A New Clustering and Nomenclature for Beta Turns Derived form High-Resolution Protein Structures

Maxim Shapovalov[1,2]
Slobodan Vucetic[2]
Roland L. Dunbrack, Jr.[1,*]

[1] Fox Chase Cancer Center
333 Cottman Avenue
Philadelphia PA 19111

[2] Temple University
1801 N Broad Street
Philadelphia PA 19122

* Corresponding Author:

Roland L. Dunbrack, Jr.
Roland.Dunbrack@fccc.edu

Abstract

Protein loops connect regular secondary structures and contain 4-residue beta turns which represent 63% of the residues in loops. The commonly used classification of beta turns (Type I, I', II, II', VIa1, VIa2, VIb, and VIII) was developed in the 1970s and 1980s from analysis of a small number of proteins of average resolution, and represents only two thirds of beta turns observed in proteins (with a generic class Type IV representing the rest). We present a new clustering of beta turn conformations from a set of 13,030 turns from 1078 ultra-high resolution protein structures (≤1.2 Å). Our clustering is derived from applying the DBSCAN and k-medoids algorithms to this data set with a metric commonly used in directional statistics applied to the set of dihedral angles from the second and third residues of each turn. We define 18 turn types compared to the 8 classical turn types in common use. We propose a new 2-letter nomenclature for all 18 beta-turn types using Ramachandran region names for the two central residues (e.g., 'A' and 'D' for alpha regions on the left side of the Ramachandran map and 'a' and 'd' for equivalent regions on the right-hand side; classical Type I turns are 'AD' turns and Type I' turns are 'ad'). We identify several new types of beta turn. Up-to-date statistics, probability densities of conformations, and sequence profiles of beta turns in loops were collected and analyzed. A library of turn types, *BetaTurnLib18,* and cross-platform software, *BetaTurnTool18,* which identifies turns in an input protein structure, are freely available and redistributable from dunbrack.fccc.edu/betaturn and github.com/sh-maxim/BetaTurn18. Given the ubiquitous nature of beta turns, this comprehensive study updates understanding of beta turns and should also provide useful tools for protein structure determination, refinement, and prediction programs.

Introduction

Ordered protein structures consist of elements of regular secondary structure, such as alpha helices and beta-sheet strands, and irregular elements of structure referred to as loops or coil regions. Due to the globular nature of folded proteins, the direction of the peptide chain often has to change radically within a few short residues within loops, which comprise half of residues in protein structures. These changes in direction are often accomplished by turns, which can occur multiple times in loop regions. A single turn consists of a segment between 2 and 6 amino acids, and is identified by the distance between the Cα atoms of the first and last residues and sometimes by the existence of specific hydrogen bonds within the segment [1-6]. These turn fragments are often hydrophilic, since loops are usually on the protein surface [7-9].

Four-residue *beta* turns are by far the most common turn type [5] and constitute 25-30% of all protein residues [10]. The first examination of beta turns was by Venkatachalam in 1968 [1]. He identified a key hydrogen bond formed between the backbone carbonyl oxygen atom of the first residue and the backbone amide hydrogen atom of the fourth residue of a beta-turn. Venkatachalam devised a system of three turn types: I, II, and III, and their mirror images I', II', and III'. He predicted that the mirror images would be disfavored due to steric clashes.

In 1973, Lewis et al. observed that the backbone 1-4 hydrogen bond is not present in a large number of peptide fragments that might still be considered turns [11]. Instead of the older hydrogen bond criterion, they established the definition of a beta turn requiring a distance between the 1 and 4 alpha-carbon atoms of less than 7 Å, while the central residues 2 and 3 are not part of a helix. This beta-turn definition expansion by Lewis et al. led to the adoption of ten different turn types: I, I', II, II', III, III' from Venkatachalam and new types IV, V, VI, and VII. Type V turns have ϕ,ψ dihedrals of the 2nd and 3rd residues around (-80°, +80°) and (+80°, -80°) respectively. Type VI turns contain a *cis* proline at the 3rd residue. Type VII turns have ψ_2 of

180° or ϕ_3 of 180°. The turns not fitting the established definitions, because two or more of the dihedrals of residues 2 and 3 were not within 40° of the defined values, were placed in a miscellaneous category, type IV.

Richardson et al. [2] kept the same 7 Å criterion used by Lewis et al. and from analysis of a larger data set, reduced the number of beta turn types to six categories with designated ϕ,ψ limits of the second and third residues: I, I', II, II', VIa, VIb, and the seventh miscellaneous category, Type IV. They merged type III with type I turns because they occupy similar regions of the Ramachandran map at positions 2 and 3, and many type III turns that could be identified at that time were part of 3_{10} helices.

In 1988 Wilmot and Thornton [3] scanned a dataset of 58 protein structures for turns, and expanded Richardson's categories by adding type VIII turns, consisting of α and β conformations at the 2nd and 3rd residues respectively [3]. To match each turn to its type, for each residue one of two backbone dihedrals had to be within 30° while the other one was allowed to be within 45° of the canonical values for that turn category. In 1990, the same researchers assigned and named all turns by the Ramachandran plot regions (α_R, β_E, β_P, α_L, γ_L, and ε) of the two central residues [4]. They found 16 types that were observed at least once in the total of 910 turns of 58 protein structures. However, eight types were observed less than 10 times (1%) and three types only once.

Panasik et al. examined the 7 Å cutoff criterion used to assemble the database of turn structures [12]. They discovered that restrained minimization perturbed the original structures only within 1 Å RMSD of their original structure and caused most turns that originally lacked Venkatachalam's 1-4 backbone hydrogen bond to form this hydrogen bond. Perskie and Rose studied 2, 3, and 4-residue segments connecting regular secondary structure types and found that a small number of structures covered 98% of the cases and that these conformations could be generated with a simple energy function containing steric constraints and hydrogen bonds

[13]. The turn type definitions of Richardson et al. [2] and Wilmot et al. [3] have also guided the design of many turn prediction algorithms [14-20]. For example, in Kountouris et al., the prediction includes types I, II, IV, VIII, and "non-specific" [16].

More recently, de Brevern performed clustering of the miscellaneous beta-turn type IV representing one third of all beta turns [21], those not fitting the criteria for types I, I', II, II', VIa1, VIa2, VIb, and VIII [22]. This additional clustering resulted in placing about half of the type IV turns in four groups of turns adjacent to the existing Type I and Type II turn boundaries. However, they do not appear to represent peaks in the density of beta turn conformations, but rather occur after the residues in a 30° x 30° box around the Type I and Type II definitions are removed.

The residue preferences for the most common turn types have been described in previous reports [2, 18, 23]. Hutchinson and Thornton in 1994 analyzed the residue preferences in 2,233 examples of the common turn types in a set of 205 protein chains [18]. They noted the preference for proline at positions in turn types that require ϕ of approximately -60°, such as position 2 of Type I, Type II, and Type VIII turns, and the preference for Gly, Asn, Asp and Ser when ϕ>0° (position 3 of Type II turns, position 2 of Type II' turns, and position 2 and 3 of Type I' turns). Hutchinson and Thornton observed the preference for Asp, Asn and Ser at positions that require a residue in the bridge region of the Ramachandran map (ϕ =-80°, ψ =0°), for instance at position 3 of Type I and Type II' turns. Some residues are preferred because they make hydrogen bonds to other residues within the turn or even to the residue before or after the turn. Some hydrophobic residues are preferred when the turn is a tight beta turn connecting two beta-sheet strands, particularly for positions 1 and 4 of Type I' and II' turns, in some cases due to their role as part of hydrophobic cores of proteins. They also noted that preference for Pro at position 4 of Type VIII turns, which prevents residue 3 from being in the alpha region of the

Ramachandran map [24]. Type VIII turns have a beta conformation residue at this position, while a Type I turn contains an alpha conformation at position 3.

Although many analyses of beta turns have been performed, there are a number of questions that remain outstanding. First, what is the nature of the Type IV beta turns, and can the existing nomenclature be extended to reduce their number? Second, what are the true modes of the density in dihedral angle space for each beta turn? For highly skewed populations, the most common conformation (i.e. modes in the density) for each beta turn would be useful in protein structure determination, refinement, and prediction. Third, what are the amino acid distributions of each beta turn type and can they be rationalized? These have not been reported in detail since the 1990s. Fourth, what is the relationship between 3_{10} helices and beta turns?

With a much larger number of ultra-high resolution structures now available, we have undertaken a fresh analysis of the nature of beta turns in protein structures. We have classified all of the turn conformations in a data set of 1,078 non-redundant protein chains from high-quality crystal structures of 1.2 Å resolution or higher. In keeping with previous analyses [2, 11], we define beta turns as four-residue segments with the Cα atoms of residues 1 and 4 having a distance of ≤ 7.0 Å and secondary structure of the 2nd and 3rd residues different from sheet (E), helix (H), π-helix (I) and 3_{10}-helix (G) according to DSSP [10]. Our clustering is performed with a distance measure based on the seven dihedral angles that connect the Cα atoms of residues 1 and 4: ω(2), φ(2), ψ(2), ω(3), φ(3), ψ(3) and ω(4), similar to what we used for the clustering of antibody CDRs [25].

To identify clusters that truly represent peaks in the density of data points in the space spanned by these dihedral angles, we have utilized a density-based clustering algorithm, DBSCAN [26]. DBSCAN locates regions of high density in data sets as defined by two parameters: *eps* and *minPts*. Data points which have *minPts* or more of other data points within a distance *eps* are termed 'core points' and become members of clusters. Points which are not

core points but also are within *eps* of a core point are termed 'border points'. All other points are 'noise points'. Clusters are formed by finding the connected graphs where the edges are between core points within *eps* of each other. Border points are added to the cluster with the core point that they are closest to. Noise points remain unclustered.

