Received: 24 December 2020 | Revised: 22 January 2021 | Accepted: 22 January 2021

DOI: 10.1002/pro.4032

FULL-LENGTH PAPERS

THE PROTEIN SOCIETY WILEY

Local sequence-structure relationships in proteins

Tatjana Škrbić[1,2] | **Amos Maritan**[3] | **Achille Giacometti**[2] | **Jayanth R. Banavar**[1]

[1]Department of Physics and Institute for Fundamental Science, University of Oregon, Eugene, Oregon

[2]Dipartimento di Scienze Molecolari e Nanosistemi, Università Ca' Foscari Venezia, Venezia Mestre, Italy

[3]Dipartimento di Fisica e Astronomia, Università di Padova and INFN, Padova, Italy

Correspondence
Tatjana Škrbić, Department of Physics and Institute for Fundamental Science, University of Oregon, Eugene, OR 97403.
Email: tskrbic@uoregon.edu

Funding information
European Cooperation in Science and Technology, Grant/Award Number: CA17139; Fondazione Cassa di Risparmio di Padova e Rovigo, Grant/Award Number: Excellence Project 2018; H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 894784; Ministero dell'Istruzione, dell'Università e della Ricerca, Grant/Award Number: MIUR PRIN-COFIN2017 Soft Adaptive Networks grant 2017Z55KCW; University of Oregon, Grant/Award Number: Knight Chair

Abstract

We seek to understand the interplay between amino acid sequence and local structure in proteins. Are some amino acids unique in their ability to fit harmoniously into certain local structures? What is the role of sequence in sculpting the putative native state folds from myriad possible conformations? In order to address these questions, we represent the local structure of each C_α atom of a protein by just two angles, θ and μ, and we analyze a set of more than 4,000 protein structures from the PDB. We use a hierarchical clustering scheme to divide the 20 amino acids into six distinct groups based on their similarity to each other in fitting local structural space. We present the results of a detailed analysis of patterns of amino acid specificity in adopting local structural conformations and show that the sequence-structure correlation is not very strong compared with a random assignment of sequence to structure. Yet, our analysis may be useful to determine an effective scoring rubric for quantifying the match of an amino acid to its putative local structure.

KEYWORDS

amino acid groupings, amino acid propensity, local structure, sequence-structure relationship

1 | INTRODUCTION

It is known that there are just a few important principles[1–6] that drive the folding process of a protein: the requirement of avoiding steric overlaps in both the folded and unfolded states, the lower conformational entropy in the folded state than in the unfolded state, the hydrophobic effect favoring a compact conformation that is able to expel water from the core of the folded state and the delicate balance of hydrogen bonds with the solvent and within the protein backbone that can tip the energetic balance between the unfolded and folded state. The fundamental issue is how nature has effectively explored the astronomically large sequence space through evolution to make proteins the molecular target of natural selection.

Here we characterize the native state folds within a simple coarse-grained representation and elucidate the role, if any, played by the repertoire of amino acids in fitting into one of these local geometries. We model a chain by just its C_α atoms and follow the coordinate representation shown in Figure 1. With the knowledge of the preceding C_α locations, we specify the position of a given C_α atom by three coordinates,[7] the bond length, b, and two angles, θ, and μ. θ is the bending angle at the given C_α location, whereas μ is the angle between

successive binormals (Figure 1). The binormal associated with a specific consecutive triplet of C_α atoms is the unit vector perpendicular to the plane of the triplets. The tangent, the normal, and the binormal, all at the middle C_α atom, form a right-handed Cartesian coordinate system. This coordinate system was introduced by Rubin and Richardson in a paper describing the Byron bender that allowed for a simple construction of protein C_α models.[8,9]

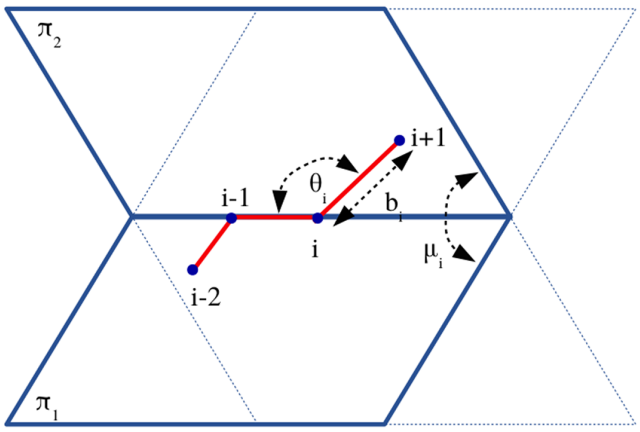

FIGURE 1 Definition of coordinate system. The bond length b at location i, b_i, is the distance between the points i and $(i + 1)$. The angle θ_i is the angle subtended at i by points $(i - 1)$ and $(i + 1)$ along the chain. The third coordinate μ_i is the dihedral angle between the planes π_1 and π_2 formed by $[(i - 2), (i - 1), i]$ and $[(i - 1),i,(i + 1)]$, respectively and is the angle between the binormals at $(i$-1) and i. Knowledge of the coordinates of the previous three points $(i - 2,i - 1,i)$ and the three variables (b_i, θ_i, μ_i) are sufficient to uniquely specify the coordinates of the point $(i + 1)$

Our analysis is carried out with a set of more than 4,000 experimentally determined protein native state structures. Starting from the Top 8,000 set proteins of the Richardson laboratory[10,11] with 70% homology level, we excluded all structures with missing atoms in the protein backbone, yielding a set of 4,416 protein native state structures that we used for our analysis (the same set was used in Reference 7) (see Table S1). We successfully validated our analysis using 478 proteins from the Dunbrack data set,[12] this time with a maximum sequence homology level of 20%. There were 205 proteins in common between the Richardson and Dunbrack sets that we used. We carried out the (θ, μ) analysis for both the Richardson and Dunbrack data sets and obtained virtually identical results with the Dunbrack data being understandably more sparse. We present here the detailed analysis for just the much larger Richardson data set.

2 | RESULTS AND DISCUSSION

A simplification arises because the vast majority of bond lengths are nearly constant (Figure 2). Figure 2a and a blown-up version, Figure 2b, depict histograms of bond lengths with two peaks centered around 3.81 and 2.95 Å. The shorter bonds are associated with a Ramachandran angle ω[1] around 0°[13] (Figure 2c). Because the fraction of short bonds is relatively small (0.3%), our analysis here is carried out with all C_α positions, each characterized by a bond length, the θ and μ angles, and the amino acid identity. An analysis of the amino acids associated with just the short bonds shows the preponderance of glycine in

