# Secbase: Database Module To Retrieve Secondary Structure Elements with Ligand Binding Motifs

Oliver Koch,[†] Jason Cole,[‡] Peter Block,[†] and Gerhard Klebe*,[†]

Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany, and The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, U.K.

Received July 16, 2009

Secbase is presented as a novel extension module of Relibase. It integrates the information about secondary structure elements into the retrieval facilities of Relibase. The data are accessible via the extended Relibase user interface, and integrated retrieval queries can be addressed using an extended version of Reliscript. The primary information about α-helices and β-sheets is used as provided by the PDB. Furthermore, a uniform classification of all turn families, based on recent clustering methods, and a new helix assignment that is based on this turn classification has been included. Algorithms to analyze the geometric features of helices and β-strands were also implemented. To demonstrate the performance of the Secbase implementation, some application examples are given. They provide new insights into the involvement of secondary structure elements in ligand binding. A survey of water molecules detected next to the N-terminus of helices is analyzed to show their involvement in ligand binding. Additionally, the parallel oriented NH groups at the α-helix N-termini provide special binding motifs to bind particular ligand functional groups with two adjacent oxygen atoms, e.g., as found in negatively charged carboxylate or phosphate groups, respectively. The present study also shows that the specific structure of the first turn of α-helices provides a suitable explanation for stabilizing charged structures. The magnitude of the overall helix macrodipole seems to have no or only a minor influence on binding. Furthermore, an overview of the involvement of secondary structure elements with the recognition of some important endogenous ligands such as cofactors shows some distinct preference for particular binding motifs and amino acids.

## INTRODUCTION

As the function of a protein is determined by its 3D structure, the spatial architecture of functionally related proteins is more conserved in evolution than their actual amino-acid sequences. It seems that about 1000 relevant protein folds are observed across all proteins.[1] Proteins with similar fold but dissimilar sequences could have evolved from the same ancestor, and although they are involved in different biochemical reactions, they could still bind similar ligands[2] or show similar biochemical activities involved in different biological functions.[1] Thus, identifying proteins from different families with a similar folding pattern could suggest unexpected cross-reactivity or could be exploited for the discovery of novel lead structures. Furthermore, local similarity in conserved motifs can occur in globally deviating structures,[3] particularly next to active sites or binding pockets which are usually targeted in drug design.

At present, we witness an exponential increase in experimentally determined crystal structures of proteins and protein−ligand complexes. Accordingly, we need powerful database tools to retrieve, analyze, and correlate the plethora of structural information becoming available. Relibase has been developed as object-oriented data management system[4] that stores structures of protein−ligand complexes deposited in the PDB. It is particularly focused on aspects considering the protein−ligand interface and gives the user easy access to information about ligand binding modes across multiple database entries. As described, often the question arises whether structural features seen in ligands are reflected by specially recurring structural elements in proteins.

We therefore embarked onto the development of Secbase, a modular extension of Relibase. It integrates information about secondary structural elements assigned to individual proteins. Secbase first provides the user means to analyze protein−ligand interactions with respect to secondary structure elements, and, second, it allows analyses and discovery of functional similarity within related folding patterns, particularly with respect to structure-based drug design and molecular modeling.

**Secondary Structure Elements.** With respect to protein architecture, there are three types of structural elements that contribute to their construction.[5] Helices and β-sheets, already proposed by Pauling and Corey in 1951,[6−8] are responsible for neutralizing the highly polar main chain with respect to its hydrogen-bonding facilities in the hydrophobic interior of proteins. They exhibit clearly defined torsion angles, and for this reason they are called "*regular secondary structure elements*". In contrast, "*irregular*" turns cover a wide range of possible torsion angle and give the protein chain the opportunity to fold back upon itself.

* Corresponding author phone: +49 6421 282 1313; fax: +49 6421 282 8994; e-mail: klebe@mailer.uni-marburg.de.
[†] Philipps-Universität Marburg.
[‡] The Cambridge Crystallographic Data Centre.

10.1021/ci900202d CCC: $40.75 
Published on Web 10/07/2009

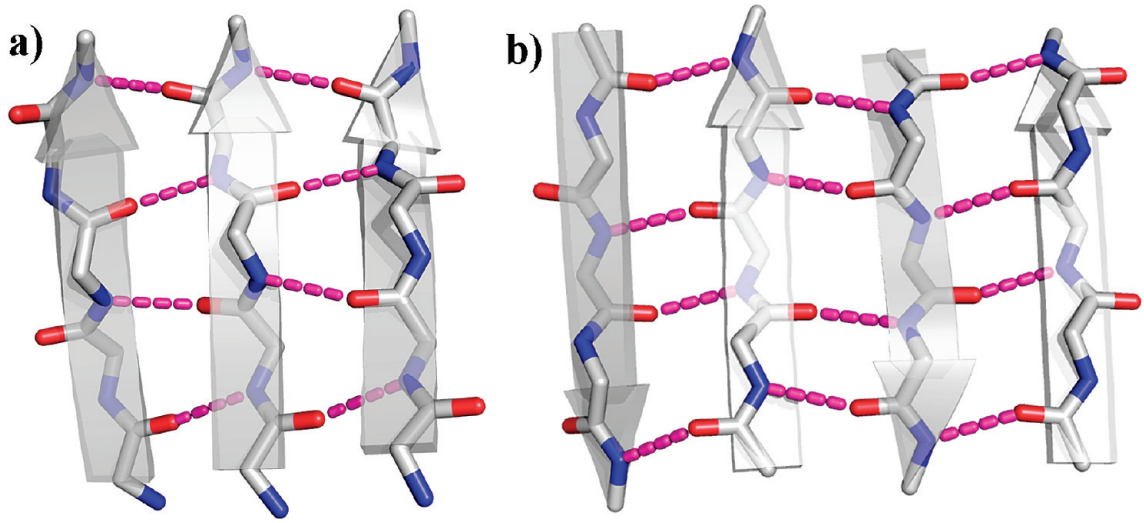



**Figure 1.** (a) Parallel and (b) antiparallel $\beta$-sheet with main chain hydrogen bonds (magenta).

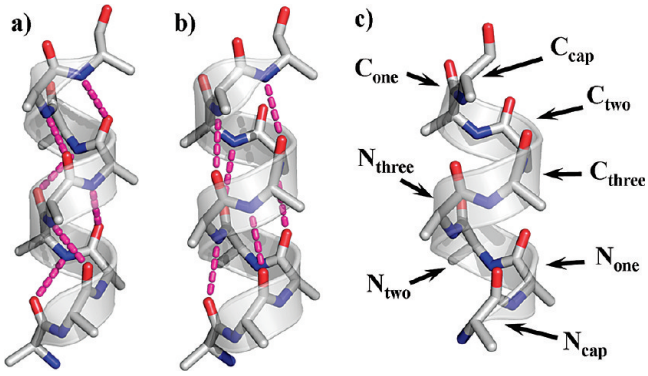



**Figure 2.** (a) $3_{10}$-helix and (b) α-helix with main chain hydrogen bonding (magenta) and (c) different positions at helix terminus.

$\beta$-Sheets are built-up from adjacent $\beta$-strands that are connected via hydrogen bonds between main chain functional groups. $\beta$-Strands normally have a length between five and ten amino acids. There are different types of $\beta$-sheets, and their classification is based on the spatial direction of the adjacent $\beta$-strands and the generated hydrogen bonding pattern. So-called parallel $\beta$-sheets (Figure 1a) contain $\beta$-strands oriented into the same spatial direction with an evenly spaced hydrogen-bonding pattern between one residue on the first $\beta$-strand and two residues on the adjacent $\beta$-strand. Within an antiparallel $\beta$-sheet (Figure 1b) $\beta$-strands show reverse orientation with a parallel hydrogen-bonding pattern between one residue on either of the neighboring strands. This leads to alternating pairs of parallel narrow- and wide-spaced hydrogen bonds.

Helices are mainly stabilized by intrahelical hydrogen bonds between the main chain CO-group of residue i and the main chain NH-group of residue i+n. This leads to three types of helices: the α-helix ($n = 4$, Figure 2a), the $\pi$-helix ($n = 5$), and the $3_{10}$-helix ($n = 3$, Figure 2b), and each of them is described by specific properties.

The side chains attached to the Cα atom are pointing away from the helix axis. The helix termini exhibit amide groups at the helix N-terminus and carbonyl groups at the C-terminus, respectively. They can be involved in hydrogen bonding. The helix boundary residues (Figure 2c) are called $N_{cap}$ and $C_{cap}$. In detail, the $N_{cap}$ position is the first residue with a nonhelical backbone conformation. The following or preceding residues are named: $N_{one}$, $N_{two}$, and $N_{three}$ at N-terminus of an α-helix and $C_{three}$, $C_{two}$, and $C_{one}$ at C-terminus of an α-helix (Figure 2c). This leads to four amide groups at the N-terminus and four carbonyl groups at the C-terminus not involved in intrahelical hydrogen-bonding.

Turns are irregular secondary structures with a hydrogen bond or a specific Cα-distance between the first and the last residue that is involved in building a turn.[9] The classification of the different turn families is based on the number of residues (2 to 6 residues) and the hydrogen bonding pattern. The normal conformation describes turns with a hydrogen bond between $CO_i$ and $NH_{i+n}$ (Figure 3b), and, in contrast, the reverse conformation contains a hydrogen bond between $NH_i$ and $CO_{i+n}$ (Figure 3a). Furthermore, conformations without a hydrogen bond but a specific $C\alpha_i$-$C\alpha_{i+n}$ distance cutoff exist. These are known as open conformations (Figure 3c). For further details about turns and the recent uniform classification see ref 9.

**Evidence for the Importance of Secondary Structure Elements with Respect to Ligand Binding.** In general, secondary structure elements are regarded as responsible for the rigid architecture of proteins and the correct orientation of side chains for ligand binding. However, they could also take significant influence on the physicochemical properties of functional groups adjacent or being part of the secondary structure elements. For example in α-helices and parallel $\beta$-sheets cooperative effects are present that result in a decrease of the mean hydrogen bond length with increasing length of the structural element.[10] Another example is the effect of α-helices on the p$K_a$ values of an amino acid at their termini. A mutagenesis study[11] of cysteines at different positions of the N-terminus of an α-helix has been performed in combination with an experimental determination of the thiol p$K_a$-values. It clearly showed a p$K_a$-shift at the $N_{cap}$ position in contrast to the p$K_a$-values at other positions or a control cysteine not involved in helices. In human thioredoxin the nucleophilic attack of a charged thiolate is essential for the functional role of this protein. The catalytic cysteine is located at the $N_{cap}$ position of a α-helix and shows a p$K_a$ of 6.3. Furthermore, p$K_a$ values of carboxylates such as aspartates and glutamates in proteins[12] reveal lower mean p$K_a$

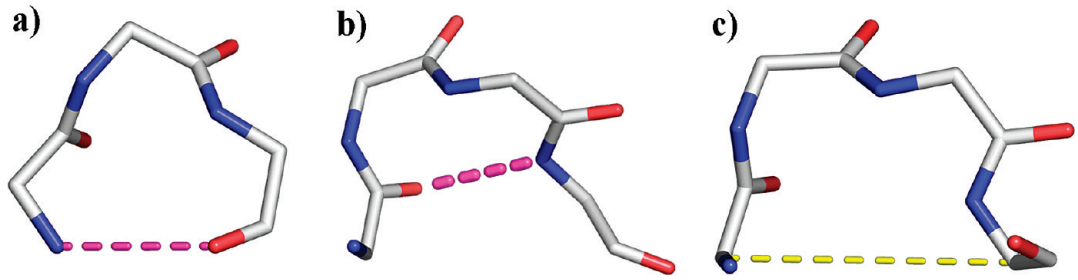



**Figure 3.** Different turn conformation (magenta: hydrogen bonding, yellow: Cα-Cα distance): (a) reverse turn with 3 residues, (b) normal $\beta$-turn (4 residues), and (c) open $\beta$-turn (4 residues).

