups His^+, Lys^+, and Arg^+, under typical physiological acidic conditions $4 \leq p\mathrm{H} \leq 6$, a partial counterion titration of carboxylic side chains (Asp^-, Glu^-) is assumed to occur, rendering their two-body repulsive coulombic interactions negligible at all spatial ranges beyond their effective van der Waals's radii. Their empirical charges defining the long-range Coulombic forces are modulated accordingly so that their repulsive energy be-

comes meaningful ($\geqslant RT/2$) under alkaline physiological conditions in the absence of counterions.

D. The nonbonded dipole-dipole contribution U_{dip}

In order to account for the structural tuning due to nonbonded dipole-dipole piling interactions in α helices, we have incorporated a sum of products of elastic contributions responsible for helix refinement of (Φ, Ψ) coordinates (cf. [27]). The ith term of the sum indicates a structure refinement propensity towards the optimal helix coordinates (Φ_h, Ψ_h) in the contour interval $(i, i+4)$. This propensity is enforced only if the aminoacids in that interval lie within a certain threshold value of the optimal coordinates. The threshold must obviously warrant that the units lying in R basin 2 will not abandon this basin throughout the optimization process. The mathematical expressions for such a refiner read

$$U_{\mathrm{dip}} = \sum_{i=1,\ldots,N} D(i) U_{\mathrm{dip}}(i, i+4), \tag{4}$$

$$D(i) = \prod_{i=1,\ldots,i+4} \chi_{i'}(\Phi_{i'}, \Psi_{i'}) \tag{5}$$

with

$$\chi_{i'}(\Phi_{i'}, \Psi_{i'}) = 1 \quad \text{if} \quad |\Phi_{i'} - \Phi_h| < \Delta$$

$$\text{and} \quad |\Psi_{i'} - \Psi_h| < \Delta$$

$$\text{and} \quad \chi_{i'}(\Phi_{i'}, \Psi_{i'}) = 0 \quad \text{otherwise;} \tag{6}$$

and

$$U_{\mathrm{dip}}(i, i+4) = K \left[\sum_{i'=1,\ldots,i+4} c(i'-i) [(\cos \Phi_{i'} - \cos \Phi_h)^2 + (\cos \Psi_{i'} - \cos \Psi_h)^2] \right]. \tag{7}$$

The threshold value $\Delta \approx 22°$ and the elastic module $K \sim -1.8$ kcal/mol are adopted to determine the $i, i+3$ and $i, i+4$ contributions to the modulation of torsional coordinates leading to helix refinement without outcompeting other structural motifs. The relative weights of such contributions [27] have been also incorporated: $c(1)=0$, $c(2)=0.5$, $c(3)=1$, $c(4)=0.4$, as indicated in Eq. (7). It should be noted that the $(i, i+4)$ cumulative N—H···O═C hydrogen bonding contribution to the fine tuning and refinement of (Φ, Ψ) helix coordinates reinforces the dipole-dipole nonbonded contribution.

E. The empirical hydrogen-bond potential U_{H}

As stated originally by Pauling and co-workers [28,29], hydrogen bonds restrain the peptide chain to its native conformation. Thus, such bonds will be regarded as structure buttresses rather than determinants of the chain topology. In this sense, we have engineered an (ij) two-body potential energy contribution,

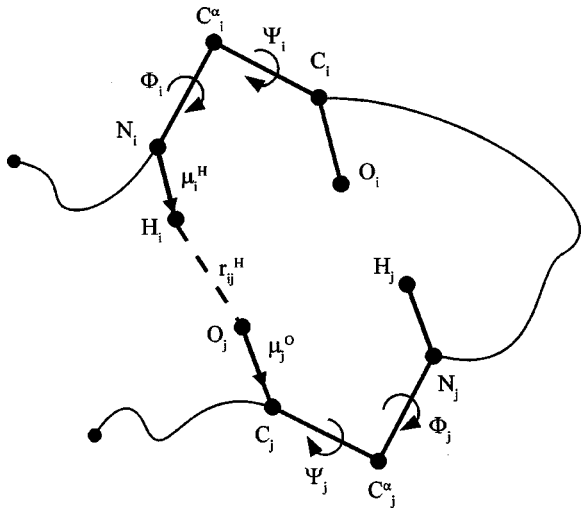


FIG. 4. Scheme of one of the two *a priori* possible H-bond interactions between residues i and j with $|j\text{-}i| \geqslant 4$. The interaction is distinguished by the identification of the amide proton donor which, in the case illustrated, is residue i.

