THERMODYNAMIC, STRUCTURAL AND DYNAMICAL ANALYSES OF RANDOM COIL STATE OF PROTEINS

by

Çiğdem Sevim Bayrak

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

Doctor of Philosophy
in
Computational Sciences and Engineering

Koç University

November 2013

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Çiğdem Sevim Bayrak

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:

Date: 27/11/2013

ABSTRACT

Proteins are dynamic biological molecules that play important roles in cells. They may adopt many conformations influenced by short and long range interactions. In the random coil state, only short range (local) interactions are accounted. It is possible to analyze the random coil state of proteins using the methods of polymer chain statistics.

The unfolded state of proteins consists of enormous number of random configurations. Furthermore, proteins adopt many different conformations under denaturing conditions. Also, the intrinsically disordered proteins that do not assume a stable, unique 3D structures display characteristics of random coils. Hence, it is important to have a full understanding of random coil states.

A statistical analysis of thermodynamic functions such as conformational free energy, mean energy, entropy and heat capacity is given to characterize the random coil state. The approach is based on the rotational isomeric states model that is developed for polymer theory. The explicit expressions for the thermodynamic properties are derived.

Several computational studies for predicting peptide conformations exist in the literature. However, most of the existing work is confined to secondary structure prediction. An approach for determination of the high probability conformations of peptides in the random coil state is lacking in the literature although some studies provide prediction of the most probable state for each residue of a chain. We present a novel scheme that considers a probability distribution that may provide a detailed analysis of the conformation space by sampling. Since this is the first approach that provides high probability conformations of peptide sequences, we have compared our results by calculating prediction accuracies for the most probable states of residues. Our approach gives 9% better predictions than the existing ones based on the most probable states of the residues.

Since proteins are dynamic structures, it is important to characterize them both structurally and dynamically to have a complete understanding. We present a computational method to analyze

conformational transitions of the twenty amino acids based on MD simulations and the DRIS model. The relaxation times are identified and presented for the twenty amino acids. Our results are in good agreement with experimental data.

ÖZET

Proteinler hücrelerde önemli rollere sahip dinamik biyolojik moleküllerdir. Kısa ve uzun aralıklı etkileşimlerden etkilenerek birçok yapı edinirler. Rastgele sarım halinde yalnızca kısa aralıklı (lokal) etkileşimler sorumludur. Proteinlerin rastgele sarım halini polimer zincirlerin istatistiği yöntemlerini kullanarak incelemek mümkündür.

Proteinlerin katlanmamış hali muazzam sayıda rastgele yapıdan oluşur. Ayrıca, proteinler doğal yapılarını bozan şartlar altında birçok farklı yapı edinir. Aynı zamanda, sabit ve eşsiz bir üç boyutlu yapısı olmayan özünde düzensiz proteinler rastgele sarım karakteristiği gösterir. Bu nedenle, rastgele sarım hallerinin tam anlaşılması önemlidir.

Rastgele sarım hallerini karakterize etmek için yapısal enerji, ortalama enerji, entropi ve ısı kapasitesi gibi termodinamik fonksiyonların istatistiksel analizi verilmiştir. Bu yaklaşım polimer teorisi için geliştirilmiş olan dönel izomeric haller modeline dayanır. Termodinamik özellikler için açık ifadeler elde edilmiştir.

Literatürde peptid yapılarını tahmin eden farklı hesaplamalı çalışmalar bulunmaktadır. Ancak, mevcut çalışmaların çoğu ikincil yapı tahmini ile sınırlıdır. Birkaç çalışma zincirdeki her bir kökün en olası halinin tahminini sağlasa da, rastgele sarım halindeki peptidlerin yüksek olasılıklı yapılarını elde eden bir yaklaşım literatürde eksiktir. Örnekleme ile yapı uzayının detaylı analizini sağlayabilecek bir olasılık dağılımını ele alan yeni bir plan sunuyoruz. Bunun peptid dizilerinin yüksek olasılıklı yapılarını sağlayan ilk yaklaşım olmasından ötürü, sonuçlarımızı köklerin en olası halleri için tahmin doğruluğunu hesaplayarak karşılaştırdık. Yaklaşımımız köklerin en olası hallerine dayanan mevcut yaklaşımlardan %9 daha iyi tahminler vermektedir.

