Received: 17 January 2019 | Revised: 21 May 2019 | Accepted: 7 June 2019

DOI: 10.1002/prot.25758

RESEARCH ARTICLE

PROTEINS WILEY

On the presence of short-range periodicities in protein structures that are not related to established secondary structure elements

Ioannis G. Riziotis | Nicholas M. Glykos

Department of Molecular Biology and Genetics, Democritus University of Thrace, University campus, Alexandroupolis, Greece

Correspondence
Nicholas M. Glykos, Department of Molecular Biology and Genetics, Democritus University of Thrace, University campus, Alexandroupolis, Greece.
Email: glykos@mbg.duth.gr

Abstract

Standard secondary structure elements such as α-helices or β-sheets, are characterized by repeating backbone torsion angles (φ,ψ) at the single residue level. Two-residue motifs of the type $(\varphi,\psi)_2$ are also observed in nonlinear conformations, mainly turns. Taking these observations a step further, it can be argued that there is no a priori reason why the presence of higher order periodicities can not be envisioned in protein structures, such as, for example, periodic transitions between successive residues of the type (…-α-β-α-β-α-…), or (…-β-α_L-β-α_L-β-…), or (…-α-β-α_L-α-β-α_L-…), and so forth, where the symbols (α,β,α_L) refer to the established Ramachandran-based residue conformations. From all such possible higher order periodicities, here we examine the deposited (with the PDB) protein structures for the presence of short-range periodical conformations comprising five consecutive residues alternating between two (and only two) distinct Ramachandran regions, for example, conformations of the type (α-β-α-β-α) or (β-α_L-β-α_L-β), and so forth. Using a probabilistic approach, we have located several thousands of such peptapeptides, and these were clustered and analyzed in terms of their structural characteristics, their sequences, and their putative functional correlations using a gene ontology-based approach. We show that such nonstandard short-range periodicities are present in a large and functionally diverse sample of proteins, and can be grouped into two structurally conserved major types. Examination of the structural context in which these peptapeptides are observed gave no conclusive evidence for the presence of a persistent structural or functional role of these higher order periodic conformations.

KEYWORDS
periodic structures, Ramachandran plot, secondary structure

1 | INTRODUCTION

The definition of secondary structure in biological macromolecules is a core principle in the study of the protein structure and function. It refers to the hydrogen bonded local folding of amino acid residues and the formation of energetically stable structural elements. The pattern of hydrogen bonding is a defining characteristic of the established secondary structure elements, such as the α-helix[1] or the β-sheet.[2] An alternative—but complementary—description of secondary structure is based on the distinct preferences of the backbone (φ,ψ) torsion angles for the various secondary structure elements. The established approach for studying the (φ,ψ) angles of protein structures is the Ramachandran

plot,[3,4] which allows the direct visualization of the torsion angles preferred by the various secondary structure elements such as α-helices, β-sheets as well as less common[5] elements like P_{II}[6-8] and α_L-helices.[9] These preferences lead to the formation of identifiable clusters on the two-dimensional (2D) Ramachandran space, and are described by Hollingsworth et al. as linear groups, that is, conformations constructed by three or more residues with repeating (φ,ψ) pairs.[10] Well-known exceptions to the single-residue (φ,ψ) periodicities are reverse turns or β-turns, as defined and classified in seven types (I, I′, II, II′, VIa, VIb, and VIII) by Venkatachalam et al.[11] and refined by J.S. Richardson.[12] In reverse turns, the central residue (φ,ψ) values of the turn diverge significantly from the values of the two neighboring residues. This distinctive (φ,ψ) value transition pattern is also commonly observed in the extended definition of β-turns by Wilmot and Thornton.[13] The transitions of (φ,ψ)-pairs between two value ranges—known as "regions" in the Ramachandran plot nomenclature—, have been thoroughly examined and generalized beyond the β-turns in a very comprehensive survey by Hollingsworth et al.[14] They are defined as $(\varphi,\psi)_2$-motifs, and refer to any possible and observed case of transition between two Ramachandran regions. The results of this survey shown that these motifs are very abundant in proteins, something that amends the classical definition of (φ,ψ) linearity in secondary structure and suggests that (φ,ψ) periodicities in the two-residue level may be present in secondary structure elements as well. Moreover, it is well-known that backbone torsion angle irregularities, which break linearity, are a very common attribute of loops.[15,16] This is something worth investigating further, in order to determine whether higher-order (φ,ψ) periodicities are present in loops.

In this study, we examine the deposited (with the PDB) protein structures for the presence of nonstandard repeating (φ,ψ)-motifs. We algorithmically searched for five-residue long protein fragments in which successive residues adopt conformations that alternate between two distinct major Ramachandran regions. For instance, motifs of the type (α-β-α-β-α) or any other possible repeating transition ensemble between distinct (φ,ψ) value pairs. To avoid recapturing common secondary structure elements such as β-turns—which also adopt $(\varphi,\psi)_2$ motifs—we completely omitted Gly and Pro residues from our search. Finally, and as will be discussed later, the reason that we have limited our search to peptapeptides is that no statistically significant results could be obtained from the longer peptides we examined.

In the following paragraphs, we describe the probabilistic algorithm we devised for analyzing the five-residue fragments derived from the PDB, and extensively discuss the statistical analyses performed aiming to meaningfully cluster the derived peptide structures. This is followed by the analysis of these pentapeptides in terms of their sequence, secondary structure preferences, functional diversity and correlation to their structural context. We conclude by discussing the limitations and implications of this work, especially with respect to the possibility that even higher dimensionality periodicities could be present in the known protein structures.

2 | METHODS

2.1 | Data preparation

A sample of 27 300 protein structures with ≤80% sequence identity, solved at 3.0 Å resolution or better, were culled from the *PDB*[17] using a list generated from the *PISCES*[18] server. We chose a rather low-resolution cut-off so that the sample is large enough for the statistical analysis to be meaningful. The (φ,ψ) dihedral angles of the proteins were extracted by the structure analysis program *PROCHECK*,[19] and used as input for our motif searching algorithm.

2.2 | Algorithmic principles

The algorithm we devised uses a probabilistic treatment to search and score five-residue fragments that are consistent with the sought structural motif of periodic transitions between two and only two distinct Ramachandran regions as shown schematically in Figure 1. In the first step, and starting from a concatenated dihedral angle list of the whole data set as input, a complete set of all observed (with the PDB) five-residue fragments which have defined φ,ψ angles and do not contain glycine or proline, is obtained (8 304 637 pentapeptides). In the second step, these fragments are scored depending on their consistency with the sought structural motif, which comprises three basic rules (see Figure 1):

1. Residues i, $i + 2$, $i + 4$ must all reside on one Ramachandran region.
2. Residues $i + 1$, $i + 3$ must reside on another Ramachandran region.
3. The two regions must be distinct.

Representing residues as data points on the 2D Ramachandran space, we define as (Δ_n) and (d_n) the Euclidean distances between residues in the same region and distinct regions respectively. The process of Euclidean distance calculation requires the φ,ψ coordinates of two residues and must take into account the circular periodicity of the

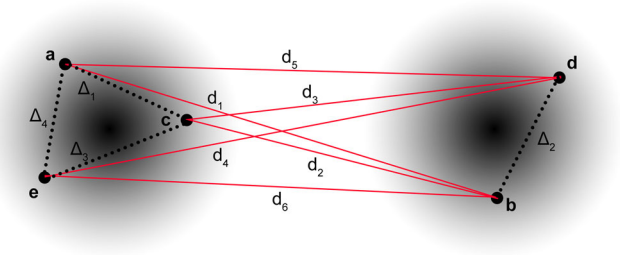

FIGURE 1 Schematic representation of the structural motif searched by the algorithm. The two 2D Gaussians represent distinct Ramachandran regions and the five points (a,b,c,d,e) represent the coordinates of five consecutive residues of a hypothetical peptide fragment. The black dotted lines are the difference vectors (Δ_n) between residues of the same region, while the red lines represent the distances (d_n) between residues of distinct regions. The algorithm calculates these distances for all PDB-derived pentapeptides and converts them to a goodness-of-fit metric [Color figure can be viewed at wileyonlinelibrary.com]

angles. In a grid of nine (three by three) identical Ramachandran plots, the distance between two residues *A* and *B*, is the minimum distance between the $A(\varphi_0, \psi_0)$—in the central plot—and all the symmetric points of $B(\varphi_i, \psi_i)$. Details of the algorithm are described in the following paragraphs.

