# Detecting Relationships between Amino Acid Residue Sequences and 3D Protein Structures based on a New Class of Rotamer Libraries

**Alexander Hinneburg**[a], **Daniel A. Keim**[a], **Wolfgang Brandt**[b,*]
(a) Instiute for Computer Science, University of Halle, D-06099 Halle, Germany
{hinneburg, keim}@informatik.uni-halle.de,
(b) Institute for Biochemistry, University of Halle,, D-06099 Halle, Germany
brandt@biochemtech.uni-halle.de
(*) to whom correspondence should be addressed

### Abstract

In the past a good number of rotamer libraries have been published, which allow a deeper understanding of the conformational behaviour of amino acid residues in proteins. Since the number of available high resolution X-ray protein structures has grown significantly over the last years, a more comprehensive analysis of the conformational behaviour is possible today. In this paper, we present a method to compile a new class of rotamer libraries for detecting interesting relationships between residue conformations and their sequential context in proteins. The method is based on a new algorithm for clustering residue conformations. To demonstrate the effectivity of our method we apply our algorithm to a library consisting of all 8000 tripetid fragments formed by the 20 native amino acids. The analysis shows some very interesting new results, namely that some specific tripeptid fragments show some unexpected conformation of residues instead of the highly preferred conformation. In the neighborhood of two asparagin residues, for example, threonin avoids the conformation which is most likely to occur otherwise. The new insights obtained by the analysis are important in understanding the formation and prevention of secondary structure elements and will consequently be crucial for improving the state-of-the-art of protein folding.

**Keywords:** rotamer library, cluster algorithm, conformational analysis, tripeptids, protein folding
**Running Head:** New Class of Rotamer Libraries

## 1 Introduction

In the recent, years the number of proteins stored in the PDB (Protein Data Bank) has grown significantly. Because of the technical progress a good number of high-resolution X-ray structures of proteins became available. Statistical methods have been applied to the PDB to extract knowledge about the conformational behavior of amino acid residues. Amino acid side chain conformations have been studied, for example, by *Chandrasekaran, Ramachandran 1970*; *Cody et al. 1973*; *James, Sielecki 1983*; *Ponder, Richards 1987*. These studies resulted in side chain rotamer libraries, which consist of a list of diskrete conformations having a weight which corresponds to their frequencey in the PDB. Since the PDB contains a multitude of high-resolution structures, it was also possible to determine rotamer preferences depending on the backbone conformation. Based on this idea, a number of weak correlations of rotamer distributions and secondary structures have been found [*Janin et al. 1978*]; [*McGregor et al. 1987*]; [*Sutcliffe et al. 1987*]; [*Schrauber et al. 1993*]. Recently, a backbone dependent side chain rotamer library has been presented by *Dunbrack, Karplus 1994*; *Dunbrack, Cohen 1997*. The effectivity of the backbone dependent rotamer libraries has been shown by *Dunbrack, Karplus 1993*; *Bower et al. 1997* for homlogy modelling and by *Kuszewski et al. 1996* for NMR and X-ray structure refinement.

Although the idea of using rotamer libraries has already been applied successfully in the past, until now only a small fraction of its potential has been revealed. The backbone dependent side chain rotamer library mentioned above, for example, only uses 132 out of the about 2000 proteins with a resolution $\leq 2A$, which are available in the PDB. To be able to better understand tertiary structures, it is highly desirable to compile comprehensive rotamer libraries which are based on all protein structures available in the PDB. Using such a rotamer library, for instance, one would be able to determine how the conformation of an amino acid residue (in particular that of a side chain) in a protein depends on its neighbours in the sequence. To find and understand such relationships, a new method is required which is able to deal with large amounts of residue conformations and to classify them effectively. Our new method presented in this paper is based on a cluster analysis in the conformation space (cf. section 2). The basic idea is to model the conformation of amino acid residues or small peptide fragments as points in the multidimensional dihedral angle space. The cluster algorithms then determines clusters by assigning an influence function to each data point, by summing up all influence functions to determine the overall density function, and by finding the maxima of the overall density function using a gradient-based hill-climbing procedure (cf. section 3). The method is used to compile a new class of rotamer libraries which allows new insights into interesting dependencies between the 3D-structure of small peptide chain fragments and their sequential context. In section 5, we evaluate the effectivity and efficiency of our new approach and provide some interesting results showing, for example, that in the neighborhood of two asparagin residues, threonin avoids the conformation which is highly preferred otherwise. Note that our new method is generally applicable to arbitrary protein fragments. In this paper, however, we restrict ourselves to the evaluation and analysis of tripeptid conformations.

