doi:10.1016/j.jmb.2009.06.058 J. Mol. Biol. (2009) 392, 1303–1314

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

The Limited Role of Nonnative Contacts in the Folding Pathways of a Lattice Protein

Brian C. Gin[1,2,3], Juan P. Garrahan[4] and Phillip L. Geissler[1,2]*

[1]*Department of Chemistry, University of California at Berkeley, Berkeley, CA 94720, USA*

[2]*Chemical Sciences and Physical Biosciences Divisions, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

[3]*School of Medicine, University of California at San Francisco, San Francisco, CA 94143, USA*

[4]*School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, UK*

Received 25 March 2009; received in revised form 19 June 2009; accepted 23 June 2009
Available online 2 July 2009

Models of protein energetics that neglect interactions between amino acids that are not adjacent in the native state, such as the Gō model, encode or underlie many influential ideas on protein folding. Implicit in this simplification is a crucial assumption that has never been critically evaluated in a broad context: Detailed mechanisms of protein folding are not biased by nonnative contacts, typically argued to be a consequence of sequence design and/or topology. Here we present, using computer simulations of a well-studied lattice heteropolymer model, the first systematic test of this oft-assumed correspondence over the statistically significant range of hundreds of thousands of amino acid sequences that fold to the same native structure. Contrary to previous conjectures, we find a multiplicity of folding mechanisms, suggesting that Gō-like models cannot be justified by considerations of topology alone. Instead, we find that the crucial factor in discriminating among topological pathways is the heterogeneity of native contact energies: The order in which native contacts accumulate is profoundly insensitive to omission of nonnative interactions, provided that native contact heterogeneity is retained. This robustness holds over a surprisingly wide range of folding rates for our designed sequences. Mirroring predictions based on the principle of minimum frustration, fast-folding sequences match their Gō-like counterparts in both topological mechanism and transit times. Less optimized sequences dwell much longer in the unfolded state and/or off-pathway intermediates than do Gō-like models. For dynamics that bridge unfolded and unfolded states, however, even slow folders exhibit topological mechanisms and transit times nearly identical with those of their Gō-like counterparts. Our results do not imply a direct correspondence between folding trajectories of Gō-like models and those of real proteins, but they do help to clarify key topological and energetic assumptions that are commonly used to justify such caricatures.

Keywords: Gō model; nonnative contacts; lattice model; protein folding; principle of minimum frustration

Edited by M. Levitt

Introduction

The current understanding of protein folding has been strongly shaped by theoretical and computational studies of simplified models.[1] Such models are typically constructed by discarding fine details of molecular structure or by making simplifying assumptions about the energies of interaction among amino acid residues. A special class of models, based on Gō's insights, asserts that only a subset of interactions (those between segments of a protein that contact one another in the native state) is crucially important for folding.[2] The Gō model further assumes a unique energy scale for these native contacts. Here, we will focus on elaborated "Gō-like" models that allow for a diversity of native contact energies.

The Gō model was originally proposed as a schematic but microscopic perspective on the stability and kinetic accessibility of proteins' native states. It accordingly provided generic insight into issues of cooperativity and nucleation, and the

**Corresponding author.* 207 Gilman Hall, Department of Chemistry, University of California at Berkeley, Berkeley, CA 94720, USA. E-mail address: geissler@berkeley.edu.

Abbreviations used: CAO, contact appearance order; CDO, contact disappearance order.

relationship between sequence and structure.[1] Results from early studies employing Gō-like models established a basis for theories that focus on gaps in the spectrum of conformational energies[3,4] and the funnel-like nature of potential energy landscapes.[5–9] Corroborated by experiments, concepts intrinsic to and inspired by Gō-like models now form a canon of widely accepted ideas about how proteins fold.[1,10,11]

The role of Gō-like models, in comparison with experimental data, has expanded beyond its original conception as a schematic tool. Neglect of nonnative contacts offers substantial computational relief to numerical simulations, allowing thorough kinetic and thermodynamic studies to be performed even for detailed molecular representations.[12–15] For example, Gō-like models have been used successfully to predict the folding rates of small single-domain proteins from their native topologies.[16] Recently, several studies have ascribed additional significance to the detailed dynamic pathways defined by Gō-like models.[14] The transition states and order of folding events in Gō-like simulations of several single-domain proteins[17] and multimeric structures[18] were found to largely correspond with those inferred from an analysis of experimental ϕ-values. A recent comparison of equilibrium and kinetic data between Gō-like models and an experiment for the villin subdomain revealed that even the simplest topology-based models were able to reproduce a wide range of experimental results, but did not provide evidence sufficient to conclude that Gō-like predictions of folding order are followed by the real protein.[19] While these results suggest that Gō-like models can provide much more than schematic physical perspectives, the advisability of drawing detailed mechanistic conclusions from their folding dynamics remains unclear. Arguments in favor of a precise correspondence with realistic molecular dynamics are often based on theories suggesting similarities between the energy landscapes of real proteins and their Gō-like representations, as afforded by topological constraints and evolutionary pressures. Nevertheless, such general principles offer only rough guidance, and few computational studies have compared the folding pathways of Gō-like models and their "full" counterparts (in which nonnative contact energies are included) in a broad context.[20]

Very favorable interactions between segments of a protein that are not adjacent in the folded state generally impede folding. They might do so by introducing detours or traps en route to the native state or simply by stabilizing the ensemble of unfolded conformations.[21–23] It is often imagined that the former possibility plagues a vast majority of nonnatural amino acid sequences, which fold sluggishly, if at all.[24,25] According to this picture, nonnative contacts should feature prominently in the convoluted folding pathways of an undesigned sequence. Such kinetic frustration could pose several biological risks *in vivo*, where aggregation and slow response can be serious liabilities. Indeed, typical proteins taken from living organisms fold reliably and with relative efficiency.[26]

These notions and observations motivate a "principle of minimum frustration" asserting that natural amino acid sequences have been "designed" by evolution to minimize the disruptive influence of nonnative contacts on the dynamics of folding.[5] One might thus apply Gō-like models to these designed sequences with confidence, since omitted interactions are precisely the ones whose effects have been mitigated by natural selection. By contrast, one might expect Gō-like models to poorly represent the folding mechanisms of slowly folding molecules, whose nonnative interactions are presumably responsible for hampering pathways to the native state.[22,23]

Testing these ideas of sequence design and kinetic frustration is made difficult by several factors. Experimentally, microscopic details of folding kinetics cannot be resolved, but only inferred, from indirect observables or effects of mutations. Furthermore, the most concrete hypotheses stemming from the principle of minimum frustration involve Gō-like models, which cannot be realized in the laboratory. Computer simulations of detailed molecular representations can generate, at great cost, dynamic information sufficient to determine a folding mechanism for only the smallest of natural proteins.[27] Although the statistical dynamics of coarse-grained or schematic representations can be readily explored, biology does not provide collections of fast-folding and slow-folding sequences to compare in these artificial contexts. Finally, even when appropriate ensembles of sequences and ensembles of folding trajectories are available, a useful comparison of Gō-like models and their full counterpart requires a compact way of characterizing the course of highly chaotic dynamics.[28] A general method for this purpose is not available, although studies of nucleation as a rate-limiting fluctuation provide a useful starting point.[29,30]

This article presents the first systematic large-scale comparison of folding pathways within Gō-like and full models. We focus on a schematic lattice representation of proteins that are well suited for this task in several ways: (a) geometrically, because contacting segments of the chain can be unambiguously identified; (b) statistically, because representative ensembles of folding trajectories can be generated for large numbers of amino acid sequences; and (c) conceptually, because the essential competition between contact energetics and chain connectivity can be isolated from the complicating effects of secondary structure, side-chain packing, and so on. While these latter effects unquestionably bear, in important ways, on the folding of real proteins, it is nevertheless imperative to understand the fundamental physical scenarios that they enrich and modify. Indeed, much of biologists' working intuition for protein folding and design was developed in the context of similarly schematic models. Our results challenge some of those notions.