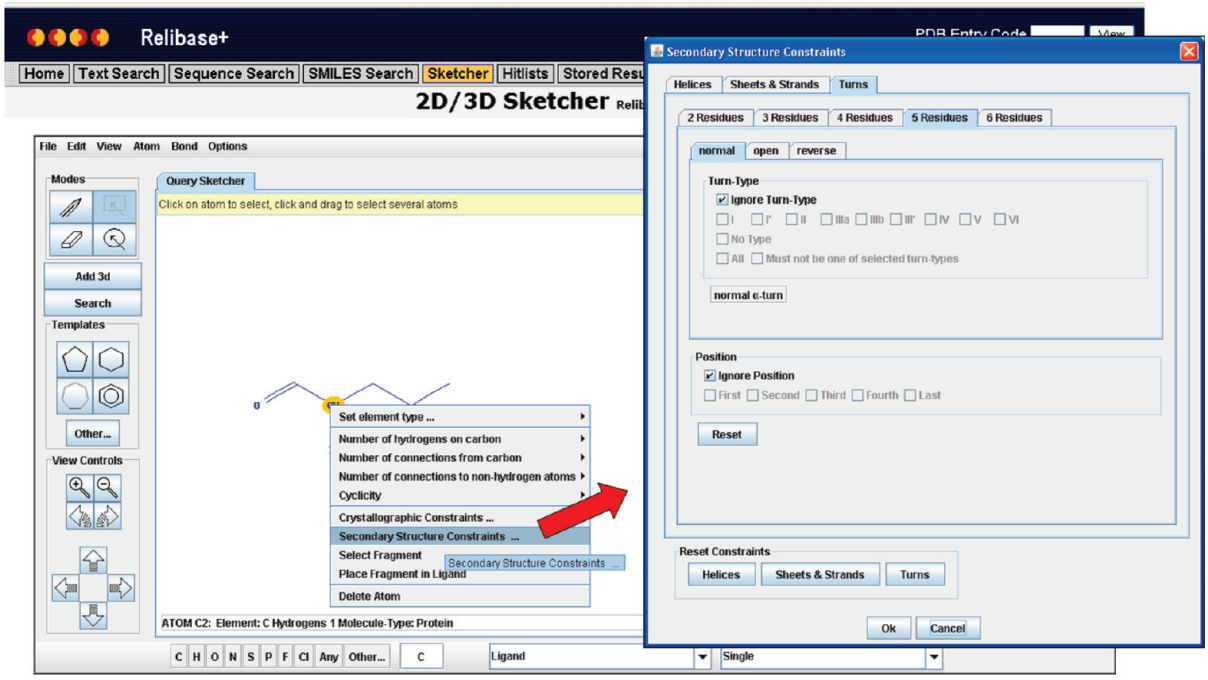



**Figure 4.** Relibase sketcher with additional Secbase constraints.

values at the N-termini of helices compared to the overall mean p$K_a$ values within proteins and model peptides.

Finally also turns are important from a functional point of view.[9] They can be involved in ligand binding, e.g., by presenting side chains in correct conformation. For example, the aspartate of the catalytic triad of trypsin-like serine proteases is part of a strong $\gamma$-turn.[13] Turns also participate in molecular recognition processes between proteins[14] or between a peptide substrate that mimics a turn conformation of a protein.[15]

These examples provide significant evidence that secondary structural elements have systematic influence on ligand binding. Furthermore, the systematic analysis of turns could possibly provide information for the design of novel ligands as turn mimetics.

## IMPLEMENTATION AND THEORY

**Secbase Implementation.** Secbase assigns the information about helices, $\beta$-sheets, and turns to existing Relibase data. Furthermore, algorithms are integrated that identify kinks and bends within helices and $\beta$-sheets. The information about $\beta$-sheets is retrieved from the corresponding pdb-file. The information about turns are processed from the protein structures and assigned based on the classification described elsewhere.[9] The hydrogen bonds of the normal and reverse turn families were determined using an implementation of the DSSP energy-function from Kabsch and Sander[16] and open turns are recognized based on a $C\alpha_i$-$C\alpha_{i+n}$ distance cutoff of 10 Å. These broad sets of turn structures were then classified into smaller sets based on the main-chain torsion angles using the previously described ESOM-Maps.[9]

For the information about α-helices two different data sources are integrated. On the one hand, the information provided by the corresponding pdb files is used. On the other hand, a new helix assignment algorithm was developed to identify helices within the protein structures that is based on the newly classified turns. This new algorithm leads to a more consistent assignment especially of the C-terminus of helices in comparison to the pdb assignment. A detailed description of this new assignment will be discussed in detail elsewhere.

The Relibase code was extended, so that the secondary structure element data can be used as constraints during substructure searches, accessed through Reliscript and viewed within the Astex viewer.

**Substructure Search.** The Relibase Web interface is a powerful tool for analyzing protein−ligand complexes. In particular the sketcher option could be used for identifying specific interactions. It provides access to 2D/3D ligand substructures and nonbonded protein−ligand interaction

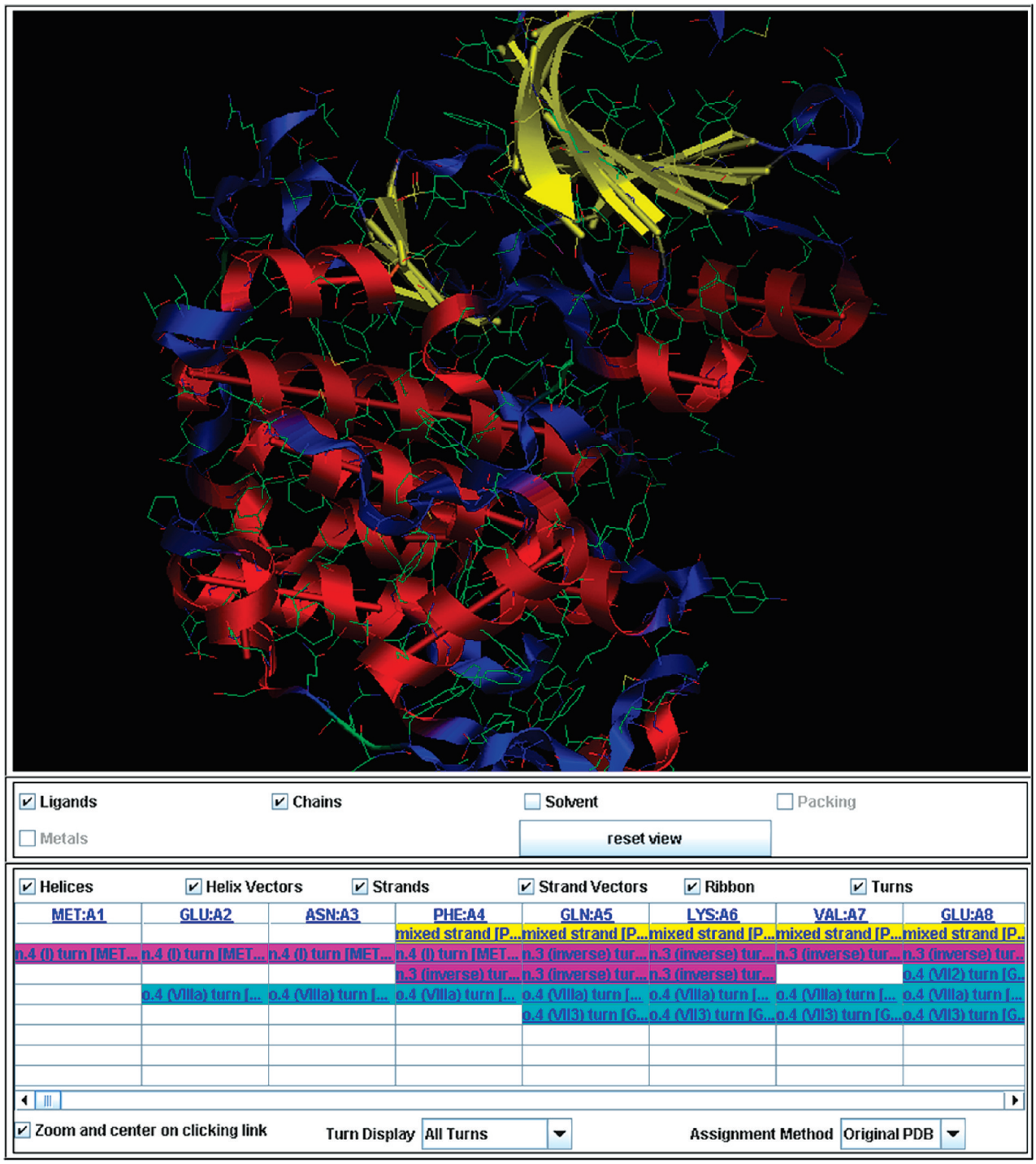



**Figure 5.** Example protein (1vft) viewed in Astex viewer.

searches. The Relibase code was extended, so that Secbase constraints could be used during substructure and nonbonded protein−ligand interaction searches. For example, one can constrain a search such that only protein substructures are hit that occur at the $N_{Cap}$ position of an α-helix with a length of e.g. ten residues. Other typical Secbase constraints that have been implemented are the following:

- helix type: α-helix, $\pi$-helix, or $3_{10}$-helix with a minimum and maximum length
- helix position: $N_{Cap}$, $N_{One}$, $N_{Two}$, $N_{Three}$, $C_{Cap}$, $C_{One}$, $C_{Two}$, $C_{Three}$, and other
- sheet type: parallel, antiparallel, or mixed $\beta$-sheet with a minimum and maximum number of $\beta$-strands
- at a kink among a α-helix or a $\beta$-strand
- specific turn families and types as described by Koch and Klebe[9]

These Secbase constraints can be easily defined within the sketcher for substructure searches (Figure 4).

**Reliscript.** Reliscript is the Python based command-line interface to access the entire functionality of Relibase. It allows more complex queries to be constructed using Python and the Relibase search functions, and it has been extended to make the Secbase information also accessible. The information about the secondary structure elements is stored in terms of Python objects. The main object type is the SSE object that stores the basic information and the other

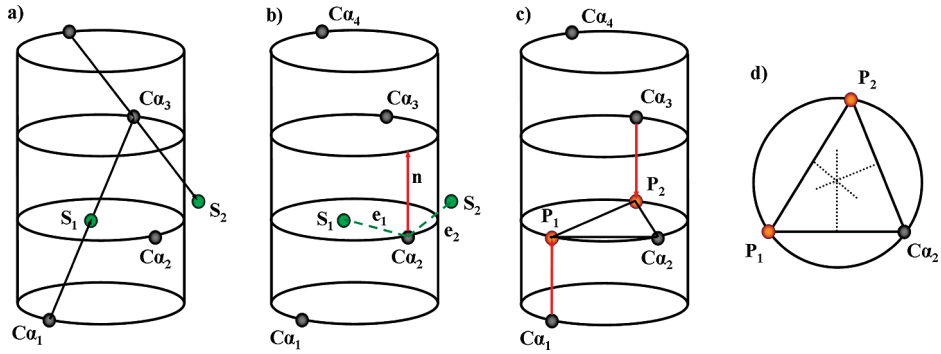



**Figure 6.** Midpoint calculation of four Cα of a α-helix: a) projection of two support points on a plane with $C\alpha_2$, b) calculation of the normal vector, c) parallel translation of two points on the surface of the helix, and d) midpoint calculation.

secondary structure element objects. The base objects are STRAND, HELIX, and TURN, and each object contains the information about an individual element. Additionally, STRAND and HELIX objects also have the described geometric description together with information about kinks. To get access to the secondary structure elements of a protein the corresponding PDB object contained a new attribute ('.sse') that comprises a list of SSE objects.

**Viewing Secondary Structure Assignments.** The secondary structure assignments of a protein can be viewed within the Web interface using the AstexViewer together with a controller that allows customization of the view (see Figure 5). In addition, a secondary structure browser permits navigation of secondary structure elements in the protein. Each cell in the scrollable table corresponds to a secondary structure assignment to a given amino acid in the protein and gives a short overview about the type of secondary structure. The different turn categories (normal, open, and reverse), helices, and strands are colored differently. Furthermore, it is possible to switch between the helix-assignment from the original PDB file and the new helix assignment.

**Algorithms to Evaluate the Geometry of Helices and Strands.** To evaluate the architecture of α-helices and *β*-strands the following descriptions were implemented to analyze their geometry in terms of vectors and to detect kinks and bends.

α-*Helices and Kinks.* To assign a central vector to each helix, the helix is assumed to be a cylinder with the Cα-atoms lying on the surface of this cylinder (Figure 6). The distances between the Cα-atoms are equal, both in horizontal and vertical directions. The first step is a projection of two points on a plane with $C\alpha_2$ and perpendicular to the helix axis (Figure 6a). The point $S_1$ is lying exactly at half distance between $C\alpha_1$ and $C\alpha_3$ and $S_2$ is the extension of the vector between $C\alpha_4$ and $C\alpha_3$, whereas $C\alpha_3$ is the midpoint between $S_2$ and $C\alpha_4$. The three points $C\alpha_2$, $S_1$, and $S_2$ span a plane that is perpendicular to the helix axis, and for this reason the normal vector $\vec{n}$ of the plane is parallel to the helix axis. Vector $\vec{n}$ is calculated as the vector product of the two vectors $\overrightarrow{S_1C\alpha_2}$ and $\overrightarrow{S_2C\alpha_2}$ (Figure 6b). Using the normal vector and a parallel translation of the points $C\alpha_1$ and $C\alpha_3$, this leads to a triangle that is perpendicular to the helix axis and whose vertices $C\alpha_2$, $P_1$, and $P_2$ are lying on the surface of the helix cylinder (Figure 6c). The perpendicular bisectors of the sides intersect each other in one point, and this point is equal to the midpoint of the helix axis that is based on the four Cα atoms (Figure 6d). An example for a calculated midpoint is shown in Figure 7a.