2.3 | Probabilistic definition of Ramachandran regions

As discussed in the previous section, what we want to estimate is the probability that a set of residues (defined by their φ,ψ angles) belong to the same or distinct regions of the Ramachandran space. Defining distinct regions with hard-coded limits on (φ,ψ) angle space leads to loss of generality and is difficult to encode probabilistically. If, however, we view the Ramachandran plot as a set of superimposed 2D Gaussians, a more general treatment is feasible as implied by Figure 1: Calculating the probability that two points reside in the same or different regions ("Gaussians" from the algorithm's point of view), only requires the coordinates of points as well as an estimate for the variance and mean of the respective distributions. To make the calculations tractable without losing generality, we have resorted to the following three simplifications: The first is the assumption that the distribution of the allowed regions on the Ramachandran plot can indeed be viewed as a superposition of symmetric 2D Gaussians. The second simplification is based on the assumption that the variance of these Gaussian approximations is identically the same for all allowed Ramachandran regions. The third simplification is that a meaningful estimate of the probability that two points belong to the same (or distinct) regions can be obtained without knowledge of the center of the respective regions if all pairwise comparisons are based on the global distribution of the difference vectors in Ramachandran space. Please do note that all these approximations are safe in the sense that by overestimating (through averaging) the variance of these Gaussian distributions, as will be discussed in the next paragraph, we minimize the probability of missing a true hit (ie, we minimize false negatives). Unavoidably, this leads to a concomitant increase of noise (ie, false positives) but we feel that this is an acceptable strategy given the subsequent steps of clustering analysis performed as will be discussed in section 2.5. We should also note at this point that the main purpose of defining Ramachandran regions in a coarse manner, is to facilitate representation of dihedral transitions between distinct angle value ranges, rather than classifying them in the established terms (core, generously allowed, allowed, and not allowed regions).[19] Having that being said, performing a fine-grained classification of dihedral transitions, would require the explicit definition of each of the classic Ramachandran regions, canceling the distance-based, probabilistic approach we used and leading to a loss of generality. On the contrary, our method is not dependent on (φ,ψ) pairs themselves, that is, no angle data enter the score calculation. Additionally, the classification of the hits is performed by straightforward pairwise comparison via Cartesian clustering, assuring structural consistency among cluster members. This focuses on the structural content of the fragments adopting these transitions, making the exact position of residues relatively to the core-regions on the Ramachandran space less meaningful. For the reasons described, our results will be presented in STRIDE terms, to indicate the transition patterns of the peptide fragments in a comprehensive way and provide better visualization.

Based on the analysis outlined above, and given a distance between two residues in Ramachandran space, our algorithm only needs four parameters to estimate the sought probabilities. The first two parameters are the mean distance and corresponding variance over all pairs of residues that belong to the same Ramachandran region. The third and the fourth parameters are the mean distance and corresponding variance for pairs of residues belonging to distinct regions. To average-out the differences between the shape and extent of the allowed Ramachandran regions, we obtained estimates for these four parameters from (a) the 1D distribution of all PDB-derived distances between residues [*i*, *i* + 1] (which according to our algorithm must reside on different regions), and, (b) the 1D distribution of the [*i*, *i* + 2] distances (which, for the motif we examine here, must reside in the same Ramachandran region). These two distributions are shown in the top panel of Figure 2. Not unexpectedly, the two distributions are quite similar and both display two distinct peaks, one at small distances, the second at a distance of approximately 180°. We assigned the first peak to the distribution of distances between residues that belong to the same cluster, and the second to the distribution for residues that belong to different Ramachandran regions. The deviation of these peaks from a normal distribution is probably due to the deviation of the actual allowed Ramachandran regions from ideally symmetric two-dimensional Gaussians.

We obtained numerical values for the means (μ) and standard deviations (σ) of those two peaks (treating them as ideal normal distributions) through a nonlinear regression fitting performed with the function *nls()* of the *R* package. Due to the deviation of these peaks from ideal normal distributions, we have used for our fitting only the right and left halves of the corresponding low and high deviation peaks.

The derived estimates were found to be:

- $\mu_1 = 0$ and $\sigma_1 = 10.55$ for the distribution of distances between residues that belong to the same Ramachandran region, and,
- $\mu_2 = 180.0$ and $\sigma_2 = 34.30$ for the distances between residues belonging to distinct regions.

These values were used to convert distances to probabilities as described in the next section.

2.4 | Calculation of log-odd scores

For every pentapeptide fragment obtained from the PDB, the algorithm calculates all the distances shown in Figure 1, and converts them into probabilities using the complementary error function *erfc (x)*.[20] Using the means and standard deviations obtained as described in the previous section (with $\mu_1 = 0$ and $\sigma_1 = 10.55$ for the same region case, $\mu_2 = 180.0$ with $\sigma_2 = 34.30$ for the vectors between distinct regions), the probabilities are calculated using the following formulas:

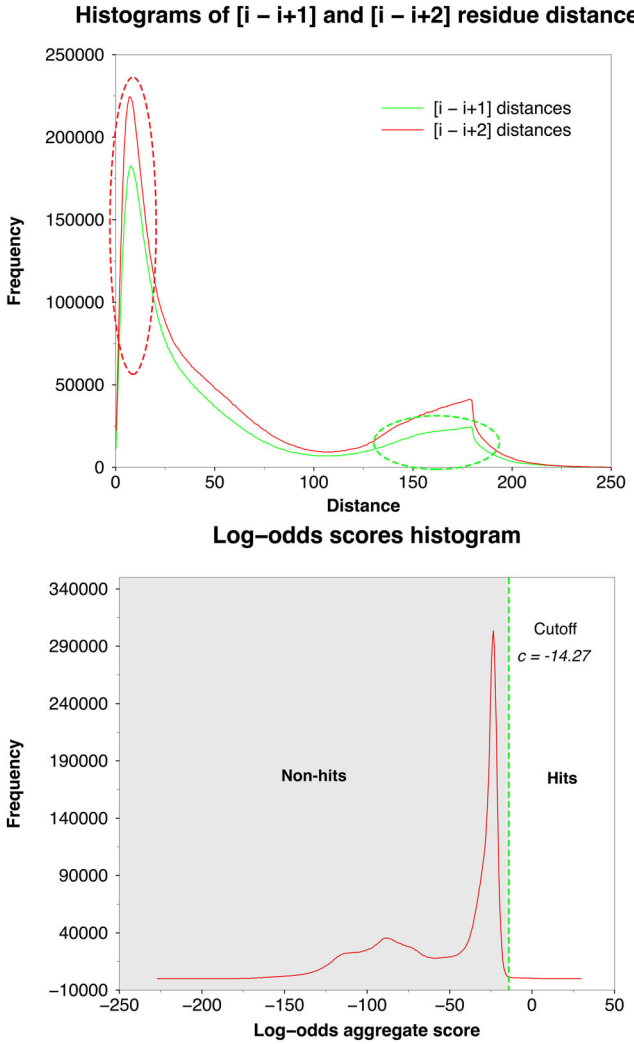

FIGURE 2 Top Panel: Histograms of inter-residue pairwise distances. The two superimposed curves are the histograms of the distances between adjacent residues [*i*, *i* + 1] (green curve) and of the distances between residues [*i*, *i* + 2] (red curve). Two peaks are observed in both distributions; the peak on the left corresponds to the distribution of distances between residues in the same Ramachandran region. The peak on the right corresponds to the distribution of distances between residues that belong to distinct Ramachandran regions. To estimate the parameters needed for the algorithm, nonlinear regression fitting was performed in the right and left halves of the respective peaks (shown in green and red circles, respectively, see text for details). Lower panel: histogram of peptide scores. The red line in this figure shows the distribution of the log-odd scores obtained from all PDB-derived pentapeptides in our sample. Approximately normally distributed major peak defines the noise level of the calculation. The vertical green dotted line, which was taken as the cutoff for a peptide to be considered as a hit, lies at four standard deviations away from the mean of the noise peak and corresponds to a (non-normalized) log-odd cutoff of ~−14 [Color figure can be viewed at wileyonlinelibrary.com]

(a) The probability $P_{\Delta n}$ that two residues belong to the same region is given by:

$$P_{\Delta_n} = erfc\left(\frac{\Delta_n - \mu_1}{2\sigma_1\sqrt{2}}\right)$$

where (Δ_n) is the distance—in Ramachandran space—between the two residues under examination, and (μ_1, σ_1) are the expected mean and variance of the difference vectors for amino acids that belong to the same Ramachandran region.