$$U_{\mathrm{H}} = \sum_{4 \leqslant (j-i)} U_{\mathrm{H},ij} = \sum_{4 \leqslant (j-i)} (u_{\mathrm{H},ij} + u_{\mathrm{H},ji}), \tag{8}$$

where $u_{\mathrm{H},ij}$ denotes the energetic contribution resulting when residue i is the donor of the amide proton and residue j is the donor of the carboxylic oxygen electron pair (see Fig. 4), while $u_{\mathrm{H},ji}$ denotes the alternative contribution in which residue j is the donor of the amide proton and residue i, the acceptor. Thus, we have

$$u_{\mathrm{H},ij} = F_{\mathrm{H}}(r_{ij}^{\mathrm{H}}) G_{\mathrm{H}}(\boldsymbol{\mu}_i^{\mathrm{H}}, \boldsymbol{\mu}_j^{\mathrm{O}}), \quad u_{\mathrm{H},ji} = F_{\mathrm{H}}(r_{ji}^{\mathrm{H}}) \cdot G_{\mathrm{H}}(\boldsymbol{\mu}_j^{\mathrm{H}}, \boldsymbol{\mu}_i^{\mathrm{O}}), \tag{9}$$

$$G_{\mathrm{H}}(\boldsymbol{\mu}_i^{\mathrm{H}}, \boldsymbol{\mu}_j^{\mathrm{O}}) = -2 \ \text{Kcal/mol} + 0.001 \ \text{Kcal/mol deg}^2 [\cos^{-1}(\boldsymbol{\mu}_i^{\mathrm{H}} \cdot \boldsymbol{\mu}_j^{\mathrm{O}})]^2. \tag{10}$$

The variables r_{ij}^{H}, $\boldsymbol{\mu}_i^{\mathrm{H}}$, $\boldsymbol{\mu}_i^{\mathrm{O}}$ denote, respectively, the distance between the amide hydrogen atom of aminoacid i and the carbonyl (C═O) oxygen atom of aminoacid j, the N—H oriented bond unit vector and the O═C oriented bond unit vector of aminoacid j (cf. Fig. 4). In this notation, the departures from collinearity are given in degrees and measured by the quantity $\cos^{-1}(\boldsymbol{\mu}_i^{\mathrm{H}} \cdot \boldsymbol{\mu}_j^{\mathrm{O}})$. In the case of a perfectly collinear N—H···O═C hydrogen bond between units i and j, both unit vectors $\boldsymbol{\mu}_i^{\mathrm{H}}$ and $\boldsymbol{\mu}_j^{\mathrm{O}}$ are colinear ($\boldsymbol{\mu}_i^{\mathrm{H}} \cdot \boldsymbol{\mu}_j^{\mathrm{O}} = 1$). The term $-F_{\mathrm{H}}(r_{ij}^{\mathrm{H}})$ represents an inverted normal bell potential well of depth -1 with two inflexion points at $r_{ij}^{\mathrm{H}} = 1.8$ Å and 2.4 Å, representing typical minimal and maximal H-bond distances, and vanishing in the region $r_{ij}^{\mathrm{H}} < 1.4$ Å, $r_{ij}^{\mathrm{H}} > 2.8$ Å [28,29].

The strain energy associated with departures from N—H···O═C collinearity is known to weaken the H bond [28]. This fact is reflected in an energy increase from the optimal $\Delta\Delta H$ value fixed at -2 Kcal/mol ($\Delta\Delta H$ refers to the enthalpy difference between amide H bonding with water and helical i-$i+4$ H-bonding in a solvent-excluded environment) [29]. As indicated in Eq. (8), a semiempirical quadratic factor has been adopted to model such an energy increase as a function of the angular departure from collinearity. The form of this semiempirical angular factor and the quadratic form of the angular distortion energy given in Eq. (10) have been adopted earlier by Pauling and Corey [28].

The stabilizing effect of hydrogen bonds within a hydrophobic environment is more pronounced for the β-sheet motif than for the α helix [29]. This enhancement is effectively modeled since the potential well $-F_{\mathrm{H}}(r_{ij}^{\mathrm{H}})$ reaches its minimum at the β-sheet value $r_{ij}^{\mathrm{H}}=1.8$ Å, while it is some 40% higher at the α-helix value $r_{ij}^{\mathrm{H}}=1.9$ Å.

F. The disulfide potential U_{SS}

This term is treated as an effective hydrophobic well (cf. Sec. III C [7–10,17]) associated exclusively with the pairs of amphiphilic Cys residues. The depth of the well can be modulated according to the chosen redox conditions in the solvent. Such a depth will in turn determine whether or not the disulfide chemistry will entrain or subordinate the folding process, a situation which, like the slow *cis-trans* isomerization of the peptide bond adjacent to proline [11], requires special modeling beyond the scope of this work.