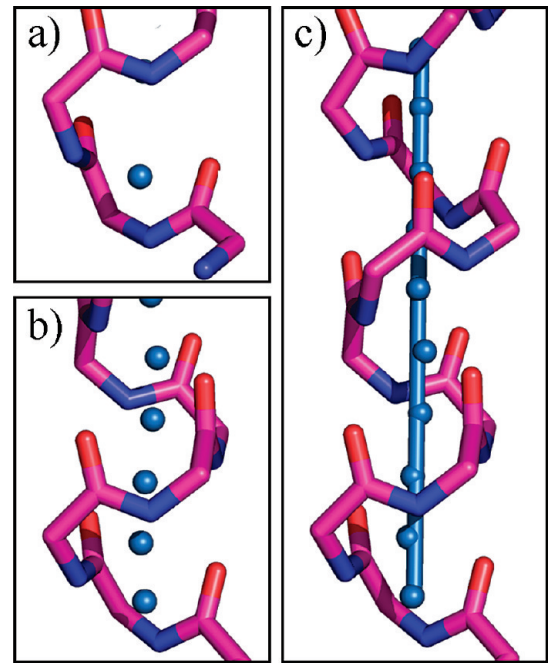



**Figure 7.** Geometric description of α-helices: a) single midpoint for 4 adjacent Cα atoms, b) all midpoints for overlapping Cα, and c) a vector description drawn through the calculated midpoints.

Proceeding with this algorithm along the whole helix with overlapping four Cα atom section leads to a description of the helix through midpoints (Figure 7b) and finally, a vector could be fitted through this midpoints (Figure 7c).

Additionally, kinks within a helix can be identified. For each midpoint the angle between the vector to the previous midpoint and the vector to the following vector is calculated. If this angle is more acute than a predefined threshold, the helix classified as kinked. This threshold angle is set to a default of 150°, but it can be set manually within Reliscript.

*β-Strands and Kinks.* The geometric description of *β*-strands is less difficult than identifying kinks and bends in a *β*-strand. The vector that is used to describe a *β*-strand is drawn through the Cα-atoms of the *β*-strand. To identify kinks and a bent *β*-strand the main chain atoms C, Cα, and N of one residue and N of the next residue are considered. The angle between the vectors through $N_i/C\alpha_i$ and $N_{i+1}/C_i$ shed light

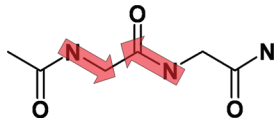



**Figure 8.** Vectors for identifying $\beta$-strand conformation.

on the conformation of the $\beta$-strand (Figure 8). If the angle is smaller than 86°, the $\beta$-strand is assigned as kinked at this position (Figure 9b). Although the linear part should be expected to have the largest angle, the $\beta$-strand is bent, if the angle is >122° (Figure 9c). Only parts of the polypeptide chain with an angle in between these two cut-offs show an appropriate conformation to build a linear $\beta$-strand (Figure 9a).

The information about kinks and bends is used to divide the $\beta$-strand into different parts. For each part a characteristic vector is calculated. A $\beta$-strand is split up if a kink can be identified (Figure 10a) or if a long bent part is observed. Bent parts are approximated using a simple geometric description: The bent part is split into different smaller parts of similar lengths (Figure 10b).

The final cutoff values for a bend (>122°) and a kink (<86°) are based on an analysis of the rmsd-values between the corresponding geometric description and the main-chain atoms. This analysis suggests, purely from an heuristic point-of-view, a minimum for the rmsd-values, and the number of vectors can be assigned using a cutoff value of 86° for kinked and 122° for bent conformations. Similarly to the helix cutoff value for a kink, these predefined threshold values can be overwritten manually in Reliscript.

## RESULTS APPLYING SECBASE

**Helix Propensities.** Knowledge about different amino acid propensities in helices, $\beta$-sheet, and random coils has been available for decades.[17] It has been adjusted several times to account for the growing number of structurally characterized proteins.

The conformational parameter P[18,19] describes the amino acid propensity within different secondary structure elements and is built up from the frequency ($f$) and the average frequency ($\langle f \rangle$)

$$P_{sse} = \frac{f_{sse}}{\langle f_{sse} \rangle}$$

The frequency of a specific residue is defined as

$$f_{sse} = \frac{n_{i,sse}}{n_{i,p}} = \frac{number\ of\ residue\ i\ in\ sse}{total\ number\ of\ residue\ i\ in\ proteins}$$

and the average frequency of finding all amino acid residues is defined as

$$\langle f_{sse} \rangle = \frac{n_{tot,sse}}{n_{tot,p}} = \frac{total\ number\ of\ all\ residues\ in\ sse}{total\ number\ of\ all\ residues\ in\ proteins}$$

With the help of Reliscript and a small Python script it is straightforward to calculate this conformational parameter P. The example script in Figure 11 counts the amino acid types within $\beta$-strands, which is needed for the calculation of the confirmative parameter P:

For the calculation of the conformational parameter of the α-helices a nonredundant data set was used that is based on the pdb-select list[20] and contains 2257 protein chains and 9163 α-helices with a minimum length of 8 residues (Tables 1 and 2). The minimum length was chosen so that the N-terminus and C-terminus in the helix do not overlap. Terminus regions span 4 amino acid residues, so in smaller α-helices residues in the middle of the α-helix could be assigned to either terminal end as either N- or C-terminus residues. For example in an α-helix with 6 residues the third residue ($N_{two}/C_{three}$) and the fourth residue ($N_{three}/C_{two}$) are belonging both to the N- and C-terminus. Table 3 shows the result for α-helices with a length of 6 residues. Comparing the values in all three tables reveals different amino acid propensities, and therefore the data extracted from smaller helices should be neglected with respect to a general conformational analysis, since the N-terminus and C-terminus are not independent.

Tables 1 and 2 compare these results to a previous survey of Penel et al.[21] including 2102 helices from 298 protein structures and a survey of Aurora et al.[22] including 1316 helices from 274 polypeptide chains. Penel et al.[21] analyzed the N-terminus and the middle part of α-helices and Aurora et al.[22] both termini. Both data sets contain only chains with a sequence homology smaller than 25%. In general, the results for the N-terminus and residues within the α-helix from Penel et al.[21] agree well with our results, with the exception of the cysteine propensity, which is found to be higher in our case. In contrast, the results from Aurora et al. differ widely from this study and the investigation of Penel et al. at the N-terminus. With respect to the analysis of the C-terminus our results also show differences to the Penel study. This may reflect the lower count of helices considered in the older analysis.

The analysis of the N-terminus reveals no surprises; most amino acid propensities are now well understood:[21] The high propensities of Asp and Asn at the $N_{Cap}$ position are based on their hydrogen bonding potential to interact with the free NH group of the $N_{two}$ position. Ser and Thr can be hydrogen bonded to the $N_{three}$ position. Glu, Gln, and Asp at the $N_{two}$ position can form a hydrogen bond to their own backbone NH groups and Glu and Gln at the $N_{three}$ position can form a hydrogen bond to the free NH group at the $N_{Cap}$ position. Additionally, side chain-side chain interactions are favorable between certain residues with high occurrence propensities ($N_{cap}$–$N_{three}$: Ser/Thr–Glu/Asp/Gln, $N_{Cap}$–$N_{two}$: Ser/Thr–Glu/Ser and Arg–Asp). The mostly low occurrence of Gly, Ile, Leu, Met, and Val can be explained by their nonpolarity and the usually high solvent exposure at N-terminus.

Ala shows a high propensity throughout the whole helix, which is due to the short side chain which shows minimum loss of configurational entropy on folding.[21] In contrast, Gly shows a low propensity throughout the whole helix, which could be explained by its enhanced backbone flexibility and, in consequence, the higher loss of backbone conformational entropy compared to other residues.[23] The only exception of this is found at the $C_{cap}$ with the highest propensity for Gly, since Gly is capable of adopting unusual backbone torsion angles. Such unusual geometries are needed to provide a free CO group for hydrogen bonding to the free NH group of the following residues.[24] Special conformations

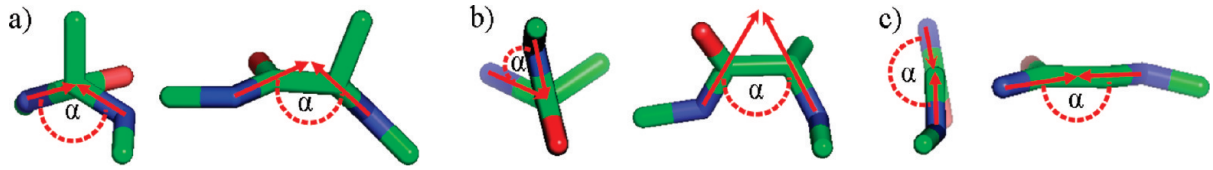



**Figure 9.** a) Linear, b) kinked, and c) bended parts of polypeptide chain.

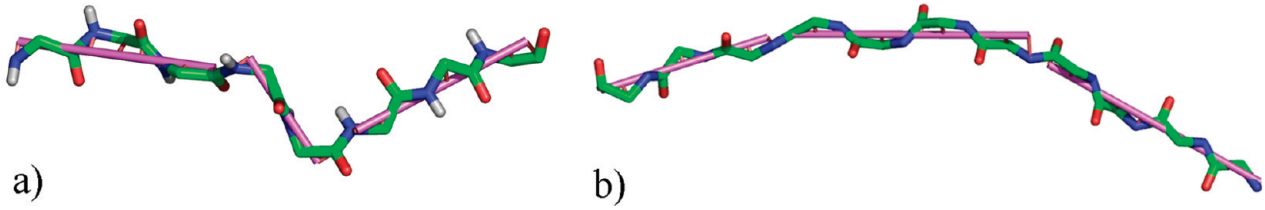



**Figure 10.** Geometric description (purple sticks) of a) a kinked and b) a bended $\beta$-strand.

```
[1] import reliscript
[2] pdb = reliscript.create('1ets')
[3] strand = {}
[4] for sse in pdb.sse:
[5]   for sht in sse.sheets:
[6]     for str in sht.strands:
[7]       for res in str.residues:
[8]         strand[res.name] = strand.get(res.name, 0) +1
```

**Figure 11.** Python script that use reliscript for calculating the amino acid distribution in $\beta$-strands.

**Table 1.** Amino Acid Propensities of the N-Terminus in Comparison to the Former Analyses of Penel et al. (P) and Aurora et al. (A)[a]

