Topology-dependent protein folding rates analyzed by a stereochemical model

Inês R. Silva[a)]
Departamento de Física e Matemática Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Av. Bandeirantes 3900, 14040-901 Ribeirão Preto, SP, Brazil

Lucila M. Dos Reis[b)] and A. Caliri[c)]
Departmento de Física e Química Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Av. do café S/N, 14040-903 Ribeirão Preto, SP, Brazil

(Received 16 May 2005; accepted 11 August 2005; published online 20 October 2005)

It is an experimental fact that gross topological parameters of the native structure of small proteins presenting two-state kinetics, as relative contact order χ, correlate with the logarithm of their respective folding rate constant κ_f. However, reported results show specific cases for which the $(\chi, \log \kappa_f)$ dependence does not follow the overall trend of the entire collection of experimental data. Therefore, an interesting point to be clarified is to what extent the native topology alone can explain these exceptional data. In this work, the structural determinants of the folding kinetics are investigated by means of a 27-mer lattice model, in that each native is represented by a compact self-avoiding (CSA) configuration. The hydrophobic effect and steric constraints are taken as basic ingredients of the folding mechanism, and each CSA configuration is characterized according to its composition of specific patterns (resembling basic structural elements such as loops, sheets, and helices). Our results suggest that (i) folding rate constants are largely influenced by topological details of the native structure, as configurational pattern types and their combinations, and (ii) global parameters, as the relative contact order, may not be effective to detect them. Distinct pattern types and their combinations are determinants of what we call here the "content of secondary-type" structure (σ) of the native: high σ implies a large κ_f. The largest part of all CSA configurations presents a mix of distinct structural patterns, which determine the $\chi \times \log \kappa_f$ linear dependence: Those structures not presenting a proper χ-dependent balance of patterns have their folding kinetics affected with respect to the pretense linear correlation between χ and $\log \kappa_f$. The basic physical mechanism relating σ and κ_f involves the concept of cooperativity: If the native is composed of patterns producing a spatial order rich in effective short-range contacts, a properly designed sequence undertakes a fast folding process. On the other hand, the presence of some structural patterns, such as long loops, may reduce substantially the folding performance. This fact is illustrated through natives having a very similar topology but presenting a distinct folding rate κ_f, and by analyzing structures having the same χ but different σ. © *2005 American Institute of Physics*. [DOI: 10.1063/1.2052607]

I. INTRODUCTION

The architecture of real proteins is very complex. However, simplified lattice and off-lattice models have been successful in capturing key aspects of the structural determinants of folding kinetics. Mimicking real protein systems, minimalist models are able to find the unique native state among an astronomical number of possible chain configurations and, through efficient simulations, many insights have been gained into general properties of residue sequences and preferred structures that are required to provide fast folding and stability to model proteins. Such models have also been used to demonstrate that structural regularities, resembling real protein secondary structures, are related to high "designability" and evolutionary stability[1] and, more recently, they have been used to investigate the relationship between gross topological parameters of the native structures and the logarithm of folding rate constants of small two-state folding proteins[2,3] as observed experimentally.[4–6]

With the progressive increase of the number of studied proteins, it has been increasingly reported that small two-state proteins present an in-water folding rate κ_f that range several orders of magnitude as their corresponding relative contact order, represented here by χ, varies from about 5 to ~30%.[4–9] Conceptually, χ is a measure of the average distance between contacting residues, represented here as the average number of residues between pairs i,j of contacting chain units, normalized by the sequence length L, that is,

$$\chi = \frac{1}{LN_c} \sum_{\{N_c\}} \Delta l_{i,j}, \tag{1}$$

where $\Delta l_{i,j}$ is the number of residues between contacting

[a)]Electronic mail: inesregina@hotmail.com
[b)]Electronic mail: imreis@glu.fcfrp.usp.br
[c)]Author to whom all correspondence should be addressed. Electronic mail: ancaliri@fcfrp.usp.br

residues i and j; the sum runs over all N_c contacting pairs $\{i,j\}$ (if two heavy atoms, one in residue i and the other in residue j, are separated by a distance $d_{i,j}$ smaller than a certain cutoff d_0, then they are considered contacting residues; distinct values for d_0 between 3.5 and 7.0 Å have been used). Generalizations of the relative contact order concept have also been experienced[8,10,11] confirming that the folding rate of proteins significantly depends on structural (topological) parameters. However, as has been shown experimentally, mutations may also affect the folding rate up to three orders of magnitude,[6] indicating that topology is not the sole determinant of the folding rate. Additionally, distinct proteins of very close χ values (and with eventual similar structure topologies) may fold with very different folding rates, while proteins of significantly different χ values (presenting distinct structural topologies) may fold with practically the same folding rates.[12–15] Therefore, there is a need to improve the understanding of how and to what extent the topology of the native structure alone can explain such experimental data.

The objective of this work is to study and identify structural determinants of the folding kinetics; we proceed by reproducing ideally the correlation between $\log \kappa_f$ and χ using a lattice model. This is not a straightforward task, since it requires a protein design rule, that is, for any given native configuration, the instruction for writing down a sequence of residues that drives the chain to the selected native structure must be known. This is the folding inverse problem. This requirement is promptly fulfilled by models that use the minimum alphabet size (two letters), as the HP or Gō model,[16–18] but such models, and even more sophisticated lattice models using larger alphabets and simple pairwise interaction schemes,[19–21] do not present a consistent correlation between the logarithm of folding rate κ_f and contact order χ. Therefore, in order to improve this correlation, a certain degree of cooperativity in the interaction scheme has been successfully proposed.[3,12]

In this work we employ a ten-letter non-native-centric lattice model, which provides a specific "syntax" for writing down the sequence for each target (native) structure. The alphabet used here[22] has exactly the size of the smallest alphabet that can reproduce fundamental aspects of the folding process, as recently suggested theoretically.[23] Native structures are represented by compact self-avoiding (CSA) configurations ($3\times3\times3$ cubic sublattice) for chains of $L=27$ residues. All structures are then characterized by basic topological attributes. Special attention is paid to those cases that deviate from the overall $\log \kappa_f$ versus χ linear tendency determined by the entire collection of structures employed.[3,20] The folding rates were obtained through the Monte Carlo (MC) simulation method. A main set of 52 representative native structures were chosen in such a way to cover the entire range of χ possible values and provide a rich variety of fold patterns; however, many other CSA configurations were selected and simulated in order to validate our propositions and conclusions.

Our results, within the scope of the present model, show explicitly how specific structural patterns and their arrangement in the native structure interfere on the folding kinetics, accelerating or slowing down the folding reaction.

II. MODEL AND METHOD

A. Stereochemical model for the chain-solvent system

The chain-solvent system is represented here by a lattice model, with the units (residues) of a 27-mer chain restricted to occupy exclusively 27 consecutive sites of an infinite cubic lattice; the other sites are all occupied by solvent molecules.[22] During the simulation, solvent molecules and chain units exchange their respective sites in such a way that all sites of the lattice always remain fully occupied. For each configuration change, say from configuration x to configuration y, only hydrophobic energy variations are considered, which is achieved through chain-solvent interactions; all other interactions, namely solvent-solvent and residue-residue interactions, are reduced to hard core-type interactions (excluded volume). For a regular lattice and uniform density, as in the present case, this interaction scheme is exactly equivalent to an additive first neighbor inter-residue pairwise potential, $h_{i,j}=h_i+h_j$, where h_i is the hydrophobic level of residue i in the chain sequence. Real amino acids are then represented here by a repertory of ten distinct units (ten-letter alphabet) characterized by distinct hydrophobic levels and a set of inter-residue specificities $\{c_{i,j}\}$. The set of distinct hydrophobicity levels represents the most general and influential chemical factor acting during the folding process, and the set of inter-residue constraints mimics steric specificities of the real amino acids by establishing which pairs of residues are forbidden to interact as first neighbors; they work just as extra pairwise specificities (hard-core-like potential) and its main consequence is to select folding and unfolding pathways through the configurational space.[24]

Therefore, the configurational energy $E([\kappa,l])$ of an arbitrary configuration ξ defined by a set $[\kappa,l]$ of N_ξ first neighbor inter-residue contacts is

$$E([\kappa,l]) = \sum_{\{i,j\}} (h_{i,j} + c_{i,j})\,\delta_{(i,j),[\kappa,l]}, \tag{2}$$

where the sum runs over all $L(L-1)/2$ residues pairs $\{i,j\}$; the factor $\delta_{(i,j),[\kappa,l]}=1$ if the pair (i,j) belongs to the set $[\kappa,l]$, and $\delta_{(i,j),[\kappa,l]}=0$ otherwise (first neighbor interaction); $N_\xi=0, 1, 2, 3, \ldots, 28$.