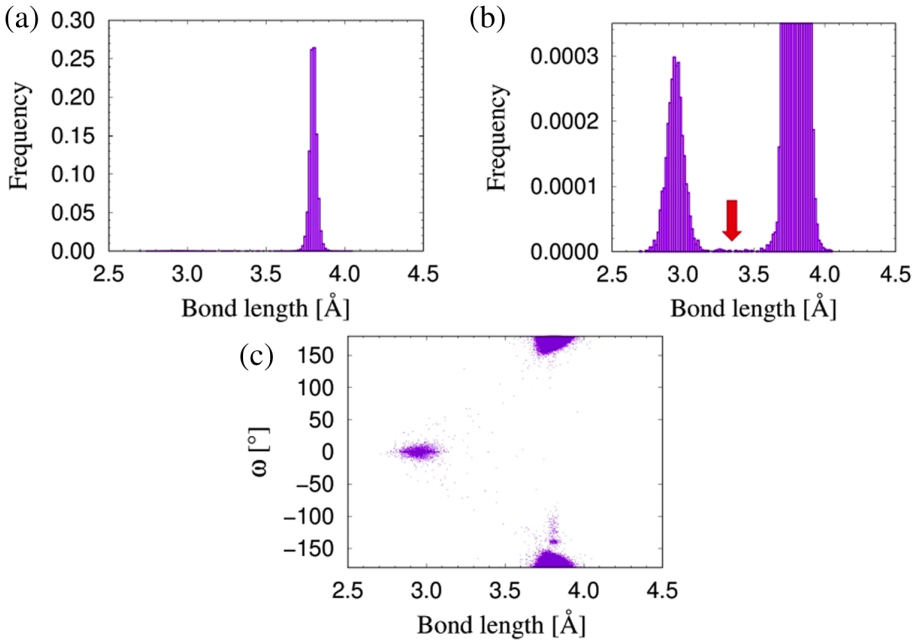

FIGURE 2 Distribution of bond lengths. (a) Shows a histogram of bond lengths in our data set. A blown up version in b shows that the distribution is bimodal with short bonds (centered around 2.95 Å) and long bonds (centered around 3.81 Å). The red arrow is the length we use for partitioning the bonds into the short and long categories. (c) Shows the link between the Ramachandran ω angle (1,13) and the bond length

the first position and proline in the second position (because of the low barrier for transitioning between its *cis* and *trans* conformations).

For a noninteracting phantom chain, one obtains a uniform distribution of points in the (θ, μ) plane (not shown as a figure). As a benchmark, we studied, using Wang-Landau Monte Carlo simulations,[14] a simple self-avoiding polymer chain model composed of 40 unit diameter tangent spheres (tethered hard spheres) subject to a self-attraction between sphere centers located within a distance of 2 units of each other. Figure 3a,b show a cross plot in the (θ, μ) plane of 17 conformations in the coil phase adopted by the chain at high temperatures and for 17 low energy conformations, respectively. The situation is dramatically different for proteins compared with a standard self-avoiding polymer model. Figure 3c is the (θ, μ) cross plot for the protein data set with a highly selective occupancy of (θ, μ) space (a version of this graph was presented earlier in Reference 7).

We binned the data in Figure 3c into squares of width 5° along θ (24 bins in the range 60°–180°) and 5° along μ (72 bins spanning the range from 0° to 360°) to determine the three highest density regions. These density peaks are shown in the figure as black X's along with three larger squares of size 10° × 10° around them. They are identified as helices (the blue region with black X at θ = 92.5° and μ = 47.5°), β-strands (the red region with black X at θ = 122.5° and μ = 192.5°), and loops (the green region with black X at θ = 92.5° and μ = 242.5°) with 184,382, 16,372, and 10,974 points, respectively. The density of points in the α-helix peak is ∼20 times that of loops and β-strands but the loop and β-strand regions are more spread out than the helical region. The other populated regions in the (θ, μ) plane correspond to variants of helices and β-strands and the loops that link them together in the native state structure.

It is important to note that the angles θ and μ are distinct from the Ramachandran[1] angles, which require the knowledge of the locations of backbone atoms besides those of the C_α atoms. The (θ, μ) pair is a coarse-grained representation of the Ramachandran angles and can be useful to describe a generic chain conformation and employed in models of statistical mechanics.[15] In fact, knowing a sequence of Ramachandran angles, one can derive the values of θ and μ. The inverse process of determining the Ramachandran angles from the (θ, μ) values

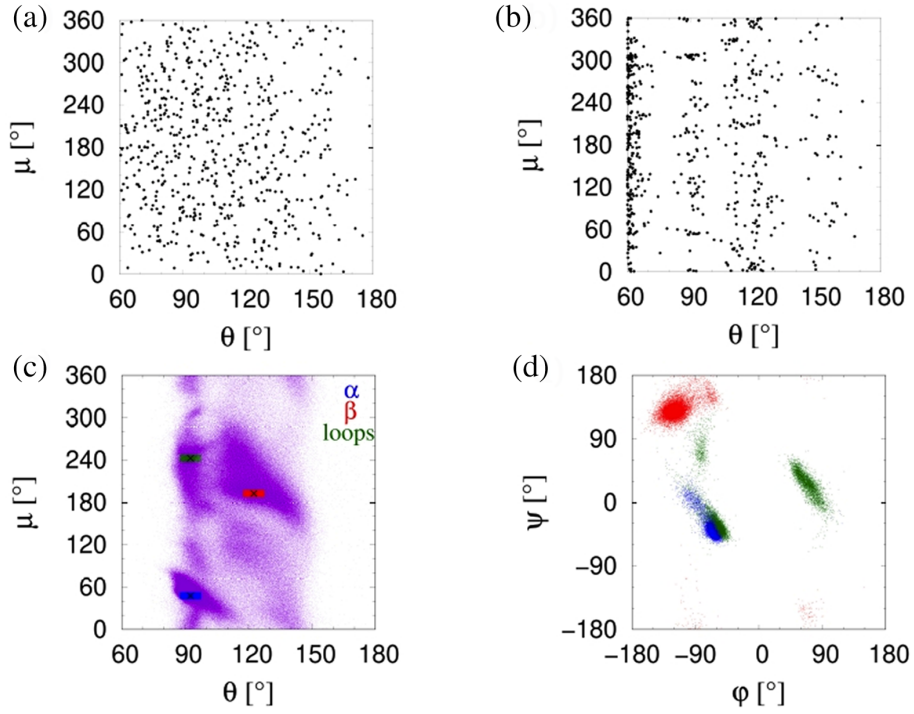

FIGURE 3 Local structure representation. (a) (θ, μ) cross plot for the high temperature coil phase of tethered hard spheres. The only constraint here is the requirement of self-avoidance of the spheres. The points are scattered across the plane with no θ angle less than 60° (a steric constraint) and few almost straight line triplets with a θ near 180°. (b) (θ, μ) cross plot for low energy states of tethered hard spheres. Here again one observes no θ angles below 60° and favored θ angles of 60°, 90°, 120°, and 150° showing that the order favors a face-centered-cubic packing locally, which would be appropriate for the close packing of untethered spheres. (c) (θ, μ) plot for the Richardson data set comprising 4,416 proteins and 972,519 residues (purple points). The three highlighted regions correspond to density peaks related to α-helices (blue region θ = 92.5° and μ = 47.5°), β-strands (red region θ = 122.5° and μ = 192.5°), and loops (green region θ = 92.5° and μ = 242.5°) (d) Plot of the Ramachandran (φ, ψ) angles for the highlighted regions in Figure (c)

do not have a unique solution. For the C_α atoms in the interior of all 4,416 proteins, we measured the (θ, μ) as well as the Ramachandran (φ, ψ) angles. We illustrate the relationship between the two coordinate systems in Figure 3d. We plot the three colored regions (blue, red, and green) of dense points in Figure 3c, but this time expressed as the (φ, ψ) Ramachandran angles color-coded in the same manner as in the (θ, μ) plot. Note that the closely packed points in the (θ, μ) plot are more dispersed in the Ramachandran plot sometimes occupying noncontiguous regions. This is because θ and μ depend on more than one set of Ramachandran angles and the relationship is complicated and nonlinear.