| aa | $N_{cap}$ | P | A | $N_{one}$ | P | A | $N_{two}$ | P | A | $N_{three}$ | P | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALA | 0.5 | 0.5 | 0.67 | 1.11 | 1.2 | 1.10 | 1.25 | 1.3 | 1.39 | 1.23 | 1.3 | 1.43 |
| ILE | 0.27 | 0.2 | 0.78 | 0.87 | 0.9 | 1.06 | 0.56 | 0.5 | 0.64 | 0.76 | 0.6 | 1.18 |
| LEU | 0.4 | 0.5 | 0.79 | 1.01 | 0.8 | 0.84 | 0.64 | 0.6 | 0.91 | 0.93 | 0.8 | 1.52 |
| MET | 0.54 | 0.5 | 0.98 | 0.77 | 0.9 | 0.90 | 0.66 | 0.5 | 1.10 | 1.1 | 1.0 | 1.68 |
| PHE | 0.49 | 0.4 | 0.96 | 0.89 | 0.9 | 0.90 | 0.69 | 0.7 | 1.00 | 0.92 | 1.0 | 1.10 |
| PRO | 1.28 | 1.5 | 1.12 | 2.46 | 2.3 | 1.67 | 0.92 | 1.0 | 0.94 | 0.49 | 0.5 | 0.15 |
| VAL | 0.31 | 0.3 | 0.67 | 0.83 | 0.8 | 0.76 | 0.62 | 0.6 | 0.70 | 0.95 | 0.9 | 1.14 |
| ARG | 0.64 | 0.6 | 0.76 | 1.05 | 0.9 | 1.05 | 0.94 | 0.8 | 0.95 | 0.79 | 0.8 | 1.33 |
| ASP | 2.46 | 2.4 | 1.58 | 0.98 | 0.8 | 1.14 | 1.59 | 1.8 | 1.64 | 1.48 | 1.6 | 0.90 |
| GLU | 0.61 | 0.6 | 0.94 | 1.4 | 1.3 | 2.30 | 2.54 | 2.6 | 2.07 | 2.0 | 2.3 | 1.70 |
| LYS | 0.63 | 0.5 | 0.84 | 1.11 | 1.0 | 1.08 | 1.05 | 0.9 | 0.80 | 0.81 | 0.8 | 0.82 |
| ASN | 2.18 | 2.0 | 1.28 | 0.54 | 0.5 | 0.72 | 0.81 | 0.8 | 0.67 | 0.63 | 0.6 | 0.55 |
| CYS | **0.98** | **0.5** | 0.37 | 0.45 | 0.6 | 0.26 | 0.42 | 0.6 | 0.52 | 0.62 | 0.6 | 0.52 |
| GLN | 0.66 | 0.4 | 1.05 | 1.16 | 1.3 | 1.31 | 1.33 | 1.4 | 1.60 | 1.64 | 1.8 | 1.43 |
| HIS | 0.99 | 1.0 | 0.83 | 0.72 | 0.7 | 0.83 | 0.91 | 1.0 | 1.36 | 0.93 | 0.8 | 0.66 |
| SER | 2.54 | 2.4 | 1.25 | 0.82 | 1.0 | 0.81 | 1.04 | 1.1 | 0.69 | 0.82 | 0.7 | 0.61 |
| THR | 2.15 | 2.1 | 1.41 | 0.77 | 0.8 | 0.77 | 0.82 | 0.7 | 0.92 | 1.02 | 1.0 | 0.75 |
| TRP | 0.5 | 0.6 | 0.94 | 1.11 | 1.2 | 1.26 | 0.82 | 0.9 | 1.10 | 0.97 | 1.1 | 1.68 |
| TYR | 0.58 | 0.6 | 0.82 | 0.93 | 1.0 | 0.99 | 0.69 | 0.5 | 0.73 | 0.95 | 0.9 | 0.65 |
| GLY | 1.18 | 1.3 | 0.98 | 0.62 | 0.7 | 0.55 | 0.72 | 0.7 | 0.56 | 0.52 | 0.6 | 0.37 |

[a] Large deviations are indicated by underlining (Aurora et al. compared to both other analysis) or in bold (Penel et al. compared to this analysis).

at the end of an α-helix are called "capping motifs", since they cause stabilizing effects at the end of an α-helix (e.g.: the Schellman motif or the αL-motif[25,26]). The low occurrence of Thr, Val, and Ile at the C-terminus can be explained by the bulky, $\beta$-branched side chain which leads to a poorer solvation of the free CO groups.

Pro cannot form a hydrogen bond, and only the $\varphi$ angle is constrained to helical conformation; the Ψ angle shows an extended, nonhelical conformation.[21] The high occurrence at the $N_{one}$ position can be explained due to these properties.

The higher amino acid propensity of Cys found in this study is not surprising since Cys can form a hydrogen bond to free NH-groups at the N-terminus and shows good $N_{Cap}$ preferences in experimental studies.[24]

**Table 2.** Amino Acid Propensities at the C-Terminus in Comparison to the Former Analysis of Aurora et al. (A) and the Residues within the α-Helix (Other) in Comparison to the Former Analysis of Penel et al. (P)[a]

| aa | other | P | $C_{three}$ | A | $C_{two}$ | A | $C_{one}$ | A | $C_{cap}$ | A |
|---|---|---|---|---|---|---|---|---|---|---|
| ALA | 1.51 | 1.6 | 1.71 | 1.73 | 1.43 | 1.33 | 1.47 | 1.87 | 1.09 | 1.19 |
| ILE | 1.3 | 1.3 | 1.12 | 1.15 | 1.08 | 1.58 | 0.71 | 0.90 | 0.49 | 0.61 |
| LEU | 1.43 | 1.5 | 1.73 | 1.80 | 1.45 | 1.63 | 1.54 | 1.65 | 1.06 | 1.36 |
| MET | 1.39 | 1.6 | 1.52 | 2.21 | 1.23 | 1.76 | 1.26 | 1.35 | 1.01 | 1.35 |
| PHE | 1.05 | 1.1 | 1.06 | 1.35 | 0.83 | 1.22 | 0.92 | 0.67 | 0.91 | 1.20 |
| PRO | 0.09 | 0.1 | 0.08 | 0.07 | 0.03 | 0.07 | 0.0 | 0.03 | 0.0 | 0.10 |
| VAL | 1.03 | 1.1 | 0.76 | 0.94 | 0.86 | 1.08 | 0.58 | 0.51 | 0.5 | 0.46 |
| ARG | 1.3 | 1.3 | 1.31 | 1.24 | 1.5 | 1.39 | 1.32 | 1.66 | 1.09 | 1.45 |
| ASP | 0.69 | 0.7 | 0.67 | 0.68 | 0.66 | 0.60 | 0.66 | 0.91 | 0.77 | 0.72 |
| GLU | 1.13 | 1.1 | 1.14 | 1.16 | 1.41 | 1.43 | 1.3 | 1.88 | 0.9 | 1.27 |
| LYS | 1.12 | 1.1 | 1.29 | 1.22 | 1.58 | 1.71 | 1.42 | 1.63 | 1.24 | 1.45 |
| ASN | 0.75 | 0.8 | 0.67 | 0.70 | 0.73 | 0.64 | 1.07 | 0.70 | 1.61 | 1.33 |
| CYS | 0.83 | 0.8 | 1.09 | 0.63 | 0.68 | 0.44 | 0.8 | 0.33 | 1.04 | 0.44 |
| GLN | 1.31 | 1.3 | 1.19 | 0.88 | 1.46 | 1.37 | 1.22 | 1.24 | 1.21 | 1.43 |
| HIS | 0.82 | 0.8 | 0.77 | 0.76 | 1.03 | 1.02 | 1.09 | 0.89 | 1.42 | 1.55 |
| SER | 0.67 | 0.7 | 0.63 | 0.65 | 0.76 | 0.42 | 1.06 | 0.71 | 1.16 | 1.02 |
| THR | 0.75 | 0.7 | 0.54 | 0.46 | 0.67 | 0.57 | 0.92 | 0.50 | 0.69 | 0.82 |
| TRP | 1.08 | 1.1 | 0.95 | 1.57 | 0.84 | 1.00 | 0.71 | 1.00 | 0.5 | 0.58 |
| TYR | 0.96 | 1.0 | 0.97 | 1.10 | 0.78 | 1.02 | 1.09 | 0.73 | 0.96 | 1.06 |
| GLY | 0.41 | 0.4 | 0.33 | 0.32 | 0.26 | 0.20 | 0.29 | 0.33 | 2.06 | 0.74 |

[a] Large deviations are indicated by underlining.

**Water and Ligand Interactions to Backbone Amides.** The Relibase Web interface with the sketcher option has been used to identify structural database entries with the searched specific interactions. Within this section, example studies are presented that show the power of Secbase to analyze the influence of secondary structure elements on ligand binding. The studies are focused on ligand interactions with backbone N−H amide groups.

An analysis of the interaction geometry between water, ligand oxygen atoms bound to phosphorus or carbon or water-mediated ligand binding to a backbone amide was performed using Relibase in combination with Secbase (Figure 12) to retrieve information about the general influence of secondary structure elements and the influence of cooperative effects in particular.

Figure 13 shows an example of water interacting with a backbone amide that fulfills the search constraints. The following Secbase constraints were applied to backbone amide groups to obtain the data with respect to a unique type of secondary structure element:

**Table 3.** Amino Acid Propensities for α-Helices with 6 Residues[a]

| aa | $N_{cap}$ | | $N_{one}$ | | $N_{two}/C_{three}$ | | $N_{three}/C_{two}$ | | $C_{one}$ | | $C_{cap}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALA | 54 | 0.71 | 83 | 1.09 | 88 | 1.15 | 69 | 0.9 | 75 | 0.98 | 93 | 1.22 |
| ILE | **21** | **0.35** | 51 | 0.85 | **31** | **0.52** | 72 | 1.2 | 68 | 1.13 | 42 | 0.7 |
| LEU | 69 | 0.75 | 96 | 1.05 | 77 | 0.84 | 104 | 1.13 | **174** | **1.9** | 97 | 1.06 |
| MET | 14 | 0.74 | 17 | 0.9 | **11** | **0.58** | **30** | **1.6** | **29** | **1.54** | 20 | 1.06 |
| PHE | 23 | 0.53 | 49 | 1.14 | 40 | 0.93 | **80** | **1.85** | **66** | **1.53** | 35 | 0.81 |
| PRO | 76 | 1.69 | **118** | **2.63** | 39 | 0.87 | **17** | **0.38** | **2** | **0.04** | **0** | **0.0** |
| VAL | **39** | **0.55** | 65 | 0.92 | 43 | 0.61 | 83 | 1.17 | 82 | 1.16 | 54 | 0.76 |
| ARG | 36 | 0.7 | 51 | 0.99 | 46 | 0.9 | 44 | 0.86 | 62 | 1.21 | 50 | 0.97 |
| ASP | **120** | **2.02** | 54 | 0.91 | **89** | **1.5** | 48 | 0.81 | 37 | 0.62 | 58 | 0.98 |
| GLU | 49 | 0.72 | 89 | 1.3 | **166** | **2.43** | 99 | 1.45 | 45 | 0.66 | 59 | 0.86 |
| LYS | 35 | 0.57 | 48 | 0.78 | 73 | 1.19 | 70 | 1.14 | 61 | 0.99 | 83 | 1.35 |
| ASN | **76** | **1.68** | **21** | **0.46** | 38 | 0.84 | 28 | 0.62 | 23 | 0.51 | 65 | 1.44 |
| CYS | **26** | **1.57** | 17 | 1.03 | 16 | 0.97 | **25** | **1.51** | 18 | 1.09 | 24 | 1.45 |
| GLN | 34 | 0.86 | 25 | 0.63 | 44 | 1.11 | 55 | 1.38 | 39 | 0.98 | 36 | 0.91 |
| HIS | 30 | 1.2 | 19 | 0.76 | 29 | 1.16 | 25 | 1.0 | 29 | 1.16 | 26 | 1.04 |
| SER | **108** | **1.76** | 69 | 1.12 | 60 | 0.98 | 50 | 0.81 | 46 | 0.75 | 73 | 1.19 |
| THR | 76 | 1.35 | 57 | 1.01 | 29 | 0.52 | 41 | 0.73 | 63 | 1.12 | **41** | **0.73** |
| TRP | 13 | 0.85 | **27** | **1.77** | 14 | 0.92 | 18 | 1.18 | 21 | 1.38 | **5** | **0.33** |
| TYR | 29 | 0.78 | 31 | 0.84 | 39 | 1.05 | 46 | 1.24 | 50 | 1.35 | 39 | 1.05 |
| GLY | 86 | 1.24 | **28** | **0.4** | 42 | 0.6 | **13** | **0.19** | **22** | **0.32** | **115** | **1.65** |
| OTH | 3 | 0.65 | 2 | 0.44 | 3 | 0.65 | 0 | 0.0 | 5 | 1.09 | 2 | 0.44 |
| ALL | 1017 | 1.0 | 1017 | 1.0 | 1017 | 1.0 | 1017 | 1.0 | 1017 | 1.0 | 1017 | 1.0 |

[a] Propensities smaller than 0.5 or bigger than 1.5 are shown in bold.

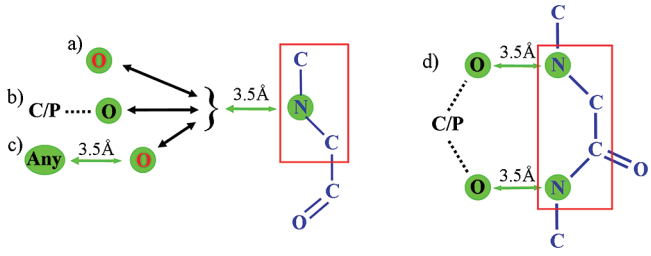



**Figure 12.** Analyzed interaction geometry of (a) a water molecule, (b) a ligand functional group, or (c) a ligand functional group mediated via a water molecule to the protein backbone N−H amide groups (blue), a maximum mutual distance of 3.5 Å (green arrows) is considered. Search query (d) involves interaction of two oxygen atoms covalently attached to C or P interacting with two neighbored backbone amides. Blue color indicates protein parts, black: ligand parts, red: water. The dotted lines represent any type of covalent bond, the red rectangle comprises the atoms used for structural superposition in the results.