Running DBSCAN on the turn data set and optimizing the *eps* and *minPts* parameters resulted in a set of 11 clusters. After performing kernel density estimates of the Ramachandran maps of residues 2 and 3, we identified multiple peaks in the density consistent with the existence of multiple clusters in some of the DBSCAN clusters. We were able to subdivide several of these turn types with *k*-medoids based on the number of peaks in the density, resulting in a set of 18 distinct turn types. We apply a refined division of the Ramachandran map into distinct populated regions developed by Hollingsworth and Karplus [27] to derive a simple nomenclature for our 18 turn types. For our turn types we provide modal conformations and residue preferences at the first, second, third and fourth positions. We also developed a cross-platform Python script for determining the beta turn types in an input PDB structure.

Results

Clustering of high-resolution data set of beta turns

We compiled a set of 1,078 protein chains with a resolution equal to or better than 1.2 Å and pairwise sequence identity less than 50% (Supplemental Data) [28, 29], comprising protein sequences with 232,865 residues (Table 1). For clustering purposes a total of 16% of all residues were excluded due to missing backbone or Cβ or γ heavy atom coordinates, adjacent chain breaks, presence of alternative conformations, or missing secondary structure information. The resulting data set is named *RefinedSet* with 195,752 residues suitable for turn type analysis. If we define beta turns as four consecutive residues where the central residues 2 and 3 are not in regular secondary structure (alpha helix, 3-10 helix, π helix or beta strands) and the 1-

4 Cα-Cα distance is less than or equal to 7 Å, we can locate 13,030 beta turns in this data set built on 43,162 residues. This is less than 13,030 x 4 or 52,110 residues because 49% of beta turns in *RefinedSet* share residues with one or more other beta turns.

	Count	**No of residues**	**Residue perc.**
Sequences	1,078	232,865	-
CompleteSet			
Residues	-	224,918	100.0
Loops	17,176	102,357	45.5
β turns	21,454	64,632	28.7
Isolated β turns	7,708	30,832	13.7
RefinedSet			
Residues	-	195,752	87.0
β turns	13,030	43,162	19.2
Isolated β turns	6,649	26,596	11.8

Table 1. Loop, beta-turn and their residue statistics in *CompleteSet* and *RefinedSet,* which only contain residues with coordinates. An isolated β turn is a β turn not sharing any of its residues with residues from other β turns.

Beta-turn conformations form overlapping clusters of varying density. Application of a single clustering method such as k-means, k-medoids, or DBSCAN failed to identify all known turn types and created merged data sets with multiple modes in the density. It would either (1) detect most known turn types but fail to separate turns of Type I and Type VIII or (2) separate these two types as two clusters but in addition produce too many small clusters and noise points. Ultimately, a procedure consisting of applying DBSCAN with optimized parameters followed by k-medoids to divide the initial DBSCAN clusters that clearly contained multiple peaks in the dihedral-angle space density proved fruitful.

A single application of DBSCAN on all available 13,030 turns produced 11 clusters. Kernel density estimates (KDEs) of residues 2 and 3 indicated that four of these clusters could productively be subdivided into more than one cluster (Figure 1). The most populous DBSCAN cluster contains 8,455 turns of the classical Type I and VIII turns but in fact exhibits five peaks in

the KDE of residue 3 (Figure 1A). Repeated application of k-medoids to subdivide this cluster, produced 5 clusters that correspond to well-known features in the Ramachandran map of most residue types, including the alpha–helical region, the gamma-prime turn (ϕ,ψ = -84°,68°), the prePro zeta conformation (ϕ,ψ = -137°,76°), and two conformations in the beta region, roughly equivalent to parallel (ϕ,ψ = -120°,125°), and anti-parallel beta sheet (ϕ,ψ = -134°,157°), regions. Each of these 5 sub-clusters exhibited different amino acid distributions indicating that they are indeed distinct turn types.

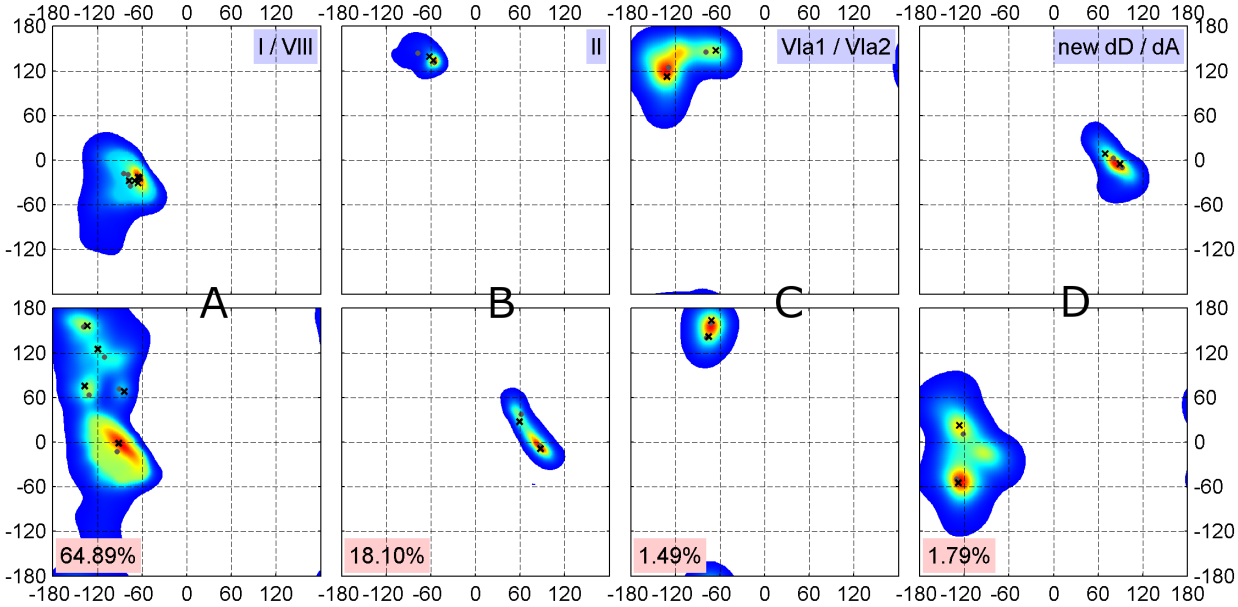


Figure 1. Two-dimensional (ϕ, ψ) kernel density estimates (KDE) of turn residue 2 (top row) and 3 (bottom row) for four multi-modal clusters returned by DBSCAN (of 11 total) during the first clustering round. These 4 clusters have multiple peaks in their densities and were further individually divided into subclusters with one or more rounds of k-medoids. Their final medoids and modes are marked with gray dots and black crosses respectively. The populations of the four clusters are shown in pink. (A) Cluster for classical Type I and VIII turns exhibit 5 peaks. Type I has the strongest peak and account for 76% of the cluster data points. (B) Classical Type II turns can be subdivided into two new clusters; (C) Classical VIa1 and VIa2 are produced by subdividing this DBSCAN cluster (D) This new turn type with residue 2 in the left-handed helical region and residue 3 adjacent to the right-handed helical region was subdivided into two new clusters. The data have 3 obvious density peaks but the smallest peak (ϕ, ψ ~ -90°,-20°) did not have a distinct amino acid profile and had a very small number of points.

The classical Type II turn cluster from the DBSCAN run contained two peaks, which

could easily be subdivided by k-medoids at opposite ends of the left-handed alpha helical region in the Ramachandran map for residue 3 (Figure 1B). The upper peak of these conformations preferred glycine at residue 3, while the other was dominated by non-glycine residues at position 3 (see Figure 2 for comparison of amino acid profiles in all clusters). The Type VIa turn cluster exhibited two peaks in residue 3 that produced equivalents to the previously defined Type VIa1 and VIa2 turns (Figure 1C). Finally, DBSCAN produced a new turn type cluster with left-handed helical residues at position 2 and right-handed helical residues at position 3. This cluster contained two dominant peaks in the density for residue 3 (Figure 1D). We subdivided it into 2 separate clusters with k-medoids. In principle, one of these clusters could be further subdivided but we did not find distinct amino acid profiles, and the smaller peak contained a small number of points.

Properties of new and existing turn types

Our clustering procedure results in a set of 18 turn types (Table 2 and Figure 2) among the 13,030 beta turns detected in *RefinedSet*. The noise group defined by DBSCAN consisted of 319 turns (2.4%). Table 2 contains the populations among our 13,030 beta turns and the modal values of ϕ and ψ for residues 2 and 3. Among our 18 turn types, 6 types are new, 8 types are created by splitting existing Type II turns (2 types), Type VIa turns (2 types), and Type VIII turns (4 types), and the remaining 4 existing clusters are updated versions of the remaining classical turn types (I, I', II', VIb). Instead of the traditional definitions consisting of 30° x 30° bins in ϕ and ψ, we define these turn types as peaks or modes in the seven-dimensional $\omega(2)$-$\phi(2)$-$\psi(2)$-$\omega(3)$-$\phi(3)$-$\psi(3)$-$\omega(4)$ density resulting from our clustering procedure. The turns in other protein structures can be labeled by calculating the distance in the seven-dimensional dihedral space from the centroids of our 18 turn types (see below for description of software).

