Journal of Computer-Aided Molecular Design (2005) 19: 551–566
DOI 10.1007/s10822-005-9006-2

Topological side-chain classification of β-turns: Ideal motifs for peptidomimetic development

Tran Trung Tran[a], Jim McKie[b], Wim D.F. Meutermans[b], Gregory T. Bourne[a,b], Peter R. Andrews[b] & Mark L. Smythe[a,b,*]
[a]*Protagonist Pty Ltd, Level 7, Queensland Bioscience Precinct, 306 Carmody Road, 4072, Brisbane, St Lucia, Australia;* [b]*Institute for Molecular Bioscience, The University of Queensland, 4072, Brisbane, Australia*

Received 20 April 2005; accepted 27 July 2005

Key words: β-turns, β-turn classification, β-turn mimetics, cluster, drug design, peptide

Summary

β-turns are important topological motifs for biological recognition of proteins and peptides. Organic molecules that sample the side chain positions of β-turns have shown broad binding capacity to multiple different receptors, for example benzodiazepines. β-turns have traditionally been classified into various types based on the backbone dihedral angles ($\phi 2$, $\psi 2$, $\phi 3$ and $\psi 3$). Indeed, 57–68% of β-turns are currently classified into 8 different backbone families (Type I, Type II, Type I′, Type II′, Type VIII, Type VIa1, Type VIa2 and Type VIb and Type IV which represents unclassified β-turns). Although this classification of β-turns has been useful, the resulting β-turn types are not ideal for the design of β-turn mimetics as they do not reflect topological features of the recognition elements, the side chains. To overcome this, we have extracted β-turns from a data set of non-homologous and high-resolution protein crystal structures. The side chain positions, as defined by C_α–C_β vectors, of these turns have been clustered using the *k*th nearest neighbor clustering and filtered nearest centroid sorting algorithms. Nine clusters were obtained that cluster 90% of the data, and the average intra-cluster RMSD of the four C_α–C_β vectors is 0.36. The nine clusters therefore represent the topology of the side chain scaffold architecture of the vast majority of β-turns. The mean structures of the nine clusters are useful for the development of β-turn mimetics and as biological descriptors for focusing combinatorial chemistry towards biologically relevant topological space.

Introduction

β-turns [1–4] are a preferred recognition motif of peptides and proteins. Examples of turns as recognition motifs can be found in high resolution crystal structures of antibody-peptide complexes [5–7] and from structure activity studies of many peptide hormones, such as angiotensin II, [8, 9] bradykinin, [10], choleocystokinin [11], gonadotrophin releasing hormone [12], and somatostatin, [13, 14] to name a few. β-turns are often conserved during evolution and are considered as initiation sites for protein folding [1]. They therefore describe biologically-relevant regions of protein and peptide structural space [1, 15–17].

Ripka et al. [18] has defined three classes of peptidomimetics. Class I mimetics often match the amide bond backbone and Class II mimetics do not necessarily mimic the structure of the parent peptide. Class III compounds are based on replacing the amide backbone of peptides by other templates or scaffolds. Examples of scaffolds are outlined in Suat Kee et al. [19] and include, but are not limited to, heterocycles, [20, 21] carbohydrates,

*To whom correspondence should be addressed. Phone: 61-7-33462975; Fax: 61-7-33462379; E-mail: m.smythe@imb.uq.edu. au

[22, 23] and pentaazacrowns [24]. This approach is derived from the premise that the chemical scaffold projects the functional units (the side chains) in similar topologies to the parent amide bonded scaffold in a β-turn conformation.

Benzodiazepines are a classic example of Class III peptidomimetics and are a protypical privileged substructure [25], a class of molecule with broad binding capacity to different unrelated receptors. They have been reported to structurally match the four C_α–C_β vectors of common β-turn types [26], and it has been suggested that this results in their observed broad binding capacity [25, 27]. Other examples of privileged substructures that match turn motifs have also been reported [28].

As previously mentioned, β-turns are an important recognition motif of proteins and peptides. Despite the importance of side chain spatial arrangements in molecular recognition, β-turns are currently classified into nine types based on the main chain dihedral angles, $\phi 2$, $\psi 2$, $\phi 3$ and $\psi 3$ [2, 29–32]. Although this classification of β-turns has been extremely useful and has been used widely to design peptidomimetics [15, 33–43], it makes very little functional sense, as it is not side-chain based. In addition, and as we will illustrate, each type of β-turn in the current classification can have two or more clusters of side chain spatial arrangements and different types of "classical" β-turns can have the same side chain spatial arrangement.

There have been two reports on the further classification of β-turns using additional characteristics or descriptors [44]. Whilst the β descriptor of Ball et al. [45] defines some global structural characteristics of the turns, it is clearly an oversimplification, as it only considers two of the possible four side chain positions and used only a small data set of 154 experimentally derived β-turns. Garland and Dean [46, 47] have clustered β-turn motifs using four different descriptors, c_α atom doublets, C_α atom triplets, C_α–C_β vector doublets and C_α–C_β vector triplets. The clustering was not based on experimental data, but was based on all possible permutations of selecting doublets or triplets out of each of the existing eleven idealized β-turn types.

This paper aims to cluster the functionally relevant side-chain positions of β-turns for the development of Class III peptidomimetics. This is in contrast with traditional classification of β-turns that is based on the backbone dihedral angles, ϕ, ψ. The classification system defines the biologically relevant regions of β-turn structural space, which is an ideal starting point for library focusing and the design and development of β-turn mimetics. Molecules that match important functional topologies may show broad binding capacities similar to privileged substructures.

Method

β-turn clustering

Extraction of β-turns from Protein Data Bank (PDB)

A high resolution and non-redundant database of β-turns are required for the determination of common β-turn motifs that exist in proteins. To ensure high quality data, only high-resolution structures with a resolution of ≤ 2 Å and an R factor $\leq 20\%$ were extracted from the Protein Data Bank [48]. Furthermore, to eliminate the biased sampling in the PDB caused by the presence of multiple structures that are minor variations of a particular protein chain, only non-homologous protein chains with $\leq 25\%$ homology with other protein chains were used. The distribution of the $C_{\alpha 1}$–$C_{\alpha 4}$ distances of the resulting 3984 four-residue segments that are β-turn like (having $C_{\alpha 1}$–$C_{\alpha 4}$ distance of ≤ 7 Å and not having consecutive $C_{\alpha 1}$–$C_{\alpha 4} < 5.5$ Å [49]) is plotted in Figure 1 and a major peak is observed at $C_{\alpha 1}$–$C_{\alpha 4}$ distances of 5.5 Å. In 1973, Lewis [31] concluded that β-turns have $C_{\alpha 1}$–$C_{\alpha 4}$ distance of ≤ 7 Å. Based on the distribution of the $C_{\alpha 1}$–$C_{\alpha 4}$ distances of only eight X-ray diffraction determined structures. To remove any possible biases caused by noisy data, the outliners of the major peak at 5.5 Å were removed by eliminating the turns with $C_{\alpha 1}$–$C_{\alpha 4}$ distance of less than 5 Å or greater than 6.2 Å, resulting in 2675 β-turns in the database.