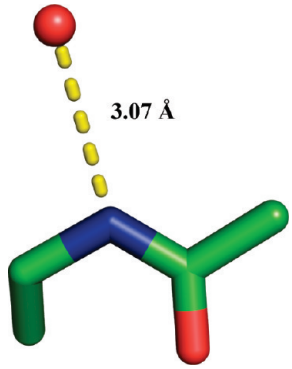



**Figure 13.** Water interacting with a backbone amide.

For random chain residues (*RCRs*):
- helix type: *must not be in any helix*
- sheet properties: *must not be in any sheet*

For α-helix:
- helix type: α *helix*
- $N_{cap}$ properties: each with one of: $N_{cap}$, $N_{one}$, $N_{two}$, or $N_{three}$

For *β*-sheets:
- sheet properties: *antiparallel* or *parallel*
- position: both *first* and *last* strand allowed

For further postprocessing steps, the resulting superimposed structures were used in combination with Python and Reliscript. These superimposed structures are automatically produced during the search process. During the postprocessing analysis, hydrogens of backbone amides were manually set according to McDonald and Thornton.[27] An in-house Python interface for Gnuplot (http://www.gnuplot.info) with standard diagram types was used to automatically generate diagrams shown in the following analysis.

*Water-Backbone Amide.* Forsyth showed[12] the lack of a correlation between the different p$K_a$ values at the helix N-terminus and an increasing magnitude of the helix macrodipole. However, it is of interest, whether the identified cooperative effects in α-helices and parallel *β*-sheets[10] have any impact on the binding geometry of adjacent water molecules. For this purpose, Relibase/Secbase retrieval software was used to extract all water molecules found in a sphere with a 3.5 Å radius next to a backbone amide group embedded into different secondary structure element (Figure 12a).

It is quite evident, that the number for each of the N-terminal positions is equal to the number of α-helices considered, but the number of water molecules considered can differ for deviating N-terminal positions (Table 4).

**Table 4.** Number of Hits for Water Molecules near a Backbone Amide with Respect to the Embedding Secondary Structure Elements and the Number of Different Chains and Different Families

| | hits | chains | families |
|---|---|---|---|
| RCR[a] | 1841292 | 51135 | 8331 |
| $N_{cap}$[b] | 39430 | 13614 | 1686 |
| $N_{one}$[b] | 70580 | 15589 | 1978 |
| $N_{two}$[b] | 54983 | 14410 | 1728 |
| $N_{three}$[b] | 20622 | 9857 | 1216 |
| antiparallel[c] | 58679 | 13912 | 1883 |
| parallel[c] | 9804 | 4430 | 604 |

[a] RCR: random chain residues. [b] $N_{cap}/N_{one}/N_{two}/N_{three}$: α-helix N-terminus position. [c] Antiparallel *β*-sheet and parallel *β*-sheet.

The $N_{one}$ position seems to accommodate a water molecule most frequently, possibly to achieve its residual solvation. In contrast, the $N_{three}$ position is the most unlikely one to interact with a water molecule. Supposedly, these observations result from the fact that the $N_{three}$ backbone amide is at an excellent position to serve as a counter group for interactions with side chain functional groups. This is in agreement with the described high propensity of particular residues at the $N_{cap}$ position (e.g., Ser and Thr, see previous study) which form a hydrogen bond to the accessible backbone N−H amide group at the $N_{three}$ position. In contrast, the $N_{one}$ position seems to lack similar interaction partners, accordingly it is available with high frequency for interactions with water molecules. Another interesting fact is the pronounced difference between parallel and antiparallel *β*-sheets which still seeks a reasonable explanation. A straightforward explanation could be that in general the solvent-accessibility of backbone amides is higher for antiparallel *β*-sheets than for parallel *β*-sheets. A second perhaps more satisfactory explanation arises from the above-

**Table 5.** Amino Acid Distribution for Water near the Free Backbone Amide Group[a]

| | RCR no. | [%] | $N_{cap}$ no. | [%] | N | $N_{one}$ no. | [%] | N | $N_{two}$ no. | [%] | N | $N_{three}$ no. | [%] | N | anti no. | [%] | para no | [%] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALA | 150177 | 8.2 | **2759** | **7.0** | **1.8** | **8660** | **12.3** | **1.5** | **7682** | **14.0** | **1.5** | **3171** | **15.4** | **1.7** | 2963 | 5.0 | 499 | 5.1 |
| ILE | 74521 | 4.0 | 1038 | 2.6 | 1.4 | 3440 | 4.9 | 1.0 | 1531 | 2.8 | 0.8 | 722 | 3.5 | 0.7 | 4066 | 6.9 | 803 | 8.2 |
| LEU | 124054 | 6.7 | 2081 | 5.3 | 1.3 | 6422 | 9.1 | 1.0 | 3063 | 5.6 | 0.9 | 1707 | 8.3 | 1.0 | 5117 | 8.7 | 1004 | 10.2 |
| MET | 26764 | 1.5 | 548 | 1.4 | 1.3 | 1217 | 1.7 | 1.2 | 817 | 1.5 | 1.2 | 356 | 1.7 | 0.8 | 1015 | 1.7 | 155 | 1.6 |
| PHE | 68457 | 3.7 | 1074 | 2.7 | 1.3 | 2929 | 4.1 | 1.1 | 1595 | 2.9 | 0.9 | 684 | 3.3 | 0.8 | 2786 | 4.7 | 775 | 7.9 |
| PRO | 45598 | 2.5 | **781** | **2.0** | **0.3** | **2490** | **3.5** | **0.3** | **295** | **0.5** | **0.1** | **163** | **0.8** | **0.4** | 474 | 0.8 | 44 | 0.4 |
| VAL | 105750 | 5.7 | 1434 | 3.6 | 1.5 | 4515 | 6.4 | 1.1 | 2350 | 4.3 | 1.0 | 775 | 3.8 | 0.5 | 6659 | 11.3 | 1539 | 15.7 |
| ARG | 88537 | 4.8 | 1535 | 3.9 | 1.2 | 3317 | 4.7 | 0.9 | 2472 | 4.5 | 1.0 | 982 | 4.8 | 1.2 | 2756 | 4.7 | 310 | 3.2 |
| ASP | 140365 | 7.6 | 5010 | 12.7 | 0.9 | 4968 | 7.0 | 1.2 | 5625 | 10.2 | 1.1 | 1571 | 7.6 | 0.9 | 2291 | 3.9 | 474 | 4.8 |
| GLU | 119367 | 6.5 | 2016 | 5.1 | 1.2 | 6672 | 9.5 | 1.0 | 7916 | 14.4 | 0.9 | 2213 | 10.7 | 0.8 | 3642 | 6.2 | 392 | 4.0 |
| LYS | 104168 | 5.7 | 1446 | 3.7 | 1.0 | 4365 | 6.2 | 1.0 | 3621 | 6.6 | 1.0 | 1045 | 5.1 | 1.0 | 3216 | 5.5 | 302 | 3.1 |
| ASN | 97141 | 5.3 | 2974 | 7.5 | 0.8 | 1897 | 2.7 | 1.2 | 1986 | 3.6 | 1.0 | 677 | 3.3 | 1.2 | 1830 | 3.1 | 405 | 4.1 |
| CYS | 22778 | 1.2 | 411 | 1.0 | 0.6 | 590 | 0.8 | 1.0 | 413 | 0.8 | 0.8 | 104 | 0.5 | 0.4 | 1342 | 2.3 | 195 | 2.0 |
| GLN | 61711 | 3.4 | 1167 | 3.0 | 1.1 | 3201 | 4.5 | 1.1 | 2790 | 5.1 | 1.0 | 743 | 3.6 | 0.6 | 2085 | 3.6 | 212 | 2.2 |
| HIS | 45967 | 2.5 | 936 | 2.4 | 1.0 | 1366 | 1.9 | 1.1 | 1120 | 2.0 | 0.9 | 429 | 2.1 | 0.9 | 1405 | 2.4 | 290 | 3.0 |
| SER | 141272 | 7.7 | 4885 | 12.4 | 0.9 | 3816 | 5.4 | 1.0 | 3594 | 6.5 | 1.0 | 1127 | 5.5 | 1.1 | 4416 | 7.5 | 533 | 5.4 |
| THR | 124215 | 6.7 | 2430 | 6.2 | 0.5 | 3061 | 4.3 | 1.0 | 2560 | 4.7 | 1.1 | 1180 | 5.7 | 1.1 | 6379 | 10.9 | 780 | 8.0 |
| TRP | 28514 | 1.5 | 527 | 1.3 | 1.5 | 1613 | 2.3 | 1.3 | 616 | 1.1 | 0.8 | 418 | 2.0 | 1.3 | 1569 | 2.7 | 107 | 1.1 |
| TYR | 60941 | 3.3 | 1313 | 3.3 | 1.5 | 2258 | 3.2 | 0.9 | 1310 | 2.4 | 0.9 | 781 | 3.8 | 1.0 | 2101 | 3.6 | 575 | 5.9 |
| GLY | 210993 | 11.5 | **5065** | **12.8** | **1.6** | **3783** | **5.4** | **1.3** | **3627** | **6.6** | **1.4** | **1774** | **8.6** | **2.6** | 2567 | 4.4 | 410 | 4.2 |

[a] RCR: random chain residues, α-helix N-terminus: $N_{cap}$, $N_{one}$, $N_{two}$, $N_{three}$, anti: antiparallel and para: parallel $\beta$-sheet; general high degree of variation is shown in bold.

mentioned differences in cooperative effects between both types of sheets that might serve as an explanation.[10]

The amino acid distribution of water-backbone interactions is shown in Table 5. The N-terminal values are normalized with respect to the percentage of expected amino acids across this distribution which is based on the overall amino acid propensity (see the previous section)

$$N = \frac{h_{i,pos}}{h_{pos}} \Big/ \frac{n_{i,pos}}{n_{pos}}$$

where $h_{i,pos}$ = number of hits for residue $i$ at position pos, $h_{pos}$ = overall number of hits at position $pos$, $n_{i,pos}$ = number of residue $i$ at position $pos$, and $n_{pos}$ = overall number of residues at position $pos$.

Unsurprisingly, small amino acids such as Ala and Gly appear as frequent interaction partners for water molecules. This is probably caused by a lack of a sterically demanding side chain. Pro cannot establish a hydrogen bond to water, and accordingly it shows generally a low interaction frequency.

As in a previously published analysis,[10] the hydrogen bond geometry was analyzed with respect to the occurrence in particular secondary structure elements and the number of neighboring cumulative hydrogen bonds within the element (data not shown). The expected cooperative effect generated by the cumulated H-bonds in α-helices and $\beta$-sheets is not indicated by the data; all geometries to accommodate water molecules are comparable. Of course, the assignment and thus accuracy of the location of water molecules in protein structures determined by X-ray diffraction has to be interpreted with some care, since not all water positions are well-defined in the density. Nevertheless, the results of this analysis do not provide significant evidence for a systematic influence of any cooperative effects in α-helices or parallel $\beta$-sheets on the binding geometry of water molecules.

For further analysis, Gly was separated from other amino acids as this residue shows quite distinct properties from those bearing a side chain. Figure 14 shows the projection of the identified water oxygens onto a plane perpendicular to the N−H bond (see the figure caption for details). All elements show a water cluster directly above or near the backbone amide (the origin of the figures). Furthermore, the α-helical N-terminal positions $N_{cap}$, $N_{one}$, and $N_{two}$ (Figure 14a-c) exhibit an additional water cluster in the direction of the subsequent position. In fact, these are waters from the water cluster directly above the backbone amide of the next residue. In contrast to the α-helical N-terminus, the $\beta$-strands do not exhibit a similar secondary water cluster (Figure 14f,g). Comparing the backbone conformations of these elements explains this observation. Different from the helix backbone, where all peptide N−H bond vectors are pointing in the same direction, the peptide N−H bonds within a strand are oriented in somewhat alternating directions (Figure 15). Furthermore, the distance between two neighbored amides at the kinked helix-terminus backbone is small enough for these secondary interactions.

Water molecules bound next to Gly residues show a less distinctive distribution; Figure 16 shows the distribution for $N_{cap}$ and random chain residues. In contrast to the broad distribution near random chain residues, the water clusters next to the N-terminal positions are shifted toward the second Cα hydrogen atom that occupies the position of the side chains found in other amino acids.

*Water-Mediated Ligand Binding to a Backbone Amide.* This analysis deals with water molecules that are involved in a water-mediated ligand-protein interaction via the backbone amide N−H group (Figure 12c). Across all secondary elements analyzed (Table 6), Gly shows the highest occurrence among all residues that are involved in this type of water-mediated ligand binding. Among these highest Gly occurrence appears at the $N_{cap}$ position of helices with over 77% followed by the $N_{three}$ position with ∼55%.