# 2 General Idea

In the backbone dependent rotamer library developed by *Dunbrack, Cohen 1997* for each residue type a probabiltiy distribution of the side chain angle $\chi_1$ is calulated for each node on an equidistant grid in the 2D $(\phi, \psi)$-space. The distributions of $\chi_2, \chi_3$ and $\chi_4$ only depend on the previous side chain dihedral angle. For detecting more global relationships this method becomes inefficent since the size of the grid grows exponentially in the number of considered angles. Another problem arises if one is interested in the probability distribution of more than one angle. Using Bayesian statistics, it is difficult to derive combined distributions of two or more angles.

The conformation of amino acid residues or small peptide fragments can also be described by data points in a multidimensional dihedral angle space. The approach we are using discretisizes the multidimensional angle space corresponding to the observed data distribution by clustering the data in the dihedral angle space. More formally, this can be described as follows.

Given is a set of protein sequences $P$. A sequence $p \in P$ is denoted as a string of linked amino acid residues $a$ from the set of natural amino acids $A$:

$$p \in P,\ p = a_1 a_2 \ldots, a_l,\ a_i \in A, i = 1, \ldots, l.$$

In our approach, we do not directly use the real tertiary structure since the mapping of the 3D coordinates of the structure to dihedral angles basically contains all relevant information about the protein structure and is much easier to handle. For clear notations, we introduce the projection of the 3D atom coordinates of a protein $p \in P$ to a sequence of vectors of dihedral angles as:

$$p \in P,\ D(p) = s_1, s_2, \ldots s_l,\ s_i \in [-180, 180)^{d_i}, i = 1, \ldots, l$$

with $d_i$ being the number of dihedral angles for the residue $a_i$. For example, $d_i$ is 3 for $a_i = G$ (glycin) and $d_i = 7$ for $a_i = A$ (argenin). The components of a vector $s_i$ are

$$s_i = \begin{cases} (\phi, \psi, \omega) & ; d_i = 3 \\ (\phi, \psi, \omega, \chi_1, \ldots, \chi_{d_i - 3}) & ; 3 \leq d_i \leq 7 \end{cases}.$$

We use the dihedral angles of one residue as the smallest unit for detecting relationships. Note that one can easily modify the structure of the data by, for example, grouping the dihedral angles.

To produce the rotamer library for detecting relation ships between the 3D-stucture of a residue and in its sequential context two steps have to be performed.

**Step 1** For all different residues do:
Determine all conformations of the residue in the protein structures of $P$ and discretisize the conformational angle space according to the observed data distribution by using a cluster algorithm which identifies highly populated areas in the multidimensional dihedral angle space.

**Step 2** For all residues in the protein sequences, replace the dihedral angle vector with the cluster-id of that vector and based on the resulting data, build frequency tables for all different residue fragments of fixed length by counting the occurences of all fragments which correspond to the same sequence of cluster-ids.

In the resulting frequency tables, we obtain significant information about dependencies between the residues in a given sequential neighbourhood and the preference for a certain conformational structure. Note that most of these correlations can only be detected if the sequential context is considered. In the following, we discuss the most important step of our approach, namely building the frequency tables by clustering the dihedral angle space, in more detail.

# 3 Cluster Analysis in the Dihedral Angle Space

In the clustering step, the densely populated areas in the conformational space of each of the 20 natural amino acid residues have to be identified. This can be done separately for each of the residues. As the first step of the clustering algorithm, the source data for the cluster analysis of a residue $a \in A$ is determined by collecting all conformations of $a$, which occur in the protein conformations in $P$. In addition to the residue conformation, the protein name and the position of the residue in the protein chain are stored. With this additional information it is possible to retrieve the sequential context of a residue conformation after the cluster analysis. The considered set of proteins $P$ contains all proteins from the PDB, which have resolutions of the X-Ray structure of $\leq 2\overset{\circ}{A}$. With this condition, $P$ contains about 2000 proteins. From this set of proteins, we get a conformational data set for each residue with a size between 48.000 (for alanin) and 9000 (for tyrosin). In order to get a good number of classified tripeptid conformations for each of the 8000 possible tripeptid fragments, we used the complete data sets in the cluster analysis.