Proteinler dinamik yapılar oldukları için tam anlaşılması için hem yapısal hem de dinamik olarak karakterize edilmeleri önemlidir. Yirmi amino asidin yapısal geçişlerini analiz etmek için MD simülasyonları ve DRIS modeline dayanan hesaplamalı bir yöntem sunuyoruz. Yirmi amino asit için gevşeme sürelerini tanımladık ve sunduk. Sonuçlarımız deneysel veri ile iyi uyum göstermektedir.

To my little daughter Defne Nil for providing me strength and hope to complete this dissertation

ACKNOWLEDGEMENTS

I owe my deepest gratitude to my advisor Prof. Burak Erman for his great guidance. Prof. Erman provided me valuable suggestions and directions throughout this research. Furthermore, he motivated me to think and research on problems as a mature scientist. This thesis would not have been completed without his encouragements.

I am also grateful to my thesis committee of Prof. Attila Gürsoy, Prof. Özlem Keskin, Assistant Prof. Mehmet Sayar and Prof. Türkan Haliloğlu for their participation and for critical reading of my thesis.

I would like to thank both present and former members of CMSE and CHBI for their friendships. This journey was also an opportunity to meet great friends who made the five years enjoyable: Özge Engin Şensoy, Evrim Besray Ünal, Nurcan Tunçbağ, Gözde Kar, Tuğba Özal, Bahar Değirmenci, Derya Aydın, Cemal Erdem, Beytullah Özgür, Cahit Dalgıçdır, Ayşe Küçükyılmaz, Ceren Tüzmen, Cemre Kocahakimoğlu, Gözde Eskici, and Deniz Aydın.

I am indebted to my parents and my lovely sister for the encouragement and love they provided me through not only this study but also my entire life. Special thanks to my sister İrem Sevim for bringing joy and support to my life.

Finally, I would like to thank to my beloved husband Ali Osman for his endless support, love and understanding. He provided uncomplaining encouragement in the crisis moments and opportunity to study whenever and wherever I needed to. He made it possible for me to complete this study.

I gratefully acknowledge the financial support of TÜBİTAK and Koç University.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

PDB	Protein Data Bank
IDP	Intrinsically Disordered Protein
RIS	Rotational Isomeric State
MD	Molecular Dynamics
DRIS	Dynamic Rotational Isomeric State

Chapter 1

INTRODUCTION

Proteins have significant roles in cellular function and disease processes. They are dynamic structures that assume a large ensemble of conformations around the average structure. It is important to analyze both the probabilities of the conformational states and the energy barriers between these states to understand proteins.

Proteins assume different stable states as folded (native), unfolded (random coil), molten globule, or pre-molten globule. Different conformations of proteins are characterized by backbone torsion angles (ϕ,ψ) of the amino acids of the chain. Some values of the (ϕ,ψ) angles are not allowed and some values are more popular than the other values. These preferences are different for different types of amino acids.

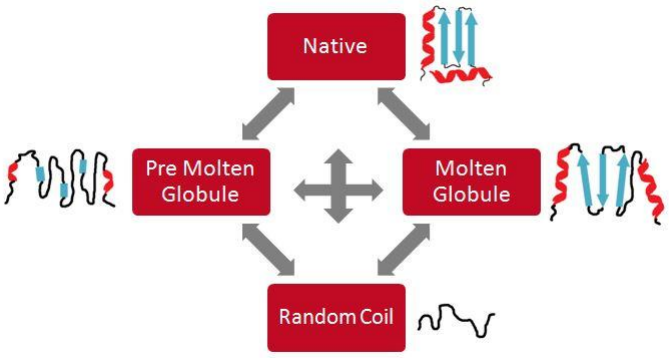


Figure 1. A schematic quartet model for different states of proteins

Protein chains adopt random configurations based on the backbone torsion angles of the residues. Configurational preferences of a given residue depend on both local and nonlocal interactions. Short-range (local) interactions result from the effects of neighboring residues whereas the long-range (nonlocal) interactions result from the non-neighbor residues. In the

random coil state, only the near-neighbor effects are taken into account. The term ‘randomly coiled proteins’ describing this state has been studied in detail by Flory and collaborators in polymer theory [1-7].

