Folding with a protein's native shortcut network

Susan Khor

Department of Computer Science, Memorial University of Newfoundland, St. John's, Newfoundland and Labrador, Canada

Correspondence
Susan Khor, Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada.
Email: slc.khor@gmail.com

Abstract
A complex network approach to protein folding is proposed, wherein a protein’s contact map is reconceptualized as a network of shortcut edges, and folding is steered by a structural characteristic of this network. Shortcut networks are generated by a known message passing algorithm operating on protein residue networks. It is found that the shortcut networks of native structures (SCNs) are relevant graph objects with which to study protein folding at a formal level. The logarithm form of their contact order (SCNlnCO) correlates significantly with folding rate of two-state and nontwo-state proteins. The clustering coefficient of SCNs (CSCN) correlates significantly with folding rate, transition-state placement and stability of two-state folders. Reasonable folding pathways for several model proteins are produced when CSCN is used to combine protein segments incrementally to form the native structure. The folding bias captured by CSCN is detectable in non-native structures, as evidenced by Molecular Dynamics simulation generated configurations for the fast folding Villin-headpiece peptide. These results support the use of shortcut networks to investigate the role protein geometry plays in the folding of both small and large globular proteins, and have implications for the design of multibody interaction schemes in folding models. One facet of this geometry is the set of native shortcut triangles, whose attributes are found to be well-suited to identify dehydrated intraprotein areas in tight turns, or at the interface of different secondary structure elements.

KEYWORDS
contact map, folding kinetics, folding pathway, logarithm contact order, triadic interactions

1 | INTRODUCTION

The contact map has long been a bedrock tool in the study of protein structure—arrangement of amino acids relative to each other in a protein molecule. In terms of graph or network theory, a contact map is an adjacency matrix; most simply, a binary one, with each entry indicating the presence or absence of a contact or an edge between a pair of nodes (amino acids) determined by a distance cutoff. A complex network approach to protein contact maps brings with it a slew of network-based measurements, largely of sociological origins, for example, short paths, cliquishness, centrality, closeness, and assortativity, whose relevance to intraprotein structure and dynamics have been posited.

Of particular interest to us is the small-world character of protein contact maps, which highlights the simultaneous presence of short paths and cliquish contacts between amino acids. We take the position that these two defining attributes of a small-world network are not coincidental or inevitable in proteins, but that the cliquishness of contacts serves to establish the short paths. This position is tenable if the intraprotein paths are generated via a local search (as opposed to a global search such as breadth first). Navigability of a small-world network, that is, the ability of a local search to construct short paths that traverse long distances in a sparse yet cliquish network, is not a given; it depends on the network having suitable cliquishness.

Previously we observed, with a local search algorithm called Euclidean-distance Directed Search (EDS) on protein contact maps called Protein Residue Network (PRN), that not only are globular proteins small-world networks, but that they are also navigable small-world networks.

A by-product of EDS is a set of PRN edges identified as shortcuts, because they help EDS to avoid backtracking. Shortcut edges are highly traversed by EDS paths, and have a significant role in the connectivity of nodes in a PRN via EDS paths. Since a protein folds to create appropriate intraprotein communication pathways to execute
its function and respond to external stimuli, shortcut edges could be a critical set of contacts that drives folding, or that needs to be made by folding. On the basis of this conjecture, a new type of protein contact map, one composed of a protein’s native shortcut edges, is its native shortcut network (SCN0), is proposed for the study of protein folding. The main goal of this work is to explore SCN0’s relevance to protein folding.

It helps that a protein’s SCN0 covers almost all of its amino acids, and is a connected graph. Furthermore, SCN0s capture essential structural characteristics of protein molecules, and produce sensible results when applied to traditional topological protein measures such as Relative Contact Order (Section 3.1). The primary evidences for SCN0s’ relevance to protein folding are the significant linear correlations they produce with folding kinetic variables (Section 3.2), and the reasonable protein folding pathways they generate within a native-centric coarse-grained diffusion-collision-esque model (Section 3.3). We believe ours is the first attempt at an explicit construction of protein folding pathways guided solely by a complex network metric (the SCN0 clustering coefficient) of native protein structures. A cursory probe of the contact attributes of SCN0 triangles (SCN0 cycles of length three) reveals traces of known protein folding cooperativity factors (Section 3.4). This unexpected agreement recalls the utility, and perhaps necessity, of considering ternary amino acid relations in physically-based folding models in addition to binary interactions, or more radically, to replace pairwise interactions with ‘suitable triple-body interactions as the basic interacting unit. A protein’s SCN0 triangles are speculated as candidates for such a set of trime interactions.

2 MATERIALS AND METHODS

2.1 Shortcut network

A shortcut network (SCN) is a sub-graph of a Protein Residue Network (PRN). The edges of a SCN comprise shortcut edges identified by the Euclidean-distance Directed Search (EDS) algorithm on a PRN. A SCN (PRN) derived from a native protein structure is denoted SCN0 (PRN0).

A PRN is a simple undirected connected graph. PRN construction is based on Refs. 16,17. A PRN node represents a protein residue. A PRN edge connects nodes u and v if and only if (1) their sequence distance |u - v| is > 1, and (2) their interaction strength, \( l_{uv} = \frac{20.5 \times 100}{R_i \times R_j} \geq 5.0. \) \( n_{uv} \) is the number of distinct pairs (i, j) such that i is an atom of residue u, j is an atom of residue v, and the Euclidean distance between atoms i and j is at most 7.5 Å. \( R_i \) and \( R_j \) are extracted from a table of normalization values by residue type (Table 1 in Ref. 16). Condition (1) is necessary since the inclusion of peptide bonds in PRN0s has a detrimental effect on shortcut edges (Supporting Information Figure S1). The cutoff values in condition (2) were settled upon through trial and error (details in Ref. 10) so that PRN0s are connected graphs without being unnecessarily dense, and that contacts between intraprotein domains are represented in PRN0s. The maximum Ca-Ca Euclidean distance of edges in PRN0s (≈14 Å) is about twice that in “conventional” Ca-Ca contact maps. Despite this, the modular structure of proteins is preserved in PRN0s. Besides, distant contacts (≥10 Å) can be functional. The maximum Ca-Ca Euclidean distance of edges in SCN0s is smaller (≈9 Å).

EDS is a greedy local search algorithm similar to Kleinberg’s, which formalizes the message-passing protocol used in Milgram’s social network experiment, but with backtracking to exploit a dynamic repository of distance information created anew with each search (Supporting Information Note S1). PRN edges that enable EDS to avoid backtracking are identified as shortcut edges.

2.2 Clustering coefficient

The clustering coefficient of a network is a measure of the network’s cliquishness. High cliquishness in a network implies a strong transitive dependency between contacts: the contact probability between two nodes which are already in contact with a common node is high. The clustering coefficient C of a graph G, is the average clustering coefficient over all of G’s nodes which number N: \( C_G = \frac{1}{N} \sum_{i} C(i) \). The clustering coefficient of a node i, C(i), is a measure of how complete the graph centered at i is. In a simple undirected un-weighted graph, \( C(i) = \frac{2e_i}{k_i(k_i-1)} \) where \( k_i \) is the number of nodes in contact with i, and \( e_i \) is the number of edges between the \( k_i \) nodes, that is, the direct neighbors of i. \( C_{SCN0} \) is the clustering coefficient of a SCN0.