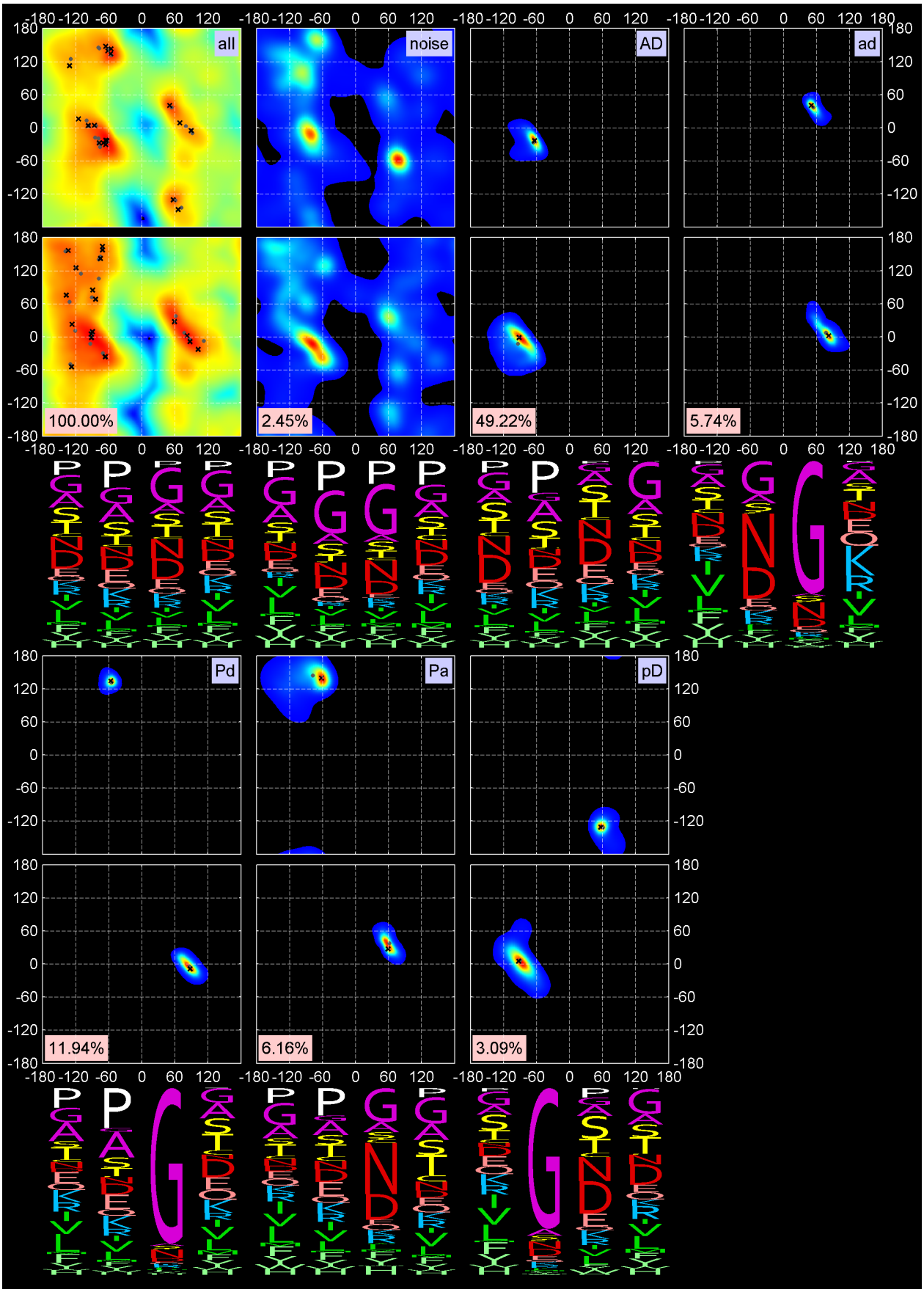
Rather than the Roman numeral system in use since 1973 with abandoned numbers and

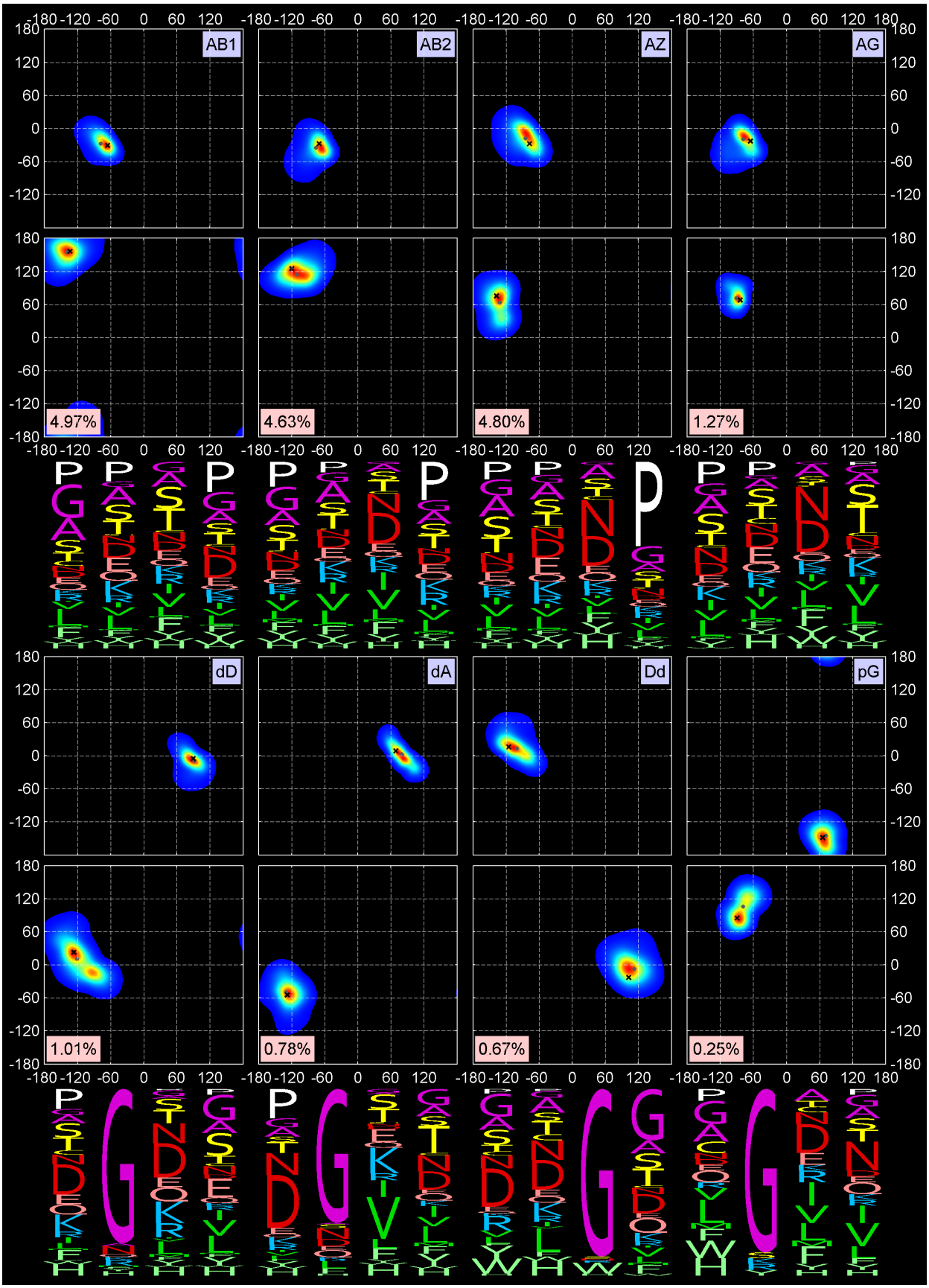
a set of intricate prime symbols and alphanumerical sub-indices, we propose a new nomenclature using the Latin alphabet with type names assigned according to the one-letter region codes of Ramachandran map for the 2nd and 3rd residues of a beta turn (Figure 3). The figure shows our designations and the positions of the medoids of residues 2 and 3 (yellow dots). Our scheme is based on a detailed analysis of the distribution of residues from very high-resolution structures by Hollingsworth and Karplus [30]. Most of the one-letter Ramachandran codes are familiar to structural biologists.

For convenience both in computer code and for people, we converted their Greek designations to the Latin alphabet, for example α to A, β to B, γ to G and so on. We reserve upper-case letters for the left side of Ramachandran map with negative φ and lower-case letters for positive φ. The upper and lower case letters have a point symmetry relative to the map center (0, 0). The only non-standard designation in this scheme is to denote the gamma turn region at $\phi = -84°$, $\psi = 68°$ as "G" in upper case. Originally, this region was designated gamma-prime (g or G') and the corresponding region with $\phi > 0$ was designated as gamma (G). But Hollingsworth and Karplus determined that the gamma region is barely populated, and only the original gamma-prime region is well populated. Thus, we reserved the designation "G" for the gamma prime region on the left side of the Ramachandran map. In only one case (AB1 and AB2), we need a third character to distinguish the two turn types. AB1 is the upper one the Ramachandran map and AB2 is the lower one, which is an easy way to remember them. In Table 2, we also provide a correspondence between the proposed and existing type nomenclatures.

No. by size	Size	Perc.	New name	Previous name	Residue 2 modal value			Residue 3 modal value		
					ω	ϕ	ψ	ω	ϕ	ψ
1	6,413	49.22	AD	I	*trans*	-64	-23	*trans*	-92	-1
2	1,556	11.94	Pd	II	*trans*	-57	134	*trans*	87	-9
3	802	6.16	Pa	new (prev. II)	*trans*	-62	139	*trans*	59	27
4	748	5.74	ad	I'	*trans*	50	40	*trans*	82	2
5	648	4.97	AB1	new (prev.	*trans*	-66	-30	*trans*	-134	157
6	625	4.80	AZ	new (prev.	*trans*	-78	-28	*trans*	-137	76
7	603	4.63	AB2	VIII	*trans*	-70	-27	*trans*	-120	125
8	402	3.09	pD	II'	*trans*	57	-131	*trans*	-93	5
9	166	1.27	AG	new (prev VIII)	*trans*	-66	-23	*trans*	-84	68
10	135	1.04	BcisP	VIa1	*trans*	-131	112	*cis*	-72	164
11	131	1.01	dD	new	*trans*	89	-5	*trans*	-127	23
12	125	0.96	PcisD	VIb	*trans*	-56	143	*cis*	-90	9
13	102	0.78	dA	new	*trans*	69	9	*trans*	-128	-55
14	87	0.67	Dd	new	*trans*	-115	16	*trans*	102	-23
15	59	0.45	PcisP	VIa2	*trans*	-66	148	*cis*	-76	142
16	52	0.40	cisDA	new	*cis*	-98	4	*trans*	-67	-36
17	32	0.25	pG	new	*trans*	66	-148	*trans*	-90	85
18	25	0.19	cisDP	new	*cis*	-86	4	*trans*	-71	158
noise	319	2.45	noise							
Total	**13,030**	**100.00**								

