A Knowledge-Based Approach to Predict Protein Torsion Angles

by

Güzin Tunca

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science
in
Computational Sciences and Engineering

Koc University
January 2007

Koc University
Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master’s thesis by

Güzin Tunca

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

____
Burak Erman, Ph. D. (Advisor)

____
Attila Gursoy, Ph. D.

____
Ozlem Keskin, Ph. D.

Date: ______________________________

ABSTRACT

The three dimensional structure of a protein can be identified in terms of its Φ-Ψ torsion angles. These torsion angles can be considered as the degrees of freedom of a protein. In this study, a method grouping these torsion angles in different rotational isomeric states and estimating their probabilities is developed. Specifically, the probabilities of the various torsion angle states in Ramachandran maps is proposed and the accuracy of the method is examined using a knowledge based approach. Statistical independence and dependence of the states of different residues along the peptide chain are analyzed. The Flory isolated pair hypothesis, near neighbor correlations, context effects and long-range correlations are discussed. In the knowledge based approach, two different protein libraries i) coil library ii) full library are constructed and information from both these libraries is used. Results showed that amino acids have propensities for some rotational isomeric states that favor the choice of the native state torsion angles and they are context dependent, preferring different torsion states determined by the amino acid sequence of the protein. Context dependency is also related to chameleon sequences and the effect of chameleon sequences is also integrated into the method.

ÖZET

Bir proteinin üç boyutlu yapısı Φ-Ψ dönme açıları (dihedral) cinsinden tanımlanabilir. Bu dönme açıları proteinin serbestlik derecesini oluturur. Bu çalışmada dönme açılarını değişik dönme izomerleri olarak guruplayan ve bu izomerlerin olasılıklarını değerlendiren bir yöntem geliştirildi. Özellikle Ramachandran haritasındaki çeşitli dönme açı değerlerinin olasılıkları kullanıldı ve yöntemin doğruluğu bilgi tabanlı bir yaklaşımla sorgulandı. Bir peptid zinciri üzerindeki amino asitlerin dönme açı değerlerinin birbirlerine bağımlılığı incelendi. Flory izole çiftler hipotezi, yakın komşu ilintisi, çevresel etkiler ve uzun mesafe etkileşimleri tartışıldı. Bilgi tabanlı yöntemde iki değişik protein veritabanı kullanıldı: i) proteinlerin düzensiz yapı gösterdiği bölgelerden alınmış veritabanı ii) tüm yapıdan elde edilen veritabanı. Sonuçlarda amino asitlerin protein doğal halinin seçimini destekleyen bazı dönme izomerleri durumlarına daha yatkın olduğunu gösterdi. Dönme açıları değerlerinin, amino asitin içinde bulunduğu ortama bağlı olduğu ve dönme açılarının amino asit dizini tarafından belirlendiği gösterildi. Ortam bağımlılığı aynı zamanda kamelyon dizinleriyle de ilişkilendirilip hesaplamalara katıldı.

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Burak Erman, who has been a great source of inspiration and patience and never let me down in the hardest times of my study.

I should not forget my friends here, thanks for the support and morale you provided in my days of need. I would like to specially thank to Pinar Karabulut and Gunes Gundem for being the coolest and best friends one can ever has.

Finally, I thank my my parents and my sister who have never lost their faith in me and beleived in me without questioning.

TABLE OF CONTENTS

Appendix

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Together with lipids, polysaccharides and nucleic acids, proteins are a class of biological macromolecules that make up the biological organismsô primary constituents. Proteins can be defined simply as polymers constituted of specific sequence of amino acids linked together with the help of peptide bonds.

A peptide bond is a chemical bond formed between two molecules when the carboxyl group of one molecule reacts with the amino group of the other molecule, releasing a molecule of water. This is a dehydration synthesis reaction, and in the case of proteins, the formation of peptide bond occurs between amino acids. In Figure 1.1, R and Rô are the two amino acids that are linked with a peptide bond. The end of polypeptide chain having a free amino group is called N-terminal while the other end owning a carboxyl group is called the C-terminal of the chain.

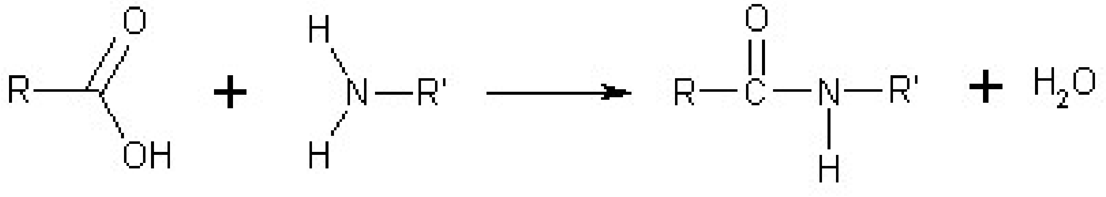


Figure 1.1: Peptide bond formation

The peptide bond shows the characteristics of a partial double bond with an estimated ratio around 40% under typical conditions [2]. Under normal pH values, the peptide bond is uncharged; however due to its double-bonded resonance form, it has an unusually large dipole moment. As a result of this dipole moment, certain secondary structures such as the alpha helix and beta sheets merge, producing a large net dipole rendering the rotation around peptide bond infeasible and fixing the rotational angle of the peptide bond around 180^o. The fixation of rotational angle of peptide bond, which is called the angle ɤ , makes the O, C, N and H atoms of a residue to lie on a rigid planar unit as shown in Figure 1.2. However the other rotational angles (ū , Ɋ) of the residue can take values in a range defined by the Ramachandran Map that will be discussed later in this chapter [3].

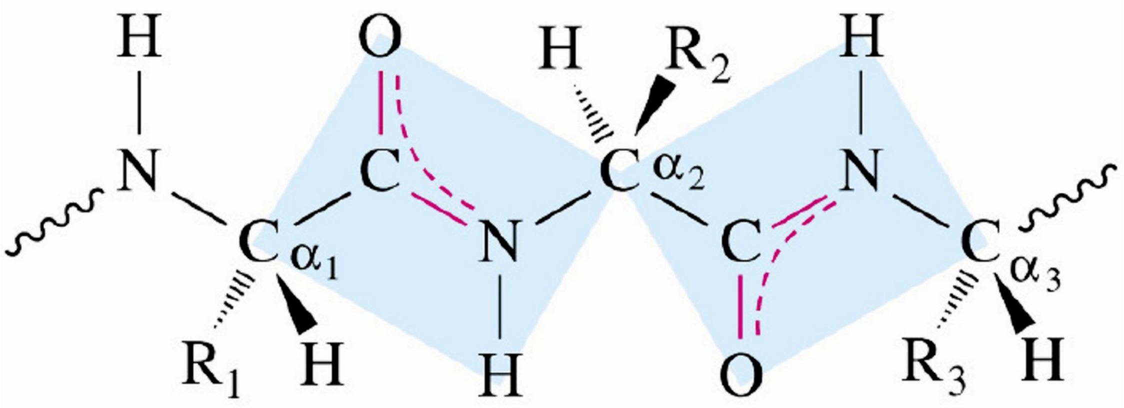


Figure 1.2: The planarity of the peptide bond and other rotational angles [1]

The calculation of a torsion angle between two atoms involves consecutive four atoms. In other words, if the torsion angle between atoms i and i+1 is to be calculated, the atoms i-1, i, i+1 and i+2 should be considered. The ū$_i$ torsion angle describes rotations about the N$_i$Cŭ bond (relevant four atoms are C$_{i-1}$,ӞN$_i$, Cŭ and C$_i$), and the Ɋ$_i$ torsion angle describes rotations about the $C_{ŭi}C_i$ bond (relevant atoms are N_i, $C_{ŭi}$, C_i and N_{i+1}).3̈The ɤ $_i$ torsion angle describes rotations about the peptide bond, C_iN_{i+1}(relevant atoms are $C_{ŭi}$, C_i, N_{i+1} and $C_{ŭi+1}$), making the first ū angle and the last Ɋ angle of a protein structure undefined.

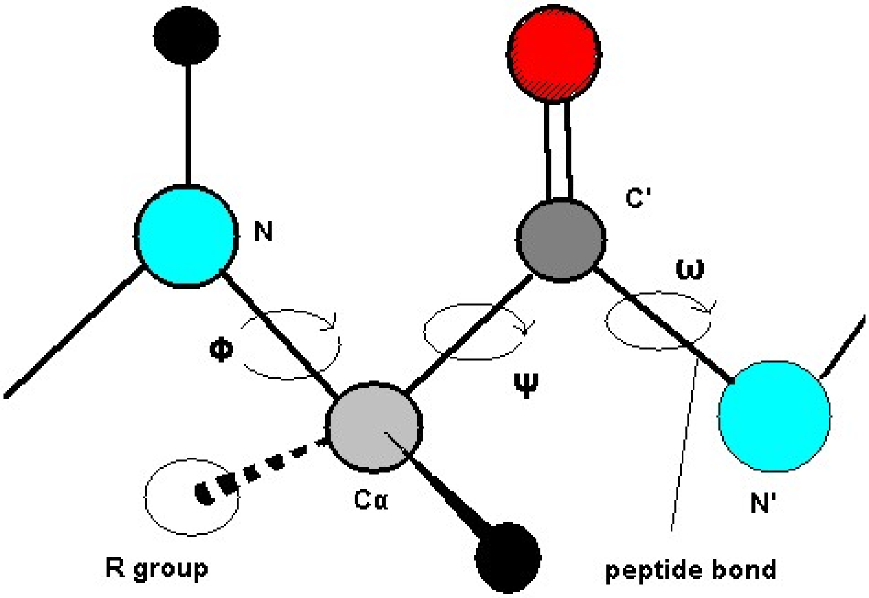


Figure 1.3: The rotational ū , Ɋ and ɤ angles of a residue [1]

Since bond lengths and bond angles are fairly unvarying in the known protein structures, the key point of protein to fold to its three dimensional native conformation lies in the torsion angles of the backbone which can be considered as the degrees of freedom of a protein structure. Therefore, the native conformation of a protein can be identified as the sequence of torsion angle pairs of its successive residues.

As referred to earlier, the principal determining factor of a protein's conformation is the rotation of its bonds. When the ū -Ɋ angle pairs for known protein structures are studied, it can be clearly seen that these angle pairs are not distributed evenly and equally among all possible angle choices. Particular ū angle values prefer to occur with specific Ɋ angle values and vice-versa. The reason of this non-uniformity of preference is that certain ū -Ɋ pairs will try to put two atoms into the same volume causing a steric clash. These collisions make that ū -Ɋ pair very improbable. The ū -Ɋ rotation angle pairs accumulate mostly at regions that span the space of distance that keeps atoms safely away from each other avoiding clashes and collisions. In addition to these, attractive forces between the two atoms of a residue also have a contribution to the choice of - state preference.

Ramachandran Maps are two-dimensional plots of ū -Ɋ angle pairs having ū angle values on the x-axis and the Ɋ angle values on the y-axis. The angle pairs plotted come from ū -Ɋ angle data retrieved from the protein sequences with known three-dimensional structures from the Protein Data Bank (PDB). The PDB is a repository for 3-D structural data of proteins and nucleic acids [4]. This 3-D data, obtained experimentally by X-ray crystallography or NMR spectroscopy, is submitted to the databank and is released to the use of researchers, and can be accessed for free. The database is the central repository for biological structural data. In other words, Ramachandran Map is a way to visualize torsion angles ū against Ɋ of amino acid residues in proteins. It has the information for all possible combinations of ū and Ɋ and therefore all possible conformations for a polypeptide chain.

Since understanding the function of proteins bears the key to understanding all cell functions and therefore all mysteries of the organism, knowing the three dimensional native structure of a protein leads to knowing its function and understanding how cells operate. Since structure of a protein determines its function, the multiplicity of functions means multiplicity of probable three-dimensional structures for proteins.

The particular amino-acid sequence of a protein leads it to fold into its native conformation or conformations and therefore many proteins fold spontaneously to their native state during or after being synthesized. Although proteins may be seen as self-folding, the characteristics of the solution in which they are found, salt concentration in the environment, the temperature range and pH greatly affects the process of folding [5,6,7]. At the basic level of folding, firstly the secondary structures, namely alpha helices and beta sheets are established and only afterwards tertiary structure.

In certain environments and under some conditions told above proteins donât fold at all. These conditions cause the protein to unfold or denature and lack to build the secondary and tertiary structures needed for the protein to be functional [8,9]. A denatured protein deficient of secondary and tertiary structures exists in a condition called random coil.

A secondary structure specifically alpha helix or beta sheet is a repeating three-dimensional form with a fixed bonding pattern. These structures are not formed by strong covalent bonding, but by weaker hydrogen bonding between backbone amide groups.

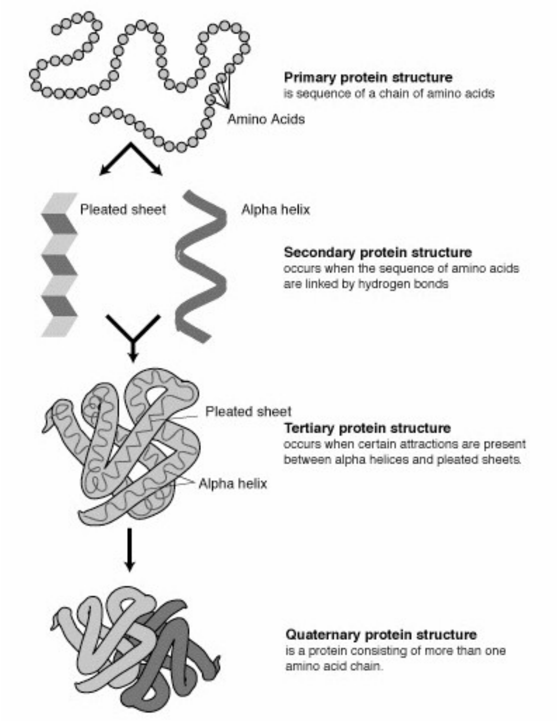

Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

Figure 1.4: Protein structure, from primary to quaternary structure [38]

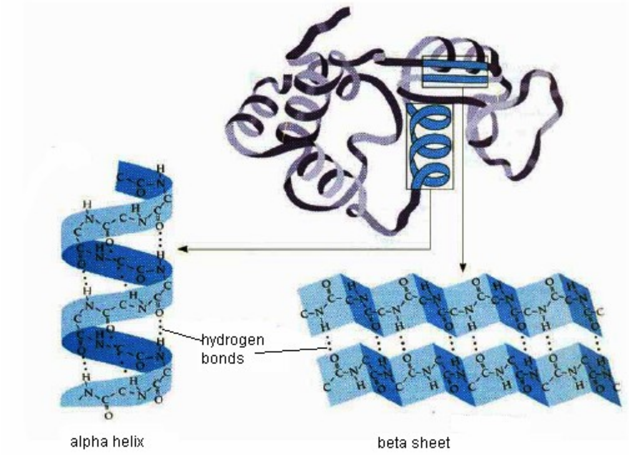


Figure 1.5: The detailed formation of secondary structures [1]

Rotational angles corresponding to specific secondary structures can be specified with the help of Ramachandran Map. Ramachandran Map is a periodic space with boundaries [-180^o, 180^o] x [-180^o, 180^o]. Correlations of co-existence of specific ū -Ɋ pairs have been investigated by many researchers [10-14]. The calculation of statistical averages and corresponding correlations of torsion angles of protein sequences with known structures opens the path to the prediction of native state of proteins sequences with unidentified tertiary structure.

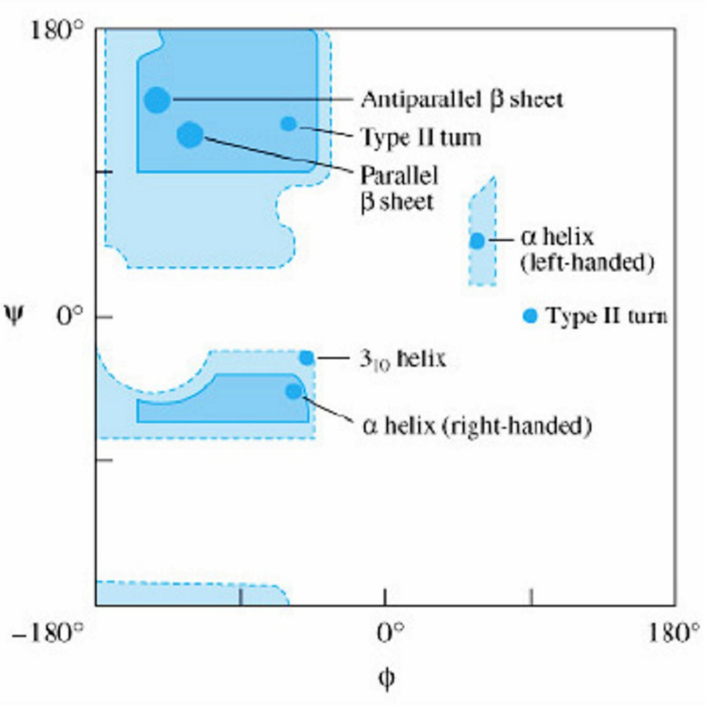


Figure 1.6: Ramachandran Map showing the corresponding regions of secondary structures[1]

Many factors determine an amino acidôs secondary structure propensity and ū -Ɋ state preference in a specific context, such as side-chain interactions, hydrophobic contacts and steric effects. However, one of the most important factors that predispose a residue to be in a specific secondary structure is the effect of its neighboring residues. This neighboring residue effect can also be seen in the ñcoil libraryò part of the PDB that has the information of residues not included in an alpha helix or beta sheet.

Most proteins fold into unique 3-dimensional structures. The shape into which a protein naturally folds is known as its native state. Apart from the native state a protein can be found in two additional forms that are consecutively random coil state and denatured state. Random coil is a state in which amino acids are oriented randomly, while still being bonded to adjacent amino acids. However, except environments with extreme pH values, proteins cannot be found to be in the random coil state. As defined by Dill et al. the denatured state of a protein is a distribution of many different molecular conformations, the averages of which are measured by experiments [15]. Our reference point involves the assumption that the denatured state of a protein can be represented by coil library since by omitting residues located in secondary structures, the regular interactions associated with these secondary structures are eliminated and the average distribution for different protein conformations defined by Dill et al. is provided.

Contribution:

Conformational preferences of amino acids are suitably described by adopting the ű and ɣ torsion angle representation, and the associated Ramachandran Maps. The free energy surfaces constructed over these maps indicate well-defined basins. A big amount of research has been done on the correlations of ū and Ɋ angle preferences of amino acids in a specific secondary structure or in a specific sequence [16,17]. In this study, the context dependency of amino acids preferences for torsion angles was shown using statistical weights of torsion states of ū -Ɋ pairs of protein fragments by using knowledge-based pair-wise dependent ū -Ɋ energy maps of coil-like structures and full structures from PDB. Results obtained using PDB data were interpreted to see the extent of correlations between adjacent torsion angle pairs belonging to both the same and different residues. These correlations favor the choice of the native state torsion angles, and they are strongly context dependent determined by the specific amino acid sequence of the protein as expected. To represent ū and Ɋ angles simply and discretely, the rotational isomeric state model was adopted. With the help of this model, the probability of a given residue with specific neighbors in a specific context to be in each isomeric state was calculated and comparisons between the predictions and actual propensities of residues were made. In addition, the probabilities of sequence fragments to adopt a secondary structure were computed and a reliability score for each prediction was calculated.

A chameleon sequence is one that may exist either in a helical or an extended configuration in a databank. This is an indication that the overall probability of occurrence of a chameleon an H or an E is close to each other. The performance of a prediction method can be evaluated by measuring how well it works on chameleon sequences. The method developed in this study does not only make estimations about the probable secondary structure of an amino acid in a chameleon sequence, but also gives a confidence value for each prediction. In other words, it is a self-evaluating method giving the reliability score for each amino acid. Consequently, what determines this method from others is that it shows how to interpret knowledge based probabilistic as well as it self evaluates reliability score for its performance.

Outline:

Chapter 2 summarizes the related work done on coil libraries, knowledge-based potentials and models adopted rotational isomeric state approximation. Chapter 3 gives the information about the methods and models used throughout this study. In that chapter, insight information about Rotational Isomeric States, knowledge-based probability functions, statistical weight matrices, chameleon sequences and Markov dependent probability functions are given. Chapter 4 encapsulates the results of probability calculations and questions the performance of the method. Analyses of results are also available in this chapter. The last chapter finalizes the study giving an overall look and offering probable future works.

Chapter 2

RELATED WORKS

Ramachandran put forward the correlations between ű and ɣ angles of a single residue by including the exclusion of steric overlaps, which hold for both denaturated and native proteins. The scope of this study is extended to the analysis of - angles to a sequence of amino acid pairs, in other words, doublets. Interactions among neighbor residues are analogous to short-range interactions along the primary sequence of a protein; however, these interactions cannot suffice alone to give information about the overall tertiary structure of proteins as was discussed by Bahar et al. [18].

The main motivation that draw the path of this study about the deeper analysis of - Ramachandran Maps and statistics of amino acids comes from one of the works of Karplus who showed that torsion angles have been distributed with a structure on the Ramachandran Map [19].

The use of Ramachandran Plots to define the states of a given residue agrees with the Rotational Isomeric State formalism was introduced by Volkenstein, Flory, and others [20-22]. The formalism defines the preferred torsion states of chain bonds, similar to the various basins of the Ramachandran Maps and uses them to predict especially the spatial dimensions of synthetic, flexible-chain polymers. An advantage of the method is that the specific chemical structure of the chains can be incorporated into the formalism by specifying bond lengths, bond angles, side groups and all interactions resulting from the interactions of these. Within this context, the formalism should also be useful for studying the dimensions of unfolded proteins, which have heterogeneous sequences and chemical structure is of importance.

The state of a residue in the absence of neighboring residues indicates the intrinsic propensity or the backbone preference of that residue to be in that state. Several researchers have investigated the intrinsic propensities of each amino acid type to prefer to be in an alpha helix, beta sheet or coil region. There are more than one factor that determines these propensities such as hydrophobic tertiary contacts, steric clashes, side chain interactions of neighboring residues and etc. on the other hand, Thornton and collaborators used a statistically based approach relying on the - preferences of residues in sequences that were assigned coil regions as the secondary structure [23].

When the residue is embedded in the polypeptide chain, its states may be correlated with those of the neighboring residues (local correlations) along the chain and those distant along the chain (long-range correlations).

The first time that neighboring residue effect was demonstrated was by Penkett et al. [24] and they introduced coupling constants of peptides by NMR studies. Jha et al. studied structural propensities for alpha helices, beta sheets in a restricted coil library and the conclusion was that these propensities are often strongly influenced by both the chemical nature and the conformation of neighboring residues, contrary to the Flory isolated residue hypothesis [25]. The physical cause of the neighboring residue effect was studied by Avbelj et al. [26].

The Flory isolated residue pair hypothesis assumes that in the random conformational state two neighboring residues along the chain are statistically uncorrelated in the absence of long-range correlations [20, 27, and 28].

This statement is based on the observation that if the chain is kept in its linear conformation and the ū$_{i}$,Ɋ$_{i}$ and ū$_{i+1}$, Ɋ$_{i+1}$ pairs are varied over all allowable values given in the

Ramachandran Maps, no combination of these four rotations will bring the residue i into interaction with residue i+2. If the rest of the chain is not fixed in its linear shape when the four bonds are being rotated as stated above, then residue i+k, for any k>2, may interact with residue i. An interaction of this type is classified as a long-range interaction. Keeping the rest of the chain in its linear form corresponds to isolating the pair i, i+1.