2.3 Restricted binary collision (RBC) model

Protein “folding” in RBC begins with a set of protein fragments that correspond to secondary structure elements. A secondary structure element (SSE) is a maximal contiguous segment of a protein sequence with the same secondary structure, that is, either H (helix), S (strand), or T (turn). In each fold step, exactly two secondary structure units (SSU) are collided to form a new SSU and a new Conformation. A secondary structure unit (SSU) comprises one or more non-Turn SSEs, and all Turn SSEs sandwiched between any of its non-Turn SSEs on the protein chain. A Conformation is a combination of one or more SSUs such that it contains all the non-Turn SSEs of a protein. This process iterates until all the SSEs of a protein coalesce into a single SSU (Supporting Information Note S2).

Residues of a SSU must occupy a contiguous segment of the protein sequence. The purpose of this restriction is to prioritize the closure of smaller over larger loops, in accord with the principal that protein folding prefers paths that reduce the probability of premature loss in conformational entropy. A similar restriction in other folding models has been criticized for being un-physical for proteins such as barnase. This restriction can be lifted in RBC, but the consequence is less reasonable \( C_{SCN0} \) folding pathways for several test proteins. From the brief discussion here, it is clear that a number of variations on RBC are possible. This flexibility is common in model design, and it leaves room for different protein folding patterns, if necessary. For now, we stick with a simple version to explore the principle of using a structural metric, \( C_{SCN0} \), to identify folding pathways for several archetypical proteins.

Throughout RBCs assembly process, SSEs and SSUs are assumed to be rigid enough such that their shortcut edges (the part of SCN0
contained within them) remain unchanged. What is essential in RBC is the set of contacts, not the extent to which secondary structures are formed; the latter does influence the former, but does not determine it. With the set of Villin-headpiece subdomain Molecular Dynamics simulation generated structures (Supporting Information Note S5), we found that the mapping between SCNs and protein configurations is degenerate (not strictly one to one). Protein configurations with different root-mean-squared-deviation (RMSD) to the native structure are not produced. So can configurations with different counts of DSSP-determined alpha-helix residues. This degeneracy helps RBC, despite its apparent bottom-up hierarchic design, to be amenable to proteins that fold via the nucleation-condensation mechanism, wherein the formation of a set of native (usually chain distant) contacts precedes the concomitant development of secondary and tertiary structures. Besides, the so-called framework model requires that secondary structures be marginally stable only. A unified view of folding mechanisms for proteins is possible through the concept of an extended folding nucleus; the nucleation-condensation and framework models maybe two extremes of a single folding principle.

### 2.4 Conformation fitness

Let \( X \) be a Conformation with \( p \) single-SSE SSUs, and \( q \) multi-SSE SSUs. Then fitness of Conformation \( X \) is:

\[
F(X) = f\left(\bigcup_{i=1}^{q} \text{SSU}_i\right) + \sum_{i=1}^{p} f(\text{RS}_i = \text{SSU}_i)
\]

where RSU is the union of the residues belonging to the \( q \) multi-SSE SSUs in Conformation \( X \), RS\(_i\) is the set of residues belonging to SSE\(_i\), and \( f \) is either the structural or energetic fitness function defined below.

#### 2.4.1 Structural fitness

The \( C_{\text{SCN}} \) value for a set of residues is the clustering coefficient obtained with the native shortcuts whose endpoints both belong to the set of residues.

#### 2.4.2 Energetic fitness

Energy of a set of residues is calculated with DeepView's (version 4.1.0) Compute Energy (Force Field) function, which is a partial implementation of the GROMOS96 force field.

While RBC prohibits formation of SSUs that violate chain continuity, it allows interactions between chain distant multi-SSE SSUs when calculating Conformation fitness (first term in Equation 1). Thus, islands of contiguous segments distant from each other on a protein sequence may not be completely isolated from one another in RBC. This feature could be critical to capture nucleation contacts, which are predominantly long-range. Interactions between chain distant multi-SSE SSUs is for fitness calculation at a single fold step only; they do not produce a new SSU, and may change in future fold steps. A comparison of Equation 1 with an additive version shows that interactions between chain distant multi-SSE SSUs are critical to produce reasonable results for several proteins (Supporting Information Note S3).

However, too much interaction between folding substructures is problematic for large multidomain proteins. An ad hoc remedy is to limit the range of interactions between multi-SSE SSUs to a maximum of 100 residues, the characteristic maximum length of small two-state proteins. This means that fitness of a newly minted (multi-SSE) SSU is evaluated with all other existing multi-SSE SSUs, \( h_1, h_2, ... h_l \) in the first term of Equation 1 provided that the leftmost and rightmost ends of the protein sequence covered by \( g \) and \( h \) is < 100 residues in length. The multi-SSE SSUs that lie outside this boundary are evaluated in isolation, just like single-SSE SSUs in the second term of Equation 1. This range limited version of Equation 1 is tested on a two-domain protein, 7DFR.

### 2.5 Folding pathway

In a Conformation search graph, the nodes denote Conformations, and the edges denote (kinetic) accessibility of one Conformation from another. A Conformation \( Y \) that can be derived from another Conformation \( X \) via a binary collision of \( X \)'s SSUs, is accessible from \( X \). A folding pathway is a path in the Conformation search graph that starts at the Conformation with the maximum number of SSUs, and ends at the Conformation with a single SSU. The challenge is to identify a fitness gradient or rule to guide the selection of intermediate Conformations that make a plausible folding pathway for proteins in general.

The pathway construction rule proposed is: choose an accessible Conformation with the largest \( C_{\text{SCN}} \) at each step. This rule is motivated by the observation that \( C_{\text{SCN}} \) for a protein increases as it folds. In Supporting Information Figure S3, Conformation (1H) (3H 5H) is selected as the intermediate Conformation because it has the largest \( C_{\text{SCN}} \) of the Conformations accessible from Conformation (1H) (3H) (5H). To record the order in which SSUs form on a folding pathway, participants of a fold event are parenthesized, and each parenthesis-pair is given a unique integer such that elements in smaller numbered parentheses coalesce earlier. 1BDDs \( C_{\text{SCN}} \) folding pathway is (1H (3H 5H)12); this is its dominant unfolding pathway under condition of high temperature or high denaturant concentration.

1BDDs \( C_{\text{SCN}} \) folding pathway coincides with its \( E \) folding pathway, that is, the path of steepest descent in energetic values. But this need not be the case in general (e.g., Supporting Information Figures S12 and S13), and a demonstration of such convergence is also not our aim. Divergence between a protein's \( C_{\text{SCN}} \) and \( E \) folding pathways highlights the uniqueness of the structural fitness function—that its results are generally not reproducible by an energetic fitness function in RBC. Indeed, multiple folding pathways are possible for a protein, and \( E \) folding pathways are sensitive to the energetic function employed. Energetic fitness of Conformations are calculated to assess whether the \( C_{\text{SCN}} \) folding pathways exhibit a downward trend in terms of energy.

### 2.6 Protein structures and datasets

Of the protein chains examined (Supporting Information Table S1), 17 fold type (\( \alpha, \beta, \) and \( \alpha\beta \)) diverse structures (Supporting Information Table S5) are selected to test RBC's capacity to produce reasonable...
3 | RESULTS AND DISCUSSION

3.1 | Native shortcut edges capture the structural essence of proteins

This section shows that native shortcut networks (SCNs) are not random contact maps, even though they are identified via the EDS algorithm. This is done by examining how several well established structural characteristics of protein molecules pertinent to protein folding manifest in SCNs.