The task of the cluster algorithm is to group the objects from the given data sets into smaller, homogeneous, and meaningful subsets which are the clusters. In our case, the objects are conformations of one residue type described by a vector of dihedral angles. To formally define the term *homogeneous*, we need an appropriate distance measure on the objects. In case of dihedral angle vectors, it is rather straight-forward to extend the Euclidian distance to measure the shortest path of transformation between two residue conformations. This can be defined as

$$x, y \in [-180, 180)^d,\ dist(x, y) = \sqrt{\sum_{i=1}^{d} \begin{cases} |x_i - y_i|^2, & |x_i - y_i| \leq 180 \\ (360 - |x_i - y_i|)^2, & \text{else} \end{cases}}$$

The impact of this distance measure is the wrap-around at the borders. The effect is shown in Figure 1 where the shaded area diplays a two-dimensional sphere around the the point $(-180, -180)$.

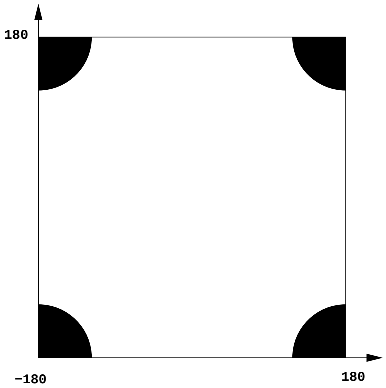


**Figure 1:** Effect of the Warp-Around

After defining the distance measure, we next have to define an adequate notion of clusters. Since the definition of clusters largely depends on the data and the application, we first tried to get a visual impression of the structure of clusters in our application. For this purpose, we used the Ramachandran-Plot of the actual conformation set, which is a projection to the $(\phi, \psi)$-plane. Figure 2 shows the $(\phi, \psi)$-plot of glycine conformations as an example. Note that the plot is only a projection of the high-dimensional dihedral angle space to the two-dimensional display space, which results in loosing some of the information. Nevertheless, the plot reveals some important properties of the clustering in our data set.

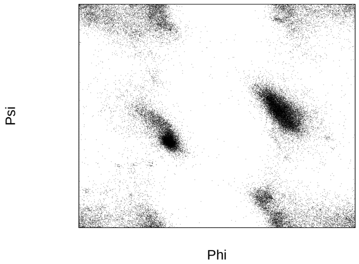


**Figure 2:** Ramachandran-Plot for Glycine

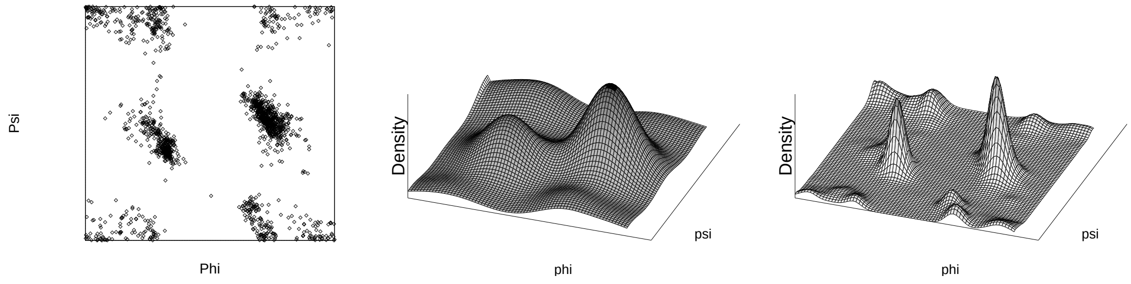
The Figure shows densely populated areas which are separated by nearly empty space. It is well known from biochemistry that preferred areas exist in the conformation space, and the clusters in the plot correspond to preferred secondary structure elements in which the residues are involved. Two further observations can be derived from the plots: First, the shape of the clusters is not fixed to certain shapes (e.g., spherical shapes) and second, the space between the clusters is filled with a significant number of outliers. Outliers are points which do not belong to any cluster. The observations lead to restrictive requirements for the cluster algorithm: The algorithm should be able to find clusters of arbitrary shape, handle a variable amount of outliers, and efficiently deal with a large multidimensional data set (up to 50.000 points). From the wide range of cluster algorithms which have been proposed in the literature ([*Duda, Hart 1973*], [*Ester et al. 1996*], [*Zhang et al. 1996*]), only few algorithms fulfill these requirements and none of them works efficiently on large amounts of multidimensional data. A new approach which has been recently proposed by the authors in the context of knowledge discovery in multimedia databases [*Hinneburg, Keim 1998*] can be adapted to meet the requirements.

