A Novel Topology for Representing Protein Folds

Mark R. Segal[*]

[*] University of California, San Francisco, mark@biostat.ucsf.edu

A Novel Topology for Representing Protein Folds

Mark R. Segal

Abstract

Various topologies for representing three dimensional protein structures have been advanced for purposes ranging from prediction of folding rates to ab initio structure prediction. Examples include relative contact order, Delaunay tessellations, and backbone torsion angle distributions. Here we introduce a new topology based on a novel means for operationalizing three dimensional proximities with respect to the underlying chain. The measure involves first interpreting a rank-based representation of the nearest neighbors of each residue as a permutation, then determining how perturbed this permutation is relative to an unfolded chain. We show that the resultant topology provides improved association with folding and unfolding rates determined for a set of two-state proteins under standardized conditions. Furthermore, unlike existing topologies, the proposed geometry exhibits fine scale structure with respect to sequence position along the chain, potentially providing insights into folding initiation and/or nucleation sites.

The protein folding problem, or problems (Dill et al. 2007), despite considerable recent headway, remains one of the greatest challenges facing computational biology. The inter-related folding problems can be described as (Dill et al. 2008): (a) *the folding code*: given a protein's amino acid sequence, how does the thermodynamic interplay of interatomic forces determine the protein's structure? (b) *structure prediction*: how can a protein's (native, 3D) structure be computationally predicted from its amino acid sequence? and (c) *the folding process*: given the vast number of conformational possibilities embodied in a protein's amino acid sequence, how does it fold so quickly to its native state? This last problem, often referred to as the Levinthal (Levinthal 1968) paradox, has been addressed using a spectrum of theoretical and experimental approaches. Some *remarkable* (Grantcharova et al. 2001) and *striking* (Huang et al. 2007) findings to emerge in recent years pertain to the fact that folding rates of two-state proteins (those folding without observable intermediates), which can vary over more than 8 orders of magnitude, from microseconds (Kim and Baldwin 1990) to hours (Kubelka et al. 2004), and include a wide range of folds and functions, are largely determined by the topology of the native structure, with relative insensitivity to features such as the details of inter-atomic interactions and protein length (Plaxco et al. 1998; Shi et al. 2008).

Making such inferences – prediction of folding rates based on protein topology – requires a quantification of topology and a number of derived summaries have been advanced for this purpose. Generally, these summaries are employed as "bulk" properties – aggregated over the protein structure – so as to relate to (overall) folding rate. However, more locally defined topological summaries may prove informative with respect to local attributes such as folding initiation and propagation sites (Dyson et al. 2006).

It has been noted that (tertiary) native structures ought reflect their folding path histories, at least for some folding mechanisms (Dill et al. 1993). This motivates our framing of a novel topological characterization of a folded protein. It is based on the permutation representation of nearest neighbors, with subsequent use of Kendall's tau distance metric to capture perturbation from the unfolded polypeptide chain. We contrast the performance of leading topologies in predicting two-state protein folding and unfolding rates, demonstrating significant prediction gains for our new measure. This performance is all the more notable since it is achieved without implicit or explicit optimization, the new topology being devoid of tuning parameters. Some preliminary exploration of local properties is also proffered.

The importance of topology in terms of prediction of folding rates was first established for relative contact order (RCO) (Plaxco et al. 1998). Let a_i designate the i^{th} residue in a protein primary sequence of length n. A (non-local) contact between two residues a_i, a_j, separated by at least l_{cut} residues along the sequence, is defined as occurring if there are two heavy (non-hydrogen) atoms, one from each residue, within a cutoff distance of R_{cut}. Standard values for l_{cut} and R_{cut} are 2 (sequence positions) and 6 Å, respectively. Assume there are n_c contacting residue pairs. Then RCO is defined as

$$\text{RCO} = \frac{1}{n \cdot n_c} \sum_{(a_i,a_j)}^{n_c} |i - j|$$

where the sum is over all contacting pairs (a_i, a_j).

Several variants of RCO have been proposed, with emphasis on sensitivity to the cutoff parameters, and the *scope* (short-, mid-, or long- range) of contacts, as discussed later. An alternate formulation, termed *effective contact order* (ECO) (Fiebig and Dill 1993; Dill et al. 1993), operationalizes contacts and scope in terms of shortest path lengths between residues that can be achieved in the presence of existing (covalent or topological) links. This formulation attempts to capture effective loop size, and hence the size of the conformational search space necessary to form a conditional (on preexisting links) contact and, as such, is postulated to relate to search (folding) speed (Dill et al. 2008). Our new topology captures such constructs but in a distinct framework.

What constitutes folding from a topological (rather than mechanistic) perspective? Clearly, any definition must be dependent on the underlying polypeptide chain, since purely 3D coordinate based definitions would give rise to a multitude of (irrelevant) "folds". One widely used primitive is based on backbone dihedral angles as depicted in Ramachandran plots (Ramachandran et al. 1963). These plots have been used for crystallographic quality control purposes to detect angular outliers since many angle combinations do not occur due to steric hindrance. Further, by modeling sequential angular dependencies along the chain using dynamic Bayesian networks, successful generative models of local protein structure have been devised (Hamelryck et al. 2006; Boomsma et al. 2008). However, unlike RCO above, topological summaries derived from angular representations have not been employed in relation to predicting attributes such as folding rates.