Representation of data

Our motivations for clustering using C_α–C_β vectors are several fold. The C_α–C_β vector describes the initiation of the side chain geometry, and is well defined experimentally as it is anchored to the backbone. This is in contrast to the more flexible penultimate side chain atoms. Importantly, most mimetic strategies involve anchoring C_α–C_β bonds to a non-peptidic scaffold, the extra atoms of the

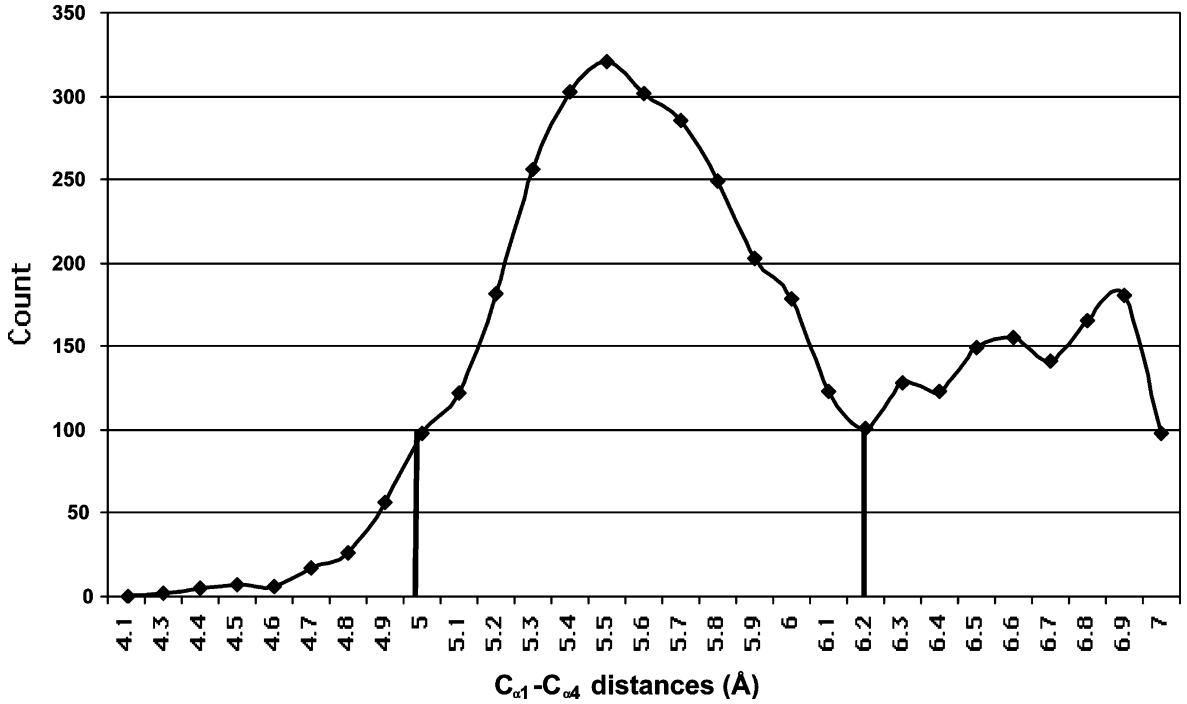

Figure 1. Distribution of $C_{\alpha 1}$–$C_{\alpha 4}$ distances of all four residues segments that are not helical nor β-sheets and that are found in high resolution and non-homologous structures in the Protein Data Bank [48].

side chain providing a degree of flexibility in molecular recognition. We therefore consider that clustering according to C_α–C_β vectors is functionally significant, when the aim is to use the motifs identified to direct peptidomimetic development.

Each of the 20 naturally-occurring amino acids, except for glycine, possesses a C_α–C_β vector due to the covalent bond between the central α carbon and the β carbon of the side chain. For β-turns that contain a glycine, the glycine residue was mutated to alanine to generate the required C_α–C_β vector. This was achieved by superimposing an ideal alanine structure onto the *n*, C_α and C′ atoms of the glycine residue. The proportion of Glycine in the four residues comprising the selected 2675 β-turns are 8%, 7%, 25% and 15%, respectively, which is quite consistent with the 12%, 10%, 20% and 17% obtained by Chou [49] from 29 proteins in 1977.

An important advance in database searching has been made by representing 3D structures in terms of the relationship between atoms located in distance space, rather than Cartesian space [50, 51]. A location in distance space is defined by distances between atoms, expressed in the form of a distance matrix. Distance matrices are therefore coordinate independent, and comparisons between distance matrices can be made without restriction to a particular frame of reference, such as is required using Cartesian coordinates. It is important to emphasize that an arrangement of atoms and its mirror image are described by identical distance matrices. A root mean squared deviation (RMSD) can be used to alleviate this ambiguity. The four C_α–C_β vectors of each β-turn are represented by a distance matrix rather than a Cartesian coordinate system. Since there are four pairs of distances between each pair of C_α–C_β vectors ($C_{\alpha 1}$–$C_{\alpha 2}$, $C_{\alpha 1}$–$C_{\beta 2}$, $C_{\beta 1}$–$C_{\alpha 2}$ and $C_{\beta 1}$–$C_{\beta 2}$) and there are six possible pairs of C_α–C_β vectors (1–2, 1–3, 1–4, 2–3, 2–4 and 3–4), then 24 distances are required to represent the 3D topography of a β-turn. The distances between $C_{\alpha i}$ and $C_{\beta i}$ were not included because these bonded distances are relatively invariant between β-turns when compared to the non-bonded distances used.

kth-nearest neighbor

The *k*th-nearest neighbor clustering algorithm [52, 53] employed here for clustering of β-turns

is basically a simple-linkage clustering algorithm [54] in which every member is initially assigned to a different cluster and clusters are subsequently merged if the minimum distance between a member of a cluster and a member of another cluster is less than some threshold. The *k*th-nearest neighbor clustering algorithm [52, 53] differs from simple-linkage clustering algorithm in that the distance between members is replaced by a dissimilarity measure defined below.

$d_k(x)$ is defined as the Euclidian distance from observation x to the kth nearest observation. $v_k(x)$ is defined as the volume enclosed by the sphere, centering at observation x and having a radius of $d_k(x)$. The scaled density at observation x, $f(x)$, is defined as $k/(N^*v_k(x))$ where N is the total number of observations. The dissimilarity measure between observations x_i and x_j, $D(x_i, x_j)$, can be calculated from the following definitions. First, x_i and x_j are said to be adjacent if the Euclidean distance between the two points is less than $d_k(x_i)$ or $d_k(x_j)$. If the observation x_i and observation x_j are not adjacent, then the dissimilarity measure, $D(x_i, x_j)$, is set to infinity. Otherwise, $D(x_i, x_j)$ is defined as the average of the inverse of the scaled density, i.e. $D(x_i, x_j) = (1/f(x_i) + 1/f(x_j))/2$. Clustering should group together regions of high density separated by regions of low density. Effectively, by defining the dissimilarity measure as the inverse of the density, this algorithm first groups together adjacent points or clusters that have high-density.

The *k*th-nearest neighbor algorithm from the commercially available SAS/STAT program [55] was used to cluster the distance matrices representing the topography of the C_α–C_β vectors of the β-turns. The option '*k*' is called the smoothing parameter. A small value of '*k*' produces jagged density estimates and large numbers of clusters, and a large value of '*k*' produces smooth density estimates and fewer clusters. A '*k*' value of two was used because, only a rough estimate of clusters is required here. The clusters obtained here are used as initial seeds for the filtered nearest centroid sorting algorithm described below.

Filtered nearest centroid sorting clustering algorithm

The nearest centroid sorting clustering algorithm by Forgy [56, 57] requires a prior estimate of some initial seeds. The algorithm assigns each observation to the nearest initial seed to form temporary clusters. The seeds are then replaced by the means of the temporary clusters and the process is repeated until no further changes occur in the clusters. After the *k*th nearest neighbor clustering of the β-turns, a modified form of the 'nearest centroid sorting algorithm' [56], filtered nearest centroid sorting clustering algorithm was used to refine the clustering. This method superimposes observations in Cartesian coordinate space and thus removes the mirror image problem inherited from the distance matrix representation in the *k*th nearest mean clustering algorithm. The reasons for this two-stage clustering process are: (1) Hierarchical clustering based on RMSD could not be used in the first place because, the number of observations is larger than the limit set by the SAS/STAT program [55] and (2) Faster and leaner approximation methods, such as nearest centroid sorting [56, 57] or *K*-means clustering algorithm [58] could not be used without prior estimate of initial seeds or the number of initial seeds by some.

The filtered nearest centroid sorting algorithm is basically the same as the nearest centroid sorting algorithm except that if the minimum RMSD of a β-turn to the mean structures is above some definable threshold, the turn is considered too remote and therefore not assigned to the temporary clusters. In latter iterations, these unassigned turns are superimposed onto the new mean structures of the new temporary clusters and if the minimum RMSD is below the threshold, then they are assigned to the new temporary cluster. The aim of this filtering is to remove the turns that are very different from the mean structures, so as not to bias the mean. Furthermore, 100% of the β-turns do not need to be clustered; only a major proportion is required. The filtered nearest centroid sorting algorithm was implemented in a C++ program entitled "fncsa_cluster_analysis.cpp".