Understanding the random coil state of proteins is important due to several reasons: Firstly, the set of random configurations covers all possible initial conformations of proteins. Depending on the primary sequence, some conformations emerge as highly probable due to the amino acid specific regions of the (ϕ,ψ) angles. Secondly, under strongly denaturing conditions, a wide range of values become available to (ϕ,ψ), and conformations are close to those of the random coil [8, 9]. Thirdly, the functionally important ‘intrinsically disordered protein’ concept where the primary sequence prohibits the folded state, may suitably be analyzed by the tools used to understand the random conformations [10, 11].

This dissertation mainly focuses on applications of the rotational isomeric states (RIS) model on proteins. The outline of this dissertation is as follows:

Chapter 2 presents a general review on the RIS model, determination of statistical properties of chains based on the RIS model, intrinsically disordered proteins (IDP), hidden Markov model and the dynamic RIS model that allows determination of transition probabilities.

Chapter 3 focuses on a statistical analysis for determination of thermodynamic properties of proteins in the random coil state. The availability of several conformations necessitates a statistical approach. Conformational free energy, mean energy, entropy and heat capacity expressions are derived using the RIS model.

Chapter 4 demonstrates a computational scheme for finding high probability conformations of peptide sequences in the random coil state. The model calculates the probability distribution of the torsion states based on the RIS formalism. High probability conformations are then obtained by using a hidden Markov model Viterbi algorithm.

Chapter 5 illustrates a computational method for determination of conformational transitions of the amino acids by the dynamic rotational isomeric state (DRIS) model. Local dynamics of amino acids resulting from rotational transitions between isomeric states are analyzed.

In Chapter 6, the results are discussed and a short summary of major conclusions of the study are presented.

Chapter 2

LITERATURE REVIEW

2.1 ROTATIONAL ISOMERIC STATES MODEL

The statistical description of the denatured states of proteins based on polymer theory goes back to the work of Flory and collaborators based on the Rotational Isomeric State (RIS) formalism[1]. The RIS model for a protein chain consists of two major components: (1) The statistical weights of the torsion states of the (ϕ,ψ) angles, (2) The proper matrix multiplication operations leading to the partition function of the chain. Thermodynamics of a single chain then follows upon matrix operations based on the partition function and its derivatives [12, 13].

The RIS formalism replaces the continuous distribution of backbone torsion angles by a distribution over several discrete states, and integrals over the energy surface are approximated by summations over these states.

The native state of a protein is obtained when each torsion angle selects a unique value. Two torsion angles around the alpha carbon, C^{α}, describe the local conformation of a residue. The applicability of the RIS model to proteins is facilitated by the Flory isolated pair hypothesis, which suggests that each pair of torsion angles is independent of the angles occupied by neighboring pairs[1, 2]. Rose and coworkers investigated the validity of the hypothesis by exhaustive enumeration and showed that the isolated pair hypothesis does not hold in general [14]. Zaman et al. demonstrated that important interactions between neighboring residues exist and thus the hypothesis is invalid [15]. On the other hand, Ohkuba and Brooks confirmed the validity of Flory's hypothesis in the context of helical peptides [16]. Generating a statistical coil model, it has been shown that the inclusion of correlations coming from conformations of neighboring residues improves the consistence with experimental data [17]. Jha et al. studied the intrinsic conformational preferences of residues in a restricted coil library and showed that there is a correlation between the conformation and chemical character of neighboring residues and the structural preferences [18]. Keskin et al. also studied the conformational propensities of amino acids and confirmed the existence of significant correlations between neighboring torsion angle

pairs [19]. Esposito at al. indicated that the value of an omega torsion angle is strongly correlated with the value of the adjacent psi angle [20]. Colubri et al. studied the contribution of neighbors in protein folding by comparing the results of simulations with and without neighbor effects [21]. In a recent work it has been shown that the usage of (ψ_i, φ_{i+1}) provides more information on backbone behavior as opposed to independent usage of residues [22].

It has been shown that some values of torsion angles are more favorable than others, and different amino acid types have different propensities to occur in different angles [23]. The dependence between the torsion states of two neighboring residues is a function of the type of the residues [19]. We elaborate further on this point in discussing the construction of energy maps.