As an alternative to backbone angles and contact orders, we can conceptualize folding as resulting in some residues being brought closer together *relative* to their positions in a denatured random coil. While related to the underpinnings of contact order, it is by operationalizing this notion, without invoking contact distances, that we arrive at our new topology.

Let $u_i = (x_i, y_i, z_i)$ denote the three dimensional coordinates of the Cα atom of residue a_i. We compute the $n \times n$ matrix of Euclidean distances between all Cα pairs:

$$D = [d_{ij}]; \qquad d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

Then, for each residue (i.e., for each row of D) we map its Euclidean distance to a "nearest neighbor" ranking:

$$R = [r_{i\cdot}]; \qquad r_{i\cdot} = \text{rank}\{d_{ij} : j = 1, \ldots, n\} \tag{1}$$

This enables use of cycle structure to capture topology with respect to an underlying chain, which is not available using Euclidean distances directly. A useful byproduct of such rank based approaches is their relative insensitivity to noise, a known concern with regard experimental (X-ray crystallography or NMR) determination of atomic coordinates (Nigham and Hsu 2008).

As a first step we treat $r_{1\cdot}$ as a *permutation* of $\{1, 2, \ldots, n\}$ and, as such, an element of the symmetric group S_n. This provides access to a wealth of techniques and theory, some of which is germane to folding. Every permutation can be written as a product of disjoint cycles. For example, corresponding to the permutation that takes the red (sequence) ordering to the blue (nearest neighbor) ordering in Figure 1A we have

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 3 & 5 & 4 & 2 & 6 & 9 & 8 & 7 \end{pmatrix} = (1)(2\ 3\ 5)(4)(6)(7\ 9)(8)$$

where $(b_1 b_2 \ldots b_k)$ means $b_1 \rightarrow b_2, b_2 \rightarrow b_3 \ldots b_k \rightarrow b_1$. Intuitively, we expect fold topology to relate to cycle structure. If we (simplistically) regard a highly denatured (unfolded) protein as an unstructured molecule, and focus on the N-terminal residue a_1, the above process gives the identity permutation composed of n 1-cycles. When folded, as depicted in the two dimensional cartoon in Figure 1A, we obtain a cycle structure that captures the loops.

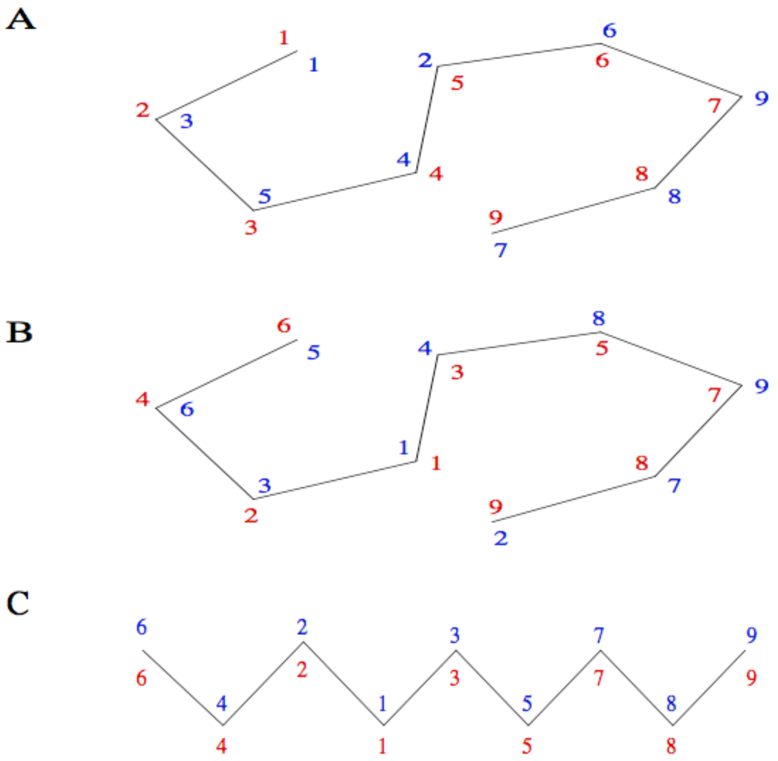

Figure 1: Sequence position (red) and nearest neighbor (blue) orderings for a cartoon fold schematic. A. Sequence numbering proceeds along the chain from position 1 (N-terminal) to 9 (C-terminal). Nearest neighbors, from position 1, are computed using ranked Euclidean distances. B. As for A but with an interior referent position: position 1 (red, original position 4). The red sequence numbering illustrates *referent N-terminal re-ordering.* Nearest neighbor ranks (blue) are computed from this new referent. The (unstandardized) Kendalls tau distance between the referent and nearest neighbor orderings for residues 1 through 9 are 7, 4, 6, 11, 12, 4, 2, 2, 11 respectively. C. A simple schematic mimicking a two-dimensional projection of an α helix. Here the referent and nearest neighbor orderings coincide for each residue, so each Kendalls tau distance is 0.