Cluster analysis

Vector plots of β-turns

It is difficult to visualize the 24 distances that represent the topography of a β-turn. A vector plot is used to aid the visualization by approximating the 24 distances with four torsional angles $\theta1$, $\theta2$, $\theta3$ and $\theta4$ (see Figure 2). The $\theta1$ is defined as torsional angle between $C_{\beta1}$, $C_{\alpha1}$, $C_{\alpha2}$ and $C_{\beta2}$, $\theta2$ is defined as the torsional angle between $C_{\beta2}$,

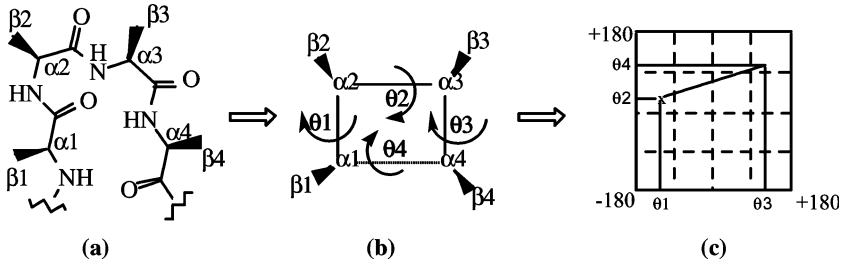

Figure 2. (a) Each β-turn is represented by four C_α–C_β vectors highlighted by the dark triangle. (b) To aid visualization of the spatial arrangement of the turn after clustering, the four torsional angles $\theta1$, $\theta2$, $\theta3$ and $\theta4$ are used as approximation to the 24 distances. (c) The four torsional angles are plotted as a vector from ($\theta1$, $\theta2$) (represented by the symbol 'x') to ($\theta3$, $\theta4$).

$C_{\alpha2}$, $C_{\alpha3}$ and $C_{\beta3}$; $\theta3$ is defined as the torsional angle between $C_{\beta3}$, $C_{\alpha3}$ $C_{\alpha4}$ and $C_{\beta4}$; and $\theta4$ is defined as the torsional angle between $C_{\beta1}$, $C_{\alpha1}$ $C_{\alpha4}$ and $C_{\beta4}$. Since the distances between adjacent C_α atoms in a β-turn are relatively constant due to the nature of the peptide bond, the four torsional angles should represent the essential conformational feature of a β-turn. The four torsional angles are plotted as a vector from ($\theta1$, $\theta2$) (represented by the symbol 'x') to ($\theta3$, $\theta4$). Effectively, this plot approximates the 24 distances of β-turns to four torsional angles ($\theta1$, $\theta2$, $\theta3$ and $\theta4$), which are plotted as a vector on a 2D graph. The torsional angles are periodic. A value of x is equivalent to $x-360$, $x+360$ and so on. To remove the graphing problem associated with the periodic nature of the torsional angles, each torsional value is transformed into a period, which is closest to the torsional angles of the first β-turn.

Visualizing β-turn clusters

Another method to visualize the clusters of β-turns is to superimpose the 3D structures of all the turns in a cluster. Superimposition is performed from the four C_α–C_β vectors of a β-turn to the four C_α–C_β vectors of the mean structure of the cluster. For glycine, the C_α–C_β vector is obtained by superimposing a standard alanine residue to the n, C_α and C′ atoms of the backbone of the glycine residue. The "fncsa_cluster_analysis.cpp" program outputs the coordinates of the superimposed structures in a multi-structure pdb file format, which is visualised using the program InsightII of Molecular Simulation Inc [59].

There are a few steps in the calculation of the mean of a cluster in Cartesian coordinate space. Firstly, an initial mean structure for a cluster is set to be the first β-turn that does not have glycine or proline residue. Then, each β-turn is superimposed to this temporary mean structure based on the coordinates of the C_α–C_β vectors. After the superimposition, a new temporary mean structure is computed by averaging the x, y and z coordinates. The latter two steps are repeated until successive mean structures differ by less than some arbitrary threshold.

Calculation of the RMSD matrix of all the clusters

RMSD matrix is calculated to examine the performance of the clustering by assessing the dissimilarity within and between clusters. Each cluster is compared with every other cluster so that the row and column number of the matrix represents the cluster number. The value in a cell at row x and column y represents the mean RMSD when all the β-turns in cluster x is superimposed to the mean structure of cluster y. The diagonal of the matrix with row x and column x represents intra-cluster RMSD while the other cells represents inter-cluster RMSD. Values in row x and column y are not necessarily similar to values in row y and column x because, the former represent the mean RMSD of the β-turns in cluster x superimposed onto the mean structure of cluster y and the latter represent the mean RMSD of the β-turns in cluster y superimposed onto the mean structure of cluster x. However, the two numbers are very similar.

Analysis of whether the clusters are overlapping or distinct

Do the resulting clusters represent overlapping-variation from a continuous spread or do they

represent distinct clusters that do not overlap in hyperspace? To answer this question, we defined that two distributions are 'distinct' if 90 or greater percentage of the data does not overlap in some dimension. There can be *n*th dimension (*nD*) where *n* is an integer greater or equal to 1. Based on this definition, a program named 'evaluate_overlap.cpp' was written which input all the conformations of each cluster and determines whether each cluster is distinct or overlapped with other clusters in 1D. The single dimensions examined were the six torsional angles between the four C_α–C_β vectors in a β-turn (*T*0–*T*5), 24 nonbonded distances (*D*0–*D*23) and six ϕ, ψ torsional angles (Φ1, Φ2, Ψ2, Φ3, Ψ3, Φ4). Distinctions basing on combinations of these dimensions, forming 2 or more dimensions, were not computed.

The maximum and minimum values, which delineate the most frequent 90% of each distribution, were not computed using the mean plus and minus some standard deviation because, some of the spreads were not binomial distributions. The maximum and minimum was computed by first 'binning' the data with respect to each of the 36 dimensions mentioned earlier. Then, the bins for each dimension are sorted from most frequent to least frequent. As the program traverse down the bins, the maximum and minimum of the dimension is stored. The traversal is stopped when the program has traversed through at least 90% of the data. Consequently, the stored maximum and minimum represent the maximum and minimum that delineates the top 90% of the distribution. This method of finding maximum and minimum works well with single peak distributions. For distributions with two or more peaks, the spread covered by the maximum and minimum are over-estimated. In such case (2% of the final analysis), the determination of overlap or distinct is performed by visual inspection of the distributions.

For a particular dimension, if the maximum and minimum of cluster *X* overlap with those of cluster *Y*, then cluster *X* and cluster *Y* is considered to be overlapping in that dimension. However, if the maximum and minimum of cluster *X* does not overlap with the maximum and minimum of cluster *Y*, then cluster *X* is considered to be distinct from cluster *Y* in the dimension.

Results

Clustering

The *k*th nearest neighbor cluster algorithm was used to cluster the 2675 β-turns in the database. The mean structure (seed) of each of the outputted 570 clusters was calculated by averaging each of the 24 distances representing the topography of β-turns. In the second cycle, *k*th nearest clustering was performed on these 570 seeds and 117 seeds were obtained. The third cycle of *k*th nearest clustering produced 25 seeds and the fourth cycle produced 7 seeds. A fifth cycle is expected to reduce the 7 seeds to 2 or 3 seeds. Consequently, both the 7 and 25 seeds were examined in more detail prior to the selection of final β-turn clusters.

To determine a reasonable value for the threshold used in the filtered nearest centroid sorting algorithm, the seven seeds obtained from the *k*th nearest neighbor clustering were refined using filtered centroid sorting algorithm with four different threshold values, 0.6, 0.65, 0.7 and infinity (no threshold at all). The results show that the lower the threshold, the higher the percentage of β-turns which are not assigned to the clusters, i.e. rejected. The percentages of rejection for the four threshold values are 19%, 14%, 8% and 0%, respectively. The RMSD matrices of the results were calculated and the averages of the mean inter-cluster RMSD are 1.05, 0.95, 1.03 and 1.15, respectively. Ideally one would like clusters to differ as much as possible, and hence have a high inter-cluster RMSD. The mean inter-cluster RMSD was lowered in going from a threshold of infinity to 0.7 and to 0.65, however it got higher in going from 0.65 to 0.6. The average of the mean intra-cluster RMSD are 0.36, 0.36, 0.40 and 0.44, respectively. In this instance, low intra-cluster RMSD is favored, therefore emphasizing that the observations in each cluster are similar. There were improvements in the intra-cluster RMSD in going from a threshold of infinity to 0.7 and from 0.7 to 0.65. However, there was no improvement in going from a threshold of 0.65–0.6. As a compromise of the conflicting interest of percentage rejection, inter-cluster RMSD and intra-cluster RMSD, a filter threshold of 0.65 was chosen.

