J Comput Aided Mol Des (2010) 24:917–934
DOI 10.1007/s10822-010-9384-y

Defining scaffold geometries for interacting with proteins: geometrical classification of secondary structure linking regions

Tran T. Tran · Christina Kulis · Steven M. Long · Darryn Bryant · Peter Adams · Mark L. Smythe

Received: 9 December 2009 / Accepted: 31 August 2010 / Published online: 23 September 2010

Abstract Medicinal chemists synthesize arrays of molecules by attaching functional groups to scaffolds. There is evidence suggesting that some scaffolds yield biologically active molecules more than others, these are termed privileged substructures. One role of the scaffold is to present its side-chains for molecular recognition, and biologically relevant scaffolds may present side-chains in biologically relevant geometries or shapes. Since drug discovery is primarily focused on the discovery of compounds that bind to proteinaceous targets, we have been deciphering the scaffold shapes that are used for binding proteins as they reflect biologically relevant shapes. To decipher the scaffold architecture that is important for binding protein surfaces, we have analyzed the scaffold architecture of protein loops, which are defined in this context as continuous four residue segments of a protein chain that are not part of an α-helix or β-strand secondary structure. Loops are an important molecular recognition motif of proteins. We have found that 39 clusters reflect the scaffold architecture of 89% of the 23,331 loops in the dataset, with average intra-cluster and inter-cluster RMSD of 0.47 and 1.91, respectively. These protein loop scaffolds all have distinct shapes. We have used these 39 clusters that reflect the scaffold architecture of protein loops as biological descriptors. This involved generation of a small dataset of scaffold-based peptidomimetics. We found that peptidomimetic scaffolds with reported biological activities matched loop scaffold geometries and those peptidomimetic scaffolds with no reported biologically activities did not. This preliminary evidence suggests that organic scaffolds with tight matches to the preferred loop scaffolds of proteins, implies the likelihood of the scaffold to be biologically relevant.

Keywords Protein interaction · Privileged structures · β-turns · Clustering · Protein loops · Molecular recognition · Scaffold · Biologically relevant · Peptidomimetics · Biological descriptors

Introduction

One of the cornerstones of drug discovery are scaffolds. Scaffolds are also referred to as templates, substructures, chemotypes, core structures or molecular frameworks. In hit discovery, the goal is to identify a series of molecules based on a scaffold that show good preliminary structure–activity relationships. In lead optimization, affinity and pharmacokinetic parameters are optimized by modulating functional groups attached to the scaffold, or by even amending elements of the scaffold. In lead hopping,

Electronic supplementary material The online version of this article (doi:10.1007/s10822-010-9384-y) contains supplementary material, which is available to authorized users.

T. T. Tran (✉) · S. M. Long · M. L. Smythe (✉)
Protagonist Pty Ltd, PO Box 6421, St Lucia, QLD 4067, Australia
e-mail: t.tran@protagonist.com.au

M. L. Smythe
e-mail: m.smythe@protagonist.com.au

T. T. Tran · C. Kulis · M. L. Smythe
Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia

S. M. Long · D. Bryant · P. Adams
Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia

the goal is to effectively replace an existing scaffold with a new one that preserves the placement of side-chains.

A scaffold is not a well defined term. Lu [1] suggests there are three types of scaffolds. To synthetic chemists, a scaffold represents a retrosynthetic intermediate down a convergent synthesis pathway [1], to a biochemist the scaffold might be the invariant element in a series of molecules that have been assayed [1], and Lu also defines a topological scaffold as one with no side-chains [1].

In drug discovery, the majority of targets are proteinaceous in character. There is evidence suggesting that some scaffolds yield biologically active molecules more than others when binding proteins, these are referred to as privileged substructures [2]. In these cases, the invariant molecular substructure may interact favorably with common elements of different proteins (by simple extension of Bondensgaard [3]) or orientate functional groups into preferred geometries for preferential binding to protein surfaces [2, 4].

We have been interested in understanding if there are preferred scaffold topologies that interact with proteinaceous surfaces. With this in mind, we have been analyzing various features of proteins and protein complexes. We have concluded that the scaffold architecture of proteins is best represented by the c_α–c_β bond of amino acid residues. This bond is shared by every amino acid residue (except glycine). In this fashion, proteins can use the same scaffold architecture (c_α–c_β shape) of its binding surface with different amino acids to achieve different functions, or different scaffold architecture (different c_α–c_β shapes) with the same or different chemistries to achieve different functions.

In this paper, we cluster the scaffold position (c_α–c_β bonds) of protein loops. Loops are defined as continuous four-residue segments of a protein chain that are not part of, and that connect, adjacent α-helix or β-strand secondary structures. Loops are important recognition motifs as they are generally found on the surface of proteins linking secondary structures [5]. Loops are critical to many biological functions as exemplified by their roles in ligand binding [6], DNA-binding [7], binding to protein toxins [8], forming enzyme active sites [9], binding of metal ions [10], binding of antigens by immunoglobulins [11], binding of mononucleotides [12] and binding of protein substrates by serine proteases [13]. In an examination of approximately 135 G-protein coupled receptors, Fairlie et al. [14] concluded that their peptidic ligands adopted turn or loop conformations for binding. Given the significance of GPCRs in drug discovery, an understanding of the scaffold shapes of loop recognition may provide a generic solution to the discovery of new scaffolds for GPCRs.

Loops are a superset of β-turns, since β-turns are loops with the special property that the $c_{\alpha1}$–$c_{\alpha4}$ distance is less than 7 Å [15]. We present data that shows that the resulting side-chain loop clusters are capable of identifying scaffolds that are biologically relevant. We conclude that the resulting loop clusters when coupled with conformational-based drug design, focused library design, medicinal chemistry and biological screening may significantly impact on the discovery of new biologically relevant scaffolds. As biological relevant shapes can be derived from non-loop structures, such as helices, sheets, and small molecule binding surfaces, it should be noted here that by focusing on loops, the identified scaffolds only mimics a subsection of the biological relevant universe.

Methods and results

Extraction of loops from protein data bank

A database of loops was created by first extracting well refined (resolution of $\leq$2.0 Å and R-factor $\leq$20%) and non-homologous ($\leq$25%) protein chains [16, 17] from the Protein Data Bank [18]. The program STRIDE [19] was then used to identify the secondary structural elements (helices and sheets) contained in these chains. The linking regions, defined as the remaining residues that link these secondary structural elements or the protein terminus, were used for further analysis. The linking regions that consisted of four or more amino acid residues were divided into four residue segments, resulting in a total of 23,650 four-residue loops. Three hundred and nineteen of those loops were rejected because the distance between backbone atoms n, c_α and c' was not appropriate ($\leq$0.8 or $\geq$2.0 Å). Each of the remaining 23,331 loops was then simplified into four c_α–c_β vectors (Fig. 1a). For the glycine residue, the c_α–c_β vector was obtained from the c_α–c_β vector of a standard alanine after it has been superimposed onto the glycine residue based on the backbone atoms n, c_α, c' and o.

Our motivation for clustering using the 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. Drug discovery initiatives are focused around attaching different functional groups to organic scaffolds. The equivalent scaffold architecture of proteins are the c_α–c_β vectors, as this is shared by every amino acid residue (except glycine).

Furthermore, Class III mimetic strategies [20] involve anchoring c_α–c_β bonds to a non-peptidic scaffold, and the extra atoms of the side-chain provide a degree of flexibility in molecular recognition. We therefore consider that clustering according to c_α–c_β vectors as functionally significant [21], especially when the identified common motifs are used to direct peptidomimetic development.

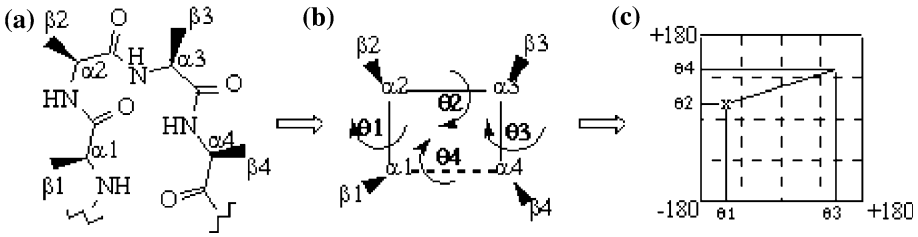

Fig. 1 Four c_α–c_β vectors of a loop. **a** Each four-residue loop is represented by four c_α–c_β vectors highlighted by the *dark triangles*. **b** To aid visualization of the topology of the loop after clustering, the four torsional angles $\theta 1$, $\theta 2$, $\theta 3$ and $\theta 4$ are used. **c** The four torsional angles are plotted as a vector from ($\theta 1$, $\theta 2$) (represented by the symbol '*x*') to ($\theta 3$, $\theta 4$)

Systematic identification of highly populated conformations (seeds)

By comparing each of the 23,331 loops with all the other loops, all the neighbors having a RMSD value of less than a constant, NEIGHBOR_LIMIT, were identified and counted for each loop. The next step was to determine for each loop, whether the number of neighboring loops is the peak when compared to its neighboring loops. The loop is classified as a peak if all the neighboring loops have a lower or equal number of neighboring loops. The numbers of peaks for various NEIGHBOR_LIMIT are shown in Fig. 2. This figure illustrates that the number of peaks varied from 7,818 to 13 as the NEIGHBOR_LIMIT varied from 0.7 to 0.25. Peaks with the number of neighbors of less than an arbitrarily chosen SEA_LEVEL of 20 are not interesting for mimetic design as they would not be well populated conformational motifs. The number of peaks with greater than the SEA_LEVEL of 20, shown in Fig. 2, ranges from 56 to 3, decreasing with increasing NEIGHBOR_LIMIT.

Since some peaks can be quite similar to each other, filtering was performed to identify a set of unique peaks to represent the data. Two peaks are considered not sufficiently unique if the fraction of shared loops between the two peaks exceeds an OVERLAP_LIMIT value that was set to 20% of the total number of loops in the two peaks. From peaks with the largest number of neighbors to peaks with the lowest number of neighbors, the peaks were filtered out if they are not sufficiently unique when compared with all the previous chosen unique peaks. This definition, which is based on the fraction of overlap, is more discerning than a definition that is based on the RMSD distance between the average structures of the peak. Figure 2 also shows the number of unique peaks as a function of NEIGHBOR_LIMIT. For NEIGHBOR_LIMIT of 0.4 and above, all the peaks were unique. However, for NEIGHBOR_LIMIT of less than 0.4, not all the peaks were unique.