In the following, we briefly introduce the algorithm and describe the adaptation to the problem of clustering the conformation space of amino acid recidues. The basic idea of the DENCLUE (DENsity based CLUstEring) algorithm is to determine clusters based on the overall density of the data in the data space. The overall density function is defined as the sum of influences, which the data points are assumed to have in the data space. The influence of a data point is usually modeled by a gaussian influence function. Definition 1 introduces a density function with gaussian influence functions (cf. [*Schnell 1964*] or [*Fukunaga, Hostler 1975*] for a similar notion of density functions).

**Def. 1: (Density Function)**
The density function is defined as the sum of the gaussian influence functions of all data points. Given $N$ conformations decribed by a set of dihedral angle vectors $D = \{x_1, \ldots, x_N\} \subset [-180, 180)^d$ the density function is defined as

$$f^D(x) = \sum_{i=1}^{N} e^{-\frac{dist(x,x_i)^2}{2\sigma^2}} \ .$$

The paramter $\sigma$ controls how far the influence of a data point is propagated into the data space. Figure 3 shows an example of a Ramachandran plot and the overall density induced from it for different values for $\sigma$.



(a) Data Set (b) $\sigma = 40$ (c) $\sigma = 10$

**Figure 3:** Example for Density Functions

For our definitin of clusters, the notion of a density attractor is needed. Informaly a density attractor is a local maximum of the density function. To determine a density attractor, we need the gradient which is given for a density function according to Definition 1 by the following equation:

$$\nabla f^D(x) = \sum_{i=1}^{N} (x_i - x) \cdot e^{-\frac{dist(x,x_i)^2}{2\sigma^2}} \ .$$

**Def. 2: (Density Attractor)**
A point $x^* \in [-180, 180)^d$ is called a *density attractor*, if $x^*$ is a local maximum of the density function $f^D$.
A point $x \in [-180, 180)^d$ is *density attracted* to a density attractor $x^*$, if $\exists k \in N : dist(x^k, x^*) \leq \epsilon$ with

$$x^0 = x, \ x^i = x^{i-1} + \delta \cdot \frac{\nabla f^D(x^{i-1})}{\|\nabla f^D(x^{i-1})\|} \ .$$

The second part of definition 2 implies that a point $x$ is density attracted to $x^*$, if the hill climbing procedure which is started at $x$ and which is guided by the gradient leads into a $\epsilon$-sphere around the density attractor $x^*$. Figure 4 shows an example of density attractors in a one-dimensional space.

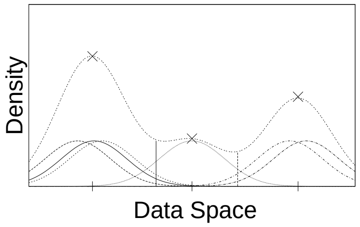


**Figure 4:** Example of Density Attractors

In real data sets, not all density attractors are significant in terms of a cluster because there may be a significant number of outliers. Outliers are points which are not influenced by "many" other data points. We can use an additional parameter $\xi$ to formalize the "many".

**Def. 3: (Cluster, Outlier)**
A *cluster* (wrt $\sigma, \xi$) for a density attractor $x^*$ is a subset $C \subseteq D$, with $x \in C$ being density attracted by $x^*$ and $f^D(x^*) \geq \xi$. Points $x \in D$ are called *outliers* if they are density attracted by a local maximum $x^*_O$ with $f^D(x^*_O) < \xi$.