A distinguished aspect of the present model is that it is not native-centric: (i) The rules for sequence designing are the same for any target structure (CSA configuration); the rule is mainly based on the "hydrophobic inside" rule[25] and on local residue topological environment in the native structure; (ii) the intermonomer constraints are fixed for each monomer pair, that is, they do not depend on the particularities of the native structure; and (iii) the native data are never employed in the course of the MC simulation to decide if a particular configuration change is accepted or not.

Lattice models based on such a hydrophobic-type potential are efficient in packing the chain and in finding the native state, but they fail to confer stability to the native state because the additive inter-residue pairwise potential, $h_{i,j}=h_i$

$+h_j$, satisfies marginally the segregation principle $2h_{i,j}-h_{i,i}-h_{j,j}\geqslant 0$ through the equal sign, that is, $2h_{i,j}-h_{i,i}-h_{j,j}=0$. However, the addition of the specific set $\{c_{i,j}\}$ of steric constraints in the interaction potential have several significant consequences; for example, it improves the overall stability condition of the globule in the native state and helps to select folding pathways. It also dramatically changes the system's configurational activity, producing a peaked and symmetric curve for the heat capacity as a function of the temperature (without steric constraints, the curve is broad and asymmetric). For more details on the hydrophobic levels $\{h_i\}$, constraints $\{c_{i,j}\}$, chain design, and on kinetic and thermodynamic properties, see Refs. 22 and 24.

The folding process is simulated through the Metropolis MC method, involving the standard elementary chain moves: crankshaft, corner, and end flips; a particular reference point along the chain is chosen at random before each move attempt and the simulation evolves without any reference to the native configuration, except to check when it is found for the first time.

B. Folding rate

All computational experiments were performed under the same "physiological" temperature $\kappa_B T_P=1.0$ in order to imitate a biological system in which the folding process for all proteins is achieved at fixed temperature.[22] Simulations were conducted on a preselected (not random) set of 52 representative structures of native conformations, enabling an extensive sample of the entire range of possible χ values for this model and, at the same time, illustrating distinct geometrical pattern components and their kinetic implications, as discussed below. For each native structure, a set of 15 runs $\{t_i\}$ was performed during a time window t_w corresponding to about 3×10^7 MC steps. For each run, the time t_i expended for finding the native structure for the first time (folding first passage time)[12] is recorded; t_i is given by the number of attempted MC moves.

The kinetic of folding of small single-domain proteins, approximated by a simple two-state kinetic, is characterized by an exponential distribution of folding times,

$$P(t)\sim \exp(-t/t_0), \tag{3}$$

with $\int P(t)dt=1$. Therefore, for large enough samples (several hundreds of independent runs), the average folding time,

$$\langle t\rangle = \int_0^\infty tP(t)dt = t_0, \tag{4}$$

is usually approximated by the mean first passage time (MFPT), that is, $t_0\cong$MFPT. The folding rate is then represented by 1/MFPT, and so $\log(\kappa_f)=-\log(\text{MFPT})$.[3]

However, in the present case, due to the relatively small set of runs employed (actually, a total of 15 independent runs were performed for each one of the 52 cases), it was used as an alternative approach to obtain a valid representation for the folding rate. We take advantage of the fact that the $\langle \ln(t)\rangle$ is linearly related to the $\ln(\langle t\rangle)$, that is,

$$\langle \ln(t)\rangle = \int_0^\infty \ln(t)P(t)dt = \ln(\langle t\rangle) - \gamma, \tag{5}$$

which can be directly verified using the result $\int_0^\infty \ln(t)\times\exp(-t)dt=-\gamma$, where $\gamma\cong 0.5772$ is Euler's constant. Therefore, $\log(\langle t\rangle)=\langle\log(t)\rangle+\gamma\log(e)$, where $\log(e)\cong 0.4343$. This property is particularly useful for small samples; the effect of data dispersion, always present in small samples, is minimized by averaging $\ln(t_i)$. Therefore, $\log\kappa_f$ is represented here by

$$\log\kappa_f = (1/15)\sum_{i=1}^{15}\log t_i^{-1} - 0.2507. \tag{6}$$

In order to validate Eq. (6), hundreds of experiments were performed for a few selected structures, and the set of 52 structures (Table I) was treated through another approach: The histogram of the time-dependent cumulative number $N(t)$ of folded proteins was fitted by

$$N(t) = N_0[1-\exp(-t/t_0)], \tag{7}$$

with $N_0=15$. The characteristic time t_0, estimated through Eq. (7) for each native structure presenting an absolute (100%) folding success (which was most of them), was then used to determine the discrepancy between $\log(t_0^{-1})$ and the corresponding $\log\kappa_f$ obtained through Eq. (6). The agreement was good enough: in most of the cases the discrepancy was smaller than 3% (in just one case it was about 11%), justifying the "economical" size of the set $\{t_i\}$. For the cases presenting a folding success below 80%, with respect to the time window t_w (ten structures out of the 52 and several other extra structures examined), the folding time distribution could not be approximated by a single exponential, because they present two significantly different characteristic times with a competitive number of "proteins" undergoing each corresponding folding route (at the established working temperature).[22]

C. Topological characteristics of the CSA configurations

Proteins in general have a large fraction of their residues conformed in secondary structures, mainly α helices and β strands. These regular repeating structural patterns are formed by local contacting units (α helices) and progressive sequencing of contacting units along the chain (β strands). Therefore, it is not surprising that protein domains tend to present small χ values. Although the structural possibilities for a real protein are far much richer than the set of all compact structures generated by a chain constituted by 27 units restricted on the sites of a cubic lattice, which in principle is a limitation of the model, these CSA configurations reflect many geometrical patterns of real chains. We start by showing in Fig. 1 how the CSA structures (unrelated by reflection, rotation, and reverse-labeling symmetries) are distributed with respect to relative contact order χ, which satisfies $0.238\,095\leqslant\chi\leqslant 0.494\,709$. There are only 97 distinct χ values, corresponding to distinct multiples of $2/(27\times 28)$. The distribution of χ values presented in Fig. 1 is slightly different from that shown in Ref. 3; the present distribution

TABLE I. Topological characteristics of the 52 native structures used in this work. The following is shown for each native configuration: the relative contact order χ and χ^c; the number of the topological element S; the number of patterns [*STTS*] and [*TT*]; the extension of the largest sequence of Ts({T…T}max); the number of patterns staple and snail; and the corresponding logarithm of its determined folding rate ($\log \kappa_f$). No folding rate was calculated for cases with a folding success ratio of below 80% in the time window t_w. The symbols identify distinct groups, as explained in Fig. 5. In general, as χ increases, $\log \kappa_f$ decreases, but the presence in the structure of patterns as [*STTS*] (as in native configuration Nos. 1, 21, or 32) or long extension of T's (structure No. 17) promotes very fast folding independently of χ, as well as the presence in the structure of patterns like snail and staple (as in structure Nos. 7 and 33) slows down the folding process independently of χ.