It has been conjectured that well-designed lattice heteropolymers fold through mechanisms that are determined solely by their native structures.[26] Were this hypothesis correct, for both full and Gō-like models, a comparison of fast-folding pathways in the two models would not be especially informative. In that case, the sequence of events that advance a molecule toward the native state (which we designate as its folding mechanism) would be exclusively a question of geometry and local mobility. We have found, to the contrary, that a wealth of folding mechanisms is possible even for a *single* native conformation.

Spanning a range of hundreds of thousands of sequences, with widely varying rates and mechanisms, the work reported in this article constitutes a thorough test of certain aspects of the principle of minimum frustration and addresses at a new level of kinetic detail the dynamic realism that can be expected from Gō-like models. Our results for the lattice heteropolymer model evidence a remarkably strong mechanistic correspondence between full and Gō-like models. Unexpectedly, this dynamic conformity holds not only for fast-folding sequences but also for the slowest sequences whose folding can be followed in practice. Close correspondence in folding mechanisms holds as long as the Gō-like approximation retains heterogeneity in the native contact energies of the full potential. These findings suggest a profound frustration invariance in the ensemble of trajectories that proceed from deep within the unfolded state all the way to the native structure.

Theory

Lattice model: Full and Gō-like representations

We focus on lattice heteropolymers, whose folding properties have been studied extensively for specific example sequences, structures, and chain lengths.[1,31] Here, a protein's conformation is described by a self-avoiding walk on a three-dimensional lattice with spacing a (see, e.g., Fig. 1a). Each vertex of this walk

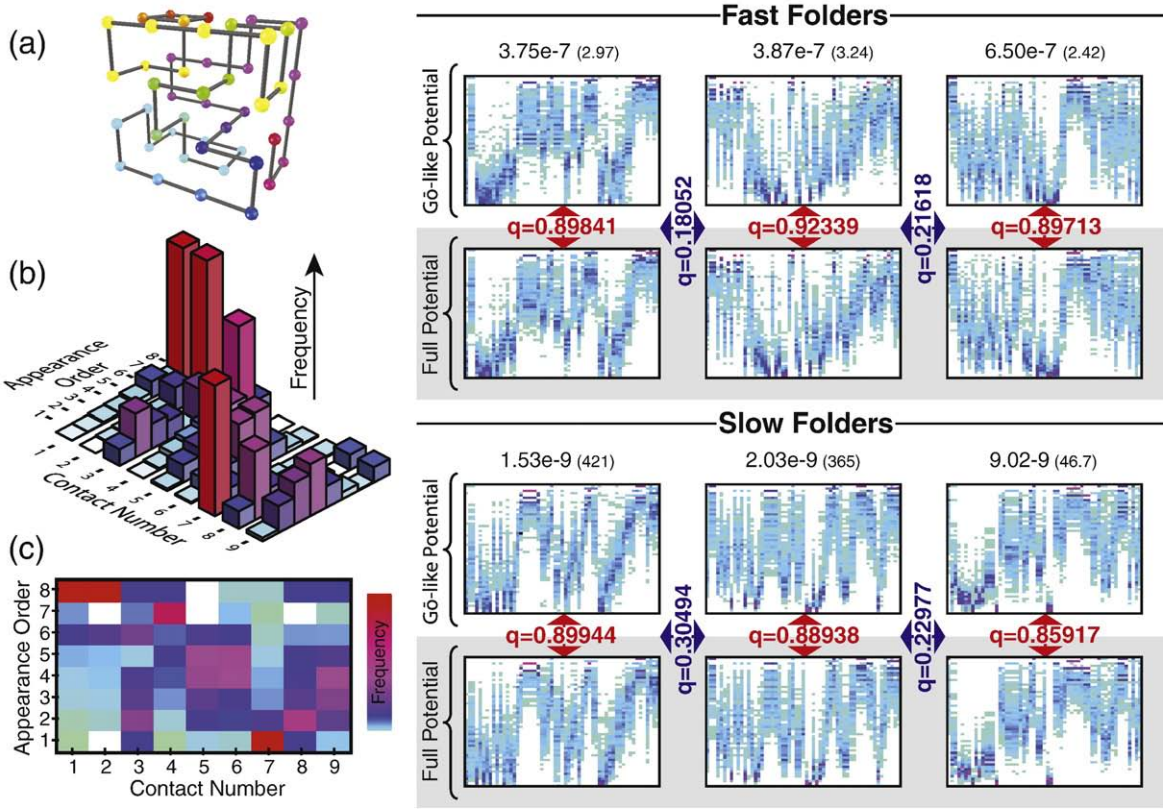

Fig. 1. (a) The 48-mer native structure of the lattice heteropolymer studied in this work. (b) Example of histograms of the order of permanent formation of native contacts (CAO) for each of the nine native contacts of a 12-mer lattice structure. Histograms are collected from the set of folding trajectories of a given amino acid sequence. (c) Same histogram set as (b), but shown as a density map. Right, top: CAO histograms of three fast-folding sequences of the 48-mer structure (a) for both the full potential energy and the Gō-like approximation (which disregards nonnative contact energies but maintains the original heterogeneity in native contact energies). The overlap parameter q quantifies the similarity of CAO histograms and, thus, topological folding pathways. The overlap of CAO histograms between full and Gō-like potentials for each sequence is close to 1 ($q \approx 0.9$), indicating the similarity of their folding mechanisms. In contrast, the overlap between CAO histograms of different sequences is much smaller ($q \lesssim 0.2$). Right, bottom: Same as before, but now for three slow-folding sequences. Again, the CAO distributions of full and Gō-like potentials are very similar, while those between different sequences are not. The folding rates of sequences in the full potential are given above their CAO histograms (in units of inverse Monte Carlo steps), and the ratio between their folding rates in the Gō-like and full potentials ($k_{\text{Gō-like}}$:k_{full}) is given in parentheses.

represents an amino acid monomer, which possesses no internal structure and interacts only with "contacting" monomers that occupy adjacent vertices. For a chain comprising N monomers, the energy of a particular configuration can thus be written as:

$$E = \sum_{i=1}^{N-1} \sum_{j>i}^{N} u_{\text{core}}(r_{ij}) + \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} B_{ij} \Delta(r_{ij} - a) \quad (1)$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$. The hard-core potential $u_{\text{core}}(r)$, which takes on a value of ∞ for $r=0$ and a value of 0 for $r>0$, imposes the constraint of self-avoidance. Interaction energies B_{ij} are determined by the sequence-dependent identities of monomers i and j, in accordance with the model of Miyazawa and Jernigan, and act only at a spatial separation of one lattice spacing [$\Delta(x)=1$ if $x=0$ and vanishes otherwise].[32] Dynamics are evolved according to a standard metropolis Monte Carlo algorithm (see Methods).