(b) The probability P_{dn} that two residues belong to distinct regions is given by:

$$P_{d_n} = erfc\left(\frac{d_n - \mu_2}{2\sigma_2\sqrt{2}}\right)$$

where (d_n) is the distance—in Ramachandran space—between the two residues under examination, and (μ_2, σ_2) are the expected mean and variance of the difference vectors for amino acids that belong to distinct Ramachandran regions.

For each peptide fragment, a total of 10 probability values were calculated corresponding to the 10 inter-residue vectors shown in Figure 1. These were then converted to a final (per pentapeptide) nonnormalized log-odd score using the *logit* function:

$$logit(P) = \log\left(\frac{P}{1-P}\right)$$

The calculations described above were performed through the application of the R statistical package in combination with locally written programs.

2.5 | Structure selection and clustering

The lower panel of Figure 2 shows the distribution of the non-normalized log-odd scores obtained from all pentapeptides in our sample. The major noise peak is distributed Gaussian-like with a mean log-odd score of $\mu = -23.67$ and a corresponding SD of $\sigma = 2.35$ (estimated via *nls* fitting). We compared the results obtained by using two different cut-off values for selecting putative hits, the first at 3σ above the mean of the noise peak and the other at 4σ above mean (corresponding to log-odd score cut-off values of −16.61 and −14.27, respectively). Examination of the resulting structures for their consistency with the sought motif showed that the 4σ cut-off significantly reduced the noise while simultaneously gave a large enough sample for the subsequent analyses to be statistically meaningful. With the selected 4σ cut-off, a total 25 643 structures were classified as hits. After omitting identical homopolymeric entries and entries containing noncanonical number of backbone atoms, a final grand total of 12 525 structures remained, that correspond to 8535 unique PDB entries.

These hits had to be grouped according to their structural similarity.[21] Five clustering methods were tested, which are extensively discussed in the Supporting Information file:

1. Cartesian RMSD hierarchical clustering with a preset RMSD cut-off.
2. Cartesian RMSD hierarchical clustering with automatic optimal cluster number estimation.

3. Torsion angle RMSD hierarchical clustering with a preset RMSD cut-off.
4. Torsion angle RMSD hierarchical clustering with automatic optimal cluster number estimation.
5. Dihedral Principal Component Analysis (dPCA).[22,23]

The two methods of automatic cluster estimation were carried on by the *pamk*[24] function of the *R* statistical package, which implements the *k-means* algorithm.[25] *R* was also used for the two non-automatic cluster number estimation methods. The RMSD methods required the construction of two RMSD matrices, one using the Cartesian coordinates of the backbone atoms, the other the respective φ,ψ dihedral angles. The cartesian RMSD and torsion RMSD matrices were calculated by the programs *CARMA*[26] and *torsionRMSD*, respectively, while the dPCA was performed with the program *GRCARMA*.[27] These RMSD matrices are shown in the lower panels of **Figures S1-S5**. The RMSD cut-offs for the two nonautomatic methods were estimated by studying the RMSD histograms in both cases. The cut-off values correspond to the local minima between the major peaks of both histograms (1.59 Å for the Cartesian and 1.44 rad for the dihedral methods respectively, the histograms are shown in **Figure S6**). The results derived from every clustering method were compared by studying the Ramachandran plots, the secondary structure assessment WebLogos[28,29] and the dissimilarity RMSD matrices of the highly-populated clusters (**Figures S1-S5**). As will be discussed in the next section, the automatic methods and the dPCA do not group the structures as accurately as the Cartesian or dihedral clustering with preset RMSD cutoffs, and tend to merge rather dissimilar clusters into two large ones. Between the two remaining clustering methods, the dihedral clustering produces numerous clusters but with only minor differences between pairs of different clusters. The conclusion derived from the above analyses was that the most accurate method is the Cartesian RMSD hierarchical clustering with set RMSD cut-off as discussed in detail in Supporting Information file.

3 | RESULTS

3.1 | Clustering results

Cartesian clustering with a preset RMSD cut-off of 1.59 Å produced a total of 20 clusters, of which only seven contained a significant number of members. Of these seven clusters, the top two account for 85% of the number of pentapeptide structures in our sample. Table 1 lists population statistics for these clusters. It is worth noting here that all five of the clustering methods that we have tried (see previous section), also returned the same two major clusters accounting for approximately 85% of the pentapeptides in our sample.

Figure 3 shows a graphical representation of the RMSD matrix between all pentapeptides in our sample after sorting them according to which cluster they belonged. The origin of the matrix is at the upper left-hand corner, warm colors (yellow, red) indicate high RMSD values and cold colors (green, blue) low values as shown in the color bar. The matrix clearly shows that the clustering method selected performs well with a clear separation between relatively homogeneous clusters as indicated by (a) the low RMSDs between members of the same cluster (dark blue/green squares centered on the matrix diagonal), and, (b) the relatively high RMSDs between structures that belong to different clusters (off-diagonal yellow/orange parallelograms). Even at this coarse level of the analysis, some additional observations can be made. For example, the lighter blue/green color of cluster 2 indicates the presence of higher structural variability in this cluster. Or, for another example, the off-diagonal green rectangle that is aligned with the location of clusters 6 and 1 indicate the presence of a structural similarity between those two clusters, while the orange rectangle connecting clusters 5 and 2 indicate significant structural differences between the corresponding clusters. These observations are placed on a solid ground with the direct structural comparison shown in Figure 4. The *second* column in Figure 4

TABLE 1 The seven most populated clusters

Cluster	Cluster members out of total 12 525 peptides	Percentage (%)
1	5779	46.1
2	4926	39.3
3	116	0.9
4	90	0.7
5	873	7
6	353	2.8
7	336	2.7
TOTAL	12 473	99.5

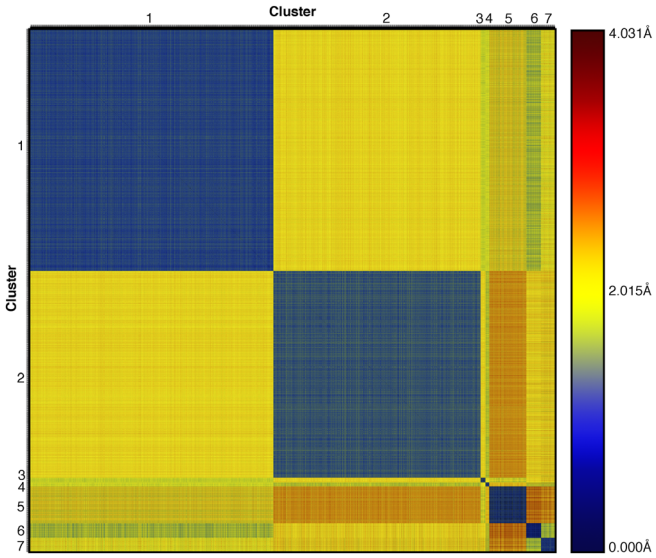

FIGURE 3 RMSD matrix of the seven prominent clusters. Color representation of the RMSD matrix for all pentapeptides belonging to the seven characterized clusters. Warm colors (red/yellow) indicate high RMSD values, cold colors (blue/green) indicate low values. The RMSD values range between 0.0 and 4.031 Å as shown in the color scale bar to the right of the diagram. See text for details [Color figure can be viewed at wileyonlinelibrary.com]