Interestingly, the positions of water molecules that act as mediators to the $N_{cap}$ position are shifted away from the projected position directly above the peptide N−H bond vector (Figure 17a), similarly to the previously described analysis. This explains the high occurrence of water molecules next to Gly, as all other residues would bear a

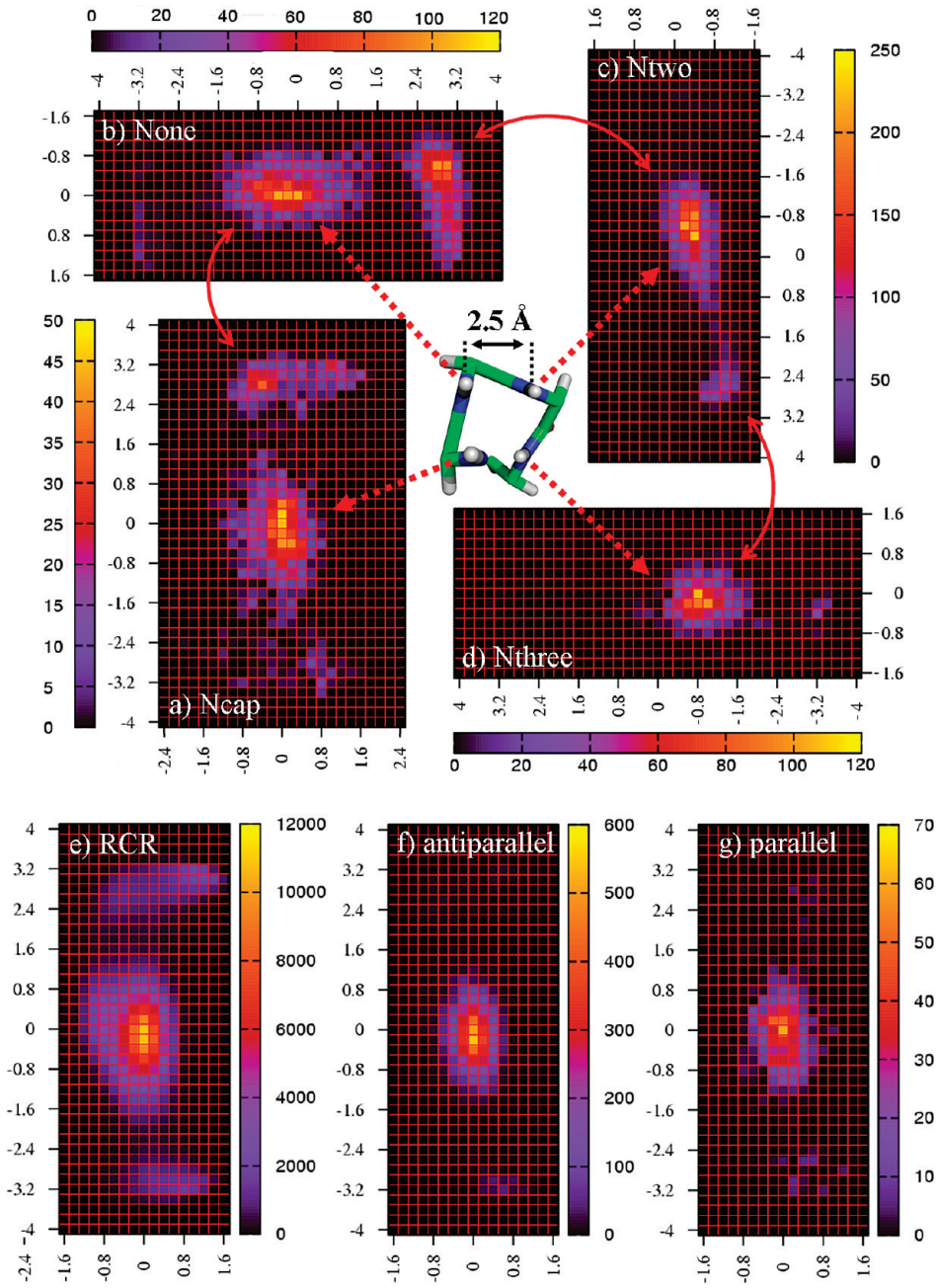
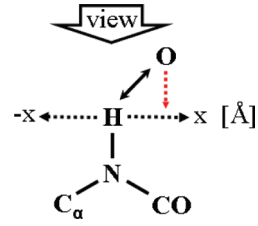



**Figure 14.** Projection of water-O onto a plane perpendicular to the N−H bond for all amino acids except Gly: a) $N_{cap}$: 2731 hits (number of hits according to color coding), b) $N_{one}$: 6445 hits, c) $N_{two}$: 5976 hits, d) $N_{three}$: 2398 hits, e) RCR: 374102 hits, f) antiparallel $\beta$-sheet: 14443, and g) parallel $\beta$-sheet: 1659 hits (H: describes the origin; positive $x$-axis points into N−C direction; axes: distance to H in Å; α-helical N-terminus positions are rotated to fit the terminus orientation as shown in the picture of α-helical N-terminus in the middle).




sterically crowded side chain that would repel a putatively bound water molecule. Comparing these results with Figure 16a indicates that the water molecules involved in ligand binding are responsible for most of the hits found near a Gly N−H amide backbone. The occurrence of water molecules near Gly that are not involved in ligand binding seems negligible. It appears that the detected water molecules operate as space fillers for the missing side chain present in the other amino acids at this position. This highlights the role of waters as a versatile particle to fill empty gaps in protein structures.

The projection of the position of water molecules mediating ligand binding to the $N_{one}$ position in all residues except Gly (Figure 17b) indicates that particularly water molecules

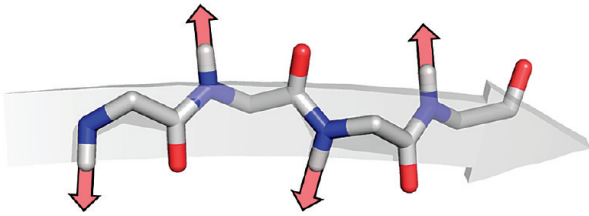


**Figure 15.** $\beta$-Strand with somewhat alternating peptide N−H bond directions (indicated by red arrows).

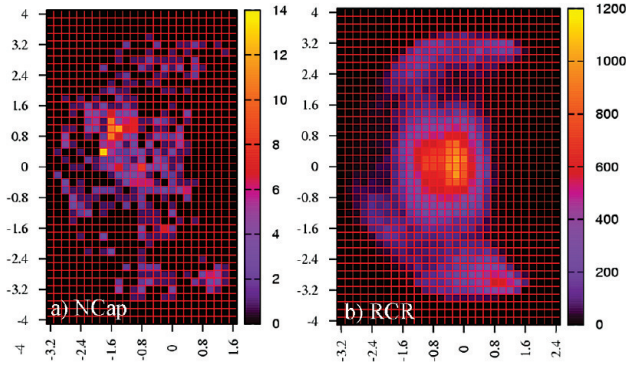



**Figure 16.** Projection of water-O for Gly, orientation, and color coding similar to Figure 14.

bound to the subsequent residue are interacting with the backbone functionality at this position. Although, the $N_{one}$ position shows the highest number of directly formed interactions to water molecules at the helix N-terminus, as apparent from Table 4, the number of water molecules directly interacting with the backbone N−H and a ligand atom at this position is quite low. This leads to the possible hypothesis that the $N_{one}$ position appears less likely to be occupied by water molecules mediating contacts to bound ligands. The remaining part of the analyzed positions show similar projections as described in Figure 14 (data not shown).

*Ligand-Backbone Amide.* The interaction of backbone amide groups not involved in inter-residue contacts to ligand oxygen atoms will be addressed in this final analysis (Figure 12b). Either carbon or phosphorus were chosen as covalently attached binding partners of the ligand's carbonyl-type oxygen. Most likely groups involving these bonds are negatively charged (e.g., carboxylate or phosphate groups). They are postulated to be stabilized by an α-helix macrodipole. Table 7 shows the distribution of oxygen at oms according to the Sybyl atom type notation for oxygens bound to central carbon atom (http://www.tripos.com/mol2/atom_types.html).

Regarding the interaction of N−H groups to oxygen atoms being part of the carboxylate groups (O.co2), the $N_{one}$ and $N_{two}$ positions show the highest relative involvement. At the $N_{cap}$ and $N_{three}$ positions the highest percentage for O.3-type oxygen atoms is found, and N−H groups in sheets prefer O.2-type oxygens. The high frequency of O.co2 possibly supports the idea of a more favorable accommodation of negatively charged groups at the N-terminus of the α-helix macrodipole. However, also another more geometric aspect should be considered as a possible explanation for the high occurrence of negatively charged groups at this site. As mentioned, secondary interactions formed to the neighbored backbone amide groups can provide an additional stabilizing effect. Accordingly, particularly the motif of two adjacent oxygen atoms as present in carboxylate groups can perfectly interact with two parallel aligned neighbored backbone amide groups (Table 8). This binding motif is not given at the rim of a $\beta$-strand of either parallel or antiparallel $\beta$-sheets. It could well be that the special geometry disposed at the terminal end of an α-helix is required to provide an appropriate binding pattern with two parallel NH groups to simultaneously recognize both oxygens of a carboxylate group. Compared to the number of helix N-terminal position, the number of chain residues in random geometries is much higher which is due to the fact that the polypeptide chain outside a helix or a $\beta$-strand can also show this specific kinked conformation.

A similar analysis of phosphate groups reveals also a high frequency of hits at the α-helix N-terminus once two oxygen atoms bound to the same phosphorus atom are requested to interact with two neighbored N−H backbone amide groups (see Table 9). Similarly to the study of carboxylate groups,

**Table 6.** Distribution of Residues That Are Involved in Water-Mediated Ligand Binding

| | RCR | | $N_{cap}$ | | $N_{one}$ | | $N_{two}$ | | $N_{three}$ | | $\beta$-sheet anti | | $\beta$-sheet para | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALA | 6993 | 8.2% | 9 | 1.5% | 14 | 3.2% | 74 | 11.7% | 44 | 11.6% | 152 | 9.3% | 10 | 6.1% |
| ILE | 4230 | 5.0% | 1 | 0.2% | 10 | 2.3% | 21 | 3.3% | 0 | 0.0% | 27 | 1.7% | 0 | 0.0% |
| LEU | 3768 | 4.4% | 7 | 1.2% | 10 | 2.3% | 13 | 2.1% | 16 | 4.2% | 112 | 6.9% | 29 | 17.6% |
| MET | 1383 | 1.6% | 1 | 0.2% | 18 | 4.1% | 18 | 2.8% | 8 | 2.1% | 42 | 2.6% | 5 | 3.0% |
| PHE | 2899 | 3.4% | 5 | 0.8% | 25 | 5.7% | 28 | 4.4% | 15 | 4.0% | 28 | 1.7% | 14 | 8.5% |
| PRO | 1174 | 1.4% | 5 | 0.8% | 21 | 4.8% | 0 | 0.0% | 0 | 0.0% | 12 | 0.7% | 0 | 0.0% |
| VAL | 4693 | 5.5% | 5 | 0.8% | 14 | 3.2% | 29 | 4.6% | 6 | 1.6% | 157 | 9.6% | 12 | 7.3% |
| ARG | 4091 | 4.8% | 17 | 2.8% | 57 | 13.0% | 22 | 3.5% | 0 | 0.0% | 96 | 5.9% | 3 | 1.8% |
| ASP | 5420 | 6.4% | 24 | 4.0% | 22 | 5.0% | 30 | 4.7% | 7 | 1.8% | 52 | 3.2% | 4 | 2.4% |
| GLU | 2306 | 2.7% | 2 | 0.3% | 4 | 0.9% | 16 | 2.5% | 2 | 0.5% | 62 | 3.8% | 0 | 0.0% |
| LYS | 2271 | 2.7% | 5 | 0.8% | 2 | 0.5% | 9 | 1.4% | 1 | 0.3% | 87 | 5.3% | 0 | 0.0% |
| ASN | 4102 | 4.8% | 6 | 1.0% | 10 | 2.3% | 82 | 12.9% | 0 | 0.0% | 81 | 5.0% | 18 | 10.9% |
| CYS | 1219 | 1.4% | 3 | 0.5% | 1 | 0.2% | 26 | 4.1% | 7 | 1.8% | 44 | 2.7% | 0 | 0.0% |
| GLN | 2264 | 2.7% | 0 | 0.0% | 4 | 0.9% | 2 | 0.3% | 2 | 0.5% | 41 | 2.5% | 0 | 0.0% |
| HIS | 1700 | 2.0% | 1 | 0.2% | 1 | 0.2% | 30 | 4.7% | 2 | 0.5% | 84 | 5.1% | 1 | 0.6% |
| SER | 7562 | 8.9% | 34 | 5.6% | 52 | 11.8% | 67 | 10.6% | 22 | 5.8% | 151 | 9.2% | 8 | 4.8% |
| THR | 6181 | 7.3% | 9 | 1.5% | 76 | 17.3% | 27 | 4.3% | 26 | 6.9% | 159 | 9.7% | 8 | 4.8% |
| TRP | 1124 | 1.3% | 2 | 0.3% | 12 | 2.7% | 4 | 0.6% | 0 | 0.0% | 9 | 0.6% | 1 | 0.6% |
| TYR | 2199 | 2.6% | 1 | 0.2% | 4 | 0.9% | 35 | 5.5% | 14 | 3.7% | 43 | 2.6% | 18 | 10.9% |
| GLY | **19575** | **23.0%** | **469** | **77.4%** | **82** | **18.7%** | **101** | **15.9%** | **207** | **54.6%** | **194** | **11.9%** | **34** | **20.6%** |
| ALL | 85154 | 100.0% | 606 | 100.0% | 439 | 100.0% | 634 | 100.0% | 379 | 100.0% | 1633 | 100.0% | 165 | 100.0% |