This caricature clearly lacks many of the chemical details underlying the function and secondary structure of real proteins. By capturing an essential interplay between diverse local interactions and constraints of polymer connectivity, it nonetheless recapitulates many nontrivial features of protein statistical mechanics: Even for chains of modest length (say, $N=27$), the number of possible conformations is sufficiently immense to motivate Levinthal's paradox (i.e., it is not obvious that they should be able to fold at all). Folding occurs in a cooperative fashion and occurs efficiently only for well-designed sequences. For a given sequence, certain residues figure much more prominently in folding kinetics than others; correspondingly, certain residues are more highly conserved than others in computer simulations of evolutionary dynamics.

The Gō-like approximation of the model of Eq. (1) is constructed simply by ignoring the energies of nonnative contacts:

$$\tilde{E} = \sum_{i=1}^{N-1} \sum_{j>i}^{N} u_{\text{core}}(r_{ij}) + \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} N_{ij} B_{ij} \Delta(r_{ij} - a) \quad (2)$$

where $N_{ij}=1$ if the monomers i and j are adjacent in the native configuration, and $N_{ij}=0$ otherwise. While disregarding the energy contribution of nonnative contacts, the energy function $\tilde{E}$ of Eq. (2) retains full heterogeneity in native contact energies of the original potential Eq. (1). We will show below that it is a crucial aspect of the Gō-like models that we study here.

The central issue addressed by our work is the nature of dynamic consequences due to neglect of nonnative interactions in Gō-like models. It is therefore important to rule out a trivial correspondence between full and Gō-like descriptions at the outset. For example, if all possible nonnative contacts of a particular sequence were less energetically stable than typical native contacts by a substantial amount, then projecting out those interactions could be expected to have little influence on folding kinetics and thermodynamics. Such a gap in the spectrum of possible contact energies exists for none of the sequences we have studied. Specifically, for most sequences, the number of possible nonnative contacts that are strongly attractive (compared to the energy of typical thermal excitations) is nearly equal to the number of strongly attractive native contacts. Even the fastest-folding sequences feature the possibility of potent nonnative interactions, whose influence on folding mechanism is not at all clear *a priori*. In the case of very slowly folding sequences, Gō-like models can lack many of the most energetically stable individual interactions of their full counterparts.

Contact appearance order

Many studies have previously suggested that lattice heteropolymers of modest length fold via a nucleation mechanism.[29,30] Formation of a handful of key contacts poises the system at a transition state, from which the chain can rapidly access the folded state or, with equal probability, return to the unfolded state. This set of crucial contacts comprises a "folding nucleus" and serves as a bare synopsis of dynamic pathways that lead to the native state.

A cogent comparison of folding mechanisms requires a means of characterizing dynamic pathways that is both thorough and computationally inexpensive. Identifying the folding nucleus satisfies neither of these necessities well. In particular, locating configurations from which the folded and unfolded states are equally accessible involves propagation of many trajectories and, by itself, does not delineate routes toward and away from the transition state.[33] We have devised an alternative measure that not only is succinct and computationally tractable but also characterizes the entire route from the unfolded state to the folded state. Specifically, we record the order in which native contacts form permanently during a protein's folding mechanism, which we term its "contact appearance order" (CAO). Our parameters thus chronicle lasting changes in the chain's "topology," understood in terms of linkages through the polymer backbone and through nonbonded contacts.

This CAO is a highly nontrivial measure of the progress toward folding and provides a detailed characterization of mechanism in the sense we have defined. It can be calculated simply from the time dependence of a trajectory spanning unfolded and folded states. Like persistence times[34] in the context of nonequilibrium systems such as glasses, it is intrinsically a multitime quantity; it can neither be computed for a single configuration nor be used to build constrained ensembles whose statistics shed light on the nature of reaction coordinates. But, also like persistence times,[34] it focuses attention on key dynamic events with unmatched precision. For the purpose of diagnosing the occurrence of lasting topological changes, CAOs serve almost ideally. For some other approaches (e.g., surveying free-energy

landscapes on which folding takes place), CAOs would serve poorly.

We have verified that the mechanistic meaning we ascribe to CAOs is consistent with more conventional characterizations of reaction progress. Most importantly, the order of a contact's appearance correlates strongly with a statistical measure of commitment to folding at the time when that contact forms permanently. We use the parameter p_{fold}—the probability that a trajectory initiated from a given configuration will reach the folded state before first relaxing to a state with few native contacts[35]—to demonstrate this fact. Figure 2c shows that the average value of p_{fold} rises steadily with CAO, from a value well below $p_{fold}=1/2$ to $p_{fold}=1$.

The point at which p_{fold} crosses 1/2 is often considered the transition state for folding. The set of contacts consistently present in such configurations is correspondingly designated as the folding nucleus. We have confirmed that the nucleus identified in this way corresponds closely with the set of contacts that have formed permanently when $p_{fold}=1/2$. Additionally, we have verified that the CAO-identified nucleus of several sequences from Mirny *et al.* is consistent with the nucleus identified in that study.[26] While this consistency check reflects favorably on the soundness of exploring folding mechanisms by scrutinizing CAOs, it does not imply that CAO analysis is predicated on putative nucleation mechanisms for folding. CAOs trace a history of conformational change that emphasizes any event with enduring topological consequences, regardless of whether the rate-determining steps in folding are uphill, downhill, or neutral in free energy; whether