The DENCLUE algorithm has two important parameters, namely $\sigma$ and $\xi$. The parameter $\sigma$ describes the influence of one data point on its neighbourhood. There are two extremas $\sigma_{max}$ and $\sigma_{min}$. If $\sigma \geq \sigma_{max}$ the influence is propagated so far that the density function $f^D$ has only one density attractor. The other extrem is $\sigma \leq \sigma_{min}$ in which case the Gaussian functions become little peaks and each data point becomes a density attractor of its own. Choosing a good $\sigma$ can be done by considering different $\sigma$ and determining the largest intervall between $\sigma_{max}$ and $\sigma_{min}$, where the number of density attractors is almost constant. The clustering resulting from that approach can be seen as naturally adapted to the data set. In Figure 5 we provide an example for the number of density attractors $m(\sigma)$ depending on $\sigma$.

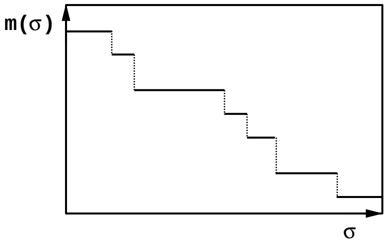


**Figure 5:** Number of Density Attractors Depending on $\sigma$

The second parameter $\xi$ describes the minimum density level above which a density attractor is considered significant. A good choice for $\xi$ helps the algorithm to focus on the densely populated areas and to save computational time. Note that the border of a cluster may be in regions with a density lower than $\xi$. Important is that the density attractor $x^*$ has $f^D(x^*) \geq \xi$. The details of the theoretical foundations and implementation of the DENCLUE algorithm are beyond the scope of this paper and are described in [*Hinneburg, Keim 1998*].

# 4 Building Fragmant Rotamer Libraries

With the results derived by the cluster analysis, we are now able to build the desired frequency table. The result from the cluster analysis is that each residue conformation in the protein structures $p \in P$ has become part in exactly one cluster (the outliers are grouped into a special cluster). Formally, we introduce a mapping function $C$ from the residue conformations into the set of cluster identificators $CI$. This function provides the identificator of the cluster for a given residue conformation.

$$p \in P,\ D(p) = s_1, \dots, s_{l_p}\ \forall j \in \{1, \dots, l_p\} : C(s_j) = c_j, c_j \in CI$$

For a more convenient notation, we write

$$p \in P,\ C(D(p)) := C(s_1), \dots, C(s_{l_p}) = c_1, \dots, c_{l_p}$$

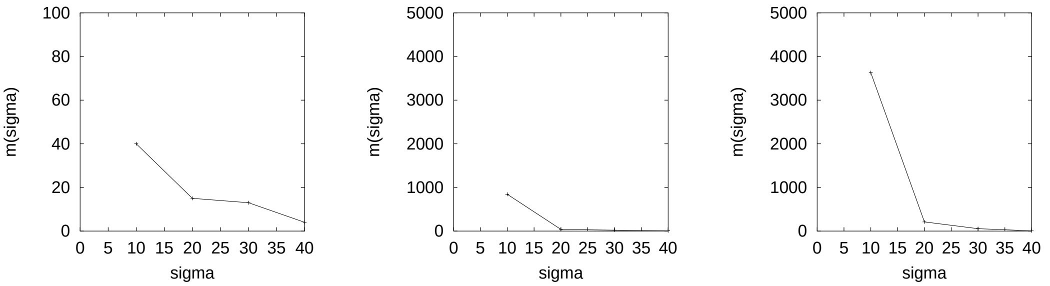
The fragment rotamer libraries can be complied based on the mapping C for different fragment sizes without recomputing the cluster analysis in step 1. Let be $t > 1$ the fragment size for the desired library. The fragment rotamer library $L$ can be considered as a relation, where each of the $20^t$ possible residue fragments of length $t$ correspond to a finite set of $t$-tupels of cluster-ids. The set $A^t$ is the set of the fragments with length $t$ based on the set $A$ of the 20 natural amino acid residues.

$$x \in A^t, L(x) = \{[(c_1^1, \dots, c_t^1), h^1], \dots, [(c_1^{l_x}, \dots, c_t^{l_x}), h^{l_x}]\}$$

Each tupel $i$ matches to $h^i \geq 1$ occurences of the corresponding fragment and its mapping given by the $t$-tupel in the protein structures $p \in P$. The frequency $h$ of each tupel is provided in $L$ as a statistical information about the preference of the corresponding fragment for the conformation space corresponding to the t-tupel of cluster-ids. The significance of such perferences depends strongly on the total number $\sum_{i=1}^{l_x} h^i$ of occurences of the examined fragment $x \in A^t$ in the protein structures $p \in P$.