2.2 STATISTICAL MECHANICS

A configuration $\{\omega\}$ of an N-bond chain is defined by torsion angles (ϕ,ψ) describing the torsional rotations of amino acids. If we define the torsional energy when bond $i-1$ is in state η and bond i is in state ζ as $E_{\eta\zeta;i-1,i}$, then the configurational energy of the chain for a given configuration is

$$E\{\omega\} = E_{12} + E_{23} + E_{34} + \cdots + E_{i-1,i} + E_{i,i+1} + \cdots + E_{2N-1,2N} \tag{2.1}$$

Here $E_{i-1,i}$ is the energy corresponding to the joint occurrence of bonds $i-1$ and i in their respective states where the indices η and ζ indicating the corresponding states are omitted for simplicity.

Statistical weights for corresponding to the torsional energies is defined by

$$u_{\eta\zeta;i} = \exp(-E_{\eta\zeta;i}/RT) \tag{2.2}$$

Statistical weights may be expressed in the form of statistical weight matrix U_i as

$$U_i = \left[u_{\eta\zeta}\right]_i \tag{2.3}$$

where the $\eta\zeta^{th}$ element indicates the statistical weight bond i is when it is in state ζ while the bond $i-1$ is in state η. The rows of the matrix represent the states for bond $i-1$ and the columns represent the states for bond i.

Then, the statistical weight of a configuration $\{\omega\}$ of a chain is calculated a serial multiplication of the statistical weights of individual bonds.

$$\Omega_{\{\omega\}} = \prod_i u_{\eta\zeta;i} \tag{2.4}$$

The statistical weight of a configuration shows the frequency of occurrence in a statistical mechanical ensemble at equilibrium.

The partition function, Z, of the chain is given by

$$Z = \sum_{\{\omega\}} \Omega_{\{\omega\}} = \sum_{\{\omega\}} \prod_i u_{\eta\zeta;i} \tag{2.5}$$

where the summations are taken over all configurations. Calculation of the partition function using this equation would be computationally difficult as the number of bonds increasing. A computationally efficient method is using matrix methods. The entire sum of products of statistical weights can be determined by matrix multiplication [1, 12, 24].

The partition function of a chain of N bonds with ν number of rotational states is calculated by

$$Z = J^* \left[\prod_{i=2}^{N-1} U_i \right] J \tag{2.6}$$

where J^* and J are the row and column vectors of order $1\times\nu$ and $\nu\times1$ respectively.

$$J^* = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}; \quad J = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{2.7}$$

The probability that a given molecule occurs in a given configuration is determined as the statistical weight divided by the sum of statistical weights for all possible configurations. Hence,

$$p_{\{\omega\}} = Z^{-1}\prod_{i=2}^{N-1} u_{\zeta\eta;i} \tag{2.8}$$

The probability $p_{\eta;i}$ that bond i is in state η is calculated as the quotient of the sum of the statistical weights for all configurations for which bond i is in state η, divided by Z

$$p_{\eta;i} = Z^{-1}J^{*}\left[\prod_{h=2}^{i-1} U_h\right]U'_{\eta;i}\left[\prod_{j=i+1}^{N-1} U_j\right]J \tag{2.9}$$

where $U'_{\eta;i}$ is the matrix obtained from U_i by equating all elements to zero except the elements of column η. In this way, we keep only the terms satisfying the condition $\omega_i = \omega_{\eta;i}$.

The probability $p_{\zeta\eta;i}$ that bonds $i-1$ and i occur simultaneously in states ζ and η respectively is given by

$$p_{\zeta\eta;i} = Z^{-1}J^{*}\left[\prod_{h=2}^{i-1} U_h\right]U'_{\zeta\eta;i}\left[\prod_{j=i+1}^{N-1} U_j\right]J \tag{2.10}$$

where $U'_{\zeta\eta;i}$ is obtained by striking all entries of U_i with the exception of $u_{\zeta\eta;i}$.

This matrix multiplication method is also applicable to determination of derivatives of the partition function. For example, first derivative of the partition function with respect to temperature is

$$dZ/dT = L^{*}\left[\prod_{i=2}^{N-1}\hat{U}_{T;i}\right]L \tag{2.11}$$

where

$$L^* = \begin{bmatrix} J^* & 0 & \cdots & 0 \end{bmatrix} \text{ and } L = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ J \end{bmatrix} \tag{2.12}$$

and $\hat{U}_{T;i}$ is the super matrix that the elements of which are matrices

$$\hat{U}_{T;i} = \begin{bmatrix} U & U_T^{'} \\ 0 & U \end{bmatrix}_i \tag{2.13}$$

Here $U_{T;i}^{'} = dU_i / dT$.