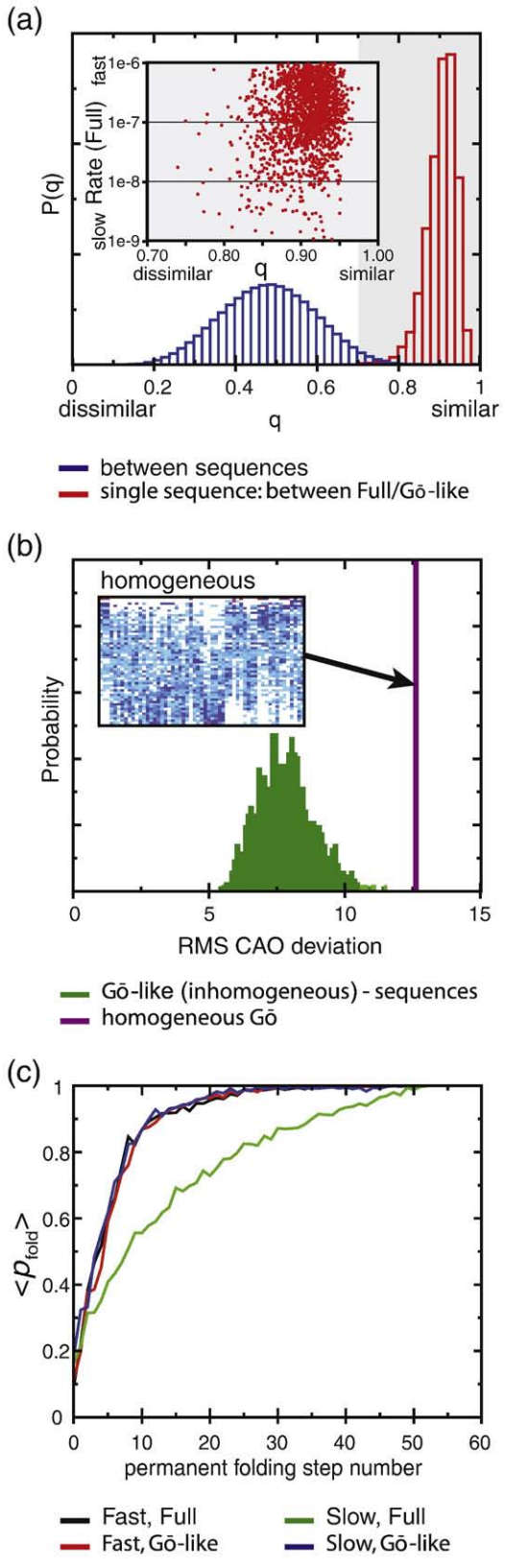

Fig. 2. (a) Distribution of CAO overlap $P(q)$ between different sequences, and between full and Gō-like potential, for 1000 sequences chosen randomly out of 10^5 sequences that fold to the 48-mer structure of Fig. 1a. The sequences in this distribution were generated by a single high T_{ev} evolutionary trajectory (see Methods). The inset shows that the similarity between full and Gō-like pathways for each sequence is independent of folding rate. Data for this inset were generated from 2000 sequences chosen randomly from five independent evolutionary runs (5×10^5 total sequences), all folding to the native 48-mer structure of Fig. 1a. (b) Distribution of the root-mean-squared fluctuations of contact order δC over the set of Gō-like sequences. CAOs in heterogeneous Gō-like potentials vary less from one folding trajectory to another than in the homogeneous Gō model. It is the heterogeneity in native contact energies that selects specific folding pathways; this selectivity is absent in a homogeneous Gō potential. The inset shows the CAO histogram for the homogeneous Gō model (cf. Fig. 1). (c) Average p_{fold} as a function of the number of permanent native contacts formed, for the full and Gō-like potentials, for a fast-folding sequence and a slow-folding sequence. In all cases, p_{fold} is close to 0 until the first permanent contacts are made, confirming that our CAO analysis captures the relevant dynamic folding regime. p_{fold} is the probability for a given conformation to reach the folded state before unfolding. For a given folding trajectory, we calculate p_{fold} in accordance with the method of Faisca *et al.* by running independent trajectories from configurations chosen at evenly spaced time intervals.[35] We regard a molecule as unfolded when the instantaneous number of native contacts drops to a value consistent with the average number of native contacts in the unfolded state. Additionally, we require that this threshold lie below any value found in equilibrium fluctuations of the native state.

folding is kinetically a two-state phenomenon; and whether the progress of folding is plagued by long-lived kinetic traps.

What CAOs do not resolve is the unproductive development of native structure. Attention is focused solely on segments of time evolution that bridge folded and unfolded basins of attraction. Occasional excursions within the unfolded state amass an atypically large number of native contacts but, due either to topology or to the presence of interfering nonnative contacts, do not in fact make progress toward folding. CAOs contain no information about these excursions. In comparing full and Gō-like models, we therefore make no statements about the character of such nonfolding dynamics. By exclusively examining the accumulation of native contacts, we also lose direct information regarding the evolution of nonnative contacts. If the rupture of a particular nonnative contact were a crucial step in the folding of a certain sequence, our methods would not detect its occurrence explicitly. We stress, however, that substantial nonnative structure is present when the first permanent native contacts are formed. We could therefore indirectly detect the significance of nonnative contact dynamics through influences on the pattern of early topological changes.

Compiling the order of permanent contact formation over many folding trajectories of a given sequence, we construct for each native contact a statistical distribution of CAO. Figure 1b and c illustrates how the set of resulting CAO histograms forms a visual fingerprint of a sequence's folding mechanism. Because the dynamic events that it chronicles span a wide range of p_{fold}, a CAO histogram characterizes not only the transition state for folding but also the dynamics of ascent to the transition state and descent from the transition state. The correspondence between an amino acid sequence and its CAO histogram is as subtle as (if not more so) the connection between sequence and native conformation that defines some of the most challenging aspects of the protein folding problem. Most of the results we will present concern a *single* native structure (shown in Fig. 1a for $N=48$), removing a potentially trivial agreement between full and Gō-like models. Even for this unique structure, sequences of the full model differing by only a few point mutations can exhibit qualitatively different CAO histograms, reflecting substantial changes in folding pathway. The distribution of contact energies can thus play a critical and complex role in determining folding mechanism, over and above dictating its end point. Given this nontrivial relationship, it would be surprising if nonnative contacts did not generally act to shape or bias CAO statistics.

The primary goal of this article is to compare the CAO statistics of sequences propagated using full and Gō-like models. In judging their similarities and differences, it is essential to establish for reference how significantly CAO histograms can vary, within either model, for sequences that fold to a common structure. As mentioned above, others have proposed that such variations are weak (i.e., that the topology of the folded structure prescribes a nearly unique topological route for folding). Using the algorithm described in Methods, we have generated an unprecedentedly diverse set of sequences that fold to the same target structure within the full model. As shown in Fig. 1, variations in CAO statistics within this set are much more substantial than previously thought. Any success of Gō-like models in reproducing the folding pathways of the full model cannot be attributed simply to their sharing a common native structure.

We quantify the similarity of CAO statistics (for two sequences within the same model, or for full and Gō-like models with the same sequence) using an "overlap" parameter[36] q such that $0 \le q \le 1$, with larger q representing greater similarity. Details of the calculation are given in Methods.

Results

In the ensemble of sequences we generated, the fastest-folding sequences access the native state more than 1000 times more rapidly than the slowest-folding sequences. CAO histograms were generated for all sequences, with each one evincing a well-defined topological pathway. Typically, the appearance order C of a given native contact varies from one trajectory to another by only a few positions (see the text below). This regularity belies substantial conformational fluctuations attending each folding event, which exert little influence on the formation of *permanent* contacts. Sharply peaked CAO histograms do not indicate a lack of complexity, but instead a successful characterization of forward progress along the reaction coordinate for folding.

Figure 1 shows CAO histograms for several sequences folding to this specific 48-mer structure (depicted in Fig. 1a). Results are presented for dynamics propagated according to both full and Gō-like models. Comparing these topological fingerprints across different sequences hints at the broad variety of possible folding pathways. Contacts essential to the early stages of folding for one sequence can be irrelevant in the pathway taken by another. This finding contrasts strongly with the "one-structure one-nucleus" hypothesis, bolstering recent reports of dissimilar folding nuclei.[30]

Strong variations in the topological folding pathways chosen from one sequence to another immediately indicate that the original homogeneous Gō model[28] cannot capture the folding behavior of a typical sequence. With a homogeneous set of native contact energies, that model can only discriminate between different native structures, not between different sequences that adopt them. In loose terms, the folding dynamics of the homogeneous Gō model resemble a superposition of those we determined for diverse sequences of the full model. Whereas a typical set of contact energies selects a well-defined folding pathway in the full model, an egalitarian set

of stabilizing energies permits a broad sampling of routes to the native state.