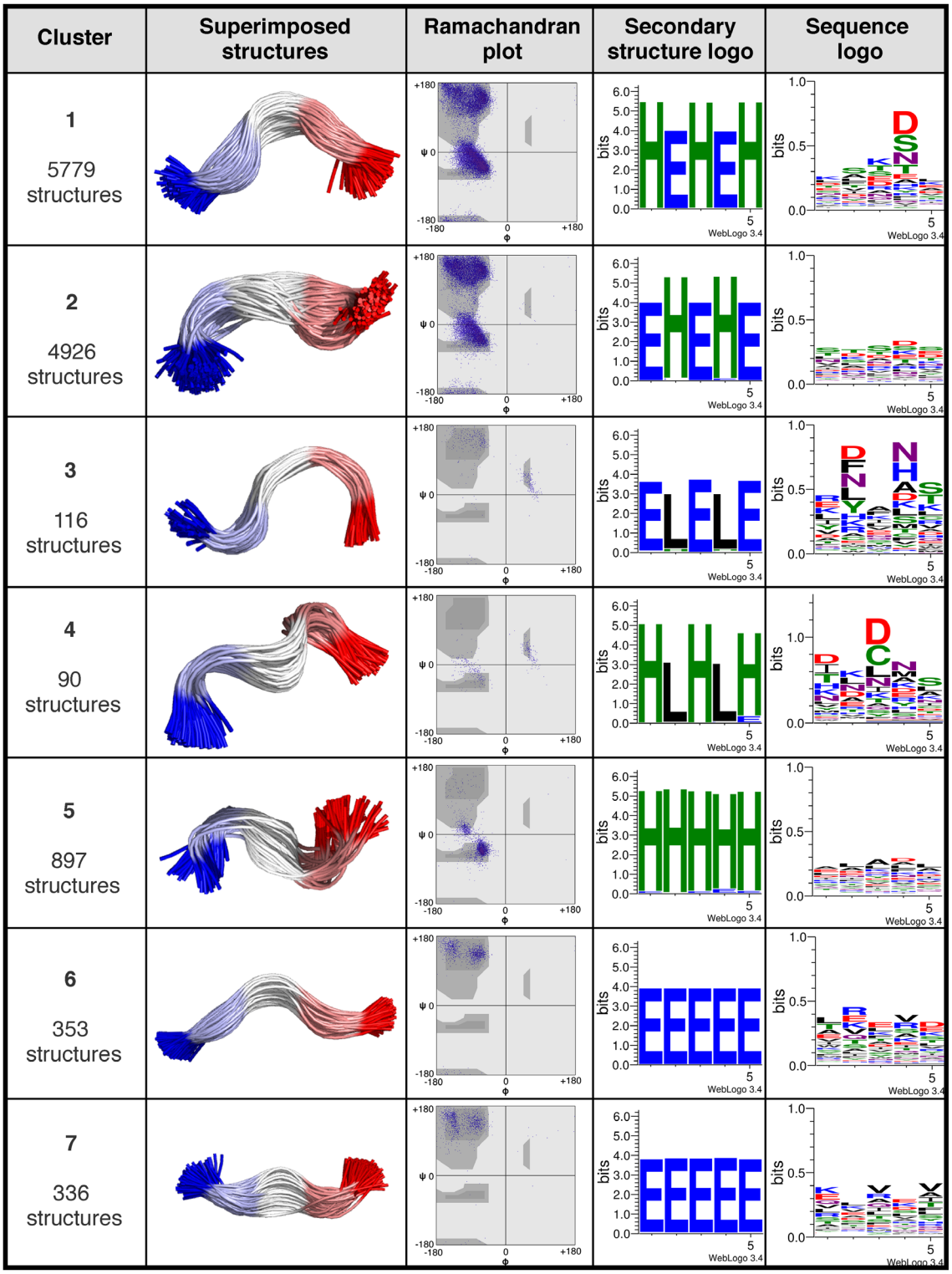

FIGURE 4 Structural and sequence-based comparison of the pentapeptides. This figure, organized in the form of a table, compares for each of the seven clusters identified the backbone 3D structures (second column), the Ramachandran plots (third column), the per-residue secondary structure assignments (fourth column), and finally, the peptides' sequence preferences in the form of a WebLogo (fifth column). The secondary structure preferences are depicted using a WebLogo-like representation (H, right-handed helix; E, extended β structure; L, αL-helix). The structural diagrams are superpositions of representative members of each cluster and what is shown here is a smoothed backbone-only representation using an N-to-C-terminus color scheme. See text for further details

compares schematic diagrams of the structures of the seven clusters using a smoothed backbone representation (with the colors indicating the N-to-C-terminal peptide polarity). To clearly show the amount of structural variability present in each cluster, what is shown in this figure is a least-squares superposition of several (equally spaced) members from each cluster. Structure graphics in this column were prepared in PyMol.[30]

Before continuing with the detailed analysis of these structures (next two sections), it is important to first demonstrate that these peptides are indeed unrelated functionally and evolutionarily, that is, they do not represent a mere repetition of sequences derived from homologous protein families. That this is indeed the case, and that the structures that belong to the different clusters are indeed evolutionarily independent, is shown by (a) the divergence of the peptide sequences as seen in the *fifth* column of Figure 4, and more importantly, (b) by the functional divergence of the proteins from which these peptides were obtained which is shown in Figure 5. This figure shows results from a gene ontology analysis using the GO enrichment tool ReViGO[31] and clearly demonstrates the large functional diversity that is present in sample of proteins from which these peptide segments were derived. Something to be commented also is the presence of hits in proteins of large and small size, >800 and <100 residues, respectively. Large proteins usually adopt different structural motifs than regular globular proteins, with a representative example being collagen, which is a Polyproline-II repeating motif, forming a fibrous triple helix structure.[32,33] Such instances make up the 1.06% of our nonredundant dataset, and only the 0.89% of unique PDB IDs in the seven most populated clusters. With respect to small proteins (<100 residues), although these make up 16,6% of the dataset, only 3.87% of small protein unique PDB IDs are in the seven clusters. As these percentages are rather low in the hits, we can safely say that neither large nor small proteins are overrepresented, and the vast majority of the hits correspond to proteins spanning from 100 to 800 residues.

One last thing that is worth noting here, is that the motifs we located have a common characteristic with the reverse β-turns in terms of φ,ψ angles, in that in both cases motifs of the type $(\varphi,\psi)_2$ are adopted.[14] The main difference is that reverse turns are four residue-long and adopt a single $(\varphi,\psi)_2$-motif in the central two residues, while in the peptides we located the $(\varphi,\psi)_2$-motifs are continuous in all five residues.

3.2 | Prominent peptide conformations correspond to two types of *a-β* transition motifs

Two transition motifs, α-β-α-β-α and β-α-β-α-β (motifs 1 and 2 hereafter) are the most highly populated clusters and are shown in the first two rows of Figure 4. Motif 1 adopts a broad turn conformation, with a tighter bulge in the center, formed by residues *i* + 2 and *i* + 3, as shown in more detail in Figure 6 (please note that the term "bulge" is used here to describe the particular backbone curvature and is not a reference to the β bulges). In contrast with the established turns, this conformation is not stabilized through hydrogen bonding. To show that this is indeed the case, we examined the peptide structures for the presence of putative hydrogen bonds between the —NH group of residue *i* + 3 and the —CO group of residue *i* + 1, and similarly for the pairs of residues (*i* + 2, *i* + 4) and (*i*, *i* + 2). By using the very generous criteria of a donor-acceptor distance of less than 3.9 Å[34,35] and an acceptor-donor-hydrogen angle of less than 63°,[36] the hydrogen bond formation frequencies for the (*i* + 3:*i* + 1), (*i* + 2:*i* + 4), and (*i*:*i* + 2) pairs were found to be only 22.3%, 3.5%, and 2% of the total number of structures that belong to motif 1. The observation that hydrogen bonding is not the main force stabilizing the structure of this motif is consistent with its rather extended form, and indicates that these loop-like structures are possibly stabilized from their structural environment in the context of their corresponding folded proteins.

Motif 2 (β-α-β-α-β, second row of Figure 4) is a conformation that could alter the orientation and direction of secondary structure elements on the —N and —C termini, as it adopts an "S"-like shape with two bulges facing opposite directions. The N-terminal bulge consists of residues *i*, *i* + 1, *i* + 2 and the C-terminal bulge consists of residues *i* + 2, *i* + 3, *i* + 4. These bulges could also be stabilized by hydrogen bonds; the presence of three potential hydrogen bonds has been examined in the cluster and shown in Figure 6: *i* + 2 → *i*, *i* + 4 → *i* + 2 and *i* + 3 → *i* + 1 with occupancies 30%, 27.9%, and 0.8%, respectively. The two favored hydrogen bonds of high occupancy (*i* + 2 → *i* and *i* + 4 → *i* + 2), can also occur concurrently, as observed in several structures of the cluster. The abundance of the α-β-α-β-α and β-α-β-α-β motifs indicates high structural conservation, while sequence analysis (sequence logos[28] in Figure 4) shows significant diversity in the primary structure level.