	χ	χ^c	*S*	[*STTS*]	[*TT*]	[*TTTT*]	$\{T\ldots T\}_{\max}$	*Staple*	*Snail*	$\log \kappa_f$	Group
1	0.23810	0.0741	9	8	0	0	2	0	0	−5.369	■
2	0.23810	0.1111	9	7	0	0	2	0	0	−5.988	●
3	0.23810	0.1481	9	6	0	0	2	0	0	−5.695	●
4	0.24339	0.1534	7	4	3	2	4	0	0	−6.585	▽
5	0.24339	0.1534	7	3	3	2	6	0	0	−6.835	▽
6	0.24339	0.1376	7	2	3	2	6	0	0	−6.780	▽
7	0.24603	0.1534	9	6	0	0	2	0	1	−6.940	▽ ○
8	0.24603	0.1931	6	0	6	3	10	0	0	−6.218	▽
9	0.25926	0.1587	7	4	2	1	4	0	0	−6.143	●
10	0.25926	0.1667	8	3	1	0	6	0	0	−6.739	▽
11	0.25926	0.1667	7	3	3	1	7	0	0	−6.539	▽
12	0.26190	0.0873	9	8	0	0	2	0	0	−5.450	■
13	0.26190	0.1534	9	7	0	0	2	0	0	−5.741	●
14	0.26190	0.1958	3	0	11	4	16	0	0	−5.649	●
15	0.26190	0.1852	4	0	8	3	10	0	0	−6.442	▽
16	0.27249	0.1905	8	6	0	0	3	0	0	−5.855	●
17	0.27249	0.1984	3	0	10	2	15	0	0	−5.494	■
18	0.30423	0.1561	7	4	3	1	2	0	0	−5.999	●
19	0.30423	0.2354	4	0	7	1	10	0	0		▼
20	0.30952	0.1296	9	8	0	0	2	0	0	−5.408	■
21	0.30952	0.2063	9	7	0	0	2	0	0	−5.479	■
22	0.30952	0.2063	3	0	7	2	11	0	0	−6.295	●
23	0.31481	0.246	7	3	4	2	6	0	0	−6.603	●
24	0.31746	0.1614	8	6	0	0	4	0	0	−5.804	△
25	0.31746	0.2487	6	0	4	1	6	1	0	−7.053	● ○
26	0.32804	0.2011	6	4	4	1	6	0	0	−5.642	△
27	0.33069	0.164	8	6	0	0	3	0	0	−5.954	△
28	0.33069	0.3095	6	1	4	0	7	0	0	−6.960	●
29	0.33069	0.2434	6	0	3	0	12	0	0		▼
30	0.33069	0.246	5	0	5	1	10	0	1		★ ○
31	0.33333	0.2698	5	0	5	1	7	0	0		▼
32	0.33333	0.1614	7	6	2	1	4	0	0	−5.465	■
33	0.33333	0.2804	7	2	2	0	6	1	0		▼ ○
34	0.34921	0.2646	6	2	3	0	9	0	0	−6.998	●
35	0.36772	0.2619	8	4	1	0	5	0	0	−6.702	●
36	0.36772	0.3175	5	0	4	1	6	0	0	−7.179	●
37	0.40476	0.3571	6	4	2	0	6	0	0		▼
38	0.40476	0.3598	4	0	9	3	12	0	0	−7.137	●
39	0.41005	0.3889	7	3	2	0	5	0	0	−6.888	●
40	0.41005	0.3333	5	0	4	1	8	0	0	−7.636	●
41	0.41270	0.3862	6	2	2	0	6	0	0		▼
42	0.41534	0.2354	7	4	1	0	5	0	0	−6.020	△
43	0.42857	0.4021	7	3	0	0	3	0	0	−7.393	●
44	0.44180	0.4206	7	2	1	0	4	0	0	−7.394	●
45	0.44180	0.3995	5	0	5	0	5	0	0		▼
46	0.44444	0.4365	6	3	1	0	7	0	0	−7.607	●
47	0.45238	0.3942	7	3	1	0	4	0	0	−7.479	●
48	0.45238	0.4021	6	0	4	1	6	0	0		▼
49	0.46561	0.455	5	2	4	1	6	0	0	−7.566	●
50	0.46561	0.3995	4	0	7	1	9	0	0	−7.161	●
51	0.47884	0.3333	5	0	5	1	10	0	0	−7.444	●
52	0.48413	0.3942	8	1	0	0	6	1	0		★ ○

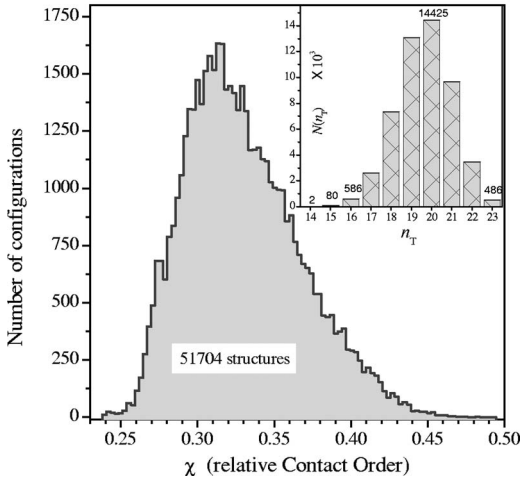

FIG. 1. Distribution of relative contact order χ of the set of all 51 704 independent CSA configurations. There are 97 distinct χ values restricted to the interval $0.238\,095 \leqslant \chi \leqslant 0.494\,709$; most parts of the structures present a contact order around $\chi=0.31$: about 50% of the configurations have a contact order in the interval $0.30<\chi<0.35$, and more than 76% of them present intermediate χ, defined roughly as $0.30<\chi<0.42$. The inset shows how configurations are distributed in function of the number n_T of turns (T).

was recalculated excluding those structures related by reverse labeling symmetries—that is, those structures in which exchanging each index of residues i along the chain by $L+1-i$ leaves the configuration unchanged—giving a total of 51 704 CSA configurations.

The bell-shaped distribution shown by Fig. 1 indicates that a very small fraction of the CSA configurations presents extremes of χ, with most of them being concentrated at about $\chi=0.31$. This property is in agreement with the fact that most of the structural domain in proteins, which are composed by α helices and β sheets, present small χ values, indicating that such domains are restricted to a very small fraction of all compact configurational possibilities for a linear chain. Therefore, it becomes easy to recognize that a random selection of CSA configurations is not necessarily a representative sample of the conformation universe of real proteins.

Since contact order is a global parameter, many configurations with very distinct local topological elements may present the same χ value, and vice versa. Therefore, it is helpful to examine topological details of the structural components that make up the final χ amount. As a first step in this procedure, we identify the basic topological elements of all target structures, as well their distinct configurational patterns,[26] in order to increment the structural characterization of each CSA configuration.

1. Basic topological elements: E, T, and S

Each particular chain unit (residue) of a CSA configuration is associated with one of the three following topological elements: It is an E unit if it is one of the chain ends, a T unit if the chain makes a turn through it, and an S unit if it is connected to other two neighbors along the chain through a straight line. Each CSA configuration is then constituted by two E units (the ends) and a certain number n_T and n_S of T and S units, respectively, so that for a 27-mer chain $n_T+n_S=25$. The inset of Fig. 1 shows the distribution of the number