The general principle applied to select the NEIGHBOR_LIMIT value, is that the larger the NEIGHBOR_LIMIT value, the more the generalization of the data, and that we seek to find the largest generalization without the loss of significant characteristic of the data set. Since the goal of this work is to find common clusters of loops, we believe the number of unique peaks with greater than twenty neighbors (SEA_LEVEL = 20) in the data set is the significant characteristic. Figure 2 shows that the number

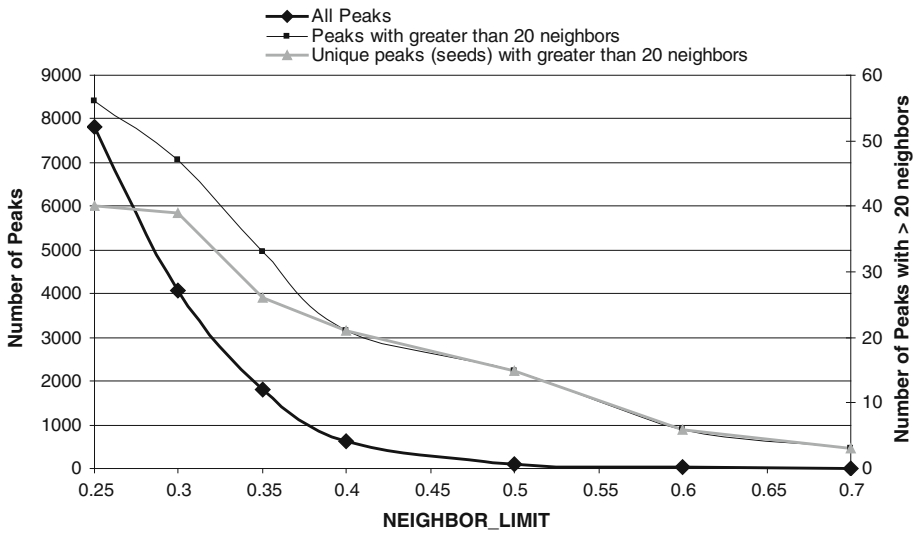

Fig. 2 Plot of peaks for various NEIGHBOUR_LIMIT. Plot of (**1**) number of peaks, (**2**) number of peaks with greater than 20 neighbors and (**3**) number of unique peaks with greater than 20 neighbors as a function of NEIGHBOR_LIMIT

of unique peaks with greater than twenty neighbors ranges from 3 to 40, decreasing with increasing NEIGHBOR_LIMIT value. There are loses of the significant characteristic (decrease in the number of unique peaks with greater than twenty neighbors) in all increases in the generalization except for the increase in generalization from NEIGHBOR_LIMIT of 0.25 to 0.3. Therefore, the NEIGHBOR_LIMIT of 0.3, that corresponding with 39 unique peaks, were considered as a largest generalization which still could represent most of the unique peaks with greater than twenty neighbors. We also determined that the unique peaks obtained using a higher NEIGHBOR_LIMIT are a subset of those obtained using a lower NEIGHBOR_LIMIT of 0.3 (See Supplementary Material, Table s1).

Filtered centroid sorting clustering

After the systematic identification of unique peaks with greater than twenty neighbors, a modified form of the nearest centroid sorting algorithm [22] called filtered nearest centroid sorting algorithm was utilized to refine the clustering. The filtered nearest centroid sorting algorithm is essentially the same as the nearest centroid sorting algorithm of Forgy [22, 23] where a prior estimate of initial seeds is required. The nearest centroid sorting algorithm assigns each observation to the nearest initial seed to form temporary clusters. After assigning all the observations to the 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. The filtered nearest centroid sorting algorithm differs from the nearest centroid sorting algorithm in that if the minimum RMSD of an observation to the mean structures is above a definable threshold, TOLERANCE, then the observation is considered too remote and therefore not assigned to the temporary clusters. In latter iterations, these unassigned observations are superimposed onto the mean structures of the new temporary clusters and if the minimum RMSD is below the TOLERANCE, then they are assigned to the new temporary cluster. The reason for this filtering is to remove the observations that differ significantly from the mean structures, so as not to bias the mean. Furthermore, not all the observations have to be clustered, only a major proportion is required.

To determine the clusters for the loops, the 39 unique peaks with greater than 20 neighbors were used as initial seeds for our filtered centroid sorting algorithm using various TOLERANCE. The percentage of representation, the average intracluster RMSD, the average intercluster RMSD and the ratio of the latter two are plotted as a function of the TOLERANCE in Fig. 3. The figure shows that as the TOLERANCE increases, the percentage representation and the intra-inter-RMSD ratio increases. At a TOLERANCE of 0.3, the percentage representation, the average intracluster RMSD, the average intercluster RMSD and ratio of the latter two are 12, 0.22, 0.95 and 0.11%, respectively; at a TOLERANCE of 0.6 RMSD they are 71, 0.42, 1.94 and 0.21%, respectively; and at a TOLERANCE of 0.9, they are 100, 0.51, 1.90 and 0.27%, respectively. Therefore, there are opposing forces in the choice of tolerance. Higher tolerance is favored because of the greater percentage of representation, but is disfavored because of the higher intra to inter cluster RMSD ratio. Lower tolerance is favored because of the lower intra to inter cluster RMSD ratio but is disfavored because of the lower percentage representation. From a loop mimetic point of view, it is more important to have a reasonable intracluster similarity (intra cluster RMSD) than to have high percentage representation. To aid in the choice of intracluster RMSD and indirectly the choice of TOLERANCE, a plot

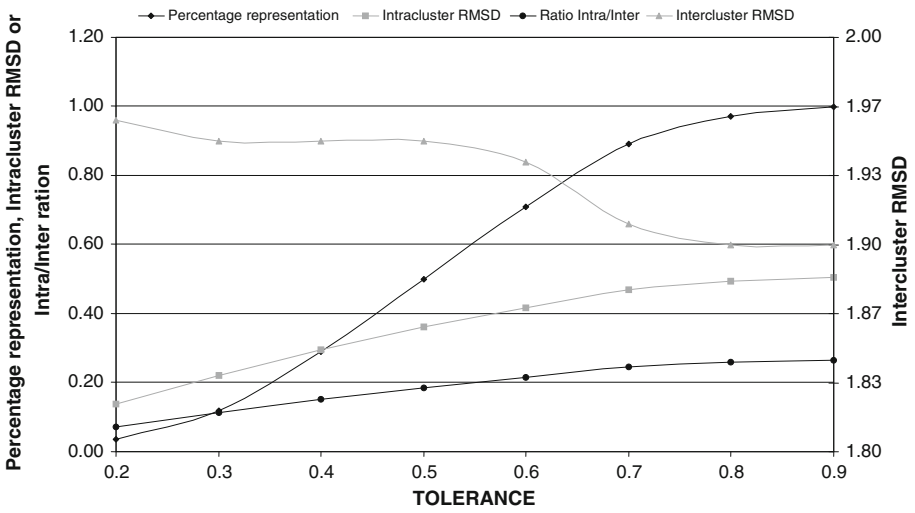

Fig. 3 Various TOLERANCE values plotted against percentage representation, intracluster RMSD, intercluster RMSD and their ratio. Filtered centroid sorting algorithm was used with various TOLERANCE to obtain the 39 clusters. The percentage representation, the intracluster RMSD, the intercluster RMSD and the ratio of the latter two are calculated and plotted as a function of TOLERANCE

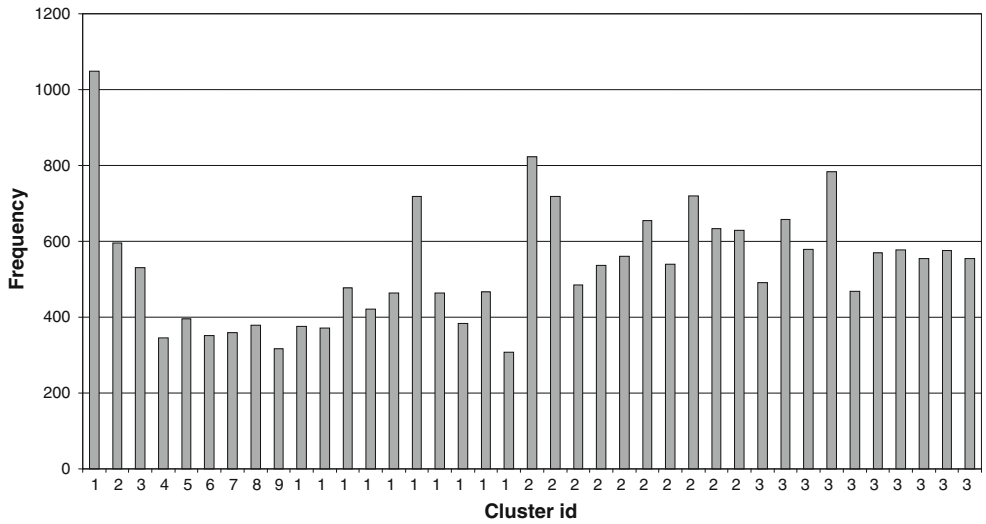

Fig. 4 Histogram of the number of loops in each of the 39 clusters

of all the loops in cluster one formulated using various tolerances are shown in Figure s2 in the supplementary material. The figure shows that with an average intracluster RMSD of 0.47, which corresponds with a TOLERANCE of 0.7 and a representation of 89%, the members in the cluster still are sufficiently similar for loop mimetic purposes. The numbers of loops in each of the 39 clusters obtained using a tolerance of 0.7 is plotted in Fig. 4. The least populated cluster has 307 loops (cluster 19) and the most populated cluster has 1,048 loops (cluster 1).

Vector plots of loops

A vector plot is used to aid the visualization of the three dimensional conformation of the four c_α–c_β vectors of a loop (Fig. 1a). Since the distances between adjacent c_α atoms are reasonably constant due to the nature of the peptide bond, the conformation can be approximately represented by the four torsional angles $\theta1$, $\theta2$, $\theta3$ and $\theta4$ (Fig. 1b). $\theta1$ is defined as the torsional angle between $c_{\beta1}$, $c_{\alpha1}$, $c_{\alpha2}$ and $c_{\beta2}$. $\theta2$ is defined as the torsional angle between $c_{\beta2}$, $c_{\alpha2}$, $c_{\alpha3}$ and $c_{\beta3}$; $\theta3$ as the torsional angle between $c_{\beta3}$, $c_{\alpha3}$, $c_{\alpha4}$ and $c_{\beta4}$; and $\theta4$ as the torsional angle between $c_{\beta1}$, $c_{\alpha1}$, $c_{\alpha4}$ and $c_{\beta4}$. The four torsional angles are plotted as a vector from ($\theta1$, $\theta2$) (represented by the symbol 'x') to ($\theta3$, $\theta4$) (Fig. 1c). The torsional angles are periodic in that a value of x is equivalent to $x \pm$ n360 where n is an integer. 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 loop in the cluster. The vector graphs for all the 39 clusters in Fig. 5 show that the loops within a cluster have reasonably similar conformations.

Calculation of the RMSD matrix of all the clusters

RMSD matrix is calculated to examine the performance of the clustering by assessing the similarity within and dissimilarity between clusters. Each cluster is compared with every other cluster so that the row and column number of the RMSD matrix represents the cluster number. The value in a cell at row x and column y represents the mean RMSD when all the loops in cluster x is superimposed to the mean structure of cluster y. The diagonal of the matrix with row x and column x represents the intra-cluster RMSD while the other cells represents the 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 loops in cluster x superimposed onto the mean structure of cluster y and the latter represent the mean RMSD of the loops in cluster y superimposed onto the mean structure of cluster x. However, the two numbers are very similar and therefore they are averaged and place in row x and column y where $x \geq y$. The RMSD matrix, showing the average intra- and inter-clusters RMSD for all the 39 clusters, is shown in Table s2 of the supplementary material. The average intracluster RMSD of 0.47 and the vector graphs for all the clusters in Fig. 5 show that the loops within a cluster have similar conformations. The loops between clusters are dissimilar as the average intercluster RMSD is 1.91.