Gō-like models that embrace variety in native contact energies, however, capture the topological pathways followed by their full model counterparts with striking accuracy. CAO histograms obtained from full and Gō-like dynamics for any particular sequence can hardly be distinguished (see Fig. 1). Not only are the average CAOs of each contact nearly equivalent but also fine details of CAO statistics are unaffected by neglect of nonnative contact energies. While previous work hypothesized a dynamic correspondence for fast folders, the topological conformity of full and Gō-like mechanisms that we observe for slow folders is highly unexpected.

For sequences with folding rates of $\lesssim 10^{-9}$, we are unable to harvest folding trajectories in sufficient numbers to construct CAO histograms. According to microscopic reversibility, however, topological routes for folding are identical with time-reversed routes of unfolding. We have therefore extended our analysis of CAO for efficiently folding sequences to one of contact *disappearance* order (CDO) for very sluggishly folding sequences. The agreement between CDO histograms of full and Gō-like models is no less striking than that of the CAO histograms plotted in Fig. 1, even in cases where the "native" state is grossly unstable. These calculations are somewhat less straightforward: The order of first disappearance (CDO) is equivalent to the order of permanent appearance (CAO), but only for trajectories reaching the unfolded state *without* revisiting the native state. As such, they require specifying when a molecule has unfolded. For this purpose, we regard a molecule as unfolded when the instantaneous number of native contacts drops to a value consistent with the average number of native contacts in the unfolded state. Additionally, we require that this threshold lie below any value found in equilibrium fluctuations of the native state. We have verified that CAO and CDO histograms indeed match for sequences folding at moderate rates.

Quantitative measures of mechanistic diversity are presented in Fig. 2a. For each pair of sequences generated by our evolutionary simulation, we computed the similarity parameter q between CAO histograms for the full model. The resulting distribution of q values is broadly peaked at $q \approx 0.4$, signifying that there is a significant diversity of CAO pathways represented by the sequences in the ensemble. For each individual sequence, we also quantified the relationship between CAO histograms generated by full and Gō-like models. These q values are distributed much more narrowly about a considerably higher average, $q \approx 0.9$. With the use of sequence-to-sequence variation in CAO pathways as yardstick, the irrelevance of nonnative contacts to the topological folding pathway is beyond doubt. The inset to Fig. 2a emphasizes that this result has little to do with folding efficiency. Typical q values for the comparison of full and Gō-like models are just as high for the slowest folders examined as for the fastest folders.

Figure 2b quantifies the variation of CAO between folding trajectories. For each sequence, we quantify the root-mean-squared fluctuation in the contact order:

$$\delta C = \frac{1}{n_{\max}} \sum_{n=1}^{n_{\max}} \sqrt{\langle C_n^2 \rangle - \langle C_n \rangle^2} \qquad (3)$$

Figure 2b shows the distribution of δC among the ensemble of Gō-like sequences. It is peaked at a value of $\delta C \approx 7.5$. In contrast, for the homogeneous Gō model, $\delta C \approx 12.5$, indicating that CAO values are much more broadly distributed between trajectories (see inset to Fig. 2b). The homogenous Gō model indeed lacks the pathway specificity exhibited when contact energies are diverse, as in heterogeneous Gō-like models.

The relevance of CAOs to the folding dynamics is illustrated in Fig. 2c. For two sequences and their Gō-like approximations, it plots p_{fold}[33,37] as a function of the total number of permanent native contacts formed, averaged over 200 folding trajectories. p_{fold} gives the probability for trajectories initiated from a particular configuration to fold completely before visiting the unfolded state and provides a standard basis for defining transition states in complex systems.[33,37] Figure 2c shows that $p_{\text{fold}} \ll 1$ when the first permanent contact is formed. Since $p_{\text{fold}} = 1$, by definition, when the last permanent contact is formed, CAO histograms chronicle nearly the entire course of folding dynamics, all the way from the unfolded basin of attraction ($p_{\text{fold}} = 0$) to the native state ($p_{\text{fold}} = 1$).

The insensitivity of topological folding pathways to nonnative contact energies by no means implies a complete dynamic equivalence of full and Gō-like models. For example, a sequence's mean first passage time for folding can differ by as much as 3 orders of magnitude for full and Gō-like models. This discrepancy is larger for sequences with slower folding rates. Such discrepancies may be due to the presence of off-pathway traps in the unfolded state and possibly nonnative stabilized intermediates along the folding pathway. However, our calculations suggest that such marked distinctions are largely limited to dynamics occurring before the value of the committer function p_{fold} increases significantly from 0 (i.e., before significant progress has been made along the folding reaction coordinate).

As illustrated in Fig. 3a, we can divide each folding trajectory into a period before any permanent contacts are made (the "prefolding phase") and the remaining period in which a lasting native structure develops (the "folding phase"). Note that this division takes place well before a molecule commits to the folded state ($p_{\text{fold}} > 1/2$); indeed, the number of nonnative contacts at the beginning of the folding phase is typically comparable to that of the unfolded state. Figure 3b shows the distribution of prefolding and folding phases' durations for two sequences that are representative of fast and slow folders. In both cases, the influence of nonnative contacts on the folding phase dynamics is weak.