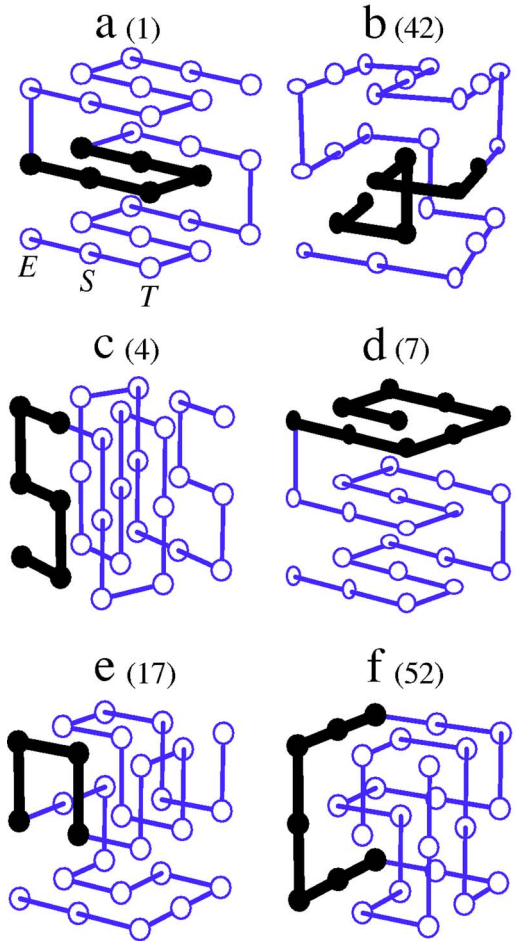

FIG. 2. Topological characteristics of CSA configurations. The six native configurations are labeled from a to f; the numbers following the letters refer to native configurations in Table I. Basic topological elements are illustrated in structure a by its correspondent symbols, that is, E for a chain end, T for a turn, and S for a straight line. The structural patterns [*STTS*], [*TT*], and [*TTTT*] are emphasized in the native configurations a, c, and e, respectively; sequence $\{T\ldots T\}$ in b; and snail and staple are illustrated in the native configurations d and f, respectively.

$N(n_T)$ of configurations with n_T units of type T; note that the number $N(n_S)$ of configurations with n_S units of type S is equal to the number $N(n_T)$ of configurations with n_T units when $n_T=25-n_S$, because each structure with n_S units must have $n_T=25-n_S$ units. As one should expect, for compact configurations, most of them present n_T at about four times the number n_S. The upper limit for S type units for all CSA configurations is $n_S^{\max}=11$, and $n_S^{\min}=2$; a small amount of configurations presents these extreme values. The three types of basic topological elements are illustrated by the structure shown in Fig. 2(a) by their respective symbols (E, T, and S).

2. Structural patterns

Some sequential combination of S and T units are notable because they resemble real protein secondary structures: Three examples are illustrated in Figs. 2(a), 2(c), and 2(e), respectively: sequence of units [*STTS*], sequence [*TTTT*], and [*TT*]. The asterisks (*) mean free assignment for the two other units in the sequence, that is, combinations of E and T for [*STTS*] and [*TT*] patterns, and combinations of E, S, and T for the [*TTTT*] pattern. The

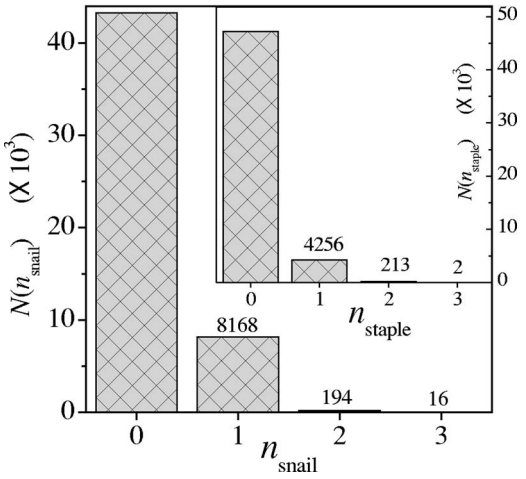

FIG. 3. Distribution of the number $N(n_{\text{snail}})$ of CSA configurations having $n_{\text{snail}}=0$, 1, 2, or 3 snail patterns in the same structure; the inset shows the same for configurations having staple patterns.

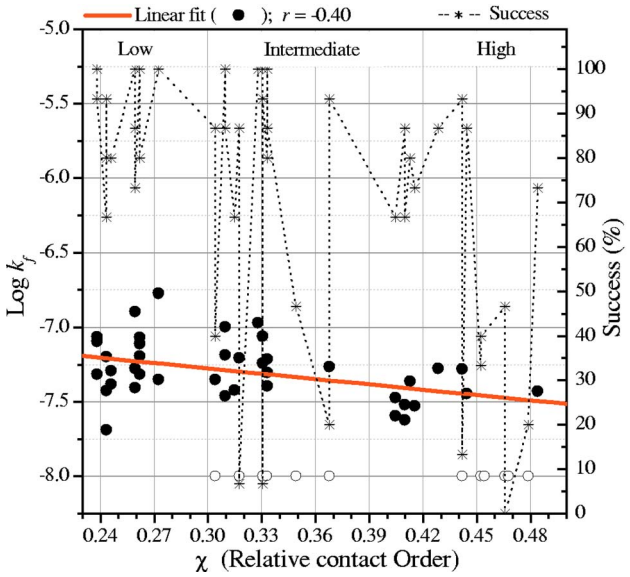

FIG. 4. Behavior of logarithm of the folding rate κ_f (solid line) and folding success (asterisks connected by a dotted line) as a function of relative contact order χ; hydrophobic potential $\{e_{i,j}=h_j+h_j\}$ only. The weak correlation between $\log \kappa_f$ and χ (solid thick line), correlation coefficient about $r=-0.4$, was calculated using all native configurations that presented a folding success ratio of greater than 66% (full circles); 12 out of 52 configurations presented a success ratio of smaller than 66%. They are represented by empty circles and their folding rates were symbolically set at $\log \kappa_f=-8.0$; superposed empty circles represent cases with the same χ.

brackets [] indicate that all units, including those represented by asterisks, are all in the same plane. The patterns [*TTTT*] and [*TT*] are special cases of general sequences of uninterrupted T's, represented here by {T…T}; such sequences may include up to 23 T's and also are potential producers of secondary-type structures; Fig. 2(b) illustrates this case for a sequence of five T's.

These patterns can be seen as basic structural elements, such as helices, and parallel and antiparallel sheets;[26] in each particular structure, they occur in different numbers and arrangements in spite of the relative small chain size. Clearly, for a fixed chain size ($L=27$), the amount of a particular pattern necessarily limits the number of other pattern types in the same structure. For instance, a structure with a higher content of [*STTS*] patterns [which may have up to eight in a single structure; see Fig. 2(a)] leave no place for [*TT*] or [*TTTT*] patterns. Combinations of such patterns work as motifs in real proteins (in the sense of assembly of secondary structures) and, as it will be shown in Sec. III, they constitute important kinetic factors favoring rapid folding.

On the other hand, the presence of some particular patterns in a structure may affect the folding in the opposite way, that is, slowing down the process; some of them resemble pseudoknots as found in real proteins.[27,28] Basically, they are constituted by large planar loops forming, for example, snail and staple-like shapes, as identified in Fig. 2(d) and 2(f), respectively. About 24% of all CSA configurations present some combinations of them, as exemplified in Fig. 3.

III. RESULTS

Figures 4 and 5 show the simulation results for representative CSA structures; the selected 52 native structures cover the entire range of χ values and were topologically classified into seven categories and identified by specific symbols, as detailed below. Both figures exhibit the dependence of $\log \kappa_f$ on χ, but under two distinct conditions: In the former only the hydrophobic potential $\{e_{ij}=h_i+h_j\}$ was used, whereas in the latter steric constraints $\{c_{ij}\}$ were added, resulting in $\{e_{ij}=h_i+h_j+c_{ij}\}$.

As shown in Fig. 4, most of the cases present a folding success ratio of greater than 66% (for simulations in the time window t_w), as indicated by the dotted line connecting asterisks (right axis); 12 out of the 52 native structures did not reach this rate level of folding success, but in just one case only the native structure was never found (0% success). In general, one expects that the success ratio would grow if the time window t_w is increased, but extra runs using an extended time window (actually $2\times t_w$) reveal that for some native structures the folding success ratio does not change. This means that the designed chain has a specific chance of being trapped in different compact structures distinct from its native configuration. For the unique case with 0% folding success ratio shown in Fig. 4 (configuration number 50; see Table I), several different CSA configurations were visited during the simulation time window t_w, but the native structure itself was never found. Therefore, this threshold condition was used for fixing t_w in order to identify those native structures that, under the present established conditions (hydrophobic potential and temperature $\kappa_B T_p=1.0$), show a lower folding success rate. The thick line (linear regression) shows the dependence of logarithm of the folding rate κ_f on χ; the linear regression was obtained by using the results from those structures presenting a folding success ratio of greater than 66%; the excluded cases (12 structures) are represented in Fig. 4 by empty circles; their folding rates were all set at $\log \kappa_f=-8.0$ by hand in order to emphasize graphically their poor folding performance. A weak correlation between $\log \kappa_f$ and χ is observed; the correlation coefficient was $r=-0.4$.