IV. THE SIDE-CHAIN ENTROPIC CONTRIBUTION ΔS_{sc}

Ours is an effective α-carbon model and as such it does not incorporate explicitly the side-chain geometry. Rather side-chain torsional motion is integrated out as an entropic contribution that depends on the backbone conformation. Thus, let $\Delta S_{\mathrm{sc}}(i,j)$ denote the change in side-chain conformational entropy associated with the (i,j)-contact formation. Then $\Delta S_{\mathrm{sc}}(i,j)$ is given by

$$\Delta S_{\mathrm{sc}}(i,j)=R\sum_{k=i\,j}\ \sum_{m=1,\ldots,\zeta(k)}\ln[W_{k,m}/\Omega_{k,m}], \qquad (11)$$

Here $m=1,\ldots,\zeta(k)$ labels the different side-chain torsional degrees of freedom for residue k, and $W_{k,m}$ represents the perimeter measure of the portion of unit circle available to the mth torsional variable for residue k when engaged in the (i,j) pair, while $\Omega_{k,m}=2\pi$ represents the "torsional volume" available to the mth degree of freedom for the free residue k. The ζ values for each kind of residue (Table III) are obtained by counting the number of unconstrained dihedrals of the side chain in the free residue. The following expression, valid only for the engaged hydrophobic residues, has been adopted to simplify the computations,

$$\prod_{m=1,\ldots,\zeta(k)}[W_{k,m}/\Omega_{k,m}]\approx q^{\zeta(k)}, \qquad (12)$$

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

Side chain entropy exponent (ζ)
Ala 1
Val 3
Leu 4
Ile 4
Gly 0
Pro 0
Cys 2
Met 4
His 2
Phe 2
Tyr 3
Trp 2
Asn 2
Gln 3
Ser 2
Thr 3
Lys 5
Arg 6
Asp 2
Glu 3

where $q\approx 2.8$ (cf. [14,30,31]) is a side-chain torsional restriction factor that holds whenever a free residue becomes engaged in a hydrophobic contact. This restriction factor has been selected so that the net free energy change associated with a single hydrophobic contact is of the order of -1 kcal/mol [31].

V. RESULTS

We carried out 60 runs, each comprised of 9.6×10^7 iterations (reaching the 960 μs timespan) of the type indicated in Fig. 2 and described in Secs. II–IV for *mammalian ubiquitin* [32–37], an $N=76$ globular protein with no disulfide bridges ($U_{\mathrm{SS}}=0$ at all times). The specified solvent conditions parametrically relevant to our model (cf. Sec. III) are $T=308$ K and $p\mathrm{H}=4.5$. The most representative coarsely defined pathway, reproduced in 22 out of the 60 runs is displayed at the CM level in Figs. 5(a)–5(f). The reproduction was never perfect but satisfactory if we allow for a coarse-grained CM space, identifying CM's lying within a 1% Hamming distance from each other, and coarse-graining time to within 128-ps intervals. Thus, the favored pathway is reproducible to within a Hamming distance of 1% between CM's of the compared pathways and to within a 128 ps time interval for a given CM within a 1% Hamming distance from the CM of the selected pathway.

The six snapshots given in Figs. 5(a)–5(f), were obtained, respectively, at 180 ns, 1 μs, 1.9 μs, 10 μs, 200 μs, and 960 μs. The first two [Figs. 5(a) and 5(b)] correspond to a highly fluctuating stage of the process where portions of secondary structure never manage to be stabilized for more than 10–