There are four important earlier papers that our work builds on. Rackovsky and Scheraga[16] considered a torsion-curvature plot (distinct from but related to the plot we studied) for 22 protein structures for two different structural groups (helices + bends and extended strands) and the amino acids present therein. Levitt[17] analyzed 13 proteins and considered a (θ, μ) plot similar to ours except that the definition of μ was shifted by one C_α position in the backward direction compared with our definition. Our own definition was motivated by defining θ and μ at a given site i that would determine the coordinates of the $(i + 1)$-th C_α coordinate. Importantly, Levitt determined an approximate empirical relationship between his θ and μ to elucidate approximate potentials for folding simulations.

Oldfield and Hubbard[18] considered two successive θ angles and one μ angle (defined for a bond joining the two C_α atoms) for a set of 83 protein structures and carried out a comprehensive study of local conformational space (but not amino acid preferences) recognizing that the two major building blocks of protein native state structures, helices, and strands, are repetitive conformations. DeWitte and Shakhnovich[19] considered 87 protein structures with a goal of deducing the pairwise potentials, in the spirit of Miyazawa and Jernigan, for the formation of secondary structures in protein simulations based on a cross-plot of two successive μ angles (this time again defined as bond variables rather than at a site) and employing Levitt's empirical relationship. Finally, the approach of Bahar, Kaplan, and Jernigan[20] is most similar to ours. They do have a (θ, μ) plot just like ours except that their μ definition is shifted by one position compared with ours. They used 302 protein structures for their analysis, they carried out an amino acid propensity estimate like we do, and they successfully developed short-ranged (along the sequence) rotational potentials for single amino acids.

In essence, our work here builds on these earlier advances. The principal distinctions are the definition of μ—our μ is defined at a site not at a bond, it is shifted with respect to other definitions, and the number of protein structures we use, many decades after the earliest work, is understandably larger and comprises over 4,000 experimentally determined and curated protein structures. Our goal in this paper is not to extract effective potentials but rather analyze, more generally, sequence-local structure relationships. Furthermore, we seek to group the 20 amino acids into distinct groups in terms of their similarity to substitute for each other in local conformational space.

Figure 4 shows histograms of θ and μ values and evidence for a clear correlation between the average values of θ and the average value of μ among all proteins.

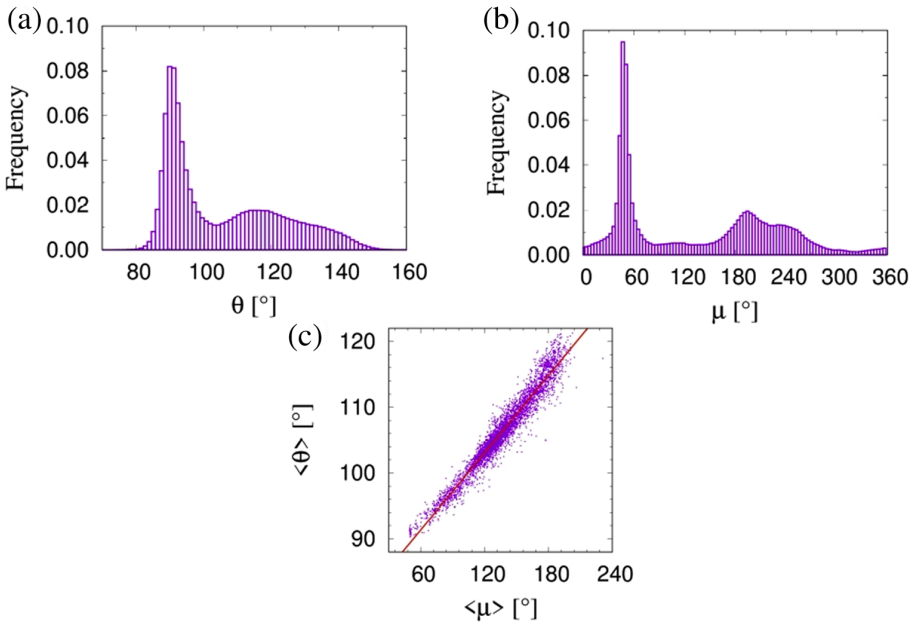

FIGURE 4 (a) and (b) Histograms of θ and μ values showing a multi-peaked structure. (c) A plot of the average value of θ versus the average value of μ for all 4,416 proteins showing a tight correlation with a Pearson correlation coefficient of .97. This may be readily understood by noting that a protein structure is primarily composed of helices and sheets with varying fractions depending on the protein being considered. The θ-μ values for an α-helix are both smaller than those of a β-strand leading to the correlation. Note that the standard deviations (not shown) are large because of the relatively large width in angle space of the regions

TABLE 1 Frequency of 20 amino acids in the set of 4,416 proteins (second column) and a measure of the localization of each amino acid in (θ, μ) space (third column)

Amino acid type	Fraction (%)	Inverse participation ratio (IPR)
ALA	8.53	3.28
ARG	4.84	3.24
ASN	4.42	4.53
ASP	5.96	4.60
CYS	1.36	3.69
GLU	6.48	3.25
GLN	3.61	3.30
GLY	7.90	11.61
HIS	2.32	4.31
ILE	5.62	2.93
LEU	8.79	2.70
LYS	5.70	3.43
MET	2.02	2.95
PHE	4.04	4.06
PRO	4.59	83.28
SER	5.88	5.14
THR	5.58	4.75
TRP	1.52	3.99
TYR	3.61	4.25
VAL	7.23	3.77

Table 1 presents data on the amino acid occurrence probability and the degree of localization in (θ, μ) space. For each amino acid, we measured the inverse participation ratio (IPR) defined as

$$\mathrm{IPR} = \frac{\left(\sum_{i=1}^{N} x_i^2\right)^2}{\sum_{i=1}^{N} x_i^4} \quad (1)$$

where x_i denotes the normalized density of occupancy of the i-th bin in (θ, μ) space and the total number of bins $N = 1728$. An IPR value of 1 indicates perfect localization in just one bin whereas the largest possible value of the IPR is $N = 1728$ for a uniform occupancy of all 1728 bins. A perfect localization (IPR = 1) is indicative of an amino acid that is always associated with the same local structure leading to a perfect sequence-structure relationship. The most localized amino acid is LEU (IPR = 2.70) while the least localized is PRO (IPR = 83.28). Figure 5 shows the occupancies of the (θ, μ) space of amino acids LEU and PRO. Interestingly, even the most localized amino acid, while being largely concentrated in just a few squares, is yet spread out over many squares indicating that there is no strong selection of local structure by amino acid identity.