Now, Fig. 5 shows a very distinct scenario: The effect of the steric constraints $\{c_{ij}\}$ on the folding kinetics is remark-

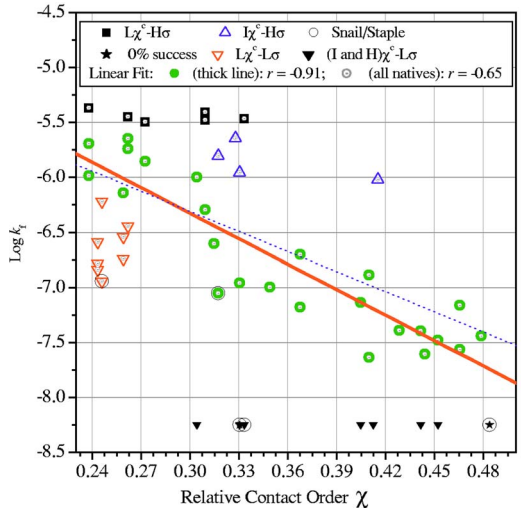

FIG. 5. Same as in Fig. 4 but now including steric constraints: (a) (■): correspond to those native configurations presenting folding rate at the highest level practically insensitive to χ; they present a low contact order, $\chi \leqslant 0.30$, and a high content of secondary-type patterns, namely, $L\chi-H\sigma$; see also Fig. 6; (b) (△): intermediate χ values ($0.30<\chi<0.42$) and a high content of secondary-type patterns: $I\chi-H\sigma$. (c) (▽): low contact order $\chi \leqslant 0.26$, but a low content of secondary-type structure: $L\chi-L\sigma$; see also Fig. 7. (d) (▼): contact order $\chi>0.3$, intermediate and high contact order, and low content of secondary-type patterns: (I and H) $\chi-H\sigma$; in this case the folding rate was intentionally set at $\log \kappa_f=-8.25$ to emphasize that the folding success for these cases was always below 80%. (e) (★) identifies those configurations with 0% of success for the time window t_w established; and (f) (◯) mark those native configurations that have structural patterns that resemble pseudoknots, named here as snail and staple; see also Fig. 2 and Table I. Finally, (●) identify the remainder of native configurations, which are characterized by a balanced mix of topological contacts. The linear regression was calculated for two conditions: first by using all successful structures (dotted line), and by the structures labeled by full circles (●, thick line).

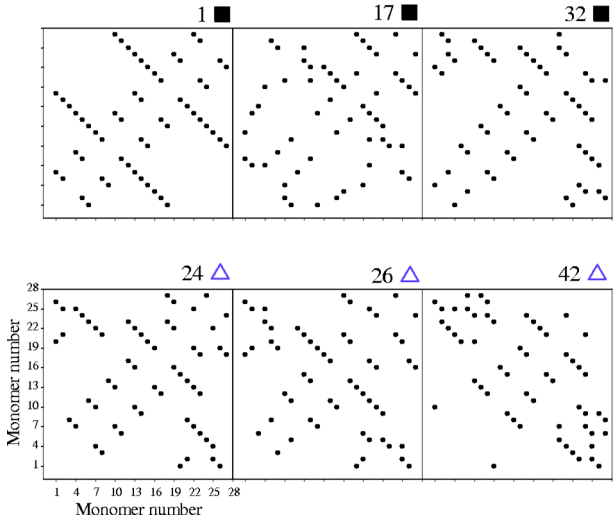

FIG. 6. Inter-residue contact maps; dots indicate contacting residues (lattice first neighbors). (a) Top three maps (■) illustrate the fastest "proteins": low contact order ($L\chi$) and high content of secondary-type structure ($H\sigma$), easily identified through the persistent lines along (and near) their main diagonal and lines parallel to the secondary diagonal, resembling α helices and β strands. (b) Down three maps (△): intermediate contact order ($I\chi$) and high content of secondary-type structure ($H\sigma$). Numbers and symbols on each map are the same as those of Fig. 5 and Table I, which provide numerical data about topological characteristics and folding rates for each structure. See also Fig. 2 for corresponding 3D structures.

able, enlarging the folding rate for more than two orders of magnitude for most of those native structures characterized by small χ. As a general rule, one may say that the effect of steric constraints on the folding kinetics enlarges as χ reduces, accentuating the correlation between $\log \kappa_f$ and χ. Indeed, taking into account all the cases with folding success ratios of greater than 80% (42 cases), the correlation coefficient was significantly increased to $r=-0.65$. However, note that the linear fitting shown in Fig. 5 ($r=-0.91$; thick line) corresponds to those structures labeled by full circles. The excluded cases, labeled by squares, stars, and triangles, constitute "notable exceptions," which are specially treated in this work. To deal with them, the influence of steric constraints on the folding kinetics was considered in detail by clustering the 52 native structures in seven groups. Six of them illustrate particularly how topological specificities can alter the overall linear dependency between $\log \kappa_f$ and χ, and the seventh group is composed of those native structures that behave closely to the mentioned linear correlation. Table I provides numerical data about topological characteristics and folding rates for all 52 native structures (stereochemical model). In the following, the topological peculiarities of each group are described and analyzed, one by one.

The first group is labeled by solid squares (■) and corresponds to those native states with a folding rate at the highest level (see Fig. 5). They have a relatively small contact order, $\chi \leqslant 0.34$, and folding rate κ_f practically independent on χ; $\log \kappa_f \simeq -5.4$. These CSA configurations have in common, what we call here, a highly effective "content of secondary-type structure," represented by σ, which is more easily visualized from their correspondent inter-residue contact maps than from their corresponding three-dimensional (3D) structures [see Fig. 6 (up three maps) and Fig. 2, respectively]. We note the following: (i) the persistent dotted lines along and close to the main diagonal, resembling α-helices (and parallel) β strands; see also Fig. 7–9); (ii)

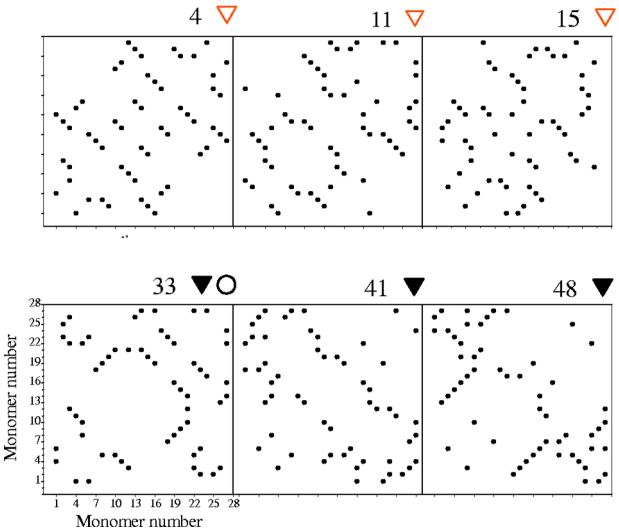

FIG. 7. Top three maps (▽): native configurations with low contact order χ and low $\sigma(L\chi-L\sigma)$. Down three maps (△): native configurations with intermediate and high contact order, and low content of secondarylike structures (I and $H\chi-L\sigma$). When compared with Fig. 6, these maps are more complex; they present diffuse and segmented regions forming bolder shapes.