To determine if the 25 seeds from the third cycle or the 7 seeds from the fourth cycle of the *k*th nearest neighbor clustering best represent the side

chain spatial arrangements of β-turns, both the results were subjected to the filtered centroid sorting algorithm followed by the calculation of the RMSD matrix. The RMSD matrix for the 7-clusters is shown in Table 1 and the RMSD matrix for the 25-clusters can be obtained by contacting the authors. Clustering processes aim to find clusters, which have low intra-cluster RMSD separated by high inter-cluster RMSD. For the 25 clusters, the average of the mean intra-cluster RMSD is 0.31, the average of the mean inter-cluster RMSD is 1.11 and the maximum mean intra-cluster RMSD is 0.42. For the 7 clusters, the average of the mean intra-cluster RMSD is 0.36, the average of the mean inter-cluster RMSD is 0.95 and the maximum mean intra-cluster RMSD is slightly higher, 0.49. The results show that the clustering into the 7 clusters is not as good as the clustering into the 25 clusters, the intra-cluster RMSD was larger (0.36 compared to 0.31) and the inter-cluster was smaller (0.95 compared to 1.11). However, since this is not a drastic difference and the 7-clusters still give reasonable intra-cluster RMSD, the more tractable 7-clusters result was preferred over the 25-clusters result.

Table 1. RMSD matrix of the results obtained from the filtered centroid sorting refinement of the 7 clusters formed from the fourth cycle of the *k*th-nearest neighbor clustering algorithm.

Cluster	1	2	3	4	5	6	7
1	**0.32**	0.70	0.78	1.10	0.71	0.84	1.39
2	0.70	**0.31**	0.65	0.60	0.63	1.09	0.93
3	0.85	0.75	**0.49**	0.86	0.79	0.99	1.10
4	1.12	0.64	0.81	**0.38**	0.82	1.47	0.67
5	0.72	0.66	0.72	0.81	**0.35**	1.15	0.89
6	0.84	1.09	0.93	1.46	1.14	**0.31**	1.69
7	1.40	0.95	1.05	0.67	0.90	1.70	**0.38**

Refinement of the clustering

Vector graphs, as described in the method section, were used to visualise the seven cluster result (Figure 3). The figure shows that all the clusters except for cluster three have a reasonable uniform distribution from a single mode. Cluster three, however, seems to have two modes; one mode with $\theta 4 \geq 70°$ and the other mode with $\theta 4 < 70°$. Furthermore, the RMSD matrix in Table 1 shows that cluster three has the most varied intra-cluster RMSD of 0.49. To determine if cluster three should remain as one cluster or should be divided into two clusters, a practical step was used in which cluster three was divided into two clusters (one cluster with $\theta 4 \geq 70°$ and another cluster with $\theta 4 < 70°$) and the new result assessed by comparison with the original result. The resulting eight clusters were refined once more using the filtered nearest centroid sorting algorithm. The RMSD matrix and the vector plot for the new eight clusters were calculated and the results are shown in Table 2 and Figure 4, respectively. In dividing cluster three into two clusters, the mean intra-cluster RMSD has not changed significantly (from 0.36 to 0.35), the maximum intra-cluster RMSD improved from 0.49 to 0.43, the minimum intra-cluster RMSD improved from 0.31 to 0.21, the

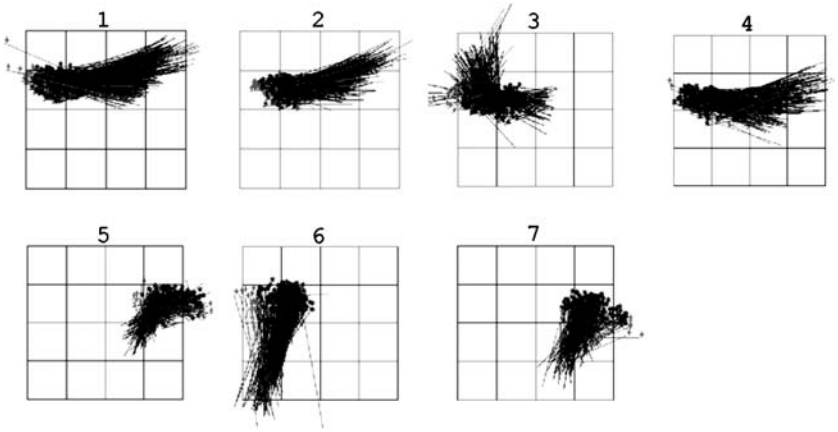

Figure 3. Vector plot of the seven clusters obtained from the *k*th nearest neighbor cluster and the filtered nearest centroid sorting algorithms. A threshold of 0.65 RMSD was used.

Table 2. RMSD matrix of the eight clusters formed from clustering algorithm and explicit division of cluster three into two clusters.

Cluster	1	2	3	4	5	6	7	8
1	0.32	0.70	0.87	1.09	0.71	0.84	1.39	0.80
2	0.69	0.29	0.70	0.59	0.64	1.09	0.94	0.71
3	0.89	0.74	0.39	0.81	0.84	1.06	1.06	0.75
4	1.11	0.62	0.80	0.36	0.80	1.47	0.66	0.87
5	0.71	0.66	0.82	0.80	0.34	1.16	0.89	0.72
6	0.84	1.09	1.04	1.46	1.15	0.31	1.69	0.94
7	1.40	0.96	1.05	0.67	0.90	1.70	0.38	1.09
8	0.84	0.77	0.77	0.89	0.76	0.97	1.10	0.43

mean intra-cluster has improved from 0.95 to 1.25 and finally the percentage of turns represented by the clusters remained the same at 86%. The vector plot in Figure 4 shows that the β-turns in each cluster distribute within a narrow range about a single mode. These results suggested that the eight clusters system is a better representation of β-turn motifs compared to the seven clusters system.

It was observed that type I′ β-turns were not included in any of the eight clusters, they were rejected in the filtered nearest centroid sorting clustering because, their RMSD with the mean of the eight clusters were more than the threshold of 0.65. This reflects a weakness in the algorithm of using kth nearest mean to cluster the mean clusters, which results in no seed near the low frequency type I′ β-turns. Since Figure 5 indicates that there is a cluster near the type I′ β-turns, the mean of the type I′ was calculated and the result was included together with the other eight initial seeds for the filtered nearest centroid sorting clustering. The RMSD matrix and the vector plot for the new nine clusters were calculated and the results are shown in Table 3 and Figure 6, respectively. In addition of the type I′ average structure into the initial seeds, the mean intra-cluster RMSD has not changed significantly (from 0.35 to 0.36), the minimum and maximum intra-cluster RMSD remained the same, the mean inter-cluster worsen (from 1.25 to 1.1) and the percentage of β-turns classified improved from 86% to 90%. The vector plots in Figure 6 shows that the β-turns in each cluster distribute within a narrow range about a single mode. These results suggest that the nine clusters system is a reasonable representation of β-turn motifs.

Mean structures

The final nine cluster result was also visualized by superimposing each β-turn in the clusters onto the clusters' mean structure (Figure 7). The visual result is consistent with the mean intra-cluster RMSD value of each cluster in Table 3. The cluster with the least amount of C_α–C_β vector spread (cluster 2, Figure 7) corresponds to the smallest mean intra-cluster RMSD. It is interesting to note that the backbone structure can vary significantly although the C_α–C_β vectors are uniform within a cluster. A top view of the most uniform cluster, cluster 2 for example, shows that different backbone conformations can have similar

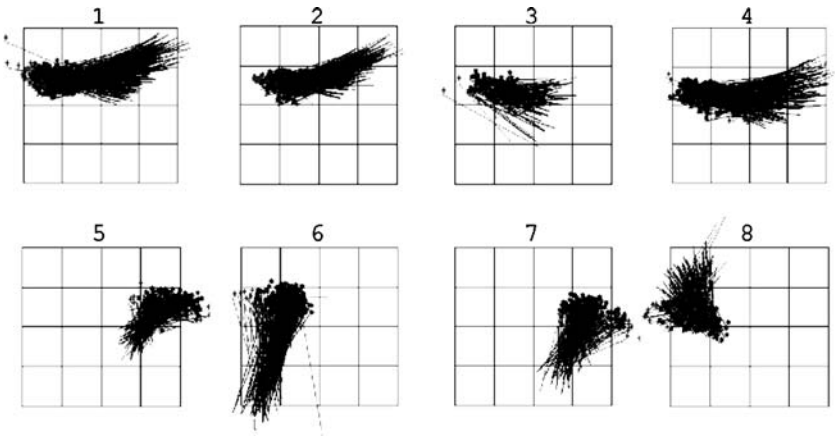

Figure 4. Vector plots of the eight clusters formed from the clustering algorithm and explicit division of cluster three into two clusters.