We carried out an analysis of triplet amino acids identities of all the 324 tight bends with θ angles less than 80°. The smallest θ angle in the data set has a value of 59.98° and the corresponding amino acid triplet is GLY-GLN-ASP. These tight turns $(i-1,i,i+1)$ have no selectivity

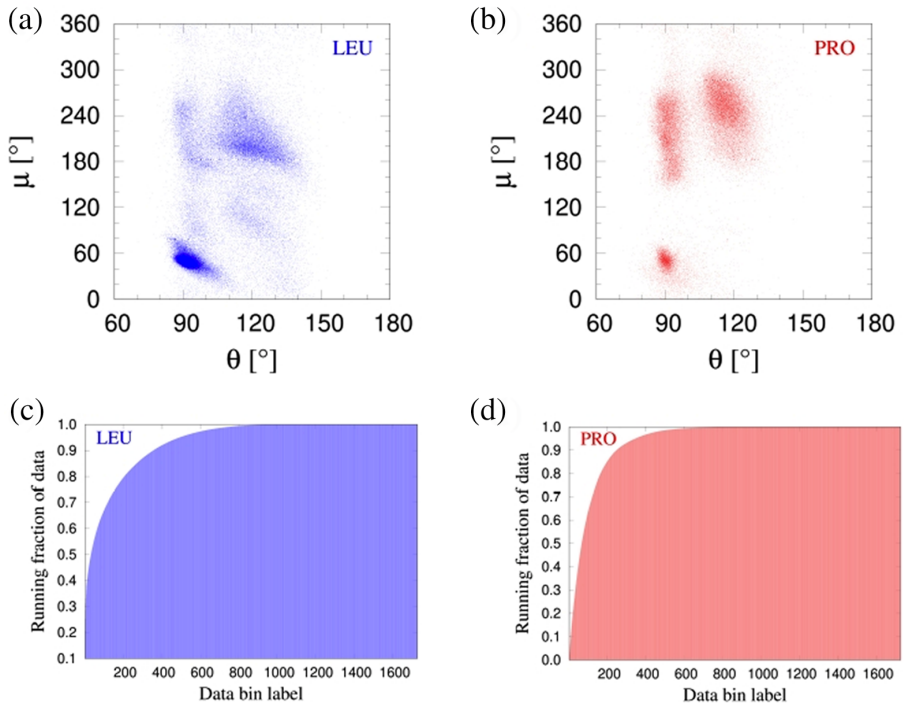

FIGURE 5 Occupancy pattern of amino acids LEU and PRO in (θ, μ) space. (a) and (b) depict the locations of the two amino acids. LEU is the most localized amino acid (IPR = 2.70) whereas PRO has the largest IPR = 83.28 value among the amino acids and is spread out the most. A rank ordered normalized occupancy fraction of the two amino acids is shown in (c) and (d). The number of bins needed to account for 50% and 90% occupancy for the two amino acids are LEU—33 and 356, and PRO—66 and 248, respectively

in μ angles. However, there is indeed a sequence-structure relationship with (GLY or SER) accounting for a total of 34% occupancy in the $i-1$ position, (PRO or SER) having 31% residency in site i, and (ALA or SER) accounting for 21% in the site $(i+1)$.

We studied histograms of the θ and μ values associated with each of the 20 amino acids. The distributions are roughly equally wide and substantially independent of amino acid identity (Figures 6, 7).

Unlike the α-helix region associated with tight local packing and hence a relatively small variation in the θ angle, there is a range of θ values associated with the β-strand region. We carried out sequence analyses of the β-strands to understand whether there is an amino acid selection principle for θ. We selected the (θ, μ) subspace consisting of μ values in the range from 175° to 185° (±5° degree interval around the ideal value of 180°) and of θ angles in the range from 105° to 145°. We divided up the relevant range of θ angles into 40 bins of width 1°. Again, we measure the IPR defined in Equation (1) with $N=40$ in this case. The extreme values of the IPR are 16.08 for the most localized amino acid, PRO, and 31.46 for the most spread out amino acid, ASP (Figure 8). The average θ value and its standard deviation for all amino acids in the β-region is 128.0° and 9.5°, respectively.

We also studied the identities of the 210 pairs of amino acids (and their associated side-chain sizes) located at sites $i-1$ and $i+1$ (these side chains stick out in roughly the same direction with a possibility of steric clashes) flanking site i in the β-region. We considered only those statistically significant pairs $(i-1,i+1)$ which occurred at least 162 times (estimated as the total number of pairs divided by 210) with beads $i-1$, i, $i+1$ all lying in the β-strand region and divided the θ range again into 40 equally spaced bins. The number of amino acid pairs that met the 162 thresholds was 52 of the 210 pairs. We find that all pairs are spread out in θ values. The most localized pair among these was ALA-THR with an IPR of 10.51 and the most spread out pair was PHE-PRO with an IPR of 22.77 (see Figure 9 for histograms of θ values associated with these pairs). A cross plot of the mean van der Waals diameter of a pair and its average θ value (not shown) results in a weak correlation and an overall negative trend. All these results indicate that the sequence does not play a significant role in determining the θ angle associated with a β-strand.

We carried out simple sequence analyses of the loop region as well, to understand whether there is a selection principle for the value of the μ angle. We select the (θ, μ)

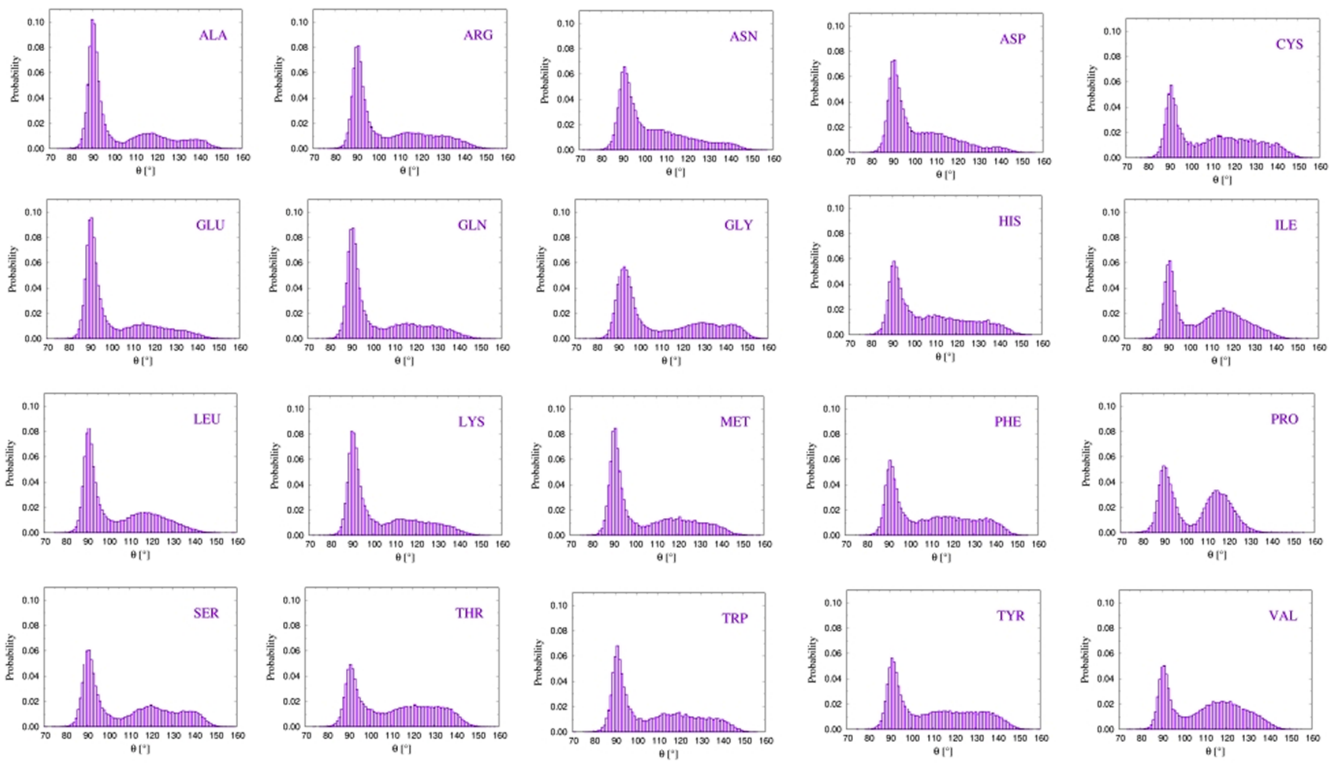

FIGURE 6 Histograms of the θ values for each of the 20 amino acids. While the shapes of the histograms vary from amino acid to amino acid, the ranges are mostly independent of amino acid identity. PRO is a bit of an outlier with a somewhat lower upper cut-off value of θ