Calculations on tripeptides and longer sequences show that the Flory isolated pair hypothesis is not strictly true [29-31]. Deviations from isolated pair hypothesis are due to near neighbor (NN) effects. More specifically, the NN effect implies that the two sets of the angles $_i$, $_i$ and $_{i+1}$, $_{i+1}$ cannot take values independently. Although the origin of the NN effect is not fully understood yet, the electrostatic screening model can explain why the can explain why the angles are shifted toward more negative values if the neighboring residues of a given residue X are aromatic or beta branched [26].

Keskin et al. used the Rotational Isomeric State model in order to calculate the correlations between the torsion angles of chymotrypsin inhibitor 2 [32]. The first was using the knowledge-based pair-wise dependent - energy maps derived from the PDB and the second way was collecting torsion angle data from a set of random coil configurations. Their study showed that knowledge-based data derived from PDB shows strong correlations between adjacent torsion angles of the same and different residues. These correlations can be thought as favoring the native state of a residue and as strong identifiers of context dependency.

When probabilities are derived by the knowledge-based approach, several ñenvironmentalô factors contribute to the configurational state of a residue. First, the neighbors of a residue along the chain exist at specific conformations in the native state. For example, a residue in a helical sequence sees a different neighborhood than if it is in a beta strand. This effect is referred to as the ñcontext effectô, which may, however, average out if the database is large enough and all possible neighborhoods are available [33]. Secondly, every protein in the database is in its native compact state, and long-range forces between residues that are spatially close but far apart along the chain contour are dominant. The differences between database statistics and molecular simulations have been addressed in several papers.

Hermans and collaborators compared the results of simulations and database statistics for five amino acids and discussed the sources of the differences between the two [34]. The influences of the local acid sequence on - probabilities were investigated by Garnier and collaborators [35]. The angle probabilities estimated from a databank were shown to be context sensitive and position dependent [36].

Serrano used a coil database and identified the real intrinsic propensities independent of context effects [33]. Similarly, Thornton and collaborators determined the intrinsic properties of residues from a coil data bank [23]. Coil libraries are constructed from residues in the non-structured regions of native proteins with the expectation that contributions from the near neighbor and resulting context effects areas small as possible.

Sippl was one of the researchers who adopted the insight that the molecular structures identified with experimental methods contained a large amount of information on the stabilizing forces within proteins, and statistical analysis had the potential to uncover the key rules in charge of protein stability [37]. He also claimed that along with statistical mechanics, statistical analysis of proteins with known three-dimensional structures is a potent tool to derive potential functions from a database of known structures.

Wodak et al. also used different types of potentials derived from a dataset of known protein structures by computing statistical relations between amino acid sequence and different descriptions of the protein conformation [38]. They deployed these potentials to formulate backbone dihedral angle preferences, pair wise distance-dependent interactions between amino acid residues, and solvation effects based on accessible surface area calculations.

Knowledge-based potentials can be used to determine whether a specific amino acid sequence is inclined to fold into a specific native tertiary structure. Stadler et al. used this idea to investigate the sequence structure relations in proteins with a method using neural networks [39].

Chapter 3

MATERIALS AND METHODS

Backbone conformation of a specific protein is completely identified by the configurations of its torsion angle pairs, namely and angles. Interactions among amino acids of a protein sequence determine the final stable three-dimensional structure of the protein. The characteristics of both short range and long-range interactions can be inferred from the torsion angle pairs of proteins with known structures. Predictions for the three dimensional structure of newly identified protein sequences can be made with the help of these torsion angle characteristics.

In the present study, we investigate the relationships between amino acid sequence and torsion angle preferences, in other words, the most probable ū , Ɋ configuration that an amino acid would prefer to obtain in a sequence with specific preceding and succeeding amino acids. This study consists of two parts that have the same basis and divert to different paths.

3.1.The Basic Computations:

At the basis of these two different parts of the study lies the derivation of torsion angle couples (ū , Ɋ) from PDB (PDB) structures. In order to set the fundamentals for the angle libraries, 2223 non-redundant and non-homologous PDB structures that are the representatives of all PDB molecules were acquired from PDBSelect (The list of non-redundant PDB structures is provided in the Appendix A.3) [40,41]. In addition to the sequence information of these PDB proteins, secondary structure information was taken from Database of Secondary Structure in Proteins (DSSP), which is a database of secondary structure assignments for all protein sequences in PDB [42,43]. Originally the DSSP has eight secondary structure assignments probable for a residue; however for the conventional concept of secondary structures being alpha helices, beta sheets and coil (or loop) regions, DSSPôs three-structure model was adopted. The usage of this three-state model also diminished the computation time and supplied more compact data for the further secondary structure predictions. The conversion from eight-state model to three-state model can be seen in Table 3.1a and 3.1b.

Type	Description
B	isolated &beta-bridge
E	extended &beta-strand
G	3_{10}-helix
H	alpha-helix
I	pi-helix
S	bend
T	turn (isolated)
U	none of the above

Table 3.1a: Eight state model of secondary structure identification

Type	Description
H	helix, (G, H, I)
E	strand, (B, E)
C	coil, (S, T, U)

Table 3.1b: Three state model of secondary structure identification

3.2. Rotational Isomeric States

Once the secondary structure assignments were completed, the other important feature of this study was to begin. The torsion angels for 2223 non-redundant PDB structures were calculated according to the IUPAC-IUB standard and the calculated angle values were converted to *states* in accordance with the rotational isomeric state model we used in order to calculate statistical averages and correlations for torsion angles of proteins [20, 21]. All unidentified angles, such as ū angles of the first residues and Ɋ angles of the last residues of each protein were fixed at 0^o.

In the rotational isomeric state approximation as defined by Flory earlier, each residue is assumed to obtain several discrete rotational states. Discrete state formalism is used for the torsion angles, where each torsion angle area is divided into intervals of 30^o. Therefore, we have 12 torsion states representing the torsion angles. The space of angles is divided into 12 regions from ï ´ to + ´ with increments of ´ /6. Then we name these regions from 1 to 12 as follows in Table 3.2:

State	Lower bound	Upper bound
1	-´	-5/6 ´
2	-5/6 ´	-4/6 ´
3	-4/6 ´	-3/6 ´
4	-3/6 ´	-2/6 ´
5	-2/6 ´	-1/6 ´
6	-1/6 ´	0
7	0	1/6 ´
8	1/6 ´	2/6 ´
9	2/6 ´	3/6 ´
10	3/6 ´	4/6 ´
11	4/6 ´	5/6 ´
12	5/6 ´	´

Table 3.2: Definition of states

The Ramachandran Map corresponding to the states defined above is as follows in Figure 3.1, blue areas representing alpha helix and beta sheet regions.

After calculating torsion angles and assigning the corresponding states for the 2223 non-redundant PDB structures, two different angle libraries were created: i) the coil library, ii) the full library. The coil library contains the set q_C of residues whose torsion angles were derived only from coil regions of the protein structures. It is known that half of folded proteins have alpha helices or beta sheets as the secondary structure.

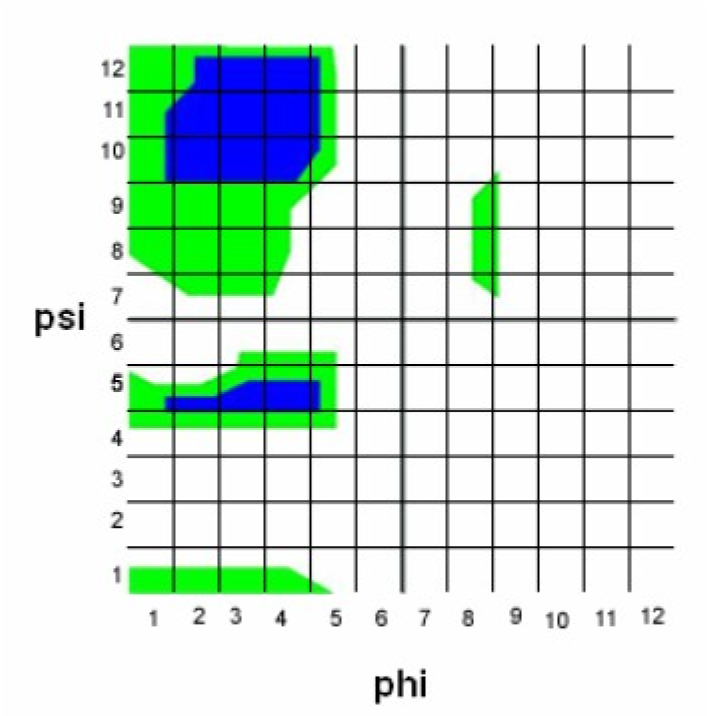


Figure 3.1: Ramachandran map showing alpha helix and beta sheet regions

The creation of coil library has been managed by removing the regions of the proteins that have alpha helix or beta sheet structures. The reason for using angles libraries derived from coil regions leans to the hypothesis that coil regions can be easily treated as frameworks for the unfolded state of proteins [44]. The coil library has the torsion angle information of 45500 residues of 2223 protein structures. On the other hand, the full library was formed from the set ɋ$_{F}$ of all residues in the non-redundant PDB with their corresponding secondary structures taken from the DSSP. The full library contains 202032 residuesô torsion angle states.

The creation of these two libraries was a prerequisite for calculating statistical averages of amino acid pairs, taking the Ɋ angle of the first residue and ū angle of the second one as the reference point.

Since we were looking for the conformational preferences of a specific amino acid in a specific sequence and its torsion bond angle correlations, the location of this amino acid in the sequence, the relation between its torsion angles and torsion angles of its predecessor and successor residues bears the key role to understand the native structure of a protein. Therefore, each consecutive two residues were grouped starting from first and were named doublets. In other words, the management of the doublets is as follows: first doublet contains 1st and 2nd residues and second doublet contains 2nd and 3rd residues and etc. Hence we have n-1 doublets where n is the number of residues in the protein sequence. Each doublet is identified as XY, X being the (i-1)st residue and Y being the ith residue.

Since there are 20 conventional amino acids in nature, there have been 20x20=400 amino acid pairs, in other words doublets, to be kept. Information of residues represented with letters ñBò, ñZò and ñXò by PDB naming conventions was excluded due to the ambiguity of residues they represent. The letter ñBò is used for representing ambiguous ASP or ASN residues, while ñZò is used when there is an ambiguity between GLU and GLN. ñXò is the way to describe a non-determined residue. In order to prevent any statistical error, these experimentally unidentified ñBò, ñZò and ñXò residues of sequences with known 3-D structure weren't used while making calculations.

For each consecutive XY doublet, torsion angles from both the full library and the coil library were selected and 400 output files having ū$_{i-1}$,Ɋ$_{i-1}$,ū$_{i}$ and Ɋ$_{i}$ consecutively were created for each library. Each computation in the first part of the research was performed separately for both libraries. These four angles would help us to estimate the conformation of the protein sequence along with knowledge-based and statistical mechanical techniques.

Among these four internal angles, the main focus is on the pair-wise occurrence of the angle pair (Ɋ$_{i}$,ū $_{i+1}$). In chain molecules, the bond vectors are bound by firm mutual correlations in the sense that the direction of a specific bond is under the influence of the directions of its neighbors in the main chain. In most chain molecules the rotation angle (in the case of proteins, the torsion angle) about a given bond is correlated with the rotations about the bondôs immediate neighbors on either side. The correlation between angle couples (ū $_{i,}$ Ɋ$_{i}$) is a measurement for intraresidual dependency while the correlation between angle pair (Ɋ$_{i}$,ū $_{i+1}$) gives the information about interresidual dependency which determines the overall characteristics of the three dimensional structure of the protein chain. Since the aim of this study is questioning the role of the interdependency of residues in finding the overall structure of a protein chain, and as Keskin et al. mentioned the interresidue correlations would improve the statistics in the non-redundant database, the dependency between the bonds of two neighboring residues may be emphasized mostly by focusing on the angle pair (Ɋ$_{i}$,ū $_{i+1}$).

3.3. First part: calculation of knowledge-based conformational probabilities of a protein sequence:

3.3.1. Knowledge-based probability function:

In the presence of interresidue interactions and correlations, a certain bondôs rotational potential depends on the rotational state of its neighbors. The rotational isomeric state approximation and statistical mechanical approach greatly mimic the exact bond rotation potentials while trying to find out the configuration of molecules and help us to develop the appropriate probability function.

The basic problem of the rational protein design is to theorize a probability function that identifies the native protein structure and stability at best. Anfinsenôs thermodynamic hypothesis that the conformation of amino acids that has the moleculeôs minimum free energy state is the native state of that given protein sets the logical background for finding the proper probability function [45]. Our motive in finding the probability function was the assumption that the state of a residue with the highest probability has the minimum energy and therefore how much of the native stateôs configurational information is contained in the denatured state [46]. However, finding an appropriate probability function is an important issue since scoring functions used in estimating free energy change due to folding are not well defined at a physical chemical level and frequently get vague even while investigating experimentally observed energetic properties of proteins. However, the dramatic growth of empirical information from sequence and tertiary structure databases enabled us to model a probability function that uses information obtained from these databases. When extracted in the form of statistical means appropriate to be used for computational algorithms, this information from databases is mentioned as knowledge-based potentials, a way to both lessen the complexity of searching the sequence space and refine the scoring functions of the prediction methods. Secondary and tertiary structure databases have the vital role in developing the knowledge-based probability function.

3.3.2. Statistical weight matrices for interdependent bonds:

As Flory stated in his book, a given conformation (ű) of a chain molecule consisting of *n* bonds can be represented by means of the rotational isomeric state scheme in a *v-digital* system, *v* being the number of rotational isomeric states that are predefined to represent torsion angles of a bond. For instance, if the number of rotational isomeric states is v=2 and the representatives of rotational isomeric states being 0 and 1, the chain molecule having n bonds may be represented as:

0 1 0 0 1 1 etc.

If it is assumed that the rotational potential of a given bond i depends only on its first neighbors, bond i-1 and i+1, and longer range interactions are not of great importance to the approximation, total configurational energy of the molecule can be induced simply to be the sum of energies of the first neighbor pairs. Therefore the total configurational energy of a chain molecule with the conformation (ű) becomes:

$$E \quad E_0 \quad E_{01} \quad E_{10} \quad E_{00} \quad E_{01} \quad E_{11} \qquad \text{Eq(3.1)}$$

The first term on the right hand side carries a single index since it doesnɑ̂t have a preceding bond in the sequence, and all terms other than the first term carries a double index , first index being the bond preceding the second index. The total energy equation can be generalized as follows:

$$E\{ \ \} \quad \overset{n \ 1}{\underset{i \ 2}{}} E_i(\ _{i \ 1}, \ _i) \quad \overset{n \ 1}{\underset{i \ 2}{}} E \ _{;i} \qquad \text{Eq(3.2)}$$

ɕ being the state of bond i-1 and ɖ being the state of bond i. The energy,

$$E \ _{;i} \quad E_i(\ _{i \ 1}, \ _i) \qquad \text{Eq(3.3)}$$

can be identified as the contribution of bond i in state ɖ while bond i-1 is in state ɕ to the total energy of the molecule. It may seem that the dependence of bond i to the bond i+1 is overlooked if this approach is followed. However, it is merely embedded in the following term of the sum. Consequently, the total energy can be encapsulated systematically as the sum of terms of energies of dependent consecutive states of torsion angles.

Statistical weight matrices were created by using the relationship

$$u_{\ ;i} \quad \exp(\ E \quad / RT) \qquad \text{Eq(3.4)}$$

U being the statistical weight matrix:

$$U_i \quad u_{\ \ i} \qquad \text{Eq(3.5)}$$

with states (ɖ) for bond i indexing the columns of the statistical weight matrix U_i and states (ɕ) of bond i-1indexing the rows.

The partition function is given by the equation:

$$Z \quad J*[\ {}^{n}_{i\ 1} \ U_i]J \qquad \text{Eq(3.6)}$$

$$J* \quad [1,1,1...1] \qquad \text{and} \quad J \quad \begin{matrix} 1 \\ 1 \\ . \\ . \\ 1 \end{matrix} \qquad \text{Eq(3.7)}$$

The relative probability of the frequency of a configuration {ū } may be represented by its statistical weight. The probability of a given sequence to adopt a specific conformation is equal to the statistical weight divided by the sum of statistical weights of all possible configurations of this molecule, which is the partition function Z.

$$P\{ \ \} \quad Z^{1} \ {}^{n\ 1}_{i\ 2} \ u_{\ ;i} \qquad \text{Eq(3.8)}$$

As Flory stated, while evaluating the partition function by taking the product of statistical weight matrices U_i for each configuration would be prohibitive for a sequence having a large number of residues. Since multiplications were done within matrices with very small entries, some multiplication results converged to zeros as the sequence length went larger. In order to avoid this problem, the sequences were cut into fragments of length 30.

For the first part of our study for calculating knowledge-based probability calculation of a protein sequence, we adopted an approach derived from Floryôs method.

3.3.3. Probability Levels:

By the knowledge-based approach, the rotational isomeric state probabilities of each ū and Ҁ angles for each residue in a given protein sequence to be in a specific state were calculated. After calculating the probabilities for each one of possible twelve states of the two torsion angles, ū and Ҁ separately, these probabilities were sorted in ascending order. Once the sorting of the probabilities of states of each torsional angle was completed, each residueôs native stateôs order was checked. The sorting process helps to find a proper scoring method which is the use of probability levels. ñThe probability levelò term is defined to score the achievement of the Markov dependence approximation of the rotational isomeric states for each residue. In order to simply define what probability level is, it can be deduced that it is the index of the native state of the residue in question in the sorted prediction results vector. The probability level ranges from 1 to 12, 1 being the state with the smallest calculated probability and 12 being the state with the largest calculated probability. The states of the torsion angles from 1 to 12 were arranged in increasing order with respect to their probabilities. For the ith residue, for instance, if the highest probability is the same as in the native state, we identify the probability level of the ith bond as 12. We calculated the probability level of each bond in this manner. These assigned probability levels were then used for scoring accuracy of the predictions made by probabilities.

The overall scoring for the accuracy of prediction was done with the help of these probability levels. All probability levels were calculated separately for each state of both torsion angles for each protein sequence separately and their mean was calculated for each protein sequence. As a result of this operation, for each protein sequence, we were left with two prediction accuracy scores: one for the ū angle prediction and one for the Ɋ angle.

3.4. Second part: calculation of knowledge-based conformational probabilities of chameleon sequences in a protein sequence:

Next step was to check whether the doubletsô and individual amino acidsô preferences have the tendency to be in certain states in the Ramachandran plot or not. However, all the individual residues and doublets couldnôt be expected to prefer the right conformational states since they could be at any conformation in different proteins, in an alpha helix in one secondary structure whereas it could be in a beta-sheet in other protein.

This non-uniform nature of amino acids led us to consider the chameleon segments which are identical sequences that adopt different secondary structural properties in different proteins [47,48]. Prior research has shown that k-mers of different lengths (k ranges from 5 to 8) with identical primary sequence can be found to be in varied conformations in unconnected proteins [49,53]. These k-mer chameleon sequences obscure the correct assignment of a secondary structural property to a residue and overall structural prediction of a protein sequence. The most challenging distinctive factor that decides whether a secondary or tertiary structure prediction method is efficient or not is the methodôs discriminating chameleon sequences correctly.

The aim of this part of the study is to find these chameleon sequences in the non-redundant

PDB and control the predicted secondary structures of these chameleon sequences with the actual frequencies extracted from the DSSP library of non-redundant PDB.

3.4.1. The creation of chameleon libraries:

Like the doublets, the tendency of singlets for each state in the Ramanchandran plot show variance and dependent probability of each residue to the neighboring residues are calculated for secondary structures: alpha helix, beta sheet and coil regions. The calculated probabilities are then compared with the actual frequency of seeing these sequences in certain secondary structures in the complete PDB. The actual frequency data is provided from the chameleon libraries created from the non-redundant PDB protein structures.

Firstly, all protein sequences in the non-redundant PDB were divided into fragments of length 4, 5, 6, 7 and 8 residues. These lengths were chosen since they are the minimum number of residues required to create a secondary structure element. In addition to this, fragments longer than 8 residues are seen very infrequently, therefore it is hard to gather statistically significant information from fragments longer than 8 residues. The sequences were divided with sliding the frame of length of fragment. For instance, if the fragments of length k are going to be extracted from a protein sequence of length n, then the number of fragments extracted from the protein sequence in question would be (n-k)+1.

Along with the residue information of the fragments, secondary structure information was also obtained from DSSP. The chameleon libraries are like the following table that shows examples of 5-residue long chameleon sequences:

Sequence of frag.	frag. freq.	secondary structures of frag.
LSSGG	5	['TTTTG', 'CCCCS', 'EESSS', 'TTTCC', 'ETTTT']
PEGLR	6	['TTSCB', 'CTTEE', 'HHHHH', 'TTCCS', 'TTCEE', 'HHHHH']
AAATA	6	['HHHHC', 'SCCCC', 'HHHTS', 'EECCC', 'HHHHT', 'SCCCE']
LGLKE	5	['TTCCS', 'CCHHH', 'CSCCB', 'TTCCG', 'EEETT']
KALEL	6	['HHHHH', 'HHHHH', 'HHHHG', 'HHHHH', 'SEEEE', 'EEEEE']

Table 3.3: Example from the 5-mer chameleon library

The first column in the table above is the sequence of the fragment, the second column is the number of times the fragment was seen in the non-redundant PDB and the following list of secondary structure identifiers in brackets is the different secondary structures the fragment sequence happened to obtain in the database.

When creation of the chameleon library was completed, the task to be done was calculating the actual probability of each residue in each chameleon sequence to be in a specific secondary structureð alpha helix, beta sheet or coil region. The actual probabilities below are calculated as the simple probability: P(actual)=(number of secondary structure in question)/total number.

Table 3.4 shows two examples of the resulting actual probabilities of two chameleon sequences of different lengths. The fragment at the upper part of the box is from 5-residue long fragment library and the second is from 6-residue long fragment library.

```
================================================================
The sequence KALEL is seen 6 times
index   aa      helix   beta    coil
1       K       0.667   0.167   0.167
2       A       0.667   0.333   0.0
3       L       0.667   0.333   0.0
4       E       0.667   0.333   0.0
5       L       0.667   0.333   0.0

The sequence  ELKKA  is seen  7  times.
Index   aa      helix   beta    coil
1       E       1.0     0.0     0.0
2       L       1.0     0.0     0.0
3       K       0.85    0.0     0.15
4       K       0.85    0.0     0.15
5       A       0.71    0.0     0.29
================================================================
```

Table 3.4: The fragments above are two examples of actual secondary structure preferences.