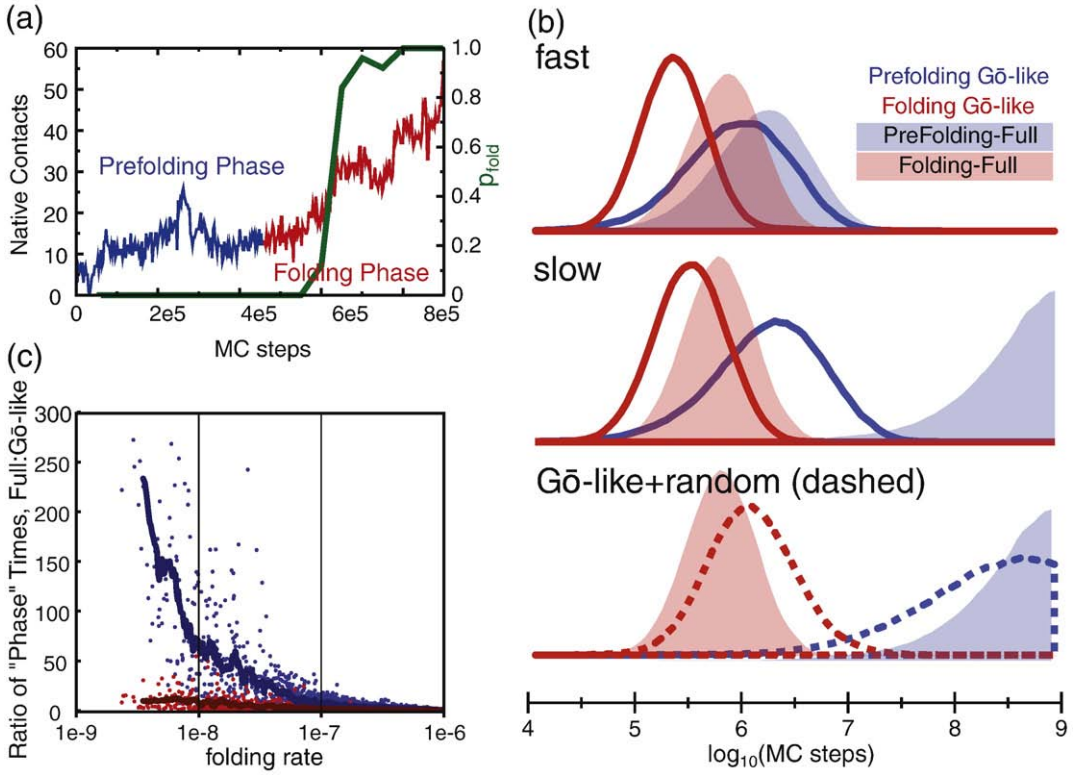

Fig. 3. (a) Number of native contact as a function of time in a folding trajectory illustrating the "prefolding" (blue) and "folding" (red) phases of the dynamics. The prefolding phase extends from the folding trajectory's start time until the time that the first permanent native contact is formed. The folding phase extends from this time to the time when the native conformation is reached. The full (green) curve shows the p_{fold}, which only departs from 0 after the folding phase has started (cf. Fig. 2). (b) Right: Distribution of the duration of prefolding and folding phases, in the full potential and its corresponding Gō-like approximation. For fast-folding sequences (top), the distributions for both folding and prefolding durations of the Gō-like model are close to those of the full potential. For slow-folding sequences (middle), the Gō-like model reproduces the distribution of folding duration, but underestimates the prefolding times. If the Gō-like potential of slow-folding sequences is supplemented by random nonnative contact energies (bottom), the prefolding distributions can be made to match without disrupting the correspondence in the folding phase distributions. (c) Ratio between full and Gō-like models' folding (red) and prefolding (blue) phase durations for all sequences ordered according to their full potential folding rate; full lines are the average ratios for each scatter plot. For fast folders, the average times, as calculated from the full and Gō-like models, are comparable both for the folding phase and for the prefolding phase. For slow folders, the prefolding time in the Gō-like model is much smaller than that in the full potential, and this difference increases with decreasing folding rate.

Nonnative contacts mildly extend the time required to complete folding after the first permanent contact is made by less than an order of magnitude. By contrast, prefolding dynamics of poorly designed sequences are quite sensitive to nonnative contact energies. For the example shown in the middle of Fig. 3b, the waiting period prior to the formation of a single permanent contact is roughly 3 orders of magnitude longer in the full model as in the Gō-like model. No such dilation is observed for sequences that fold quickly in the full model.

Because CAO is a sensitive measure of approach to the dynamic bottleneck for folding, our division of prefolding and folding phases is a kinetically meaningful one. Most importantly, $p_{fold} \ll 1$ throughout the prefolding dynamics, as seen in Fig. 3a, indicating that the system remains well within the unfolded basin of attraction. Only when permanent contacts are made does p_{fold} rise significantly, so that the folding phase entirely encompasses departure from the unfolded state and transit to the native structure. It is remarkable that nonnative contacts, which can substantially prolong dwell times in the unfolded state, exert no discernible influence on the topological folding order and only a small effect on the duration of folding phase dynamics.

Our simulations suggest that progress toward the native state is essentially orthogonal to the formation and rupture of nonnative contacts. A number of such contacts are certainly present over much of the course of folding, but they do little to decide what conformational rearrangements bring a chain closer to its transition state for folding. To further test this idea, we studied the folding dynamics governed by potential energy functions that combine aspects of full and Gō-like models. Specifically, we selected a set of nonnative contact energies at random from a Gaussian distribution (see Fig. 3b). The "frustrating" influence of these random energies matches precisely the behavior we have reported for the full

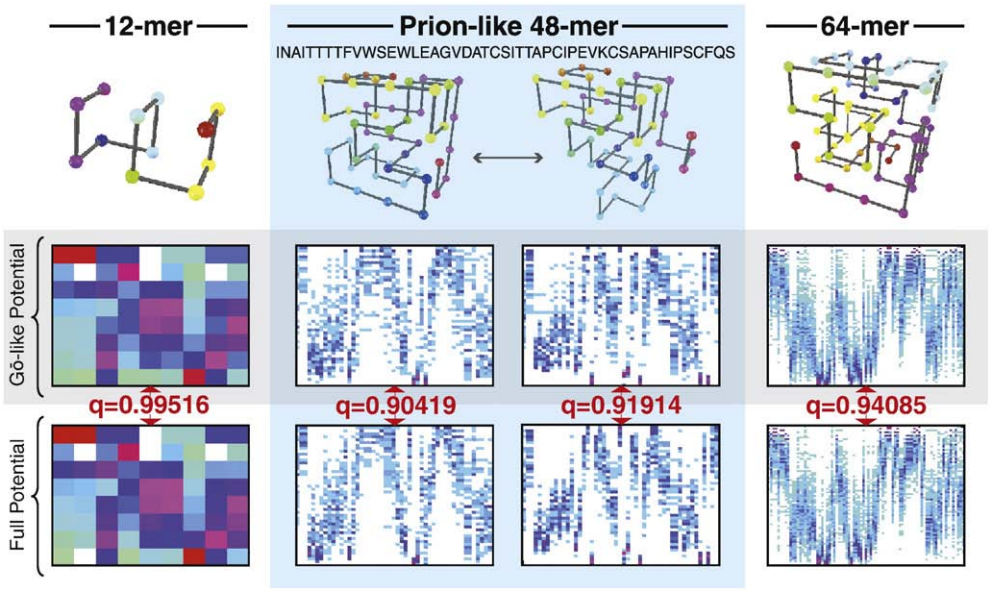

Fig. 4. The CAO correspondence between the full potential and the Gō-like approximation is robust to changes in chain length or target native structure. Left: CAO histogram for a 12-mer structure folding to the structure shown in the figure. Center: A sequence of the 48-mer structure of Fig. 1 that has a secondary stable configuration. Each target structure defines a Gō-like approximation from the set of their native contacts. Each Gō-like model predicts accurately the CAO histogram for folding to the corresponding structure. Right: Correspondence of Gō-like/full CAO histograms in a 64-mer structure.

model: CAO histograms are completely insensitive to the average strength and variance of nonnative attractions, while overall folding rates decrease with increasing nonnative attraction strength.

The observation of correspondence between the dynamics of the full lattice model and that of a heterogeneous Gō-like approximation does not noticeably depend on chain length or on details of the native structure. We have generated sequences with a range of folding rates for several native conformations of chains with lengths 8, 12, 48, and 64. For the two shortest chains, we used each maximally compact lattice structure as a folded state. For the two longest chains, we studied several native structures varying significantly in compactness and contact order.[38] Typical results shown in Fig. 4 highlight that the fidelity of Gō-like folding mechanisms is a very general feature of these lattice heteropolymers.