The second step is to move beyond the extreme N-terminal residue and to define permutations for each residue from the perspective of its position in the chain. Thus, instead of regarding $r_{i\cdot}$ as a permutation of $\{1, \ldots, n\}$ it is treated as a permutation of

$$\begin{pmatrix} \ldots & 4 & 2 & 1 & 3 & 5 & \ldots \\ \ldots & r_{i,i-2} & r_{i,i-1} & r_{i,i} & r_{i,i+1} & r_{i,i+2} & \ldots \end{pmatrix} \tag{2}$$

where the top row in (2), termed the *referent N-terminal re-ordering* and designated ρ_i, represents nearest neighbor ordering for residue a_i in the unfolded state. This schema is illustrated in Figure 1B. Note that $r_{ii} = 1$ by definition (each Cα is closest to itself), so that every cycle representation will contain the 1-cycle (1). Of course, we could equally utilize a C-terminal based re-ordering, as discussed later.

In this manner we obtain a permutation and its attendant cycle representation for each residue. We can then entertain characterizing a folded protein structure using properties or summaries of this collection. For example, we could summarize each residue by maximal cycle length, and then further summarize a structure by the maximum (over all residues) of these maxima. However, this summary proves to be not very useful. At the residue level, maximal cycle length is strongly dependent on whether a referent N-terminal or C-terminal re-ordering is employed, an arbitrariness to avoid. And, on the protein level, we obtain maximal cycle lengths of $\approx n-2$ across a wide range of structures. This is consistent with modal cycle length under *random* permutation.

So, we take a more direct approach to capturing the difference between 3D nearest neighbor and sequence orderings. The referent N-terminal re-ordering is obviously a permutation of $\{1, \ldots, n\}$, and so an element of S_n. Now, a variety of metrics have been defined on S_n (Diaconis 1988). Here, as recommended (Diaconis 1988), we focus on Kendall's tau ($K\tau$) which, for π, σ permutations in S_n, is defined as

$$K\tau(\pi, \sigma) \overset{\text{def}}{=} \text{minimum number of pairwise adjacent transpositions taking } \pi^{-1} \text{ to } \sigma^{-1}$$

the inverses being used to make the metric right invariant. Then, as our third and final step to operationalizing a nearest neighbor : sequence position based topology, we define our Kendall's tau - nearest neighbor (Kτ-NN) summary for residue i, designated Γ_i, as $\Gamma_i = \Gamma_i(\rho_i, r_{i\cdot}) = K\tau(\rho_i, r_{i\cdot})/\binom{n}{2}$ where division by $\binom{n}{2}$ standardizes such that $\Gamma_i \in [0, 1]$. We do not employ optimization in arriving at a bulk summary, but simply take the average over all residues: $\bar{\Gamma} = 1/n \sum_{i=1}^{n} \Gamma_i$.

Importantly, Γ_i is insensitive as to whether an N-terminal or C-terminal referent re-ordering is employed. This is a simple consequence of the triangle inequality: let η_i represent the referent C-terminal re-ordering as given by the top row of (3)

$$\begin{pmatrix} \ldots & 5 & 3 & 1 & 2 & 4 & \ldots \\ \ldots & r_{i,i-2} & r_{i,i-1} & r_{i,i} & r_{i,i+1} & r_{i,i+2} & \ldots \end{pmatrix} \tag{3}$$

Then by a version of the triangle inequality $|\Gamma_i(\rho_i, r_{i\cdot}) - \Gamma_i(\eta_i, r_{i\cdot})| \leq \Gamma_i(\rho_i, \eta_i) = O(n^{-1})$. This agreement is exemplified in Figure 2 which showcases near perfect agreement for N-terminal and C-terminal reorderings and illustrates the dependence on sequence length, the structures possessing 65 and 294 Cα's respectively.