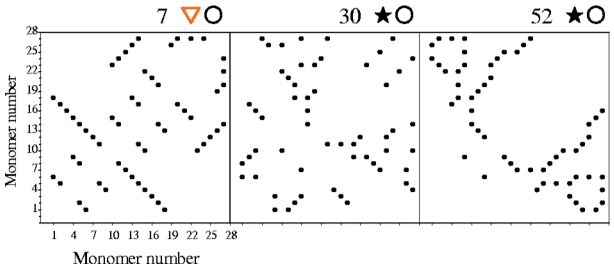

FIG. 8. Same as the two previous figures: ▽ and △ corresponding, respectively, to $L\chi$–$L\sigma$ and (I and H) χ–$L\sigma$.

dotted lines parallel to the secondary diagonal resembling antiparallel β strands; and (iii) a combination of the two, which should correspond to the composition of secondary structures (motifs). In the present lattice model, the final content of the secondary-type structure, σ, results from peculiar combinations of structural patterns, such as [*STTS*] and {T....T}. The native structures in this group are then characterized by their low contact order ($L\chi$) and high content of secondary-type patterns ($H\sigma$), shortly represented as $L\chi$–$H\sigma$. The structural [*STTS*] pattern is one of the ones responsible for $H\sigma$: Some native structures in this group have six or more [*STTS*] patterns (it is possible to have up to eight [*STTS*] patterns in a single structure). Also, uninterrupted sequences of T's can be folding facilitators, such as the case illustrated by structure 17, which shows a sequence of 15 uninterrupted turns, {T....T} as indicated in Fig. 2(e); it starts at the emphasized pattern [*TT*]. This sequence of T's is responsible for the long line along the main diagonal of its inter-residue contact map (see Fig. 6, configuration 17). Independently of χ, the number of [*STTS*] patterns in a particular structure is an important factor favoring the folding kinetics (see Table I). Extra runs involving native structures with a high density of [*STTS*] patterns or a long

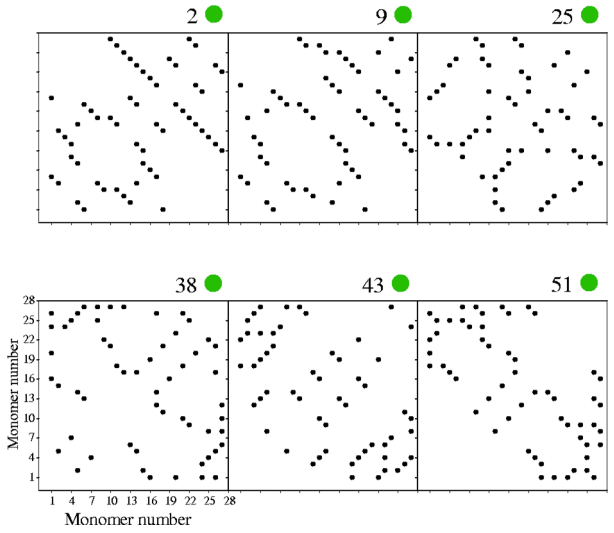

FIG. 9. Native structures presenting a balanced mix of topological contacts (short- and long-range contacts) according to its contact order χ. The configurations are shown in χ-crescent order from the top left; see Table I. Note that the relative content of the secondary-type structure (number and length of dotted lines parallel to the main and secondary diagonal) decreases as with the progressive change of χ but the secondary-type structures are always present.

uninterrupted sequence of turns (up to 23 turns) confirm this result: All of them presented folding rates at the highest level.

Some native structures presenting a contact order of intermediate values ($I\chi$: $0.30<\chi<0.42$) may also have a high content of secondary-type structures, $I\chi$-$H\sigma$ [see Fig. 6, down three maps (△)]. Accordingly, they also present folding kinetics at superior level, with log κ_f between −6.0 and −5.6, and are fairly independent of χ [see Fig. 5, (△)]. As in the previous case, the high σ is very manifest regardless of the χ value. It is interesting to observe in Fig. 6, and Table I that the folding rate decreases as the appearance of the inter-residue contact maps transfigure from a very simple skinny form, as simple as crossing perpendicular lines, into a more complex bold shape.

On the other hand, there are native structures with a low σ despite having a low contact order, $\chi\leqslant 0.26$, that is, $L\chi$-$L\sigma$, as illustrated by Fig. 7 (the top three native structures (▽)]. Indeed, in this case, the appearance of the inter-residue contact maps definitively changes into a much more complex shape: The dots are more spread out and the clear lines parallel to the main and secondary diagonal are now segmented or diffuse, producing kinds of inflated forms. Consistently with the premise that folding kinetics are dependent on σ, these native structures present a significantly smaller folding rate than should be expected if only their contact order was taken into account, as emphatically shown by Fig. 5 (▽); see also Table I.

Those structures presenting a folding success below 80% were considered as unsuccessful folding structures; all of them have a contact order $\chi>0.3$ (intermediate and high contact order) and a low content of secondary-type patterns ($I\chi$-$L\sigma$ and $H\chi$-$L\sigma$), as illustrated in Fig. 7, down three maps. The unsuccessful runs seem to be the result of taking particular pathways that drive the chain into dead-end conformations with respect to the native structure; anyway, for those runs that finally reached the native structure, they did it through folding rates at about the lowest level. Therefore, in order to emphasize their relative low folding performance, their folding rates were indiscriminately set at log $\kappa_f=-8.25$ [see Fig. 5 (▼)]. Note that they do not correspond necessarily to those native structures that presented low folding rates when using only the hydrophobic potential (see Fig. 4). This means that although the steric constraints improve the folding kinetics for most of the structures, for some structures it may slow down the folding rate and even prevent the folding at all—mainly for larger χ—as it will be considered as follows.

Indeed, in two cases the native structures never were found during the prefixed time window t_w; they are symbolically indicated at the bottom of Fig. 5 by filled stars (★). One of them has an intermediate contact order, namely $\chi=0.330\,69$, and the other is the structure with the highest contact order, $\chi=0.484\,13$; the respective inter-residues contact maps are shown in Fig. 8. Their folding failure is due to the low σ reinforced by the presence of one snail, structures number 7 and 30, and one staple, structure number 52; see Table I.

The two aforementioned target structures also belong to

the group labeled by an empty circle (◯), which identifies those configurations that present structural patterns tending to affect negatively the folding kinetics (snail or staple). Five native structures out of the selected 52 are in this category; they present distinct χ values and in all cases the corresponding folding rate is pushed to a lower folding kinetic level. We first analyze the one with low contact order, namely $\chi=0.246\,03$; it is identified as structure number 7 in Table I and in Fig. 2 as structure "d." In fact, comparing the native structures "a" and "d" of Fig. 2, one may note that the structural difference between them is due only to the change of two [*STTS*] patterns, structure "a," into a snail pattern, structure "d." This native structural change increases the χ value by just a small amount, namely from $\chi=0.238\,10$ to $0.246\,03$ ($\sim 3\%$), but affects the folding rate by almost two orders of magnitude (see Fig. 5). For all cases in this group [labeled with empty circles (◯)], the folding kinetics were severely affected by increasing their folding time with respect to what should be expected from their respective χ values, or even absolutely preventing their folding success, particularly for larger χ values. Such long-loop patterns involve a sequence of seven monomers without internal contact (that is, without contacting themselves), and so present much lower ways to create favorable conditions for a cooperative folding process than other patterns with a richer density of "local" contact along the chain.

The last group is labeled with full circles (●) and identifies the remaining native structures that are characterized by an equilibrated mix of distinct structural patterns: They present topological contacts that change continuously from short- to long-range contact, as χ increases. This particular balance among local and long-range contacts is better visualized through inter-residue contact maps, as illustrated in Fig. 9. It happens that most CSA configurations present such a mix of structural patterns (local and long-range contacts), and so they constitute a large subset of configurations that dominates the linear $(\chi, \log \kappa_f)$ fit. Therefore, our premise is that such structures are truly responsible for the linear dependence of $\log \kappa_f$ on χ; any structure breaking this χ-dependent equilibrium has its folding kinetics affected, increasing or reducing the folding rate with respect to the linear relation between $\log \kappa_f$ and χ. We return to this point later, at the end of this section.