Discussion

Several arguments have been presented in the literature to justify the use of Gō models for studying the folding mechanisms of real proteins. Specifically, it has been asserted, based on the principle of minimum frustration, that evolutionary optimization of natural amino acid sequences removes kinetic barriers, rendering the energy landscape smoothly funneled and, therefore, Gō-like.[7,11] Biases due to topological features of the native state, unchanged in a protein's Gō-like representation, have also been invoked to argue for mechanistic fidelity.[39,40]

Our results largely support arguments drawn from the first assertion. For lattice heteropolymers of computationally manageable length, they strongly negate those of the second assertion.

In accordance with the principle of minimum frustration, we find that a Gō-like approximation can be quite accurate for the fastest-folding sequences: Not only their topological folding mechanisms but also the duration of their prefolding and folding phases are hardly changed by neglect of nonnative interactions. While nonnative contacts do slightly prolong the duration of both phases, the overall weak impact of nonnative contacts that we observed resonates with expectations from an energy landscape perspective on fast folders,[5] whose dynamics are imagined to proceed in a smoothly funneled way, free of significant frustration. Slower (less optimized) sequences exhibit substantially prolonged prefolding phases during which traps and intermediates are significantly populated, consistent with the dynamics of highly frustrated systems.

Surprisingly, however, we find that the topological folding mechanisms by which these slower sequences escape traps are identical with the unfrustrated folding pathways of their Gō-like counterparts. The duration of trajectory segments that span folded and unfolded states is also well captured by Gō-like models, despite the fact that a substantial nonnative structure must be disrupted en route. While neither predicted nor contradicted by the principle of minimum frustration, this finding may reflect a fundamental insensitivity of transition paths to nonnative interactions and merits further attention.

Lattice heteropolymers are perhaps the crudest representation of protein mechanics to which our analysis could be meaningfully applied. The correspondence between full and Gō-like folding mechanisms that we have revealed might break down in more detailed models. As a first step, it will be important to assess the validity of our results in simplified models involving explicit side chains. Kubelka *et al.* have suggested that side-chain fluctuations may effect a reduced diversity of interaction strengths, possibly enhancing the role of topology in shaping pathways for folding.[19] Additionally, it has been reported that lattice heteropolymers do not exhibit glassy folding dynamics even at very low temperatures, while non-Arrhenius temperature dependence naturally arises in slightly elaborated models that describe side-chain packing in addition to backbone conformation.[25] Gō-like energetics could alter folding pathways by abating the frustration underlying such glassy relaxation. This possibility, which merits further investigation, does not, however, negate the significance of our findings. Our primary purpose is not to justify the use of Gō-like models for a detailed study of real proteins' folding mechanisms but to establish the influence of nonnative interactions on dynamics intrinsic to the fundamental interplay between chain connectivity and heterogeneous contact interactions. That interplay, whose understanding is central to any instructive physical picture of protein folding, is not just present in simple lattice models but is also the exclusive source of their complexity. The results we have presented therefore establish an important point: Mechanistic aspects of protein folding that arise from the basic physics of heteropolymer freezing are remarkably insensitive to nonnative structure.

Methods

Lattice model dynamics

The standard dynamic rules for evolving a chain molecule proceed from a metropolis Monte Carlo algorithm. Trial moves, in which one or two randomly selected monomers move in an "edge-flip" or "crank-shaft" fashion, are accepted with probabilities that generate a Boltzmann distribution at a temperature $T=0.16\varepsilon_0/k_B$, where ε_0 sets the energy scale of the Miyazawa–Jernigan model. For example, the strongest attractive interaction (between two cysteines) has an energy $\varepsilon_{CC}=-1.06\varepsilon_0$; for lysine–lysine, $\varepsilon_{KK}=0.25\varepsilon_0$. Folding trajectories are initiated from swollen configurations drawn from a high-temperature ($k_BT/\varepsilon_0=100$) equilibrium distribution in which contact energies are negligible compared to typical thermal excitations.

Evolutionary sequence generation

Our method of sequence generation, which effects a biased random walk in the space of all possible sequences, is an extension of the method of Mirny *et al.*[26] To generate ensembles of sequences folding to a specific native structure, we introduce random point mutations and accept them with a metropolis probability:

$$P_{acc}=\min\left[1,\exp\left(-\frac{\Delta F^{\ddagger(\beta)}-\Delta F^{\ddagger(\alpha)}}{T_{ev}}\right)\right] \tag{4}$$

that generates a Boltzmann-like distribution. Here, $\Delta F^{\ddagger(\alpha)}$ is an estimated activation free energy for the folding of sequence α, $k^{(\alpha)}=k_0\exp(-\Delta F^{\ddagger(\alpha)}/k_BT)$. We estimate the folding rate constant $k^{(\alpha)}$ for sequence α, relative to the rate of basic microscopic motions k_0, by computing the fraction of trajectories $\langle h_{fold}\rangle_\tau\approx 1-\exp(-k^{(\alpha)}\tau)$ that fold within a fixed amount of time τ [with $k^{(\alpha)}\ll\tau^{-1}\ll k_0$]. This strategy offers two distinct advantages: (1) the evolutionary temperature T_{ev}, which governs the stringency of selection for efficient folding, can be controlled systematically; and (2) estimates of folding efficiency via $\langle h_{fold}\rangle_\tau$ can converge much more rapidly than the mean first passage time calculations employed in Mirny *et al.*[26]

Our evolutionary simulations, conducted at a moderate "temperature" $T_{ev}=0.008\varepsilon_0/k_B$, demonstrate that, in fact, many folding pathways can provide efficient access to a single native state. It is therefore not at all self-evident that a particular well-designed amino acid sequence should arrive at its native structure via similar routes in full and Gō-like versions of the lattice heteropolymer model.

Using this method, we have generated hundreds of thousands of sequences that fold to given structures (e.g., that of Fig. 1a) through a variety of folding mechanisms. This is the ensemble of sequences we use in this article. Further details of the evolutionary dynamics used to generate these large ensembles of sequences will be given in a forthcoming publication.