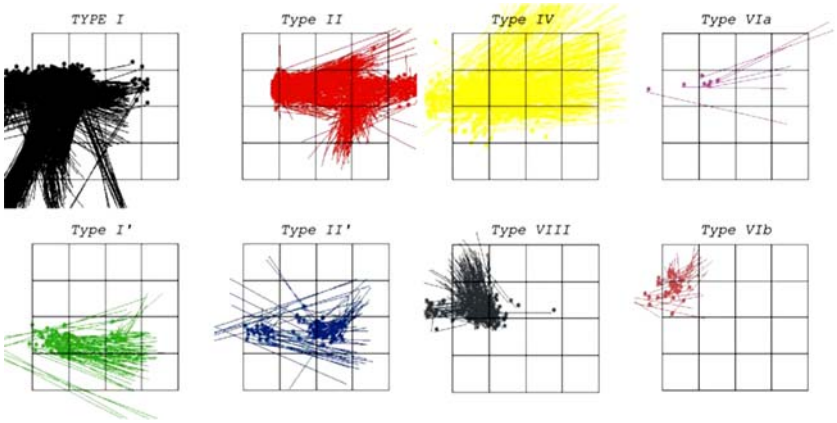

Figure 5. Vector plots of the eight β-turn types defined by Hutchinson and Thornton [32]. The order of the plots is: type I, II, IV, VIa, I′, II′, VIII and VIb.

C_α–C_β vector spatial arrangement (Figure 8). In this instance, type I and type II β-turns are presenting the same C_α–C_β vector spatial arrangement. To appreciate the difference between the clusters, the mean structures of each cluster were superimposed based on the $C_{\alpha1}$, $C_{\alpha2}$ and $C_{\alpha3}$ atoms. The result of this superimposition is displayed in Figure 9. The lowest inter-cluster RMSD (0.59) in Table 3 is between cluster 2 and cluster 4. The result in Figure 9 also demonstrates that cluster 2 (red) and cluster 4 (green) are most similar and furthermore, provides a visual aid to understanding the meaning of an inter-cluster RMSD value of 0.59. The highest inter-cluster RMSD (2.38) exists between cluster 6 and 9 (Table 3). The result in Figure 9 also demonstrates that cluster 6 (dark blue) and cluster 9 (grey) differ significantly.

Discussion

Combinatorial chemistry and drug discovery are now undergoing a significant shift away from maximizing diversity towards focusing diversity around biologically relevant regions [60]. This is driven by the enormous chemical diversity accessible to chemists [25] and the ever widening ''productivity gap'' of the pharmaceutical industry [28, 61, 62]. Typically, such biological focusing involves designing combinatorial libraries towards a protein family based on a known active compound or chemotype that binds to a member or members of the target family.

A molecular recognition event is dependent on the electrostatic and steric complementarities of ligand and receptor, and for small molecules this recognition surface is dictated or defined by the

Table 3. RMSD matrix of the nine clusters formed from clustering algorithm, explicit division of cluster three into two clusters and explicit inclusion of the mean of type I′ in the initial seeds.

Cluster	1	2	3	4	5	6	7	8	9
1	0.32	0.70	0.86	1.09	0.71	0.84	1.39	0.79	2.14
2	0.69	0.29	0.70	0.59	0.64	1.09	0.93	0.71	1.67
3	0.88	0.74	0.39	0.81	0.84	1.06	1.06	0.76	1.61
4	1.10	0.62	0.80	0.36	0.81	1.47	0.67	0.87	1.24
5	0.71	0.66	0.82	0.80	0.34	1.16	0.88	0.72	1.76
6	0.84	1.09	1.04	1.46	1.15	0.30	1.69	0.93	2.37
7	1.40	0.96	1.05	0.67	0.89	1.70	0.38	1.09	1.09
8	0.83	0.77	0.77	0.90	0.76	0.96	1.10	0.43	1.77
9	2.15	1.69	1.61	1.24	1.77	2.38	1.10	1.77	0.43

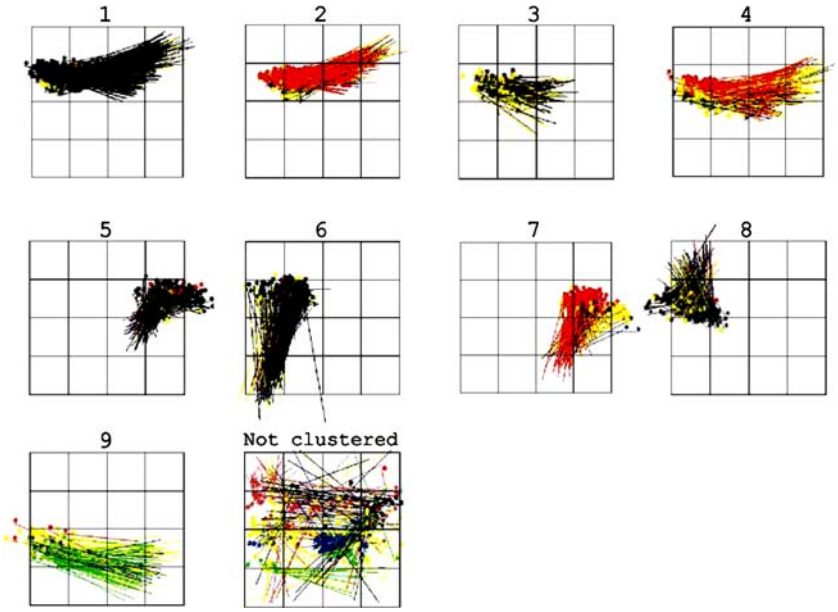

Figure 6. Vector plots of the nine clusters formed from the clustering algorithm and explicit division of cluster three into two clusters and inclusion of the average structure of type I′ in the initial seed. The last plot represents the conformations that were rejected and previous plots represent the nine clusters. The Clusters are colored according to conventional β-turns type definition. (Type I-black, Type II-red, Type I′-green, Type II′-dark blue, Type IV-yellow, Type VIII-grey, Type VIa1-magenta, Type VIa2-light blue and Type VIb-thin red).

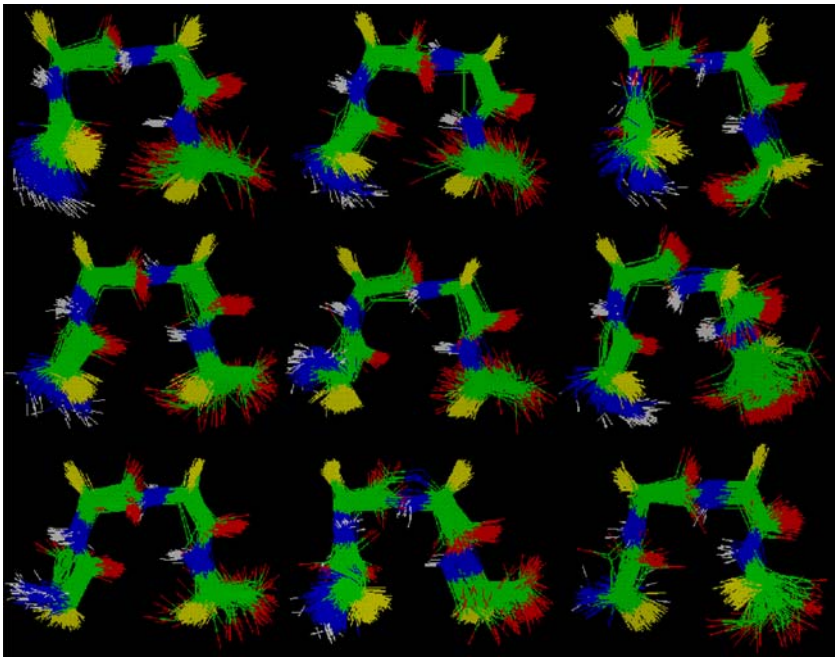

Figure 7. The β-turns within each of the nine clusters were superimposed onto the cluster's mean structure. The colouring schemes are: N: Blue, O: Red, H: white, C: green and C_β: yellow.