3.4.2. Markov-dependent probability calculation with knowledge-based statistical weight matrices:

The calculation of actual probabilities was a prerequisite for comparison with the predicted probabilities to measure how well the method performs in distinguishing chameleon sequences. The probability calculations were done in a similar manner as the first part of the study with the statistical weight matrices extracted from only full library of residues. The frequency of an amino acid being seen in a state of a (ū ,Ɋ) couple is denoted by $f(a)_{i,j.}$ $f(a)_{i,j}$ is simply the number of seeing a specific amino acid a in the specific state (i,j). We go on to calculate internal energy of each amino acid in each state for further U matrix building.

$$E_{i,j} \qquad RT\ln\frac{f_{i,j}}{\underset{allaa}{f_{i,j}}} \qquad\qquad \text{Eq(3.9)}$$

$$U \qquad \begin{matrix} e^{E_{1,1}} & \ldots & e^{E_{1,12}} \\ . & . & . \\ e^{E_{12,1}} & \ldots & e^{E_{12,12}} \end{matrix} \qquad\qquad \text{Eq(3.10)}$$

The table below shows the results of the predictions done with the help of using knowledge-based statistical weight matrices for the two example chameleon sequences in the Table 3.4.

```
=================================================================
Sequence :  KALEL
index  aa    helix prob           beta prob           coil prob
1      K     0.944                4.09e-06            0.056
2      A     0.844                0.023               0.132
3      L     0.843                0.156               3.56e-06
4      E     0.035                0.006               0.958
5      L     0.198                0.801               1.123e-05

Sequence  :  ELKKA
index  aa    helix prob           beta prob           coil prob
1      E     0.0071               0.0007              0.992101217072
2      L     0.8872               0.1126              9.14533406293e-06
3      K     0.9636               1.109e-05           0.0363073774115
4      K     0.9715               1.502e-05           0.0284796672393
5      A     0.3835               0.0625              0.553925991115
=================================================================
```

Table 3.5: Example of prediction results for two chameleon sequences of length 5 and 6 respectively

3.4.3. The calculation of secondary structure probabilities:

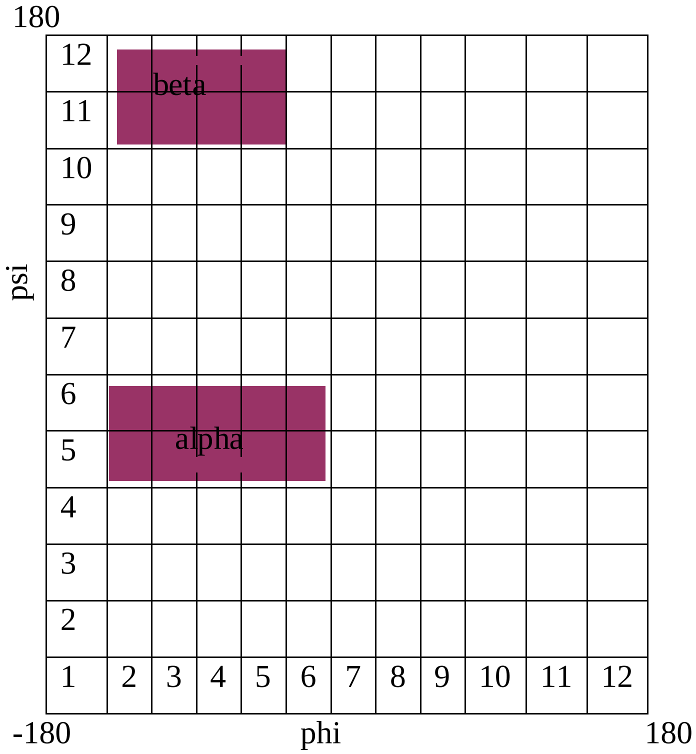


Figure 3.2: states representing secondary structures

The Figure 3.2 shows the states assigned to specific secondary structures. The probability of each secondary structure is simply calculated as the sum of probabilities of states that was predefined as being a given secondary structure. Helix probabilities were calculated by adding the probabilities of regions shown in Figure 4, changing ū from 1 to 6 and Ɋ from 3 to 6. Beta probabilities are calculated in the same manner, changing ū from 1 to 5, Ɋ from 10 to 12. Coil probability is simply 1-(helix prob + beta prob).

3.4.4. The comparison of actual probabilities with the predictions:

To discriminate the accuracy of the knowledge-based conformational probabilities, a scoring method was developed. The method simply measures the correlation between the actual secondary structure probabilities of a given residue in a given chameleon fragment and the predicted probability of the residue in question. In order to get a score, each residue was treated as a three dimensional vector having alpha helix probability, beta sheet probability and coil probability as x, y and z coordinates.

[x, y, z]=[P(alpha helix), P(beta sheet), P(coil)] Eq(3.11)

Therefore each residue in each fragment has two three dimensional vectors, one containing information of the actual probabilities and the other of the predicted probabilities. Each amino acid has these vector sets as many as the times it was encountered in the chameleon database. To come up with a single overall actual and prediction vector, the mean values of all coordinates of the vector were calculated and overall vectors are created. If the actual probability vector is shown with P and predicted probability with Pô, and the overall mean vectors as $\overline{P}$ and $\overline{P}'$:

$$\overline{P}(a) \quad [\overline{x},\overline{y},\overline{z}] \text{ and } \overline{P}'(a) \quad [\overline{x}',\overline{y}',\overline{z}'] \qquad \text{Eq(3.12)}$$

a being the residue in question and $\overline{x}$ being the mean of actual helix probabilities, $\overline{y}$ being the mean of actual beta sheet probabilities and $\overline{z}$ being the mean of actual coil region probabilities. $\overline{x},\overline{y},\overline{z}$ are the values for prediction results.

After having the mean vectors, the correlation C between the actual and predicition results was computed as follows:

$$C(a) \quad \frac{\overline{P}(a) \quad \overline{P}'(a)}{\left\|\overline{P}(a)\right\| * \left\|\overline{P}'(a)\right\|} \qquad \text{Eq(3.13)}$$

$|P(a)|$ and $|P'(a)|$ being the lengths of two probability vectors. The C value ranges between 0 and 1, 1 representing 100% accuracy of the prediction.

The correlation scores for each amino acid are separately calculated for the five different chameleon sequence lengths in addition to the calculation of the overall correlation score.

In addition to the fragment length specific evaluation, secondary structure specific evaluations were done. In this kind of evaluation, the accuracy of predicting the native secondary structure for each residue in the database was calculated. However, these calculations are slightly different than the the calculations above. This time each amino acid has a set of six probability vectors.

$$\begin{array}{l} [P_{act}(a)]_{helix} \quad [p_{helix_1}, p_{helix_2}, p_{helix_3}, \ldots, p_{helix_m}] \\ [P'_{pred}(a)]_{helix} \quad [p'_{helix_1}, p'_{helix_2}, p'_{helix_3}, \ldots, p'_{helix_m}] \end{array} \qquad \text{Eq(3.14)}$$

with a as the amino acid in question, m as the number of times seeing the amino acid a in the chameleon database. The equation above shows the probability vectors only for alpha helices. Each amino acid has 4 more of these vectors for beta sheetsô and coil regionsô actual and prediction values.

The correlation calculation for each secondary structure was done between the actual probability vector and prediction vector as follows:

$$[C(a)]_{helix} \quad \frac{[P_{act}(a)]_{helix} \quad [P'_{pred}(a)]_{helix}}{\|[P_{act}(a)]_{helix}\| \quad \|[P'_{pred}(a)]_{helix}\|} \qquad \text{Eq(3.15)}$$

The correlations also for beta sheets and coil regions are calculated and they can be seen in the following Chapter 4, Results and Discussion.

Chapter 4

RESULTS AND DISCUSSION

4.1. Results for the first part of the study:

After calculating the probabilities with the methods explained and equations detailed in chapter Models and Methods, the next step was measuring how well our method works on sequences whose torsion angle preferences were known. Since we were using a rotational isomeric state model having 12 distinct states for representing all possible values for each torsion angle, the probability levels' values range from 1 to 12, 1 for the smallest prediction value and 12 for the largest prediction value for each residue. Then, the probability level of the native state was found. For instance if the native state was predicted best then it had the probability level 12, however if it was the second best its probability level would be 11.

For instance, Figure 4.1 shows the Φ angle state preferences of amino acid alanine (ALA), the statistics was derived from non-redundant full library and represents the overall propensity of alanine independent from the sequence. As it can be concluded from Figure 4.1, the Φ angle of residue ALA has a great inclination for states 4 and 5 which are included in the helix regions in the Ramachandran Map. For the following parts of the study the real values of Φ and Ψ torsion angles of the protein in question will be referred as actual values. In other words, the actual preference of a residue is the (Φ,Ψ) angle couple that the residue in question occupied in a specific protein sequence. The further comparisons will be done between the actual preference of the residue in a given sequence and the computationally predicted states for this residue in the same sequence.

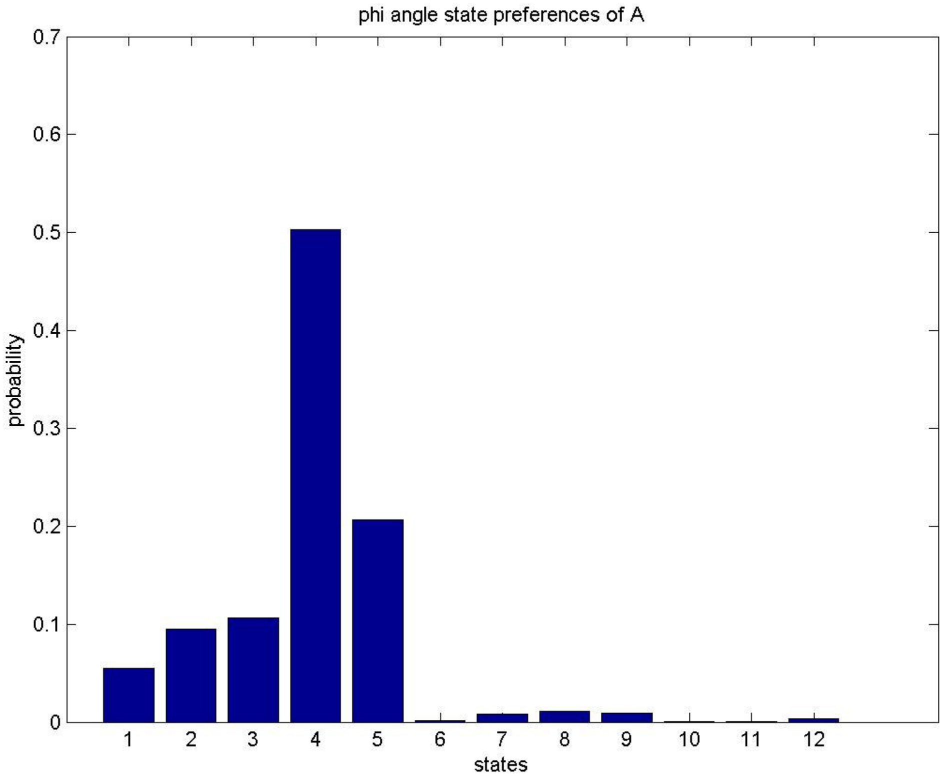


Figure 4.1: Φ angles state preferences of alanine from full library of non-redundant PDB.

On the other hand, in order to show the importance of "context effect", the torsion angle state preferences of an amino acid in a specific sequence should be examined. Figures 4.2, 4.3 and 4.4 show the calculation results of the prediction method with the information from full library. The native states of all residues are marked with a black square in the graphs showing prediction results.

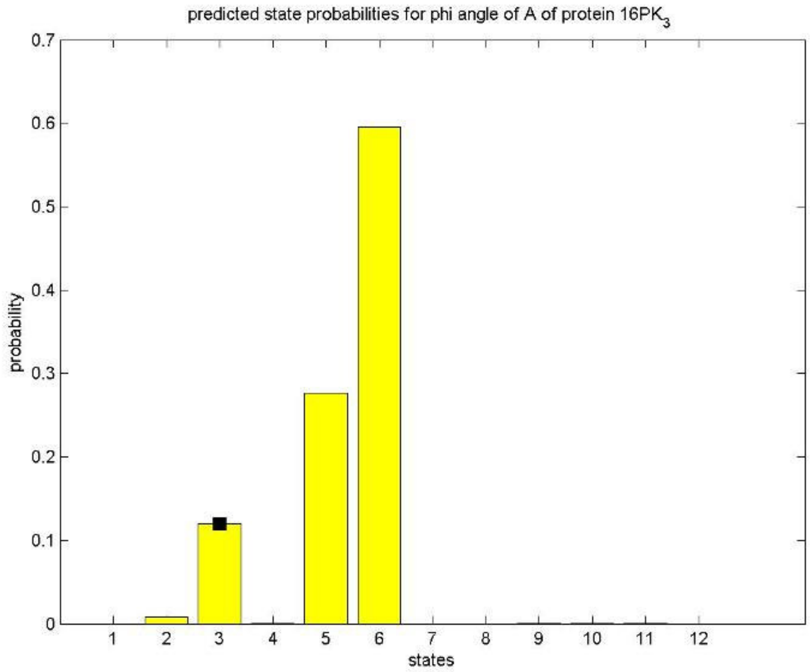


Figure 4.2: The calculated probabilities for Φ angle of the 107th alanine residue of 16PK.

In the native three-dimensional structure, 107th ALA residue of 16PK has the secondary structure beta sheet. This ALA residue’s Φ angle’s native state is state 3 and it was predicted with a probability level of 10 since it is the third best guess.

If Figures 4.1 and 4.2 are to be compared, it can be clearly seen that amino acid ALA prefers to have its Φ angle in state 4 in the overall, however, in the specific case 107th residue of 16PK, amino acid ALA has its Φ angle in the third state.

The reason of this difference is that the Figure 4.2 shows the Φ preferences of alanine specifically in the sequence of 16PK while the Figure 4.1 shows the Φ preferences of alanine in the overall. The difference between the two figures emphasizes the importance of environment that a specific residue occupies.

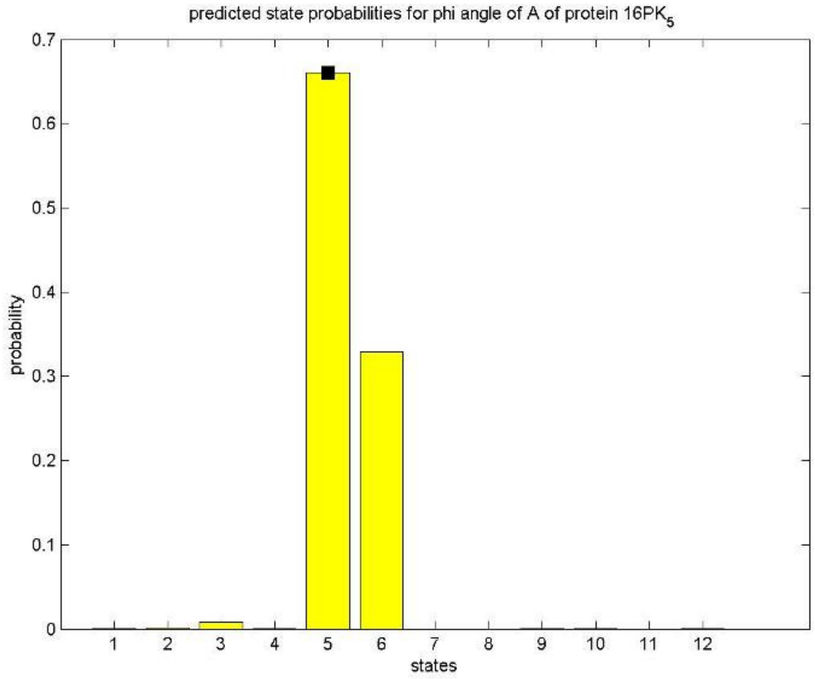


Figure 4.3: The calculated probabilities for Φ angle of the 153th alanine residue of 16PK.

Figure 4.3 shows the calculated Φ angle preferences 153th ALA residue of 16PK whose native state is state 5 and is actually a part of an alpha helix. The native state of this specific angle is in accordance with Figure 4.1 and has been predicted with the best score and has the probability level 12.

The following Figure 4.4 is the graph of prediction results for 123rd ALA residue of protein 135L. This ALA residue has its Φ angle in state 5 and is in a coil region in the native state. It has been predicted with the highest probability and therefore has the probability level 12.

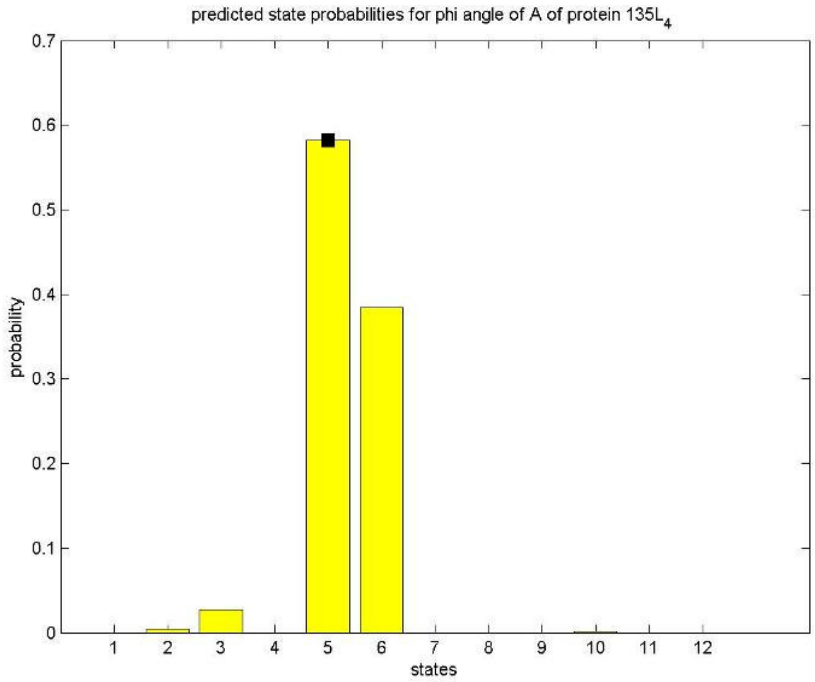


Figure 4.4: The calculated probabilities for Φ angle of the 123rd alanine residue of 135L.

From these figures, it can be concluded that the preceding and following neighbors of a residue and the sequence in which that residue lies have the greatest impact in determining its torsion angle preferences; in other words, three dimensional structure. Due to this interdependency of bonds, the preferences for the native state of each bond are favored by the environment of the residue in other words, amino acid sequence of the protein. Therefore, within a specific sequence, an amino acid may probably occupy a state different than its most probable intrinsic state. Further comparison graphs of amino acid alanine and actual torsion angle state preferences of other amino acids can be found in the Appendix A.1 and A.2.

The determination of probability level for each residue in every fragment was done in this manner. As explained in Chapter 3, 30-residue-long fragments from non-redundant PDB were used for testing the performance of our knowledge-based approach. The mean values of probability levels for each 30-residue-long fragments were taken and only one score for each fragment was assigned as a result. The results for both coil and full libraries were obtained. Firstly, there was only coil library from which angles had been extracted. Since using coil library doesn't seem to boost up prediction results, using only the full library for the further calculations was decided. The results obtained with the fragmented sequences had better overall scores than the results obtained using whole sequence of the proteins. The graphs below show the results from 300 fragmented sequences of length 30 amino acids.

Each point in Figure 4.5 represents a 30 aa-long fragment and the corresponding probability levels are the results for the whole length fragments. The overall probability level of a fragment is calculated simply as the mean of probability levels of each residue in the fragment. The abscissa values are sorted to have increasing probability level values grouped from left to right on the figure.

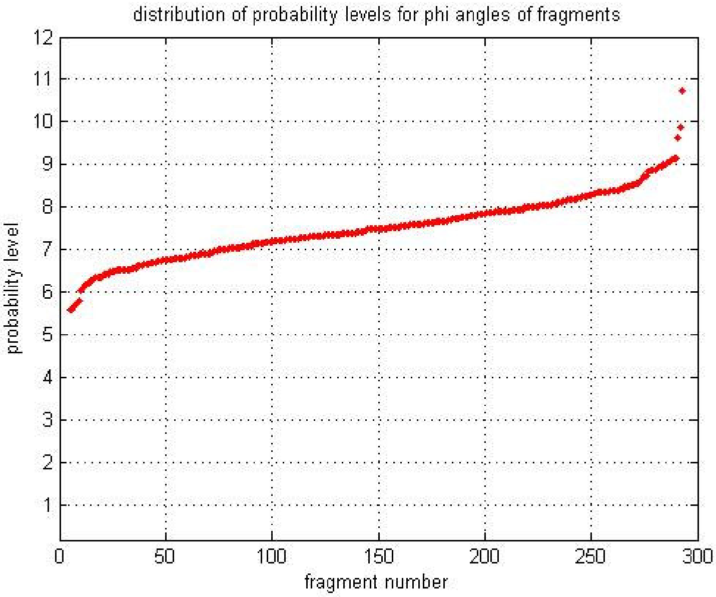


Figure 4.5: Probability level distribution for Φ angles of 300 different fragments calculated with knowledge-based statistics coming from full library.

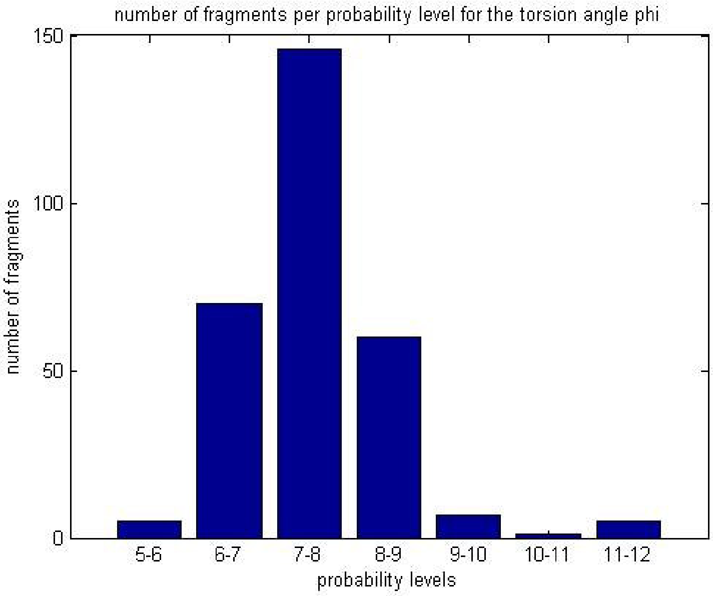


Fig 4.6: Probability level vs. number of fragments from full library for the torsion angle Φ.

Among 300 predicted sequence fragments, the graph above has probability levels on the x-axis and the number of sequences predicted with the corresponding probability level on the y- axis. The graphs above show the distribution of prediction scores for Φ angle from the fragmented sequences of the full library. The prediction scores deviate between 6 and 12, mostly accumulating in the area between 7 and 8 giving a mean around 7.5 out of 12 which means an accuracy around 62.5%. At some points of the graphs there are zero values which mean the probability levels for these fragments couldn't be calculated. The reason is that the fragmenting was done automatically, and therefore there were some fragments shorter than 30-residue-long. Fragments shorter than three residues couldn't be predicted with the knowledge-based statistical weight matrix method because a fragment having at most 2 residues lacks two torsion angles, the first Φ angle and the last Ψ angle.

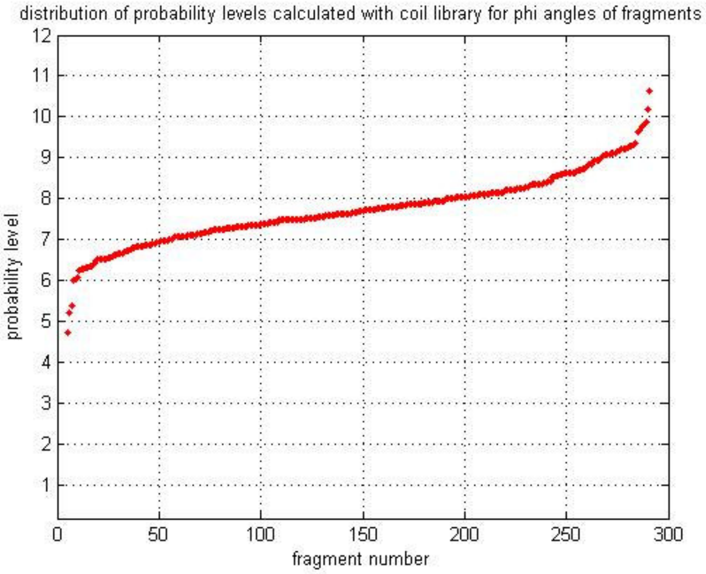


Figure 4.7: Probability level distribution for Φ angles of 300 different fragments calculated with knowledge-based statistics coming from coil library.