A natural question that arises at this point is the following. Do these loop-like peptide structures have a preferential structural context in which they occur? For example, do they connect other secondary structure elements on either side of the peptide, and if yes, are there preferences as to what type of secondary structure elements they connect? To answer this question, we show in Figure 7 the WebLogo representation of the STRIDE-derived secondary structure assignments for the five-residue motifs along with assignments of 10 residues preceding and following the corresponding motifs. Clearly, and at least for motifs 1 and 2, notably strong preferences are indeed present: Motif 1 appears to predominantly connect helical segments, whereas Motif 2 demonstrates a preference for β structures. To place this observation on a firm ground, we show in **Figures S7** and **S8** a graphical representation of the raw secondary structure data used for calculating the WebLogo representation of Figure 7. These two figures leave little doubt: motif 1 is indeed a mostly α-helical-connecting structure, whereas motif 2 is dominantly associated with β structures. To put this in numbers we have performed for motif 1 a cluster analysis of the STRIDE assignments (the cluster analysis was performed as described in ref. 37). What we have found is that ~44% of the cluster members adopt a helix-loop-helix structure, with an additional 15% and 6% having only a C- or N-terminal helix respectively. These preferences, however, are not strict as shown in **Figure S12**. This figure depicts a representative collection of actual structural schematics for few of the structures that belong to these two motifs. Clearly, and although most of the structures for motif 1 are helical and for Motif 2 extended beta, notable exceptions are present in both cases. An interesting observation here is the connection between these

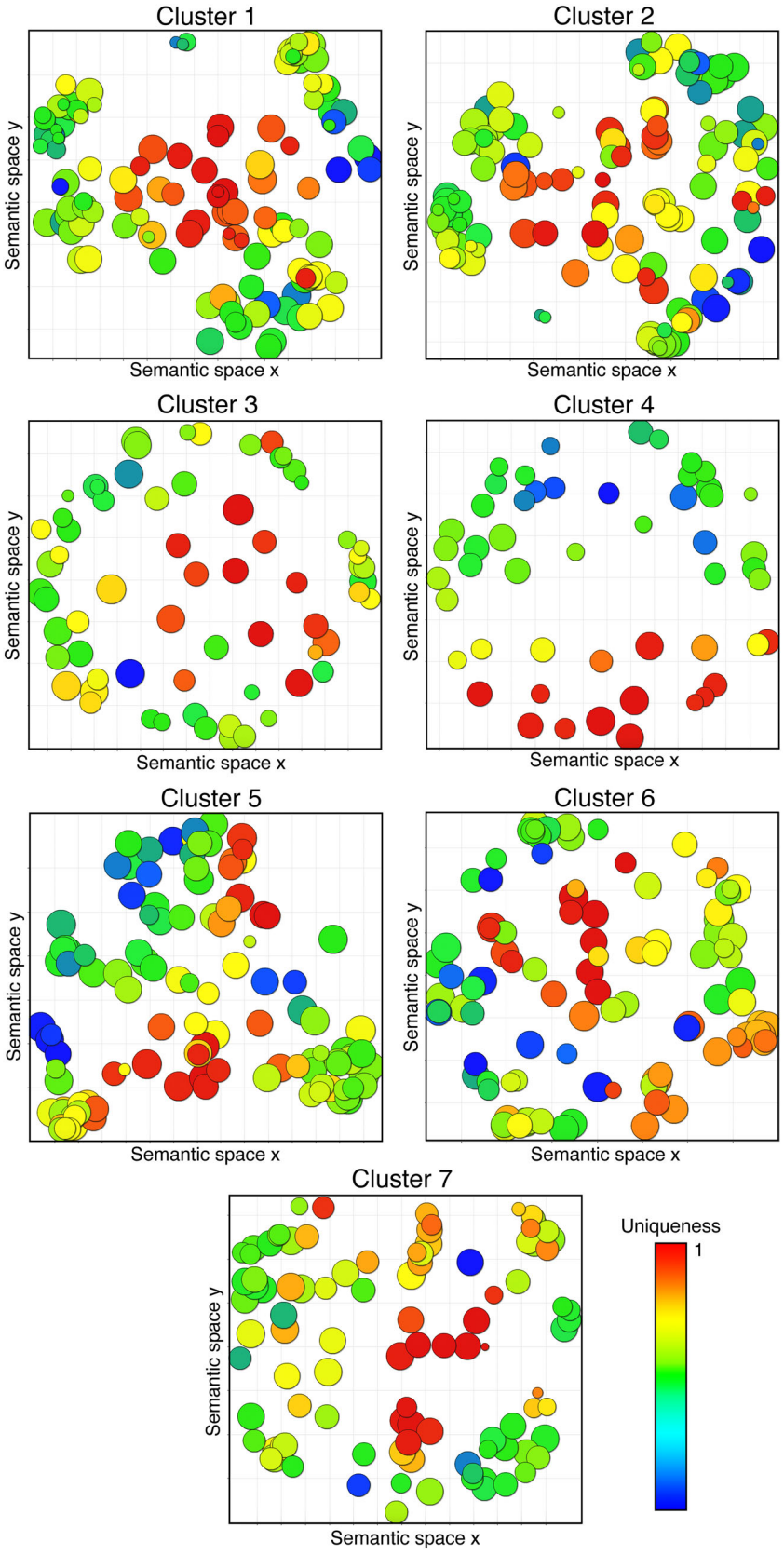

FIGURE 5 Visualization of gene ontology enrichment of each cluster. These scatter plots represent the diversity of biological functions of the proteins in each cluster and were created by the program ReViGO[31] using the list of GO terms as derived from the PDB.[17] The radius of each bubble is proportional to the generality of the corresponding GO term (smaller bubbles indicate more specific GO terms) and the color scaling represents the uniqueness of each term in the list. Note that every cluster is heterogeneous in terms of biological function, a clear indication that the peptides in our sample are not evolutionarily related repetitions derived from homologous protein families [Color figure can be viewed at wileyonlinelibrary.com]

secondary structure preferences and the motifs per se: motif 1 is of the ***α***-*β*-*α*-*β*-***α*** type and mostly connects αhelical segments. Motif 2 is of the ***β***-*α*-*β*-*α*-***β*** type and predominantly connects β structures. Seen in this light, it can be argued that what these motifs essentially represent is structurally conserved three-residue-long loops connecting α or β structures respectively. Having said that, and as shown in