Average linkage clustering algorithm

Based on the above mentioned RMSD matrix, we applied the average linkage clustering algorithm [24–26] to determine the structural relationship between the 39 clusters. In average linkage clustering algorithm [24–26], each structure is initially assigned to its own cluster of size one.

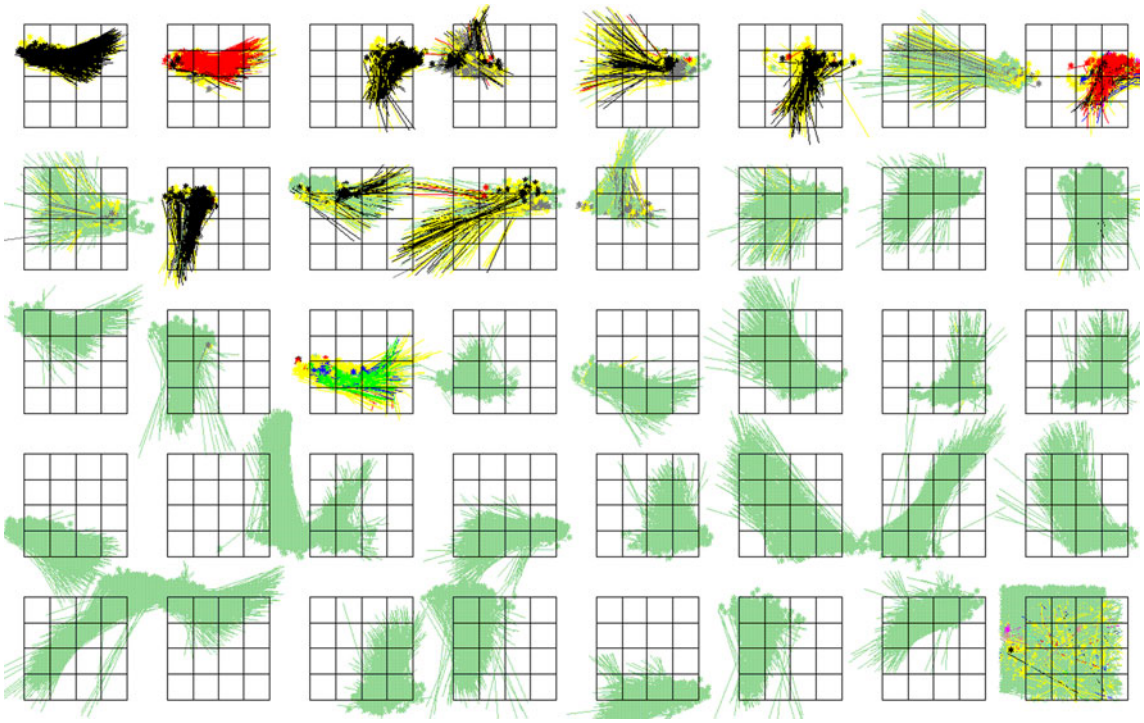

Fig. 5 Vector plots of the 39 clusters. Vector plots of the 39 clusters, with cluster numbering starting at number one on *top left* and counting across each row. The last cluster, cluster 40, represents all the loops that have not been clustered according to our filtered centroid sorting algorithm. The Clusters are colored according to conventional β-turn type definition (Type I-*black*, Type II-*red*, Type I'-*dark green*, Type II'-*dark blue*, Type IV-*yellow*, Type VIII-*grey*, Type VIa1-*magenta*, Type VIa2-*light blue*, Type VIb-*thin red* and for loops that are not β-turns, *light green*)

Subsequently, clusters are merged if the average distance between all the structures in the two clusters fall within some threshold. The resulting hierarchical tree, obtained by applying the average linkage clustering algorithm on the 39 clusters, is shown in Fig. 6. All the cluster numbers used in this paper follow the order from left to right of this hierarchical tree.

Analysis of whether the clusters are overlapping or distinct

Do the 39 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 clusters are 'distinct' if the most frequent 80% of the data in one cluster does not overlap with the most frequent 80% of the data in the other cluster. The overlaps were computed for each of the following 32 descriptors of the loop conformation. Each c_α–c_β vector pair has 4 distances ($c_{\alpha1}$–$c_{\alpha2}$, $c_{\alpha1}$–$c_{\beta2}$, $c_{\beta1}$–$c_{\alpha2}$, $c_{\beta1}$–$c_{\beta2}$), so the 6 possible c_α–c_β vector pairs (1–2, 1–3, 1–4, 2–3, 2–4, 3–4) of the four c_α–c_β vectors in a loop result in 24 distance descriptors. Furthermore, we also utilized all six possible torsional angle descriptors between the four c_α–c_β vectors and two torsional descriptors $c_{\alpha1}$–$c_{\alpha2}$–$c_{\alpha3}$–$c_{\alpha4}$ and $c_{\beta1}$–$c_{\beta2}$–$c_{\beta3}$–$c_{\beta4}$. Distinctions were made based on each of these 32 descriptors.

The maximum and minimum values which delineate the most frequent 80% of each distribution were not computed using the mean plus and minus some standard deviation because some of the spreads were not always a 'Normal' distribution. The maximum and minimum was computed by first 'binning' the data with respect to each of the 32 descriptors mentioned earlier. Then, the bins for each descriptor were sorted based on the frequency, from most frequent to least frequent. As the program traverses down the bins, the maximum and minimum of the descriptor are stored. The traversal is stopped when the program has traversed through at least 80% of the data. Consequently, the stored maximum and minimum represent the values that delineate the top eighty percentage 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, the determination of distinct is under-estimated.

For a particular descriptor, if the maximum and minimum of cluster X overlap with those of cluster Y, then cluster X and cluster Y are considered to be overlapping. On the other hand, if the maximum and minimum of cluster

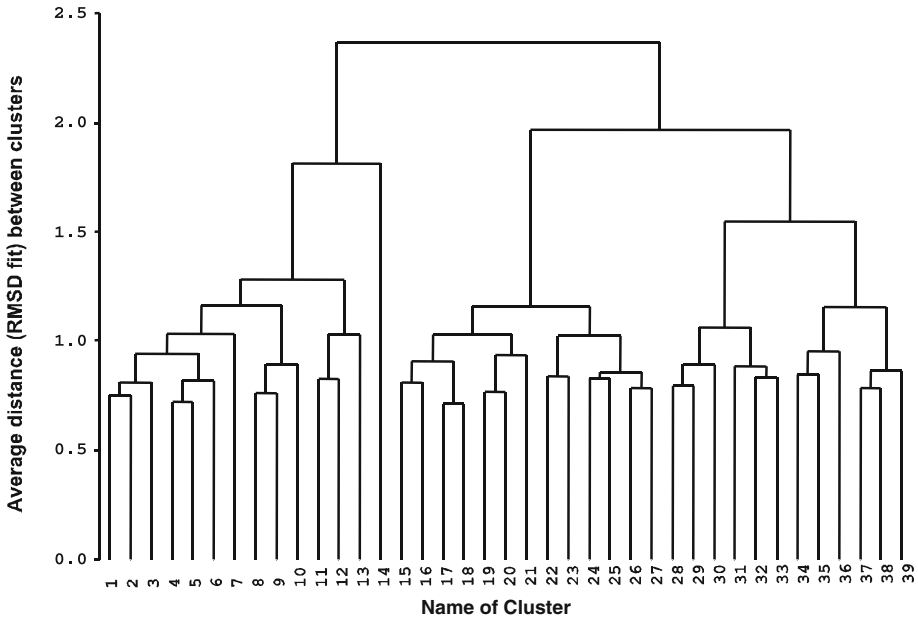

Fig. 6 Tree diagram obtained from average linkage clustering of the 39 clusters

X does not overlap with the maximum and minimum of cluster Y, then cluster X is considered to be distinct from cluster Y.

β-Turn mimetics and loop clusters

The resulting 39 side-chain loop clusters reflect the scaffold architecture of protein loops. We will examine whether the loop clusters, as biological descriptors, will be effective at identifying organic-based scaffolds that mimic the biologically relevant scaffold shapes of the loops. The β–turn is one subset of the loops. A thorough literature search reviewed 19 β-turn scaffolds (see Fig. 7, **1** & **2** [27–29], **3** [28, 30, 31], **4** [32, 33], **5** [30, 34, 35], **6** & **7** [28, 29, 36], **8** [30], **9** [37], **10** & **11** [38, 39], **12** [40], **13** [41–43], **14** [37], **15** [37, 41, 44–46], **16** [37, 44], **17** [47–49], **18** [30, 50], **19** [51]) with heterocyclic backbones, the presence of four or more R-groups, and the ability of the molecule to confer overall rigidity.

Each β-turn mimetic was sketched using version 7.1 of the Sybyl software package [66]. Each structure was refined and minimized using the Concord program [67]. The ConFirm [68, 69] conformational search program, from the Catalyst software package, was then used to generate a subset of low energy conformers for each β-turn scaffold. The *Best* mode was selected with the following parameters: A maximum 2,000 conformers are allowed, and only conformers within 6 kcal of the standing lowest energy conformer were collected (calculated using the CHARMm force field [52–54]). The completeness of the conformational search was evaluated by checking whether the scaffold structure changed using RMS fits of each conformer to its corresponding input conformation, and by examining the differences in the torsional angle at each R-group.

Each conformer was searched against each of the 39 loop cluster c_α–c_β positions using our in-house Vectrix program. The equivalent c_α–c_β positions of the β-turn mimetic scaffolds nominated for the search can be defined by identifying the 'R' atom of the scaffold as the c_β atom. Then the atom directly attached to this 'R' atom is defined as the c_α atom. These c_α–c_β positions describe the initiation of variable side-chains from an invariant scaffold. (Scaffolds 2 and 3, and 7 and 8 differ in the c_α–c_β positions searched at the R1 position.) A β-turn scaffold was considered a match with a loop cluster if it satisfies two criteria, distance tolerance criterion and RMS criterion. Distance tolerance criterion is characterized by c_α–c_α, c_α–c_β and c_β–c_β distances of 0.8, 0.9, and 1.0 Å, respectively, of a β-turn scaffold to a corresponding average loop cluster. For each of these distances, the RMS criterion is defined by the overall RMS of all distances less than or equal to 0.7 (the threshold for determine whether a loop is within a particular cluster).

Discussion

Drug discovery is heavily reliant on combinatorial chemistry and high throughput screening (HTS). Combinatorial chemistry is the process used to prepare compounds based around a scaffold. Therefore, scaffolds are critical to drug development.