Table 2. A set of 18 beta turn types derived from clustering of the *RefinedSet* of beta turns. The new nomenclature (column 4) based on the Ramachandran map positions (Figure 3) of residues 2 and 3 is provided next to the classical nomenclature. “cis” precedes conformations that contain cis peptide bonds for ω(2) or ω(3). The modal values of the dihedrals ϕ and ψ and the cis/trans configuration of residues 2 and 3 are shown. Complete information for 18 beta turns and beta-turn assignment Python tool are at http://dunbrack.fccc.edu/betaturn or http://github.com/sh-maxim/BetaTurn18.


all
noise
AD
ad
100.00%
2.45%
49.22%
5.74%
Pd
Pa
pD
11.94%
6.16%
3.09%


AB1
AB2
AZ
AG
4.97%
4.63%
4.80%
1.27%
dD
dA
Dd
pG
1.01%
0.78%
0.67%
0.25%

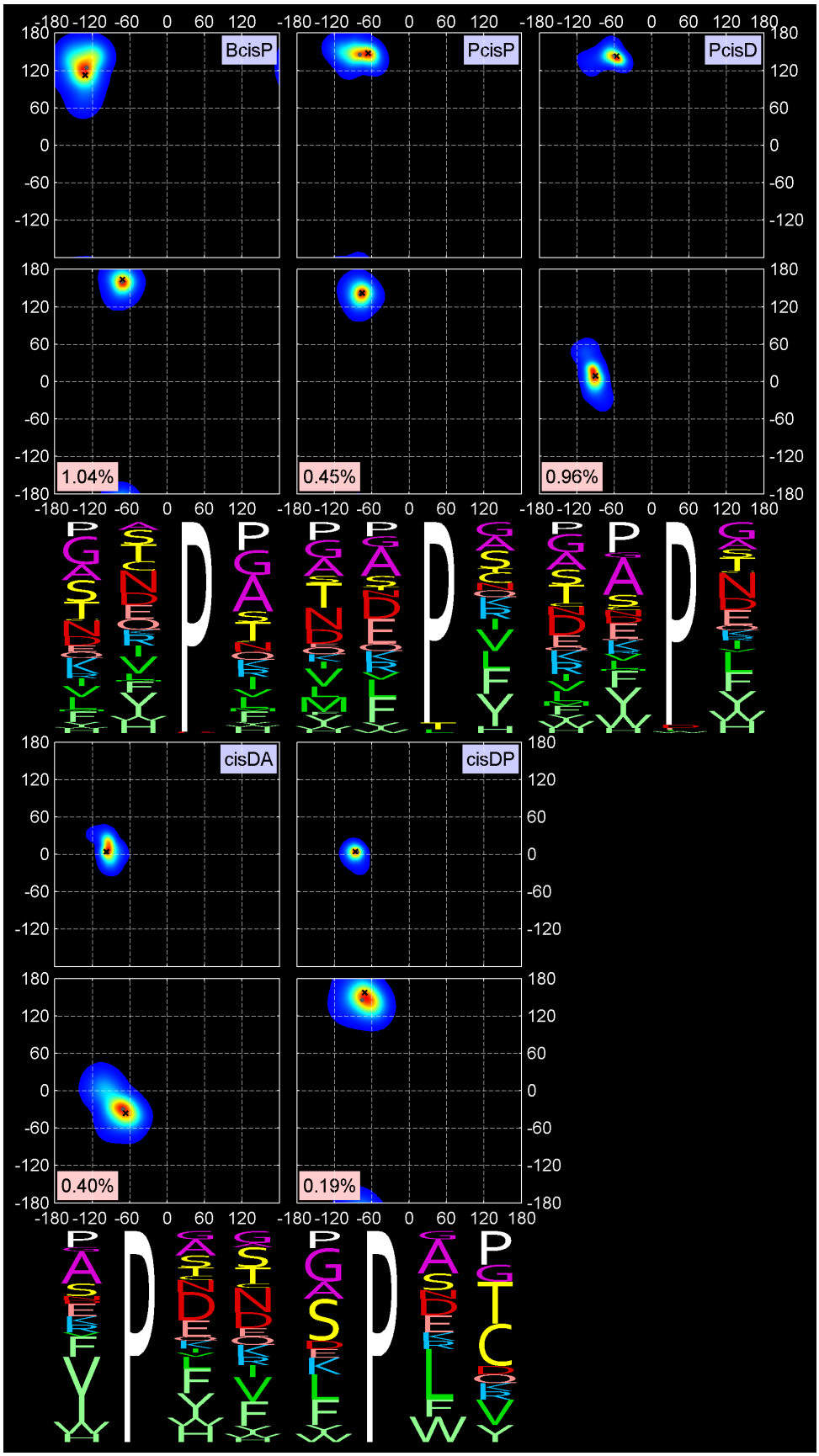

BcisP
PcisP
PcisD
1.04%
0.45%
0.96%
cisDA
cisDP
0.40%
0.19%

Figure 2 (previous pages). Two-dimensional KDEs for (ϕ_2, ψ_2) and (ϕ_3, ψ_3) and amino-acid sequence profiles for 18 turn types, all data points, and noise. The mode and centroid of each cluster are shown with a black cross and gray dot respectively. First row: All data points, noise data points, AD (classical Type I) and ad (Type I') turns. "All" data points KDE is shown on a log scale. Second row: Pa and Pd turns (Type II) and pD turns (Type II'). Third row: AB1, AB2, AZ, and AG turns (all formerly Type VIII). Fourth row: new turn types dD and dA and their approximate inverse, Dd, and new turn type pG. Fifth row: cis3 turn types, BcisP (Type VIa1), PcisP (Type VIa2), and PcisD (Type VIb). Sixth row: new cis2 turns, cisDA and cisDP.

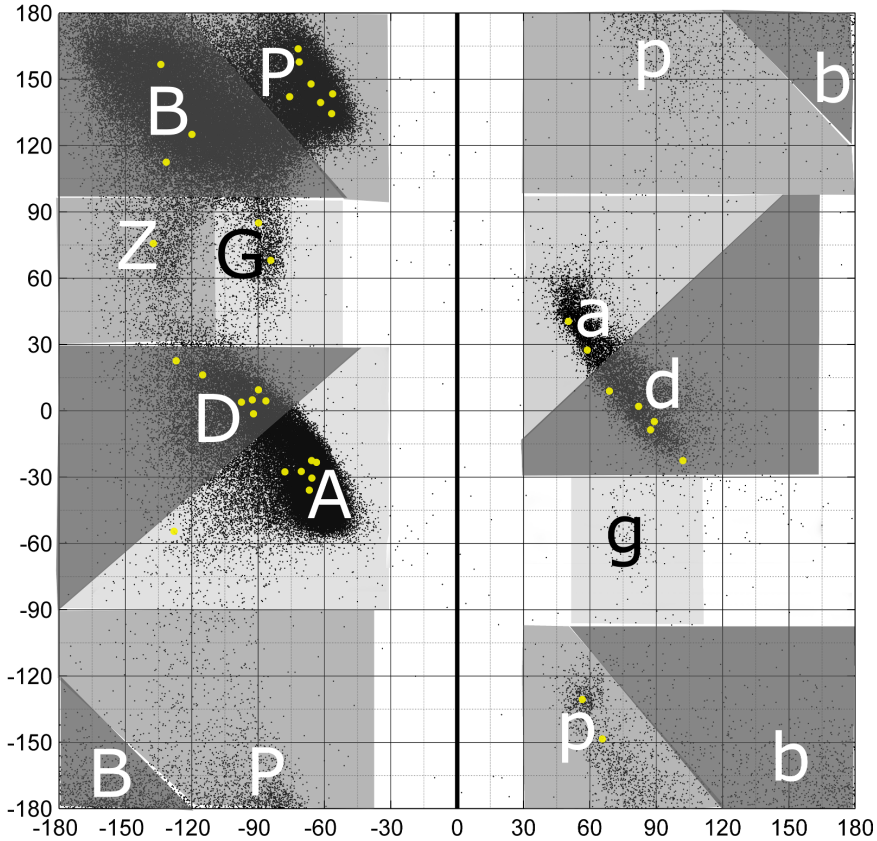


Figure 3. Single-letter region codes for the Ramachandran map. Greek letter designations were converted to Roman letter counterparts. The left side of the map with negative ϕ is encoded in upper-case letters, while the right side of the map with positive ϕ is in lower-case letters. Regions related by symmetry through the origin are denoted with the same letter (e.g., A and a). Points are from all amino acids in the *RefinedSet*, including regular secondary structures and loops. All cluster medoids of residues 2 and 3 are marked with yellow dots.

Structure quality analysis with EDIA [31] based on electron-density maps demonstrated that the 18 turn clusters are of high quality (Figure 4). EDIA integrates electron density in a

sphere around each heavy atom and penalizes both positive and negative density in electron density difference maps. Turns in each cluster have very small populations with substantial inconsistencies (0-2.4%). The noise cluster has a higher rate of substantial inconsistencies with 6.1% of these turns having poor electron density.