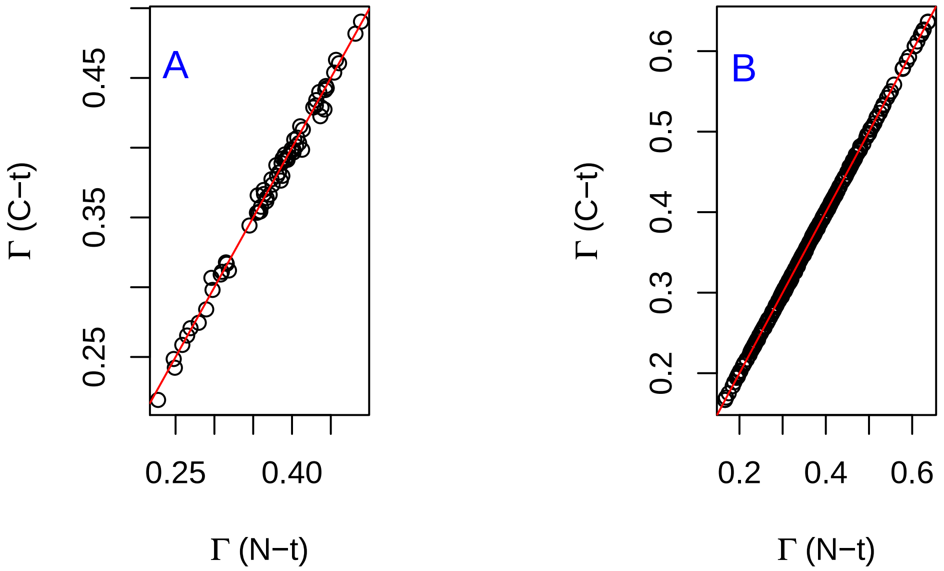

Figure 2: Kτ-NN contrasting N-terminal and C-terminal referent re-orderings: A. PDB 2CI2 which has 65 Cα's; B. PDB 1L8W (Chain B) with 294 Cα's.

Some distinctions between Kτ-NN and RCO and related topologies are worth highlighting. It is important to recognize the complete absence of tuning (parameters) in computing Γ. While refinements involving incorporation of tuning possibilities are discussed subsequently, all results presented herein use the (untuned) formulation described above. In contrast, use of RCO requires specification as to what physical (Euclidean) distance constitutes a contact, and what minimal sequence separation should be imposed. For the former a value of 6 Å is commonly used, this being the original specification (Plaxco et al. 1998), but other choices (e.g., 8 Å: Gromiha and Selvaraj 2001; Bonneau et al. 2002) have been advocated. While some studies indicate that results are insensitive to this specification (Plaxco et al. 1998; Yuan 2005), others suggest that the choice has a strong influence (Mirny and Shakhnovich 2001). Additionally, there are disparate ways of operationalizing contact order *scope*: local, mid, and long range contacts being distinguished (Gromiha and Selvaraj 2001; Zhou and Zhou 2002). Similarly, for example, the geometric distance based on Delaunay tessellation (DT) (Ouyang and Liang 2008) requires specification of sequence and spatial separation parameters.

Kτ-NN topology attempts to capture chain deformation / structural information *between* the referent and contacting residues, whereas this is ignored in computing RCO and DT. Conversely, Kτ-NN also incorporates such information *beyond* contacting residues. It could be argued that inclusion of such remote (from the referent residue) nearest neighbor rankings is at best irrelevant, and at worst distorting, for a topological summary. The following considerations are germane: (i) applications of topologies are typically to *bulk* protein attributes, as opposed to residue specific. Summarization over the entire chain serves to downweight these distant contributions; (ii) a constraint on the range of sequence positions prior to ranking (1) is a putative tuning parameter; and (iii) transforming Γ_i to corresponding *proximities* via $\text{prox}(\Gamma_i) = 1 - \exp(-\Gamma_i)$ can be used to achieve such downweighting (Diaconis et al. 2008) without specifying tuning parameters. Notably, despite these concerns, the performance of Γ_i-based summaries in predicting two-state folding rates exceeds that of alternates as described next.

The dataset used to assess performance of the competing topologies was obtained from a recent compilation (Maxwell et al. 2005). Critically, this paper was the first to derive and assemble folding (k_f), and unfolding (k_u), rate constants obtained under standard experimental conditions, necessary for meaningful comparisons. The data provides rate constants for 30 proteins, 27 of which have PDB identifiers. RCO for these 27 structures was obtained using the Baker lab perl script `http://depts.washington.edu/bakerpg/contact_order/`, as well as an online calculator `http://www.copredictor.ca/`. Calculation of $\bar{\Gamma}$ and DT made recourse to custom R (R Development Core Team 2007) code.

Associations between the respective topologies and log folding rates are presented in Figure 3. Note that the (absolute) correlations attained using $\bar{\Gamma}$ are substantially and significantly greater than those achieved by RCO for both folding and unfolding rates.

Delaunay tessellations have been used in several contexts to capture protein structural attributes, so it is natural to relate correspondingly defined topologies to folding rates. Doing so for both two- and multi- state proteins (Ouyang and Liang 2008) yielded impressive results, with DT (Nα in Ouyang and Liang 2008) outperforming RCO in both settings. However, DT is strongly correlated with chain length. For multi-state proteins length is known to be a significant determinant of k_f (Ivankov et al. 2003). But, for two-state proteins, results generally show no association between folding rates and length (Plaxco et al. 1998; Grantcharova et al. 2001). Indeed, for the present set of two-state proteins, with folding rates determined under standardized conditions, neither length nor DT are well correlated with k_f: absolute correlations being 0.19 and 0.41 respectively, again significantly less than for $\bar{\Gamma}$.