Similarly,

$$d^2Z/dT^2 = M^* \left[\prod_{i=2}^{N-1} \hat{U}_{T;i} \right] M \tag{2.14}$$

where

$$M^* = \begin{bmatrix} J^* & 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 \end{bmatrix} \text{ and M=column} \begin{bmatrix} 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & J \end{bmatrix} \tag{2.15}$$

$$\hat{U}_{T;i} = \begin{bmatrix} U & U_T^{'} & U_T^{'} & U_T^{''} \\ 0 & U & 0 & U_T^{'} \\ 0 & 0 & U & U_T^{'} \\ 0 & 0 & 0 & U \end{bmatrix} \tag{2.16}$$

Here $U_{T;i}^{'} = dU_i / dT$ and $U_{T;i}^{''} = d^2U_i / dT^2$.

2.3 INTRINSICALLY DISORDERED PROTEINS

A protein whose native state has undergone a main change as noncovalent, cooperative or reversible is described as a denatured protein. The structural characteristics of denatured proteins are directly related to the characteristics of intrinsically disordered proteins (IDP). Disordered proteins are described as partially or completely unfolded proteins without a unique native structure. IDPs do not have a unique stable three dimensional structure upon crowding and the

rapid conversion between multiple states remains [25]. The lack of structure may yield functional advantages like easily adapting different conformations and ability of binding to several targets.

Disordered proteins are observed ubiquitously, ranging from the totally random coil to molten globular [26], or even they may be exchanging conformations between various states such as the prion, a protein responsible in the mad cow disease, α-synuclein, a mutated form of which is responsible for the Parkinson's disease, etc (see for example DisProt: the Database of Disordered Proteins [27]). For example the disordered α-synuclein takes a partially helical conformation when in contact with the cell wall. The transition between the partially unfolded to partially helical states is necessary for its function.

In many cases, disordered proteins switch their conformations from a completely or partially disordered conformation to a more ordered structure upon binding. Dunker et al. suggested that these disorder-to-order transitions disconnect binding energy (affinity) and the ability to discriminate a substrate (specificity) and provide advantage to adapt different binding targets [28]. Romero et al. showed that thousands of proteins contain disordered regions supporting the idea that function-structure relationships need to be extended to contain unfolded or disordered proteins [29]. In the same year, Garner et al. suggested that disordered regions are distinct from ordered proteins and thus form a separate category [30]. Wright and Dyson pointed out the contributions of disordered structures in cellular functions and suggested that the general assumption of the close relation between the protein function and its folded structure should be considered again [31]. Uversky et al. analyzed the properties of natively disordered proteins by means of hydrophobicity and charge and made it possible to predict if the given protein sequence leads to a structured or an unstructured protein [26]. They showed that low mean hydrophobicity together with high net charge indicates the lack of structure in proteins. Dunker et al. classified and analyzed disorder-function relationships. The observed functions are molecular recognition, molecular assembly/disassembly, protein modification, and entropic chain activities [32]. In another study, Dunker et al. predicted disorder of proteins based on the amino acid sequence and showed the commonness of disorder in protein structure [33]. As a result, they showed that eukaryotes have more disordered regions than archaea or bacteria suggesting that unstructured proteins are more frequent in complex organisms.

Tompa predicted existence of repeat regions in some known unstructured proteins and showed that these repeat segments are functionally essential and widely occur in disordered proteins [34]. The classification of the functional information of disordered proteins based on the Gene Ontology (GO) classification has been given in five functional areas; (i) transcription and transcription regulation, (ii) signal transduction and the regulation of cell cycle, (iii) the biogenesis and functioning of nucleic acid containing organelles, (iv) messenger ribonucleic acid (mRNA) processing, and (v) the organization and biogenesis of cytoskeleton . Also, the classification of functional modes of disordered proteins has been given in seven categories; (i) entropic chain functions, (ii) display site functions, (iii) chaperone functions, (iv) effector functions, (v) scavenger functions, (vi) assembler functions, and (vii) prion functions.