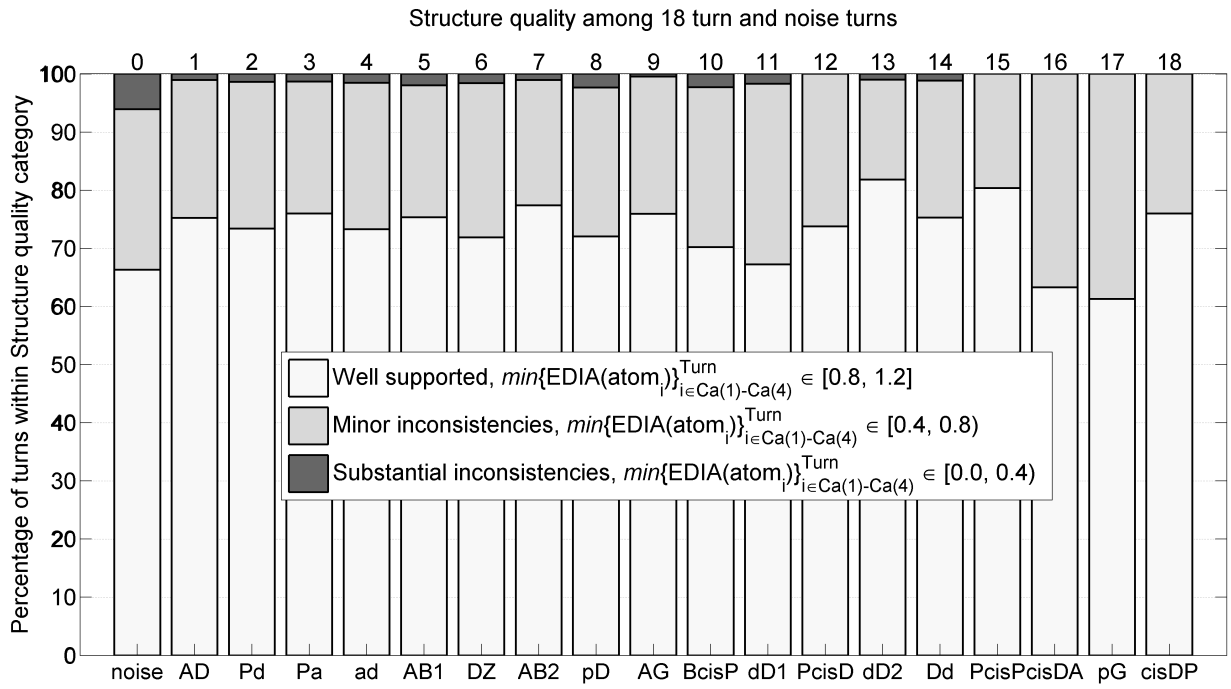


Figure 4. Structure quality of noise turns and 18 turn types based on agreement between atom positions and electron-density maps measured by EDIA software with the minimum of *EDIA* values for all 13 backbone atoms (N, Cα, C, O) in a turn connecting Cα1 to Cα4.

Figure 2 contains the amino acid distributions at all four positions of each beta turn type, including the noise group. As expected, glycine and proline play important roles in many turn types. Proline is required for the two cis2 turn types, cisDA and cisDP, and the three cis3 turn types, BcisP, PcisP, and PcisD. It is also important in position 4 of the AZ turn where it is responsible for the zeta conformation at position 3, which occurs for residues immediately before proline. Proline also prevents the alpha conformation at the residue before it, and is therefore significant in the AB1 and AB2 turns at position 4 and completely absent from AD turns at positions 3 and 4.

Glycine allows amino acids to occupy the “d” and “p” regions of the Ramachandran map, as exhibited by the ad, Pd, dD, dA, and pG type turns. As noted above, the former Type II turns have been divided into Pa and Pd; glycine is dominant at position 3 in the Pd cluster but not so in the Pa cluster, which is instead dominated by Asn and Asp at position 3. Asp and Asn are also dominant at position 2 of the ad turn (former Type I’) and position 3 of the AG and AZ turns.

There are sometimes subtle but statistically significant differences in the sequence profiles of very similar clusters. For instance, residue 3 of AB1 and AB2 has slightly different positions in the Ramachandran map. The upper peak (AB1) prefers Gly, Ala, Ser, and Thr at position 3, while the lower peak (AB2) prefers Asn and Asp. The dD and dA clusters are different at position 3, with dD preferring Asn and Asp, which are almost absent from position 3 of the dA turns. The latter contain beta branched, hydrophobic residues at position 3 (Val and Ile).

A census of loops, turns, and 3_{10} helices

With a high-resolution data set and a new set of beta turn types, we were interested in the frequency and distribution of beta turns in protein loops of various lengths. For this purpose, we cannot exclude beta turns based on poor electron density or multiple conformations. For residues in our set of 1,078 proteins with alternative conformations, we selected a primary conformation with the highest occupancy. This dataset is named *CompleteSet* containing 224,918 residues with reported coordinates (Table 1).

To identify loops, we needed to consider how to treat 3_{10} helices since three-residue 3_{10} helices occur frequently within longer loops and might be considered part of the loop rather than significant elements of regular secondary structures. To investigate this, we examined kernel density estimates of the Ramachandran maps of residues in 3_{10} helices of different lengths (Figure 5). A large majority of 3_{10} helices are very short – 81% are length 3 and 12% are

length 4 (Table 3). The kernel density estimates demonstrate an interesting phenomenon. At positions 1, 2, and 3 of the length-3 3_{10} helices, the mode in the φ-ψ density moves progressively

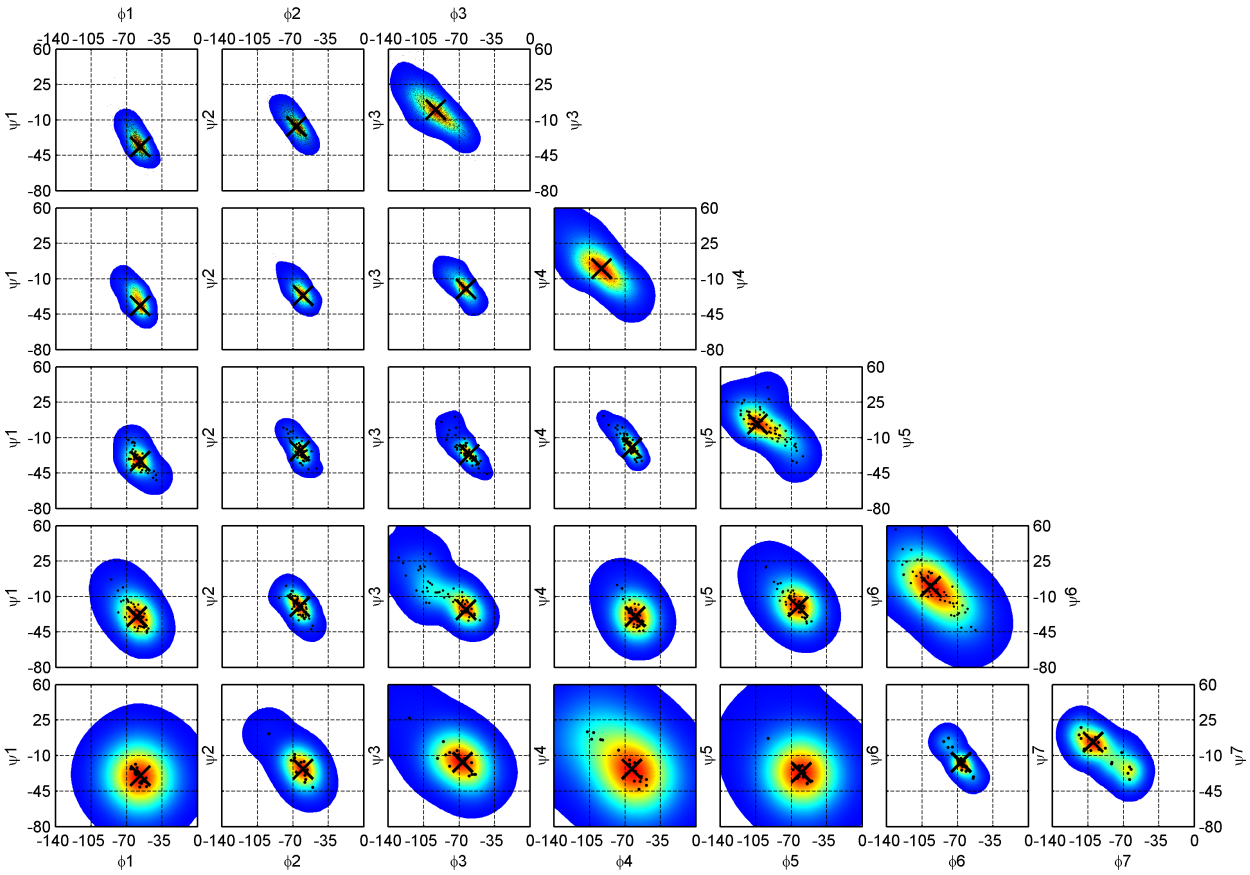


Figure 5. Ramachandran maps for 3-10 helices of increasing length, *n* from 3 (GGG, top row) to 7 (GGGGGGG, bottom row). Optimized kernel density estimates overlaid with (ϕ,ψ) sample scatter plots are drawn for each 3-10 helix residue starting from the first left position, 1 (first column) to the last right position, 7 (last column). Modes (peaks) for each 2D distribution are shown with a large black cross. The first (1) and last (*n*) residue conformations of 3-10 helices are *A* (-64°, -23°) and *D* (-92°, -1°) conformations respectively. The intervening residues are in between, sometimes with conformations spanning both populations. The exact values for (ϕ,ψ) modes of 3-10 helices are reported in Table 3.

upwards and leftwards: φ,ψ = [-57°,-37°], [-67°,-17°], and [-93°,0°]. This is roughly equivalent to two consecutive AD turns with the middle residue in an intermediate position between A and D. The same is true of length-4 3_{10} helices. For the longer 3_{10} helices, the last residue is always

clearly in a delta position [-90°,0°] and the first residue is clearly alpha [-60°,-30°], while the intervening residues are in between, sometimes with points spanning both populations. This is consistent with the early analysis of beta turns by Lewis et al. that defined Type I turns as [ϕ(2),ψ(2) = -60°,-30°; ϕ(3),ψ(3) = -90°,0°] and Type III turns as [ϕ(2),ψ(2) = -60°,-30°; ϕ(3),ψ(3) = -60°,-30°] [11], which were later identified by Richardson as mostly restricted to 3_{10} helices [2]. The pattern we observe is slightly more complicated with the modal value shifting from A to D along the 3_{10} helix. The exact modal (ϕ,ψ) values for 3-10 helices of varying length are reported in Table 3.