So far, we have utilized the new topology only in terms of a bulk property: its average, $\bar{\Gamma}$, over a given protein. Examination of numerous traces of individual (residue level) Γ_i versus sequence position reveals notable fine structure and variation. To illustrate, we showcase behavior for the two proteins from (Maxwell et al. 2005) with extreme k_f values: PBD IDs 1APS and 1LMB. Figure 4 contrasts profiles of three topological summaries over sequence position for the two proteins. The superposed smooths (1APS in red, 1LMB in green) were obtained using lowess (R Development Core Team 2007). The measures are, respectively, Kτ-NN (Γ_i), residue level relative contact order (RCO$_i$), and average area buried under folding (AABUF), a refinement of hydrophobicity incorporating residue size (Nishimura et al. 2005; Dyson et al. 2006). For both Γ_i and RCO$_i$ we see clear differences in terms of overall level between the two proteins, indicative of their bulk (mean) summaries ability to predict folding rate. No such separation is evidenced for AABUF which, indeed, is not associated with folding rates, at least for the proteins considered here (not shown). Further, AABUF does not

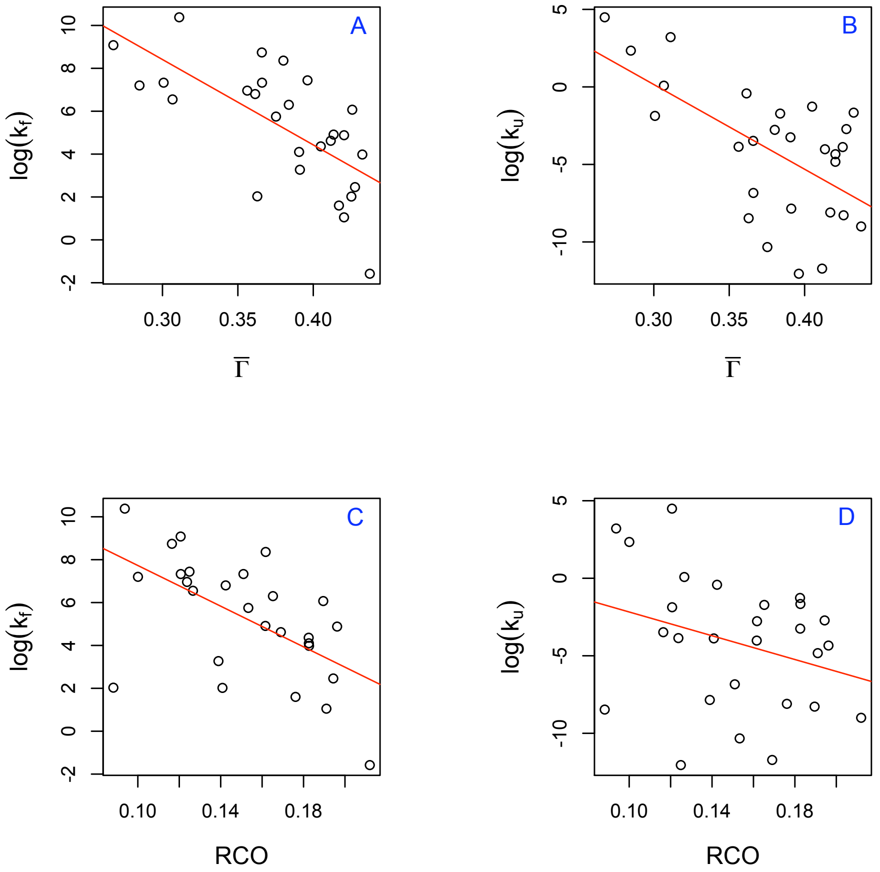

Figure 3: Folding k_f and unfolding k_u rates vs Kτ-NN ($\bar{\Gamma}$) and relative contact order (RCO) topologies. summarized. The respective correlations are A. k_f vs $\bar{\Gamma}$: **-0.68**; B. k_u vs $\bar{\Gamma}$: **-0.61**; C. k_f vs RCO: **-0.58**; D. k_u vs RCO: **-0.30**.

behave smoothly with respect to sequence position. This is in contrast to RCO$_i$ and, to a much greater extent, Γ_i, which exhibits well defined local minima and maxima.

The existence of these well defined Γ_i extremes begs the question as to whether they relate to attributes of the folding process or properties of the three dimensional protein structure. Unfortunately, there is a dearth of data, at the residue level, for making such assessments. Features of interest, but for which insufficient data is available, include *nucleation* as measured by a residue's Φ value (Mirny and Shakhnovich 2001; Larson et al. 2002), which provides a measure of the extent to which the residue participates in native-like interactions during the rate limiting folding step, and folding *initiation* sites. We speculate that residues with small Φ values (not part of the folding nucleus) will have small Γ_i values, the logic being that these small values of Γ_i correspond to residues that are relatively "unperturbed" by the folding process. This pertains for Villin 14T