The nodes of a SCN cover almost all of the residues in a protein chain, $N_{SCN} \approx N$. The number of SCN edges scales linearly with $N$, $|SC| \approx 2N$ (Supporting Information Figure S1). A similar linear relationship between protein chain length and number of contacts is observed with Weikl's contact maps.22

The breakdown of a SCN's edges into short-range and long-range reflects a protein's secondary structure makeup (Supporting Information Figure S4). An edge is deemed long-range (LE) if its sequence distance is >10 residues apart. This LE cutoff is used so that contacts within secondary structure elements would generally be considered short range. On average, an $\alpha$-helix is 11 residues in length, and a $\beta$-strand, 6 residues.35 The proportion of native shortcuts that are short-range is larger in $\alpha$-rich proteins. This finding holds even when the LE cutoff is increased or decreased by four.

Native shortcut edges are significantly more enriched with hydrogen bonds26 than any other subset of PRN edges examined (Supporting Information Figure S5). The arrangement of native shortcut edges within ordered secondary structures ($\alpha$-helix and $\beta$-sheet) is more regular than within disordered regions (loops) (Supporting Information Figures S6 and S7).

Another way to assess the descriptiveness of SCNs of protein structure is with relative contact order (RCO), which was proposed as a complexity measure of a protein's native “topology”.11,37 $RCO = CO/N$ where CO is the average sequence distance of contacts, and N is the number of residues. SCN0_RCO, which is RCO calculated with native shortcut edges only, can classify proteins by their secondary structure composition (Supporting Information Figure S8). $\alpha$-rich proteins tend to have smaller SCN0_RCO values than $\beta$-rich proteins. This tendency bodes well for SCN0 as a graph object with which to study protein folding.

3.2 | SCN0 based structural metrics and protein folding kinetics

This section finds that structural metrics computed on SCNs are relevant to folding kinetics. Evaluated are the linear correlation between several known protein structural metrics, computed on native shortcut edges only, and four folding kinetic variables: (un)folding rate of two-state and nontwo-state folders, and transition-state placement and stability values of two-state folders.

The known structural metrics evaluated are: (1) relative contact order (RCO), (2) long-range order (LRO),38 (3) total contact distance (TCD),39,40 (4) logarithm of contact order (lnCO),41 and (5) clustering coefficient. Following are the original definitions of these metrics applied to a protein's native shortcut edges. SCN0 based RCO (SCN0_RCO) is defined in Section 3.1. SCN0 based LRO (SCN0_LRO) is $|SCL|/N$, where $|SCL|$ is the number of long-range edges in a SCN, and N is the length of the protein chain. SCN0 based TCD (SCN0_TCD) is $\frac{1}{N} \sum_{i=1}^{N} C_{SC} \left( S_{i} \right)$, where $|SC|$ is the number of SCN0 edges, and $S_{i}$ is the sequence distance of the $i$th native shortcut. SCN0 based lnCO (SCN0_InCO) is $\frac{1}{N} \sum_{i=1}^{N} \ln \left( S_{i} \right)$. SCN0 based clustering coefficient ($C_{SCN0}$) is defined in Section 2.2.

### TABLE 1 Pearson correlations between SCN0 based metrics and protein kinetic variables

| Variable                  | SCN0_RCO | SCN0_LRO | SCN0_TCD | SCN0_InCO | $C_{SCN0}$ |
|---------------------------|----------|----------|----------|-----------|------------|
| Two state folding rate    | -0.6111  | -0.7579  | -0.6550  | -0.7742   | 0.3447     |
| Nontwo state folding rate | 0        | 0        | 0        | 0         | 0.0100     |
| Two and nontwo state fold| -0.0782  | -0.4662  | 0.0724   | -0.5039   | 0.2743     |
| Transition-state placement| 0.5001   | 0.4713   | 0.5212   | 0.5009    | -0.5158    |
| Stability                 | 0.0151   | 0.0232   | 0.0108   | 0.0149    | 0.0118     |
| Nontwo state folding      | -0.1739  | -0.2965  | -0.1700  | -0.2203   | 0.5131     |
| (N > 100)                 | 0.4510   | 0.1919   | 0.4613   | 0.3373    | 0.0174     |

*p* values are reported below the coefficients. Bolded entries denote significant linear relationships (p-value < .05).
SCN0_inCO is the only metric that correlates significantly with most of the kinetic variables tested (Table 1). SCN0_inCO correlates significantly with folding rate of both two-state and nontwo-state folders, separately and when combined. This result is remarkable because, for the datasets including nontwo-state folding rate (Kamagata), SCN0_inCO is not significantly correlated (p-value \( \approx 0.6 \)) with \( N \), but \( N \) is significantly correlated (\( \approx 0.6 \)) with folding rate. \( N \) is a known determinant of folding rate for nontwo-state folders.42,43 The significant correlation between SCN0_inCO and combined folding rate, and the insignificant correlation between SCN0_inCO and \( N \), are robust to subset testing (Supporting Information Note S4). Hence, together with its CO-based design, SCN0_inCOs correlation with nontwo-state folding rate can be attributed to the influence of native-state topology. This attribution implies that the native-state topology hypothesis posited for small two-state folders through RCO11,23,37 can be extended to nontwo-state proteins as well, via SCN0_inCO. This extension hints at the possibility of a unified theory of globular protein folding based on native-topology. Kamagata et al.44 reached a similar conclusion, but their analysis is adulterated with protein size concerns, and depends on a small set of highly correlated data (Supporting Information Figure S9 middle right). A folding rate determinant that can accommodate mixed folding behaviors is ideal since it not only simplifies matters, but a hard divide between two-state and nontwo-state folding may not exist.43

SCN0_inCO is also significantly correlated with transition-state placement of two-state folders. Transition-state placement values reflect how similar the transition-state ensemble is to the native structure in terms of burial of hydrophobic surface areas.11 This finding reaffirms the hydrophobic effect on protein folding,45 and asserts its harmony with the influence of native-state topology on the transition state of two-state proteins.40,46

The CO metric was introduced to capture the relative difficulties posed by local and nonlocal contacts to protein folding.1,12,23 SCN0_inCO reconciles CO with polymer theory which theorizes that the entropy associated with the formation of a (long, length > 5) loop is related to the logarithm of its length.23 If the physical basis of folding rate dependence on CO is to be explained by the loss of configurational entropy as a result of loop formation, that is, the loop-closure principle,22,41 then InCO is more appropriate than CO as a single-value structural descriptor of proteins for folding kinetics. The numerator of SCN_inCO can be viewed as a "dynamic" topological descriptor, similar in concept to ECO (effective contact order).22 As a protein folds, contacts which are distant in a completely stretched out sequence, gradually become less distant with the help of contacts already formed.

SCN0_RCO, SCN0_LRO, and SCN0_TCD do not fare as well as SCN0_inCO. Namely, none of them are significantly correlated with nontwo-state folding rate. This poor performance can be reversed by limiting the dataset to large proteins (\( N > 100 \) residues) (Table 1). SCN0_RCO undergoes the most dramatic change, and SCN0_TCD the least. This reversal recalls the obfuscating effect improper handling of \( N \) can have on a metric's linear correlation with the folding rate variable, and highlights how size-independence of a topological based metric (or alternatively a dataset) can let the topological influence of a protein's native structure on folding shine forth more clearly.