Length	Count	Perc.	ϕ1, ψ1	ϕ2, ψ2	ϕ3, ψ3	ϕ4, ψ4	ϕ5, ψ5	ϕ6, ψ6	ϕ7, ψ7
3	2,313	80.7	-57,-37	-67,-17	-93, 0				
4	347	12.1	-57,-37	-60,-27	-63,-20	-93, 0			
5	90	3.1	-57,-33	-63,-23	-63,-23	-63,-20	-103, 3		
6	78	2.7	-60,-30	-67,-17	-63,-23	-60,-30	-63,-20	-97, 0	
7	27	0.9	-57,-30	-60,-23	-67,-17	-63,-23	-60,-27	-67,-17	-100, 3
≥8	12	0.4	-	-	-	-	-	-	-
Total	2,867	100.0							

Table 3. Modal values for backbone dihedral angles of 3_{10} helices of lengths from 3 to 7.

Given these data, we established two criteria with regard to 3_{10} helices for our definition of protein loops. First, short 3_{10} helices often directly abut alpha helices in DSSP, and represent distortions detected in the *i*, *i+3* hydrogen bonding pattern of the alpha helix; therefore we exclude such 3_{10} helices immediately adjacent to alpha helices from loops. For example, a continuous fragment, HHHHGGGG**CCTTCCC**GGGHHHH, has only **CCTTCCC** as a loop while the flanking Gs are still a continuation of an alpha helix region before and after the loop. Second, we treat 3_{10} helices of length 3 not adjacent to alpha helices as loop residues, since they are identified by DSSP with a single internal *i*, *i+3* hydrogen bond within the loop between the residue that precedes the 3_{10} helix and the last residue of the 3_{10} helix. These short 3_{10} helices are common in long loops and may productively be considered part of the loop rather

than an intervening element of regular secondary structure. Indeed, in some definitions of secondary structure, a 3_{10} helix requires two or more hydrogen bonds [6] and these three-residue 3_{10} helices would not be identified as such. For example, **GGG** in HHHHCCCC**GGG**CCCCHHHH is still considered a part of the loop CCCC**GGG**CCCC. We define longer 3_{10} helices (≥4) as elements of secondary structure, which therefore separate their neighboring segments into two loop regions. For instance, **GGGGG** in HHHHCCCC**GGGGG**CTTCHHHH produces two loops, CCCC and CTTC.

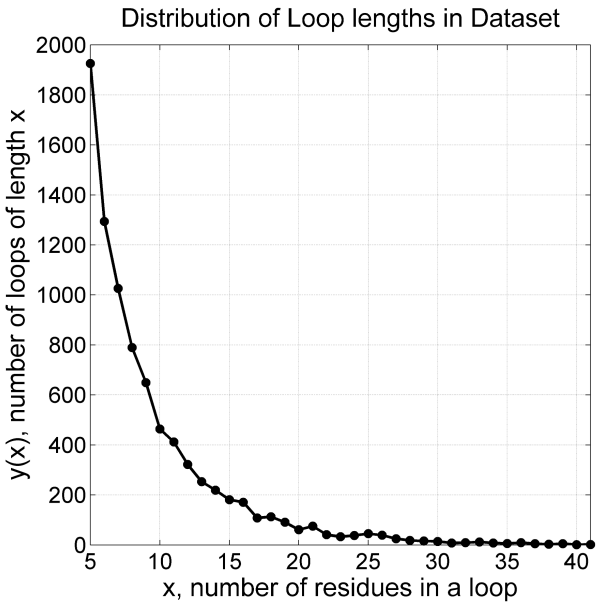


Figure 6. Distribution of loop length in *CompleteSet* of 17,176 loops. The frequency of a loop decreases exponentially after the length of 5 with approximately a 10-fold reduction every time a loop adds 10 more residues. Between loop lengths 2 and 5 (not shown), there is a frequency fluctuation with two peaks at length 2 and 4 observed due primarily to beta turns between two beta sheet strands.

With these definitions, in *CompleteSet* we detected 21,454 turns (65% more than in *RefinedSet*) and assigned them to the closest turn type in Table 2 with the distance measure in the seven-dimensional dihedral angle space. Beta turns comprise 29% of residues in the *CompleteSet* and 63% of the residues in loops between regular secondary structures (Table 1). In *CompleteSet*, we located 17,176 loops of length from 1 to 81. The frequency of a loop

logarithmically decreases with a longer loop; approximately there is a 10-fold reduction every time a loop becomes 10 residues longer (Figure 6).

We also calculated an average number of beta turns identified in each loop of a particular length and estimated the ± one standard deviation confidence range (Figure 7). There are 21,454 beta turns among 102,357 loop residues of any length, producing on average one beta turn per 4.8 residues. If we consider longer loops with 9 or more residues, there are 11,270 beta turns among 49,299 loop residues, leading to approximately the same value of one beta turn per 4.4 loop residues.

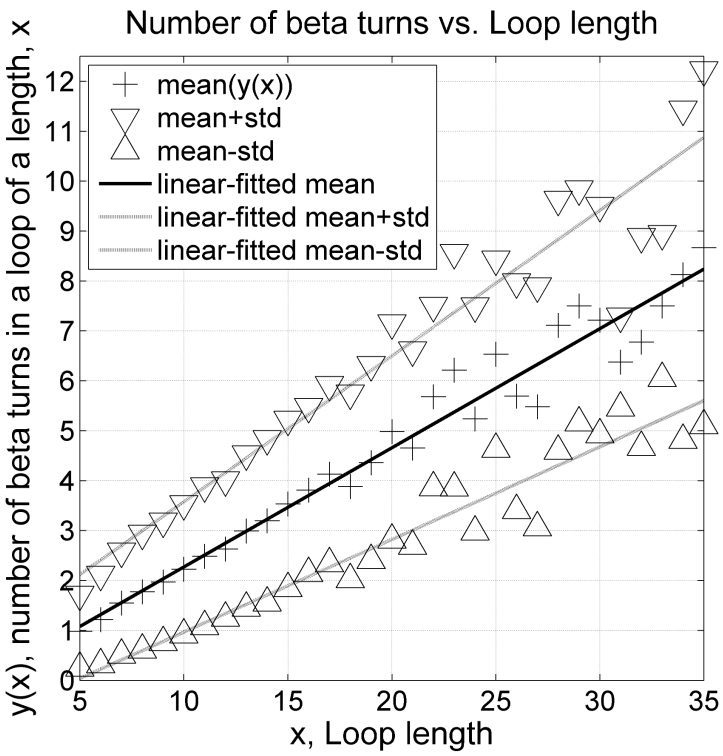


Figure 7. Number of turn residues in a loop as a function of length. For each 4.8 loop residues, there is on average a single beta turn.

A total of 64% (Table 1) of beta turns overlap other beta turns by one or more residues. 3_{10} helices of length 3 plus the preceding residue usually comprise two overlapping beta turns. In addition, other beta turn types can also combine to produce immediately overlapping,

consecutive beta turns. We define double turns as a set of 5 residues with residues 1-2-3-4 and 2-3-4-5 forming individual beta turns. These double turns occur most frequently; we found 5,372 double turns having 82 combinations of our 18 turn types (Table 4). The most common double turn (62%) is AD-AD made of two consecutive AD turns, which are the most prevalent turn type in the data set (49%, Table 2). The 10 most frequent turn type combinations account for 90% of all double turns.

Type numbing	Double turn type	Count	Perc.	Cumulative Perc.
1	AD-AD	3,307	61.6	61.6
2	AD-AZ	470	8.7	70.3
3	AD-AB2	223	4.2	74.5
4	Pa-ad	199	3.7	78.2
5	AD-AB1	137	2.6	80.7
6	AD-AG	121	2.3	82.9
7	pD-AD	119	2.2	85.2
8	Pd_dD	103	1.9	87.1
9	Pd_dA	84	1.6	88.7
10	PcisD_cisDA	76	1.4	90.1
	other	533	9.9	100.0
	All	**5,372**	**100%**	**100%**

Table 4. Statistics for 10 most common double turn type combinations in *CompleteSet*. Double turns overlap by three residues (i.e., 1-2-3-4 and 2-3-4-5).

Type numbing	Triple turn type	Count	Perc.	Cumulative Perc.
1	AD-AD-AD	486	48.6	48.6
2	AD-AD-AZ	143	14.3	62.9
3	AD-AD-AB2	68	6.8	69.7
4	AD-AD-AG	37	3.7	73.4
5	Pd_dA_AD	34	3.4	76.8
6	AD_AD_AB1	29	2.9	79.7
7	PcisD_cisDA_AD	20	2.0	81.7
	other	183	18.3	100.0
	All	**1,000**	**100%**	**100%**

Table 5. Seven most common triple turn types in *CompleteSet* among 1,000 observed total. A triple turn is formed by 6 residues with 1-2-3-4, 2-3-4-5, and 3-4-5-6 positions.

For 1,000 triple turns (Table 5) defined as a stretch of 6 residues with turns at positions

1-2-3-4, 2-3-4-5, and 3-4-5-6, the 7 most frequent combinations out of the observed 65 comprise 82% of triple turns. As for single and double turns, AD-AD-AD is again the most common triple turn, representing half of all cases. Of course turns can overlap by one or two residues (e.g. 1-2-3-4 and 3-4-5-6) but we did not study this separately.

Turn library and software

We have compiled a plain-text library file that contains data on all 18 established beta-turn types (Supplemental Data). For each turn type we provide the following information: the absolute number of dataset points and percentage for each type cluster, the older nomenclature for existing types, our new nomenclature, centroid and modes for dihedral angles pulled from the *Refined* dataset with the following information: <ω(2), ϕ (2), ψ (2), ω(3), ϕ (3), ψ (3), ω(4)>, <aa(1), aa(2), aa(3), aa(4)>, <sec(1), sec(2), sec(3), sec(4)>, res_id(1), chain_id and pdb_id. Our library, *BetaTurnLib18* is freely available to be incorporated into any third-party software for detection, type prediction, and modeling of beta turns in proteins.