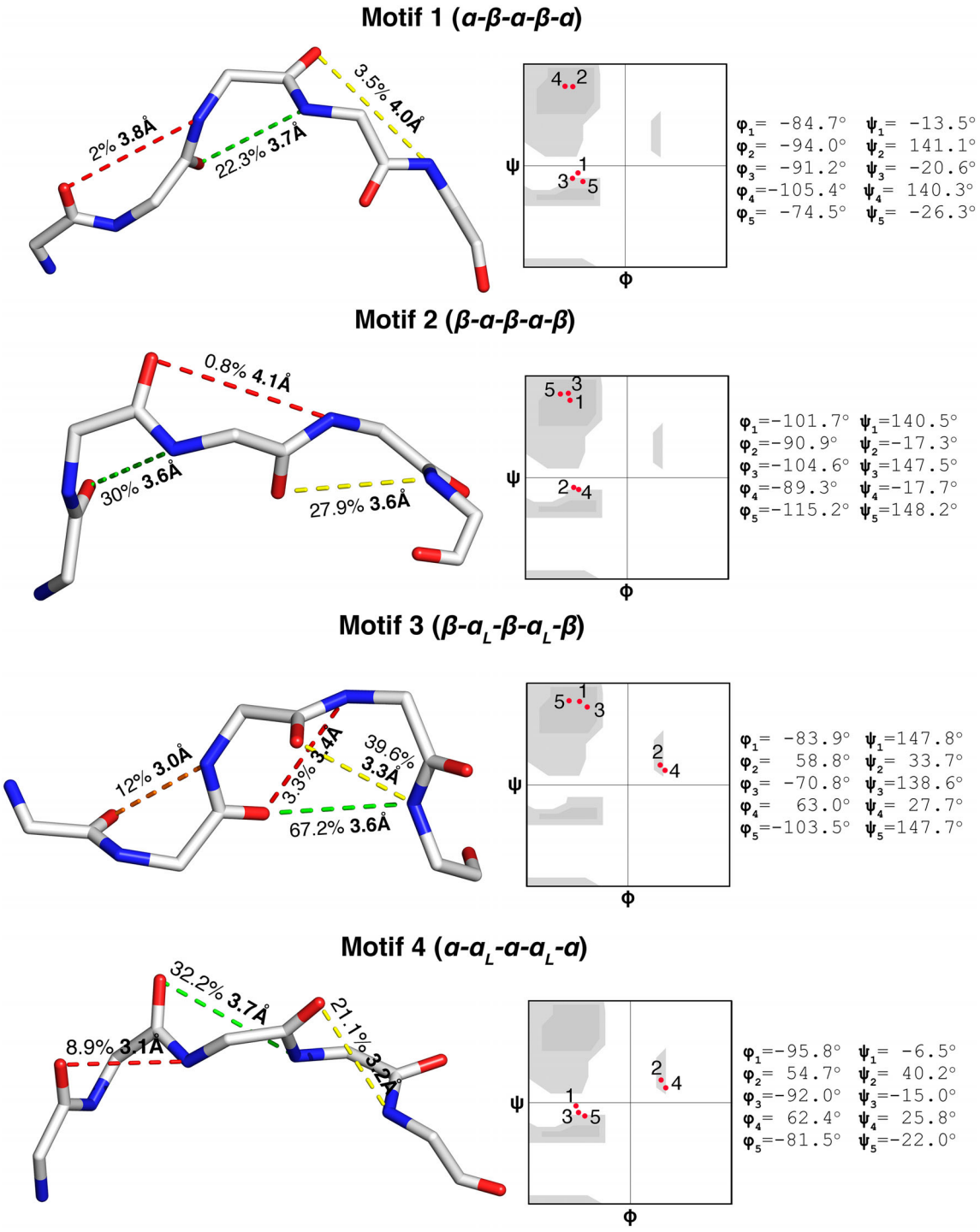

FIGURE 6 Representative backbone structures and Ramachandran plots of motifs 1-4. Comparison between the representative backbone structures and their corresponding Ramachandran plots for Motifs 1-4. The representative structures shown here are those members of the corresponding motifs, which have the lowest RMSD from the average structure of the motif (ie, the structure calculated by averaging the Cartesian coordinates of all superimposed structures that belong to the motif). Putative hydrogen bonds together with their frequencies and distances are also depicted. The Ramachandran plots shown on the right column of this figure were calculated from the representative structures shown, are marked with the respective residue numbers, and their (φ,ψ) angles are listed [Color figure can be viewed at wileyonlinelibrary.com]

Figures **S7**, **S8**, and **S12**, this observation should be seen as a statistical preference and not as a rule.

Something that should be clarified at this point is that STRIDE[38] secondary assignments as shown in Figure 7 are mostly meaningful for the N- and C- terminal sequences of the pentapeptides, rather than the pentapeptides themselves. This is because STRIDE considers hydrogen bonding and structural context besides backbone torsion angles to estimate short-range conformations. In the case of the

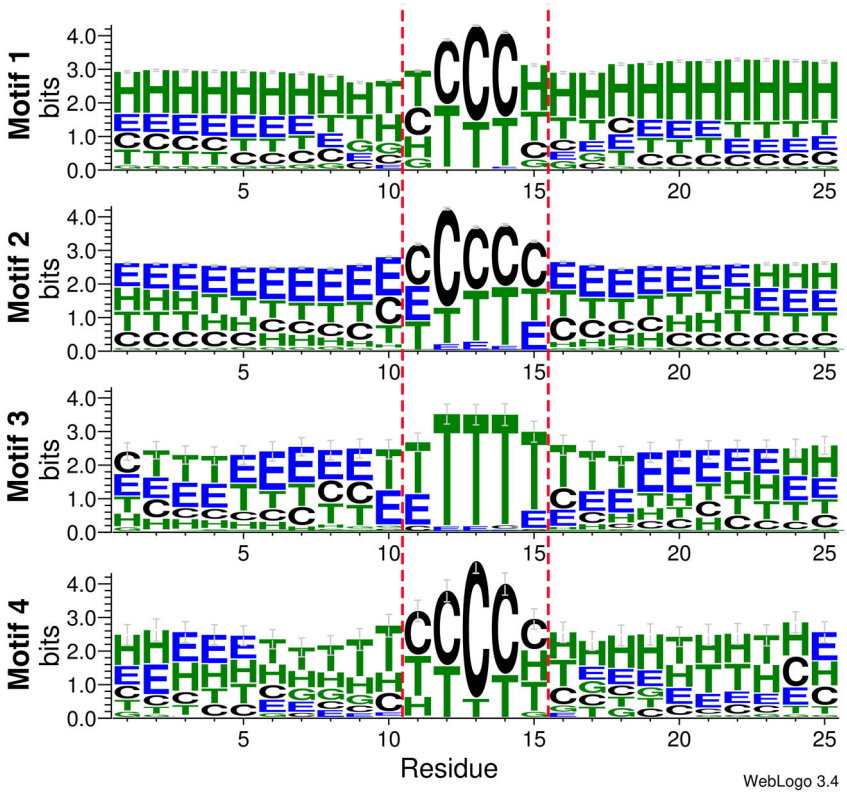

FIGURE 7 Structural context of the four major motifs. Weblogo-like representations of the per residue secondary structure preferences for the four major motifs within the structural context in which they occur. The STRIDE[38] assignments for each of the four motifs correspond to the five central residues (marked as 11-15 and enclosed in the red dashed lines). The assignments for the 10 residues preceding and following the corresponding motifs are marked as residues 1-10 and 16-25. The letter code for the secondary structure assignments are: H-G, helical; C, coil; T, turn; E, extended β structure [Color figure can be viewed at wileyonlinelibrary.com]

pentapeptides we found, we use the terms α, β and α_L to describe the position of each residue on the Ramachandran space, and not to make secondary structure assignment, as this cannot be defined in single-residue level. Due to that, there is seemingly no agreement observed between the WebLogos in the fourth column of Figure 4 and the five central residues in the WebLogos of Figure 7. In the first case, the logos are utilized to show the transition type consistency among the cluster members, whereas in the second case, the logos show the secondary structure variation of the context, assigning the pentapeptides prominently as coils and turns.

3.3 | α-α_L and β-α_L transition motifs are observed in lower frequencies

In addition to motifs 1 and 2 described above, our algorithm also located two lower frequency motifs of the types β-α_L-β-α_L-β and α-α_L-α-α_L-α (marked as motifs 3 and 4 in Figure 4). Representative structures for these two clusters can be seen in Figure 6. In motif 3, an inverse "S" conformation, with a small N-terminal bulge and a larger C-terminal bulge is observed. In contrast with motifs 1 and 2, hydrogen bonding is important for these structures: Using the same hydrogen bond criteria mentioned in the previous section, four hydrogen bonds (colored dashes in Figure 6) are formed in motif 3 between residues $i + 4 \rightarrow i + 1$, $i + 4 \rightarrow i + 2$, $i + 2 \rightarrow i$ and $i + 3 \rightarrow i + 1$ with respective frequencies of 67.2%, 39.7%, 12%, and 3.5%. The hydrogen bonding pattern together with the torsion angle preferences makes the C-terminal bulge of motif 3 very similar to a type II β-turn, thus allowing the metaphorical description of motif 3 as a "double turn."

Motif 4 is unusual in that it demonstrates a continuous transition between the α and α_L regions of the Ramachandran diagram. Although it is observed in low frequency (~0.7%), its structure shows some interesting characteristics: the C_α atoms are arranged in a bow-like broad conformation, with the carbonyl oxygen atoms oriented in such a way as to point away from the structure's curvature (clearly seen in Figure 6). As with the previous motifs, we have examined the structures for the presence of hydrogen bonds (using again the generous criteria described in section 3.2). For the residue pairs $i + 3 \rightarrow i + 1$, $i + 4 \rightarrow i + 2$ and $i + 2 \rightarrow i$ the observed frequencies were found to be 32.2%, 21.1%, and 8.9%, respectively, indicating that again that the main stabilizing force for these structures is not intra-peptide hydrogen bonding.