Now, in order to emphasize the impact of some structural details on the folding kinetics, four native structures were considered (structure numbers 27–30; Table I); they are composed of distinct structural patterns, but all have the same (intermediary) $\chi=0.330\,69$. The first one (No. 27) has six repeated [*STTS*] patterns: Structures rich in such patterns show a folding rate at a higher level, like this one with $\log \kappa_f \simeq -5.95$; it is labeled with Δ in Fig. 5. Contrasting with this case, the next one (No. 28) has just one such [*STTS*] pattern, but now its folding rate is reduced to $\log \kappa_f \simeq -6.96$; it is identified with ●. The next two structures (Nos. 29 and 30) have no [*STTS*] pattern at all: The former shows low folding performance (▼) and the latter, which presents a snail pattern in its structure (◯), was never reached (★) during the simulation time window t_w. Several extra target structures presenting snails, staples, and similar shapes identified by large loops were additionally simulated; in most cases the folding kinetics were significantly slowed down. Exceptions include those cases in which more than one long loop was coordinated spatially, forming a long antiparallel β sheet.

The basic physical mechanism relating the content of secondary-type structures and folding rate involves the concept of cooperativity, in the sense of localness of the intrachain interactions. If the native structure presents suitable combinations of structural patterns rich in short-range contacts in a repeated fashion (as α helices and β sheets), then there are significant chances for a properly designed chain to undertake a fast folding process, because thermal fluctuations generate many favorable situations involving local contacts (sequence-local or topological-local contacts).[26,29] Therefore, a particular arrangement of patterns is cooperative if after one or few native contacts are formed, other native contacts follow with a much higher chance, as in the case of β sheets. The cooperative process is consummate if an adequately designed sequence is provided, in order to optimize the system entropy-energy competition. In the same sense, one can understand why for certain patterns and combinations of patterns a poor folding performance is verified.

Contact maps constitute a good manner to identify the amount of cooperative pattern arrangements, particularly by visual inspection. Therefore, based on such maps, and in order to (partially) incorporate more structural details (as the amount of secondary-type structures σ) in a single parameter, we rewrite the contact order expression, Eq. (1), as the contribution of two terms, that is,

$$\chi = \frac{1}{LN_c}\left(\sum_{\{n_s\}} \Delta_{ij} + \sum_{\{N_c - n_s\}} \Delta_{ij}\right), \tag{8}$$

where the first sum is taken complementarily over all contacting monomers belonging to secondary structures. It is presumed here that the number n_s [on the first sum of Eq. (8)] of units belonging to secondary structures is truly determinant of the folding reaction "velocity," as it has been shown explicitly through a selection of conformers whose native states are rapidly accessible from the denatured state.[30] Therefore, we reweigh the first sum of Eq. (8) in order to unify the concepts inherent to χ and σ parameters. Specifically, we consider any pair (i,j) of units belonging to an α helix or β-sheetlike structural patterns as having indistinctly the same weight $\delta=2$. For example, for the case of an antiparallel β strand with n' contacting residues, it contributes to χ according to the amount $\sigma' = \Sigma \Delta_{i,j} = 2+4+6 \ldots 2n'_s = (1+n'_s)n'_s$, but using our proposal its cooperative contribution would be $\delta=2$ for each contacting pair, resulting in $\sigma' = 2n'_s$. Then, an alternative relative contact order parameter, namely χ^c, may be written as

$$\chi^c = \frac{1}{LN_c}\left(n_s \delta + \sum_{\{N_c - n_s\}} \Delta_{ij}\right). \tag{9}$$

This new χ^c parameter should fix, up to a certain degree, the limitation of the relative contact order χ in detecting the details of a cooperative arrangement of structural patterns of target structures, as shown by Fig. 10. Indeed, as one can see,

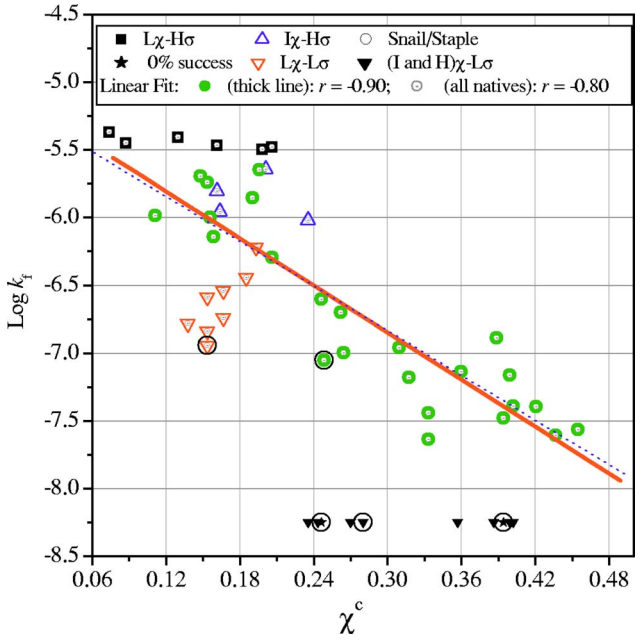

FIG. 10. Behavior of the logarithm of the folding rate κ_f as function of χ^c. The thick line indicates the linear regression determined by structures belonging to the group labeled with full circles (●); note that its behavior is unchanged with respect to Fig. 5. However, the correlation coefficient (dotted line) was significantly increased, actually from $r=-0.65$ (Fig. 5) to $r=-0.80$.

in comparing Figs. 5 and 10, the distinct groups of structures representing all different combinations of $H\sigma$ and $L\sigma, L\chi, I\chi$ and $H\chi$, are now clustered much more closely along the line determined by the overall successful structures. Particularly, a substantial increase on the correlation coefficient from $r=-0.65$ to -0.80 is achieved. Nevertheless, the correlation coefficient for the group labeled with full circles (●) is left unchanged; $r\simeq-0.90$ in both situations described by Figs. 5 and 10. This means that the remedy introduced by Eq. (11) is effective only for those cases in that the cooperative arrangement of patterns ($H\sigma$) is not reflected by its corresponding χ value. All other cases, namely structures presenting a balance among local and long-range contacts (group ●), or those with $H\chi$ are practically unchanged with respect to their linear (χ, log κ_f) behavior.

The sum on the right-hand side of Eq. (9) is the term responsible to produce distinct χ^c for structures with the same n_s (that is, the same σ). For n_s close to zero, then $\chi^c \approx \chi$; on the other hand, when n_s approaches N_c, then $\chi^c \simeq (LN_c)^{-1} n_s \delta$, with its minimum value being reached for $n_s = 28$ (see Table I, structure No. 1; Fig. 2 and 6), that is, $\chi^c_{\min} = \delta/L$; numerically, for $L=27$, it corresponds to $\chi^c_{\min} \cong 0.0741$. It is interesting to note, additionally, that the contribution for χ^c due to patterns that tend to affect negatively the folding kinetics is also contained in the sum on the right-hand side of Eq. (9). Therefore, even for structures with some significant H-σ, the folding rate may be slowed down by patterns like snail or staple.