In addition, we provide Python software, *BetaTurnTool18* supported on Linux, Mac, and Windows platforms coded with back-compatible Python 2. It reads a PDB-formatted or mmCIF-formatted coordinate file of a protein structure, automatically runs DSSP to assign secondary structure in the input file, and parses our turn library file. With this information available, the Python script detects beta-turn positions and types. Beta-turn location is determined by verifying chain connectivity of each four-residue stretch, satisfying the 7.0 Å Cα(1)-Cα(4) distance constraint calculated from the Cartesian coordinates and checking the secondary structure of the 2nd an 3rd residues assigned by the DSSP program [10]. Since some beta turns are similar to more than one of our beta turn types, the first and second most probable turn types are reported based on the distance to the first and second closest centroids calculated with the same distance metric as used in the clustering. We estimate confidence levels for the first and second

most probable types, which sum up to 1, as follows:

$$Confidence(1st\ type) = 1 - \frac{D(closest\ turn\ type)}{D(closest\ turn\ type) + D(2nd\ closest\ turn\ type)}$$

$$Confidence(2nd\ type) = 1 - \frac{D(2nd\ closest\ turn\ type)}{D(closest\ turn\ type) + D(2nd\ closest\ turn\ type)}$$

Our script prints these confidence levels as integers in the [0, 9] range. 0 and 9 correspond to the [0, 10)% and [90, 100]% confidence intervals respectively. A border-case flag ("+") is returned when the integer confidence level is 6 or below, and a "." (dot) is returned when it 7 or higher. In border cases we recommend considering both turn types for modeling. Our tool accounts for chain breaks in coordinates when it prints a single-character alignment of the complete amino acid sequence, its DSSP secondary structure, turn types, integer confidence levels, and border-case flags. Downloads and complete documentation of our tool or library are available at dunbrack.fccc.edu/betaturn and github.com/sh-maxim/BetaTurn18.

Discussion

From a set of 13,030 beta turns identified in 1,078 proteins of resolution 1.2 Å or better, we have identified 18 beta turn types, compared to the 8 beta turn types in common use (I, I', II, II', VIa1, VIa2, VIb, and VIII). Our nomenclature is based on positions in the Ramachandran map of residues 2 and 3. Six of the existing turn type remain essentially unchanged but have been renamed: Type I is now "AD" and its complement Type I' is now "ad"; Type II' is now "pD"; Types VIa1, VIa2, and VIb with cis peptide bonds at residue 3 are now "BcisP", "PcisP", and "PcisD" respectively. Four of the 10 new turn types are subtypes of the remaining two existing turn types: Type VIII turns have been replaced with 4 subtypes ("AB1", "AB2", "AZ", and "AG") with distinct peaks in the density in the Ramachandran map of residue 3 and distinct residue

preferences at one or more of the 4 turn positions. Type II turns now consist of two subtypes, “Pd” and “Pa”, defined by the presence or absence of Gly at residue 3 respectively.

Finally, there are six completely new turn types. Two new types of turns with left-handed residue 2 and right-handed residue 3 (“dD” and “dA”), and the new complementary turn type with right-handed residue 2 and left-handed residue 3 (“Dd”). There is a new turn type “pG” and two new turn types with a cis peptide bond at position 2: “cisDA” and “cisDP.”

Our decisions in clustering were based on two criteria: keeping the total number of turn types at a minimal practical level for existing and future applications; and justifying new cluster formation either by a significant distance from other formed clusters or uniqueness of a four-residue turn profile. Thus, we abandoned formation of a new cluster if its amino acid turn profile was not unique enough or its conformation spread could be justified by continuous, low-energy torsion angle variation. After running DBSCAN and subdividing some clusters, 2.4% of the turns were left as noise. We conclude that the 18 types cover 98% of turn conformations and that we have not overlooked any major turn conformations. We have also taken a census of beta turns in loops in our data set and determined that there is one turn per 4.8 residues, regardless of loop length. Many turns overlap other turns, such that 63% of residues in loops are in beta turns.

We provide a tab-delimited library file of the new 18 turn types. For each type, we list proposed and existing names, percentage, and geometrical conformation expressed with torsion angles of a real dataset sample. The conformations are given for both a centroid and a sample closest to a mode of each cluster. This library may be used in third-party software for modeling and prediction of beta turns. In addition, we prepared an easy-to-use Python tool (supported on Linux, Mac and Windows platforms) that reads a PDB-formatted or mmCIF formatted coordinate file of a protein structure as input. The tool assigns secondary structure with included redistributable DSSP program, uploads our turn library, and finally locates turns and identifies

their types. The output includes the first and second most probable types with assigned confidence levels for each detected beta turn.

Beta turns in proteins have been studied for 50 years and many classification schemes have been presented. Ultimately a consensus has arisen consisting of 8 turn types: I, I', II, II', VIa1, VIa2, VIb, VIII, and a catch-all noise turn type IV. We have found that some of these turn types represent more than one mode in the density and should be subdivided into further turn types. One of the more interesting results of the new clustering are the turn types that represent combinations of right-handed helical and left-handed helical conformations of residues 2 and 3. We find both combinations: left-right (types dA and dD) and right-left (Dd). These are *not* the same as the type V and V' turns defined by Lewis et al. [11], which had dihedrals of (-80°,+80°) and (+80°,-80°). The dihedral angle modes for turn types dA, dD, and Dd are substantially different from the Type V and V' values (Table 2).

Our new clustering will be useful in determining and refining structures with X-ray crystallography, NMR, and cryo-EM methods. Structure determination, especially at low-resolution, is dependent on templates of common conformations and statistical distributions of backbone and side-chain distributions [32]. Loops are particularly difficult to model at low resolution [33] and our clear definitions of beta turns may be useful in refining these elements of protein structures. Long loops are also difficult to model during the process of structure prediction by comparative modeling [34, 35], and better sampling and modeling of beta turns may prove useful. Finally, the interpretation of missense mutations, either inherited or those that arise somatically in cancer, remains challenging [36]. Structural approaches have contributed to solving this problem, and mutations in beta turns may be particularly disruptive to loop conformations [37]. Given the ubiquitous nature of beta turns, the clustering, nomenclature, and identification script we have developed should provide useful tools for protein structure determination, refinement, and prediction programs as well as useful tools for protein structure,

function, and evolutionary analysis.

Methods

Dataset

A dataset of high-resolution 1,078 protein chains was compiled with PISCES [28, 29]. It consists of PDB chains of length 40 or more amino acids, from X-ray structures with resolution 1.2 Å or better, R-factor of 0.2 or better, and 50% or lower inter-chain sequence identity, which is now determined by local HMM-HMM alignment with HHSearch [38]. Their 8-label DSSP secondary structure was established from a large file holding DSSP secondary structure for the entire PDB which was downloaded from the RCSB website. This raw dataset had 232,865 residues in amino acid sequences, 224,918 residues with known backbone coordinates, and 242,887 (including alternative) conformations. From this raw dataset we compiled two datasets: *CompleteSet* and *RefinedSet*. For *CompleteSet* we extracted residues with at least backbone coordinates for N, Cα, and C atoms and chose a primary conformation for each residue. For a single-conformational residue, there is only one conformation available and it was pulled into *CompleteSet*. For a multi-conformational residues, a primary conformation was pulled into *CompleteSet*. The primary residue conformation has the highest occupancies for its atoms (in PDB files, it is predominantly the '*A'* alternative conformation but not always). *CompleteSet* was used for beta-turn statistical analysis which is representative of the general protein population.

Next we generated *RefinedSet* which is a subset of *CompleteSet*. *RefinedSet* contains a set of higher-quality residues to generate a set of higher-quality beta turns for clustering to identify reliable beta-turn types. The following steps were taken to generate *RefinedSet*: (1) We skipped all residues with any missing backbone, Cβ, and/or γ heavy atoms; (2) we skipped residues with more than one conformation and required the atom occupancy to be 1.0 for the backbone atoms; (3) we combined the DSSP secondary structure information with a set of PDB

coordinates and skipped any residues with undefined DSSP secondary structure. These actions produced *RefinedSet* of a total of 195,752 residues.

We detected beta turns in *RefinedSet* according to the following beta-turn definition. For the clustering procedure, a contiguous four-residue segment is a beta turn if the Cα atoms of the 1st and 4th residues have a distance of ≤ 7.0 Å and the 8-label secondary structure of the 2nd and 3rd residues is different from beta-sheet (E), alpha-helix (H), π-helix (I) and 3_{10}-helix (G) according to DSSP [10].

For the census of beta turns within loops in *CompleteSet*, (1) we counted three-residue 3_{10}-helices within loops (CCCC**GGG**CCCC) as part of the loop and the residues were allowed to be labeled beta turns; (2) we excluded 3_{10}-helices immediately adjacent to alpha helices (HHHHGGG, GGGHHHH, etc.).

Clustering

For beta-turn conformation clustering, we have to define a distance metric – a discriminative measure of how different two beta-turn conformations are. Our beta turn definition is defined by the Cα(1)-Cα(4) distance. Starting from Cα(1), the Cα(4) atom can be built from standard bond lengths and bond angles and seven dihedral angles: ω(2), ϕ(2), ψ(2), ω(3), ϕ(3), ψ(3), ω(4). Inclusion of two flanking torsion angles: ψ(1) and ϕ(4) created too many sub-clusters with no distinct amino acid signatures (not shown).