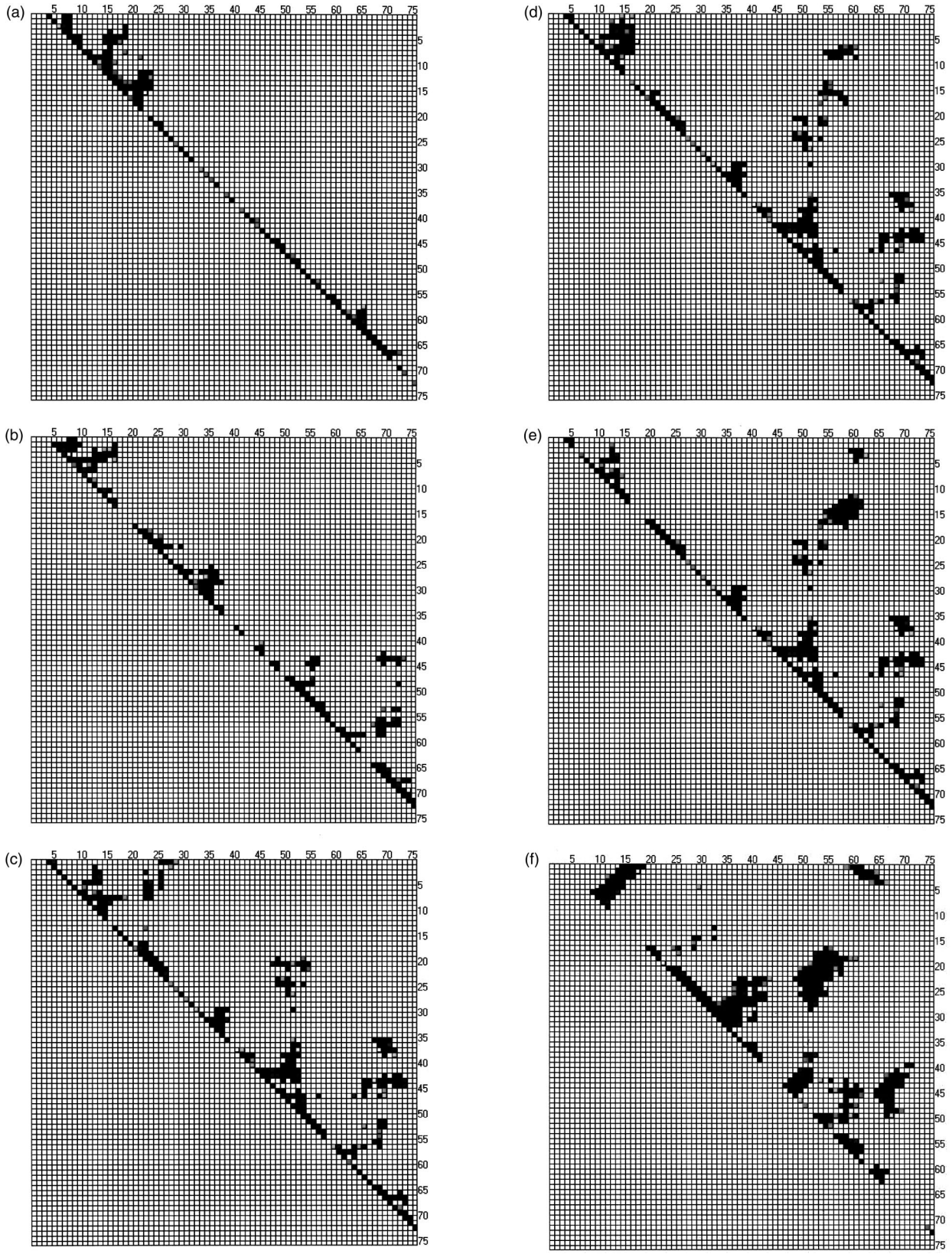
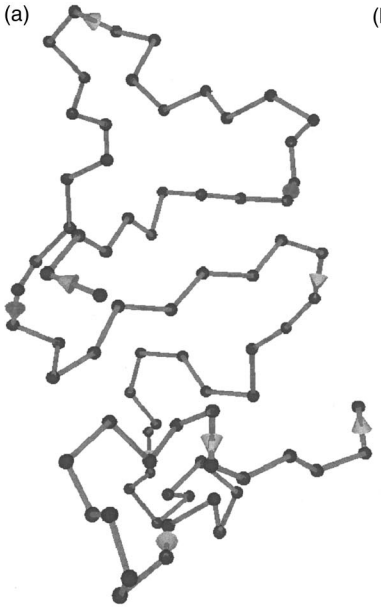


FIG. 5. (a)–(f) Six snapshots of the time evolution of the contact matrix for (*mammalian*) *ubiquitin* ($N=76$, pH$=4.5$, $T=308$ K) obtained, respectively, at 180 ns, 1 μs, 1.9 μs, 10 μs, 200 μs, and 960 μs. Dark square entries indicate distances of less than 7 Å while gray entries indicate distances in the range $7<r\leqslant 8.2$ Å.