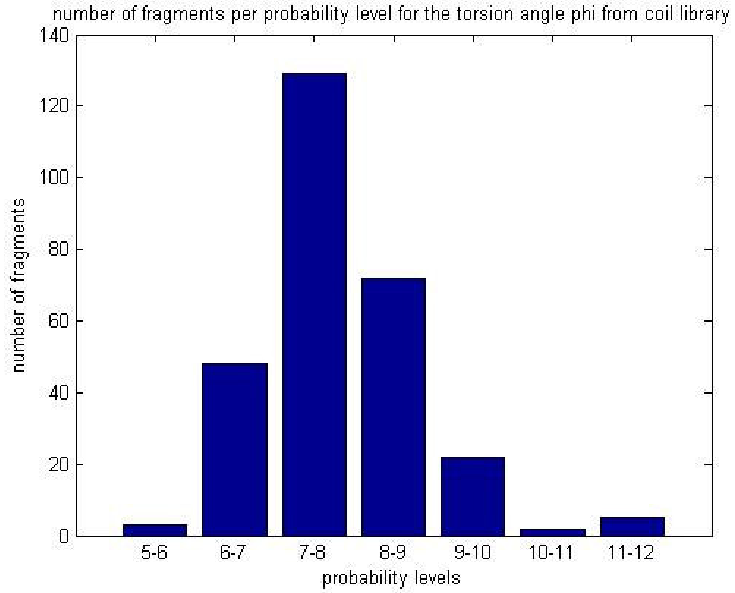


Fig 4.8: Probability level vs. number of fragments from coil library for the torsion angle Φ.

Compared to the prediction results of Φ angles from the full library, the average values of probability levels for the prediction of torsion angle Φ extracted from coil libraries show a similar distribution in means of upper and lower boundaries. However, the points are more scattered, the results deviate more with the average values around 7.7 out of 12 which means 64% accuracy in predictions. For torsion angle Φ, it won't be wrong to make the conclusion that using coil libraries won't boost up or lessen the performance of the prediction method.

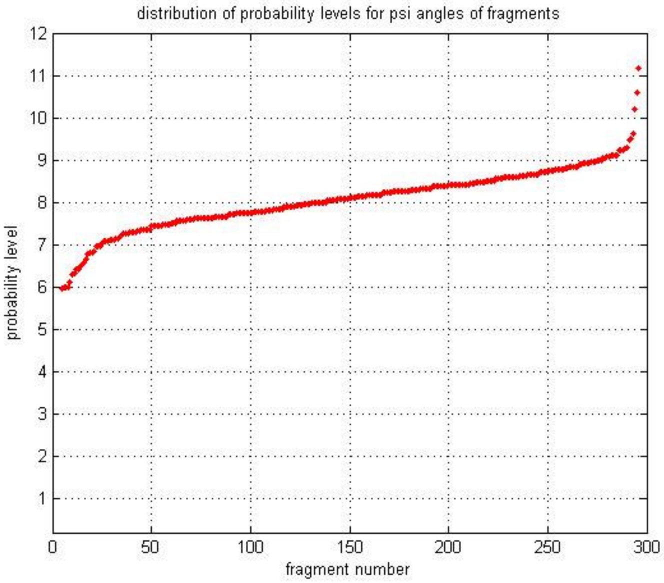


Figure 4.9: Probability level distribution for Ψ angles of 300 different fragments calculated with knowledge-based statistics coming from full library.

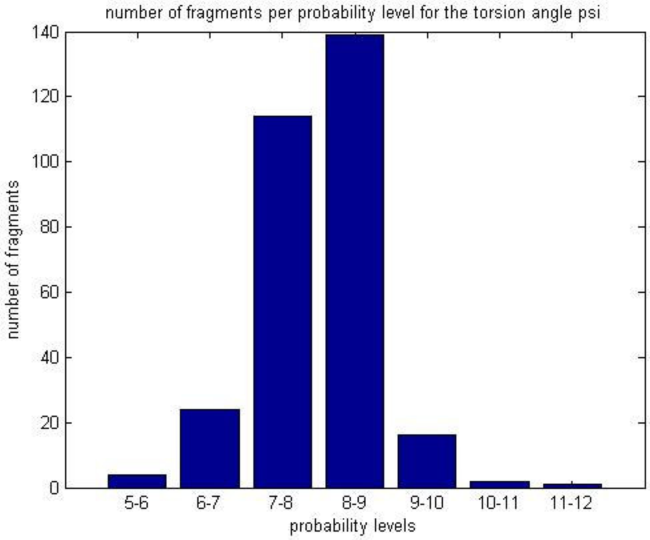


Fig 4.10: Probability level vs. number of fragments from full library for the torsion angle Ψ

As can be seen from the graph above, the score of predictions for the Ψ angle deviate between 6 and 10 mainly around 8 and 9 for all library fragment sequences. The average value of probability levels for Ψ angles calculated with the full library statistics is around 8 out of 12 and it is around 67% accuracy.

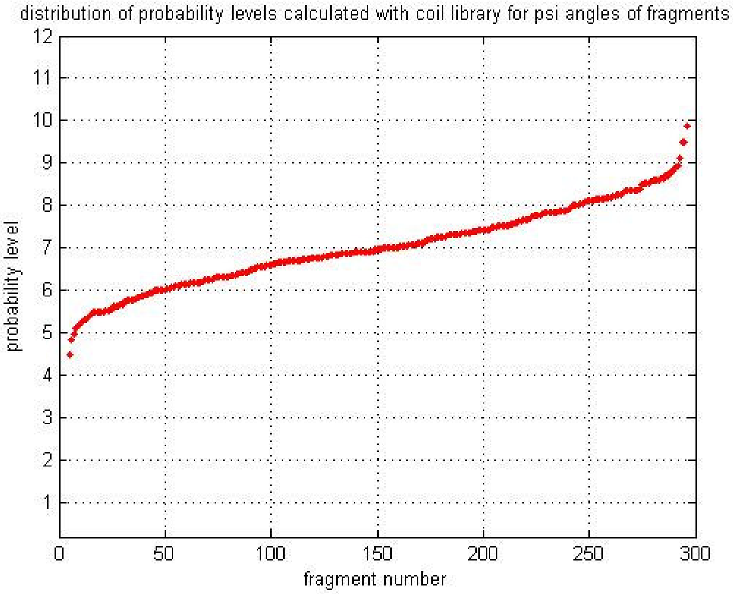


Figure 4.11: Probability level means for Ψ angles of 300 different fragments calculated with knowledge-based statistics coming from coil library.

When the graph of prediction scores for Ψ angles from coil libraries are examined, it can be seen that the results deviate between 5 and 10 mainly between 6 and 8 having an average value around 6.9 which means 57.5% accuracy, and we can conclude that these results are poorer compared to the predictions done with data from all library.

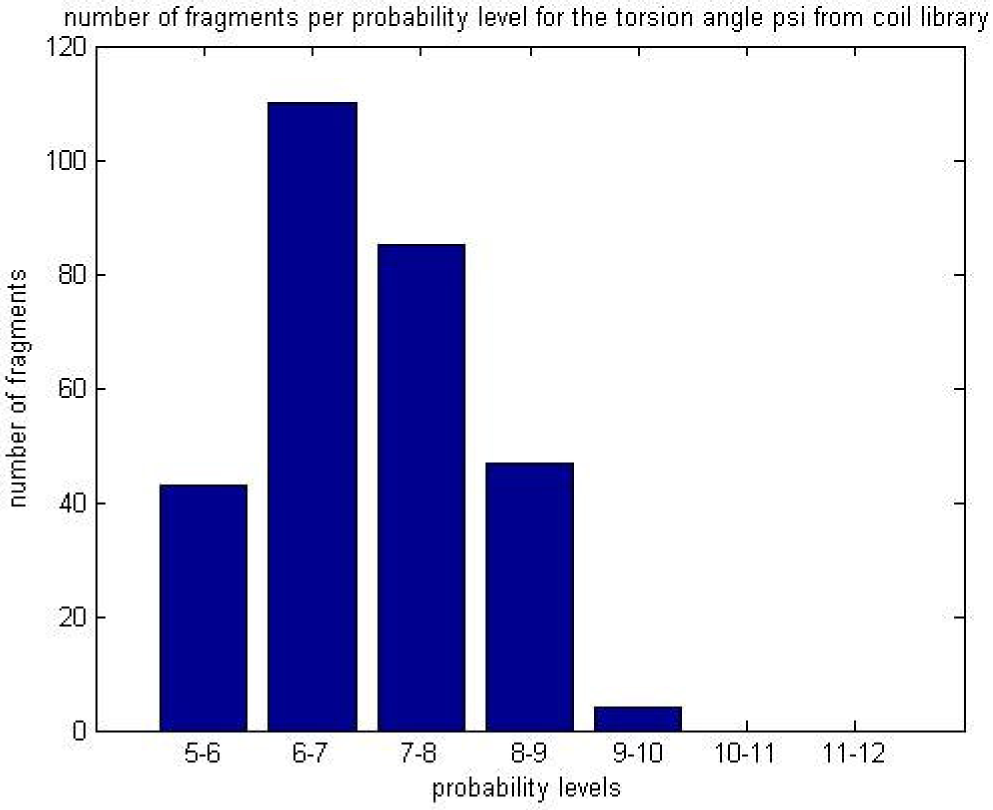


Fig 4.12: Probability level vs. frequency from full library for the torsion angle Ψ.

When we looked at the doublets' preferences for certain states in the Ramachandran plot, we saw that these doublets didn't always occupy the regions defining a certain secondary structure all the time. This result led to the consideration of chameleon sequences that occupy different secondary structures in different proteins of PDB. Our study went into a second part that finding probability of a given sequence to be both in an alpha helix and beta sheet became one of the goals.

4.2. Results for the second part of the study:

The distribution of single residues lay at the basis of this part of the study. After creating a database of chameleon sequences from non-redundant PDB database, the probabilities of these chameleon sequences to obtain each of the three secondary structure types were evaluated with the method explained in the chapter Methods and Models.

Each residue in each chameleon sequence in the database has two 3-dimensional probability vectors, P and P'; P having the actual probabilities for alpha helix, beta sheet and coil region extracted from non-redundant PDB, on the other hand P' has probabilities predicted with the knowledge-based approach used throughout this study. Therefore each amino acid has these vector pairs as many as the times it was encountered in the chameleon database. For instance, the amino acid C was seen 388 times in all chameleon sequences of all fragment lengths. Therefore amino acid C has 388 pairs of these 3-dimensional vectors with alpha helix probability on the x-axis, beta sheet probability on the y-axis and coil region probability on the z-axis. For each amino acid, overall alpha helix, beta sheet and coil region prediction scores in other words the accuracy of prediction method for the secondary structures were calculated by using correlation calculation explained in Chapter 3.

Secondly, according to fragment lengths ranging between 4 and 8, each amino acid's distance were extracted from the calculated distances and what was left was 20 vectors for 20 amino acids with differing lengths. The root mean square of the values in each of these 20 vectors was calculated to get a single score for each amino acid in five different fragment lengths.

Correlation values closer to 1 means a good prediction while those closer to 0 means poor prediction results. For fragment length 4, the predictions are slightly better than the ones for the fragment length 5 because 4-residue long fragments were more frequently seen than 5 residue long fragments and therefore a denser knowledge-base was used to predict secondary structure preferences of chameleon sequences of length 4. More statistical data led to higher correlation results for amino acids. Fragments of length 6 and longer are predicted better than fragments of length 4 and 5. In addition to that, fragments with higher frequencies are predicted with a higher accuracy. This is expected because logically, if the number of secondary structure elements for each fragment grows, a better distribution of these secondary structures is obtained and a better estimation can be made. Since the frequency of fragments is important for prediction, the coil regions were predicted with a higher accuracy than the helical and strand regions. The reason of this higher accuracy is the higher frequency of these fragments. These sequences that have coil regions as secondary structure are seen with more frequently and therefore are better for making estimations and predictions.

The predictions made for chameleon sequences of length 8 show very good results because these sequences are mainly coil regions and as explained above coil regions can be predicted better because of having broader statistical data.

Figure 4.13 was created by with the method summarized with the equation Eq.3.13. As it can be seen from Figure 4.13, except proline, which is a secondary structure breaking amino acid, all amino acids have correlation values above 0.5. Also correlation values for each secondary structure were calculated, however this time with a different method. To find how well the prediction method works for each secondary structure, alpha helix, beta sheet and coil region, actual and predicted results for each amino acid were kept separately. Hence each amino acid has a pair of vectors of different lengths, one having the actual probabilities and the other one having prediction results. In other words, each amino acid has two sets of probability results for three secondary structures separately.

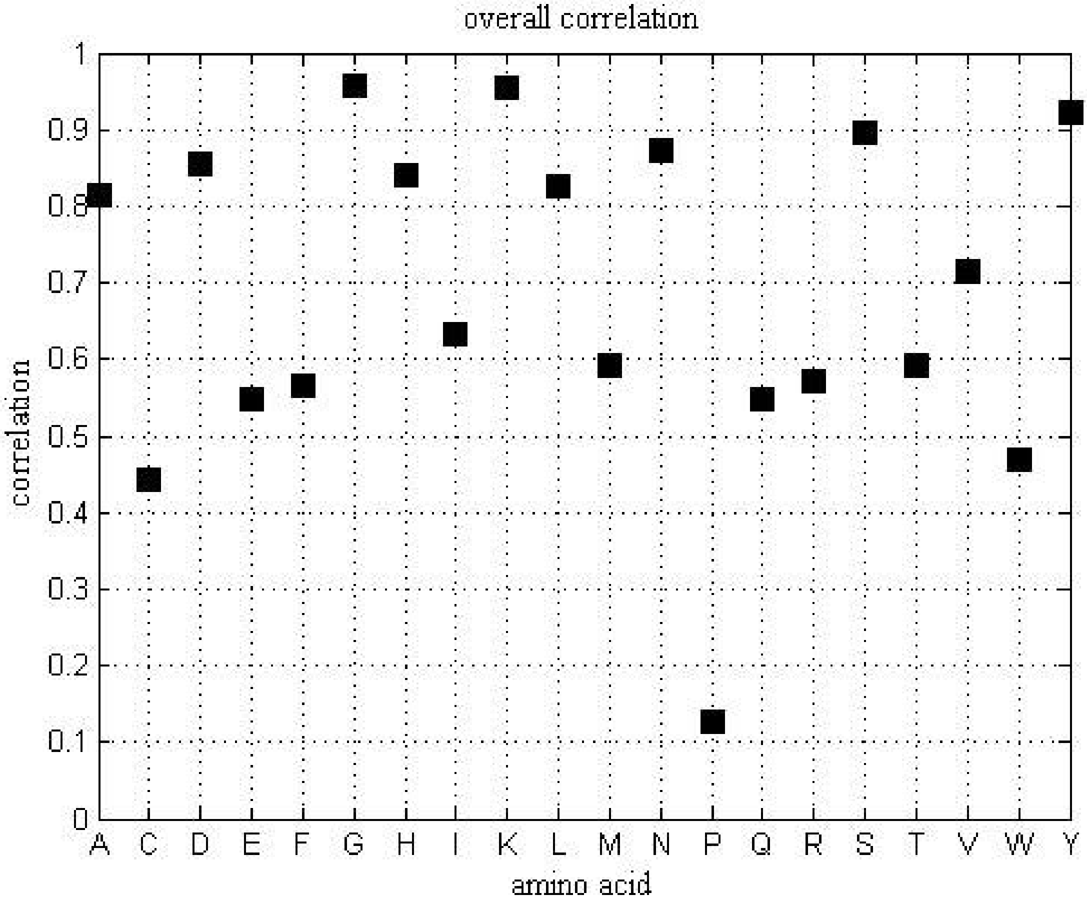


Figure 4.13: Graphs of correlation between the actual probabilities and predicted probabilities for all secondary structures for the chameleon sequences of all lengths combined.

These vectors have the probabilities of each residue whenever that specific residue is encountered in a specific sequence in the library. For instance, if the non-redundant library has n ALA residues, both probability vectors for ALA will have length n with elements having alpha helix probability, beta sheet probability and coil probability ([P(alpha), P(beta), P(coil)]). For the next step, the correlations between actual and predicted results of each amino acid for each secondary structure were calculated. From these correlation results, Figures 4.14(a), 4.14(b) and 4.14(c) have been created. Figure 4.14 shows the graphs of correlation between the actual probabilities and predicted probabilities for (a) beta sheets (b) alpha helices and (c) coil regions for the chameleon sequences of all lengths combined.

If the Graphs 4.14(a) and 4.14(b) are compared, it can be seen that the method works slightly better while assigning alpha helix structure. If a threshold of 60% accuracy is assigned, the overall performance of the method in predicting a specific amino acid is summarized below in Table 4.1.

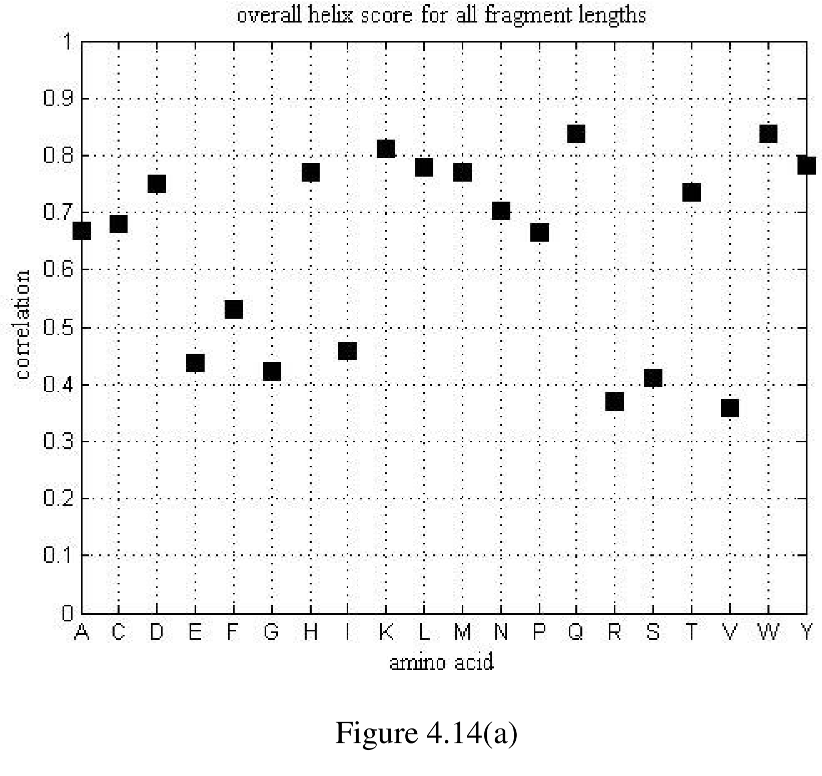
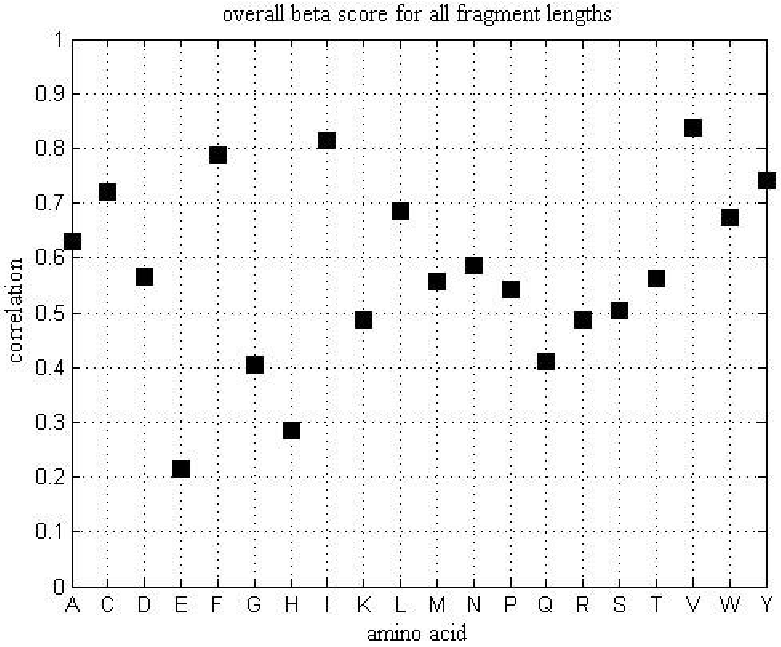


Figure 4.14(a)



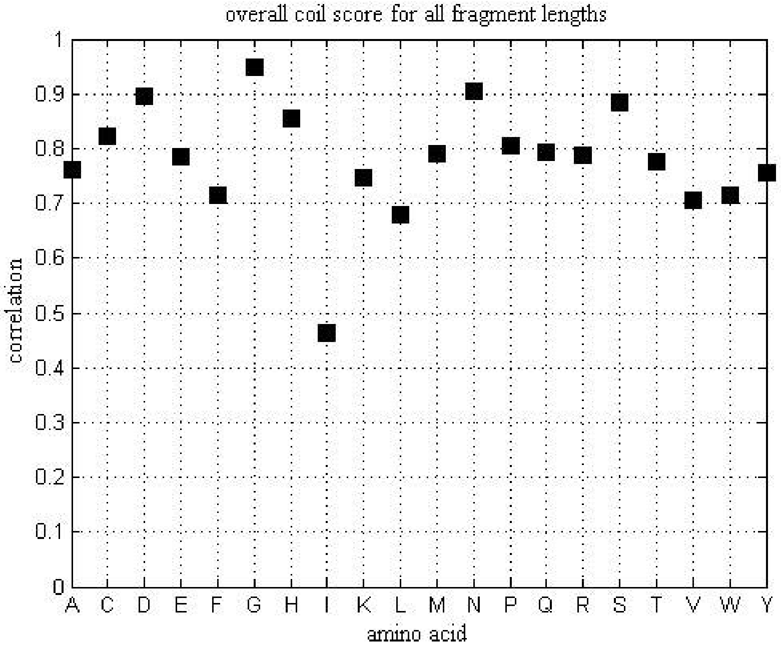


Figure 4.14(b) and Figure 4.14(c)

Alpha Helix	A, C, D, H, K, L, M, N, P, Q, T, W, Y
Beta Sheet	A, C, F, I, L, V, W, Y
Coil Region	All amino acids except I

Table 4.1: calculated secondary structure propensities for amino acids with the knowledge based method

Helix-favoring	M, A, L, E, Q, K
Beta-favoring	T, I, V, F, Y, W
Coil-favoring	G, S, P, N, D

Table 4.2: Amino acids' secondary structure propensities

If Tables 4.1 and 4.2 are to be compared, it can be concluded that the method results are greatly in accordance with the amino acids' intrinsic propensities. However, making the comparison only based on secondary propensities of individual amino acids leads to overlooking the context effect and the influence of neighboring residues on the torsion angles of a residue For instance, a given protein might have a glycine at a given position, which by itself might suggest a random coil there. However, neighboring residue and context effects, might reveal that helix-favoring amino acids occur at that position. Taken together, these factors would suggest that the glycine of the original protein adopts α-helical structure, rather than random coil.