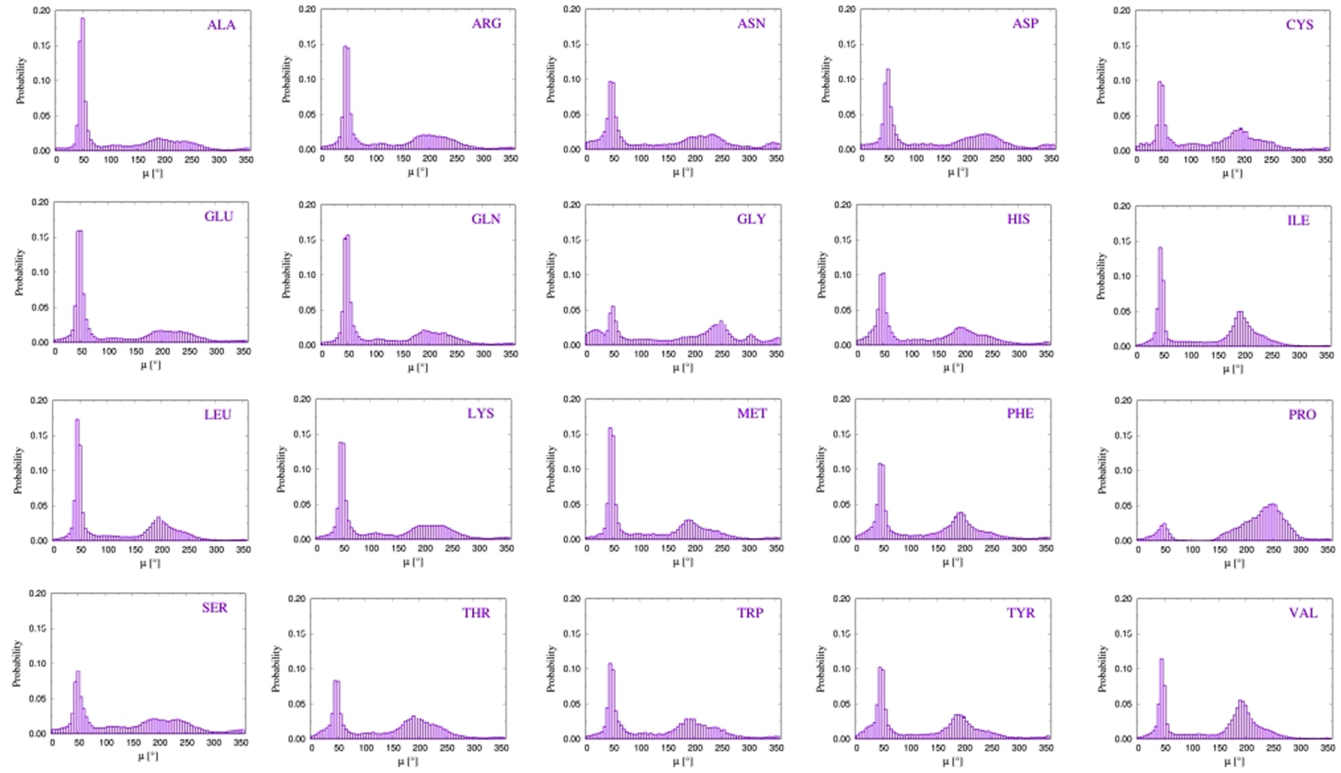

FIGURE 7 Histograms of the μ values for each of the 20 amino acids. Even though the shapes of the histograms vary from amino acid to amino acid, the ranges are mostly independent of amino acid identity

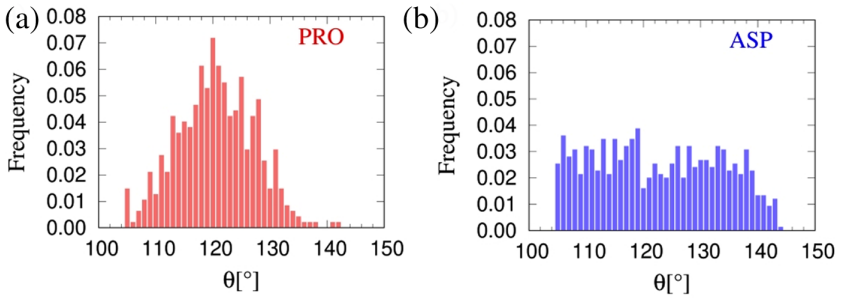

FIGURE 8 Distribution of θ angles in the β-region for PRO (a) and ASP (b). PRO is the most localized amino acid, yet exhibits some spread of θ angles

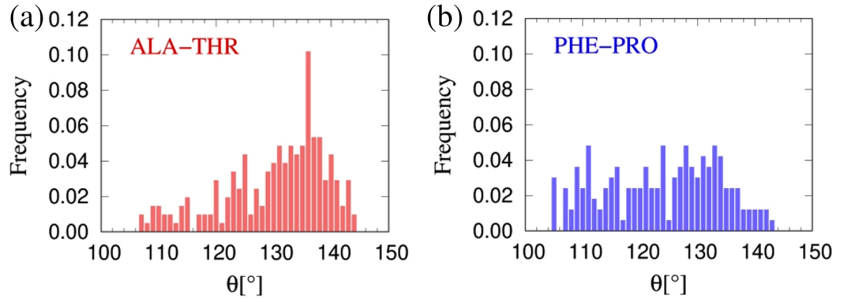

FIGURE 9 Distribution of θ angles in the β-region for (ALA-THR) and (PHE-PRO) amino acid pairs in positions $(i-1, i+1)$ respectively. ALA-THR is the most localized pair in θ space, yet is spread out. PHE-PRO is the most spread out pair

subspace consisting of θ angles in the range from 87.5° to 97.5° (±5° interval around the value 92.5°, identified as the peak density green region in Figure 3c) and μ values in the range from 90° to 360° to ensure that there is no overlap with the α-helix region. We divided up the range of μ angles into 54 bins of width 5°. We measured the IPR value for the 20 amino acids and we find that the most localized amino acid is GLY with a value of 8.49, whereas the most delocalized amino acid is PHE with an IPR equal to 28.42 (Figure 10). Note that μ = 180° and

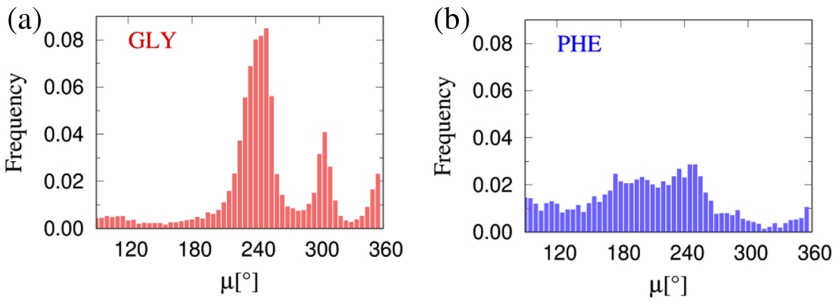

FIGURE 10 Distribution of μ angles in the loop region for GLY and PHE. GLY is the most localized amino acid, yet exhibits a spread of angles