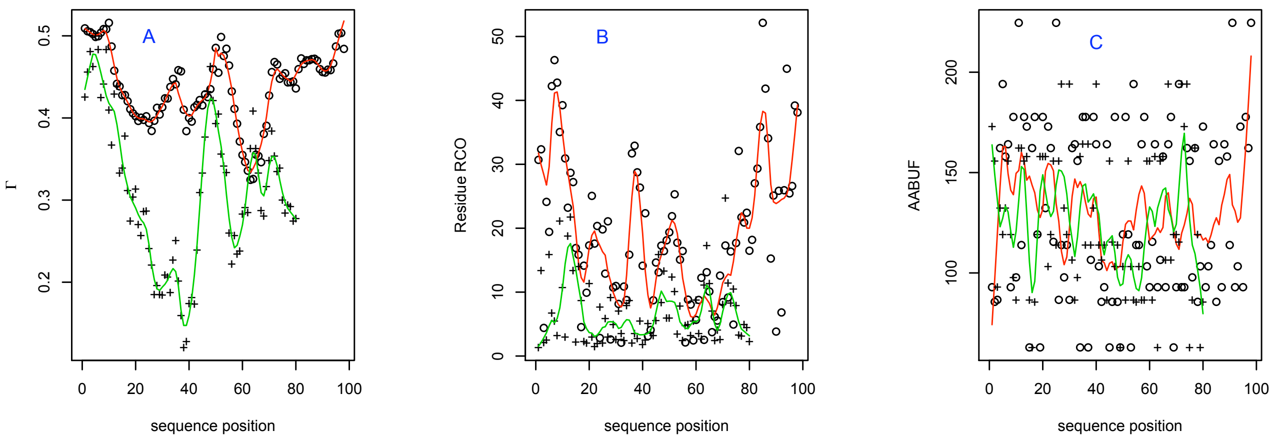

Figure 4: Superposed profiles for the proteins with minimal (1APS: red, ∘) and maximal (1LMB: green, +) k_f values: A. Kτ-NN (Γ_i); B. residue RCO; C. Average Area Buried Under Folding.

(Choe et al. 2000) for which we observe (Figure 5(A)) a significant correlation of 0.66 between Γ_i and Φ values after trimming 4 negative and near zero instances for which Φ is less than its standard error (cf Plaxco et al. 2000). The Γ_i sequence position profile (Figure 5(B)) reveals considerable fine structure. A ribbon diagram of Villin 14T, colored according to Γ_i value (Figure 6), with select extreme Φ and Γ_i residues (3, 7, 43, 84: see Figure 5(A)) highlighted, showcases the positive association. Similar to the toy example (Figure 1) we generally observe lower values Γ_i within helices and strands, and higher values at loop inflections, consistent with zipping and assembly mechanisms (Dill et al. 2008). However, it is important to recognize the highly presumptive nature of these putative associations, in some part attributable to the considerable uncertainties in, and sparsity of, measured Φ values.

Now, focusing exclusively on the protein backbone and disregarding side-chains, the folding process that transforms a highly denatured random coil, even containing residual sequence-local structure (Wang et al. 2007), can be coarsely viewed as a mapping $T : \mathbb{R}^3 \to \mathbb{R}^3$ that is a *contraction*: there is a real $q, 0 \leq q < 1$ such that $d(Tu_i, Tu_j) \leq q \cdot d(u_i, u_j)$ for all residues i, j where $d(\cdot,\cdot)$ is Euclidean distance, and u_i gives the coordinates of the i^{th} Cα atom. Then, from the Banach fixed point theorem (Khamsi and Kirk 2001), we have that T has a unique fixed point u^* such that $Tu^* = u^*$. Now, consequences of this result are moot since T is unknown, as are atomic coordinates in the unfolded state. However, by mapping from $\mathbb{R}^3$ to S_n and invoking the Kτ-NN topology, we can identify the residue closest to the fixed point as $i^* : \Gamma_{i^*} = 0$. So, applying this speculation to 1LMB (green trace, Figure 4A) we can surmise that residue 38 (Γ minima) and neighbors (in view of smoothness of Γ) are (relatively) fixed and, accordingly, are removed from nucleation or initiation sites. Future possibilities include incorporation of such predictions into structure prediction algorithms (Bonneau et al. 2002) as well as developing improved characterizations,