(a)



(b)

Residue	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
R-basin	3	2	1	1	1	2	2	1	1	2	2	1	2	1	1	2	1	3	2
Φ	76.1	-75.8	-123	-119	-118	-82.8	-84	-77.6	-141	-84.3	-78.7	-143	-62.6	-125	-61	-77.2	-77.8	93.2	-66.6
Ψ	-1.81	1.18	111	113	110	4.93	-1.59	135	136	1.49	1.26	132	-32	117	122	-3.3	138	-6.69	-47

Residue	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
R-basin	1	2	2	1	1	1	3	1	1	1	2	2	1	2	1	1	1	1	1	1
Φ	-73.9	-56.9	-82.9	-80.6	-77.4	-141	77.9	-59.4	-135	-77.6	-81	-55.1	-75.9	-50.3	-138	-75.2	-138	-136	-136	-120
Ψ	133	-44.2	-3.97	132	134	140	-4.4	114	138	133	0.55	-27.6	134	-48.5	134	138	130	135	132	114

Residue	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
R-basin	3	1	1	1	2	1	1	4	1	1	2	1	2	2	2	3	3	3	3	2
Φ	89.5	-136	-75.3	-78.2	-55.8	-143	-141	63.5	-136	-76.4	-61.1	-118	-57	-55	-59	58	58.4	55.9	55.2	-76.7
Ψ	-1.84	134	133	136	-28	133	139	-120	140	138	-33.3	113	-24.7	-30.4	-53.5	42.4	46.9	58.5	26.4	3.94

Residue	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
R-basin	1	1	2	1	3	2	1	3	2	2	2	2	1	1	1	4	3
Φ	-77.3	-139	-55.7	-71.5	57	-57.6	-62.5	61.6	-82.3	-55.1	-86.6	-80.8	-138	-143	-119	76.3	57.7
Ψ	136	132	-46.1	133	47.9	-40.5	122	48	2.53	-46.8	-6.65	-1.48	138	133	118	-86.2	58

FIG. 6. The optimized backbone geometry (a) and its associated LTM (b) for the collapse-inducing kernel conformation of *ubiquitin* obtained at 1.9 μs.

100 ns. For instance, the native (20-33) α helix stabilized by ion pairs, does form in the 100 ns–1 μs range but, due to the absence of tertiary structure buttressing, it does not prevail for over a tenth of 1 μs. This fluctuating state of folding extends over to the submillisecond, and has been probed experimentally using proton exchange labeling [32–34] and circular dichroism techniques [35]. The former probes reveal no significant amide H-bond (N—H···O═C) formation (Fig. 4) up until the 5–10 μs range, suggesting that the peptide backbone is highly exposed to the solvent even during early folding stages traditionally attributed to secondary structure formation. Indeed, we find that no stable secondary structure forms until a collapse-inducing topology [Figs. 5(c) and 6] is formed. This nucleating state contains portions of secondary structure stabilized by tertiary scaffolding and favors the hydrophobic collapse since it involves significant tertiary buttressing of kernels of secondary structure. The occurrence of this nucleating event is marked by the pronounced decrease in structural fluctuations [38–42], as marked by the time-dependent cardinal of the set $I(t)$, denoted $\#I(t)$, which undergoes a drastic decrease in the 1–10 μs timescale (Fig. 7).

Furthermore, these findings seem to fit with earlier experiments [38,39] that reveal that structure formation is induced by an initial search for the ''right'' (collapse-competent) topology. We may add that secondary structure formation is not an all-or-none process, where such motifs might be found in isolation. Rather, secondary structure should be viewed in isolation as a fluctuating entity with a highly exposed backbone prone to proton exchange within experimental time resolution. Such fluctuating objects can only be stabilized once the buttressing provided by tertiary structure comes into place concomitantly with the formation of kernels for secondary structure development [Figs. 5(c) and 6]. In the light of our own findings, nucleation models [40–42], and relevant experimental work supporting the topological collapse scenario [38,39], we believe that the hierarchical picture of local propensities biasing the subsequent large-scale organization [6,13] might need revision, at least for small globular proteins such as *mammalian ubiquitin*: The local structures cannot be sustained by themselves to bias long-range organization.

The most striking feature of the $\#I(t)$ plot is the fact that a drastic quenching of structural fluctuations occurs at 1.9 μs. This quenching coincides with the formation of the nucleating topology that induces the hydrophobic collapse [Figs. 5(c) and 6]. These facts are in full agreement with experimental observations and related paradoxes [38,39] in