Empirical observations of biologically active compounds suggests that some scaffolds are biologically relevant and have the capacity of providing active molecules

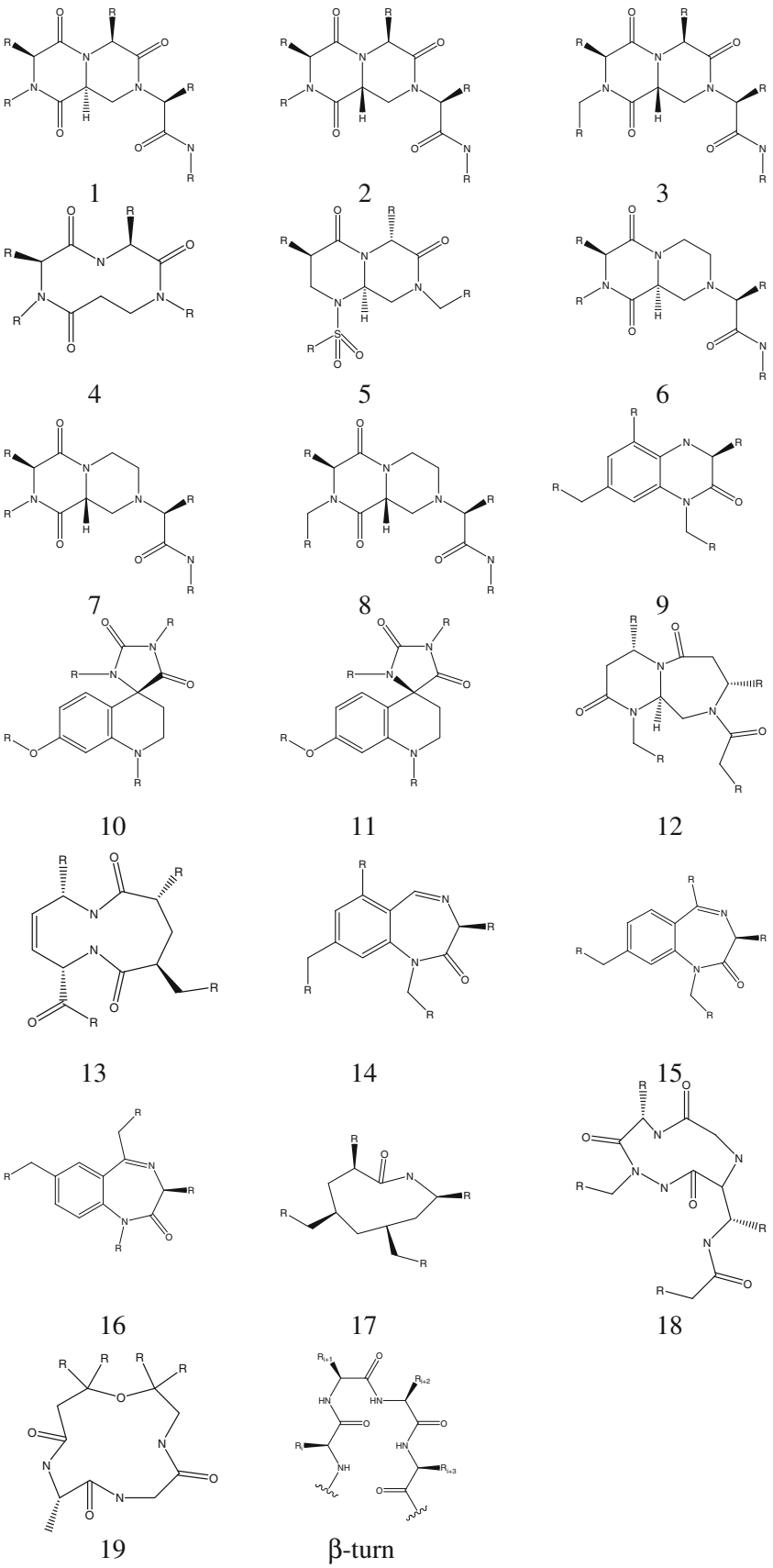

Fig. 7 Structures of β-turn mimetics with 4 or more R-groups from a recent survey of literature. The R-groups are referred to as R1, R2, R3 R4 or R5, with R1 at the *bottom left* corner, increasing in clockwise direction to R4 or R5

at diverse receptors [2]. One reason for this may be that the scaffold is presenting its side-chains in biologically relevant shapes.

The majority of drug targets are proteinaceous in character. We are interested in deciphering the preferred scaffold shapes, if any, that are required to bind to protein surfaces. To decipher these preferred shapes, we are analyzing crystal structures of protein complexes. Such a set of scaffolds would represent the functional shapes that proteins and peptides use to bind other proteins. Such shapes could then be used as biological descriptors to trawl large databases of compounds to identify organic-based scaffolds that match these biologically relevant scaffold shapes.

Using a novel clustering process, we have identified the preferred scaffold architecture of loops of proteins, well known molecular recognition motifs. The scaffold shapes of peptides and proteins are defined by the c_α–c_β bond.

We have derived 39 templates that reflect the c_α–c_β bond shapes of 89% of the loops comprised in the dataset. We have examined whether the 39 clusters are distinct.

Distinct clusters

The distinct cluster analysis described in the method section shows that 737 out of a total of 741 combinations (99%) of clusters are distinct. Since the analysis was based on individual descriptors, those non-distinct clusters could be overlapping or distinct if the analysis was extended to include combinations of the descriptors.

Relationship between the 39 clusters

To determine the conformational relationship between the 39 clusters, the average linkage clustering algorithm [24–26] was used and the resulting hierarchical tree is shown in Fig. 6. Moreover, the mean structures of the 39 clusters were superimposed based on the $c_{\alpha1}$, $c_{\alpha2}$, $c_{\alpha3}$ and $c_{\alpha4}$ atoms and the result is shown in Fig. 8. The characteristics of the major branches in the hierarchical tree in Fig. 6 were determined by examining the frequency distributions of the previously stated 32 descriptors for each of the 39 clusters. The most interesting correlation between member clusters of branches and the descriptors occurs at the $c_{\alpha1}$–$c_{\alpha2}$–$c_{\alpha3}$–$c_{\alpha4}$ torsional angle descriptor (Figure s1 in supplementary material).

The first branch, which contains cluster 1–14, is named the 'β-branch' because all of the member clusters have high percentage (10 out of the 14 clusters have $\geq$73% and the rest have $\geq$18%) of what traditionally been defined as β-turns while the rest of the clusters in the other branches have very low percentages of β-turns (cluster 30 with 10%, and the remaining having less than 2%, see Table 1). Interestingly, the clusters in the 'β-branch', colored green in Fig. 8, have relatively small $c_{\alpha1}$–$c_{\alpha2}$–$c_{\alpha3}$–$c_{\alpha4}$ torsional angles, with peaks between –55° and 85° (see Figure s1 in supplementary material). The small angle is geometrically associated with the shorter $c_{\alpha1}$–$c_{\alpha4}$ distance of β-turns. In addition, cluster 14 is the only cluster in the 'β-branch' with all negative $c_{\alpha1}$–$c_{\alpha2}$–$c_{\alpha3}$–$c_{\alpha4}$ torsional angles, occupying a very distinctive position in both the hierarchical tree in Fig. 6 and the superimposed average structures in Fig. 8.

The second branch, containing cluster 15–27 and colored red in Fig. 8, is named the 'negative branch' because it predominantly consists of negative $c_{\alpha1}$–$c_{\alpha2}$–$c_{\alpha3}$–$c_{\alpha4}$ torsional angle distributions with peaks between $-170°$ and $-80°$ (Figure s1). Visually, Fig. 8 shows that cluster 25 is the most distinctive in the negative branch and this is also reflected in Figure s1 in the supplementary material which shows the distribution for cluster 25 is the most deviant in the branch.

The third branch, containing cluster 28 to cluster 39 and colored blue and yellow in Fig. 8, has $c_{\alpha1}$–$c_{\alpha2}$–$c_{\alpha3}$–$c_{\alpha4}$ torsional angle distribution with peaks in the $\geq 70°$ or $\leq -160°$ regions. Furthermore, the $c_{\alpha1}$–$c_{\alpha2}$–$c_{\alpha3}$–$c_{\alpha4}$ torsional angle descriptor was able to subdivide the third branch into two sub-branches. The first sub-branch, which contains cluster 28–33 and is colored blue in Fig. 8, is named the 'positive branch' because it has peaks between 70° and 130°. The other sub-branch, which contains cluster 34–39

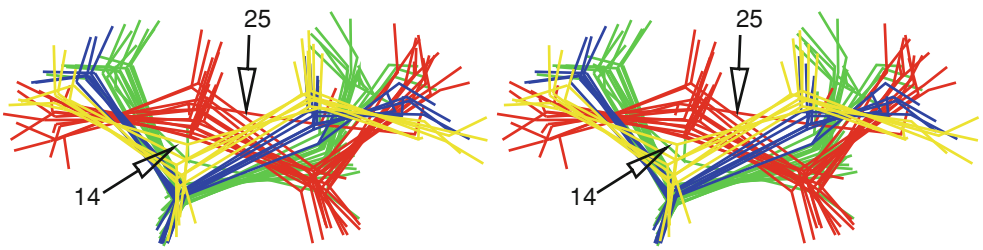

Fig. 8 Stereo diagram of the mean structures for the 39 clusters. The mean structures are superimposed based on the $c_{\alpha1}$, $c_{\alpha2}$, $c_{\alpha3}$ and $c_{\alpha4}$ atoms. The mean structures of the 'β-branch' (cluster 1–14) are colored *green*, 'negative-branch' (cluster 15–27) are colored *red*, 'positive-branch' (cluster 28–33) are colored *blue* and 'extended-branch' (cluster 34–39) are colored *yellow*

and is colored yellow in Fig. 8, is named the 'extended branch' because it has peaks in the $\geq 140°$ or $\leq -160°$ regions.

β-Turn types

β-turns, a subset of the loops being studied here, are an important recognition element of peptides and proteins. A great deal of scientific effort has been applied to classify, design and synthesize β-turn mimetics [47, 55–68]. To compare our clustering of loops with the traditional β-turn type classification as defined by Hutchinson and Thornton [69], the loops in the vector plots in Fig. 5 are colored according to the type and distribution of β-turns in each of the 39 clusters, summarized in Table 1. One-to-one correspondence between a particular β-turn type and a particular cluster requires that the primary composition of a particular cluster to be of a particular β-turn type and the primary composition of the β-turn type to be of the particular cluster. We have previously argued that particular β-turn types can have more than one peak or mode structures and that members of different β-turn types can have similar side-chain shape [70]. Therefore, it was not unexpected that there was no one-to-one correspondence between β-turn types, which were classified based on the backbone dihedral angles and the clusters found here, which were classified based on the conformations of the side-chain. Nonetheless, there seems to be some preferences for certain β-turn types towards certain clusters and vice versa.