CAO overlap parameter *q*

To quantitatively compare the similarity between two CAO histograms, we define an "overlap parameter" q such that $0\le q\le 1$, with larger q values representing greater similarity. An analogy with spin glasses would assign an overlap $q^{(\alpha,\beta)}$ between the CAO distributions for two sequences α and β proportional to:

$$\frac{1}{n_{max}}\sum_{n=1}^{n_{max}}\sum_{C=1}^{n_{max}-1}P_n^{(\alpha)}(C)P_n^{(\beta)}(C) \tag{5}$$

where $P_n^{(\alpha)}(C)$ is the probability that native contact n is made permanently at order C in a folding trajectory of sequence α, and n_{max} is the total number of native contacts. An accurate numerical estimate of the quantity in Eq. (5), however, is problematic to obtain, requiring the generation of an inordinate number of folding trajectories. As an alternative, we define q using a closely related quantity:

$$q^{(\alpha,\beta)}=\frac{1}{n_{max}}\sum_{n=1}^{n_{max}}\left[\sqrt{2\left(\frac{\sigma_n^{(\alpha)}\sigma_n^{(\beta)}}{\left(\sigma_n^{(\alpha)}\right)^2+\left(\sigma_n^{(\beta)}\right)^2}\right)}\times\exp\left(-\frac{\left(\langle C\rangle_n^{(\alpha)}-\langle C\rangle_n^{(\beta)}\right)^2}{\left(\sigma_n^{(\alpha)}\right)^2+\left(\sigma_n^{(\beta)}\right)^2}\right)\right] \tag{6}$$

where $\langle C\rangle_n^{(\alpha)}=\sum_{n=1}^{n_{max}}P_n^{(\alpha)}(C)C$ is the average CAO of contact n for sequence α and $(\sigma_n^{(\alpha)})^2=\sum_{n=1}^{n_{max}}P_n^{(\alpha)}(C)\left(C-\langle C\rangle_n^{(\alpha)}\right)^2$ is its variance. Equations (5) and (6) are completely equivalent in the case of Gaussian-distributed CAOs. Even for non-Gaussian

statistics, $q^{(\alpha,\beta)}$ remains a useful, computationally tractable, and similarly bounded measure of how similarly two sequences fold.

Acknowledgements

We wish to thank D. Chandler, J. Chodera, K. DuBay, R. Jack, and S. Whitelam for useful discussions, and W. Eaton, E. Shakhnovich, and A. Szabo for critical reading of the manuscript.

This research used the resources of the National Energy Research Scientific Computing Center, which was supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231. This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, Chemical Sciences and Physical Biosciences Divisions, of the US Department of Energy under contract no. DE-AC02-05CH11231. In carrying out this work, B.C.G. was supported by the Ruben/Fatt Memorial Endowment, and J.P.G. was supported by Engineering and Physical Sciences Research Council grant GR/S54074/01. J.P.G. was a Visiting Pitzer Professor at the University of California at Berkeley during the time that this work was initiated.

References

1. Shakhnovich, E. (2006). Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev.* **106**, 1559–1588.
2. Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.
3. Shakhnovich, E. I. & Gutin, A. M. (1990). Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature*, **346**, 773–775.
4. Sali, A., Shakhnovich, E. & Karplus, M. (1994). Kinetics of protein folding. a lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614–1636.
5. Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 7524–7528.
6. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z. & Socci, N. D. (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl Acad. Sci. USA*, **92**, 3626–3630.
7. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Folding Des.* **1**, 441–450.
8. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600.
9. Socci, N. D., Onuchic, J. N. & Wolynes, P. G. (1998). Protein folding mechanisms and the multidimensional folding funnel. *Proteins*, **32**, 136–158.
10. Pande, V., Grosberg, A., Tanaka, T. & Rokhsar, D. S. (1998). Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **8**, 68–79.
11. Onuchic, J. N. & Wolynes, P. G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70–75.
12. Shimada, J., Ishchenko, A. V. & Shakhnovich, E. I. (2000). Analysis of knowledge-based protein–ligand potentials using a self-consistent method. *Protein Sci.* **9**, 765–775.
13. Shimada, J., Kussell, E. L. & Shakhnovich, E. I. (2001). The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation. *J. Mol. Biol.* **308**, 79–95.
14. Takada, S. (1999). Go-ing for the prediction of protein folding mechanisms. *Proc. Natl Acad. Sci. USA*, **96**, 11698–11700.
15. Shoemaker, B. A. & Wolynes, P. G. (1999). Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble. *J. Mol. Biol.* **287**, 657–674.
16. Muñoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA*, **96**, 11311–11316.
17. Karanicolas, J. & Brooks, C. L. (2003). Improved go-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.* **334**, 309–325.
18. Levy, Y. & Onuchic, J. N. (2006). Mechanisms of protein assembly: lessons from minimalist models. *Acc. Chem. Res.* **39**, 135–142.
19. Kubelka, J., Henry, E. R., Cellmer, T., Hofrichter, J. & Eaton, W. A. (2008). Chemical, physical, and theoretical kinetics of an ultrafast folding protein. *Proc. Natl Acad. Sci. USA*, **105**, 18655–18662.
20. Clementi, C. & Plotkin, S. S. (2004). The effects of non-native interactions on protein folding rates: theory and simulation. *Protein Sci.* **13**, 1750–1766.
21. Sali, A., Shakhnovich, E. & Karplus, M. (1994). How does a protein fold? *Nature*, **369**, 248–251.
22. Paci, E., Vendruscolo, M. & Karplus, M. (2002). Validity of Go models: comparison with a solvent-shielded empirical energy decomposition. *Biophys. J.* **83**, 3032–3038.
23. Paci, E., Vendruscolo, M. & Karplus, M. (2002). Native and non-native interactions along protein folding and unfolding pathways. *Proteins*, **47**, 379–392.
24. Shakhnovich, E. I. & Gutin, A. M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA*, **90**, 7195–7199.
25. Gutin, A., Sali, A., Abkevich, V., Karplus, M. & Shakhnovich, E. I. (1998). Temperature dependence of the folding rate in a simple protein model: search for a "glass" transition. *J. Chem. Phys.* **108**, 6466–6483.
26. Mirny, L. A., Abkevich, V. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA*, **95**, 4976–4981.
27. Schaeffer, R. D., Fersht, A. & Daggett, V. (2008). Combining experiment and simulation in protein folding: closing the gap for small model systems. *Curr. Opin. Struct. Biol.* **18**, 4–9.
28. Pande, V. & Rokhsar, D. S. (1999). Folding pathway of a lattice model for proteins. *Proc. Natl Acad. Sci. USA*, **96**, 1273–1278.
29. Abkevich, V., Gutin, A. M. & Shakhnovich, E. I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry*, **33**, 10026–10036.
30. Sutto, L., Tiana, G. & Broglia, R. A. (2006). Sequence of events in folding mechanism: beyond the go model. *Protein Sci.* **15**, 1638–1652.
31. Shakhnovich, E. I. (1994). Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* **72**, 3907–3910.
32. Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein

crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.
33. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2000). Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* **296**, 1183–1188.
34. Ritort, F. & Sollich, P. (2003). Glassy dynamics of kinetically constrained models. *Adv. Phys.* **52**, 219–342.
35. Faisca, P. F. N., Travasso, R. D. M., Ball, R. C. & Shakhnovich, E. I. (2008). Identifying critical residues in protein folding: insights from phi-value and pfold analysis. *J. Chem. Phys.* **129**, 095108.
36. Fischer, K. & Hertz, J. (1993). *Spin Glasses.* Cambridge University Press, Cambridge.
37. Du, R., Pande, V., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. I. (1998). On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334–350.
38. Weikl, T. R. & Dill, K. A. (2003). Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol.* **329**, 585–598.
39. Oliveira, L. C., Schug, A. & Onuchic, J. N. (2008). Geometrical features of the protein folding mechanism are a robust property of the energy landscape: a detailed investigation of several reduced models. *J. Phys. Chem. B*, **112**, 6131–6136.
40. Hills, R. D. & Brooks, C. L. (2008). Coevolution of function and the folding landscape: correlation with density of native contacts. *Biophys. J.* **95**, L57–L59.