Previously, RCO, LRO, and TCD (on non-SCN0 contact maps) have produced significant, and sometimes stronger, correlations with two-state folding rate than SCN0_inCO (Supporting Table S7). However, the linear correlation measure can be sensitive to differences in datasets, and there are major differences in the quality of the three datasets used (Supporting Information Table S8). The Zou-Ozkan dataset47 comprises more proteins in total and by fold type, a more even spread of proteins across fold types, and a greater range in folding rate per fold type. There are at least 16 data points per fold type in the Zou-Ozkan dataset, compared with a minimum of four in the other two datasets.

\( C_{SCN0} \) is the only metric that correlates significantly with all three kinetic variables: folding rate, transition-state placement and stability (Table 1). In this sense, \( C_{SCN0} \) is a more complete structural descriptor for two-state folding kinetics than SCN0_inCO. However, the \( C_{SCN0} \) correlations are weak, and are not observed on a single set of two-state folders constructed from the proteins common to the three datasets. The significant positive correlation between clustering coefficient and two-state folding rate, observed here and in Ref. 2, likely depends on the consideration of side-chain atoms38 during contact map construction. Working with \( \alpha_{C \alpha} \) distance <8.0 Å contact maps of 30 structurally diverse two-state proteins, Bagler and Sinha6 did not observe a significant correlation between clustering coefficient and folding rate. However, they did find a significant negative correlation between these two variables when only long-range (sequence distance > 12) contacts are considered. Nonetheless, this finding could be an artifact of the dominant influence of long-range contacts on folding rates.

\( C_{SCN0} \) is unique as the only metric (in Table 1) that correlates significantly with stability (a variable not examined in Ref. 2). The stability values are the experimentally determined free-energy to unfold a protein.22 We attribute this positive correlation to triadic relationships between nodes, which are not captured by the other metrics because they consider contacts in isolation and are therefore dyadic in nature. Intuitively, the third contact, which completes a triangle (the smallest clique possible), is seen as reinforcing the two contacts already present, and supplying an alternative path of communication amongst the three nodes should either of the original two contacts fail.49 This feature of geometrically induced contacts in highly cliquish networks led Micheletti2 to associate cliquishness with folding cooperativity and speed: highly cliquish contact maps facilitate more folding cooperativity and thus faster folding. However, this 'cooperative formation of native interactions'2 is a hindrance to transition-state placement, statistically speaking. The relationship between \( C_{SCN0} \) and transition-state placement values, which is a negative one, may be related to the principle of minimizing entropy loss. Early loop formation promotes early and possibly premature loss in configurational entropy, and so contacts in transition-state ensembles tend to be those that minimize entropy loss given the structured segments already present.42

\( C_{SCN0} \) is not significantly correlated with nontwo-state folding rate. This deviates from a previous result where the clustering coefficients computed on contact maps in Ref. 2 are significantly correlated (\( \approx 0.65 \)) with two-state and nontwo-state folding rates, separately and when combined. The nontwo-state folding rate dataset in Ref. 2 is smaller than our dataset (Kamagata), is less fold diverse (1 \( \alpha \), 2 \( \beta \), 9 \( \alpha \)\( \alpha \)),
and two-thirds (8/12) of its structures are large (N > 100). In contrast, the Kamagata nontwo-state folding rate dataset has 23 data points (5 α, 3 β, 15 αβ), and N > 100 for ∼61% (14/23) of its structures. When the 13 largest structures, which is enriched with αβ proteines (2 α, 11 αβ), in the Kamagata dataset are examined separately, SCN0 becomes significantly correlated (0.72, p-value = .01) with nontwo-state folding rate.

The number of long-range native shortcut edges (SCLE) correlates significantly with nontwo-state folding rate separately (-0.6940), and when combined with two-state folding rate (-0.7110). However, SCN0 is strongly correlated with N (r=0.7). Therefore, this finding merely shows SCN0’s relevance to folding rates, but broader implications, such as those implied by SCN0_lnCO, cannot be made. The geometric contact metric20 produces an even stronger correlation (-<0.8) with two-state and nontwo-state folding rates, but is very strongly correlated (0.91) with N. SCLE is not a significant determinant of transition-state placement, which is expected since transition-state placement is largely independent of size, but related to topological complexity of the native state.11

SCLE is not significantly correlated with stability, despite the role long-range contacts play in boosting stability of secondary structures in a protein’s core. SCLE produce negligible cliques by themselves; their contribution to stability becomes evident with their exclusion from SCN0, which weakens SCN0 correlation with stability to 0.5067 (p-value = .0191). This weakening is a consequence of disrupted critical interplay between short- and long-range contacts in a protein’s native structure.

Stability is significantly correlated with the number of short-range native shortcut edges (SCSE) (0.4944, p-value = .0227), but less so than when the triadic relationships in SCSE (SCN0 calculated with SCSE only) are considered (0.5067, p-value = .0191).

The SCLE and SCSE correlations with stability highlight the importance of triadic relationships, in particular those that involve a mix of short- and long-range contacts, in the study of protein structure for folding. Mutually reinforcing local and nonlocal contacts play a pivotal role in two-state folding.51 Three-body interactions can provide contextual information to distinguish such cooperative situations for pairwise interactions.19 Contacts in triadic relationships in SCN0s are characterized in Section 3.4, after witnessing how SCN0 directs assembly of protein segments along preferred folding pathways in the next section.

3.3 SCN0 folding pathways

This section reports on the reasonableness of the SCN0 folding pathways proposed by RBC for several well studied proteins (Supporting Information Table S9). In general, the SCN0 folding pathways are energetically favorable; Conformations on a SCN0 folding pathway decrease in energy with each fold step (Supporting Information Figure S10).

Proteins G (2IDG, 1GB1) and L (2PTL) are structural homologs with different preferred folding pathways.52 Both have one α-helix packed against a four-stranded β-sheet comprising two β-hairpins (Supporting Information Figure S11). Protein G initiates folding with the C-terminus β-hairpin (β3-4)2 while protein L begins with the N-terminus β-hairpin (β1-2).52,53 The G mutants (1MHX and 1MI0) were designed so that folding initiates at the N-terminus β-hairpin (β1-2).54,55 The SCN0 folding pathways for these five structures are reasonable as they correctly identify the expected first fold event. The folding idiosyncrasies of proteins in this canonical set also proved a stringent test-bed for ruling out alternatives to SCN0 for identifying initial fold events, such as number of edges, number of triangles, number of hydrogen bonds, a random subset of edges, SSE communication strength,56 and contact density.