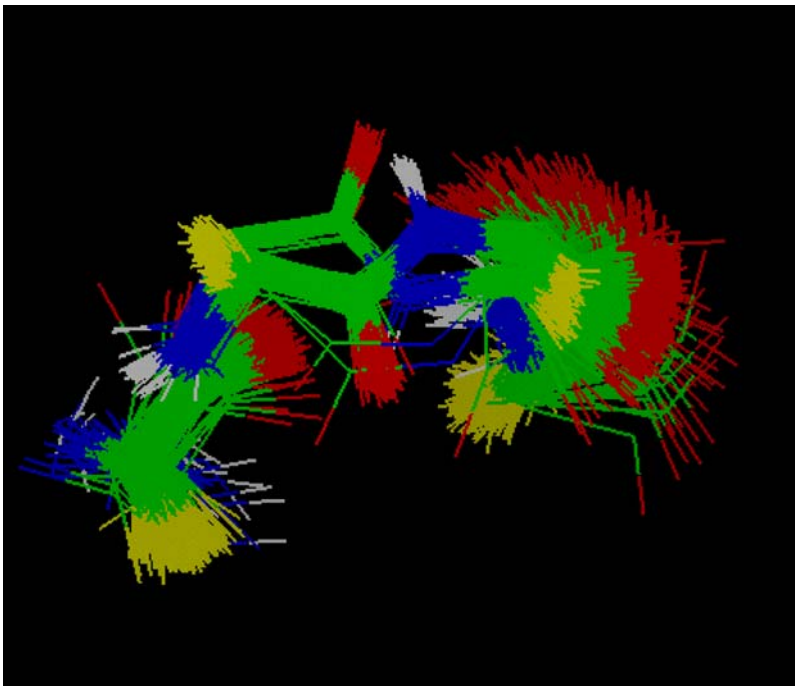

Figure 8. Top view of the β-turn structures in cluster two superimposed onto its mean structure. The figure shows that the backbone structures can vary significantly even-though the C_α–C_β vectors are distributed uniformly. The colouring schemes are: N: Blue, O: Red, H: white, C: green and C_β: yellow.

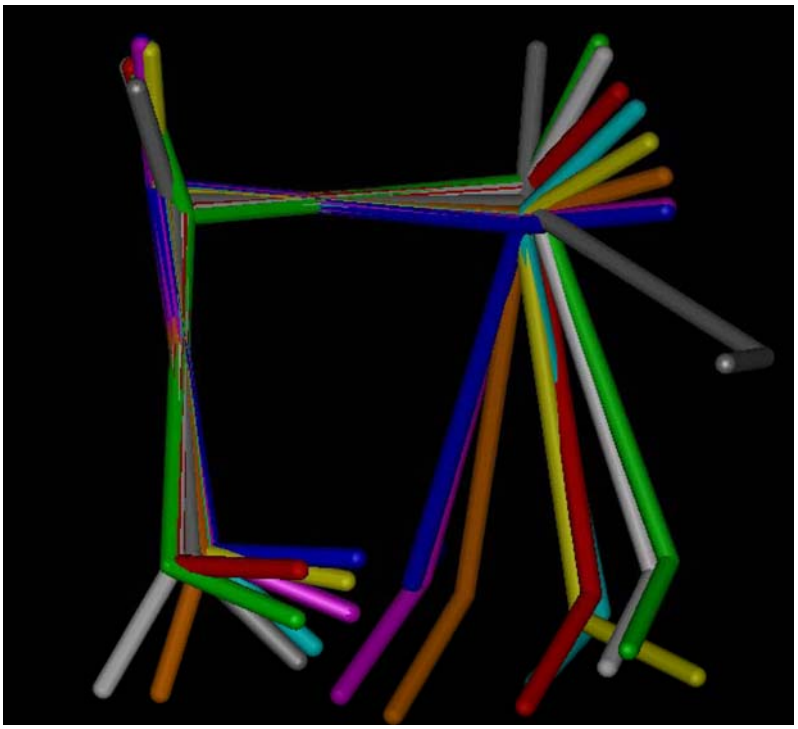

Figure 9. Superimposition of the mean structures of the nine clusters. The superimposition is based on the three atoms $C_{\alpha 1}$, $C_{\alpha 2}$ and $C_{\alpha 3}$. The color code for each cluster is described below: 1-magenta, 2-red, 3-yellow, 4-green, 5-orange, 6-dark blue, 7-white, 8-light blue and 9-grey.

geometry of appended functional groups. Selecting and synthesizing molecules that match biologically-relevant regions of structure space should also allow focusing to biologically-relevant regions. To achieve this, we need to topographically define this biologically-relevant structure space.

For proteins, surface features such as shape and electrostatic potential, hydrophobic patches [63–67], and about 76% of all hydrogen bonds in protein complexes [63, 68] are associated with side chains, suggesting the need to define common side chain topologies on protein surfaces. Since the side chains are considered to be the important recognition elements in protein–protein interactions [69], the abandonment of the backbone amides in peptidomimetic strategies will not significantly effect biological activity, and indeed is a common strategy when improving pharmacokinetics of peptide derived compounds.

β-turns are an important recognition element of peptides and proteins, and a great deal of scientific effort has been applied to classifying, designing and synthesizing β-turn mimetics [15, 33–43]. However, the traditional classification of β-turns is based on backbone criterion, which has little reflection on the position of the important recognition elements, the side chains. Figure 5 shows the vector plots of each of the traditional β-turn types using our dataset of 2675 β-turns. The figure shows that some of the β-turn types do not have uniform distribution of C_α–C_β vectors about a single peak or mode. Type I, type II and type II′ have at least two modes in the distribution. Type I′ and type VIII probably have one mode in the distribution. Type IV, which represents the β-turns that do not belong to the other turn types, is most widely spread. In addition to the fact that each traditional β-turn type can have multiple side-chain spatial arrangements, different types of β-turns can have similar side-chain spatial arrangement. A top view of cluster 2 shows that different backbone conformations (type I and type II β-turns) can display similar C_α–C_β vector topologies (Figure 8). These results suggest that the current β-turn type definition does not represent common motifs for side chain arrangements and therefore does not represent an optimal classification to focus the design of β-turn mimetics.

To alleviate this problem and to provide the required β-turn pharmacophores to drive mimetic development, we have re-classified β-turns according to side chain positions. The C_α–C_β vectors are a useful simplification of side chain position. They are anchored to the backbone and are therefore well defined experimentally, in comparison to the penultimate side chain position. They are useful in the design of peptidomimetics as the topography of the C_α–C_β vectors can be used in database searching strategies to identify appropriate mimetics. The extra atoms of the side chain can provide some element of induced fit.

Many authors have defined β-turn types [2, 29, 32, 70] based on the backbone dihedral angle ϕ_{i+1}, ψ_{i+1}, ϕ_{i+2} and ψ_{i+2}. By allowing three dihedral angles to vary by 30° from the standard angle and one dihedral angle to vary by 40° from the standard angle. Hutchinson and Thornton [32] have classified 57% of β-turns, leaving 43% unclassified (type IV). A broader definition used by Hutchinson and Thornton [32], where three dihedral angles are allowed to vary by 40° and one dihedral angle allowed to vary by 50°, classified about 68% of β-turns. Our classification algorithm was able to classify 90% of the 2675 β-turns into nine clusters.

It is simple to categories a new mimetic, a peptide or a protein loop into a turn family. This can be done by superimposing the C_α–C_β vectors (or equivalent) of the new turn or mimetic onto the mean structure of the nine families. If all of the RMSDs are greater than the filter threshold 0.65, then it does not belong to any cluster. Otherwise, the β-turn is grouped into the cluster with the lowest RMSD.

To compare our clustering with the traditional β-turn type classification, each β-turn in the vector plot is coloured according to the type of β-turn (Figure 6). Furthermore, Table 4 shows the distribution of β-turns for each combination of our cluster number and the traditional β-turn types. There seems to be some preference for certain β-turn types with certain clusters and vice versa. Type I is distributed primarily in cluster 1 (42%), 3 (8%), 5 (17%), and 6 (23%). Cluster 1, 5 and 6 are composed of predominantly type I (92%, 86% and 88%, respectively) and cluster 3 of type I (64%) and type IV (29%). Thus effectively, the non-modal distribution of type I is split into four clusters, 1, 3, 5 and 6 and some previously unclassified β-turns (type IV) have been clustered in cluster 3. On the other hand, type 2 is distributed in the other three clusters

Table 4. Comparing the traditional classification of β-turn types with the new classification described in this paper: (a) the number of β-turns in each combination of type and cluster; (b) each β-turn type is composed of various percentage of clusters; (c) each cluster is composed of various percentage of β-turn types.