3.4 | Motifs 5, 6, and 7 populate transitions within the same Ramachandran regions

A by-product of the distance-based algorithm that we implemented (see section 2), is the identification by our programs of motifs that involve transitions within the same Ramachandran region. These motifs, although consistent with our algorithmic design, are not fully consistent with the sought target of our analyses. Having said that, and for completeness, we show in Figure 4 the structural characteristics of these three additional motifs (motifs 5, 6, and 7, last three rows of Figure 4). Motif 5, and as can be seen in its corresponding Ramachandran plot (third column of Figure 4), mainly populates transitions within the helical region of the plot, with the several of these transitions being of the type α-3_{10}-α-3_{10}-α and 3_{10}-α-3_{10}-α-3_{10}.[39] Motifs 6 and 7 (two last rows in Figure 4) both demonstrate transitions

within the β region of the Ramachandran plot, with motif 6 populating transitions of the type P_{II}-β-P_{II}-β-P_{II}, while cluster 7 occupies the complementary β-P_{II}-β-P_{II}-β motif. The main structural difference between these two transitions concerns the orientation of the carbonyl oxygen atoms, which are pointing to opposite directions when the other backbone atoms are structurally aligned. Although these intra-region transitions may have some structurally interesting characteristics, the rather minor structural fluctuations that they demonstrate between successive residues makes them fall outside the scope of this communication and will be no further analyzed.

3.5 | Sequence and functional diversity

The last (fifth) column of Figure 4 shows WebLogo representations of the peptide sequences corresponding to the seven characterized motifs. Clearly, the similarity of the peptides at the structural level (second column) is not due to the presence of any form of conservation in sequence space. The implication of this finding is clear: these structurally conserved motifs are not evolutionarily related and they can be observed in a diverse set of functionally and evolutionarily unrelated proteins. Having said that, we have examined in structural terms some of the most pronounced preferences seen in these WebLogo diagrams. For example, there is a weak preference for residues D, S, N, T in position 4 of motif 1, while motif 2 shows no particular preference. In motif 3, preferences are observed in positions 2 and 4 (D,F,N,L and D,H,A, respectively) while in motif 4 there is a relatively highly conserved position 3 with D,C,L residues being the most frequently observed. A natural question is whether these weak sequence preferences have a structural basis,[21] for example a side chain forming electrostatic interactions with the peptide backbone. We have examined in structural terms such putative preferences and in **Figure S11** we show an example of such an analysis for Motif 1. This figure depicts representative Motif 1 structures of each of the four residue preferences at position 4. The structures shown in this diagram are the representative structures produced by Cartesian PCA analysis with side chains added only in position 4 and only for residues D,S,N, and T. The orientation of the side chains is towards the exterior of the backbone curve, indicating that it is unlikely for these side chains are involved in direct interactions with the main chain atoms. However, WebLogo and per-sequence STRIDE analyses have shown that when these four residues are present, the formation frequency of C-terminal α-helices is higher. A possible explanation of this finding is that interactions between the fourth residue side-chain of the motif and the following α-helix could lead to the stabilization of the latter.

Although the absence of detectable sequence conservation in the seven characterized motifs clearly points to the absence of functional relationships between the corresponding proteins, we felt that a proper analysis in functional terms of the proteins involved was necessary in order to reach a definitive conclusion. To this end, we performed a gene ontology enrichment procedure using a GO[40] database obtained directly from the PDB which assigns every chain of a PDB entry to its corresponding GO terms. For each cluster, all parent polypeptide chains were assigned to one or more GO terms, and the results were plotted using the tool ReViGO.[31] The GO terms bubble plots shown in Figure 5. Each bubble represents one GO term from the provided list, and the dissimilarity among them is assessed statistically and defined by two principal components which are denoted as "semantic space x" and "semantic space y" in the ReViGO nomenclature. The color scale and radius of each bubble in these diagrams represent the uniqueness and generality of each term respectively. The dispersion and heterogeneity of these plots clearly indicate that the proteins from which these peptide fragments have been derived correspond to a functionally heterogeneous ensemble of parent polypeptide chains, and are not just repeating occurrences from structurally and functionally similar subunits, in good agreement with the results obtained from the WebLogo analysis discussed in the previous paragraph.

3.6 | Shorter and longer-range periodicities

A natural question that arises at this point is whether periodicities similar to those observed for pentapeptides are also present in the case of shorter or longer peptides. With respect to dihedral transitions between three or four consecutive residues, these comprise a well-established and characterized ensemble of structural motifs, commonly known as tight turns. As already mentioned in the introduction, β-turns comprise four residues, of which the two central ones (i + 1 and i + 2) reside in distinct regions on the Ramachandran space. Eight types of β-turns have been identified, each one adopting a different $(\varphi,\psi)_2$-motif.[13,14,41] A similar case is seen in γ-turns, which include three residues instead of four, with (φ,ψ) pairs $(\varphi_1,\psi_1) = (172^\circ,128^\circ)$; $(\varphi_2,\psi_2) = (68^\circ,-61^\circ)$; $(\varphi_3,\psi_3) = (-131^\circ,162^\circ)$[42]; that is, two transitions on the Ramachandran space involving one favored and one non-favored region. Due to dihedral transitions being very common motifs in turns, we have chosen to exclude three and four-residue peptide segments from our search, as the algorithm hits would contain all those well-characterized types of turns. This is also the reason why the Gly and Pro residues were omitted from our search, as these are usually observed in the central position of several types of β-turns.[13] Moreover, the main question to be addressed through this study is the existence of—consecutive—dihedral transitions in proteins, so for this to be satisfied, the peptide fragment length should be at least four residues.

Regarding fragments longer than five residues, we have performed a search for transitions (followed by the same Cartesian clustering method as the one used in the case of pentapeptides) for peptide fragments ranging from 6 to 10 residues. Although several such longer peptides have been located by the search algorithm, the number of hits drops so fast with increasing peptide length that we feel that the results lack statistical significance. To put this in numbers, for peptides ranging from 5 to 10 residues we have located the following number of hits (expressed as the number of unique PDB entries from our non-redundant dataset): 5-peptides: 8535; 6-peptides: 1644; 7-peptides: 326; 8-peptides: 80; 9-peptides: 10; 10-peptides: 5. The prominent conformations observed for these fragments are presented in **Figure S13** in the form of superimposed structures along with the corresponding transition patterns (only the most highly populated clusters are shown). Note how the number of hits drops by almost a

factor of five for each added residue, leading to sparsely populated clusters for peptides longer than 7 residues, especially for the 10-residue peptides where hits were located in only five unique PDB entries. Naturally, no meaningful clustering could be performed for these longer than 7-mer fragments, and only some representative structures are shown in **Figure S13**. Comparison of the structures shown in Figure 4 with the structures of these longer peptides, indicates the presence of a persistent set of conformations for the central region of the peptides which remain relatively unaltered as the length increases, while the terminal regions show slightly more variance. Although the clustering results shown in **Figure S13** could possibly be meaningful for the hexapeptides, we are convinced that the much higher sample size for pentapeptides, together with the structural similarity between the peptapeptides and the longer peptides justifies our choice to restrict our analysis to five-residue fragments. This peptide size appears to be adequate for characterizing short-range periodicities in the two-residue level, while maintaining the structural content that is also observed in the longer fragments.

4 | SUMMARY AND CONCLUSIONS

We have shown that motifs of the type X-Y-X-Y-X, where X and Y can be any of the three major regions of the Ramachandran plot, are represented in significant populations in the set of known protein structures. Our analysis indicated that out of all possible permutations of this type of transitions, the α-β-α-β-α and β-α-β-α-β transitions (motifs 1 and 2 in Figure 4) are the prominent ones based on their frequencies of occurrence, followed by the β-α_L-β-α_L-β and α-α_L-α-α_L-α motifs. Both of the two major motifs (α-β-α-β-α and β-α-β-α-β) demonstrate an extended loop-like structure that frequently serves as a connecting element in helix-loop-helix and β-hairpin supersecondary structures, respectively. However, as **Figures S7**, **S8**, and especially **Figure S12** show, these are statistical preferences only (and not a defining characteristic of the motif and the corresponding structures). Nevertheless, and regarding Motif 1, we can probably safely conclude that the presence of N- and C-terminal α-helices is clearly preferred. Combining the information from hydrogen bonding and the occurrence of the hydrogen bonding patterns previously discussed in the center of the pentapeptide, indicates that these peptides may serve as structurally stable connecting regions in helix-loop-helix motifs. Similarly, in Motif 2, whole-molecule STRIDE[38] assignments show frequent β-strands on both sides of the pentadeptides (**Figure S8**). This together with the observed hydrogen-bonding frequencies may indicate their loop role in β-hairpins. However, representative supersecondary structures cannot be estimated in Motif 2, as there is high secondary structure diversity on either side of the pentapeptides, in contrast to Motif 1 where there is higher structural context consistency among the cluster members.