In order to measure how well the prediction method for chameleon sequences works, examples of sequence fragments of length five are chosen from PDB. The first part of these sequences shown in the first column of Table 4.3(a) are originally pieces of alpha helices, while the second part sequence fragments from Table 4.3(b) are beta sheets. The probabilities of all these sequences to be both alpha helices and beta sheets are calculated seperately and the results can be seen in the second and third columns of the table. For the last part for the performance validation of the method, we defined the term "reliability" which is simply the multiplication of alpha helix or beta sheet correlation scores from Figures 4.14(a) and 4.14(b) of each amino acid in the sequence.

S being a sequence fragment from table 4.13 and a_i being an amino acid type, the fragmen S can be identified as S=$a_1a_2a_3\ldots a_n$.

If reliability for sequences and correlation values for amino acids are defined as follows: $[R(S)]_{helix}$=helix reliability fo fragments S, $[R(S)]_{beta}$=beta reliability for fragment S, $[C(a)]_{helix}$=helix correlation for amino acid a, and $[C(a)]_{beta}$=beta correlation for amino acid a,

then the alpha helix and beta sheet reliability values can be formulized in the following manner:

$$[R(S)]_{helix} = [C(a_1)]_{helix} \times [C(a_2)]_{helix} \times ... \times [C(a_n)]_{helix} \qquad \text{Eq. (4.1)}$$

$$[R(S)]_{beta} = [C(a_1)]_{beta} \times [C(a_2)]_{beta} \times ... \times [C(a_n)]_{beta} \qquad \text{Eq (4.2)}$$

The method simply calculates the probabilities of sequences to be a part of an alpha helix or beta sheet and assigns the secondary sturucture with the higher probability to the sequence. From the tables 4.3(a) and 4.3(b), it can be seen that the method clearly and correctly identifies helical and stranded sequences and the a higher reliability of the correctly assigned secondary structure for the most of the time.

helix sequences				
sequence	helix probability	beta probability	helix reliability	beta reliability
AWAAA	1.69E-06	1.19E-07	0.16576564	0.107078376
RKLER	1.55E-06	3.62E-07	0.03756741	0.017042199
HALHY	1.57E-07	8.60E-08	0.242302364	0.025956696
EAEMK	1.50E-06	9.31E-08	0.07976025	0.008000494
LTELK	3.17E-06	1.38E-06	0.158336461	0.027893217
LVDLG	1.91E-06	1.67E-06	0.068748603	0.090718001
WSEAE	4.48E-07	1.22E-07	0.043979611	0.010047648
LREAT	2.71E-06	5.74E-07	0.061560033	0.025549068
TFRHA	2.56E-07	1.62E-07	0.073709029	0.038519953
LCMLA	3.96E-07	7.60E-08	0.212492115	0.119329657
PQELE	1.54E-06	4.39E-07	0.083149935	0.00714563
DEELA	5.28E-06	6.79E-07	0.074569516	0.011431955
KTTLS	1.26E-06	1.16E-06	0.139945969	0.053372469
KPTVK	2.89E-07	1.54E-06	0.114565901	0.060530771
PKVAA	1.48E-06	1.17E-06	0.085645964	0.088143653
VHTLL	5.66E-07	9.68E-07	0.12323957	0.063167443
KKELI	3.34E-06	8.19E-07	0.102632569	0.028713301
NEELL	3.23E-06	5.92E-07	0.08195343	0.01293025
MEDYL	5.07E-07	1.18E-07	0.154147484	0.034551942
YQRYL	4.02E-07	1.37E-07	0.147308922	0.075206134
EEEIN	2.30E-06	5.55E-07	0.026985732	0.004828845
ADKAR	2.18E-06	4.61E-07	0.099466269	0.05322939
DTINT	4.20E-07	4.29E-07	0.130015758	0.085246108
CEDFL	4.19E-07	1.90E-07	0.091913416	0.04744563
KTWRM	8.54E-08	5.45E-08	0.141730785	0.049889113
PHKYR	1.44E-07	7.96E-08	0.119839482	0.026959865
DWVTE	2.27E-07	1.99E-07	0.072101183	0.038744057
HQAKF	2.58E-07	7.57E-08	0.18498031	0.02832549
DLNRK	8.87E-07	2.97E-07	0.122783662	0.053848683
TQLLD	1.49E-06	7.18E-07	0.280211246	0.061562465
KAAET	3.11E-06	4.60E-07	0.115707569	0.023571634
PYEYE	2.20E-07	1.87E-07	0.077826366	0.013856701

Table 4.3(a): List of helical sequences from PDB and their calculated alpha helix and beta sheet probabilities and corresponding reliabilities of these probabilities

beta sequences				
sequence	helix probability	beta probability	helix reliability	beta reliability
RLKIY	7.67E-07	9.22E-07	0.083439538	0.098012094
MWQLY	2.72E-08	2.86E-08	0.330416475	0.078566731
DIEVG	8.60E-07	1.32E-06	0.022637629	0.033878942
WISLD	3.01E-07	2.86E-07	0.092186565	0.107703117
TGFIT	3.34E-07	8.79E-07	0.055108456	0.082206144
RILYS	3.88E-07	5.17E-07	0.042314768	0.101693219
NLFEV	6.45E-07	7.32E-07	0.045461101	0.05747956
EVQWS	1.93E-07	2.32E-07	0.045155807	0.025415839
VAVVA	2.15E-06	2.34E-06	0.020325913	0.234938874
RVIIT	3.43E-07	1.23E-06	0.020284391	0.151878524
AIVCN	1.82E-07	2.91E-07	0.052180519	0.182264236
TIYIN	2.07E-07	5.34E-07	0.084705537	0.162167835
VVDIV	6.64E-07	3.00E-06	0.015678905	0.271277014
FKVYG	1.38E-07	8.51E-07	0.050620381	0.096686542
FEFIN	2.88E-07	3.45E-07	0.039520414	0.064010138
KITFT	2.70E-07	1.71E-06	0.105848039	0.09858184
NRTVP	9.29E-08	4.24E-07	0.045243922	0.072673264
NLYTA	4.69E-07	5.15E-07	0.210091233	0.105846697
SFVLK	1.19E-06	1.21E-06	0.049179141	0.111592623
VWATF	1.02E-07	2.76E-07	0.077683196	0.157714147
FYVCP	5.94E-08	2.63E-07	0.066844379	0.190274008
ITVDN	3.64E-07	1.01E-06	0.06333796	0.127247492
VGWVK	1.90E-07	6.85E-07	0.036649077	0.09382721
LVVNT	6.15E-07	1.18E-06	0.051504685	0.159222269
QVLVR	8.66E-07	9.15E-07	0.03079138	0.096469447
FLGTY	3.20E-07	4.58E-07	0.100020523	0.091294415
TCYLF	8.79E-08	1.30E-07	0.161008885	0.161699332
GEIHP	1.27E-07	7.39E-07	0.043368818	0.011016607
GKILN	5.65E-07	1.41E-06	0.085940188	0.065048198
TPIVF	9.33E-08	8.41E-07	0.042298956	0.163501228
CTFKE	2.12E-07	2.45E-07	0.093574789	0.033488041
TVKRC	2.07E-07	4.58E-07	0.053235367	0.080173174

Table 4.3(a): List of extended sequences from PDB and their calculated alpha helix and beta sheet probabilities and corresponding reliabilities of these probabilities.

Chapter 5

CONCLUSION

Understanding the path of protein folding has been one of the most crucial points in both experimental and computational biological sciences. Numerous studies investigating the effect of short-range and long-range interactions in proteins have been published by both experimental and theoretical researchers. The motivation adopted throughout this study is that comprehending the preferences of torsion angles of each amino acid in a protein in the denatured state bears the key to discover the three dimensional structure of this protein and the way it folds.

Just as with the native state, the structure of this biologically important denatured state appears to depend on the amino acid sequence. [15] Much of the initial interest in non-native protein conformations concentrated on the information these denatured states can give into the process of protein. Therefore our starting point involves predicting the native state preferences of torsion angle in a protein by using denatured state as well as native state information of each residue of the protein.

The RIS model was used actually for protein chains calculations [47, 61]. Although the RIS model is generally used for polymer chains, the method can be easily adapted to polypeptide chains and protein sequences. Keskin et al. proposed the proper way of representing stochastic weights of a polypeptide chain via knowledge-based potentials. [32] In this study, we derived the stochastic weights from knowledge-based libraries and evaluated them via RIS over the chain. Doublets showed different secondary structural preferences. Due to interdependency of the bonds, these preferences favor the choice of the native state torsion angles for each protein sequence and they are context dependent, determined by the amino acid sequence of the protein. This approach is adopted due to Dill's and coworkers' conclusion that proteins are polymers, therefore theories and models of polymers can be used as starting point for treating proteins. As a consequence, the RIS model used for polymer chains may be applied to protein structures.

In this study, the RIS model was adopted to show the context dependency of amino acids' torsion angle state preferences by using statistical weights of these states derived from knowledge-based pair-wise dependent Φ-Ψ maps from non redundant PDB.

The predictions were calculated separately for Φ and Ψ angles by using both the full library and the coil library. Using two different libraries enabled us to see differences of results obtained from the coil and full libraries. The tests were done on 300 fragments that are 30 residues long. Φ angles were predicted to prefer their native state with accuracy around 62.5% with the full library while using coil library didn't boost up the results, the accuracy obtained using coil library remained around 64%. On the other hand, for the Ψ angle predictions, the full library gave a far better accuracy around 67% than the coil library accuracy that is around 57.5%.

However, the presence of chameleon sequences that reside in different secondary structures in different proteins may explain the reason of poor accuracy levels for coil libraries. The prediction results for more than half of twenty amino acids are above 60% accurate, while the amino acids with poor accuracies are proline which is a secondary structure breaker, and arginine, cysteine, glutamine, glutamic acid, tryptophan and phenylalanine that are bulky amino acids and have varying propensities to be in an alpha helix or beta sheet due to their chemical properties. As the extent of calculated correlations between torsion angle pairs shows, the choice of native state of torsion angles strongly depends on the environment of the residue, in other words, sequence of the protein.

We expect that the discussion of the computational basis of probabilities in this work will serve as a guide in interpreting knowledge-based probabilities. However, several key questions brought up are not answered conclusively and awaits further work. Do the context effects average out in calculating probabilities on sufficiently large databases? If so, do we recover the probabilities for the isolated singlets and pairs? The answers to these two questions are important because if they are both affirmative, then the determination of probabilities from isolated singlets and doublets, a relatively easy task that may be carried out computationally, will allow characterization of conformations of full proteins.

APPENDIX

A.1 The actual state preferences for both torsion angles and 10 examples of amino acid A prediction results for torsion angle phi.

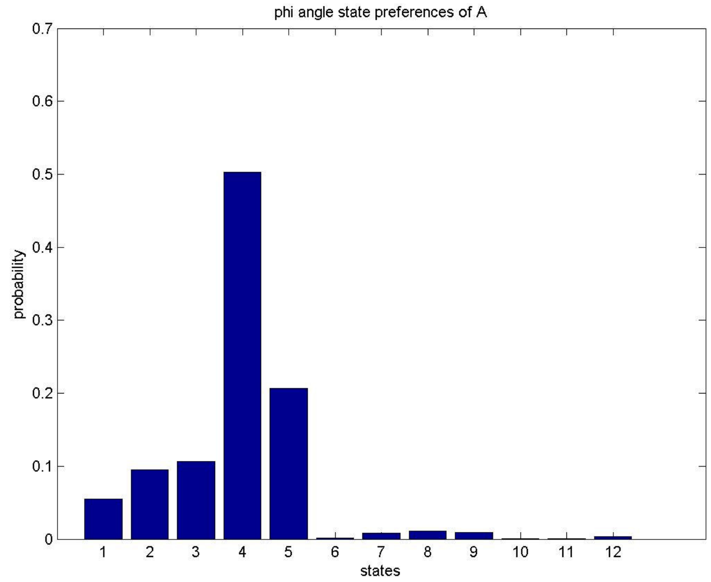


Figure A1.1: Phi angle state preferences of individual amino acid A independent of sequence

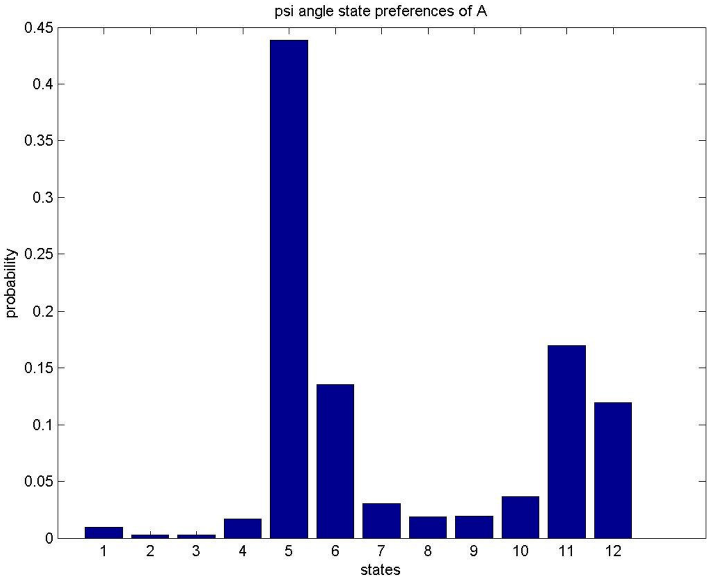
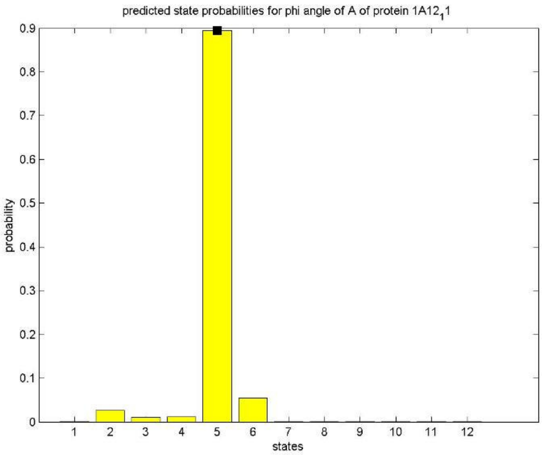
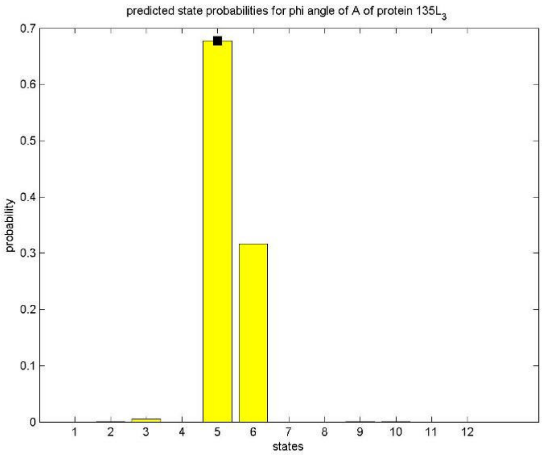


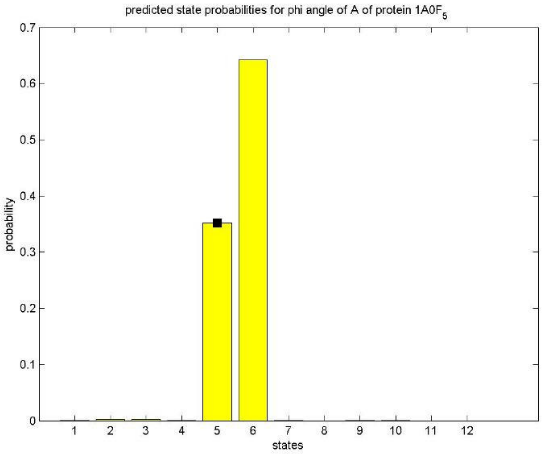
Figure A1.2: Psi angle state preferences of individual amino acid A independent of sequence



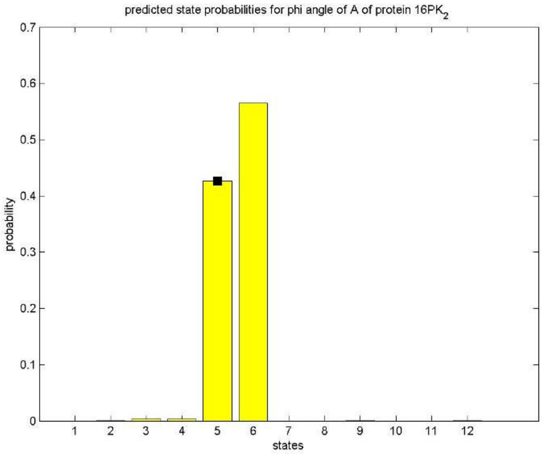
predicted state probabilities for phi angle of A of protein 1A12$_1$1



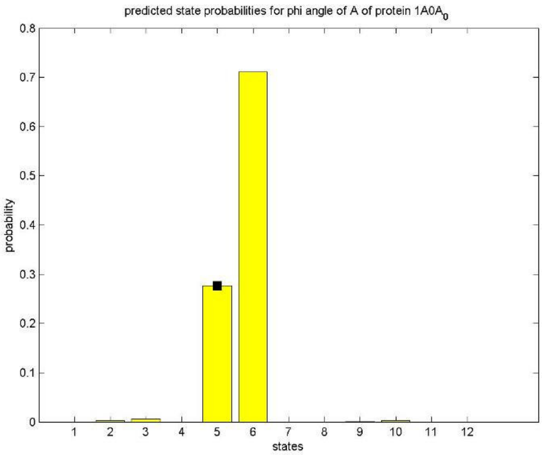
predicted state probabilities for phi angle of A of protein 135L$_3$



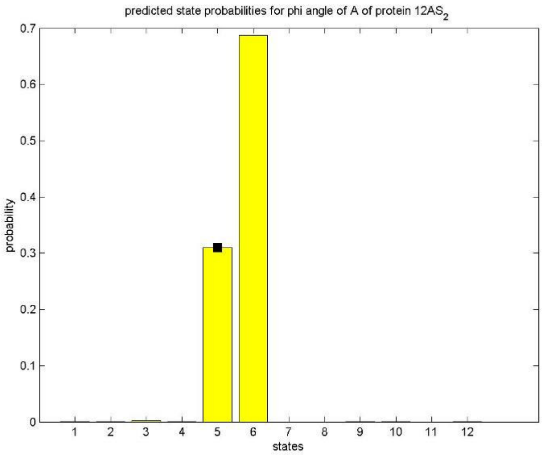
predicted state probabilities for phi angle of A of protein 1A0F$_5$



predicted state probabilities for phi angle of A of protein 16PK$_2$



predicted state probabilities for phi angle of A of protein 1A0A$_0$



predicted state probabilities for phi angle of A of protein 12AS$_2$

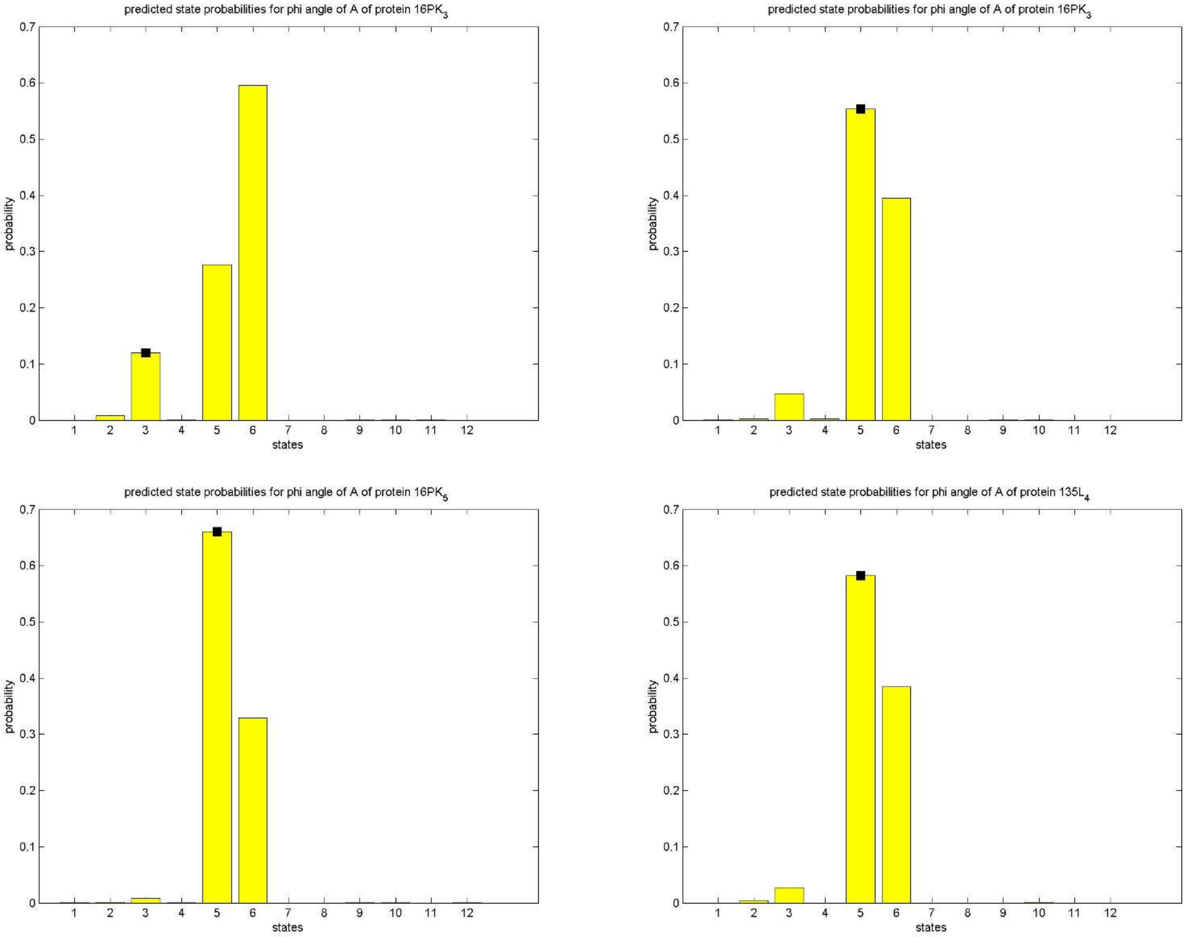


Figure A1.3: 10 examples of phi angle prediction for the amino acid A dependent to the sequence. Black squares in the figures represent native state of that residue in the sequence.

Appendix 2: Actual Φ and Ψ state preferences for all amino acids.