Gunasekaran et al. gave another argument for existence of large proportion of disordered proteins in eukaryotic cells in terms of physical constraints [35]. For protein stability, large protein size is needed for having large intermolecular interfaces. However, this would cause some critical problems. Hence, by comparing the interface sizes of ordered and disordered proteins, it was shown that proteins in their extended states are good solutions for having large interfaces with smaller protein sizes. On the other hand, an argument for efficient induced folding mechanism of disordered proteins has been given as the preformed elements which enables them an easy and fast formation. Fuxreiter et al. proposed a folding model for disordered proteins in which the preformed elements provide an advantage for an effective interaction by making the first contacts [36]. Tompa and Csermely reviewed high existence of disordered regions in RNA and protein chaperones, and their functions in chaperon action [37].

Tompa reviewed recent advances in disordered proteins in 2005 [38]. The functional modes of disordered proteins such as entropic chains, display sites, chaperons, effectors, assemblers and scavengers, and their residual structure that explains specialized modes, and the functional advantages provided by flexible structure are given in the review. Also, a functional classification scheme of disordered proteins has been given such that the function originates either from fluctuations or ability to transiently or permanently bind partner molecules. On the other hand, Dyson and Wright reviewed the biological role of intrinsically disordered proteins and discussed the importance of folding into an ordered structure upon binding to a target, called as coupled folding and binding process, and conformational flexibility [39].

Dunker et al. explained the contribution of intrinsic disorder in protein-protein interactions by emphasizing the relation between hub connectivity and the protein structure. It has been proposed that disorder in one or both proteins may either yield a nonselective structure for connections, or flexible links between domains. In addition, important roles of interactions between ordered hubs and disordered proteins have been illustrated [40].

DisProt database has extensive data on disordered proteins in which various experimental techniques (i.e., X-ray crystallography, NMR, and circular dichroism (CD)) are used to characterize the disordered regions [27]. Since it is difficult to identify and characterize disordered proteins by experiments, a number of bioinformatics tools for predicting disorder based on amino acid sequence have been developed. The links of many predictors can also be found in the DisProt (http://www.disprot.org). The characteristics of the amino acid sequences of disordered proteins and ordered proteins show differences in terms of amino acid composition and sequence complexity [41, 42]. Disordered proteins contain less nonpolar, more charged amino acids then ordered proteins and are characterized by low sequence complexity [26, 42]. Evolution of disordered proteins in terms of amino acid substitutions is also different then ordered proteins [43]. Aromaticity, charge, hydrophobicity, and flexibility are the other differences between the amino acid sequences of disordered and ordered proteins [44]. Although different predictors use different approaches, they are all based on these sequence differences and work well. Dosztanyi et al. presented a small survey of prediction methods of identification disordered proteins or regions and discussed the methodologies [45]. An analysis on the main ideas of different prediction methods and the difficulties of them has been given in a recent review [46].

Disorder in proteins has also been studied in relation to drug design. It has been shown that there are some examples indicating that protein-protein interactions with one disordered and one ordered partner are good targets for drug discovery [44, 47]. Some proteins show significant levels of disorder in human diseases such as cancer, cardiovascular diseases, diabetes, autoimmune diseases, neurodegenerative diseases (i.e., Alzheimer’s disease, Parkinson’s disease, Huntington’s disease, and Prion disease) [44, 48]. Cheng et al. indicated the role of disorder on finding new small drug molecules that can modulate protein-protein interactions [49]. One of the

binding proteins undergoes a disorder-to-order transition upon binding to its structured complement and makes an ideal druggable target.

2.4 HIDDEN MARKOV MODEL

A Markov system is a system that can be in one of several states, and can pass from one state to another each time step according to fixed probabilities. If a Markov system is in state i, there is a fixed probability, p_{ij}, of it going into state j at the next time step. The probability p_{ij} is called a transition probability. All transition probabilities $p_{ij}, i, j = 1,\ldots N$ in a system with N possible states may be represented as a $N \times N$ matrix. Then, the entries in each row add up to 1 and the matrix is called the state transition matrix. The initial distribution of the states defines the probability of the system being in each of the states at time 0. This vector of initial probabilities is called the π vector.