(a)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Unclustered	Total
Type I	510	26	91	14	201	283	19	37	0	29	1210
Type II	5	292	6	182	12	4	140	7	4	15	667
Type I′	0	0	0	0	0	0	0	0	55	29	84
Type IV	26	25	42	78	15	21	49	35	33	68	392
Type II′	0	0	0	9	0	0	2	0	16	66	93
Type Via1	0	1	0	0	0	0	1	0	0	5	7
Type Via2	0	0	0	0	0	0	0	0	0	1	1
Type VIII	14	1	4	8	6	12	3	119	0	12	179
Type Vib	0	1	0	0	0	0	0	5	0	29	35
Total	555	346	143	291	234	320	214	203	108	254	2668
(b)	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)	Cluster 4 (%)	Cluster 5 (%)	Cluster 6 (%)	Cluster 7 (%)	Cluster 8 (%)	Cluster 9 (%)	Unclustered	
Type I	42	2	8	1	17	23	2	3	0	2	
Type II	1	44	1	27	2	1	21	1	1	2	
Type I′	0	0	0	0	0	0	0	0	65	35	
Type IV	7	6	11	20	4	5	13	9	8	17	
Type II′	0	0	0	10	0	0	2	0	17	71	
Type Via1	0	14	0	0	0	0	14	0	0	71	
Type Via2	0	0	0	0	0	0	0	0	0	100	
Type VIII	8	1	2	4	3	7	2	66	0	7	
Type Vib	0	3	0	0	0	0	0	14	0	83	
(c)	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)	Cluster 4 (%)	Cluster 5 (%)	Cluster 6 (%)	Cluster 7 (%)	Cluster 8 (%)	Cluster 9 (%)	Unclustered	
Type I	92	8	64	5	86	88	9	18	0	11	
Type II	1	84	4	63	5	1	65	3	4	6	
Type I′	0	0	0	0	0	0	0	0	51	11	
Type IV	5	7	29	27	6	7	23	17	31	27	
Type II′	0	0	0	3	0	0	1	0	15	26	
ype Via1	0	0	0	0	0	0	0	0	0	2	
Type Via2	0	0	0	0	0	0	0	0	0	0	
Type VIII	3	0	3	3	3	4	1	59	0	5	
Type Vib	0	0	0	0	0	0	0	2	0	11	

2 (44%), 4 (27%) and 7 (21%). Cluster 2 is dominated by type II (84%) whereas cluster 4 and cluster 7 comprise type II (63% and 65%, respectively) and type IV (27% and 23%, respectively). Therefore, the non-modal distribution of type II is split into three clusters, 2, 4 and 7 and some previously unclassified β-turns (type IV) have been clustered in cluster 4 and 7 by our approach. Type VIII is primarily distributed in cluster 8 (66%) and cluster 8 comprises type VIII (59%), type I (18%) and type IV (17%). 65% of Type I′ are in cluster 9 and cluster 9 is composed of 51% of type I′, 31% of type IV and 15% of type II′. Type IV, which is the unclassified type, is distributed quite evenly in most clusters. The algorithm did not classify Type VI (VIa1, VIa2 and VIb) and type II′ turns into cluster, those turns were predominantly (71%, 100% and 83%, respectively) in the unclustered group.

Do the resulting clusters represent overlapping-variation from a continuous spread or do

Table 5. Comparison of each cluster with all the other clusters to determine whether they are distinct or overlapping. The row and column headings refer to the cluster number and the content of the cells in the table refers to one of the 36 dimensions, which distinguish the two associated clusters. The dimensions are abbreviated using the symbols *Dn*, *Tn*, Φn and Ψn. The symbol *Dn* refers to distance *n* where *n* is an integer between 0 and 23. The symbol *T* refers to torsional angle *n* where *n* is an integer between 0 and 5. The symbol Φn or Ψn refer to the Φ and Ψ angle *n* where *n* is an integer between 1 and 4. Since the evaluate_overlap.cpp program does not handle distributions with more than one peak, those distributions are judge to be distinct or overlap by visual inspection. The cells that have been distinguished by visual inspection are highlighted by underscoring.

Clusters	1	2	3	4	5	6	7	8	9
1		Ψ2, Φ3	T4	D3, D9	D8, D9	D5, D11, T2, T4, T5	D3, D8, D9, T1, T2	T5	D3, D9, D19, T1, T2, T3, T4
2			D15, D19	--	T0, T1, Ψ0	D2, D5, D11, T2, T4, T5	D9, T0, T1, T2	T5	D19, T1, T2, T3, T4
3				D9, D15, D19	T0, Ψ0	T4, T5	D8, D9, D19, T0, T1	T4, T5	D18, D19, T1, T3
4					D3	D3, D5, D9, D11, T4, T5	Ψ1	Ψ2, Φ3	Φ2, Ψ2
5						D5, D8, D9, D11, T1, T4, T5, Ψ0	Ψ2, Φ3	T5	D19, T3, T4
6							D3, D5, D8, D9, D11, T1, T4, T5	T2	D3, D5, D9, D11, D19, T1, T3, T4, T5
7								D9, T2, T4	Φ2, Ψ2
8									D19, T2, T4, T5
9									

they represent distinct clusters that do not overlap in hyperspace? This is analyzed as described in the Method section. The result of such analysis by the 'evaluate_overlap.cpp' program is displayed in Table 5. From this table, it can be concluded that 35 out of a total of 36 combinations of clusters are distinct. The only combination, which is not distinct, is cluster 2 and cluster 4. Consistently, this cluster combination also has the lowest inter-cluster RMSD (Table 3). We have decided not to combine the two clusters because combining does not bring any advantage to the β-turns mimetics strategy and because only 1D analysis has been performed, cluster 2 and cluster 4 could be distinct if 2D or other higher dimensional analysis were performed.

Conclusions

The traditional classification of β-turns has little reflection on the spatial arrangement of the important recognition elements, the side chains. Therefore, we have re-clustered the β-turns based on side chain conformations, as defined by C_α–C_β vectors. Nine clusters were found which represent 90% of the 2675 β-turns. The mean structures of the nine clusters can be used to identify which of the nine clusters a new turns belongs to by the use of superimposition. More importantly, we are now able to trawl large databases of potential scaffolds in order to discover new scaffolds that present side chains in biologically-relevant topographical space. Thus, the motifs can be used as biological descriptors to identify new scaffolds or chemotypes that sample biologically-relevant space. Since our clustering protocol simply requires the arrangement of C_α–C_β vectors in space we are now extending this process to loops, helices, sheets and protein surfaces in general.

Acknowledgements

We would like to thank Drs Peter Cassidy, Paul Doyle, John Harris, Mike Hann, Peter Seale, Garland Marshall and Rich Head for stimulating discussions. We would also like to thank both GlaxoSmithKline and the Australian Research Council for financial support.