The chymotrypsin inhibitor 2 protein (2CI2) is a classic model of the nucleation-condensation folding mechanism.57,58 It has one α-helix, a reactive site loop and a four-stranded β-sheet (Supporting Information Figure S12). The SCN0 folding pathway for 2CI2 is reasonable because it instigates folding with formation of the parallel β3-β2 substructure, consistent with simulation studies.59 The parallel β2-β3 substructure, which is stabilized by seven hydrogen bonds (incidentally there are seven native shortcuts that coincide with hydrogen bonds in Supporting Information Figure S6), is partially formed in the transition state,57 and together with residues from the α-helix, compose the major hydrophobic core of 2CI2.58 The β1 and β4 strands are unstructured at the transition state,57,58 although native β3-β4 interactions are observed with high probability at the transition state.25

Acylphosphatase (1APS) is a slow folding two-state protein comprising two α-helices packed against a five stranded antiparallel β-sheet. Like 2CI2, protein engineering (ε-analysis) study of 1APS reveals that most of its native contacts are only partially formed in the transition state.60 The SCN0 folding pathway for 1APS is reasonable because the first three fold steps involve β1, β3, and β5, which are the three most stable SSEs in the transition state,60 and where the three amino acids (Y11, P54, F94) sufficient to define 1APS’s nucleus,61 reside. Further, the first fold step (β2 β3)1 is supported by the well-structured loop between β2 and β3 in the transition state,60 and the β2-β3 substructure is present in the transition state.62 The last fold step which coalesce β4 β5 with β1 α2 β3 α2 refflects the formation of the minor hydrophobic core (V36, I86, L89), whose late consolidation is expected.60

The src-homology 3 (SH3) domains 1SHG and 1SRM are structural homologs with the same preferred folding pathways.62 Both comprise a five-stranded β-sheet (Supporting Information Figure S13). The SCN0 folding pathways for these two structures are reasonable because they identify the early formation of the central three-strand β-sheet (β2-β3-β4). The distal β-hairpin (β3-4) is the best formed structural element in the transition state.62,63 The diverging turn (just before β2), the n-src loop (between β2 and β3), and the distal-loop (between β3 and β4) are highly populated in computer simulations.64 The protein engineering method (ε-analysis) revealed that the diverging turn and β2 are partially ordered in the transition state and interact with the distal β-hairpin (β3-4) to form the central β-sheet.63 A hydrogen bond network,64 and a network of interactions between hydrophobic residues have been detected between the diverging turn and the distal β-hairpin in the folding transition state. Mutations on the central β-sheet significantly decreased the folding rate of the src SH3 domain.63
Protein S6 (2KJV) and its p54–55 circular permutant (2KJW) have the same tertiary structure even though the SSEs in 2KJV are rearranged in 2KJW. Their folding pathways are guided by the two ‘structural foldons’ present in S6 which work cooperatively with each other, that is, the development of one triggers the formation of the other.  

The $C_{SCN0}$ folding pathways for these two structures are reasonable according to existing experimental and simulation studies.  

On 2KJVs $C_{SCN0}$ folding pathway, elements of the first structural foldon ($\beta_1, \alpha_1,$ and $\beta_3$) are involved in the first two fold events, and the first structural foldon is formed by the third fold step. On 2KJWs $C_{SCN0}$ folding pathway, elements of the second structural foldon ($\beta_1, \alpha_2,$ and $\beta_4$) are involved in the first two fold events, and the second structural foldon is formed by the third fold step. $\beta_2$ contributes little to S6s folding, and the arrangement of SSEs in 2KJW permits it to be the last SSE to coalesce in RBC.

Depending on experimental conditions, the four $\alpha$-helix bundle ACBP (2ABD and 1ST7) can exhibit two-state or nontwo-state folding behavior. Its folding reaction has a rate-limiting native-like structure (RLNLS) where the N- and C-termini helix segments lie almost parallel to each other with the rest of the chain sparsely ordered with isolated structured fragments of low stability. Much effort has gone into illuminating the folding events that could lead up to this RLNLS. The $C_{SCN0}$ folding pathways for these two structures follow the sequence of folding events proposed in Ref. 73, that is, folding proceeds with formation of the $\alpha_2$-$\alpha_3$-$\alpha_4$ complex. The native $\alpha$-$\alpha$-$\alpha$-$\alpha$ 4 interactions in the RLNLS is actually the result of $\alpha_1$ interacting with the $\alpha_2$-$\alpha_4$ part of this complex.

Top7 (1QYS) is a small (N < 100 residues) de novo protein designed to exhibit noncooperative folding behavior. Its folding trajectory passes through a very stable intermediate stage where the C-terminus fragment (CFr) comprising one $\alpha$-helix ($\alpha_2$) and three $\beta$-strands ($\beta_3$, $\beta_4$, $\beta_5$) is well formed, while the N-terminus fragment remains disordered most of the time. On its $C_{SCN0}$ folding pathway, folding progresses sequentially from the C-terminus and forms the CFr by the forth fold step.

DHFR (dihydrofolate reductase) has two structural domains: ABD (adenine binding domain) between residues 38 and 88 ($\beta_3$, $\alpha_2$, $\beta_2$, $\beta_1$, $\alpha_3$), and a discontinuous domain (DD) spanning residues 1...37 ($\beta_5$, $\alpha_1$) and 89...159 ($\beta_4$, $\alpha_4$, $\beta_6$, $\beta_7$, $\beta_8$). At 159 residues, 7DFR (crystallized DHFR from E coli) is the largest protein to test our RBC with $C_{SCN0}$ method, hitherto. Due to its large size and multidomain composition, the range-limited version of Equation 1 (Section 2.4) is applied.

Evaluation of 7DFR’s $C_{SCN0}$ folding pathway relies on the analysis in Ref. 79, which found that protein fragments sliced by computational cuts (C-cut) are more stable than those carved by experimental cuts (E-cut). SSUs on the $C_{SCN0}$ folding pathway for 7DFR align more readily with the C-cut fragments (Supporting Information Figure S14). In the first nine steps, RBCs folding action concentrates on developing ABD and the N-terminus of DD. The domains ABD and DD need to cooperate with each other during folding as they are unstable independently. However, without the C-terminus of DD, these two domains can associate and form a stable conformation. Thus, the SSU in step 9 is a stable fragment.

The C-terminus of DD is unstable, and is the penultimate SSU to form. Without the 100-residue limit modification to Equation 1, this SSU forms earlier and produces SSU(52 H 45 5H) in the penultimate step, which splits ABD. Therefore, the 100-residue limit does have a significant positive effect for 7DFR. The result for 7DFR presented here is sensitive to SSE delineations. More analysis (perhaps of the C-cut dataset) and tests are required to generalize this method for large multidomain proteins. This approach has several advantages:

1. the so-called folding domains need not be known in advance,  
2. the self-folding steps of a protein sequence is identified from the bottom-up instead of from the top-down, and  
3. the guiding role of structure in the folding process is explicit. It also conforms to the idea of scaling up a folding mechanism that works for small single-domain proteins.

The Villin-headpiece subdomain (2F4K) is a very short (35 residue) three $\alpha$-helix protein. Its $C_{SCN0}$ folding pathway begins with the coalescence of the first and second $\alpha$-helices. This is reasonable since situated on them are three conserved phenylalanine hydrophobic core residues (F47, F51, and F58) that are crucial for the stability of the native wild-type villin structure.

For $C_{SCN0}$ to be a convincing structural descriptor of native-state topology to guide folding, its signal needs to be detected in non-native structures as well, especially in structures with high propensity to fold such as those at the transition state and beyond. On average, the non-native 2F4K structures generated by Molecular Dynamics simulation exhibit the same $C_{SCN0}$ bias as the native 2F4K structure, that is, $C_{SCN0}$ for Conformation(1H 3H) (5H) is larger than $C_{SCN0}$ for Conformation (1H) (3H 5H). Crucially, this bias is significantly stronger in the successful than the unsuccessful runs (Supporting Information Figure S16).