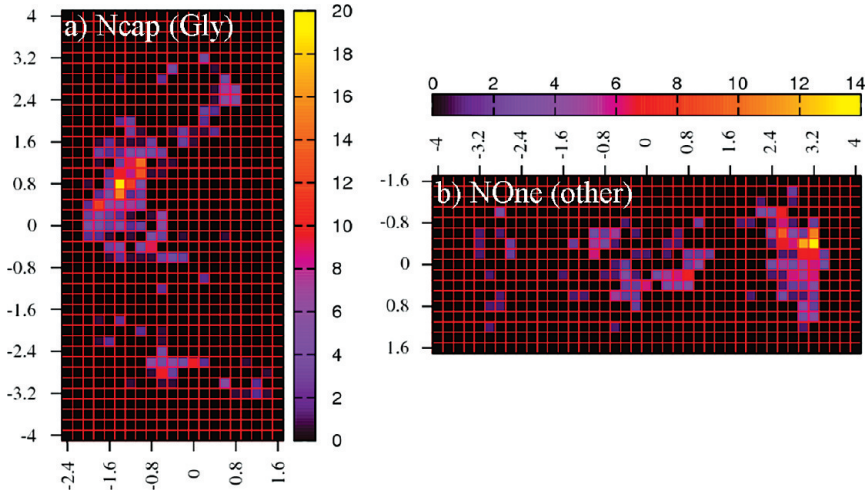



**Figure 17.** Projection of water molecules involved in water-mediated ligand binding: a) the $N_{cap}$ position at Gly and b) the $N_{one}$ position at all residues except Gly, orientation and color coding similar to Figure 14.

**Table 7.** Distribution of Oxygen Atoms, Bound to Carbon, with the Indicated Sybyl Atom-Type Notation

| | O.2 | | O.3 | | O.co2 | |
|---|---|---|---|---|---|---|
| | no. | % | no. | % | no. | % |
| RCR | 14840 | 34.48 | 18271 | 42.45 | 9930 | 23.07 |
| $N_{cap}$ | 309 | 19.36 | 864 | 54.14 | 423 | 26.5 |
| $N_{one}$ | 567 | 30.67 | 604 | 32.67 | 678 | 36.67 |
| $N_{two}$ | 158 | 18.72 | 355 | 42.06 | 331 | 39.22 |
| $N_{three}$ | 34 | 14.47 | 139 | 59.15 | 62 | 26.38 |
| antiparallel | 659 | 52.93 | 385 | 30.92 | 201 | 16.14 |
| parallel | 73 | 52.9 | 61 | 44.2 | 4 | 2.9 |

**Table 8.** Number of Observations Where Both Oxygen Atoms of a Carboxylate Group Are Involved in H-Bonds to Two Neighbored N−H Backbone Amides (See Figure 12c)

| | $N_{cap}-N_{one}$ | $N_{one}-N_{two}$ | $N_{two}-N_{three}$ | antiparallel | parallel | RCR |
|---|---|---|---|---|---|---|
| no. of hits | 48 | 126 | 40 | 0 | 0 | 1097 |

**Table 9.** Number of Hits of Phosphate Groups with Two Neighbored Backbone Amides[a]

| | $N'-N_{cap}$ | $N_{cap}-N_{one}$ | $N_{one}-N_{two}$ | $N_{two}-N_{three}$ | antiparallel | parallel |
|---|---|---|---|---|---|---|
| no. of hits | 1076 | 1789 | 1414 | 229 | 0 | 0 |

[a] Figure 12c, N′: position before $N_{cap}$.

the interaction motif of two parallel backbone N−H groups present at a kinked backbone position obviously also applies to bidentate contacts with phosphate groups.

The high frequency of Lys, Ser, Thr, and Gly involved in binding of phosphate groups (Table 10) appears reasonable, since they are known to be part of highly conserved phosphate binding motifs.[28] Interestingly, Ser and Thr are found with nearly no preference at any of the analyzed positions, whereas Lys only occurs at the $N_{one}$ position in a high number of cases. Presumably, Lys at the $N_{one}$ position is part of the phosphate-binding loop (P-Loop),[29] highly conserved in the sequence motif GXXXXGKT and GXXXXGKS, respectively. Ser or Thr, however, are detected with no preference for a particular structural element. They seem to be a preferred interaction partner for phosphate groups without the need for a specific structural environment as given by the P-loop for Lys. The overall interaction pattern for these structures appears quite similar at all different positions (Figure 18). In most cases, the oxygen atoms interact with both the backbone amide and the Ser oxygen. The analysis of carboxylate oxygens interacting with the backbone amide and the Ser oxygen and Thr oxygen, respectively, reveals a similar high occurrence of this interaction motif at the α-helix N-terminus (Ser: 598 hits [13,2%], Thr: 430 [9.5%]). Therefore, this seems to be a general interaction motif for oxygen attached to phosphate or carboxylate.

Finally, it is interesting to examine whether certain positions at the α-helix N-terminus are involved in binding of specific protein ligands such as cofactors. Detected hits were grouped according to the ligands classified in Table 11. An interaction from the $N_{three}$ backbone amide to ADP, GDP, and ATP is highly favored as reflected by the high frequency found for interactions to Thr. In contrast, Ser at the $N_{three}$ position tends to recognize PLP and FAD more frequently. Thr at the $N_{cap}$ position is highly involved in binding PLP, whereas Ser is more frequently involved in binding FAD and NADPH. PLP has been demonstrated to bind to the so-called CαNN structural motif.[28] This CαNN motif is known to exhibit the above-described interaction pattern of an oxygen atom that forms a bifurcated H-bond with either the backbone N−H and the side chain Oγ H of Thr.

This analysis might be biased due to the higher frequency of occurrence of certain classes of proteins. Nevertheless, half of all known proteins interact with structural motifs containing phosphates.[28] Accordingly, in light of this fairly large distribution the features of local interaction patterns are assumed to be statistically significantly distributed.

## DISCUSSION

The overall α-helix macrodipole as a putative feature to stabilize charges at the α-helix termini is often considered as a possible explanation for various electrostatic opportunities. This argument is particularly used in the context with the function of ion channels. For example, in the potassium channel a prominent binding position for solvated cations is provided at the apex of four helices arranged with their carbonyl terminal ends in a 4-fold circular arrangement. In

**Table 10.** Involvement of Specific Residue in Binding of P═O Ligand Functional Groups[a]

| | RCR | | $N_{cap}$ | | $N_{one}$ | | $N_{two}$ | | $N_{three}$ | | anti | | para | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALA | 2111 | 6.0% | 84 | 2.5% | 207 | 4.3% | 293 | 8.9% | 55 | 6.8% | 15 | 9.3% | 1 | 9.1% |
| ILE | 769 | 2.2% | 69 | 2.1% | 235 | 4.9% | 218 | 6.6% | 9 | 1.1% | 0 | 0.0% | 0 | 0.0% |
| LEU | 717 | 2.0% | 42 | 1.3% | 111 | 2.3% | 112 | 3.4% | 40 | 4.9% | 1 | 0.6% | 0 | 0.0% |
| MET | 456 | 1.3% | 24 | 0.7% | 94 | 2.0% | 68 | 2.1% | 7 | 0.9% | 1 | 0.6% | 0 | 0.0% |
| PHE | 712 | 2.0% | 16 | 0.5% | 211 | 4.4% | 52 | 1.6% | 12 | 1.5% | **26** | **16.1%** | 0 | 0.0% |
| PRO | 178 | 0.5% | 12 | 0.4% | 95 | 2.0% | 1 | 0.0% | 0 | 0.0% | 0 | 0.0% | 0 | 0.0% |
| VAL | 1282 | 3.7% | 65 | 1.9% | 231 | 4.8% | 169 | 5.1% | 29 | 3.6% | 0 | 0.0% | 0 | 0.0% |
| ARG | 2031 | 5.8% | 105 | 3.1% | 261 | 5.4% | 41 | 1.2% | 12 | 1.5% | 9 | 5.6% | 0 | 0.0% |
| ASP | 928 | 2.6% | 117 | 3.5% | 53 | 1.1% | 25 | 0.8% | 17 | 2.1% | 0 | 0.0% | 0 | 0.0% |
| GLU | 645 | 1.8% | 28 | 0.8% | 103 | 2.1% | 25 | 0.8% | 33 | 4.1% | 5 | 3.1% | 0 | 0.0% |
| LYS | 2047 | 5.8% | 195 | 5.8% | **1010** | **21.0%** | 27 | 0.8% | 6 | 0.7% | 1 | 0.6% | 6 | 54.5% |
| ASN | 942 | 2.7% | 85 | 2.5% | 108 | 2.3% | 85 | 2.6% | 8 | 1.0% | 0 | 0.0% | 4 | 36.4% |
| CYS | 192 | 0.5% | 35 | 1.0% | 32 | 0.7% | 51 | 1.6% | 49 | 6.0% | 1 | 0.6% | 0 | 0.0% |
| GLN | 538 | 1.5% | 18 | 0.5% | 82 | 1.7% | 25 | 0.8% | 9 | 1.1% | 5 | 3.1% | 0 | 0.0% |
| HIS | 516 | 1.5% | 13 | 0.4% | 33 | 0.7% | 58 | 1.8% | 1 | 0.1% | 10 | 6.2% | 0 | 0.0% |
| SER | **4531** | **12.9%** | **228** | **6.8%** | **557** | **11.6%** | **733** | **22.3%** | **180** | **22.2%** | **49** | **30.4%** | 0 | 0.0% |
| THR | **5020** | **14.3%** | **211** | **6.3%** | **447** | **9.3%** | **974** | **29.6%** | **317** | **39.0%** | 15 | 9.3% | 0 | 0.0% |
| TRP | 93 | 0.3% | 1 | 0.0% | 12 | 0.3% | 30 | 0.9% | 0 | 0.0% | 0 | 0.0% | 0 | 0.0% |
| TYR | 435 | 1.2% | 61 | 1.8% | 75 | 1.6% | 31 | 0.9% | 14 | 1.7% | 0 | 0.0% | 0 | 0.0% |
| GLY | **10980** | 31.3% | 1948 | 58.0% | 843 | 17.6% | 270 | 8.2% | 14 | 1.7% | **23** | 14.3% | 0 | 0.0% |
| ALL | 35123 | 100.0% | 3357 | 100.0% | 4800 | 100.0% | 3288 | 100.0% | 812 | 100.0% | 161 | 100.0% | 11 | 100.0% |

[a] RCR: random chain residues, $N_{cap}$–$N_{three}$: α-helix N-terminus, anti: antiparallel, and para: parallel β-sheets.