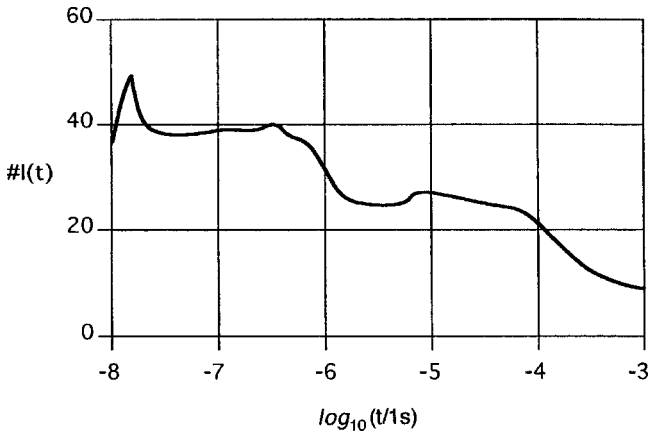


FIG. 7. The level of structural fluctuations at different stages of folding, as marked by the cardinal of $I(t)[\#I(t)]$ averaged every 1000 iterations (or resolved at the 10-ns level). This quantity gives the number of residues changing their R basin at iteration t, here plotted as a function of real time. The results correspond to the most favored and most reproducible pathway.

TABLE IV. Predicted stable LTM's for three natural proteins with PDB accesion codes $1kpt$, $1bqv$, and $1b0g$ obtained using the geometry-mediated π-ρ algorithm described in this work. At this level of topological resolution, the predicted structures are *identical* to the native structures.

File name: pdb1kpt.ent-Sequence: A-Model: 1-Number of units: 105

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

File name: pdb1bqv.ent-Sequence: No name-Model: 1-Number of units: 110

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

File name: pdb1b0g.ent-Sequence: B-Model:1-Number of units: 100

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

the sense that the backbone exposure to the solvent in the microsecond timescale is as high as in a random coil conformation, in sharp contrast with what would be expected for typical time scales of secondary structure formation. Thus, the results given in Fig. 7 corroborate the ''topological nucleus scenario'' for *ubiquitin* and lead us to infer that folding might not proceed hierarchically from local to large-scale organization, nor does it operate through a diffusion-collision expedient in which secondary structure is stabilized by large-scale events [43]. Rather, it proceeds by loosely searching for the right topology in the form of a collapse-inducing nucleus whose formation produces the quenching of structural fluctuations.

Figures 5(d) and 5(e) contain most of the native secondary and tertiary structure, especially the β-sheet motif contained in the $(2\text{-}8)\times(10\text{-}16)$ region, the (20-33) α helix, the β-sheets containing the (35-38), (40-45), (47-51) strands, as well as the native tertiary interactions in the region $(14\text{-}27)\times(52\text{-}65)$ and the tertiary native contacts in the region $(43\text{-}48)\times(63\text{-}71)$. The level of structural fluctuation associated with this stage of organization of the chain (10–200 μs) is about half of the level of fluctuation before the collapse-inducing topology had formed (Fig. 7). This fact is marked by a quasiplateau in the submillisecond range indicating no major structural rearrangement. The final stage of the folding process is the consolidation and refinement of secondary structure with concomitant formation of the most entropically expensive tertiary interaction: The parallel β sheet engaging the (60-67) β strand and the initial extremity (2-6) of the peptide chain [Fig. 5(f)].

The success of our method may be assessed by observing that the CM predicted to correspond to the stable fold [Fig. 5 (f)] is *identical* to within a Hamming distance of 0.58% to that of the native fold obtained from Protein Data Bank (PDB) file for the same primary sequence (accession code $1ubi$). Other predictions for moderately large proteins ($N \sim 100$) under the same conditions reveal similar levels of success at the level of contact pattern resolution and even higher at the topological level of resolution: The stable LTM's for the species with PDB accession codes $1kpt$, $1bqqv$, and $1b0g$ are *identical* to those of their native folds (Table IV).

VI. CONCLUSION

Any approach to the protein folding problem must reconcile the vast spectrum of structural detail of the peptide chain with the known expediency of the folding process. This crucial property suggests an initial easy-to-rectify loose search for a ''correct topology'' in the form of a roughly defined nucleus with few intramolecular hydrogen bonds and a highly exposed backbone, which is nevertheless competent in inducing the ultimate hydrophobic collapse. This scenario is discernible in recent kinetic experiments [38,39] as well as in theoretical models [40–43]. This work dealt rigorously with this picture by underlying its physical foundations and turning it into an ansatz upon which an algorithm has been implemented to infer *ab initio* conductive folding pathways, reproduce the folding kinetics, and predict native folds of proteins.