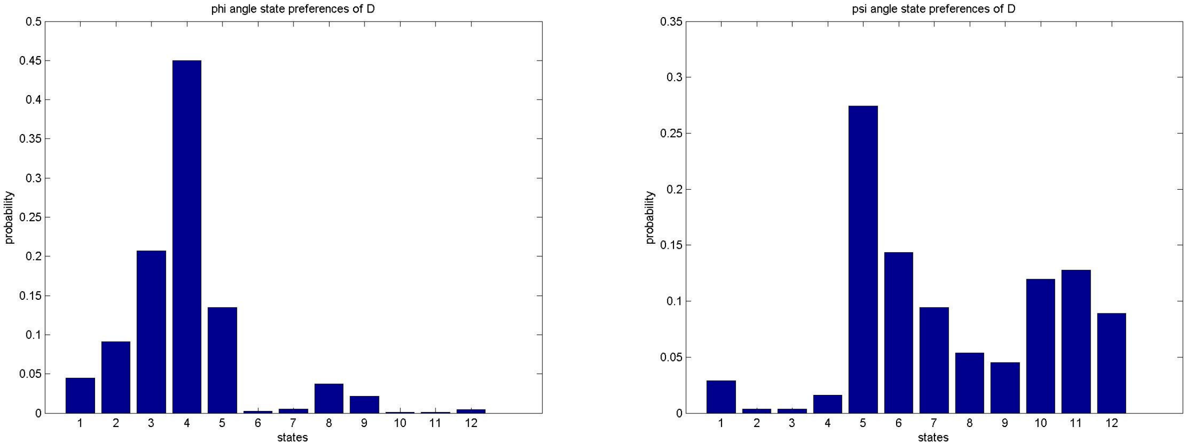


Figure A2.1: Torsion angle state preferences of individual amino acid D independent of sequence

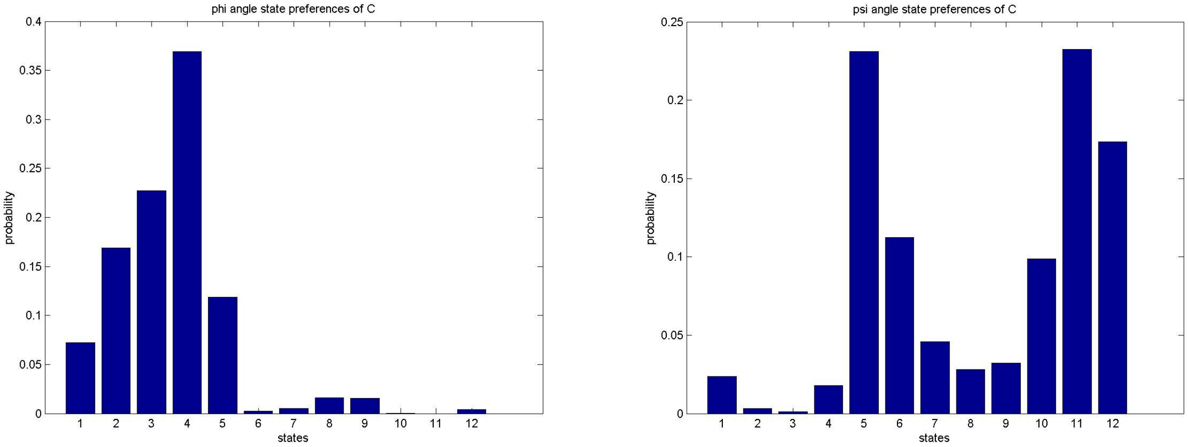


Figure A2.2: Torsion angle state preferences of individual amino acid C independent of sequence

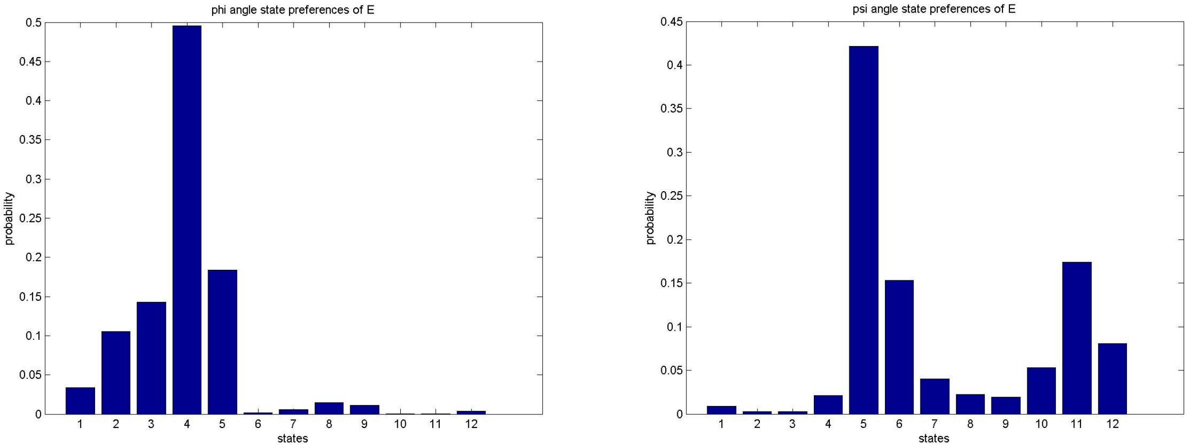


Figure A2.3: Torsion angle state preferences of individual amino acid E independent of sequence

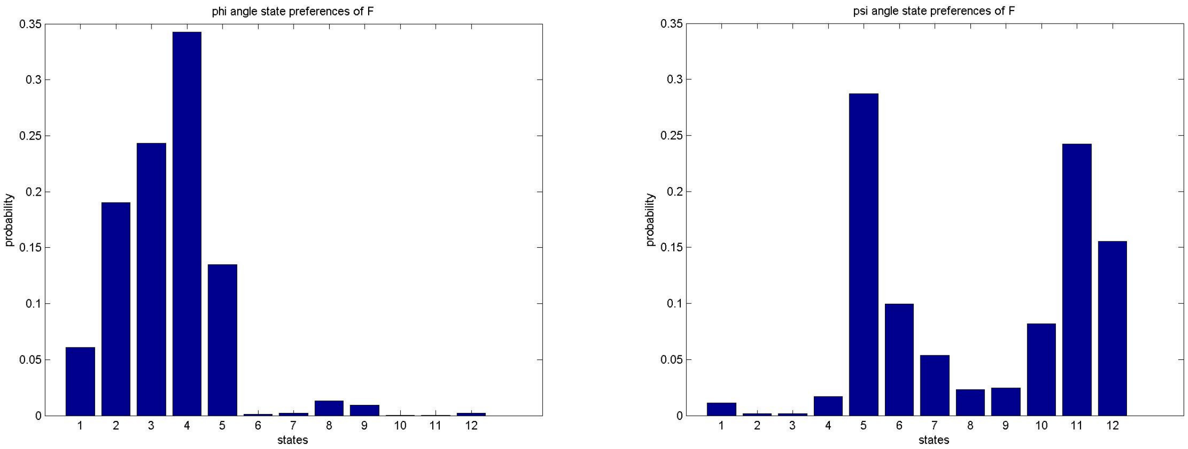


Figure A2.4: Torsion angle state preferences of individual amino acid F independent of sequence

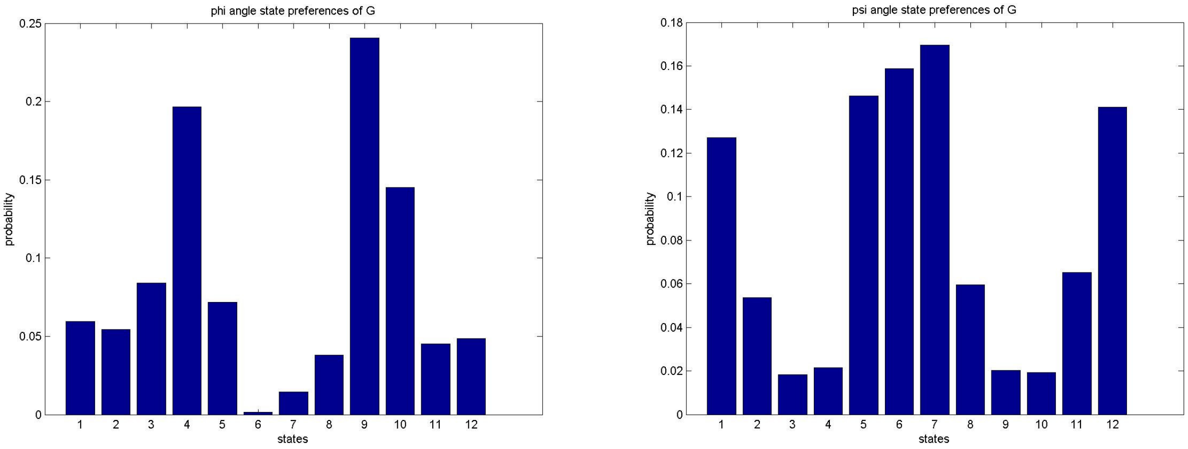


Figure A2.5: Torsion angle state preferences of individual amino acid G independent of sequence

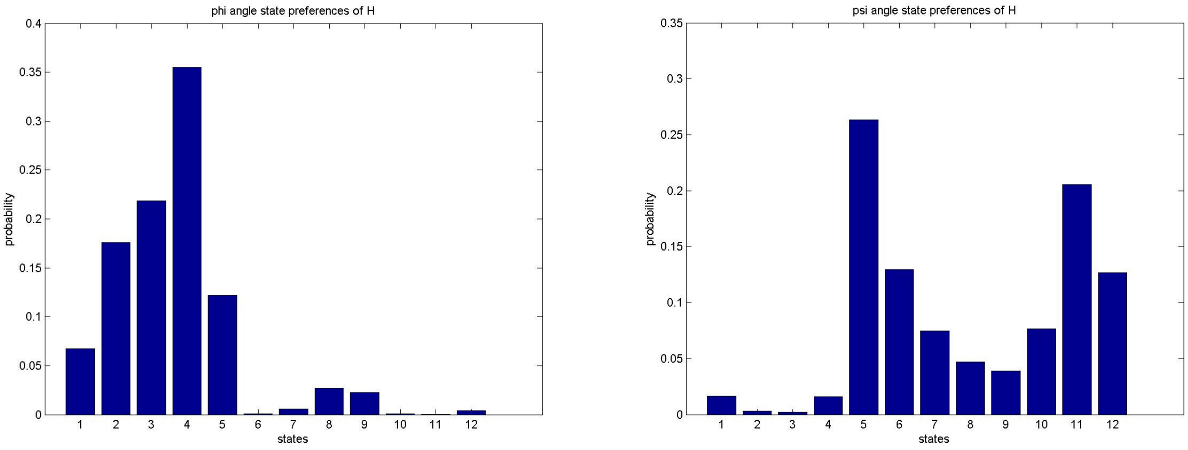


Figure A2.6: Torsion angle state preferences of individual amino acid H independent of sequence

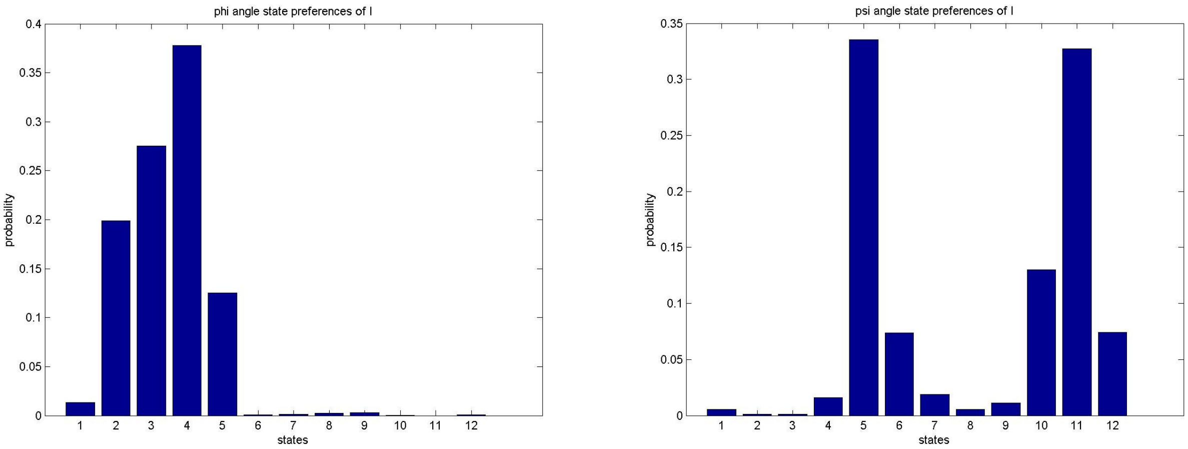


Figure A2.7: Torsion angle state preferences of individual amino acid I independent of sequence

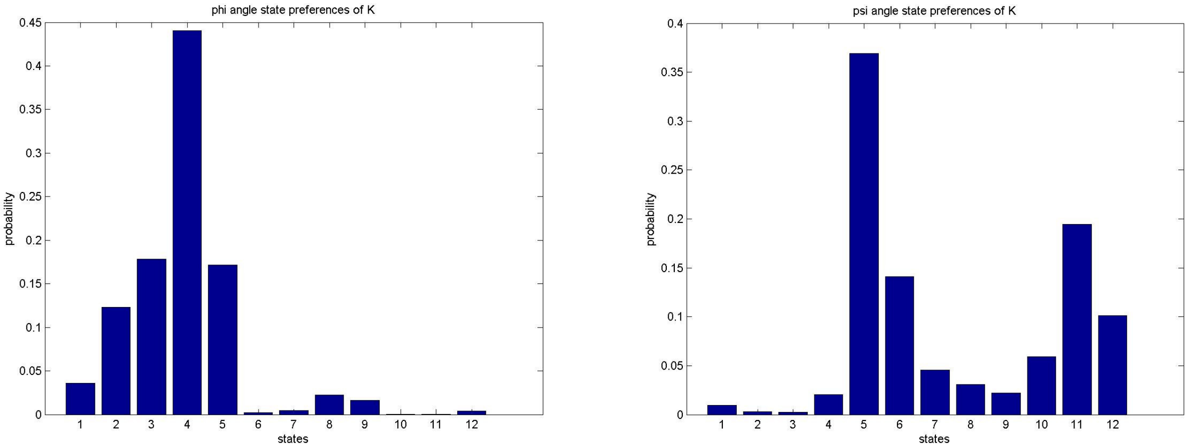


Figure A2.8: Torsion angle state preferences of individual amino acid K independent of sequence

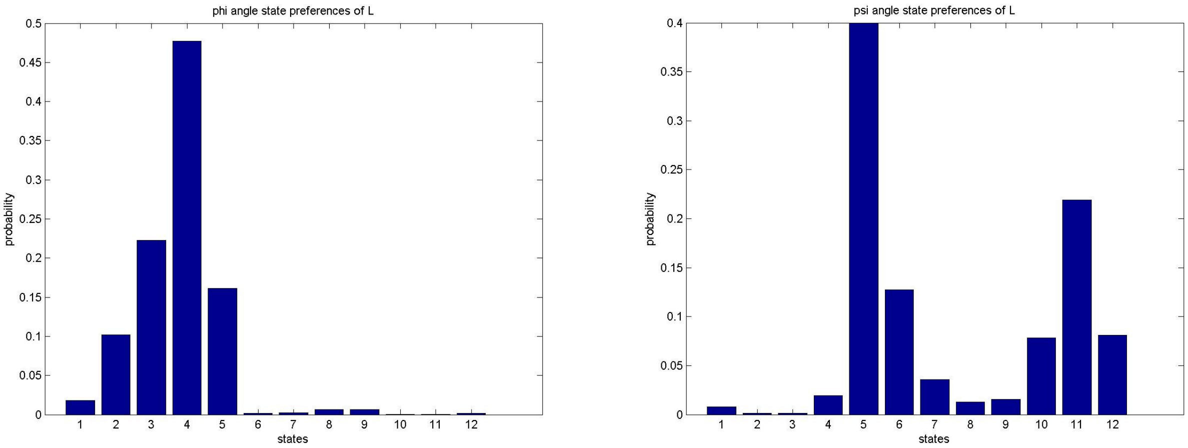


Figure A2.9: Torsion angle state preferences of individual amino acid L independent of sequence

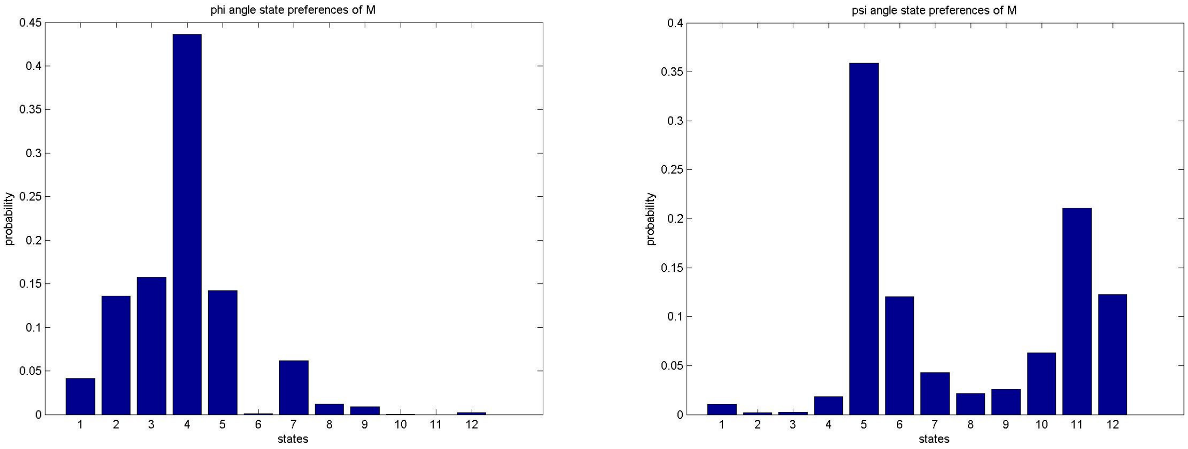


Figure A2.10: Torsion angle state preferences of individual amino acid M independent of sequence

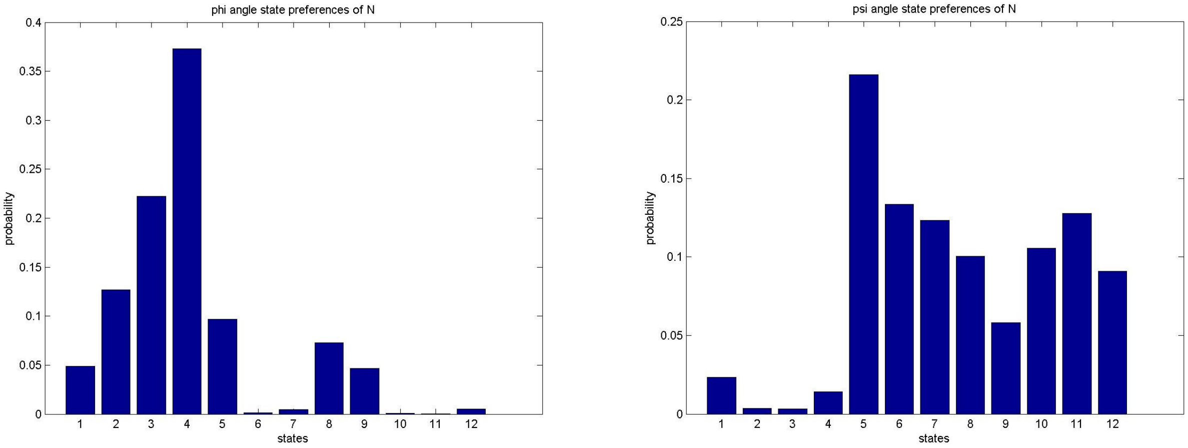


Figure A2.11: Torsion angle state preferences of individual amino acid N independent of sequence

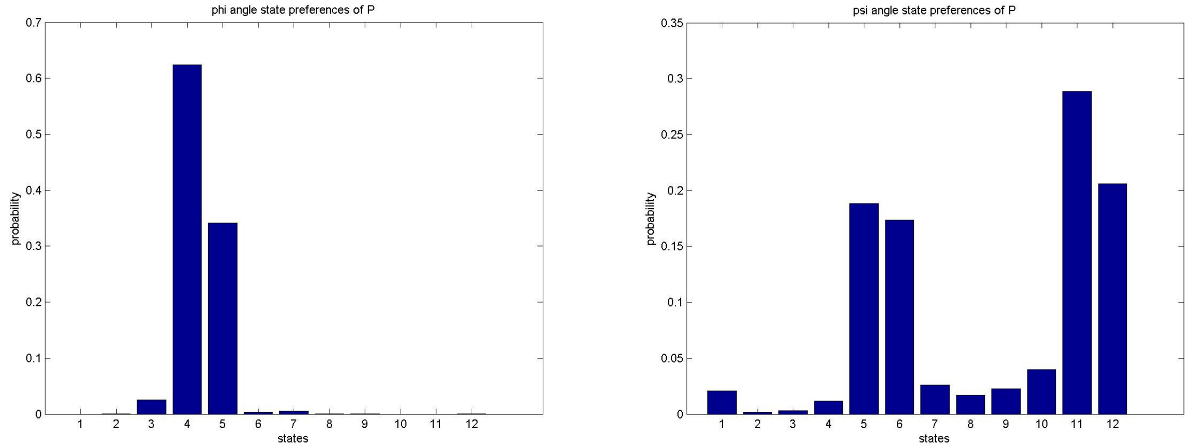


Figure A2.12: Torsion angle state preferences of individual amino acid P independent of sequence

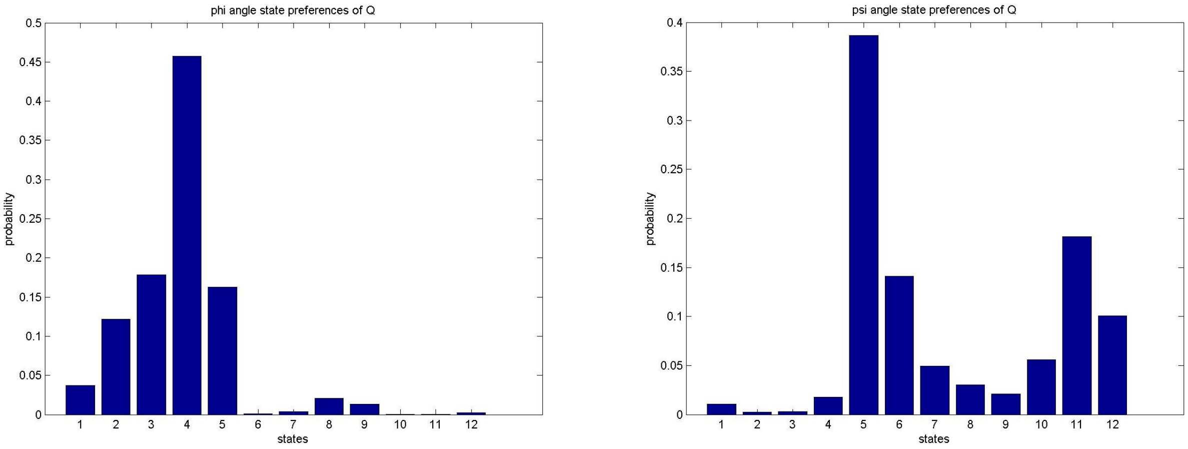


Figure A2.13: Torsion angle state preferences of individual amino acid Q independent of sequence

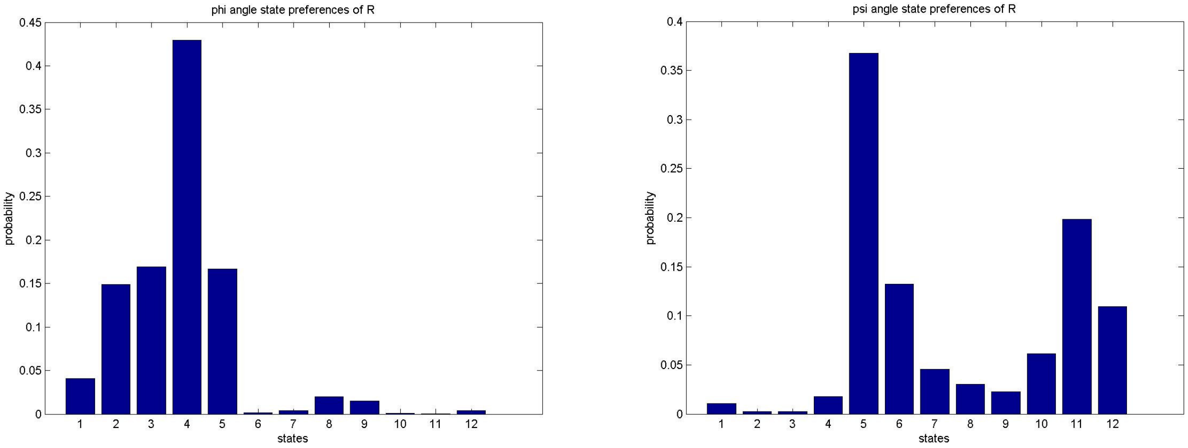


Figure A2.14: Torsion angle state preferences of individual amino acid R independent of sequence