3.3.1 | Concluding section remarks

Both two-state and nontwo-state proteins are confronted by Levinthal’s paradox: the problem of searching a vast conformation space for the native-state conformation within biologically functional times. Existence of an intrinsic bias for the native-state in such a search is necessary to avoid a golf-course hole-in-one scenario. The findings of this section suggest the geometrical nature of this bias, which can be detected by $C_{SCN0}$ for proteins of varied fold types and folding behaviors.

We believe that stability, defined here by its general meaning as that of a measurable quality of persistence over time in the face of perturbations, underlies the bias detected by $C_{SCN0}$. There is a significant positive correlation between $C_{SCN0}$ and stability, not seen with the other metrics (Section 3.2). Stability is a theme that runs throughout protein folding theory, and unsurprisingly, the discussion in this section. The $\beta_1$-$\beta_2$, $\alpha_1$, and $\beta_3$-$\beta_4$ substructures of proteins G. L, and G’s mutants, coincide with substructures in the formation order suggested by rigidity analysis. The switch in protein G’s dominant folding pathway exhibited by its mutants are the result of increasing the stability of the N-terminus $\beta$-hairpin in protein G. $C_{SCN0}$ Folding events, especially initial ones, on the $C_{SCN0}$ pathways of 2CI2, 1APS, and the SH3 domain proteins are justified by the stability of substructures and persistence of (long-range or nonlocal) contacts at the transition state.
The role of stability in folding large proteins is demonstrated quite explicitly in 7DFR’s $C_{SCN0}$ folding pathway. Ensign et al. proposed that α helix to increase its stability to ensure fast folding of 2F4K. This proposal is confirmed by an analysis of SCN0 triangles (Supporting Information Table S12).

Stability, if only marginally, is a cornerstone of the structure formation dependency motif in proteins. The nucleation mechanism is principled on the stability of contacts in a folding nucleus to trigger order in the rest of a protein structure. The framework model is principled on the stability of isolated secondary structure elements. A subtle but important form of structure formation dependency is the concept of folding cooperativity: the ability of one segment undergoing structural increase to stimulate structural development in a sequence distant segment. This ability, which is epitomized in this section by the two foldons of the S6 protein, and by the first and second α-helices in 2F4K, is key to how biological protein sequences overcome their vast random search spaces to fold at the rates they do.

3.4 Attributes of native shortcut triangles

The geometrically induced ternary relations between amino acids in a SCN0 form (Euclidean) triangles. This section investigates the defining characteristics of native shortcut triangles which underlie the $C_{SCN0}$ folding pathways, to discern if they apply known principles of folding cooperativity. The sides of a native shortcut triangle are composed of three distinct edges that form a cycle in a SCN0.

A source of cooperative folding behavior in two-state proteins is spatial-proximity of nonbonded amino acids. The quality of simulated folding-rates for 52 two-state proteins improved when the attractive range of a contact in a 12-10 Lennard-Jones potential is limited to within 1.2 times its distance in the native structure. In this simulation, a pair of residues makes a native contact if their Cα-Cα distance <6.4 Å, or if each of them has a nonhydrogen atom that is <4.5 Å apart from the other. Using these cutoff values as guidelines, we label a native shortcut triangle spatially-proximal (SP) if its three sides are composed of contacts complying with the aforementioned attractive range of interaction. In our study of 12 representative proteins, at least 80% of triangles in a SCN0 are spatially-proximal (Supporting Information Table S13). This means that, by and large, native shortcut edges abide by the distance restriction for attractive interaction stipulated in Ref. 88.

SP triangles are not exclusive to SCN0s, but are present in PRN0s as well. Of the larger set of PRN0 SP triangles, at most 15% are SCN0 SP triangles (Supporting Information Table S13). A distinctive feature of SCN0 SP triangles is their significantly smaller circum-radius and significantly smaller in-radius, relative to PRN0 SP triangles. A circum-radius is the radius of a circle circumnavigating a triangle. An in-radius is the radius of a circle inscribing a triangle. Averaged over the 12 representative proteins, the SCN0 SP triangles have an average circum-radius of 3.58 Å, and an average in-radius of 1.67 Å, while the PRN0 SP triangles have an average circum-radius of 5.30 Å, and an average in-radius of 1.93 Å (Supporting Information Table S14).

When the points of a triangle (amino acids) are close to each other in sequence distance, a small circum-radius indicates high curvature in that sequence area, for example, a turn within a β-hairpin. When the points of a triangle (amino acids) are not close to each other in sequence distance (one or more long-range contacts are involved), a small circum-radius indicates a close encounter between two sequence distant protein segments, for example, between two SSEs. These two cases cover common scenarios in productive protein folding. The second case exemplifies another source of protein folding cooperativity: multibody interaction via context dependent evaluation of nonlocal contacts, that is, a nonlocal contact (i, j) is favored (energetically) if the sequence local neighborhoods centered on i, and on j, both show native-like proclivities.

A small in-radius is expected of SP triangles. Crucially, the average maximum in-radius of SCN0 SP triangles is 1.97 Å (Supporting Information Table S14). The volume of a water molecule in contact with an amino acid is 25.8 Å³, giving a radius of 2.05 Å. This implies that SCN0 SP triangles tend to be dry, that is, they are unlikely to contain water molecules. The dryness of SCN0 SP triangles recalls the ability of a solvation/desolvation barrier to stimulate protein cooperativity, evidenced by increased diversity in simulated two-state folding rate.

In conclusion, native shortcut triangles exhibit attributes associated with known principles of folding cooperativity. This association could be the reason for their success in identifying reasonable folding pathways in Section 3.3. Note that, both the size (number of nodes), and the number of triangles in a SSU determine a SSU’s $C_{SCN0}$. Based on these findings, we propose that SCN0 SP triangles identify dehydrated intraprotein areas in tight turns, or at the interface of different SSEs, crucial for productive protein folding. It would be interesting to see how the inclusion of triadic interactions embodied by SCN0 SP triangles affects a folding model. Such an attempt was made in Ref. 14 which reported improved results, but with triples of heavy atoms within a cutoff distance of 4.8 Å.

4 Conclusion

The network of shortcut edges present in a native-state protein (SCN0) is proposed as an effective structural abstraction of protein molecules for folding purposes. Shortcut edges are identified by the EDS algorithm on a Protein Residue Network of a native-state protein (PRN0) (Section 2.1). SCN0s are able to capture the essential structural characteristics of proteins (Section 3.1), and to produce single-value descriptors of protein topology that correlate significantly with protein folding kinetics data (Section 3.2). Notably, the logarithm form of SCN0 contact order ($C_{SCN0\\text{InCO}}$) correlates significantly with folding rate of both two-state and nontwo-state folders, and the clustering coefficient of SCN0 ($C_{SCN0}$) correlates significantly with stability of two-state folders. Because these two SCN0 based structural metrics are size-independent (they do not correlate strongly with protein size, N), they can produce significant correlations with transition-state placement data from two-state folders. Native topological influence on folding reveals itself most clearly in the presence of weak N effects.

The relevance of SCN0s to protein folding does not end with statistics, but with specifics. Within the Restricted Binary Collision (RBC) model, $C_{SCN0}$ could identify expected initial fold steps, and suggest...
reasonable folding pathways for 17 proteins (Section 3.3). RBC starts with isolated protein segments fragmented along secondary structure elements, which are incrementally coalesced to form the whole native structure (Section 2.3). The coalescence (pathway construction) rule seeks to maximize the sum of the clustering coefficients of the shortcut networks \( C_{SCNO} \) within substructures of a partially formed native structure, whilst allowing for restricted interaction between substructures (Section 2.4). Preliminary work with DHFR suggests that it is possible to apply our method, with additional length restriction, to multidomain proteins, and that stability of intermediate substructures is a key to its success.