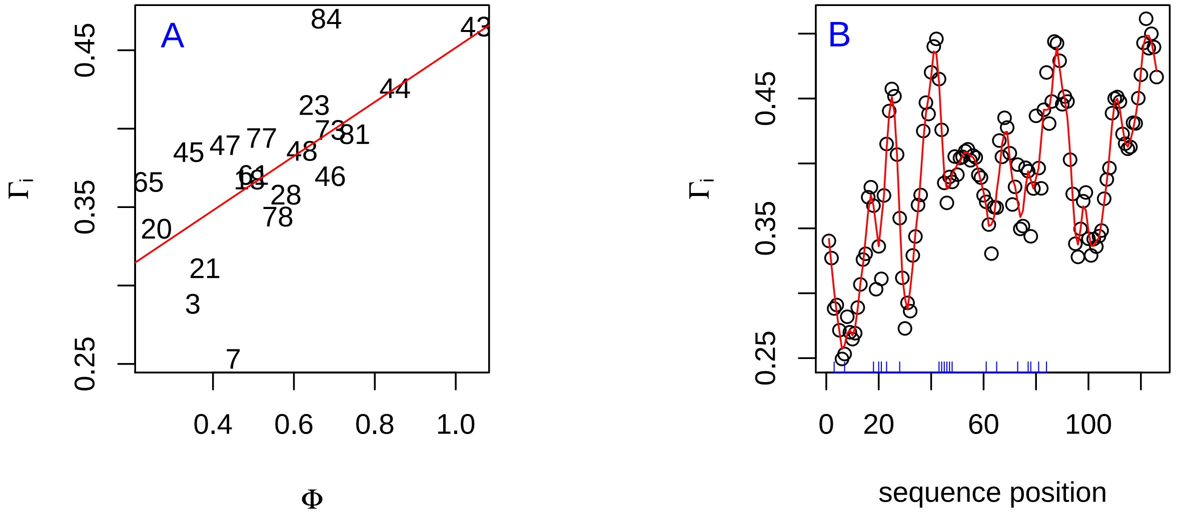

Figure 5: Relationship between Φ and Γ_i values for Villin 14T (PDB ID 2VIK): A. Points for which both Φ and Γ_i values are available (see text) are plotted using their sequence position. Correlation = **0.66**; B. Γ_i profile showing local structure the rug (blue) gives the sequence position for which Φ values were available.

refinements, and applications of the newly proposed topology.

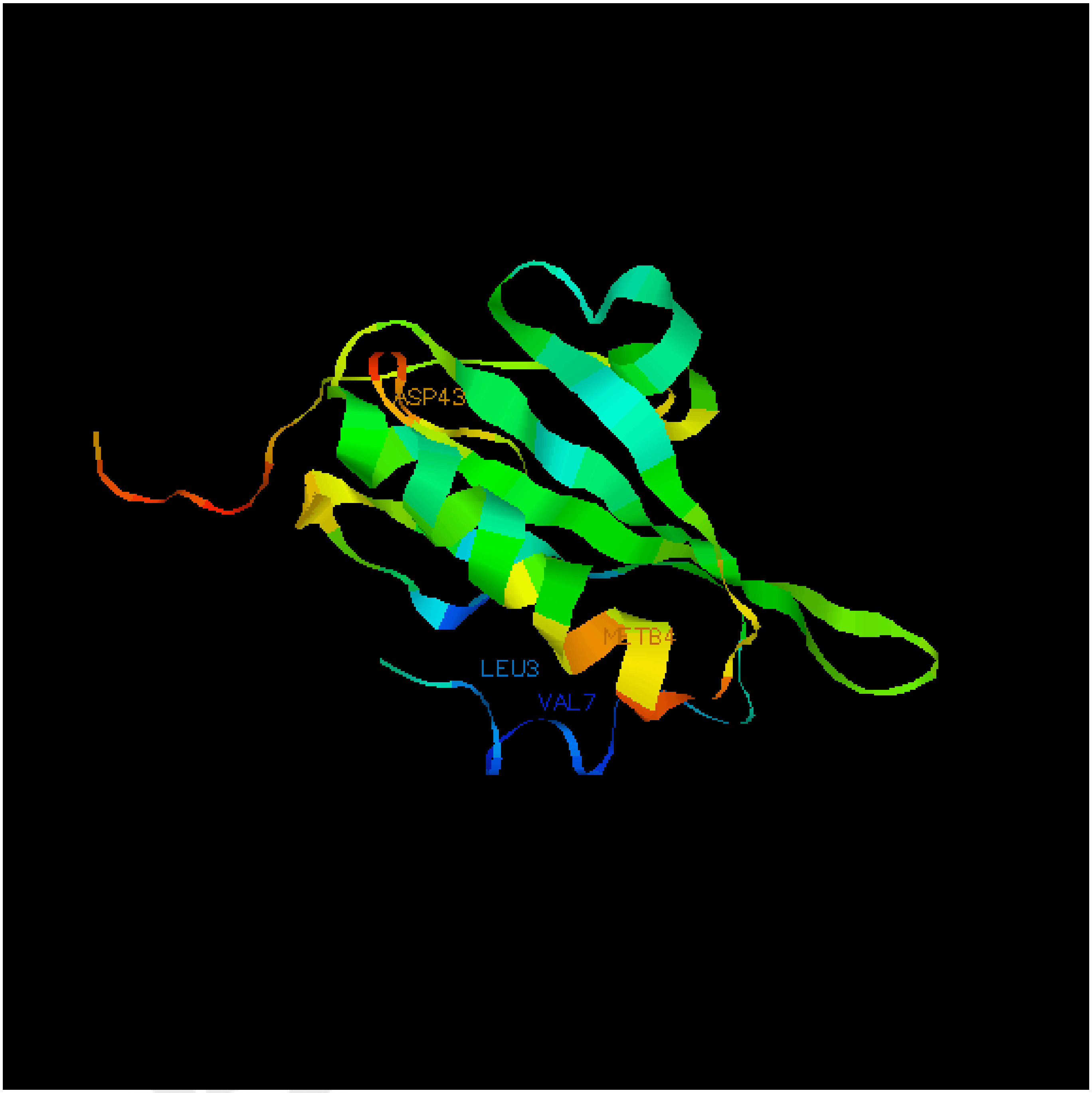

Figure 6: Ribbon diagram colored according to Γ_i value (blue: low; red: high) and highlighting extreme (Φ, Γ_i) sequence positions (3, 7, 43, 84: see Figure 5(A)).