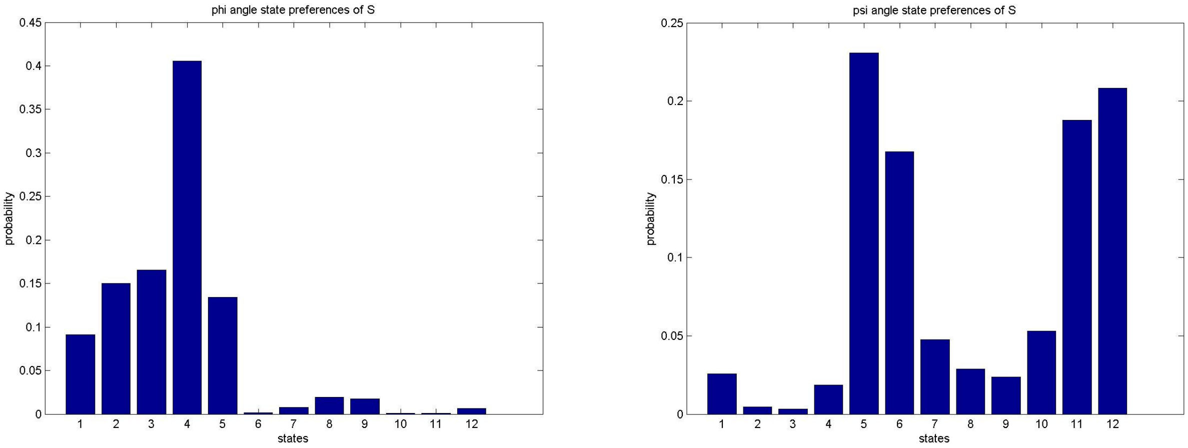


Figure A2.15: Torsion angle state preferences of individual amino acid S independent of sequence

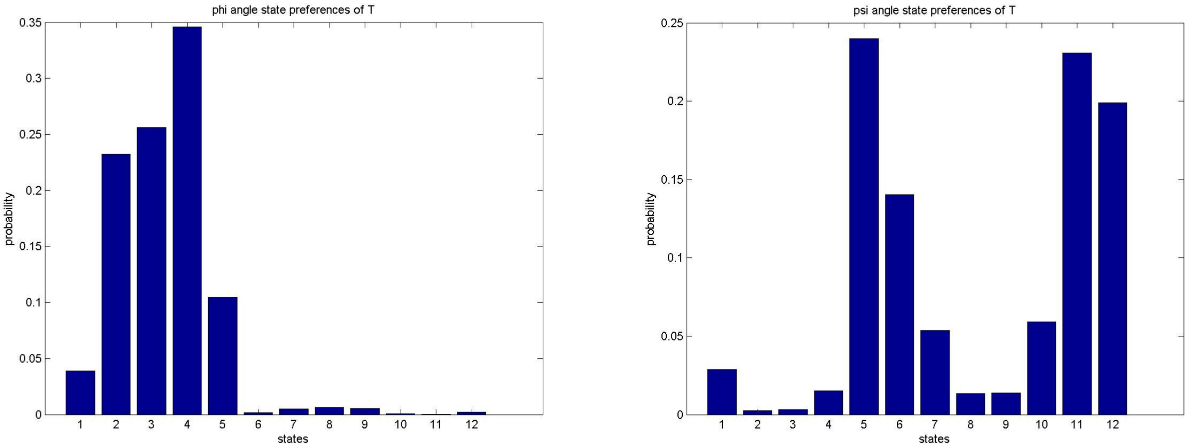


Figure A2.16: Torsion angle state preferences of individual amino acid T independent of sequence

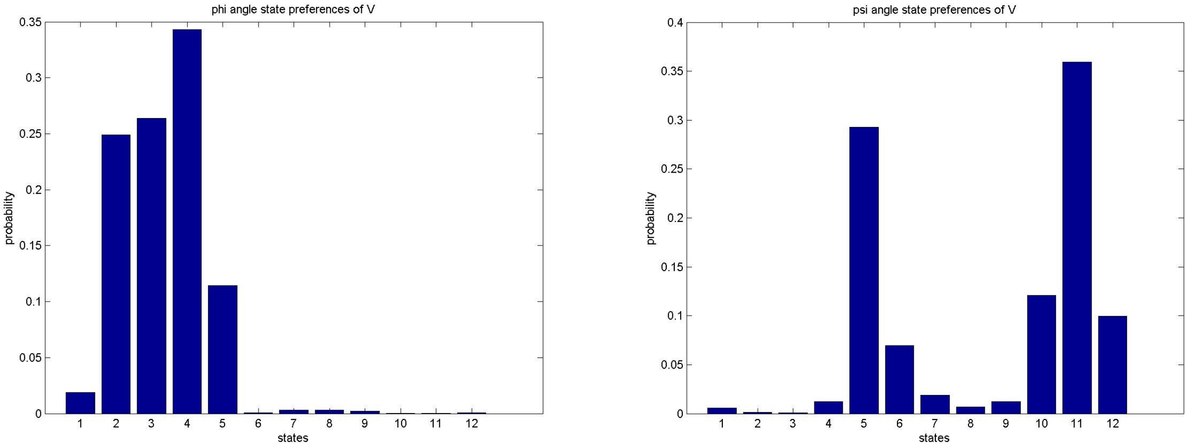


Figure A2.17: Torsion angle state preferences of individual amino acid V independent of sequence

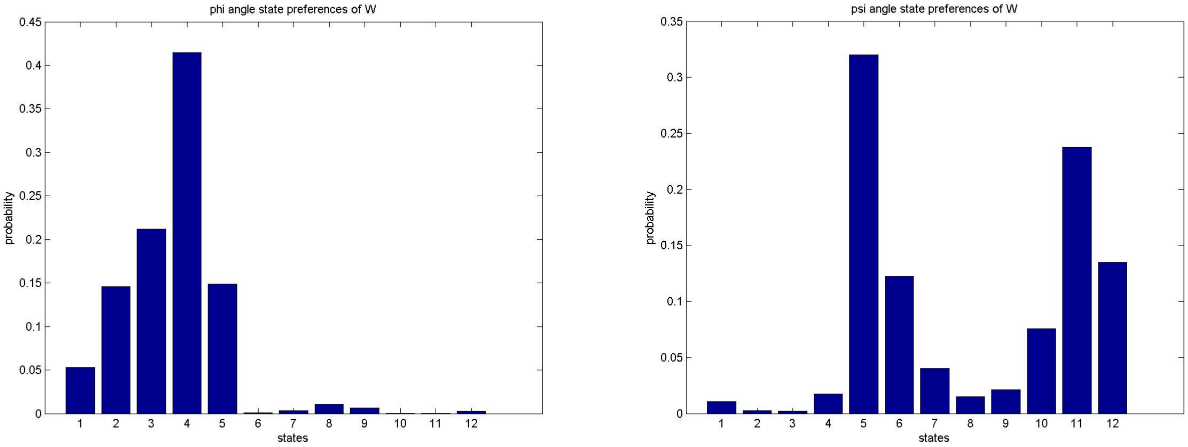


Figure A2.18: Torsion angle state preferences of individual amino acid W independent of sequence

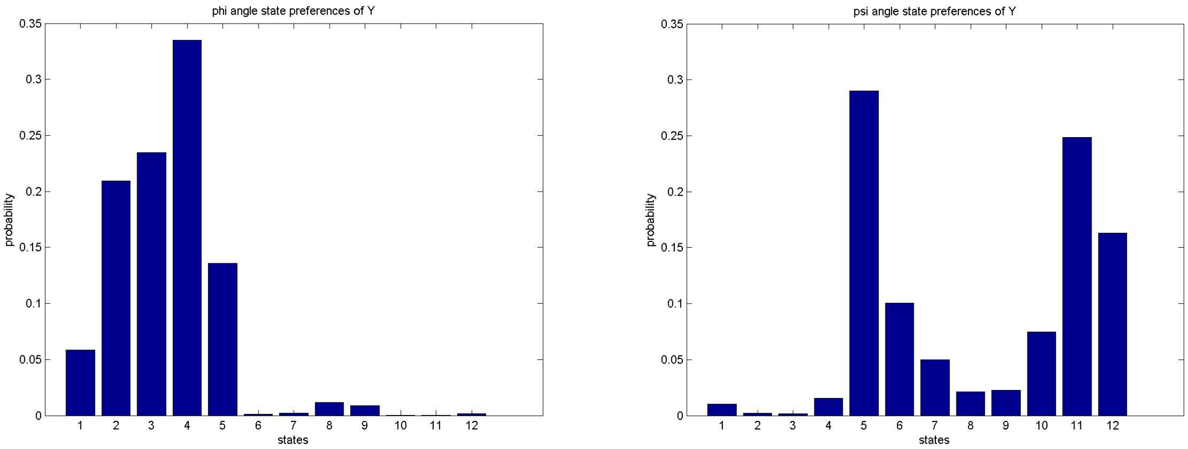


Figure A2.19: Torsion angle state preferences of individual amino acid Y independent of sequence

A.3 The list of structures from non-redundant PDB

12AS	135L	154L	16PK	1A0A	1A0F	1A12	1A1T	1A1W
1A6M	1A6S	1A73	1A76	1A79	1A8H	1A8R	1A8Z	1A92
1ADE	1ADN	1ADO	1ADR	1AE3	1AEP	1AF7	1AF8	1AFJ
1AIW	1AJS	1AK0	1AK1	1AKL	1AKO	1AL3	1ALY	1AMF
1AOZ	1AP0	1AP8	1APQ	1AQ0	1AQ5	1AQT	1ARB	1AS7
1AY2	1AYJ	1AYO	1AYR	1AZP	1AZW	1B0P	1B11	1B2P
1B64	1B69	1B6T	1B74	1B77	1B87	1B8A	1B8T	1B8W
1BCV	1BDC	1BDO	1BDS	1BE1	1BEF	1BEO	1BET	1BFF
1BK5	1BKC	1BL1	1BL8	1BLE	1BM4	1BM8	1BM9	1BMQ
1BQS	1BQV	1BR0	1BRV	1BRZ	1BS2	1BS9	1BSM	1BSY
1BY1	1BY6	1BYK	1BYL	1BYS	1BYY	1BZB	1BZG	1BZK
1C3Y	1C44	1C4Z	1C52	1C5E	1C6W	1C75	1C8P	1C8U
1CDH	1CDR	1CDZ	1CEL	1CEM	1CEU	1CF7	1CFB	1CFM
1CKV	1CKX	1CL4	1CLI	1CLQ	1CM5	1CMC	1CMO	1CN8
1CSH	1CTF	1CTJ	1CTT	1CV8	1CVM	1CVR	1CW0	1CWV
1D2O	1D2R	1D3C	1D4O	1D4V	1D6B	1D6G	1D7B	1D7M
1DDB	1DDF	1DDV	1DDZ	1DE5	1DEA	1DEB	1DEC	1DEO
1DJ0	1DJ7	1DJ8	1DJN	1DKC	1DKQ	1DL0	1DLC	1DLJ
1DPK	1DPM	1DPQ	1DPS	1DPT	1DPU	1DQ3	1DQC	1DQE
1DUJ	1DV5	1DVH	1DVO	1DW0	1DWN	1DXE	1DXG	1DYN
1E4U	1E54	1E5D	1E5K	1E8P	1E8R	1E91	1E9K	1E9M
1ECY	1ED7	1EDG	1EDH	1EDN	1EDX	1EE6	1EE8	1EEJ
1EIJ	1EIW	1EJ0	1EJE	1EJF	1EJJ	1EJP	1EL6	1EM8
1ETE	1ETP	1EUE	1EUV	1EV0	1EW4	1EW6	1EWI	1EWS
1F08	1F0K	1F0Z	1F1Z	1F2D	1F2U	1F2V	1F39	1F3U
1FC3	1FCD	1FCE	1FCF	1FCT	1FCU	1FDM	1FE4	1FE6
1FLC	1FMH	1FN9	1FNF	1FOA	1FOB	1FOF	1FP2	1FP3
1FU9	1FUG	1FUI	1FUO	1FUS	1FV5	1FVL	1FW5	1FW9
1G2R	1G31	1G47	1G5T	1G5V	1G5Z	1G6G	1G6X	1G6Z
1GAB	1GAH	1GAK	1GCB	1GCC	1GCI	1GCN	1GD5	1GD8
1GLN	1GME	1GNC	1GND	1GNH	1GNY	1GO5	1GOF	1GP6
1GUP	1GUR	1GVP	1GWM	1GXC	1GXL	1GXU	1GXY	1GYF
1H5P	1H67	1H6H	1H6Q	1H6W	1H70	1H75	1H7A	1H7D
1HD6	1HDO	1HE1	1HF9	1HG3	1HGH	1HHN	1HHS	1HI9
1HP8	1HP9	1HPG	1HPH	1HQ0	1HQI	1HRD	1HRE	1HS6
1HYP	1HYW	1HZ4	1HZE	1HZM	1I17	1I1J	1I25	1I26
1I8N	1I8T	1IAG	1IAP	1IAZ	1IB8	1IBA	1IBY	1ICA
1IIJ	1IIO	1IJA	1IJC	1IJV	1IJX	1IL6	1ILK	1ILO
1IQ4	1IQO	1IR6	1IRF	1IRS	1IRY	1IRZ	1ISU	1ITH
1IWC	1IWM	1IWO	1IXD	1IXT	1IYC	1IYG	1IYM	1IZN
1J57	1J5S	1J75	1J7L	1J7Q	1J9I	1J9L	1JAJ	1JAU
1JFM	1JFR	1JFX	1JG5	1JGS	1JH5	1JH8	1JHJ	1JI8
1JLI	1JLX	1JLZ	1JMC	1JMU	1JMV	1JO0	1JO6	1JOT
1JW3	1JWE	1JWQ	1JXC	1JYH	1JYO	1JZG	1K0H	1K0S
1K6W	1K81	1K85	1K8H	1K8V	1KA2	1KAF	1KAQ	1KBE
1KJ6	1KJK	1KJS	1KKE	1KKG	1KLO	1KLX	1KMD	1KN0
1KQR	1KS9	1KSA	1KSR	1KTB	1KTG	1KTX	1KU0	1KU7
1L2M	1L2P	1L2Y	1L3P	1L3Y	1L4T	1L5A	1L5J	1L5P

1LAM	1LBA	1LBE	1LBJ	1LBU	1LC3	1LDD	1LFW	1LGH
1LMR	1LMZ	1LN1	1LNS	1LOI	1LPE	1LPL	1LPV	1LQ7
1LU8	1LUA	1LUP	1LV3	1LV4	1LVF	1LVM	1LWD	1LWR
1M3U	1M3V	1M3W	1M3Y	1M42	1M4F	1M4I	1M4J	1M4L
1MBY	1MC0	1MDY	1MEQ	1MFG	1MGS	1MH9	1MHD	1MHL
1MOG	1MOL	1MOT	1MP1	1MP6	1MPM	1MPY	1MQW	1MR0
1MWW	1MWZ	1MXI	1MXM	1MZK	1MZM	1N0Z	1N25	1N26
1N6U	1N6Z	1N7U	1N81	1N87	1N8L	1N8N	1N91	1N9L
1NE5	1NE8	1NE9	1NEI	1NEP	1NEQ	1NEW	1NF9	1NFK
1NKR	1NKS	1NLQ	1NLS	1NLX	1NMT	1NN4	1NNV	1NNW
1NVM	1NW3	1NXI	1NY8	1NY9	1NYB	1NYC	1NYH	1NYO
1OAP	1OCK	1OE4	1OEF	1OEJ	1OF9	1OFG	1OFZ	1OGD
1ON8	1ONE	1ONR	1OO0	1OOH	1OPM	1OQJ	1OR4	1ORD
1OW5	1OWT	1OXJ	1OYG	1OYI	1OZ2	1P0G	1P0J	1P0L
1P90	1P94	1P97	1P9C	1P9E	1P9I	1P9K	1PA4	1PA7
1PF5	1PFJ	1PFK	1PFS	1PFT	1PGY	1PI4	1PII	1PJ5
1POA	1POC	1POH	1POI	1POQ	1POZ	1PP5	1PPN	1PPT
1PV6	1PVZ	1PWM	1PX8	1PYA	1PYS	1PYV	1PZ4	1PZD
1Q38	1Q3J	1Q3K	1Q46	1Q4F	1Q53	1Q56	1Q59	1Q5F
1QC7	1QCX	1QDB	1QDP	1QEX	1QFD	1QFE	1QFT	1QFZ
1QJV	1QK7	1QK9	1QKF	1QKJ	1QKL	1QKS	1QLM	1QLO
1QSA	1QSD	1QSP	1QTS	1QTW	1QU5	1QU6	1QU7	1QUL
1QZN	1QZT	1R02	1R1B	1R21	1R2A	1R2M	1R3N	1R44
1R7J	1R7M	1R8I	1R9L	1RBL	1RCB	1RCF	1RDZ	1REO
1RKI	1RKL	1RKU	1RL1	1RL6	1RLA	1RLJ	1RLW	1RMD
1RSY	1RTH	1RTT	1RU4	1RUW	1RW2	1RW7	1RWJ	1RWR
1S68	1S6D	1S6W	1S79	1S7H	1S7I	1S8K	1S9H	1SA3
1SFT	1SGJ	1SGM	1SGO	1SH8	1SHC	1SHE	1SHI	1SHS
1SMN	1SMZ	1SNL	1SO9	1SOP	1SPK	1SPP	1SQR	1SR2
1STZ	1SU2	1SUI	1SUR	1SUW	1SV6	1SVB	1SVF	1SW5
1T4Y	1T50	1T5J	1T5Q	1T6C	1T6F	1T6S	1T71	1T8H
1TER	1TF4	1TF7	1TG7	1TGQ	1THG	1THJ	1THW	1TIB
1TNS	1TO6	1TOT	1TP6	1TPM	1TQ1	1TQ6	1TSR	1TTA
1TVS	1TWI	1TWU	1TYG	1U0I	1U0S	1U2F	1U55	1U5M
1UDM	1UDN	1UEN	1UEO	1UFB	1UFI	1UFM	1UFW	1UFZ
1UIL	1UJ2	1UJ8	1UJC	1UJL	1UJR	1UJT	1UJX	1UK5
1USC	1USY	1UT1	1UT3	1UTA	1UTE	1UTG	1UTX	1UTY
1V2Y	1V30	1V31	1V32	1V38	1V4R	1V58	1V5A	1V5J
1V73	1V74	1V77	1V85	1V87	1V88	1V8C	1V8H	1V92
1VCT	1VD2	1VD4	1VD5	1VD7	1VD8	1VD9	1VDA	1VDD
1VHR	1VHZ	1VI7	1VID	1VIE	1VIG	1VK5	1VKR	1VLS
1W2L	1W33	1W4X	1W53	1W9C	1W9R	1WAB	1WBA	1WD2
1WFB	1WFD	1WFE	1WFF	1WFI	1WFJ	1WFK	1WFQ	1WFR
1WGR	1WGS	1WGU	1WGW	1WGX	1WGY	1WH0	1WH2	1WH4
1WHU	1WHV	1WHX	1WHY	1WI0	1WI1	1WI3	1WI5	1WI9
1WIR	1WIV	1WIX	1WIZ	1WJ2	1WJ3	1WJ5	1WJ6	1WJB
1WK0	1WK1	1WLG	1WMI	1WN4	1WN8	1WNH	1WO3	1WOQ
1X6M	1X8Z	1X93	1X9B	1X9L	1XAK	1XAU	1XBD	1XBR
1XN8	1XN9	1XNA	1XNB	1XNL	1XO3	1XO4	1XO8	1XO9
1XUT	1XV3	1XVA	1XWE	1XX1	1XX7	1XXO	1Y03	1Y0H
1YD6	1YDU	1YEL	1YEM	1YEW	1YGE	1YGH	1YGM	1YIF

1YTS	1YUA	1YUB	1YUI	1YVC	1YWL	1YWZ	1YX7	1YZS
2AK3	2ALC	2BAA	2BEM	2BGO	2BN2	2BNH	2BOP	2BOS
2END	2ERL	2EZE	2EZI	2EZL	2EZX	2FCB	2FMR	2FUA
2LEF	2LEU	2LFB	2LZM	2MAG	2MAS	2MBR	2MCM	2MHR
2PSP	2PTH	2PTL	2PVB	2QIL	2RGF	2RN2	2SAK	2SAS
3EBX	3ECA	3EIP	3ENG	3EZM	3HSC	3MBP	3MDE	3MSI
5ACN	5EAT	5R1R	5ZNF	6CRO	6FD1	6MHT	6PAX	6RLX
1A1X	1A26	1A2Z	1A34	1A3H	1A44	1A4M	1A5R	1A63
1A93	1AA7	1AAC	1AAZ	1AB3	1ABE	1ABV	1ABZ	1ACA
1AFO	1AFR	1AG2	1AGG	1AGJ	1AH7	1AH9	1AHJ	1AHK
1AML	1AMM	1AMP	1AMX	1AN2	1ANS	1ANU	1AOC	1AOL
1ASH	1AST	1ASU	1AUA	1AUO	1AUU	1AUZ	1AVO	1AVP
1B2V	1B34	1B35	1B3A	1B3T	1B3U	1B4B	1B4R	1B4U
1B94	1B9H	1B9L	1B9P	1B9U	1B9W	1BA5	1BAL	1BB1
1BFM	1BGF	1BGK	1BGL	1BGY	1BH9	1BHE	1BHI	1BHU
1BMT	1BNB	1BO4	1BOE	1BOM	1BOR	1BOY	1BP1	1BP7
1BT5	1BTN	1BU7	1BUO	1BV1	1BVB	1BVQ	1BW3	1BW6
1C01	1C05	1C0F	1C15	1C17	1C1K	1C20	1C25	1C3D
1C8Z	1C94	1C9S	1CA4	1CB6	1CBH	1CC5	1CC8	1CCH
1CFR	1CFZ	1CG2	1CHC	1CHK	1CI6	1CIX	1CJC	1CJW
1CO4	1COF	1COI	1COK	1COO	1COU	1COZ	1CPO	1CPQ
1CWX	1CWY	1CX8	1CZ4	1CZ6	1D0N	1D0Q	1D1H	1D1N
1D7Q	1D8B	1D8J	1D9C	1D9J	1DAB	1DAK	1DAP	1DAT
1DF1	1DFE	1DFN	1DG9	1DGN	1DH3	1DHN	1DHR	1DI2
1DLW	1DMC	1DME	1DMT	1DMU	1DNP	1DNY	1DOC	1DOQ
1DQR	1DQW	1DQZ	1DR9	1DS1	1DSB	1DSQ	1DSX	1DTC
1DZ1	1DZL	1E0B	1E0G	1E0N	1E19	1E1A	1E29	1E2A
1EAF	1EB0	1EB6	1EB9	1EBF	1EBP	1EC5	1ECE	1ECI
1EER	1EF4	1EFV	1EG4	1EG7	1EGF	1EGX	1EH1	1EH2
1EMW	1EMZ	1ENH	1EO1	1EOM	1EQ1	1EQ6	1EQ7	1EQK
1EWW	1EX1	1EX2	1EXG	1EXK	1EXT	1EY1	1EYH	1EYQ
1F3V	1F52	1F53	1F62	1F6V	1F81	1F8Y	1FAD	1FAF
1FEH	1FEW	1FEX	1FEZ	1FGJ	1FHO	1FI2	1FIL	1FIU
1FPO	1FQ0	1FQ1	1FQT	1FQV	1FR3	1FRE	1FRY	1FSZ
1FWK	1FWO	1FWQ	1FX2	1FX8	1FXD	1FY7	1FYC	1FZA
1G7E	1G7O	1G8A	1G8E	1G8F	1G8L	1G8Q	1G92	1G99
1GDT	1GEA	1GEF	1GH8	1GH9	1GHH	1GJS	1GJW	1GK7
1GP8	1GPC	1GPE	1GPR	1GPS	1GQI	1GS5	1GSA	1GT7
1GYH	1GYJ	1GYM	1GYZ	1GZJ	1GZT	1H0X	1H0Z	1H21
1H8C	1H8M	1H8U	1H9F	1HA8	1HBG	1HBW	1HCD	1HCN
1HJ0	1HJR	1HJZ	1HK6	1HKA	1HKQ	1HKY	1HN3	1HN6
1HS7	1HSL	1HST	1HTP	1HTW	1HUF	1HUL	1HUX	1HW1
1I27	1I2U	1I35	1I3J	1I42	1I4U	1I4W	1I5G	1I78
1ICH	1ICI	1ICM	1ID1	1IDA	1IEN	1IFR	1IGD	1IGU
1ILY	1IMJ	1IMT	1IMU	1IN0	1INP	1IO0	1IO1	1IOJ
1ITP	1ITU	1ITW	1ITX	1IU4	1IUF	1IUH	1IUK	1IUR
1J03	1J0F	1J0S	1J0T	1J1T	1J24	1J26	1J27	1J36
1JAY	1JBI	1JC7	1JCF	1JCL	1JDM	1JDW	1JEI	1JEK
1JJD	1JJF	1JJG	1JJO	1JJU	1JK3	1JKG	1JKN	1JKV
1JOV	1JOY	1JPY	1JQE	1JR5	1JRA	1JRJ	1JRM	1JTK
1K12	1K19	1K1G	1K1V	1K1Z	1K24	1K2E	1K2F	1K32