References

1. Rose, G.D., Gierasch, L.M. and Smith, J.A., Protein Chem., 37 (1985) 1.
2. Wilmot, C.M. and Thornton, J.M., Protein Eng., 3(6) (1990) 479.
3. Kabsch, W. and Sander, C., Biopolymers, 22 (1983) 2577.
4. Wilmot, C.M. and Thornton, J.M., J. Mol. Biol., 203 (1988) 221.
5. Stanfield, R.L., Fieser, T.M., Lerner, R.A. and Wilson, I.A., Science, 248 (1990) 712.
6. Rini, J.M., Schulze-Gahmen, U. and Wilson, I.A., Science, 255 (1992) 959.
7. Garcia, K.C., Ronco, P.M., Verroust, P.J., Brunger, A.T. and Amzel, L.M., Science, 257 (1992) 502.
8. Nikiforovich, G.V. and Marshall, G.R., Biochem. Biophys. Res. Commun., 195 (1993) 222.
9. Plucinska, K., Kataoka, T., Yodo, M., Cody, W.L., He, J.X., Humblet, C., Lu, G.H., Lunney, E., Major, T.C., Panek, R.L., Schelkun, P., Skeean, R. and Marshall, G.R., J. Med. Chem., 36 (1993) 1902.
10. Kyle, D.J., Blake, P.R., Smithwick, D., Green, L.M., Martin, J.A., Sinsko, J.A. and Summers, M.F., J. Med. Chem., 36 (1993) 1450.
11. Walford, S.P., Campbell, M.M. and Horwell, J.C., J. Pharm. Pharmacol., 48 (1996) 188.
12. Reddy, D., Jagannadh, B., Dutta, A. and Kunwar, A., Int. J. Pept. Prot. Res., 46 (1995) 9.
13. Nutt, R.F., Veber, D.F. and Saperstein, R.J., Am. Chem. Soc., 102 (1980) 6539.
14. Brady, S.F., Paleveda, W.J., Arison, B.H., Saperstein, R., Brady, E.J., Raynor, K., Reisine, T., Veber, D.F. and Freidinger, R.M., Tetrahedron, 49 (1993) 3449.
15. Freidinger, R.M., Veber, D.F., Perlow, D.S. and Brooks, J.R., Science, 210 (1980) 656.
16. Li, S.Z., Lee, J.H., Lee, W., Yoon, C.J., Baik, J.H. and Lim, S.K., Eur. J. Biochem., 265 (1999) 430.
17. Andrianov, A.M., Mole. Biol., 33 (1999) 534.
18. Ripka, A.S. and Rich, D.H., Curr. Opin. Chem. Biol., 2 (1998) 441.
19. Suat Kee, K. and Seetharama, D.S., J. Curr. Pharm. Des., 9 (2003) 1209.
20. Nagai, U., Sato, K., Nakamura, R. and Kato, R., Tett. Lett., 49 (1993) 3577.
21. Cornille, F., Slomcynska, U., Smythe, M.L., Beusen, D.D., Moeller, K.D. and Marshall, G.R., J. Am. Chem. Soc., 117 (1995) 909.
22. Hirschmann, R., Nicolaou, K.C., Pietranico, S., Leahy, E.M., Salvino, J., Arison, B.H., Cicyy, M.A., Spoors, P.G., Shakespeare, W.C., Sprengeler, P.A., Hamley, P., Smith, A.B., Reisine, T., Raynor, K., Maechler, L., Donaldson, C., Vale, W., Freidinger, R.M., Cascieri, M.R. and Strader, C.D., J. Am. Chem. Soc., 115 (1993) 12550.
23. Hirschmann, R., Yao, W., Sascieri, M.A., Strader, C.D., Maechler, L., Cichy-Knight, M.A., Hynes, J., vanRijn, R.D., Sprengeler, P.A. and Smith, A.B., J. Med. Chem., 39 (1996) 2441.
24. Reaka, A.J. H., Ho, C.M. W. and Marshall, G.R., J. Comp. Aided. Mol. Des., 16 (2002) 585.
25. Horton, D.A., Bourne, G.T. and Smythe, M.L., Chem. Rev., 103 (2003) 893.
26. Ripka, W.C.D.L., , G.V., Bach, II, , A.C., Pottorf, R.S. and Blaney, J.M., Tetrahedron, 49 (1993) 3593.

27. Breinbauer, R., Vetter, I.R. and Waldmann, H., Agnew. Chem. Int. Ed., 41 (2002) 2878.
28. Muller, G., DDT, 8 (2003) 681.
29. Richardson, J.S., Adv. Protein Chem., 34 (1981) 167.
30. Venkatachalam, C.M., Biopolymers, 6 (1968) 1425.
31. Lewis, P.N., Momany, F.A. and Scheraga, H.A., Biochem. Biophys. Acta., 303 (1973) 211.
32. Hutchinson, E.G. and Thornton, J.M., Protein Sci., 3 (1994) 2207.
33. Ball, J.B. and Alewood, P.F., J. Mole. Recognit., 3 (1990) 55.
34. Takeuchi, Y. and Marshall, G.R., J. Am. Chem. Soc., 120 (1998) 5363.
35. Douglas, A.J., Mulholland, G., Walker, B., Guthrie, D.J.S., Elmore, D.T. and Murphy, R.F., Biochem. Soc. Trans., 16 (1988) 175.
36. Li, W. and Burgess, K., Tetrahedron Lett., 40 (1999) 6527.
37. Halab, L. and Lubell, W.D., J. Org. Chem., 64 (1999) 3312.
38. Terrett, N., Drug Discov. Today, 4 (1999) 141.
39. Rosenquist, S., Souers, A.J., Virgilio, A.A., Schurer, S.S. and Ellman, J.A., Abstracts of papers of the American chemical society 1999, 217 (1999) 212.
40. Gardner, R.R., Liang, G.B. and Gellman, S.H., J. Am. Chem. Soc., 121 (1999) 1806.
41. Mer, G., Kellenberger, E. and Lefevre, J.F., J. Mol. Biol., 281 (1998) 235.
42. Lombardi, A., D'Auria, G., Maglio, O., Nastri, F., Quartara, L., Pedone, C. and Pavone, V., J. Am. Chem. Soc., 120 (1998) 5879.
43. Fink, B.E., Kym, P.R. and Katzenellenbogen, J.A., J. Am. Chem. Soc., 120 (1998) 4334.
44. Ball, J.B., Hughes, R.A., Alewood, P.F. and Andrews, P.R., Tetrahedron, 49 (1993) 3467.
45. Ball, J.B., Andrews, P.R., Alewood, P.F. and Hughes, R.A., FEBS J., 273 (1990) 15.
46. Garland, S.L. and Dean, P. M., J. Comput. Aided Mol. Des., 13 (1999) 469.
47. Garland, S.L. and Dean, P.M., J. Comput. Aided Mol. Des., 13 (1999) 485.
48. Berstein, F.C., Koetzle, T.F., Williams, G.J.B., Edgar, F., Meyer, J., Brice, M.D., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
49. Chou, P.Y. and Fasman, G.D., J. Mol. Biol., 115 (1977) 135.
50. Ho, C.M. and Marshall, G.R., J. Comput. Aided Mol. Des., 185 (1993) 3.
51. Jakes, S.E. and Willett, P., J. Mol. Graph., 4 (1986) 12.
52. Wong, A. and Lane, T., J. R. Statist. Soc. B, 45 (1983) 362.
53. Wong, M.A. and Lane, T., J. R. Statist. Soc. B, 45 (1983) 362.
54. Sokal, R.R. and Michener, C.D., University of Kansas Science Bulletin, 38 (1958) 1409.
55. SAS/STAT User's guide, Volume 1, ANOVA-FREQ, Version 6, 4 ed., 1999.
56. Anderberg, M.R. Cluster Analysis for Applications. Academic Press, New York and London, 1973.
57. Forgy, E.W., Biometrics, 21 (1965) 768.
58. MacQueen, J.B., Proc. Symp. Math. Statist. Probability., 1967, 1, 281–297.
59. InsightII user guide, 95.0 ed., Biosym/MSI/Acceryls: San Diego, CA92121–2777, 1995.
60. Rose, S. and Stevens, A., Curr. Opin. Chem. Biol., 7 (2003) 331.
61. Drews, J. and Ryser, S., Drug Inf. J., 30 (1996) 97.
62. Drews, J. and Ryser, S., Nat. Biotechnol., 15 (1997) 1318.
63. Archakov, A.I., Govorun, V.M., Dubanov, A.V., Ivanov, Y.D., Veselovsky, A.V., Lewi, P. and Janssen, P., Proteomics, 3 (2003) 380.
64. Tsai, C.-J., Lin, S.L., Wolfson, H.J. and Nussinov, R., Protein Sci., 6 (1997) 53.
65. Tsai, C.-J., Lin, S.L., Wolfson, H.J. and Nussinov, R., CRC Crit. Rev. Biochem., 31 (1996) 127.
66. Wells, J.A., Proc. Natl. Acad. Sci. USA, 93 (1996) 1.
67. Lijnzaad, P. and Argos, P., Proteins: Struct. Funct. Genet., 28 (1997) 333.
68. Xu, D., Tsai, C.-J. and Nussinov, R., Protein Eng., 10 (1997) 999.
69. Hruby, V.J., Life Science, 31 (1982) 189.
70. Lewis, P.N., Momany, F.A. and Scheraga, H.A., Biophys. Acta, 303 (1973) 211.