An examination of SCNO contacts arranged in cycles of length three as a unit (SCNO triangles) finds that their characteristics are in accord with known principles of folding cooperativity (Section 3.4). This finding suggests the utility of including three-body interactions identified by SCNO triangles in native-centric coarse-grained computational models of protein folding.

**ACKNOWLEDGEMENTS**

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada. Thanks to C.N. Rowley for helpful discussions.

**ORCID**

Susan Khor http://orcid.org/0000-0002-9200-8063

**REFERENCES**

1. Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. Biophys J. 2004;86(1 Pt 1):85-91.
2. Micheletti C. Prediction of folding rates and transition-state placement from native-state geometry. Proteins. 2003;51(1):74-84.
3. Bagler G, Sinha S. Network properties of protein structures. Phys A: Stat Mech Appl. 2005;346(1-2):27-33.
4. Amitai G, Shemesh A, Sitbon E, et al. Network analysis of protein structure identifies functional residues. J Mol Biol. 2004;344(4):1135-1146.
5. Li J, Wang J, Wang W. Identifying folding nucleus based on residue contact networks of proteins. Proteins. 2008;71(4):1899-1907.
6. Bagler G, Sinha S. Assortative mixing in protein contact networks and protein folding kinetics. Struct Bioinform. 2007;23(14):1760-1767.
7. Song J, Takemoto K, Shen H, Tan H, Gromiha MM, Akutsu T. Prediction of protein folding rates from structural topology and complex network properties. IPSJ Trans Bioinform. 2010;3:40-53.
8. Vendruscolo M, Dockholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. Phys Rev E. 2002;65(6 Pt 1):061910.
9. Kleinberg J. Navigation in a small world. Nature. 2000;406(6768):845.
10. Khor S. Protein residue networks from a local search perspective. J Compil Netw. 2016;4(2):245-278.
11. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol. 1998;277(4):985-994.
12. Islam SA, Karplus M, Weaver DL. Application of the diffusion collision model to the folding of three-helix bundle proteins. J Mol Biol. 2002;318(1):199-215.
13. Yang Q, Sze SH. Predicting protein folding pathways at the mesoscopic level based on native interactions between secondary structure elements. BMC Bioinformatics. 2008;9(1):320.
14. Ejtehadi MR, Avall SP, Plotkin SS. Three-body interactions improve the prediction of rate and mechanism in protein folding models. PNAS 2004;101(42):15088-15093.
15. Banavar JR, Maritan A, Michelelli C, Trovato A. Geometry and physics of proteins. Proteins 2002;47(3):315-322.
16. Kannan N, Visheshwara S. Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol. 1999;292(2):441-464.
17. Chatterjee S, Ghosh S, Visheshwara S. Network properties of decoys and CASP predicted models: a comparison with native protein structures. Mol BioSyst. 2013;9(7):1774-1788.
18. Khor S. Comparing local search paths with global search paths on protein residue networks: allosteric communication. J Compl Netw. 2017;5(3):409-432.
19. Milgram S. The small-world problem. Psychol Today 1967;1(1):61-67.
20. Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature 1998;393(6684):440-442.
21. Fleibig KM, Dill KA. Protein core assembly processes. J Chem Phys. 1993;98(4):3475-3478.
22. Weiβ TR, Dill KA. Folding rates and low-entropy-loss routes of two-state proteins. J Mol Biol. 2003;329(3):585-598.
23. Alm E, Baker D. Matching theory and experiment in protein folding. Curr Opin Struct Biol. 1999;9(2):189-196.
24. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Bio- polymers. 1983;22(12):2577-2637.
25. Faisca PFN. The nucleation mechanism of protein folding: a survey of computer simulation studies. J Phys Condens Matter. 2009;21(37):373102.
26. Baldwin RL, Rose GD. Is protein folding hierarchical? I. Local structure and peptide folding. Trends Biochem Sci. 1999;24(1):26-33.
27. Baldwin RL, Rose GD. Is protein folding hierarchical? II. Folding intermediates and transition states. Trends Biochem Sci. 1999;24(2):77-83.
28. Fersht AR. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability and the extended nucleus mechanism. PNAS 2000;97(4):1525-1529.
29. Gianni S, Guydosh NR, Khan F, et al. Unifying features in protein-folding mechanisms. PNAS 2003;100(23):13286-13291.
30. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PDBViewer: an environment for comparative protein modeling. Electrophoresis 1997;18(15):2714-2723. http://www.expasy.org/spdbv/
31. Inanami T, Terada TP, Sasai M. Folding pathway of a multidomain protein depends on its topology of domain connectivity. PNAS 2014;111(45):15969-15974.
32. Itoh K, Sasai M. Flexibly varying folding mechanism of a nearly symmetrical protein: B domain of protein A. PNAS 2006;103(19):7298-7303.
33. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol. 1997;4(1):10-19.
34. Karplus M. The Levinthal paradox: yesterday and today. Fold Des. 1997;2(4):S69-S75.
35. Greene LH, Higman VA. Uncovering network systems within protein structures. J Mol Biol. 2003;334(4):781-791.
36. McDonald IK, Thornton JM. Satisfying hydrogen bonding potentials in proteins. J Mol Biol. 1994;238(5):777-793. http://www.earlham.edu/jak/structure_projection/realistic.html/
37. Plaxco KW, Simons KT, Ruczinski I, Baker D. Topology, stability, sequence and length: defining the determinants of two-state protein folding kinetics. Biochemistry 2000;39(37):11177-11183.
38. Gromiha MM, Selvaraj S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J Mol Biol. 2001;310(1):27-32.
39. Zhou H, Zhou Y. Folding rate prediction using total contact distance. Proteins 2004;58(1):113-121.
40. Bai Y, Zhou H, Zhou Y. Critical nucleation size in the folding of small apparently two-state proteins. Prot Sci. 2004;13(5):1173-1181.
41. Weikl TR. Loop closure principles in protein folding. Arch Biochem Biophys. 2008;469(1):67-75.
42. Koga N, Takada S. Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. J Mol Biol. 2001;313(1):171-180.

43. Ivanov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV. Contact order revisited: influence of protein size on the folding rate. Prot Sci. 2003;12(9):2057-2062.

44. Kamagata K, Arai M, Kuwajima K. Unification of the folding mechanisms of non-two-state and two-state proteins. J Mol Biol. 2004;339(4):951-965.

45. Fersht AR. Protein folding and stability: the pathway of folding of barnase. FEBS Lett. 1993;325(1-2):5-16.

46. Faisca PFN, Travasso RDM, Parisi A, Rey A. Why do protein folding rates correlate with metrics of native topology? PLoS ONE 2012;7(4):e35599.

47. Zou T, Oznak SB. Local and non-local native topologies reveal the underlying folding landscape of proteins. Phys Biol. 2011;8(6):066011.

48. Clementi C, Garcia AE, Onuchic JN. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atm representation study of protein L. J Mol Biol. 2005;352(3):933-954.