Other than the pure structural interest of characterizing previously unknown motifs, the work reported here may find putative applications mainly in the field of protein structure prediction algorithms. For instance, a deep neural network could be trained to predict the conformation of such motifs, among others, using sequence-level data to refine the overall predicted structure. Moreover, and as discussed previously, a conclusion drawn from our results is that these short motifs are frequently observed in loops. Research focusing on such elements is increasingly gaining popularity, especially since loops have been shown to serve major functional roles[43] such as the hypervariable antibody regions.[44] Providing additional structural insight on how loops adopt to varying structural motifs can also be useful for creating or improving existing methodologies for their structural detection and characterization.[45]

We will conclude this communication with a discussion of what we perceive is the major limitation of the work reported here. In addition, this is no other than the idea already presented in the introduction of this manuscript: There is a multitude of possible higher order periodicities that could involve any imaginable type of transitions involving multiple Ramachandran regions and multiple successive residues. For example, we could envision periodicities of the type …-α-β-α_L-α-β-α_L-… or …-α-α-β-β-α-α-… or …-α_L-β-β-α_L-α_L-β-β-α_L-… and so forth. Of all these putative periodicities, here we examined the deposited with the PDB structures only for a very specific type involving successive transitions between two and only regions of the Ramachandran space. The fact that we have located several thousand fragments demonstrating such transitions indicates that a more complex search for higher order periodicities may well be a worthwhile exercise. Performing such a generalized high dimensionality search may not be trivial however. Standard tools for discovering periodicities such as the Fourier transform spring to mind but their application would not be straightforward, especially if a probabilistic treatment is required. Additionally, if numerous residues are considered for such a search, the curse of dimensionality will almost certainly present itself, together with its implied computational complexity. Finally, such a search for higher order periodicities—even if successful—is bound to only locate "rare" substructures in the sense that if such periodicities were common in protein structures they would already have been characterized. Having said that, we believe that a search for such previously uncharacterized protein periodicities would be an interesting exercise in expanding our view of protein structures.

ORCID

Ioannis G. Riziotis https://orcid.org/0000-0002-4035-1839
Nicholas M. Glykos https://orcid.org/0000-0003-3782-206X

REFERENCES

1. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*. 1951;37(4):205-211.
2. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A*. 1951;37(5):251-256.
3. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J Mol Biol*. 1963;7:95-99.
4. Ramakrishnan C, Ramachandran GN. Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophys J*. 1965;5(6):909-933.

5. Hollingsworth SA, Karplus PA. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts*. 2010;1(3–4):271-283.
6. Arnott S, Dover SD. The structure of poly-L-proline II. *Acta Crystallogr B Struct Sci Cryst Chem*. 1968;24(4):599-601.
7. Adzhubei AA, Sternberg MJ. Left-handed polyproline II helices commonly occur in globular proteins. *J Mol Biol*. 1993;229(2):472-493.
8. Adzhubei AA, Sternberg MJ, Makarov AA. Polyproline-II helix in proteins: structure and function. *J Mol Biol*. 2013;425(12):2100-2132.
9. Eisenberg D. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc Natl Acad Sci U S A*. 2003;100(20):11207-11210.
10. Hollingsworth SA, Berkholz DS, Karplus PA. On the occurrence of linear groups in proteins. *Protein Sci*. 2009;18(6):1321-1325.
11. Venkatachalam CM. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers*. 1968;6(10):1425-1436.
12. Richardson JS. The anatomy and taxonomy of protein structure. *Adv Protein Chem*. 1981;34:167-339.
13. Wilmot CM, Thornton JM. Beta-turns and their distortions: a proposed new nomenclature. *Protein Eng*. 1990;3(6):479-493.
14. Hollingsworth SA, Lewis MC, Berkholz DS, Wong WK, Karplus PA. (phi,psi)(2) motifs: a purely conformation-based fine-grained enumeration of protein parts at the two-residue level. *J Mol Biol*. 2012;416(1):78-93.
15. Thukral L, Shenoy SR, Bhushan K, Jayaram B. ProRegIn: a regularity index for the selection of native-like tertiary structures of proteins. *J Biosci*. 2007;32(1):71-81.
16. Bonet J, Planas-Iglesias J, Garcia-Garcia J, Marin-Lopez MA, Fernandez-Fuentes N, Oliva B. ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Res*. 2014;42(Database issue):D315-D319.
17. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242.
18. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics*. 2003;19(12):1589-1591.
19. Laskowski RA, MacArthur MW, Moss DS, TJ M. PROCHECK - a program to check the stereochemical quality of protein structures. *J Appl Crystallogr*. 1993;26:283-291.
20. Weisstein EW. *Erf MathWorld--A Wolfram Web Resource*. Florida: Chapman and Hall/CRC. Available from: http://mathworld.wolfram.com/Erf.html.
21. Krissinel E. On the relationship between sequence and structure similarities in proteomics. *Bioinformatics*. 2007;23(6):717-723.
22. Mu Y, Nguyen PH, Stock G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins*. 2005;58(1):45-52.
23. Altis A, Nguyen PH, Hegger R, Stock G. Dihedral angle principal component analysis of molecular dynamics simulations. *J Chem Phys*. 2007;126(24):244111.
24. Hennig C. A method for visual cluster validation. In: Weihs C, Gaul W, eds. *Classification — the Ubiquitous Challenge: Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation eV University of Dortmund, March 9–11*. Vol 2004. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005:153-160.
25. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. *J Stat Softw*. 2014;61(6):1-36.
26. Glykos NM. Carma: a molecular dynamics analysis program. *J Comput Chem*. 2006;27(14):1765-1768.
27. Koukos PI, Glykos NM. Grcarma: a fully automated task-oriented interface for the analysis of molecular dynamics trajectories. *J Comput Chem*. 2013;34(26):2310-2312.
28. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*. 1990;18(20):6097-6100.
29. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188-1190.
30. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.
31. Supek F, Bosnjak M, Skunca N, Smuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*. 2011;6(7):e21800.
32. Ramachandran GN, Kartha G. Structure of collagen. *Nature*. 1954;174(4423):269-270.
33. Shoulders MD, Raines RT. Collagen structure and stability. *Annu Rev Biochem*. 2009;78:929-958.
34. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. *J Mol Biol*. 1994;238(5):777-793.
35. Torshin IY, Weber IT, Harrison RW. Geometric criteria of hydrogen bonds in proteins and identification of "bifurcated" hydrogen bonds. *Protein Eng*. 2002;15(5):359-363.
36. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577-2637.
37. Adamidou T, Arvaniti KO, Glykos NM. Folding simulations of a nuclear receptor box-containing peptide demonstrate the structural persistence of the LxxLL motif even in the absence of its cognate receptor. *J Phys Chem B*. 2018;122(1):106-116.
38. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins*. 1995;23(4):566-579.
39. Toniolo C, Benedetti E. The polypeptide 310-helix. *Trends Biochem Sci*. 1991;16(9):350-353.
40. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. *Gene Ontol Consort Nat Genet*. 2000;25(1):25-29.
41. de Brevern AG. Extension of the classical classification of beta-turns. *Sci Rep*. 2016;6:33191.
42. Némethy G, Printz MP. The γ turn, a possible folded conformation of the polypeptide chain. Comparison with the β turn. *Macromolecules*. 1972;5(6):755-758.
43. Cortes J, Simeon T, Remaud-Simeon M, Tran V. Geometric algorithms for the conformational analysis of long protein loops. *J Comput Chem*. 2004;25(7):956-967.
44. Regep C, Georges G, Shi J, Popovic B, Deane CM. The H3 loop of antibodies shows unique structural characteristics. *Proteins*. 2017;85(7):1311-1318.
45. Shenkin PS, Yarmush DL, Fine RM, Wang HJ, Levinthal C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers*. 1987;26(12):2053-2085.

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: Riziotis IG, Glykos NM. On the presence of short-range periodicities in protein structures that are not related to established secondary structure elements. *Proteins*. 2019;87:966–978. https://doi.org/10.1002/prot.25758