# 5 Results

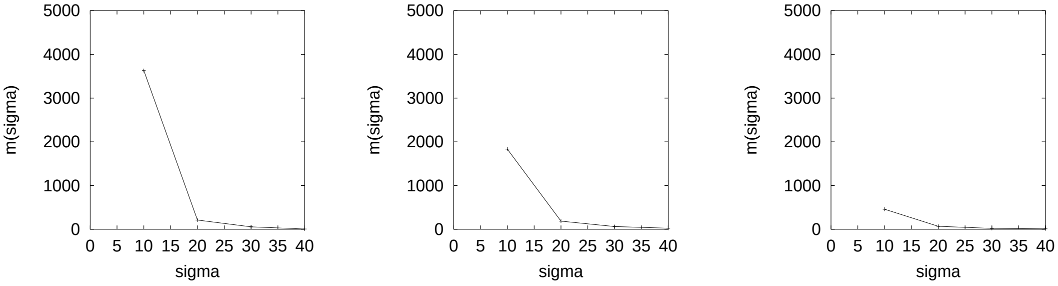
In this section, we focus on the application of the cluster algorithm to conformational data of amino acid residues and provide as an exapmle a rotamer library for tripeptids. The algorithm was applied to each of the 20 conformation data sets for different $\sigma$ ($\sigma_1 = 10$, $\sigma_2 = 20$, $\sigma_3 = 30$ and $\sigma_4 = 40$). To get a first impression of the results, we ploted the number of density attractors (not clusters) depending on $\sigma$. Figure 6 shows three typical cases which occured in our analysis: (a) there exits an interval where $m(\sigma)$ is almost constant (b) $m(\sigma)$ slowly decreases (c) $m(\sigma)$ rapidly decreases (note the scaling).



(a) cystein (b) asparagin acid (c) argenin

**Figure 6:** Number of Density Attractors Depending on $\sigma$

Residues with a behaviour such as the one presented in case (c) of Figure 6 are mostly hydrophil and have long side chains decribed by $\chi_1, \dots, \chi_4$. Hydrophil residues often occur at the surface of a protein and the side chain reaches into the water where no stable conformation is adapted. As a consequence, the data points are uniformly distributed in these dimensions and no preference can be detected. It seems to be more realistic to neglect $\chi_3$ and $\chi_4$ and postulate them as freely rotatable. Figure 7 shows the effect of neglecting $\chi_3$ and $\chi_4$ on the number of density attractors. From Figure 7 it is clear that the assumption that $\chi_3$ and$\chi_4$ can rotate freely leads to more realistic clusterings - a result which is also supported by [*Dunbrack, Cohen 1997*].



(a) argenin (b) argenin without $\chi_4$ (c) argenin without $\chi_4$, $\chi_3$

**Figure 7:** Number of Density Attractors Depending on $\sigma$

With these results we can build the mapping function $C$ from the residue conformations to the cluster-ids we formaly inroduced in section 4. A cluster identificator consists of the residue name and the cluster number. The cluster are numbered in the order of decreasing size. In Table 1, we provide an example of a clustering and in Table 2 an example of the mapping for the clusterings with $\sigma = 40$.

**Table 1:** Cluster for Threonin (not complete), $\sigma = 40$

- cluster (T,1); size: 8904/35212, 25.29%
  center: $\phi = -107.1$ $\psi = 129.2$ $\omega = 178.7$ $\chi_1 = -58.1$
- cluster (T,2); size: 8303/35212, 23.58%
  center: $\phi = -113.4$ $\psi = 160.6$ $\omega = 178.4$ $\chi_1 = 61.3$
- cluster (T,3) size: 8293/35212, 23.55%
  center: $\phi = -88.3$ $\psi = -14.5$ $\omega = -179.6$ $\chi_1 = 59.6$