A hidden Markov model (HMM) is a discrete-time Markov model with some additional features. That is, when a state is visited by the Markov chain, the state 'emits' a letter from a fixed time-independent alphabet. In a HMM, there is an observed sequence of emitted symbols which is defined by $O = O_1, O_2, O_3 \ldots$, and a sequence of states visited which is defined by $Q = q_1, q_2, q_3, \ldots$ [50, 51].

An HMM consists of the following components:

1) $S = S_1, S_2, \ldots, S_N$ A set of N states
2) $A = \{a_1, a_2, \ldots, a_M\}$ An alphabet of M distinct observation symbols
3) $P = (p_{ij})$ The transition probability matrix where $p_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$
4) $b_i(a) = P(S_i$ emits symbol a) for each S_i and a, The emission probabilities.
5) $\Pi = (\pi_i)$ An initial distribution vector where $\pi_i = P(q_1 = S_i)$

Hidden Markov models were introduced in a series of statistical papers by Leonard E. Baum and others in the late 1960s. The first application of HMM is speech recognition [52]. The system is assumed to be a Markov process with some unknown parameters in an HMM. HMMs are being used in speech recognition, natural language processing, and bioinformatics. The model is

applied to the analysis of biological sequences in the late 1980s [53]. Since then, HMMs have been widely applied in several areas such as multiple sequence alignment, gene finding, modeling protein or DNA families, predicting protein coding regions in genome sequences, predicting secondary structure elements in proteins, etc. of computational biology.

There are three types of problems that are frequently represented with an HMM. (i) Calculating the probability of some given sequence of observed symbols $P(O|\lambda)$ given the parameters $\lambda = (P, B, \pi)$, (ii) Finding the parameter set $\lambda = (P, B, \pi)$ that maximize $P(O|\lambda)$, (iii) Determining the hidden sequence of states Q that is most likely to have occurred, given sequence O. That is $\arg\max_{Q} P(Q|O)$. The latter can be carried out by an efficient algorithm. The Viterbi algorithm is introduced by Andrew Viterbi in 1967 [54].

This algorithm is a dynamic programming algorithm that efficiently searches for a state sequence Q that has the highest probability $P(Q|O)$ [50, 52, 55]. The Viterbi algorithm has two main steps; (i) finding $\max_{Q} P(Q|O)$, (ii) backtracking to find a Q that satisfies this maximum.

2.5 DYNAMIC ROTATIONAL ISOMERIC STATES MODEL

The dynamic rotational isomeric states (DRIS) model has been defined to predict local dynamics of polymer chains by Bahar and Erman [56]. The model has been used to calculate different dynamic properties associated with the transitions between the isomeric states [56-63].

DRIS is a mathematical model that applies the rotational isomeric states scheme to chain dynamics for calculating the effects of neighbor correlations based on the model introduced by Jernigan [64]. The model provides determination of the internal time correlation functions of a chain by matrix methods that has been used in calculation of chain statistics by Flory [1].

Internal dynamics of a polyethylene chain was characterized by calculating the autocorrelation of a vector affixed to the middle of a sequence in the chain [56]. Orientational correlation functions, correlation times, and spectral densities of poly(ethylene oxide) segments were determined [57]. The method was also used to determine the activation energies of the local conformational transitions for a polyethylene chain [58]. DRIS is used to obtain the conformational autocorrelation functions, and first and second orientational autocorrelation

functions of a polyethylene chain and it has been shown that the results agree satisfactorily with NMR experiments [59]. A mathematical scheme to examine the stochastic process of conformational transitions between rotational isomeric states in polymer chains was developed by Bahar [65]. According to the model, stochastic weights are assigned to the configurational transitions and stochastic weight matrices are defined. A matrix multiplication scheme is adopted to analyze the isomeric transitions by serial multiplication of stochastic weight matrices. Then, this approach is extended to efficiently calculate the first and second orientational autocorrelation functions [61]. The mechanism of local conformational transitions in poly(dialkylsiloxanes) which is a typical example that has a highly flexible backbone with bulky and highly articulated side groups is analyzed [62]. The results of DRIS method has been compared with molecular dynamics simulations and it has been shown that DRIS provides an efficient way of analyzing the mechanism of the local relaxation of a polymer chain rather than long simulations [24, 63].