Some of the β-turn types are distributed mainly in a few clusters (Table 1): 71% of type I' are in cluster 14; 86% of type II are in cluster 3 and cluster 7; 51% of type II' are in cluster 7 and 14 and the rest of type II' are un-clustered; 78% of type I are in cluster 1, 4, 5 and 8; more than half of the type VI is not clustered; and finally, type IV and type VIII are widely distributed in the β-branch cluster (1–14). In contrast, some of the clusters are preferentially distributed in a few β-turn types. The first 14 clusters and cluster 30 have a significant ($\geq$10%) percent of β-turns in their composition. The remaining 24 clusters have less than 2% of β-turns in them. Of the 15 clusters that have a significant portion of β-turns: clusters 1, 4, 5, 8 and 9 are composed mainly of type I and type IV β-turns; cluster 3 and 7 are composed mainly of type II, IV and some type I; cluster 11, 12, 13 and 30 are composed mainly of type VIII and IV; cluster 2, 6 and 10 are composed mainly of type VIII and IV and some type I; and finally, cluster 14 is composed mainly of type I', II' and IV. The lack of one-to-one mapping between the important recognition elements, the spatial arrangement of the side-chains, and the traditional β-turn types suggests that although the traditional classification of β-turns that is based on backbone dihedral angles has been useful, the resulting β-turn types are non-ideal for the design of β-turn mimetics.

In 1973, Lewis [15], based on the distribution of the $c_{\alpha 1}$–$c_{\alpha 4}$ distances of eight X-ray diffraction determined structures, concluded that β-turns have $c_{\alpha 1}$–$c_{\alpha 4}$ distances of $\leq$7 Å. Some individual clusters from this study have a significant proportion of members with a $c_{\alpha 1}$–$c_{\alpha 4}$ distance of $\leq$7 Å as well as >7 Å. This suggests that the Lewis definition of β-turns is arbitrary with respect to the side-chain spatial arrangement. Rather, to determine which of the newly derived 39 clusters a four-residue fragment belongs to, the c_α–c_β vectors of the fragment can be superimposed onto the 39 mean structures. If all of the RMSD values are greater than the filter threshold of 0.7, then it does not belong to any cluster. Otherwise, the fragment is grouped into the cluster with which it has lowest RMSD fit.

β-Turn clusters

Previously, we have clustered the β-turns based on the orientation of the four c_α–c_β vectors using a different methodology: *k*th nearest neighbor algorithm to find the initial seeds, followed by filtered nearest centroid sorting [70]. To compare our clustering of loops with our previous clustering of β-turns, the distribution of 'β-turn clusters' in each of the 39 loop clusters are summarized in the supplementary material, Table s3. It is reassuring to find, despite the different data set, methodologies and threshold values, six of the nine previously derived β-turn clusters are primarily composed of one specific loop cluster, and the remaining three β-turn clusters, 2, 3 and 4, are classified into the single loop cluster 3.

Biologically relevant descriptors

To determine if the loop clusters would be useful descriptors in identifying biologically relevant scaffolds, a test-set of heterocycles that mimic β-turns was chosen because by definition, β-turns is a subset of loops and therefore, they are the most relevant test for the loop clusters. A structural search of the conformers of the 19 β-turn mimetic scaffolds with 4 or more side-chain attachment sites (Fig. 7) was performed against the 39 mean loop cluster structures and the results are summarized in Table 2. Each loop cluster structure is characterized by four c_α–c_β vectors spatially arranged within three-dimensional space. A match is defined as a scaffold side-chain, or R-group, position that matches a minimum of three c_α–c_β vectors of a loop cluster. Overall, 14 out of 19 β-turn mimetic scaffolds matched with the loop clusters. None of the conformers generated for the scaffolds **4**, **6**, **7**, **8** and **19** produced matches.

Table 1 Comparing the traditional classification of β-turn types with the new classification of loops described in this paper

Loop Clusters	β -turns types									Turns total	Non-turn	Loops total	Turns/loops (%)
	I	II	I'	IV	II'	VIal	VIa2	VIII	VIb				
1	789	26	0	223	0	0	0	9	0	1,047	1	1,048	100
2	81	11	0	108	0	0	0	116	8	324	21	345	94
3	61	337	0	170	3	5	1	12	2	591	5	596	99
4	334	5	0	183	0	0	0	7	0	529	1	530	100
5	148	3	0	192	1	0	0	1	0	345	6	351	98
6	81	7	0	171	0	0	0	94	0	353	43	396	89
7	43	108	1	183	13	9	1	7	0	365	14	379	96
8	261	2	0	79	0	0	0	27	0	369	7	376	98
9	69	0	0	100	0	0	0	2	0	171	200	371	46
10	75	4	0	233	0	0	0	35	0	347	131	478	73
11	0	0	0	57	0	0	0	45	0	102	319	421	24
12	0	1	0	97	0	0	0	50	0	148	211	359	41
13	0	0	0	40	0	0	0	16	0	56	261	317	18
14	1	5	72	165	25	0	0	0	0	268	39	307	87
15	0	0	0	1	0	0	0	0	0	1	821	822	0
16	0	0	0	10	1	0	0	0	0	11	707	718	2
17	0	0	0	0	0	0	0	0	0	0	720	720	0
18	0	0	0	0	0	0	0	0	0	0	539	539	0
19	0	0	0	0	0	0	0	0	0	0	633	633	0
20	0	0	0	0	0	0	0	0	0	0	485	485	0
21	0	0	0	0	0	0	0	0	0	0	579	579	0
22	0	0	0	10	0	0	0	0	0	10	526	536	2
23	0	0	0	0	0	0	0	0	0	0	560	560	0
24	0	0	0	0	0	0	0	0	0	0	654	654	0
25	0	0	0	0	0	0	0	0	0	0	658	658	0
26	0	0	0	0	0	0	0	0	0	0	629	629	0
27	0	0	0	0	0	0	0	0	0	0	491	491	0
28	0	0	0	2	0	0	0	0	0	2	716	718	0
29	2	0	0	10	0	0	0	0	0	12	452	464	3
30	1	0	0	31	0	0	0	14	0	46	417	463	10
31	0	0	0	0	0	0	0	0	0	0	576	576	0
32	0	0	0	5	0	0	0	1	0	6	460	466	1
33	0	0	0	2	0	0	0	0	0	2	382	384	1
34	0	0	0	0	0	0	0	0	0	0	468	468	0
35	0	0	0	0	0	0	0	0	0	0	570	570	0
36	0	0	0	0	0	0	0	0	0	0	555	555	0
37	0	0	0	0	0	0	0	0	0	0	578	578	0
38	0	0	0	0	0	0	0	0	0	0	783	783	0
39	0	0	0	0	0	0	0	0	0	0	554	554	0
Unclustered	14	11	28	482	32	19	5	11	37	639	1,854	2,493	26

The number in each cell represents the number of loops belonging to cluster x (row x) and β-turn type y (column y). The column 'Non-Turn' represents the loops that are not β-turns. The column 'Turns/Loops' represents the percentage of β-turns in the cluster

The hypothesis we wished to examine is if the scaffolds that structurally match one or more loop clusters could be classified as biologically relevant scaffolds. To substantiate or challenge this hypothesis, detailed SciFinder [71] chemical synthesis, biological assays, and activity data were collected for each of the β-turn mimetic scaffolds.

Table 2 A summary of the top 5 RMSD matches (C3, Column 3) between the β-turn scaffolds (C1) and the loop clusters (C2)

β-turn template	Loop cluster	RMS	β-turn R-group	β-turn cα–cα bond distance	Number with correct bond distance 3	Number with correct bond distance 6	Number with correct bond distance 9	Loop cluster R-group	Loop cluster cα–cα bond distance	Number with bond distance 3	Number with bond distance 6	Number with bond distance 9
1	13	0.24	2, 3, 4	3,3,6	2	1	0	1,2,3	3,3,6	2	1	0
1	39	0.24	2, 3, 4	3,3,6	2	1	0	1,2,3	3,3,6	2	1	0
1	6	0.26	2, 3, 4	3,3,6	2	1	0	1,2,3	3,3,6	2	1	0
1	21	0.26	2, 3, 4	3,3,6	2	1	0	1,2,3	3,3,6	2	1	0
1	35	0.26	2, 3, 4	3,3,6	2	1	0	1,2,3	3,3,6	2	1	0
2	5	0.25	3, 2, 4	3,6,3	2	1	0	3,2,4	3,6,3	2	1	0
2	6	0.27	3, 2, 4	3,6,3	2	1	0	2,1,3	3,6,3	2	1	0
2	13	0.36	3, 2, 4	3,6,3	2	1	0	3,2,4	3,6,3	2	1	0
2	13	0.36	3, 2, 4	3,6,3	2	1	0	2,1,3	3,6,3	2	1	0
2	30	0.36	3, 2, 4	3,6,3	2	1	0	2,1,3	3,6,3	2	1	0
3	6	0.3	3, 2, 4	3,6,3	2	1	0	3,2,4	3,6,3	2	1	0
3	2	0.36	3, 2, 4	3,6,3	2	1	0	3,2,4	3,6,3	2	1	0
3	5	0.39	3, 2, 4	3,6,3	2	1	0	3,2,4	3,6,3	2	1	0
3	13	0.39	3, 2, 4	3,6,3	2	1	0	3,2,4	3,6,3	2	1	0
3	15	0.42	3, 2, 4	3,6,3	2	1	0	2,1,3	3,6,3	2	1	0
5	8	0.55	2, 3, 4, 1	3,3,5,6,3,6	3	2	*0*	2,3,4,1	3,3,9,6,3,6	3	2	*1*
5	14	0.61	2, 3, 4, 1	3,3,5,6,3,6	3	2	*0*	3,2,1,4	3,3,9,6,3,6	3	2	*1*
5	36	0.22	2, 3, 4	3,3,6	2	1	0	4,3,2	3,3,6	2	1	0
5	15	0.23	2, 3, 4	3,3,6	2	1	0	3,2,1	3,3,6	2	1	0
5	19	0.23	2, 3, 4	3,3,6	2	1	0	3,2,1	3,3,6	2	1	0
5	31	0.23	2, 3, 4	3,3,6	2	1	0	4,3,2	3,3,6	2	1	0
9	7	0.46	2, 3, 1	3,6,3	2	1	0	2,3,1	3,6,3	2	1	0
9	32	0.28	2, 3, 1	3,6,3	2	1	0	3,2,4	3,6,3	2	1	0
9	8	0.29	2, 3, 4	3,3,4(6)	2	1	0	2,3,4	3,3,6	2	1	0
9	10	0.29	2, 3, 4	3,3,4(6)	2	1	0	1,2,3	3,3,6	2	1	0
9	21	0.29	2, 3, 1	3,6,3	2	1	0	2,1,3	3,6,3	2	1	0
9	22	0.29	2, 3, 4	3,3,4(6)	2	1	0	1,2,3	3,3,6	2	1	0
10	8	0.61	4, 2, 1	4,6,4	*0*	1	*0*	1,2,4	3,6,9	*1*	1	*1*
11	7	0.53	4, 2, 1	4,6,4	*0*	1	*0*	4,3,1	3,6,9	*1*	1	*1*
11	4	0.63	4, 2, 1	4,6,4	*0*	1	*0*	4,3,1	3,6,9	*1*	1	*1*
12	1	0.41	2, 1, 4	4,6,6	*0*	1	*0*	2,1,4	3,9,6	*1*	1	*1*
12	3	0.46	2, 1, 4	4,6,6	*0*	1	*0*	2,1,4	3,9,6	*1*	1	*1*
12	8	0.48	2, 1, 4	4,6,6	*0*	1	*0*	4,3,1	3,6,9	*1*	1	*1*
12	14	0.49	3, 1, 4	5,6,3	1	*0*	*0*	2,4,1	6,9,3	1	*1*	*1*
12	1	0.59	3, 1, 4	5,6,3	1	*0*	*0*	2,4,1	6,9,3	1	*1*	*1*