**Table 2:** Example of the Mapping Function $C$

| AC | $\phi$ | $\psi$ | $\omega$ | $\chi_1$ | $\chi_2$ | | AC | Cluster Id |
|---|---|---|---|---|---|---|---|---|
| S | -82.54 | -23.18 | 179.45 | -36.58 | | | S | (S,3) |
| C | -101.64 | 22.62 | 176.75 | -46.90 | | | C | (C,1) |
| T | -83.39 | 159.20 | 179.77 | 53.62 | | | T | (T,2) |
| H | 63.44 | 17.12 | -178.77 | -48.42 | -72.51 | $\xRightarrow{C}$ | H | (H, out) |
| F | -112.55 | 140.25 | 0.66 | 174.56 | -82.41 | | F | (F, out) |
| P | -90.11 | 8.31 | -177.86 | 37.66 | -23.44 | | P | (P,2) |
| G | -54.72 | -24.68 | -179.60 | | | | G | (G,3) |
| N | -89.86 | 2.58 | -178.75 | 89.54 | 131.35 | | N | (N, out) |
| L | -61.54 | -43.83 | -179.54 | -76.65 | -172.62 | | L | (L,1) |

We stored these mappings and scaned them through to count the frequency of different mappings for each tripeptid fragment occuring in the protein sequences. Due to the large number of possible tripeptid fragments (8000), the set of proteins $P$ also provides a very large number of occuring tripeptid fragments. In our example, $P$ consists of about 2000 proteins and the number of occuring tripeptid fragments has been about 240000. In the derived frequency table, for each tripeptid fragment a list of the different tripels of cluster-ids and their freqencies are stored. Table 3 shows a portion of the frequency table for $\sigma = 40$.

**Table 3:** Part of the Frequency Table for $\sigma = 40$

| | | | |
|---|---|---|---|
| ... | | | |
| (N,8) | (T,3) | (P,1) | 2 |
| ASN | THR | ASN | |
| (N,2) | (T,3) | (N,9) | 109 |
| (N,4) | (T,2) | (N,5) | 13 |
| (N,5) | (T,5) | (N,3) | 12 |
| (N,1) | (T,4) | (N,1) | 11 |
| (N,3) | (T,4) | (N,1) | 7 |
| (N,7) | (T,3) | (N,4) | 6 |
| (N,3) | (T,4) | (N,3) | 5 |
| (N,5) | (T,5) | (N,5) | 4 |
| (N,4) | (T,5) | (N,5) | 3 |
| (N,1) | (T,3) | (N,1) | 3 |
| (N,2) | (T,3) | (N,9) | 3 |
| (N,1) | (T,3) | (N,1) | 2 |
| (N,4) | (T,3) | (N,7) | 2 |
| (N,1) | (T,4) | (N,3) | 2 |
| (N,1) | (T,4) | (N,3) | 1 |
| (N,3) | (T,3) | (N,1) | 1 |
| **(N,5)** | **(T,1)** | **(N,2)** | **1** |
| (N,out) | (T,2) | (N,1) | 1 |
| (N,6) | (T,3) | (N,1) | 1 |
| (N,2) | (T,3) | (N,1) | 1 |
| (N,5) | (T,3) | (N,1) | 1 |
| (N,1) | (T,4) | (N,2) | 1 |
| (N,2) | (T,3) | (N,2) | 1 |
| (N,3) | (T,4) | (N,3) | 1 |
| (N,2) | (T,3) | (N,7) | 1 |
| (N,4) | (T,5) | (N,8) | 1 |
| (N,2) | (T,3) | (N,2) | 1 |
| (N,1) | (T,3) | (N,3) | 1 |
| (N,3) | (T,3) | (N,3) | 1 |
| (N,5) | (T,2) | (N,3) | 1 |
| ASN | THR | GLU | |
| (N,4) | (T,2) | (E,4) | 6 |
| ... | | | |

Table 3 contains a full list of cluster-id tripels for the tripeptid fragment $N\ T\ N$. Two observations can be derived from the table. First, there is a significant preference for the first tripel, and second, cluster 1 for threonin (T,1) occurs only once in the list. Cluster 1 is the cluster with the largest size in the clustering for the middle amino acid residue threonin and can be considered a good preference for threonin in the conformation space. The list shows, that in the neighborhood of two asparagin residues threonin avoids the usually perferred conformation space of cluster 1.