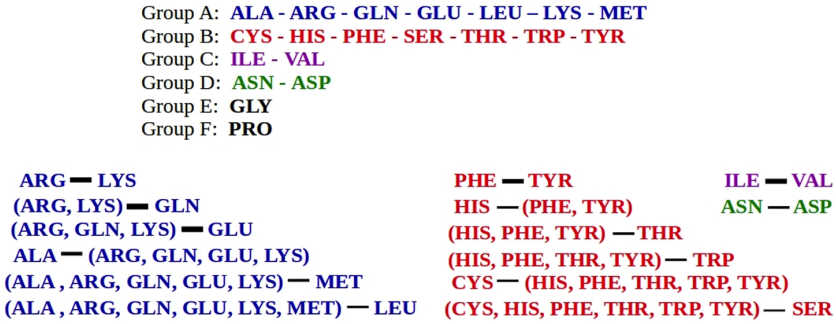

FIGURE 11 Clustering of amino acids into groups. The six amino acid groups obtained based on their similarity in occupying the local structural (θ, μ) space are shown. Six is a natural choice because the closeness for the next collapse into five groups is approximately twice as large as the previous closeness measure. A five member group would result in the merger of the two largest groups, Group A and Group B. If one were to retain seven groups, SER would detach from Group B and remain isolated as its own group. The sequences of hierarchical clustering for the first four Groups A (blue), B (red), C (purple), and D (green) is shown with the link thickness quantitatively representing the closeness measure

360° correspond to planar configurations of four consecutive C_α atoms, with the former corresponding to zigzagging and the latter to rotation in the same sense.

Based on the normalized density of occupancy of the amino acids in (θ, μ) space, one can assess the mutual similarity of the 20 amino acids by measuring the Cartesian distance between the 190 pairs of amino acids, which serves as a proxy of similarity. We have employed the Bhattacharyya coefficient[21] in order to calculate the degree of closeness of the (θ, μ) distributions of amino acids. We carried out hierarchical clustering by rank-ordering the closeness—the two closest amino acids were placed into a single group thereby now having effectively 19 groups of amino acids. This procedure was repeated recursively to reduce the effective groups of amino acids by one each time. A natural stopping point for this hierarchical clustering is when there is a relatively large jump in the measure of closeness of the remaining groups. The result of this analysis is shown in Figure 11 and yields six different groups comprising 7, 7, 2, 2, 1, and 1 amino acids each. Figure 12 shows the occupancy in (θ, μ) space of the six amino acid groups.

We alert the reader that this grouping is distinct from the more familiar groupings of amino acids based on their nonlocal interactions.[22–29] Here, instead, it is entirely based on the similarity of their propensity to adopt specific local conformations.

We defined three significantly occupied regions of (θ, μ) space corresponding to α-helix ($\theta \in [90°,95°]$, $\mu \in [45°,50°]$), β-strand ($\theta \in [105°,145°]$, $\mu \in [175°,185°]$), and loop ($\theta \in [87.5°,97.5°]$, $\mu \in [90°,360°]$). The amino acid occupancies of the three regions are normalized by their frequencies in the entire (θ, μ) space of all 4,416 proteins and they are shown in Table 2. Amino acids having a normalized occupancy greater than 1 are over-represented in a given region and vice versa compared with the expectation from random considerations. The over-represented amino acids in the α-helix region (second column of Table 2) are all members of Groups A and C of amino acids with the top four being LEU (1.56), MET (1.46), and ALA/GLU both having 1.42 normalized occupancy. The amino acids over-represented in the β-strand region (third column of Table 2) are all members of amino acid Groups B and C, the top three being VAL (1.93), ILE (1.55), and TYR (1.51). Finally, the most over-represented amino acids in the loop region correspond to those that are the most under-represented in both the α-helix and β-strand regions: PRO (2.49), GLY (1.76), ASP

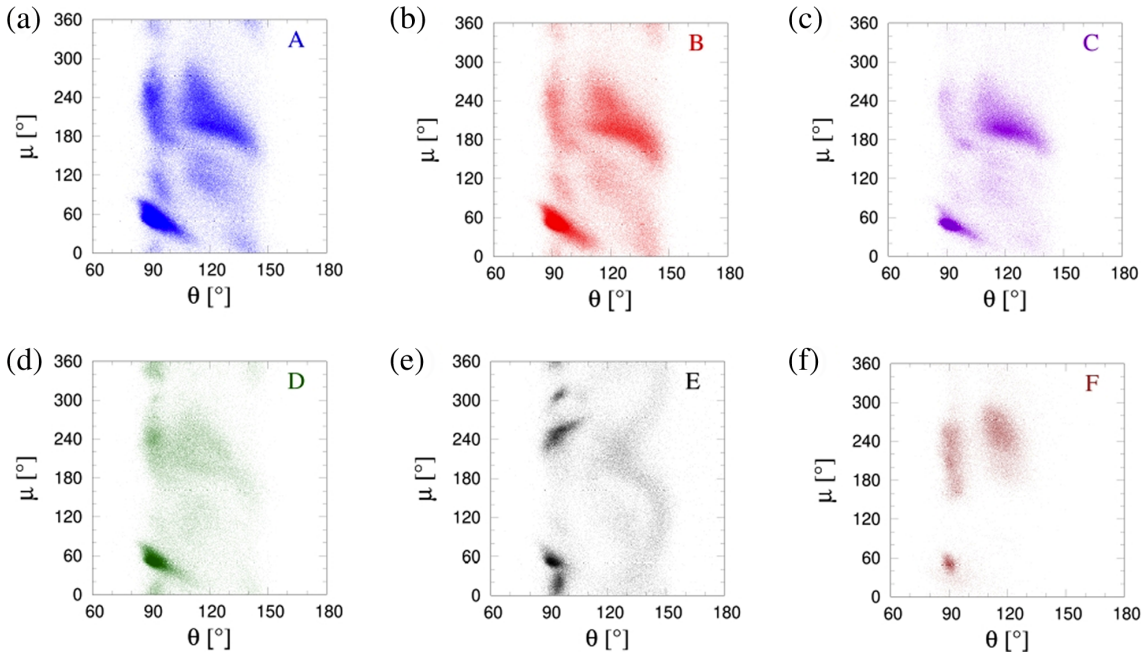

FIGURE 12 Occupancy of the six amino acid groups in (θ, μ) space. Groups A and B are somewhat similar with the main difference being the relative weights of the α-helix and β-strand regions. The most distinctive groups are E and F corresponding to GLY and PRO respectively. We remind the reader (see Figure 3c) that the density peaks occur at (θ = 92.5° and μ = 47.5°) for α-helices, (θ = 122.5° and μ = 192.5°) for β-strands, and (θ = 92.5° and μ = 242.5°) for loops

TABLE 2 Propensity of the 20 amino acids to occupy the α-helix, β-strand, and loop regions in (θ, μ) space

Amino acid type	Normalized occupancy in the α-helix region	Normalized occupancy in the β-strand region	Normalized occupancy in the loop region
ALA	1.42	0.85	0.89
ARG	1.29	1.02	0.89
ASN	0.77	0.60	1.31
ASP	0.77	0.48	1.33
CYS	0.87	1.41	0.67
GLU	1.42	0.63	0.97
GLN	1.38	0.78	0.88
GLY	0.34	0.60	1.76
HIS	0.80	0.98	0.81
ILE	1.26	1.55	0.51
LEU	1.56	0.94	0.70
LYS	1.21	0.75	1.08
MET	1.46	1.21	0.69
PHE	0.88	1.38	0.64
PRO	0.14	0.40	2.49
SER	0.61	1.05	1.04
THR	0.67	1.43	0.76
TRP	0.89	1.23	0.89
TYR	0.80	1.51	0.63
VAL	1.00	1.93	0.50

Note: The numbers shown have been normalized by the amino acid occurrences in all the (θ, μ) space.