49. Attigian AR, Attigian C. Local motifs in proteins combine to generate global functional moves. Brief Funct Genomics. 2012;11(6):479-488.

50. Ouyang Z, Liang J. Predicting protein folding rates from geometric contact and amino acid sequence. Prot Sci. 2008;17(7):1256-1263.

51. Kaya H, Chan HS. Contact order dependent protein folding rates: kinetic consequences of a cooperative interplay between favorable nonlocal interactions and local conformational preferences. Proteins 2003;52(4):524-533.

52. McCallister EL, Alm E, Baker D. Critical role of $\beta$-hairpin formation in protein G folding. Nat Struct Biol. 2000;7(8):669-673.

53. Kim DE, Fisher C, Baker D. A breakdown of symmetry in the folding transition state of protein L. J Mol Biol. 2000;298(5):971-984.

54. Nauli S, Kuhlman B, Le Trong I, Teller D, Baker D. Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2. Prot Sci. 2009;11(12):2924-2931.

55. Nauli S, Kuhlman B, Baker D. Computer-based redesign of a protein folding pathway. Nat Struct Biol. 2001;8(7):602-605.

56. Karami Y, Laine E, Carbone A. Dissecting protein architecture with communication blocks and communication segment pairs. BMC Bioinform. 2016;17(S2):13.

57. Otzen DE, Izhaki LS, EIMasry NF, Jackson SE, Fersht AR. Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding. PNAS 1994;91(22):10422-10425.

58. Izhaki LS, Otzen DE, Fersht AR. The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. J Mol Biol. 1995;254(2):260-288.

59. Kmiecik S, Kolinski A. Characterization of protein-folding pathways by reduced-space modeling. PNAS 2007;104(30):12330-12335.

60. Chiti F, Taddei N, White PM, et al. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. Nat Struct Biol. 1999;6(11):1005-1009.

61. Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues in the folding of ACBP. Nature 1999;6(11):1005-1009.

62. Martinez JC, Serrano L. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. Nat Struct Biol. 1999;6(11):1010-1015.

63. Riddle DS, Grancharova VP, Santiago JV, Alm E, Ruczinski I, Baker D. Experiment and theory highlight role of native state topology in SH3 folding. Nat Struct Biol. 1999;6(11):1016-1024.

64. Grancharova VP, Riddle DS, Baker D. Long-range order in the src SH3 folding transition state. PNAS 2000;97(13):7084-7089.

65. Lindorff-Larsen K, Vendruscolo M, Paci E, Dobson CM. Transition states for protein folding have native topologies despite high structural variability. Nat Struct Mol Biol. 2004;11(5):443-449.

66. Lindberg MO, Oliveberg M. Malleability of protein folding pathways: a simple reason for complex behavior. Curr Opin Struct Biol. 2007;17(1):21-29.

67. Lindberg MO, Haglund E, Hubner IA, Shakhnovich EI, Oliveberg M. Identification of the minimal protein-folding nucleus through loop-entropy perturbations. PNAS 2006;103(11):4083-4088.

68. Sugita M, Kikuchi T. Analyses of the folding properties of ferredoxin-like fold proteins by means of a coarse-grained Go model: relationship between the free energy profiles and folding cores. Proteins. 2014;82(6):954-965.

69. Corrales M, Cusco P, Usmanova DR, et al. Machine learning: how much does it tell about protein folding rates? PLoS ONE. 2015;10(11):e0143166.

70. Kragelund BB, Osmark P, Neergaard TB, et al. The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. Nat Struct Biol. 1999;6(6):594-601.

71. Teilmann K, Kragelund BB, Knudsen J, Poulsen FM. Formation of hydrogen bonds precedes the rate-limiting formation of persistent structure in the folding of ACBP. J Mol Biol. 2000;301(5):1307-1314.

72. Fiefer W, Kristjansdottir S, Poulsen FM. Short-range, long-range and transition state interactions in the denatured state of ACBP from residual dipolar couplings. J Mol Biol. 2004;339(5):1191-1199.

73. Brum SW, Leisamantavicus V, Danielsson J, Poulsen FM. Cooperative formation of native-like tertiary contacts in the ensemble of unfolded states of a four-helix protein. PNAS 2010;107(30):13306-13311.

74. Thomesen J, Kristjansdottir S, Kragelund BB, Teilmann K, Knudsen J, Poulsen FM. Transient intermediary states with high and low folding probabilities in the apparent two-state folding equilibrium of ACBP at low pH. J Mol Biol. 2002;318(3):805-814.

75. Lindorff-Larsen K, Kristjansdottir S, Teilmann K, et al. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme A binding protein. J Am Chem Soc. 2004;126(10):3291-3299.

76. Kristjansdottir S, Lindorff-Larsen K, Fiefer W, Dobson CM, Vendruscolo M, Poulsen FM. Formation of native and non-native interactions in ensembles of denatured ACBP molecules from paramagnetic relaxation enhancement studies. J Mol Biol. 2005;347(5):1053-1062.

77. Zhang Z, Chan HS. Native topology of the designed protein Top7 is not conducive to cooperative folding. BiophysJ. 2009;96(3):L25-L27.

78. Gegg CV, Bowers KE, Matthews CR. Probing minimal independent folding units in dihydrofolate reductase by molecular dissection. Prot Sci. 1997;6(9):1885-1892.

79. Isom T, Ma B, Tsai C-J, Nussinov R. Molecular dynamics simulation of Escherichia coli dihydrofolate reductase and its protein fragments: relative stabilities in experiment and simulations. Prot Sci. 2001;10(1):135-148.

80. Tsai C-J, Nussinov R. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. Prot Sci. 2008;6(1):24-42.

81. Tsai C-J, Maizel JV, Nussinov R. Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. PNAS 2000;97(22):12038-12043.

82. Frank BS, Vardar D, Buckley DA, McKnight CJ. The role of aromatic interactions in ensembles of denatured ACBP molecules from paramagnetic enhancement relaxation studies. J Mol Biol. 2005;347(5):1053-1062.

83. Ensign DL, Kasson PM, Pande VS. Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the Villin headpiece subdomain. Proteins 2009;71(3):680-687.

84. Levinthal C. How to fold graciously. In: Debrunner JTP, Munck E, eds. Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House. Monticello, IL: University of Illinois Press; 1969:22-24.

85. Wallin S, Chan HS. A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling. Prot Sci. 2009;14(6):1643-1660.

86. Thomas S, Tang X, Tapia L, Amato NM. Simulating protein motions with rigidity analysis. J Comp Biol. 2007;14(6):839-855.
87. Chan HS, Zhang Z, Wallin S, Liu Z. Cooperativity, local-nonlocal coupling and nonnative interactions: principles of protein folding from coarse-grained models. *Annu Rev Phys Chem.* 2011;62(1):301-326.

88. Kaya H, Uzunoglu Z, Chan HS. Spatial ranges of driving forces are a key determinant of protein folding cooperativity and rate diversity. *Phys Rev E.* 2013;88(4):044701.

89. Perkins SJ. Protein volumes and hydration effects. The calculations of partial specific volumes, neutron scattering matchpoints and 280 nm absorption coefficients for proteins and glycoproteins from amino acid sequences. *Eur J Biochem.* 1986;157(1):169-180.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Khor S. Folding with a protein’s native shortcut network. *Proteins.* 2018;86:924–934. [https://doi.org/10.1002/prot.25524](https://doi.org/10.1002/prot.25524)