Acknowledgments

I thank Ken Dill, Kevin Plaxco, Ingo Ruczinski, and Yuanyuan Xiao for helpful comments and data provision. Funding support was received from the UCSF Center for Bioinformatics and Molecular Biostatistics.

References

Bonneau R, Ruczinski I, Tsai J, Baker D. (2002). Contact order and ab initio protein structure prediction. *Protein Science*, 11:1937-1944.

Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T. (2008). A generative, probabilistic model of local protein structure. *Proc Natl Acad Sci*, 105(26):8932-7.

Cleveland WS. (1979). Robust locally weighted regression and smoothing scatterplots. *J Amer Statist Assoc*, 74: 829-836.

Diaconis P. (1988). *Group Representations in Probability and Statistics.* Hayward, CA: Institute of Mathematical Statistics.

Diaconis P, Goel S, Holmes S. (2008). Horsehoes in multidimensional scaling and local kernel methods. *Ann Appl Stat*, 2:777-807.

Dill KA, Fiebig KM, Chan HS. (1993). Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci*, 90:1942-1946.

Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. (2007). The protein folding problem: when will it be solved? *Curr Op Struct Biol*, 17:342-346.

Dill KA, Ozkan SB, Shell MS, Weikl TR. (2008). The protein folding problem. *Annu Rev Biophys*, 37:289-316.

Dyson HJ, Wright PE, Scheraga HA. (2006). The role of hydrophobic interactions in itiation and propogation of protein folding. *PNAS*, 103:13057-13061.

Fiebig KM, Dill KA. (1993). Protein core assembly processes. *J Chem Phys*, 98:347587.

Grantcharova V, Alm EJ, Baker D, Horwich AL. (2001). Mechanisms of protein folding. *Curr Op Struct Biol*, 11:70-82.

Gromiha MM, Selvaraj S. (2001). Role of medium–and long-range interactions in discriminating globular and membrane proteins. *Int J Biol Macromol*, 29:25-34.

Hamelryck T, Kent JT, Krogh A. (2006). Sampling realistic protein conformations using local structural bias. *PLoS Comput Biol*, 2:e131.

Huang JT, Cheng JP, Chen H. (2007). Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. *Proteins*, 67:12-7.

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

Khamsi MA, Kirk WA. (2001). *An Introduction to Metric Spaces and Fixed Point Theory.* New York: Wiley.

Kim PS, Baldwin RL. (1990). Intermediates in the folding reactions of small proteins. *Ann Rev Biochem*, 59:631-660.

Kubelka J, Hofrichter J, Eaton WA. (2004). The protein folding “speed limit”. *Curr Opin Struct Biol*, 14:76-88.

Larson SM, Ruczinski I, Davidson AR, Baker D, Plaxco KW. (2002). Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J Mol Biol*, 316:225-233.

Levinthal C. (1968). Are there pathways for protein folding? *J Chem Phys*, 65:44-45.

Maxwell KL, Wildes D, Zarrine-Afsar A et al. (2005). Protein folding: defining a “standard” set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci*, 14:602-16.

Mirny L, Shakhnovich E. (2001). Protein folding theory: From lattice to all-atom models. *Ann Rev Biophys Biomol Struct*, 30:361396.

Nigham A, Hsu D. (2008). Protein conformational flexibility analysis with noisy data. *J Comp Biol*, 15:813-828.

Nishimura C, Lietzow MA, Dyson HJ, Wright PE. (2005). Sequence determinants of a protein folding pathway. *J Mol Biol*, 351:383-392.

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

Plaxco KW, Simons KT, Baker D. (1988). Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, 277:985-994.

Ramachandran GN, Ramakrishnan C, Sasisekharan V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol*, 7:9599.

R Development Core Team. (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

Shi Y, Zhou J, Arndt D, Wishart DS, Lin G. (2008). Protein contact order prediction from primary sequences. *BMC Bioinformatics*, 9:255.

Wang Z, Plaxco KW, Makarov DE. (2007). Influence of local and residual structures on the scaling behavior and dimensions of unfolded proteins. *Biopolymers*, 86:321-328.

Yuan Z. (2005). Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics*, 6:248.

Zhou H, Zhou Y. (2002). Folding rate prediction using total contact distance. *Biophys J*, 82:458-463.