Ours is a self-consistent model that does not truncate conformational detail or discard potential energy contributions. Rather, it regards them hierarchically (which by no means imply that we are assuming that folding itself is hierarchical in going from local to large-scale organization). In essence, our model subsumes fast-evolving variables into conformational constraints that in turn serve as the framework for large-scale organization. Thus, our approach takes into account the fingerprint of folding dynamics: The enslavement or subordination of torsional motion to a coarsely defined flow in conformation space in which local torsional states are viewed modulo the basins of attraction (R basins) to which they belong in the Ramachandran maps. In simple words, we make use of the fact that local torsional exploration is heavily constrained by the R basin the residue is in at a given time. Thus, the torsional dynamics must in fact rely on the slower process of interbasin hopping.

This coarse ''topological'' dynamics that underlies the actual conformational search of the protein is generated by a pattern-recognition-and-feedback iterative sequence in which the roughly defined state of the chain is periodically evaluated to detect patterns that are topologically compatible with structural forms. Once any such pattern is identified, its own time evolution is slowed down with respect to free residues of the chain. For complex proteins such as the one adopted as an illustration in this work, the pattern identification becomes unambiguous only if its recognition is mediated through geometric realizations of the topology. These realizations have been obtained by optimizing a semiempirical potential. In turn, this potential subsumes conformational detail of side chains, which is not explicitly present in the torsional dynamics we intend to reproduce.

The power of our method as well as its physical soundness is evidenced by its predictive potential to generate native folds, expedient pathways and reproduce experimentally probed kinetic features.

The research reported in this work suggests that for many proteins a crucial step in triggering hydrophobic collapse consists in finding the right topology that may potentially scaffold the otherwise flickering secondary structure dictated by local or middle-range propensities. This experimentally probed scenario is far more generic than originally thought [36,38,39]. Thus, a concurrent issue that will be addressed in futhure work is the identification of the residues that participate in the formation of the collapse-inducing nucleus. These hot spots may be probed by site-directed mutagenesis, as done with CI2 (cf. [44]). Our strategy to detect the hot spots may be sketched as follows.

(a) First find out whether there is a drastic quenching in structural fluctuations along the folding process, as in the system described in this work. This event will be marked by a sudden decrease in the number of residues changing R basin ($\#I(t)$). If this decrease takes place, record the time $t=t^*$ at which it occurs.

(b) For residue n, determine the quantity $t(n)=$time it takes for residue n to cease performing interbasin hopping.

(c) Identify those residues such that $t(n)\leqslant t^*$. These are

the residues topologically ordered at the time when the nucleus is formed. Thus, they should be identified as hot spots *vis-à-vis* site-directed mutagenesis.

Forthcoming work will make use of these ideas to actually identify the core residues determinant of hydrophobic collapse in systems such as CI2, where mutagenesis data is available [44], as well as for the *ubiquitin* system presented in this work. Furthermore, the engineering relevance of such studies will be assessed.

ACKNOWLEDGMENTS

This research was partially supported by the U.S. Information Agency, the National Research Council of Argentina, and the Alexander von Humboldt Foundation. One of us (A.F.) wishes to express his thanks to Professor R. Huber, Professor R. S. Berry, Professor T. R. Sosnick, and Professor J. K. Percus for enlightening discussions and kind hospitality.