Since torsion angles are periodic, we used a metric from directional statistics [39], which is equal to the squared chord length between a pair of torsion angle values placed on the unit circle:

$$D(\theta_1,\theta_2) = 2\left(1-\cos\left(\theta_1-\theta_2\right)\right)$$

The factor of 2 is unnecessary but we have included it as we did in earlier work on the clustering

of conformations of complementarity-determining regions of antibodies [25]. For two beta turns, the distance between them is:

$$D = D(\Delta\omega_2) + D(\Delta\phi_2) + D(\Delta\psi_2) + D(\Delta\omega_3) + D(\Delta\phi_3) + D(\Delta\psi_3) + D(\Delta\omega_4)$$

where the differences in angles are taken between the dihedral angles in the two turns.

Clustering of the *RefinedSet* was performed with the DBSCAN algorithm. We manually optimized the *eps* and *minPts* in order to minimize the number of noise points and to produce clusters with [ϕ(2), ψ(2)] and [ϕ(3), ψ(3)] restricted to well known regions of the Ramachandran map for each cluster. This was possible for all the clusters returned by DBSCAN with parameters *eps* = 14° and *minPts* = 25, with the exception of Type I and Type VIII turns. These turns could not readily be separated by DBSCAN since the density for residue 3 is more or less continuous from the alpha to the beta regions of the Ramachandran map. Kernel density estimates of this cluster showed 5 peaks in the density. Three other clusters also showed more than one peak in the density of residue 3. The k-medoids algorithm was applied sequentially to these four clusters in order to find clusters associated with each peak in the [ϕ(3), ψ(3)] density.

For each of resulting 18 clusters, we identified and report both centroid and mode of a cluster density in the seven-dimensional ω(2)-ϕ(2)-ψ(2)-ω(3)-ϕ(3)-ψ(3)-ω(4) dihedral space. The centroid and mode serve as representative conformations for each turn type. We scan a distance matrix of each cluster to find a sample with the most number of neighbors within the smallest radius covering at least 20% of cluster members and at least 20 data points. We approximate a mode with such a turn sample closest to it. The centroid conformation is more sensitive for turn type detection while the mode conformation is more useful in modeling of an unknown protein structure.

Kernel density estimates

We calculated 1D and 2D kernel density estimates (KDE) for a single circular variable (a single torsion angle) and two circular variables (two torsion angles) respectively with a von Mises kernel which is suitable for circular statistics. For example, for the Ramachandran map distribution:

$$\hat{\rho}(\phi,\psi) = \frac{1}{N}\sum_{i=1}^{N} K_h\left(\left\|\phi-\phi_i\right\|\right) K_h\left(\left\|\psi-\psi_i\right\|\right)$$
$$= \frac{1}{4\pi^2 N}\sum_{i=1}^{N} \frac{1}{\left(I_0\left(\kappa\right)\right)^2} \exp\left(\kappa\left(\cos\left(\phi-\phi_i\right)+\cos\left(\psi-\psi_i\right)\right)\right)$$

We utilized a 10-fold maximum-likelihood cross-validation to establish the concentration parameter κ for each kernel density estimate.

Software and Libraries used

Our Python tool for beta-turn type assignment depends on widespread python2 language, numpy [40] and biopython [41] libraries and DSSP software [10, 42]. These are free redistributable software and are easy to download and install. In addition to biopython and numpy libraries we used sklearn [40] and scipy [43] libraries for dataset preparation, residue pruning, and clustering. These libraries allowed PDB parsing, bond and torsion angle calculation, clustering, and scatter plotting. We downloaded electron density maps from European Bioinformatics Institute at www.ebi.ac.uk and ran EDIA software to analyze agreement between turn conformations and their electron density [31].

References

1. C. Venkatachalam. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6**:1425-1436 (1968).
2. J.S. Richardson, *The anatomy and taxonomy of protein structure*, in *Adv. Protein Chem.* 1981, Elsevier. p. 167-339.
3. C. Wilmot, J. Thornton. Analysis and prediction of the different types of β-turn in proteins. *J. Mol. Biol.* **203**:221-232 (1988).
4. C.M. Wilmot, J.M. Thornton. Beta-turns and their distortions: a proposed new nomenclature. *Protein Eng.* **3**:479-93 (1990).
5. O. Koch, G. Klebe. Turns revisited: a uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. *Proteins* **74**:353-67 (2009).
6. K.-C. Chou. Prediction of tight turns and their types in proteins. *Anal. Biochem.* **286**:1-16 (2000).
7. I. Kuntz. Protein folding. *J. Am. Chem. Soc.* **94**:4009-4012 (1972).
8. G.D. Rose. Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature* **272**:586-590 (1978).
9. A.M.C. Marcelino, L.M. Gierasch. Roles of β-turns in protein folding: From peptide models to protein engineering. *Biopolymers* **89**:380-391 (2008).
10. W. Kabsch, C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**:2577-2637 (1983).
11. P.N. Lewis, F.A. Momany, H.A. Scheraga. Energy parameters in polypeptides. VI. Conformational energy analysis of the N-Acetyl N'-methyl amides of the twenty naturally occurring amino acids. *Israeli Journal of Chemistry* **11**:121-152 (1973).
12. N. Panasik, P.J. Fleming, G.D. Rose. Hydrogen-bonded turns in proteins: The case for a recount. *Protein Sci.* **14**:2910-2914 (2005).
13. L.L. Perskie, G.D. Rose. Physical–chemical determinants of coil conformations in globular proteins. *Protein Sci.* **19**:1127-1136 (2010).
14. A.J. Shepherd, D. Gorse, J.M. Thornton. Prediction of the location and type of β-turns in proteins using neural networks. *Protein Sci.* **8**:1045-1055 (1999).
15. H. Kaur, G.P.S. Raghava. Prediction of β-turns in proteins from multiple alignment using neural network. *Protein Sci.* **12**:627-634 (2003).
16. P. Kountouris, J.D. Hirst. Predicting β-turns and their types using predicted backbone dihedral angles and secondary structures. *BMC Bioinformatics* **11**:407 (2010).
17. M. Meissner, O. Koch, G. Klebe, G. Schneider. Prediction of turn types in protein structure by machine-learning classifiers. *Proteins* **74**:344-52 (2009).
18. E.G. Hutchinson, J.M. Thornton. A revised set of potentials for beta-turn formation in proteins. *Protein Sci.* **3**:2207-16 (1994).
19. P.F. Fuchs, A.J. Alix. High accuracy prediction of β-turns and their types using propensities and multiple alignments. *Proteins: Structure, Function, and Bioinformatics* **59**:828-839 (2005).
20. A. Bornot, A.G. de Brevern. Protein beta-turn assignments. *Bioinformation* **1**:153 (2006).
21. A.G. de Brevern. Extension of the classical classification of β-turns. *Scientific reports* **6**:33191 (2016).
22. K. Guruprasad, M. Prasad, G. Kumar. Analysis of γβ, βγ, γγ, ββ multiple turns in proteins. *Chem. Biol. Drug Des.* **56**:250-263 (2000).

23. E.C. Ohage, W. Graml, M.M. Walter, S. Steinbacher, B. Steipe. Beta-turn propensities as paradigms for the analysis of structural motifs to engineer protein stability. *Protein Sci.* **6**:233-41 (1997).
24. D. Ting, G. Wang, M. Shapovalov, R. Mitra, M.I. Jordan, R.L. Dunbrack, Jr. Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLOS Comput. Biol.* **6**:e1000763 (2010).
25. B. North, A. Lehmann, R.L. Dunbrack, Jr. A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* **406**:228-56 (2011).
26. M. Ester, H.-P. Kriegel, J. Sander, X. Xu. *A density-based algorithm for discovering clusters in large spatial databases with noise*. in *Kdd*. 1996.
27. S.A. Hollingsworth, P.A. Karplus. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomolecular concepts* **1**:271-283 (2010).
28. G. Wang, R.L. Dunbrack, Jr. PISCES: a protein sequence culling server. *Bioinformatics* **19**:1589-91 (2003).
29. G. Wang, R.L. Dunbrack, Jr. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **33**:W94-8 (2005).
30. S.A. Hollingsworth, M.C. Lewis, D.S. Berkholz, W.-K. Wong, P.A. Karplus. (phi,psi)(2) Motifs: A Purely Conformation-Based Fine-Grained Enumeration of Protein Parts at the Two-Residue Level. *J. Mol. Biol.* **416**:78-93 (2012).
31. A. Meyder, E. Nittinger, G. Lange, R. Klein, M. Rarey. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. *Journal of chemical information and modeling* **57**:2437-2447 (2017).
32. P.V. Afonine, B.K. Poon, R.J. Read, O.V. Sobolev, T.C. Terwilliger, A. Urzhumtsev, P.D. Adams. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallographica Section D: Structural Biology* **74** (2018).
33. M.V. Petoukhov, N.A. Eady, K.A. Brown, D.I. Svergun. Addition of missing loops and domains to protein models by x-ray solution scattering. *Biophys. J.* **83**:3113-3125 (2002).
34. H. Park, G.R. Lee, L. Heo, C. Seok. Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments. *PLoS One* **9**:e113811 (2014).
35. K. Tang, J. Zhang, J. Liang. Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth Monte Carlo method. *PLOS Comput. Biol.* **10**:e1003539 (2014).
36. L.G. Martelotto, C.K. Ng, M.R. De Filippo, Y. Zhang, S. Piscuoglio, R.S. Lim, R. Shen, L. Norton, J.S. Reis-Filho, B. Weigelt. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome biology* **15**:484 (2014).
37. T.G. Kucukkal, M. Petukh, L. Li, E. Alexov. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Curr. Opin. Struct. Biol.* **32**:18-24 (2015).
38. J. Söding. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**:951-60 (2005).
39. K.V. Mardia, P.E. Jupp, *Directional Statistics*. Wiley Series in Probability and Statistics. 2000, London: Wiley.
40. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg. Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**:2825-2830 (2011).
41. P.J. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**:1422-1423 (2009).

42. W.G. Touw, C. Baakman, J. Black, T.A. te Beek, E. Krieger, R.P. Joosten, G. Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**:D364-D368 (2014).
43. E. Jones, T. Oliphant, P. Peterson. SciPy: Open source scientific tools for Python. *URL http://www.scipy.org*