This kind of frecquency tables, which can be considered a new class of rotamer liberaries, provides an easy way of detecting of a-priori unknown relationships. The next steps in our further research will be the development of an adequate visualisation of such libraries, allowing scientists a fast exploration of the libraries and enabling them to find rules which are hidden in the data set. Further investigations are intended to explore possible applications to the protein folding problem.

# 6 Conclusions

We showed that our new method for cluster analyses of amino acid residues occurring in X-ray structures of proteins taking into account backbone as well as side chain dihedral angles is appropriate for classification of preferred conformations of amino acids. Based on our new method, a new type of rotamer libraries for tripeptide fragments has been developed. This library has been shown to be useful in finding unknown dependencies between amino acid residue sequences and the favored or disfavored formation of 3D-structures in small peptide fragments. Expected results of a careful analysis of all of the 8000 tripeptides conformations will allow contributions to a better understanding of protein folding phenomenons which are not yet completely understood. The described method allows a general application in the investigation of conformational preferences of peptide fragments in proteins since, in principle, it can be extended to work also on larger fragments.

# References

**Bower,M., Cohen,F.E. and Dunbrack,R.L.,Jr. (1997)** Homology modeling with a backbone-dependent rotamer library, *J. Mol. Biol.*, **267**, 1268-1282.

**Chandrasekaran,R and Ramachandran,G.H (1970)** Studies on the conformation of amino acids. XI. Analysis of the observed side group conformations in proteins. *Int. J. Pept. Prot. Res.*, **2**, 223-233.

**Cody,V., Duax,W.L. and Hauptman,H. (1973)** Conformational analysis of aromatic amino acids by X-ray crystallography. *Int. J. Pept. Prot. Res.*, **5**, 297-308.

**Duda,R.O. and Hart,P.E., (1973)** Pattern Classification and Scene Analysis. *Wiley and Sons.*

**Dunbrack,R.L. and Karplus,M. (1993)** Backbone-dependent rotmer library for proteins: Application to side-chain prediction. *J. Mol. Biol.*, **230**, 543-571.

**Dunbrack,R.L. and Karplus,M. (1994)** Conformational analysis of the backbone-dependent rotamer preferences of protein side chains. *Nature Struct. Biol.*, **1**, 334-340.

**Dunbrack,R.L. and Cohen,F.,E. (1997)** Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, **6**, 1661-1681.

**Ester,M., Kriegel,H.P., Sander,J. and Xu,X. (1996)** A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining, Portland, Oregon, *AAAI Press.*

**Fukunaga,K. and Hostler,L.D. (1975)** The estimation of the gradient of a density function, with application in pattern recognation. *IEEE Trans. Info. Thy.*, **IT-21**, 32-40.

**Hinneburg,A. and Keim,D.A. (1998)** An Efficient Approach to Clustering in Large Multimedia Databases with Noise. To appear in the Proc. 4rd Int. Conf. on Knowledge Discovery and Data Mining, *AAAI Press*, 1998.

**James,M.N.G. and Sielecki,A.R. (1983)** Structure and refinement of penicillopepsin at 1.8 A resolution. *J. Mol. Biol.*, **125**, 299-361.

**Janin,J., Wodak,S., Levitt,M. and Maigret,B. (1978)** Conformations of amino acid side chains in proteins. *J. Mol. Biol.* **125**, 357-386.

**Kuszewski,J., Gronenborn,A.M. and Clore,G.M. (1996)** Improving the quality of NMR and crystallographic protein structures by means of conformational database potential derived from structure databases. *Protein Sci.*, **5**, 1067-1080

**MCGregor,M.J., Islam,S.A. and Sternberg,M.J.E (1987)** Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.*, **198**, 295-310.

**Ponder,J.W. and Richards,F.M. (1987)** Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, **93**, 775-792.

**Schnell,P. (1964)** A method to find point-groups. *Biometrika*, **6**, 47-48.

**Zhang,T., Ramakrishnan,R. and Linvy,M.** BIRCH: An Efficient Data Clustering Method for very Large Databases. Proc. ACM SIGMOD Int. Conf. on Management of Data, *ACM Press*, pp. 103-114.