Table 2 continued

β-turn template	Loop cluster	RMS	β-turn R-group	β-turn cα–cα bond distance	Number with correct bond distance 3	Number with correct bond distance 6	Number with correct bond distance 9	Loop cluster R-group	Loop cluster cα–cα bond distance	Number with bond distance 3	Number with bond distance 6	Number with bond distance 9
13	6	0.48	3, 2, 4	3,6,3	2	1	0	3,2,4	3,6,3	2	1	0
13	2	0.5	3, 2, 4	3,6,3	2	1	0	3,2,4	3,6,3	2	1	0
13	36	0.5	3, 2, 4	3,6,3	2	1	0	2,1,3	3,6,3	2	1	0
13	2	0.51	3, 1, 4	6,5,3	1	1	*0*	2,4,1	6,9,3	1	1	*1*
13	4	0.53	3, 1, 4	6,5,3	*1*	1	0	1,3,2	6,3,3	*2*	1	0
14	29	0.37	3, 2, 1	4,3,6(7)	*1*	1	0	4,3,2	3,3,6	*2*	1	0
14	33	0.38	3, 2, 1	4,3,6(7)	*1*	1	0	4,3,2	3,3,6	*2*	1	0
14	4	0.43	3, 4, 2	3,4(6),4	*1*	1	0	2,3,1	3,6,3	*2*	1	0
14	30	0.44	3, 2, 1	4,3,6(7)	*1*	1	0	3,2,1	3,3,6	*2*	1	0
14	9	0.45	3, 2, 1	4,3,6(7)	*1*	1	0	4,3,2	3,3,6	*2*	1	0
15	10	0.33	3, 4, 1	3,5,6(7)	1	*0*	*0*	1,2,4	3,6,9	1	*1*	*1*
15	6	0.42	3, 4, 1	3,5,6(7)	1	*0*	*0*	1,2,4	3,6,9	1	*1*	*1*
15	3	0.47	2, 4, 1	4(5),5,5	*0*	*0*	*0*	3,4,1	3,9,6	*1*	*1*	*1*
15	5	0.5	3, 4, 1	3,5,6(7)	1	1	*0*	2,1,4	3,9,6	*1*	1	*1*
15	8	0.51	2, 4, 1	4(5),5,5	*0*	*0*	*0*	2,1,4	3,9,6	*1*	*1*	*1*
16	3	0.4	4, 2, 1	4(5),5,5	*0*	*0*	*0*	2,1,4	3,9,6	*1*	*1*	*1*
16	1	0.43	4, 2, 1	4(5),5,5	*0*	*0*	*0*	2,1,4	3,9,6	*1*	*1*	*1*
16	8	0.45	4, 2, 1	4(5),5,5	*0*	*0*	*0*	2,1,4	3,9,6	*1*	*1*	*1*
16	2	0.53	3, 2, 1	3,5,6	1	1	*0*	2,1,4	3,9,6	1	1	*1*
16	2	0.54	4, 2, 1	4(5),5,5	*0*	*0*	*0*	2,1,4	3,9,6	*1*	*1*	*1*
17	10	0.41	2, 3, 1	3,5(6),3	2	1	0	3,2,4	3,6,3	2	1	0
17	29	0.41	2, 3, 1	3,5(6),3	2	1	0	2,1,3	3,6,3	2	1	0
17	36	0.42	2, 3, 1	3,5(6),3	2	1	0	2,1,3	3,6,3	2	1	0
17	28	0.44	2, 3, 1	3,5(6),3	2	1	0	2,1,3	3,6,3	2	1	0
17	4	0.45	2, 3, 1	3,5(6),3	2	1	0	3,2,4	3,6,3	2	1	0
18	4	0.27	2, 1, 3	3,5,6	1	1	*0*	2,1,4	3,9,6	1	1	*1*
18	6	0.39	2, 1, 3	3,5,6	1	*0*	*0*	4,3,1	3,6,9	1	*1*	*1*
18	5	0.43	2, 3, 4	6,3,9	1	1	1	4,2,1	6,3,9	1	1	1
18	28	0.44	2, 1, 4	3,8,9	1	*0*	1	4,3,1	3,6,9	1	*1*	1
18	6	0.48	2, 1, 3	3,5,6	1	1	*0*	2,1,4	3,9,6	1	1	*1*
				Percentage		83%	82%	8%	Total		108	67

26

The R-groups involved in the matches are shown in C4 and C9. The numbers of bonds between the c_α atoms of the matching R-groups are shown in C5 and C10. For 3 R-group (R1, R2, R3) matches, the number of bonds between the c_α atoms connecting R1–R2, R2–R3 and R1–R3 are shown. For 4 R-group (R1, R2, R3, R4) matches, the number of bonds between the c_α atoms connecting R1–R2, R2–R3, R3–R4, R1–R3, R1–R4, and R2–R4 are shown. The number in parenthesis represents bond distance of alternative pathway. The mismatched bond distances for β-turn scaffolds are italicized and underlined

The hypothesis is substantiated if: (1) a structural match correlates with the compound having been synthesized, assayed, and showing biological activities; and (2) lack of structural match correlates with the compound having been synthesized, assayed, but not showing biological activities. The hypothesis is challenged if (3) a structural match correlates with the compound having been synthesized, assayed, but not showing biological activities; and (4) lack of structural match correlates with the compound having been synthesized, assayed, and showing biological activities. Results of the analysis are summarized in Fig. 9.

Nine β-turn mimetic scaffolds (green bars in Fig. 9) **5**, **9**, **10**, **11**, **13**, **14**, **15**, **16** and **17** substantiate the hypothesis. β-turn mimetic scaffolds, **5** and **9** produced the best matches with the loops with average RMSD of 0.23 and 0.29, respectively. Scaffold **5** is the only β-turn mimetic with all four side-chain matches to the loops, the remaining scaffolds only have three side-chain matches (Table 2). Scaffold **5** was designed as a type I β-turn mimetic [30, 35] and all four of its c_α–c_β positions matches the loop clusters, 8 and 14. Cluster 8 is prominently β-turns of type I (69%), IV (21%), and VIII (7%). Cluster 14 is predominantly type I' β-turn (23%), type IV β-turn (54%), type II' β-turn (8%), and non-β-turn (13%). A number of compounds based on **5** were synthesized [34, 35] and assayed and some were found to have μM IC_{50} against D-pen [3], D-pen [5]-enkphalin (DPDPE) binding to δ and μ opioid receptors [30, 72, 73]. Scaffold **9** was designed by Ripka et al. [37] as a β-turn mimetic and matches with loop cluster 7. The scaffold mimetic, **9**, has been found to have some functionality as or in herbicide Saftener [74], Endothelin receptor antagonist [75], AMPA receptors [76], and N-methyl-D-aspartate-stimulated 3H-GABA release [77]. The result from these two β-turn mimetics support the hypothesis, with tight matches to the loop clusters and having reported synthesis, assays and biological activity.

The hypothesis is also supported by data for β turn mimetic scaffold **13**, which showed an average RMSD match with the loops of 0.50 (Table 2), and showed biological activity as mimetics of enkephalin [41–43].

Three β-turn mimetics, **14** [37], **15** [37, 41, 44–46], and **16** [37, 44] contain a benzodiazepine scaffold. These mimetics match the loops with an average RMSD of 0.41 0.45 and 0.45, respectively (Table 2). In support of the hypothesis, they are found to be active [78] against numerous targets, including cholecystokinin (CCK), gastrin and central benzodiazepine receptors, neurokinin-1, κ-secretase, farnesyl transferase, ion channel ligands and as a mimetic of Gramicidin S [37, 46]. Evans et al. [79] in 1988 first used the term "privileged structure" to describe a structure that is capable of binding to a relatively large number of receptors, in reference to this benzodiazepine scaffold. Interestingly, of the 14 spatial matches between the benzodiazepine scaffold and the loop structure, 11 belong to the β-turn branch of the loop, in agreement with the previous report that benzodiazepine scaffold is β-turn mimetic [44, 79]. It was gratifying that the loop clusters were able to identify the prototypical privileged substructure.

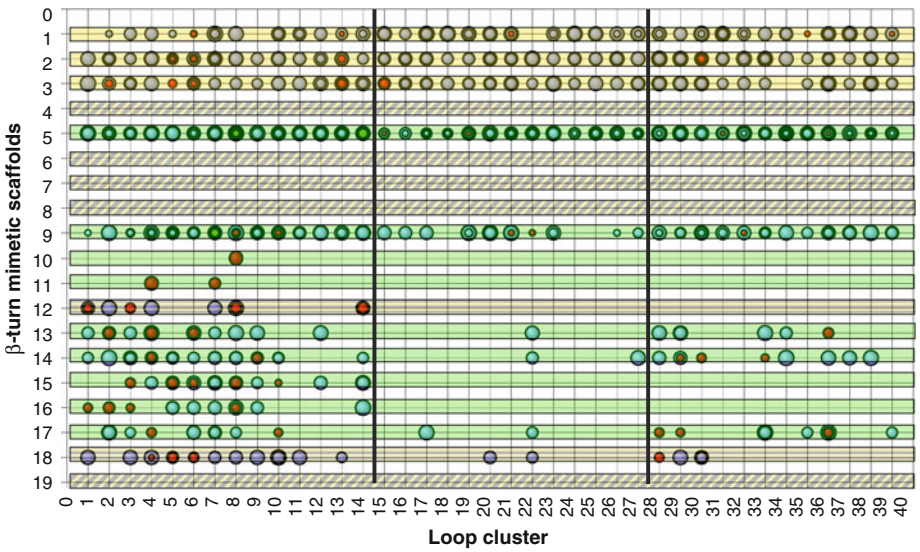

Fig. 9 Structural match between the 19 β-turn mimetic scaffolds and the loop clusters based on RMSD. Matches are represented by RMSD values (size of the *bubble*) that are ≤0.7. Multiple RMSD values for different combinations of R group within a scaffold are represented by *concentric bubbles*. Only two β-turn mimetics have matches with all four R-groups of each scaffold (*green bubble*), the remaining match with only three R-groups. The top 5 structural matches are highlighted by the *red bubbles*. The *green horizontal bars* represent data which support the hypothesis of structural matches that correlate with biological activity; the *yellow bars* are neutral and the *red bar* represents data that challenge the hypothesis

The β-turn mimetic scaffold **17** [47–49] has an average RMSD of 0.43 and corresponds to one reported synthesis and activity in relation to corticotropin releasing factor (CRF) [80]. The β-turn mimetics **10** [38, 39] and **11** [38, 39] matched the loop clusters. The R-groups, R4, R2 and R1, of **10** are the only match (high RMS of 0.61) with the R-groups, R1, R2 and R4, of loop cluster 8. The β-turn mimetic **11** is an enantiomer of **10** and the R-groups, R4, R2 and R1, of **11** match (RMS 0.63 and 0.53, respectively) with the R-groups, R4, R3 and R1, of loop cluster 4 and 7. The data from these two β-turn mimetics support the hypothesis as their structural matches correspond with reported biological activity on the somatostatin target [38, 81, 82].