IV. COMMENTS AND CONCLUSION

The effect of the steric constraints on the folding kinetics becomes evident when comparing the simulation results of the hydrophobic model ($h_{i,j}=h_i+h_j$) against the results achieved by the simulation of the stereochemical model ($e_{i,j}=h_{i,j}+c_{i,j}$): the folding rate is substantially increased when steric constraints are present, mainly for native structures presenting a small relative contact order χ, affecting the overall correlation between log κ_f and χ. It is remarkable that a set of fixed pairwise steric specificities $\{c_{i,j}\}$ conjugated with a native properly designed sequence of residues may help to drive the chain through an efficient folding process, especially for those configurations with a high content of secondary-type structures. Actually, since the findings of Ramachandran *et al.*,[31] the hard sphere model has been considered a good approximation for the chain configurational packing and other aspects of protein molecules;[32] however, steric constraints still have not been explored in great detail with respect to their effects on folding kinetics. In this context, it is interesting to compare the results of the present work with that of a recent work.[3] The effect of steric constraints (stereochemical model), as is the influence of the local-nonlocal cooperative coupling on the folding rate κ_f, are both globally χ dependent; they produce opposite results: Steric constraints always tend to enlarge the folding rate with respect to the hydrophobic model $\{c_{i,j}=0\}$, which is progressively accentuated as χ is reduced. On the other hand, local-nonlocal cooperative coupling reduces progressively the folding rate as χ increases, with respect to the original Gō model. This contrast seems to come from the fact that, while in the present model, the chain is efficiently guided to the native structure because steric restrictions minimize the nonnative interresidue contact[24] in the environment-dependent interaction, many more MC move attempts are refused than in the original Gō model, because the interacting strength is weakened if one or more chain segments are not nativelike.

There are CSA configurations for which the relative contact order χ does not correspond at all to the expected linear behavior between χ and log κ_f. Indeed, the existence of configurations with a high content of secondary-type structures (σ) for very distinct contact orders, ranging from $\chi=0.23$ to about 0.42, is the first evidence that the mix of distinct configurational patterns can be made in such a way that it may favor a cooperative folding process. The opposite is also true: Some combinations of configurational patterns, even presenting low χ, may interfere very significantly on folding kinetics, resulting in much lower rates. Particularly, this is emphasized by groups of native configurations with the same χ but that present folding rates that change drastically in function with particular combinations of configurational patterns in their corresponding structures. However, most of the CSA structures present a certain content of secondary-type structures σ that reduces as χ increases; because of this, a random sampling tends to reproduce the linear correlation between log κ_f and χ.

Therefore, the results presented here indicate that folding rate constants are largely determined by topological details of the native structure, as configurational pattern types and their combinations, and that global parameters, as the relative contact order, are not effective to detect them. This evidence cannot be directly extended to larger proteins, mainly due to the limitations of the lattice model used in this work;

however, since structures of large and small proteins are built up by combinations of well-defined hierarchical structural elements (secondary structures, motifs, and domains), it is tempting to assume that the same effect should be significantly influential in the determination of folding rate constants of larger proteins. Indeed, larger proteins may present hierarchical geometrical arrangements that are not possible in small proteins. Also, some topological concepts show intrinsic limitations to deal with proteins of different sizes. For example, our approach proposes that the term "content of secondary-type structures" be understood as a measure of the topological conditions for effective interplay of local and nonlocal intrachain interactions during the folding reaction, favoring the folding kinetics. This interplay is rather better understood from inter-residue contact maps of the correspondent native structure than from its relative number of residues belonging to α helices and β sheets. A second example is related to the relative contact order: For a single uniform helix, χ scales as $1/L$ (L is the protein length), which means that, once the linear correlation between $\ln \kappa_f$ and χ is assumed, a helical domain should fold faster as its length L increases, indicating that the protein size is not properly weighed up. In this aspect, changes on the contact order concept, treating topologically antiparallel β strands at the same level as α helix (the zipper hypothesis[33]) gives improvements in order to include σ in one single parameter.

As a final remark, we point to the interesting property of the set of steric constraints $\{c_{i,j}\}$ to group the native structures according to their respective effective σ, such as shown in Figs. 5 and 10. Its influence on the folding kinetics is in agreement with experimental results: (i) studies through ϕ-value analysis suggest a transition state for folding those that resemble the native structure, inasmuch as it is predominantly composed of elements from the native secondary structure;[34] (ii) extensive works using proteins of different sizes show that the transition state represents the rate-limiting step of the folding reaction,[8] that is, during the folding process, most of the time is expended to reach the main transition state, and after that the chain structure evolves through a rapid passage into the native state.[35] Therefore, at the end, one may equivalently say that folding rate constants are largely determined by main transition states, that in turn are mainly determined by σ. This result, together with additional properties of the stereochemical model,[24] suggest that there is an ubiquitous role for steric constraints on the protein folding problem, as has already been pointed out by several other results.[36,37]

ACKNOWLEDGMENTS

The authors thank Dr. Marco A.A da Silva for help on the enumeration of the CSA configurations. Part of this work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Proc. 00/12107-4.

[1] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[2] K. T. Plaxco, K. D. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[3] H. Kaya and H. S. Chan, Proteins: Struct., Funct., Genet. **52**, 524 (2003).
[4] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, and A. V. Finkelstein, Protein Sci. **12**, 2057 (2003).
[5] D. E. Makarov, C. A. Keller, K. W. Plaxco, H. Metiu, Proc. Natl. Acad. Sci. U.S.A. **99**, 3535 (2002).
[6] S. E. Jackson, Folding Des. **3**, R81 (1998).
[7] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **99**, 8637 (2002).
[8] B. Nölting, W. Schälike, P. Hampel, F. Grundig, S. Gantert, N. Sips, W. Bandlow, and P. X. Qi, J. Theor. Biol. **223**, 299 (2003).
[9] K. Kamagata, M. Arai, and K. Kuwajima, J. Mol. Biol. **339**, 951 (2004).
[10] L. Mirny and E. Shakhnovich, Annu. Rev. Biophys. Biomol. Struct. **30**, 361 (2001).
[11] H. Zhou and Y. Zhou, Biophys. J. **82**, 458 (2002).
[12] A. I. Jewett, V. S. Pande, and K. W. Plaxco, J. Mol. Biol. **326**, 247 (2003).
[13] D. E. Makarov and K. W. Plaxco, Protein Sci. **12**, 17 (2003).
[14] R. Bonneau, I. Ruczinski, J. Tsai, and D. Baker, Protein Sci. **11**, 1937 (2002).
[15] S. S. Jaswal, M. E. Truhlar, K. A. Dill, and D. A. Agard, J. Mol. Biol. **347**, 355 (2005).
[16] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[17] H. Abe and N. Go, Biopolymers **20** 1013 (1981).
[18] G. Srinivas and B. Bagchi, Theor. Chem. Acc. **109**, 8 (2003).
[19] E. Sali, E. Shakhnovic, and M. Karplus, Nature (London) **369**, 248 (1994).
[20] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).
[21] P. F. N. Faisca and R. C. Ball, J. Chem. Phys. **117**, 8587 (2002).
[22] R. A. da Silva, M. A. A. da Silva, and A. Caliri, J. Chem. Phys. **114**, 4235 (2001).
[23] K. Fan and W. Wang, J. Mol. Biol. **328**, 921 (2003).
[24] M. E. P. Tarragó, L. F. O. Rocha, R. A. da Silva, and A. A. Caliri, Phys. Rev. E **67**, 031901-1-7 (2003).
[25] K. A. Dill, Biochemistry **29**, 7133 (1990).
[26] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).
[27] M. L. Mansfield, Nat. Struct. Biol. **1**, 213 (1994).
[28] W. R. Taylor, Nature (London) **406**, 916 (2000).
[29] A. Caliri and M. A. A. da Silva, J. Chem. Phys. **106**, 7856 (1997).
[30] A. Maritan, C. Micheletti, and J. R. Banavar, Phys. Rev. Lett. **84**, 3009 (2000).
[31] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, J. Mol. Biol. **7**, 95 (1968).
[32] N. C. Fitzkee and G. D. Rose, Protein Sci. **13**, 633 (2004).
[33] K. A. Dill, K. M. Fiebig, and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **90**, 1942 (1993).
[34] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York, 1999).
[35] B. Nölting and K. Andert, Proteins: Struct., Funct., Genet. **41**, 288 (2000).
[36] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).
[37] J. R. Banavar, M. Cieplak, and A. Maritan, Phys. Rev. Lett. **93**(3), 238101 (2004).