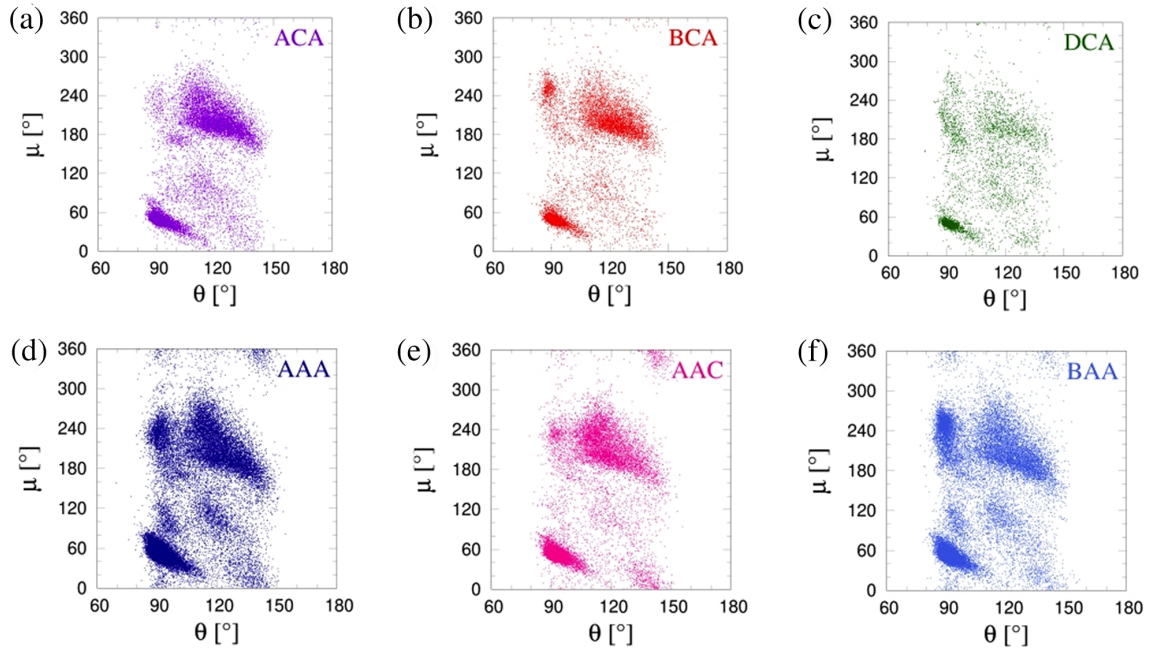

FIGURE 13 The six panels show the distributions of the six most localized triplets in the (θ, μ) plane. They all occupy the α-helix region predominantly. But they are spread out considerably underscoring the weak role of the amino acid sequence in matching with the local structure. We remind the reader (see Figure 3c) that the density peaks occur at (θ = 92.5° and μ = 47.5°) for α-helices, (θ = 122.5° and μ = 192.5°) for β-strands, and (θ = 92.5° and μ = 242.5°) for loops

TABLE 3 Identities of three amino acids with the highest propensities to occupy the α-helix, β-strand, and loop regions in (θ, μ) space (taken from Table 2)

α-Helix propensity		β-Sheet propensity		Loop propensity	
Our study	**Levitt[38]**	**Our study**	**Levitt[38]**	**Our study**	**Levitt[38]**
LEU (1.56)	MET (1.47)	VAL (1.93)	VAL (1.49)	PRO (2.49)	PRO (1.91)
MET (1.46)	GLU (1.44)	ILE (1.55)	ILE (1.45)	GLY (1.76)	GLY (1.64)
GLU (1.42)	LEU (1.30)	TYR (1.51)	PHE (1.32)	ASP (1.33)	ASP (1.41)

Note: The Table also shows the winning amino acids from Levitt's analysis of 1978.[38] There is excellent accord between our results and those of Levitt. The key difference is the identity of one of the top three amino acids in the β-sheet propensity group. PHE scores third in Levitt's analysis with a normalized probability of 1.32 whereas PHE scores fifth in our analysis with a similar probability score of 1.38. TYR scores third in our study and fourth in Levitt's analysis.

(1.33), and ASN (1.31). These four amino acids are members of the amino acid groups D (ASN and ASP), E (GLY), and F (PRO)—see amino acid grouping analysis and Figure 11. The strong correlation observed between the values of normalized occupancies of amino acids in the three regions and the results of the amino acid groupings suggest that amino acid Group A can be interpreted as the "α-helical" group, amino acid Group B as the "β-strand" group, while Group C is over-represented in both α-helix and β-strand regions. Finally, amino acid Groups D, E, and F can be described as "loop" groups, since they are strongly over-represented in loops and under-represented in both α-helix and β-strand regions. These findings are in good accord with the observed amino acid propensities in proteins previously reported in the literature.[3,5,30–33]

With the identification of just six groups, we proceeded to an analysis of correlating the local structure (θ, μ) at bead i to the identity of the triplet of amino acid groups at positions $(i-1,i,i+1)$. The simplicity now is that the total number of distinct triplets is 216 instead of 8,000. We considered each of these triplets and studied the number of times these occurred. Obviously, one would expect that triplets containing the amino acids in groups C, D, E, and F would be fewer than those occurring in Groups A and B. Indeed, the number of triplets which occurred more than 4,461 times (deduced by dividing the total number of triplets = 963,681 and the total number of types of triplets = 216) was just 57 and we used these for our analysis because of their statistical significance. The results are summarized in Figure 13.

3 | CONCLUSION

We conclude with the lessons learned from our analysis. Our goal here was to characterize the local structures associated with protein native state folds using the simple representation of just two angles (θ and μ) for each C_α position (Figure 1). This simplification is made possible because the vast majority of bond lengths is substantially constant (Figure 2). The (θ, μ) variables are a coarse-grained representation of successive Ramachandran angles. The local structures adopted by proteins are captured by simple patterns of points in the (θ, μ) plane. This reveals that protein native state structures (even at the local level) are highly structured unlike the behavior of a generic chain. Even though there is a great deal of spread in the θ and μ values, there is a tight correlation in the plot of the mean θ versus mean μ for the 4,416 proteins (Figure 4).

Armed with insights on the local structural pattern, we explored a potential sequence-structure relationship in multiple ways. We considered the propensity of the 20 amino acids to occupy certain regions of local structural space. We also divided the 20 amino acids into six groups based on their similarity to each other in being associated with regions in the (θ, μ) space. We explored singlets and triplets based on grouping. The basic result of our analysis is that any sequence-local structure relationship is not very strong and there is flexibility in the ability of the amino acids to adapt to the local structure. This is consistent with the prevalence of neutral evolution where neither the native state fold nor the ability to function changes under many amino acid substitutions. It serves to underscore the pioneering results of Brian Matthews[34,35] and his team who "used the lysozyme from bacteriophage T4 to define the contributions that different types of interaction make to the stability of proteins." One of their key findings was that "the protein is, in general, very tolerant of amino acid replacement." Our findings also are in accord with more recent experimental studies on proteins[36,37] which showed that, while protein structures are highly tolerant of amino acid substitutions, a few key alterations can yield distinct structure and function. An interesting challenge is to be able to predict, in a transparent and reliable manner, the identity of these key amino acids.