[1] C. B. Anfinsen, Science **181**, 223 (1973).
[2] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[3] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[4] R. Zwanzig, A. Szabo, and B. Bagchi, Proc. Natl. Acad. Sci. U.S.A. **89**, 20 (1992).
[5] M. Karplus, Folding Des. **2**, S69 (1997).
[6] R. Baldwin and G. Rose, Trends Biochem. Sci. **24**, 26 (1999).
[7] A. Fernández, K. Kostov, and R. S. Berry, Proc. Natl. Acad. Sci. U.S.A. **96**, 12991 (1999).
[8] A. Fernández and R. S. Berry, J. Chem. Phys. **112**, 5212 (2000).
[9] A. Fernández, K. Kostov, and R. S. Berry, J. Chem. Phys. **112**, 5223 (2000).
[10] (a) A. Fernández and A. Colubri, J. Math. Phys. **41**, 2593 (2000); (b) A. Fernández and A. Colubri, *ibid.* **39**, 3167 (1998).
[11] C. Cantor and P. Schimmel, *Biophysical Chemistry* (W. H. Freeman, New York, 1980).
[12] J. Thornton, in *Protein Folding*, edited by T. E. Creighton (W. H. Freeman, New York, 1992), pp. 59–63.
[13] R. Baldwin and G. Rose, Trends Biochem. Sci. **24**, 77 (1999).
[14] A. Fernández and A. Colubri, Phys. Rev. E **60**, 4645 (1999).
[15] P. A. Laskowski, M. W. MacArthur, D. S. Moss, J. M. Thornton, J. Appl. Crystallogr. **26**, 283 (1993).
[16] D. Luenberger, *Linear and Nonlinear Programming* (Addison-Wesley, New York, 1989).
[17] A. Fernández, Phys. Chem. Chem. Phys. **2**, 1375 (2000).
[18] T. Schlick, in *Mathematical Approaches to Biomolecular Structure and Dynamics*, IMA volumes in Mathematics and Applications, edited by J. Mesirov, K. Schulten, and D. Summers (Springer, New York, 1996), pp. 219–247.
[19] (a) M. H. Hao and H. A. Scheraga, Acc. Chem. Res. **31**, 433 (1998); (b) D. Thirumalai and D. Klimov, Curr. Opin. Struct. Biol. **9**, 197 (1999).
[20] S. Lifson, in *Methods in Structural Molecular Biology*, edited by D. B. Davies, W. Saenger, and S. Danyluk (Plenum, London 1981), pp. 359–385.
[21] I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, J. Biomol. Struct. Dyn. **7**, 391 (1989); **7**, 421 (1989).
[22] P. A. Kollman and K. A. Dill, J. Biomol. Struct. Dyn. **8**, 1103 (1991).
[23] O. Sinanoglu and A. Fernández, Biophys. Chem. **21**, 157 (1985).
[24] P. L. Privalov, in *Protein Structure and Protein Engineering*, edited by E.-L. Winnacker and R. Huber (Springer, Berlin, 1988), pp. 6–15.
[25] (a) F. M. Richards, Annu. Rev. Biophys. Bioeng. **6**, 151 (1977); (b) G. D. Rose and J. E. Dworkin, in *Prediction of Protein Structure and Principles of Protein Conformation*, edited by G. D. Fasman (Plenum, New York, 1989), pp. 625–634.
[26] (a) S. Marqusee and R. L. Baldwin, Proc. Natl. Acad. Sci. U.S.A. **84**, 8898 (1987); (b) G. I. Makhatadze, M. M. Lopez, J. M. Richardson, and S. T. Thomas, Protein Sci. **7**, 689 (1998).
[27] (a) D. A. Brant, Macromolecules **1**, 291 (1968); (b) K. Shoemaker, P. Kim, E. J. York, J. M. Stewart, and R. L. Baldwin, Nature (London) **326**, 563 (1987).
[28] (a) L. Pauling and R. B. Corey, Proc. Natl. Acad. Sci. U.S.A. **37**, 251 (1951); (b) **37**, 729 (1951); c) **39**, 253 (1953).
[29] G. I. Makhatadze, G. M. Clore, A. M. Gronenborn, and P. L. Privalov, Biochemistry **33**, 9327 (1994).
[30] (a) A. Fernández, Ann. Phys. (Leipzig) **4**, 600 (1995); (b) A. Fernández, J. Stat. Phys. **92**, 237 (1998).
[31] D. Eisenberg and A. D. McLachlan, Nature (London) **319**, 199 (1986).
[32] M. S. Briggs and H. Roder, Proc. Natl. Acad. Sci. U.S.A. **89**, 2017 (1992).
[33] S. Khorasanizadeh, I. D. Peters, T. R. Butt, and H. Roder, Biochemistry **32**, 7054 (1993).
[34] S. Khorasanizadch, I. D. Peters, and H. Roder, Nat. Struct. Biol. **3**, 193 (1996).
[35] S. T. Gladwin and P. A. Evans, Folding Des. **1**, 407 (1996).
[36] B. A. Kranz, L. B. Moran, A. Kentsis, and T. R. Sosnick, Nat. Struct. Biol. **7**, 62 (2000).
[37] J. Sabelko, J. Ervin, and M. Gruebele, Proc. Natl. Acad. Sci. U.S.A. **96**, 6031 (1999).
[38] T. R. Sosnick, L. Mayne, R. Hiller and S. W. Englander, in *Peptide and Protein Folding Workshop*, edited by W. F. DeGrado (International Business Communications, Philadelphia, 1995), pp. 52–80.
[39] T. R. Sosnick, L. Mayne, and S. W. Englander, Proteins **24**, 413 (1996).
[40] D. Thirumalai and Z. Guo, Biopolymers **35**, 137 (1995).
[41] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10026 (1994).
[42] A. R. Fresht, Proc. Natl. Acad. Sci. U.S.A. **92**, 10869 (1995).
[43] C. L. Brooks III, M. Karplus, and B. Montgomery Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*, Advances in Chemical Physics, Vol. LXXI (Wiley, New York, 1988).
[44] V. Muñoz and W. Eaton, Proc. Natl. Acad. Sci. U.S.A. **96**, 11311 (1999).