1KBH	1KCM	1KCN	1KCO	1KDL	1KDX	1KFR	1KG1	1KHH
1KN6	1KNC	1KNY	1KO5	1KO6	1KOE	1KOY	1KOZ	1KP6
1KUH	1KUU	1KV4	1KV8	1KVD	1KVN	1KWG	1KWH	1KXL
1L6H	1L6P	1L6R	1L6Z	1L7A	1L7L	1L7Y	1L8D	1L8Y
1LGQ	1LI1	1LIS	1LJ9	1LJO	1LK5	1LKI	1LKT	1LL8
1LQ9	1LQP	1LR1	1LRE	1LRH	1LRR	1LRV	1LSL	1LSS
1LX8	1LXA	1LXL	1LXY	1LY1	1LY7	1LYP	1LZW	1M12
1M4O	1M4R	1M4U	1M55	1M5Z	1M6U	1M7J	1M8Z	1M98
1MIO	1MJC	1MJD	1MK0	1MK4	1MKA	1MKE	1MKF	1MLA
1MRJ	1MRP	1MSL	1MSP	1MSZ	1MT6	1MUG	1MUN	1MVH
1N27	1N2F	1N2S	1N2Z	1N35	1N3J	1N3K	1N45	1N4C
1NAR	1NAW	1NBA	1NBC	1NBF	1NBW	1NC3	1NC5	1NCS
1NG1	1NG6	1NG7	1NGN	1NGR	1NH1	1NIJ	1NIX	1NIY
1NO4	1NOF	1NOX	1NP4	1NPU	1NR3	1NSC	1NSO	1NTC
1NYT	1NZE	1NZP	1O06	1O0P	1O0U	1O1Z	1O4Y	1O7I
1OGQ	1OGS	1OH1	1OIG	1OIL	1OIO	1OJG	1OKC	1OLG
1ORG	1ORJ	1ORO	1OSY	1OTC	1OTF	1OTG	1OTR	1OTW
1P1M	1P2X	1P3C	1P42	1P4Q	1P4T	1P57	1P5F	1P5K
1PB5	1PB6	1PBU	1PC0	1PC6	1PCF	1PCN	1PD3	1PD6
1PJM	1PJN	1PJV	1PJZ	1PKH	1PKP	1PLP	1PLQ	1PMC
1PQS	1PQX	1PRH	1PRN	1PSE	1PSM	1PSY	1PTQ	1PU1
1PZQ	1PZR	1PZW	1Q02	1Q0G	1Q0R	1Q0W	1Q1V	1Q2F
1Q5W	1Q5Z	1Q60	1Q68	1Q6A	1Q7L	1Q7S	1Q8M	1Q8R
1QG3	1QGI	1QGK	1QGM	1QH4	1QH5	1QHD	1QHF	1QHK
1QLW	1QM9	1QNR	1QNX	1QOY	1QP6	1QQ5	1QQV	1QR0
1QVA	1QW9	1QWT	1QXF	1QXM	1QXN	1QXR	1QYC	1QYP
1R48	1R4G	1R57	1R5E	1R5R	1R5S	1R5Z	1R61	1R6R
1RGE	1RGS	1RHZ	1RI5	1RI6	1RI9	1RIF	1RIJ	1RIP
1RMG	1RMK	1ROC	1ROO	1RPB	1RPR	1RPX	1RQ6	1RQJ
1RY9	1RYA	1RYT	1RZS	1S04	1S0P	1S12	1S1D	1S2O
1SAC	1SAY	1SB6	1SBP	1SCU	1SCY	1SDF	1SE9	1SED
1SIS	1SJG	1SJQ	1SJR	1SJW	1SKF	1SKH	1SKZ	1SLC
1SR4	1SR8	1SRA	1SRK	1SRO	1SRS	1SRZ	1SS3	1SSE
1SZA	1SZH	1T06	1T0I	1T0Y	1T16	1T17	1T1H	1T23
1T92	1T9F	1TAF	1TBA	1TBD	1TBG	1TCA	1TCG	1TD6
1TIF	1TIG	1TIT	1TIV	1TJL	1TJY	1TKB	1TKN	1TL2
1TTW	1TU1	1TU9	1TUA	1TUH	1TUL	1TUM	1TUW	1TUZ
1U5T	1U5U	1U7P	1U84	1U8V	1U96	1UAI	1UB9	1UC2
1UG0	1UG1	1UG2	1UG7	1UG8	1UGJ	1UGL	1UHE	1UHM
1UKF	1UKX	1UL5	1ULD	1ULO	1UMH	1UMZ	1UNK	1UOR
1UUN	1UW0	1UW1	1UW2	1UW4	1UWD	1UX5	1UXD	1UYP
1V5M	1V5N	1V5P	1V5R	1V5T	1V61	1V64	1V65	1V66
1V95	1V9V	1V9W	1V9X	1VA1	1VA9	1VAE	1VAV	1VCA
1VDF	1VDL	1VE6	1VEA	1VEE	1VEG	1VEH	1VEK	1VFI
1VMB	1VMO	1VNS	1VPU	1VSG	1VTP	1VYB	1VYI	1VYX
1WD3	1WDJ	1WEO	1WEQ	1WER	1WEU	1WEV	1WEX	1WEY
1WFT	1WFW	1WFY	1WG1	1WG4	1WG7	1WGD	1WGH	1WGL
1WH5	1WH8	1WH9	1WHB	1WHD	1WHI	1WHL	1WHM	1WHN
1WIA	1WIB	1WIC	1WID	1WIE	1WIH	1WII	1WIJ	1WIK
1WJH	1WJI	1WJJ	1WJK	1WJN	1WJP	1WJR	1WJT	1WJU
1WOT	1WPB	1WQD	1WQE	1WQK	1WTE	1WTU	1WUB	1WVK

1XDX	1XFK	1XHH	1XI1	1XI7	1XIF	1XJS	1XKM	1XKR
1XOY	1XPA	1XPJ	1XQ6	1XQ8	1XQO	1XR0	1XRD	1XRS
1Y32	1Y4E	1Y66	1Y6D	1Y6U	1Y7Q	1Y7Y	1Y9W	1YBZ
1YKE	1YLQ	1YN3	1YNI	1YOP	1YPY	1YQE	1YQF	1YRG
1YZY	1Z0R	1Z23	1ZDH	1ZEC	1ZFD	1ZTN	1ZTO	1ZXQ
2CAS	2CBL	2CMD	2CMK	2CTH	2CUT	2CWG	2DKB	2DOR
2GAT	2GMF	2GST	2HGS	2HPA	2HRV	2HVM	2I1B	2IF1
2MLP	2MLT	2NR1	2NSY	2OCC	2OMF	2OVO	2PDD	2PGD
2SCP	2SQC	2TGI	2TRX	2U1A	2UP1	2VSG	3CAO	3CHY
3PCH	3PMG	3PVA	3PYP	3SIL	3TMK	3VUB	4AAH	4BLC
1A66	1A6C	1A6F	1ACW	1AD2	1AD6	1AHL	1AHU	1AIL
1BJ8	1BJA	1BJX	1BPV	1BQC	1BQF	1BWZ	1BXD	1BXY
1D1R	1D2E	1D2N	1DBO	1DBT	1DCQ	1DIO	1DIP	1DIV
1EHS	1EHX	1EI9	1ERD	1ESC	1ESJ	1EZG	1EZJ	1EZW
1G9L	1G9P	1GA3	1GKS	1GL2	1GL4	1GTQ	1GU7	1GUI
1I7Q	1I82	1I85	1IHO	1II7	1IIE	1IOW	1IP9	1IPS
1JU8	1JUV	1JW2	1K40	1K42	1K5H	1KHI	1KHM	1KIT
1LSU	1LTS	1LU4	1M1C	1M1L	1M36	1M9O	1MAI	1MBM
1NJQ	1NKG	1NKL	1NTH	1NTV	1NV8	1O7V	1O8R	1OAO
1PMI	1PNF	1PNJ	1PUD	1PUZ	1PV0	1Q2H	1Q2J	1Q2Y
1R6Y	1R79	1R7C	1RIS	1RJI	1RKB	1RQP	1RR7	1RSO
1T2Y	1T33	1T4W	1TDP	1TE5	1TE7	1TLJ	1TM9	1TNR
1UZC	1V05	1V0E	1V6B	1V6F	1V70	1VCB	1VCC	1VCL
1WHO	1WHQ	1WHR	1WIL	1WIM	1WIN	1WJV	1WJW	1WJZ
1YTB	1YTK	1YTL	256B	2A0B	2ACY	2DRI	2DTB	2EBN
1AOO	1AOR	1AOY	1AVQ	1AXH	1AXJ	1B4V	1B5T	1B63
1C3E	1C3G	1C3R	1CCV	1CCZ	1CD8	1CKN	1CKQ	1CKU
1DOS	1DP3	1DPG	1DU1	1DU2	1DU9	1E2B	1E2T	1E4T
1FAQ	1FAZ	1FBR	1FJ7	1FJN	1FJR	1FT1	1FTR	1FU3
1H2W	1H3Q	1H3Z	1HCR	1HCX	1HD2	1HNR	1HO2	1HOE
1IV0	1IVZ	1IW4	1J3G	1J3M	1J54	1JF8	1JFG	1JFL
1KPP	1KPT	1KQ5	1KZQ	1KZU	1L0O	1L9K	1L9L	1L9V
1MM0	1MNT	1MO7	1MVJ	1MWK	1MWP	1N4T	1N5G	1N67
1OM2	1OMC	1ON4	1OUP	1OVQ	1OVX	1P5L	1P65	1P68
1QA6	1QAX	1QAZ	1QHV	1QHX	1QJ8	1QRD	1QRR	1QS2
1S3A	1S4F	1S5R	1SEI	1SES	1SFE	1SLJ	1SLQ	1SML
1TV0	1TV8	1TVG	1UCP	1UDH	1UDK	1UHS	1UHT	1UHW
1VG5	1VGH	1VHH	1VZS	1W0B	1W1N	1WF0	1WF1	1WF6
1WWZ	1WYK	1WYU	1XN4	1XN6	1XN7	1XSF	1XSJ	1XU6
2JHB	2KIN	2LBP	2PII	2POL	2POR	3CLA	3CRD	3DAA
1BBI	1BBP	1BBY	1CQ0	1CQ3	1CQQ	1ECP	1ECR	1ECS
1ND9	1NDO	1NE3	1PEA	1PEH	1PEI	1QZ4	1QZH	1QZM
4CRX	4KBP	4THI	1LM0	1LMI	1LMM	1YCC	1YCO	1YCQ
1G12	1G25	1G2H	1HW7	1HY9	1HYI	1JKW	1JKZ	1JL5
1SSF	1STM	1STN	1UOY	1UQV	1URF	1WGM	1WGO	1WGP
6STD	7RSA	8CHO	8GPB	8RXN	8TFV			

Table A3.1: List of PDB structures

BIBLIOGRAPHY

[1] Horton, H.R., L.A. Moran, R.S. Ochs, J.D. Rawn, Principles of Biochemistry, Prentice-Hall, Inc. , 2000.

[2] Pauling L. The Nature of The Chemical Bond, 3rd. Ed., Cornell University Press, 1960.

[3] Jacoboni, I., Martelli, P. L., Fariselli, P., Compiani, M., Casadio, R., Predictions of Protein Segments With The Same Amino acid Sequence And Different Secondary Structure: A Benchmark For Predictive Methods, Proteins: Structure, Function And Genetics, 41 2000, 535-544.

[4] Berman, H.M., Et Al., The Protein Data Bank. Nucleic Acids Res, 28 2000, P. 235-242.

[5] Staniforth, R.A., Et Al., The Energetics And Cooperativity of Protein Folding: A Simple Experimental Analysis Based Upon The Solvation of Internal Residues, Biochemistry, 32 1993, 3842-3851.

[6] Ghelis, C., J. Yon, Protein Folding, New York: Academic Press, 1982.

[7] Bryngelson, J.D., J.N. Onuchic, N.D. Socci, P.G. Wolynes, Funnels, Pathways, And The Energy Landscape of Protein Folding: A Synthesis, Proteins: Structure, Function And Genetics, 21 1995, 167-195.

[8] Dill, K.A. And H.S. Chan, From Levinthal To Pathways To Funnels, Nature Structural Biology, 4 1997, 10-19.

[9] Baldwin, L.R., G.D. Rose, Is Protein Folding Hierarchic? Folding Intermediates And Transition States, Trends In Biochemical Sciences, 24 1999, 77-83.

[10] Walther D., Cohen F.E., Conformational Attractors On The Ramachandran Map. Acta Crystallographica Section D, 55 1999, 506 –517.

[11] Kleywegt Gj., Jones Ta., Phi/Psi-Chology: Ramachandran Revisited, Structure, 4 1996, 1395–1400.

[12] Gunasekaran K., Ramakrishnan C., Balaram P., Disallowed Ramachandran Conformations of Amino Acid Residues In Protein Structures, Journal of Molecular Biology, 264 1996, 191–198.

[13] Pal D., Chakrabarti P., On Residues In The Disallowed Region of The Ramachandran Map, Biopolymers, 63 2002, 195–206.

[14] Lovell S.C., Davis I.W., Arendall W.B., Debakker, P.I.W., Word J.M., Prisant M.G., Richardson J.S., Richardson D.C., Structure Validation By Cα Geometry: Φ,Ψ And Cβ Deviation, Proteins: Structure, Function And Genetics 50 2003, 437-450.

[15] Dill K.A., Shortle D., Denatured States of Proteins. Annual Review of Biochemistry, 60 1991, 795-825.

[16] Ramakrishnan, C., Ramachandran G.N., Stereochemical Criteria For Polypeptide And Protein Chain Conformations. Ii. Allowed Conformations For A Pair of Peptide Units. Biophysical Journal, 5 1965, 909-33.

[17] Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V., Stereochemistry of Polypeptide Chain Conformations. Journal of Molecular Biology, 7 1963, 95-99.

[18] Bahar, I., Kaplan, M., Jernigan, R.L., Short-Range Conformational Energies, Secondary Structure Propensities, And Recognition of Correct Sequence-Structure Matches, Proteins: Structure, Function And Bioinformatics, 29 1997, 292- 308.

[19] Karplus, P.A., Experimentally Observed Conformation-Dependent Geometry And Hidden Strain In Proteins, Protein Science, 5 1996, 1406-1420.

[20] Flory, P., Statistical Mechanics of Chain Molecules. New York: Wiley, 1969.

[21] Mattice, W., Suter U., Conformational Theory of Large Molecules, New York: Wiley Interscience, 1994.

[22] Volkenstein, M., Configurational Statistics of Polymer Chains, New York: Interscience, 1963.

[23] Swindells, M.B., Macarthur, M.W., Thornton, J.M., Intrinsic Fi Mu Psi Propensities of Amino Acids, Derived From The Coil Regions of Known Structures, Nature: Structural Biology, 2 1995, 596-603.

[24] Penkett, C.J., Et Al., Nmr Analysis of Main-Chain Conformational Preferences In An Unfolded Fibronectin-Binding Protein, Journal of Molecular Biology, 274 1997, 152-159.

[25] Jha, A.K., Colubri, A., Zaman, M.H., Koide, S., Sosnick, T.R., And Freed K.F., Helix, Sheet, And Polyproline Ii Frequencies And Strong Nearest Neighbor Effects In A Restricted Coil Library, Biochemistry, 44 2005, 9691-9702.

[26] Avbelj, F., Baldwin, R.L., Origin of The Neighboring Residue Effect On Peptide Backbone Conformation, Pnas, 101 2004, 10967-10972.

[27] Brant, D.A., Flory, P.J., The Role of Dipole Interactions In Determining Polypeptide Configurations, Journal of The American Chemical Society, 87 1965, 663-664.

[28] Brant, D.A., Flory, P.J., The Configuration of Random Polypeptide Chains, Theory, Journal of The American Chemical Society, 87 1965, 1175-1184.

[29] Zaman, M.H., Shen, M.Y., Berry, R.S., Freed, K.F., Sosnick, T.R., Investigations Into Sequence And Conformational Dependence of Backbone Entropy, Inter-Basin Dynamics And The Flory Isolated-Pair Hypothesis For Peptides, Journal of Molecular Biology, 331 2003, 693-711.

[30] Hu, H., Elstner, M., Hermans, J., Comparison of A Qm/Mm Force Field And Molecular Mechanics Force Fields In Simulations of Alanine And Glycine “Dipeptides” (Ace-Ala- Nme And Ace-Gly- Nme) In Water In Relation To The Problem of Modeling The Unfolded Peptide Backbone In Solution, Proteins: Structure, Function And Genetics, 50 2003, 451-463.

[31] Mu, Y., Kosov, D., Stock, G., Conformational Dynamics of Trialanine In Water Ii: Comparison of Amber, Charmm, Gromos, And Opls Force Fields To Nmr And Infrared Experiments, Journal of Physical Chemistry, 107 2003, 5064-5073.

[32] Keskin, O., Yuret, D., Gursoy, A., Turkay, M., Erman, B., Relationships Between Aminoacid Sequence And Backbone Torsion Angle Preferences, Proteins: Structure, Function And Bioinformatics, 55 2004, 992-998.

[33] Serrano L., Comparison Between The F Distribution of The Amino Acids In The Protein Database And Nmr Data Indicates That Amino Acids Have Various Phi-Propensities In The Random Coil Conformation, Journal of Molecular Biology, 254 1995, 322-333.

[34] O’connell, T.M., Wang, L., Tropsha, A., Hermans, J., The ‘‘Random-Coil’’ State of Proteins: Comparison of Database Statistics And Molecular Simulations, Proteins: Structure, Function, And Genetics, 36 1999, 407-418.

[35] Gibrat, J.F., Robson, B., Garnier, J., Influence of The Local Amino Acid Sequence Upon The Zones of The Torsional Angles F And C Adopted By Residues In Proteins, Biochemistry, 30 1991, 1578-1586.

[36] Hong, S.K., Kurochkina, N.A., Lee, B., Estimation And Use of Protein Backbone Probabilities, Journal of Molecular Biology, 229 1993, 448-460.

[37] Sippl, J., Knowledge-Based Potentials For Proteins, Current Opinion In Structural Biology, 5 1995, 229-235

[38] http://www.genome.gov//pages/hyperion/dir/vip/glossary/illustration/protein.cfm

[39] Babajide A, Farber R, Hofacker Il, Inman J, Lapedes As, Stadler Pf. Exploring Protein Sequence Space Using Knowledge-Based Potentials. J Theor Biol. 2001 Sep 7;212(1):35-46.

[40] Hobohm, U., Scharf, M., Schneider, R., Sander C., Selection of A Representative Set of Structures From The Brookhaven Protein Data Bank, Protein Science, 1 1992, 409-417.

[41] Hobohm, U., Sander, C., Enlarged Representative Set of Protein Structures, Protein Science, 3 1994, 522-524.

[42] Hooft, R.W.W., Sander, S., Vriend, G., The Pdbfinder Database: A Summary of Pdb, Dssp And Hssp Information With Added Value, Cabios, 12 1996, 525-529.

[43] Kabsch, W., Sander, C., Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded And Geometrical Features, Biopolymers, 22 1983, 2577-2637.

[44] Fitzkee, N.C., Rose, G.D., Reassessing Random-Coil Statistics In Unfolded Proteins, Pnas, 101 2004, 12497-502

[45] Anfinsen, C.B., Principles That Govern The Folding of Protein Chains, Science, 181 1973, 223-230.

[46] Russ, W.P., Ranganathan, R., Knowledge-Based Potential Functions In Protein Design, Current Opinion In Structural Biology, 12 2002, 447-452.

[47] Crooks, G.E., Brenner, S.E., Protein Secondary Structure: Entropy, Correlations And Prediction, Bioinformatics, 20 2004, 1603–1611.

[48] Mezei, M., Chameleon Sequences In The Pdb, Protein Engineering, 11 1998, 411-414.

[49] Kabsch, W., Sander, C., On The Use of Sequence Homologies To Predict Protein Structure: Identical Pentapeptides Can Have Completely Different Conformations, Pnas, 81 1984, 1075-1078.

[50] Argos, P., Analysis of Sequence Similar Pentapeptides In Unrelated Protein Tertiary Structures: Strategies For Protein Folding And A Guide For Site-Directed Mutagenesis, Journal of Molecular Biology, 197 1987, 331-348.

[51] Cohen, B.I., Presnell, S.R., Cohen, F.E., Origins of Structural Diversity Within Sequentially Identical Hexapeptides, Protein Science, 2 1993, 2134-2145.

[52] Mezei, M., Chameleon Sequences In The Pdb, Protein Engineering, 6 1998, 411-414.

[53] Sudarsanam, S., Structural Diversity of Sequentially Identical Subsequences of Proteins: Identical Octapepides Can Have Different Conformations, Proteins, 30 1998, 228-231.