We conclude by revisiting a seminal paper by Levitt[38] more than four decades ago in which he very carefully measured the Chou-Fasman propensity[39] of the 20 amino acids to be housed in three secondary structures. He noted that, generally, the preferences of the individual amino acids for secondary structure are rather weak. He provided a physical interpretation of his results by noting that "the chemical structure and sterochemistry of the amino acid plays a major part in determining its preference and dislike for secondary structure... Bulky amino acids, namely, those that are branched at the β-carbon or have a large aromatic side chain, prefer β-sheet. The shorter polar side chains prefer reverse turns, as do Gly and Pro, the special side chains. All other side chains prefer α-helix, except Arg which has no preference." Table 3 shows a side-by-side comparison of the results of Levitt obtained with less than a 100 protein structures and our findings with entirely different methods and more than 4,000 protein structures. Our results match those of Levitt[38] confirming the adage—*old is gold*.

4 | MATERIALS AND METHODS

The PDB codes of 4,416 proteins are presented in Table S1 as Supporting Information.

ACKNOWLEDGMENTS

We are indebted to George Rose for his collaboration and inspiration. We are grateful to Pete von Hippel for his warm hospitality and to him and Brian Matthews for stimulating conversations. We are very thankful to two anonymous reviewers for their constructive suggestions.

AUTHOR CONTRIBUTIONS

Tatjana Skrbic: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; validation; visualization; writing-review & editing. **Amos Maritan:** Funding acquisition; writing-review & editing. **Achille Giacometti:** Funding acquisition; writing-review & editing. **Jayanth R. Banavar:** Conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; writing-original draft; writing-review & editing.

CONFLICT OF INTEREST

The authors declare they have no conflict of interest.

ORCID

Tatjana Škrbić https://orcid.org/0000-0002-8947-8216
Amos Maritan https://orcid.org/0000-0002-3535-7873
Achille Giacometti https://orcid.org/0000-0002-1245-9842
Jayanth R. Banavar https://orcid.org/0000-0002-9752-6871

REFERENCES

1. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem. 1968;23:283–438.

2. Fitzkee NC, Rose GD. Steric restrictions in protein folding: An α-helix cannot be followed by a contiguous β-strand. Protein Sci. 2004;13:633–639.
3. Lesk AM. Introduction to protein science: Architecture, function and genomics, Oxford: Oxford University Press, 2004.
4. Rose GD, Fleming PJ, Banavar JR, Maritan A. A backbone-based theory of protein folding. Proc Natl Acad Sci U S A. 2006;103:16623–16633.
5. Bahar I, Jernigan RL, Dill KA. Protein actions, New York: Garland Science, Taylor & Francis Group, 2017.
6. Rose GD. Protein Folding – Seeing is Deceiving, preprint.
7. Škrbić T, Maritan A, Giacometti A, Rose GD, Banavar JR. Building blocks of protein structures – Physics meets biology. *bioRxiv* doi: https://doi.org/10.1101/2020.11.10.375105
8. Rubin B, Richardson JS. The simple construction of protein alpha-carbon models. Biopolymers. 1972;11:2381–2385.
9. https://proteopedia.org/wiki/index.php/Byron%27s_Bender
10. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins. 2000;40:389–408.
11. http://kinemage.biochem.duke.edu/databases/top8000.php.
12. Wang G, Dunbrack RL Jr. PISCES: A protein sequence culling server. Bioinformatics. 2003;19:1589–1591.
13. Matthews BW. How planar are peptide bonds? Protein Sci. 2016;25:776–777.
14. Wang F, Landau D. Efficient, multiple-range random walk algorithm to calculate the density of states. Phys Rev Lett. 2001;86:2050–2053.
15. Flory PJ. Statistical mechanics of chain molecules. New York: Wiley & Sons, 1969.
16. Rackovsky S, Scheraga HA. Differential geometry and polymer conformation. Macromolecules. 1981;14:1259–1269.
17. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol. 1976;104:59–107.
18. Oldfield TJ, Hubbard RE. Analysis of C_α geometry in protein structures. Proteins. 1994;18:324–337.
19. DeWitte RS, Shakhnovich EI. Pseudodihedrals: Simplified protein backbone representation with knowledge-based energy. Protein Sci. 1994;3:1570–1581.
20. Bahar I, Kaplan M, Jernigan RL. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. Proteins. 1997;26:292–308.
21. Bhattacharyya A. On a measure of divergence between two statistical populations defined by their probability distributions. Bull Calcutta Math Soc. 1943;35:99–109.
22. Regan L, DeGrado WE. Characterization of a helical protein designed from first principles. Science. 1988;241:976–978.
23. Miyazawa S, Jernigan RL. Residue-residue potentials with favorable contact pair-term and an unfavorable high packing density term, for simulation and threading. J Mol Biol. 1996;256:623–644.
24. Riddle DS, Santiago JV, Bray-Hall ST, et al. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol. 1997;4:805–809.
25. Wolynes PG. As simple as can be? Nat Struct Biol. 1997;4:871–874.
26. Li H, Tan C, Wingreen NS. Nature of driving force for protein folding: A result from analyzing the statistical potential. Phys Rev Lett. 1997;79:765–768.
27. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol. 1999;6:1033–1038.
28. Chan HS. Folding alphabets. Nat Struct Biol. 1999;6:994–996.
29. Cieplak M, Holter NS, Maritan A, Banavar JR. Amino acid classes and the protein folding problem. J Chem Phys. 2001;114:1420–1423.
30. Rose GD, Gierasch LM, Smith JA. Turns in peptides and proteins. Adv Protein Chem. 1985;37:1–109.
31. Rose GD, Presta LG. Helical signals in proteins. Science. 1988;240:1632–1641.
32. Petuhov M, Uegaki K, Yumoto N, Serrano L. Amino acid intrinsic α-helical propensities III: Positional dependence at several positions of C-terminus. Protein Sci. 2002;11:766–777.
33. Bhattacharjee N, Biswas P. Position-specific propensities of amino acids in the β-strand -strand. BMC Struct Biol. 2010;10:29.
34. Alber T, Bell JA, Dao-Pin S. et al. Replacements of Pro-86 in phage T4 lysozyme extend an α-helix but do not alter protein stability. Science. 1988;239:631–635.
35. Matthews BW. Structural and genetic analysis of protein stability. Annu Rev Biochem. 1993;62:139–160.
36. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. The design and characterization of two proteins with 88% sequence identity but different structure and function. Proc Natl Acad Sci U S A. 2007;104:11963–11968.
37. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci U S A. 2009;106:21149–21154.
38. Levitt M. Conformational preferences of amino acids in globular proteins. Biochemistry. 1978;17:4277–4285.
39. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry. 1974;13:222–245.

SUPPORTING INFORMATION

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

How to cite this article: Škrbić T, Maritan A, Giacometti A, Banavar JR. Local sequence-structure relationships in proteins. *Protein Science*. 2021;30:818–829. https://doi.org/10.1002/pro.4032