The β-turn mimetic scaffolds, **1**, **2** and **3** (pure yellow bar in Fig. 9) share the same diketopiperazine framework [27, 29–31, 36], a well known privileged substructure, [83] and match loop clusters closely. SciFinder show there are 22 synthesized derivatives, however all the associated references do not report any assay for biological activities for the specific bicyclic diketopiperazine scaffold, and therefore, no conclusions can be drawn from these three β-turn mimetics templates.

Interestingly, for scaffolds that did not match any of the loop clusters, no reported biological activity could be identified. The β-turn mimetic scaffolds **4** [32, 33], **6** [28, 29, 36, 84], **7** [28, 29, 36, 84], **8** [30] and **19** [51, 85, 86] (diagonal yellow lines Fig. 9) do not have structural matches with the loops and interestingly, no biological activities have been reported, despite some of them being synthesized by commercial groups.

β-turn mimetic scaffold **18** [30, 50] has an average RMSD match with the loops of 0.38 (Table 2). However, no conclusion can be drawn regarding the hypothesis because there is no reported synthesis and consequently no biological activity in SciFinder (Horizontal lined yellow in Fig. 9). β-turn mimetic scaffold **12** [40] has an average RMSD match with the loops 0.46 (Table 2). No conclusion can be drawn regarding the hypothesis because there is a mere mention on an attempt at synthesis by Min et al. [40] and consequently there were no biological activity in SciFinder (Horizontal lined yellow in Fig. 9).

Unfortunately, the sample size of 19 β-turn mimetics suggests that concrete conclusions of the hypothesis cannot be made. In addition, it has been estimated that less than 1% of the biological activity of molecules has been reported [87], which also hinders the analysis. There is also certainly a trend not to publish compounds or scaffolds that are not biologically active.

Therefore, the results obtained from the analysis tentatively support the conclusion that the loop clusters, as descriptors, have the capacity to identify biologically relevant scaffolds. Analysis of the data reveals some interesting trends with respect to the development of scaffolds to mimic secondary structures. In particular, is the required distance between adjacent c_α atoms. Distances between the c_α–c_α atoms of all 39 loop clusters range from 3.73 to 3.80 Å. This narrow range is attributed to the conformationally restricted rotation about the amide bond of each loop cluster. Consequently, a β-turn scaffold must contain a c_α–c_α bond distance within this range in order to successfully match a loop cluster. The c_α–c_α distance between 3.73 and 3.80 Å is equivalent to the direct distance across three bonds.

All of the five β-turn mimetic scaffolds that did not match a loop, **4**, **6**, **7**, **8** and **19**, and therefore not shown in Table 2, contain c_α–c_α bond distances that are less than the minimum loop cluster distance of 3.73–3.80 Å. Each of these scaffolds has less than three bonds between corresponding Cα atoms. Scaffold **19** contains two 0 Å c_α–c_α distances because of two instances in which two R-groups share the same c_α atom (see Fig. 7). Scaffolds, **4**, **6** and **7**, each contain one short 1.5 Å c_α–c_α distance which is separated by only one bond and another short 2.5 Å c_α–c_α distance which is separated by only two bonds (see Fig. 7). Scaffold **8** contains two 2.5 Å c_α–c_α distances which are only two bonds apart. Subsequently, it is not possible for these scaffolds to form the minimum of three matches to a loop cluster.

For the remaining 14 matched β-turn mimetic scaffolds, the R-groups involved in the matches and their c_α–c_α bond distances are shown in Table 2. Eighty-three percent of the 108 c_α–c_α bond distances of 3 in the matched loop clusters, corresponds with a c_α–c_α bond distances of 3 in the matched β-turn scaffolds. All the mismatched c_α–c_α bond distances in the β-turn scaffold have c_α–c_α bond distances of 4 except for one which has 5 bonds (β-turn scaffold **13**, matching with loop cluster 4). Thus, the data seems to suggest that a bond distance of 3 and 4 are almost the only tolerable bond distances for structural matches with the bond distance of 3 in the loop clusters.

Similarly, 82% of the 67 c_α–c_α bond distances of 6 in the matched loop clusters, corresponds with the c_α–c_α distances of 6 in the matched β-turn scaffolds (Table 2). All the mismatched c_α–c_α bond distances in the β-turn scaffold have c_α–c_α bond distances of 5 except for one which has 8 bonds (β-turn scaffold **18**, matching with loop cluster 28). Thus, the data seems to suggest that a bond distance of 5 and 6 are almost the only tolerable bond distances for a structural match with a bond distance of 6 in the loop clusters. This information is important for the future design of β-turn mimetic scaffolds to match the biologically relevant loop clusters.

Conclusions

Scaffolds are a critical component of drug discovery. There are limited ways of focusing drug discovery efforts using

biologically relevant scaffolds. In comparison to random or synthetically driven chemical libraries, we have been pursuing the rational design and synthesis of libraries that incorporate scaffolds that present side chains in biologically relevant shapes. Thereby, greatly increasing the prospect that the developed library contains a hit. To this end, this paper presents the identification of 39 clusters of loops based on the side-chain spatial arrangement. The 39 clusters represent about 89% of the 23,331 loops.

Preliminary evidence was presented to support the hypothesis that a scaffold with large number of tight spatial match with the loops implies the likely hood of matching with the β-turn portion of the biologically relevant universe. It would be interesting to determine, in further works, if the 39 clusters is useful in identifying a bigger portion of the biologically relevant universe by using a test set that includes a variety of different scaffolds (not just β-mimetics). This would also alleviate the difficulty of finding bioactivity data because many scaffolds and bioactivity data could be extracted from databases such as MDDR, DrugBank, CMC and Chembl.

Further work is being carried out in our laboratory to identify structural scaffolds that give tight spatial match those 39 loop clusters for the formation of libraries of compounds that are biased toward the scaffold geometries of loops for the discovery of GPCR ligands. It is now well recognized that peptidergic GPCRs bind their ligands in reverse turn conformations [14].

Acknowledgments We would like to thank Greg Bourne for stimulating discussions and the Australian Research Council for financial Support.

References

1. Xu J (2002) A new approach to finding natural chemical structure classes. J Med Chem 45:5311–5320
2. Horton DA, Bourne GT, Smythe ML (2003) The combinatorial synthesis of bicyclic privileged structures or privileged substructures. Chem Rev 103:893–930
3. Bondensgaard K, Ankersen M, Thogersen H, Hansen BS, Wulff BS, Bywater R (2004) Recognition of privileged structures by g-protein coupled receptors. J Med Chem 47:888–899
4. Muller G (2003) Medicinal chemistry of target family-directed masterkeys. Drug Discov Today 8:681–691
5. Kuntz ID (1972) Protein folding. J Am Chem Soc 94:4009–4012
6. Joseph D, Petsko GA, Karplus M (1990) Anatomy of a protein conformational change: hinged "lid" motion of the triosephosphate isomerase loop. Science 249:1425–1428
7. Jones S, van Heyningen P, Berman HM, Thornton JM (1999) Protein-DNA interactions: a structural analysis. J Mol Biol 287:877–896
8. Wu SJ, Dean DH (1996) Functional significance of loops in the receptor binding domain of *Bacillus thuringiensis* CryIIIA δ-endotoxin. J Mol Biol 255:628–640
9. Wlodawer A, Miller M, Jakolski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent SBH (1989) Conserved folding in retrovial protease: crystal structure of a synthetic HIV-protease. Science 245:616–621
10. Lu Y, Valentine JS (1997) Engineering metal-binding sites in proteins. Curr Opin Struct Biol 7:495–500
11. Bajorath J, Sheriff S (2001) Comparison of an antibody model with an X-ray structure; the variable fragment of BR96. Proteins 24:152–157
12. Kinoshita K, Sadanami K, Kidera A, Go N (1999) Structural motif of phosphate binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes. Protein Eng 12:11–14
13. Perona JJ, Craik CS (1995) Structural basis of substrate specificity in the serine proteases. Protein Sci 4:337–360
14. Tyndall JDA, Pfeiffer B, Abbenante G, Fairlie DP (2005) Over one hundred peptide-activated g protein-coupled receptors recognize ligands with turn structure. Chem Rev 105:793–826
15. Lewis PN, Momany FA, Scheraga HA (1973) Chain reversals in proteins. Biochim Biophys Acta 303:211–229
16. Hobohm U, Scharf M, Schneider R, Sander C (1992) Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1:409–417
17. Hobohm U, Sander C (1994) Enlarged representative set of protein structure. Protein Sci 3:522–524
18. Berstein FC, Koetzle TF, Williams GJB, Edgar F, Meyer J, Brice MD, Kennard O, Shimanouchi T, Tasumi M (1977) The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542
19. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins Struct Funct Genet 23:566–579
20. Ripka AS, Rich DH (1998) Peptideomimetic design. Curr Opin Chem Biol 2:441–452
21. Damewood JR (1996) Peptide mimetic design with the aid of computational chemistry. In: Lipkowitz KB, Boyd DB (eds) Reviews in computational chemistry, 9th edn. VCH Publishers, New York, pp 1–79
22. Anderberg MR (1973) Cluster analysis for applications. Academic Press, New York and London
23. Forgy EW (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics 21:768
24. Sokal RR, Michener CD (1958) A statiscal method for evaluating systematic relationships. Univ Kans Sci Bull 38:1409–1438
25. SAS/STAT (1999) User's guide, Volume 1, ANOVA-FREQ, Version 6. 4 edn
26. Manly BFJ (1994) Multivariate statistical method, a primer, 2nd edn. Chapman & Hall, London
27. Belov VN, Funke C, Labahn T, Es-Sayed M, de Meijere A (1999) Cyclopropyl building blocks in organic synthesis, 50—an easy access to bicyclic peptides with an octahydro[2H]pyrazino[1, 2-a]pyrazine skeleton. Eur J Org Chem 6:1345–1356
28. Golebiowski A, Klopfenstein SR, Chen JJ, Shao X (2000) Solid supported high-throughput organic synthesis of peptide b-turn mimetics via Petasis reaction/diketopiperazine formation. Tetrahedron Lett 41:4841–4844
29. Golebiowski A, Jozwik J, Klopfenstein SR, Colson A-O, Grieb AL, Russell AF, Rastogi VL, Diven CF, Portlock DE, Chen JJ (2002) Solid-supported synthesis of putative peptide β-turn mimetics via Ugi reaction for diketopiperazine formation. J Comb Chem 4:584–590
30. Souers AJ, Ellman JA (2001) β-Turn mimetic library synthesis: scaffolds and applications. Tetrahedron 57:7431–7448
31. Kim H-O, Nakanishi H, Lee MS, Kahn M (2000) Design and synthesis of novel conformationally restricted peptide secondary structure mimetics. Org. Lett. 2:301–302