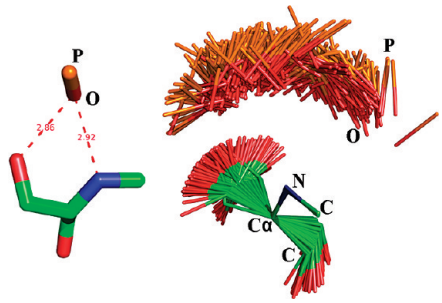



**Figure 18.** Interaction pattern for P═O substructure toward Ser at the α-helix $N_{one}$ position, left: one representative geometry, right: superposition of the 557 entries on $C_{n-1}$, $N_n$, and $C\alpha_n$.

contrast, the CIC chlorid channel presents two facing helices with their NH termini toward a preferred binding site of chloride ions.[30,31] Even though these examples suggest for charge stabilizing effects, there is still no clear-cut evidence provided to correlate the strengths of such macrodipoles with the absolute length of such α-helices. As described in this section, the high occurrence of polar and charged amino acids or ligand functional groups such as carboxylate or phosphate groups at the α-helix N-terminus can also be evidenced by a particular hydrogen bond pattern only provided by the adjacent and parallel oriented N−H backbone amide groups. Possibly this aspect is equally determining as the charge stabilizing argument. Nevertheless, it provides a convenient,

**Table 11.** Distribution of Special Ligands Such As Cofactors Interacting with Backbone Amide Group (Overall and Particularly for Ser and Thr Residues)[a]

| | $N_{Cap}$ | | | $N_{one}$ | | | $N_{Two}$ | | | $N_{Three}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ligand | no. | [%] | ligand | no. | [%] | ligand | no. | [%] | ligand | no. | [%] |
| overall | ADP | 846 | 12.65 | FAD | 946 | 10.05 | FAD | 784 | 11.93 | **ADP** | **356** | **21.93** |
| | GDP | 578 | 8.64 | NAD | 862 | 9.16 | ADP | 756 | 11.5 | **GDP** | **246** | **15.16** |
| | NAD | 388 | 5.8 | ADP | 782 | 8.31 | GDP | 512 | 7.79 | ATP | 184 | 11.34 |
| | PO4 | 386 | 5.77 | PLP | 700 | 7.44 | NAD | 478 | 7.27 | GNP | 106 | 6.53 |
| | PLP | 370 | 5.53 | PO4 | 514 | 5.46 | PLP | 466 | 7.09 | PLP | 92 | 5.67 |
| | FAD | 348 | 5.2 | FAD | 946 | 10.05 | ATP | 340 | 5.17 | FAD | 76 | 4.68 |
| Ser | **FAD** | **120** | **26.32** | FAD | 142 | 12.77 | **GDP** | **270** | **18.44** | **PLP** | **68** | **18.89** |
| | **NDP** | **80** | **17.54** | FMN | 136 | 12.23 | **ADP** | **268** | **18.31** | **FAD** | **56** | **15.56** |
| | NAD | 44 | 9.65 | PLP | 98 | 8.81 | FAD | 148 | 10.11 | ATP | 42 | 11.67 |
| | NAP | 36 | 7.89 | PO4 | 94 | 8.45 | GNP | 110 | 7.51 | GDP | 34 | 9.44 |
| | PO4 | 34 | 7.46 | ADP | 74 | 6.65 | PLP | 74 | 5.05 | GNP | 32 | 8.89 |
| | FAD | 120 | 26.32 | DUT | 50 | 4.5 | ATP | 62 | 4.23 | ADP | 26 | 7.22 |
| Thr | **PLP** | **108** | **25.59** | ADP | 130 | 14.54 | **ADP** | **332** | **17.11** | **ADP** | **174** | **27.49** |
| | FAD | 40 | 9.48 | A3P | 104 | 11.63 | GDP | 210 | 10.82 | **GDP** | **118** | **18.64** |
| | FAB | 32 | 7.58 | FAD | 102 | 11.41 | FAD | 146 | 7.53 | ATP | 88 | 13.9 |
| | NAP | 26 | 6.16 | FMN | 86 | 9.62 | PLP | 96 | 4.95 | GSP | 22 | 3.48 |
| | PAL | 24 | 5.69 | ATP | 78 | 8.72 | FMN | 90 | 4.64 | ANP | 21 | 3.32 |
| | ATP | 16 | 3.79 | PO4 | 44 | 4.92 | PO4 | 84 | 4.33 | POP | 20 | 3.16 |

[a] Three-letter-codes: ATP: adenosine-5′-triphosphate, ADP: adenosine-5′-diphosphate, GDP: guanosine-5′-diphosphate, NAD: nicotinamide-adenine-dinucloetide, NAP: nicotinamide-adenine-dinucleotide phosphate (NADP), NDP: dihydro-nicotinamide-adenine-dinucleotide phosphate (NADPH), PLP: pyridoxal-5′-phosphate; FAD: flavin-adenine dinucleotide, PO4: phosphate, DUT: deoxyudridine-5′-triphosphate, PAL: N-(phosphonacetyl)-L-aspartic acid, A3P: adenosine-3′-5′-diphosphate, ANP: phosphoaminophosphonic acid-adenylate ester, POP: pyrophosphate.

purely geometrical explanation for the high occurrence frequency of carboxylate and phosphate groups at this position. The analysis of the interaction geometry of water molecules bound to the amide backbone group also reveals no significant correlation in terms of bond lengths in consequence of any possibly superimposed cooperative effects. They also do not correlate with the actual length of the adjacent α-helix possibly producing a macrodipole of increasing strength. Nevertheless, this result has to be interpreted with some care due to the limited accuracy achieved for the determination of water position in X-ray diffraction studies of proteins.

Two theoretical studies analyzing the stabilization of charged residues at the α-helix terminal ends involved in capping motifs also provide no evidence of a correlation between helix stability and the actual helix length.[32,33] Both studies conclude that either the first or last turn of a helix provides the major contribution to the charge stabilizing effects and is mainly determined by providing hydrogen bonds. This suggests that the following turn takes only minor influence in terms of dipolar groups, and all additional turns take hardly any effect. Possibly the charge stabilizing effect is mainly provided by the terminal turns, and the overall length of a helix is only required for structural reasons. Only a helix with a particular minimal length of several turns can be oriented with the required accuracy toward a particular binding site locus to allow for reliable recognition of a given ligand on an ion. All pictures showing images of molecules were created using Pymol.[34]

## SUMMARY AND CONCLUSION

Here, a new database module is presented that integrates the information about secondary structure elements into Relibase. The data are accessible via the extended Relibase user interface and an extended version of Reliscript. In addition to the information about $\beta$-sheets and helices provided by the PDB, a uniform classification of all turn families, based on recent clustering methods, and a new helix assignment that is based on this turn classification is added. Furthermore, algorithms to analyze the geometric features of helices and $\beta$-strands were implemented.

A protein can be visualized in combination with the secondary structure elements using the Astex viewer within the Relibase Web interface. This gives the opportunity to analyze the specific secondary structure elements of a certain protein in more detail. Furthermore, the secondary structure element information can be used for constraining substructure searches to analyze, for example, protein−ligand interactions in the context of specific turn-types. Reliscript was also extended which allows to access the data and construct more complex queries with respect to secondary structure elements.

Several analyses are provided that exemplify the use of the Relibase search facilities in combination with the new information about secondary structure elements. It was shown that Reliscript is a perfect tool to calculate the α-helix amino acid propensity. Using a python-based script it is straightforward to regularly recalculate such properties once an increasing amount of structural data becomes available. Furthermore, the Web interface was used to retrieve the information about protein−ligand interactions of the protein backbone amide with respect to the α-helix N-terminus, the first or last strand in an $\beta$-sheet or random chain residues. A further analysis of the data suggests that most structural features observed at α-helix termini can be explained by the specific geometry provided by the first turn and the accessibility of parallel oriented backbone N−H amide groups. In particular, the preferred binding of carboxylate and phosphate groups can be observed. No clear-cut correlation can be substantiated that the magnitude of an overall helix macrodipole takes prominent influence on binding strength. Furthermore, a detailed analysis of endogenous ligands showed Ser, Thr, and Lys as preferred interaction partners for phosphate groups. Although there is no structural preference for Ser and Thr (e.g., at specific position of the α-helix N-terminus) interacting with phosphate groups in general, different endogenous ligands showed different preferences for Ser/Thr as an interaction partner at a specific position of the α-helix N-terminus.

Hopefully, this manuscript demonstrates the power of Secbase as a versatile tool to analyze protein−ligand interactions with respect to involved secondary structure elements. The reported examples underline the importance to integrate retrieval and analysis tools in a database such as Relibase that supports structure-based ligand design. Especially the new turn classification should provide new opportunities to identify similar structural motifs within different proteins and protein families.

Secbase, as described in this contribution, is being made available within the new release of Relibase+ and the free-Web version of Relibase.

## ACKNOWLEDGMENT

The authors thank Dr. Robin Taylor and Dr. Greg Shields (CCDC, Cambridge) and The Cambridge Crystallographic Data Centre for financial support of O.K. We would like to acknowledge the help of Dr. Andreas Heine (Univ. of Marburg) for critical reading of this contribution.

## REFERENCES AND NOTES

(1) Koonin, E. V.; Wolf, Y. I.; Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **2002**, *420*, 218–223.
(2) Breinbauer, R.; Vetter, I. R.; Waldmann, H. From protein domains to drug candidates-natural products as guiding principles in the design and synthesis of compound libraries. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 2879–2890.
(3) Grishin, N. V. Fold change in evolution of protein structures. *J. Struct. Biol.* **2001**, *134*, 167–185.
(4) Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607–620.
(5) Branden, C., Tooze, J. *Introduction to Protein Structure*, 2nd ed.; Garland: New York, 1999.
(6) Pauling, L.; Corey, R. B.; Branson, H. R. The structure of proteins, two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 205–211.
(7) Pauling, L.; Corey, R. B. The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 251–256.
(8) Eisenberg, D. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11207–11210.
(9) Koch, O.; Klebe, G. Turns revisited: A uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. *Proteins* **2008**, *74*, 353–367.
(10) Koch, O.; Bocola, M.; Klebe, G. Cooperative effects in hydrogen-bonding of protein secondary structure elements: a systematic analysis of crystal data using Secbase. *Proteins* **2005**, *61*, 310–317.
(11) Miranda, J. J. Position-dependent interactions between cysteine residues and the helix dipole. *Protein Sci.* **2003**, *12*, 73–81.

(12) Forsyth, W. R.; Antosiewicz, J. M.; Robertson, A. D. Empirical relationships between protein structure and carboxyl pKa values in proteins. *Proteins* **2002**, *48*, 388–403.
(13) Milner-White, E. J. Situations of gamma-turns in proteins. Their relation to alpha-helices, beta-sheets and ligand binding sites. *J. Mol. Biol.* **1990**, *216*, 386–397.
(14) Kee, K. S.; Jois, S. D. Design of $\beta$-turn Based Therapeutic Agents. *Curr. Pharm. Des.* **2003**, *9*, 1209–1224.
(15) Brakch, N.; El Abida, B.; Rholam, M. Functional role of $\beta$-Turn in Polypeptide Structure and its use as Template to Design Therapeutic Agents. *Cent. Nerv. Syst. Agents Med. Chem.* **2006**, *6*, 163–173.
(16) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
(17) Chou, P. Y.; Fasman, G. D. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **1974**, *13*, 211–222.
(18) Chou, P. Y.; Fasman, G. D. Beta-turns in proteins. *J. Mol. Biol.* **1977**, *115*, 135–175.
(19) Wilmot, C. M.; Thornton, J. M. Analysis and prediction of the different types of beta-turn in proteins. *J. Mol. Biol.* **1988**, *203*, 221–232.
(20) Hobohm, U.; Sander, C. Enlarged representative set of protein structures. *Protein Sci.* **1994**, *3*, 522–524.
(21) Penel, S.; Hughes, E.; Doig, A. J. Side-chain structures in the first turn of the alpha-helix. *J. Mol. Biol.* **1999**, *287*, 127–143.
(22) Aurora, R.; Rose, G. D. Helix capping. *Protein Sci.* **1998**, *7*, 21–38.
(23) Cochran, D. A.; Penel, S.; Doig, A. J. Effect of the N1 residue on the stability of the alpha-helix for all 20 amino acids. *Protein Sci.* **2001**, *10*, 463–470.
(24) Doig, A. J.; Baldwin, R. L. N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Sci.* **1995**, *4*, 1325–1336.
(25) Schellmann, C. , The αL conformation at the ends of helices. In *Protein Folding*; Jaenicke, R., Ed.; Elsevier: Amsterdam, The Netherlands, 1980; pp 53−61.
(26) Baker, E. N.; Hubbard, R. E. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **1984**, *44*, 97–179.
(27) McDonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238*, 777–793.
(28) Hirsch, A. K.; Fischer, F. R.; Diederich, F. Phosphate recognition in structural biology. *Angew. Chem., Int. Ed. Engl.* **2007**, *46*, 338–352.
(29) Saraste, M.; Sibbald, P. R.; Wittinghofer, A. The P-loop - a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **1990**, *15*, 430–434.
(30) Doyle, D. A.; Morais Cabral, J.; Pfuetzner, R. A.; Kuo, A.; Gulbis, J. M.; Cohen, S. L.; Chait, B. T.; MacKinnon, R. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science* **1998**, *280*, 69–77.
(31) Dutzler, R. Structural basis for ion conduction and gating in ClC chloride channels. *FEBS Lett.* **2004**, *564*, 229–233.
(32) Tidor, B. Helix-Capping Interaction in $\lambda$ Cro Protein: A Free Energy Simulation Analysis. *Proteins* **1994**, *19*, 310–323.
(33) Aqvist, J.; Luecke, H.; Quiocho, F. A.; Warshel, A. Dipoles localized at helix termini of proteins stabilize charges. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 2026–2030.
(34) DeLano, W. L. *The PyMOL Molecular Graphics System*; DeLano Scientific: San Carlos, CA, U.S.A., 2002.

CI900202D