32. Wels B, Kruijtzer JAW, Liskamp RMJ (2002) Synthesis of cyclic ($\alpha 2\beta$)-tripeptides as potential peptide turn mimetics. Org Lett 4:2173–2176
33. Pinnen F, Zanotti G, Lucente G (1984) Ten-membered cyclotripeptides: influence of the ring-flexibility on intramolecular reactions. Tetrahedron Lett 25:5201–5204
34. Eguchi M, Lee MS, Nakanishi H, Stasiak M, Lovell S, Kahn M (1999) Solid-phase synthesis and structural analysis of bicyclic β-turn mimetics incorporating functionality at the i to i +3 Positions. J Am Chem Soc 121:12204–12205
35. Eguchi M, Lee MS, Stasiak M, Kahn M (2001) Solid-phase synthesis and solution structure of bicyclic β-turn peptidomimetics: diversity at the i position. Tetrahedron Lett 42: 12347–1239
36. Golebiowski A, Klopfenstein SR, Shao X, Chen JJ, Colson A-O, Grieb AL, Russell AF (2000) Solid-supported synthesis of a peptide b-turn mimetic. Org Lett 2:2615–2617
37. Ripka WCDL, GV, Bach II AC, Pottorf RS, Blaney JM (1993) Protein β-turn mimetics I. Design, synthesis, and evaluation in model cyclic peptides. Tetrahedron 49:3593–3608
38. Chianelli D, Kim Y-C, Lvovskiy D, Webb TR (2003) Application of a novel design paradigm to generate general nonpeptide combinatorial scaffolds mimicking beta turns: synthesis of ligands for somatostatin receptors. Bioorg Med Chem Lett 11: 5059–5068
39. Im I, Webb TR, Gong Y-D, Kim J-I, Kim Y-C (2004) Solid-phase synthesis of tetrahydro-1, 4-benzodiazepine-2-one derivatives as a β-turn peptidomimetic library. J Comb Chem 6:207–213
40. Min BJ, Gu XY, Lee YS, Petrov RR, Mayorov AV, Hruby VJ (2006) Design and synthesis of bicyclic internal beta-turn mimetics and their applications toward biologically interesting ligands. Biopolymers 80:506
41. Egner U, Muller-Fahrnow A, Eckle E (1999) Turn mimetics for peptide design. Sci Meet 51:95–99
42. Kahn M, Wilke S, Chen B, Fujita K (1988) Nonpeptide mimetics of β-turns: a facile oxidative intramolecular cycloaddition of an azodicarbonyl system. J Am Chem Soc 110:1638–1639
43. Su T, Nakanishi H, Xue L, Chen B, Tuladha S, Johnson ME, Kahn M (1993) Nonpeptide β-turn mimetics of enkephalin. Bioorg Med Chem Lett 3:835–840
44. Fecik RA, Frank KE, Gentry EJ, Menon SR, Mitscher LA, Telikepalli H (1998) The search for orally active medications through combinatorial chemistry. J Comb Chem 18:149–184
45. Horton DA, Bourne GT, Smythe ML (2000) Exploring privileged structures: The combinatorial synthesis of cyclic peptides. Mol Divers 5:289–304
46. Ripka WC, De Lucca GV, Bach AC, Pottorf RS, Blaney JM (1993) Protein β-turn mimetics ii: design, synthesis, and evaluation in the cyclic peptide gramicidin S. Tetrahedron 49:3609–3628
47. Ball JB, Alewood PF (1990) Conformation constraints: nonpeptide β-turn mimics. J Mol Recognit 3:55–64
48. Kahn M, Chen B (1987) Methodology for the synthesis of mimetics of peptide β-turns. Tetrahedron Lett 28:1623–1626
49. Kahn M, Chen B, Zieske P (1987) The design and synthesis of nonpeptide mimic of erabutoxin. Heterocycles 25:29–31
50. Gardner B, Nakanishi H, Kahn M (1993) Conformationally constrained nonpeptide β-turn mimetics of enkephalin. Tetrahedron 49:3433–3448
51. Reddy DS, Vander Velde D, Aube J (2004) Synthesis and conformational studies of dipeptides constrained by disubstituted 3-(aminoethoxy)propionic acid linkers. J Org Chem 69:1716–1719
52. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: A program for macromolecular energy, minimization and dynamics calculations. J Comput Chem 4:187–217
53. Smellie A, Kahn SD, Teig SL (1995) Analysis of conformational coverage. 1. Validation and estimation of coverage. J Chem Inf Comput Sci 35:285–294
54. Smellie A, Kahn SD, Teig SL (1995) Analysis of conformational coverage. 2. Applications of conformational models. J Chem Inf Comput Sci 35:295–304
55. Douglas AJ, Mulholland G, Walker B, Guthrie DJS, Elmore DT, Murphy RF (1988) The preparation of a C-terminal gastrin peptide containing a synthetic B-bend mimetic. Biochem Soc Trans 16:175–176
56. Li W, Burgess K (1999) A new solid-phase linker for Suzuki coupling with concomitant macrocyclization: synthesis of beta-turn mimics. Tetrahedron Lett 40:6527–6530
57. Halab L, Lubell WD (1999) Use of steric interactions to control peptide turn geometry. Synthesis of type VI beta-turn mimics with 5-ter-butylproline. J Org Chem 64:3312–3321
58. Terrett N (1999) Combinatorial chemistry—combinatorial sarcodictyin libraries. Drug Discov Today 4:141
59. Rosenquist S, Souers AJ, Virgilio AA, Schurer SS, Ellman JA (1999) Synthesis, screening, and optimization of libraries of medium ring heterocyclic beta-turn mimetics. Abstr Papers Am Chem Soc 217:212
60. Gardner RR, Liang GB, Gellman SH (1999) Beta-turn and beta-hairpin mimicry with tetrasubstituted alkenes. J Am Chem Soc 121:1806–1816
61. Lombardi A, D'Auria G, Maglio O, Nastri F, Quartara L, Pedone C, Pavone V (1998) A novel rigid beta-turn molecular scaffold. J Am Chem Soc 120:5879–5886
62. Fink BE, Kym PR, Katzenellenbogen JA (1998) Design, synthesis, and conformational analysis of a proposed type I beta-turn mimic. J Am Chem Soc 120:4334–4344
63. Tran TT, Treutlein HR, Burgess AW (2001) Conformational analysis of thiopeptides: derivation of Sp2 sulfur parameters for the CFF91 force field. J Comput Chem 22:1010–1025
64. Tran TT, Treutlein HR, Burgess AW (2001) Conformational analysis of thiopeptides: (ϕ, ψ) maps of thio substituted dipeptides. J Comput Chem 22:1026–1037
65. Tran TT, Treutlein HR, Burgess AW, Perich J (2001) Synthesis, X-ray crystallographic structures of thio substituted N-acetyl N'-methylamide alanine and testing of sp2 sulfur parameters of the CFF91 force field. J Pept Res 58:67–78
66. Tran TT, Burgess AW, Treutlein HR, Zeng J (2001) Conformational analysis of thiopeptides: (ϕ, ψ) conformational free energy map of thio substituted alanine dipeptides. J Mol Graph Model 20:247–258
67. Tran TT, Burgess AW, Treutlein HR, Zeng J (2002) Effects of thioamide substitution on the conformation and relative stability of α- and 3_{10}- helices. J Am Chem Soc 124:5222–5230
68. Tran TT, Treutlein HR, Burgess AW (2006) Designing amino acid residues with single-conformations. Protein Eng Des Sel 19:401–408
69. Hutchinson EG, Thornton JM (1994) A revised set of potentials for β-turn formation in proteins. Protein Sci 3:2207–2216
70. Tran TT, McKie J, Meutermans WDF, Bourne GT, Andrews PR, Smythe ML (2005) Topological side-chain classification of β-turns: ideal motifs for peptidomimetic development. J Comput Aided Mol Des 19:551–566
71. SciFinder Scholar (2006) Copyright by The American Chemical Society: Washington
72. Poupaert J, Carato P, Colacino E (2005) 2(*3H*)-benzoxazolone and bioisosters as "privileged scaffold" in the design of pharmacological probes. Curr Med Chem 12:877–885
73. Flohr S, Stengelin S, Gossel M, Klabunde T, Stahl P, Safar P, Spoonamore J, Smrcina M, Klein JT, Merriman GH, Whiteley BK, Lanter C, Bordeau KJ, Yang Z (2004) Preparation of

hexahydropyrazino[1,2-a]pyrimidine-4,7-diones as anorectic agents. WO 2004-EP770 20040129
74. Schaper W, Willms L, Rosinger C, Hacker E, Rose E, Schmutzler D (2005) Preparation of quinoxalin-2-one derivatives as herbicide safeners. WO 2005112630; US 2005256000; DE 102004023332
75. Unger L, Raschack M, Wernet W, Boehm H-J, Riechers H (1995) Preparation of annelated 2-oxopiperazine endothelin antagonists. DE 4341663
76. Epperson JR, Hewanwasam P, Meanwell NA, Boissard CG, VK, G (1993) Post-Munson, D. Synthesis and excitatory amino acid pharmacology of some novel quinoxalinediones. Bioorg Med Chem 3:2801–2804
77. Joergensen AS, Stidsen CE, Faarup P, Groenvald FC (1991) Preparation of 1-carboxyalkyl-2,3-dioxoquinoxalines as glycine antagonists. WO 9113878
78. Patchett AA, Nargund RP (2000) Privileged structures—an update. Annu Rep Med Chem 35:289–298
79. Evans BE, Rittle KE, Bock MG, Diapardo RM, Freidinger RM, Whitter WL, Lundell GF, Veber DF, Anderson PS, Chang RSL, Lotti VJ, Cerino DJ, Chen TB, Kling PJ, Kunkel KA, Springer JP, Hirshfield J (1988) Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. J Med Chem 31:2235–2246
80. Amblard M, Martinez J, Berge G (2001) Preparation of oligomers of nonpeptide restricted mimetics of dipeptides or tripeptides and their use in the synthesis of proteins and polypeptides. WO 2001051506
81. Rolland C, Gozalbes R, Nicolay E, Paugam M-F, Coussy L, Bardosa F, Horvath D, Revah F (2005) G-protein-coupled receptor affinity prediction based on the use of a profiling Dataset: QSAR design, synthesis, and experimental validation. J Med Chem 48:6563–6574
82. Webb T, Chianelli D, Kim Y-C (2003) Preparation of quinoline-spiro-imidazolidinedione somatostatin analogs as agonists, partial agonists, or antagonists of somatostatin receptors. WO 20030 90677
83. Horton DA, Bourne GT, Smythe ML (2002) Exploring privileged structures: the combinatorial synthesis of cyclic peptides. J Comput Aided Mol Des 16:415–430
84. Golebiowski A, Klopfenstein SR (2001) Preparation of peptide β-turn mimetic compounds. WO 2000-US34832 20001220
85. Feng Y, Burgess K (1999) Solid-phase SNAr macrocyclizations to give turn-extended-turn peptidomimetics. Chem Eur J 5: 3261–3672
86. Feng Y, Wang Z, Jin S, Burgess K (1998) SNAr cyclizations to form cyclic peptidomimetics of β-turns. J Am Chem Soc 120: 10768–10769
87. Dobson CM (2004) Chemical space and biology. Nature 432: 824–828