For a given chain of N bonds, $\{\omega\}_k$ represents a given configuration where $k = 1, 2, \ldots, \nu^N$ if the chain has ν discrete rotational isomeric states. Then, $P^{(N)}(t)$ is defined as a ν^N-dimensional vector of time-dependent probabilities of all possible configurations $\{\omega\}_k$. Then the "master equation" that gives the time rate change of $P^{(N)}(t)$ is defined as

$$dP^{(N)}(t)/dt = A^{(N)}P^{(N)}(t) \tag{2.17}$$

where $A^{(N)}$ is the transition rate matrix of size $\nu^N \times \nu^N$. $A_{ij}^{(N)}$, the ijth element of the rate matrix, gives the rate of transition from state $\{\omega\}_j$ to $\{\omega\}_i$. The diagonal elements of the rate matrix are equal to the negative of summation of off-diagonal elements of the corresponding column that is $A_{ii}^{(N)} = -\sum_{j \neq i} A_{ji}^{(N)}$.

The formal solution of the master equation gives

$$P^{(N)}(t) = \exp\left\{A^{(N)}t\right\} P^{(N)}(t=0) \tag{2.18}$$

where $P^{(N)}(t=0)$ is the vector of the equilibrium probabilities. The term $\exp\left\{A^{(N)}t\right\}$ is defined as the time-dependent conditional (or transition) probability matrix $C^{(N)}(t)$ and represents the conditional probability of transitions. The ijth element of the conditional probability matrix, $C_{ij}^{(N)}$, denotes the conditional probability of the occurrence of configuration $\left\{\omega\right\}_i$ at time t, given the initial configuration $\left\{\omega\right\}_j$ at $t=0$. The summation of each column of $C^{(N)}$ is unity since they represent all possible transitions from a given initial configuration.

The eigendecomposition of the rate matrix $A^{(N)}$ gives

$$A^{(N)} = B^{(N)}\Lambda^{(N)}\left[B^{(N)}\right]^{-1} \tag{2.19}$$

where $B^{(N)}$ is the matrix whose ith column is the ith eigenvector of $A^{(N)}$, $\Lambda^{(N)}$ is the diagonal matrix whose diagonal elements are the eigenvalues λ_i of $A^{(N)}$, and $\left[B^{(N)}\right]^{-1}$ is the inverse of $B^{(N)}$.

The exponentiation of a diagonalizable matrix $A = UDU^{-1}$ yields $\exp\{A\} = U\exp\{D\}U^{-1}$. Using this property and the eigendecomposition of $A^{(N)}$, the time-dependent conditional probability matrix $C^{(N)}(t)$ may be calculated in terms of $B^{(N)}, \Lambda^{(N)}$, and $\left[B^{(N)}\right]^{-1}$ as

$$C^{(N)}(t) = \exp\left\{A^{(N)}t\right\} = B^{(N)}\exp\left\{\Lambda^{(N)}t\right\}\left[B^{(N)}\right]^{-1} \tag{2.20}$$

The total time-dependent joint probability matrix is defined as

$$P^{(N)}(t) = C^{(N)}(t)\text{diag}\,P^{(N)}(t=0) \tag{2.21}$$

The ijth element of the time-delayed joint probability matrix, $P_{ij}^{(N)}$ represents the joint probability of occurrence of configuration $\left\{\omega\right\}_i$ at time t and $\left\{\omega\right\}_j$ at $t=0$. Hence,

$$P_{ij}^{(N)} = \sum_k B_{ik}^{(N)} \exp\{\lambda_k t\} \left[B^{(N)} \right]_{kj}^{-1} P_j^{(N)}(0) \qquad (2.22)$$

Knowledge of the time-dependent joint probability matrix $P^{(N)}(t)$ gives a full description of the stochastic of configurational transitions.

Chapter 3

STATISTICAL MECHANICS OF PROTEINS IN THE RANDOM COIL STATE

Random configurations of protein chains are obtained under the constraints imposed by chain connectivity and the torsion states of the backbone torsion angles ϕ and ψ in the absence of sequence-distant long-range interactions. We present a statistical analysis on thermodynamic properties to describe and characterize the random coil state of proteins. Conformational free energy, energy, entropy and heat capacity expressions are derived using the Rotational Isomeric States model of polymer theory. The state space and the probabilities of each state are comprised from a coil database. Properties for the random coil state are obtained for a sample set of proteins taken from the Protein Data Bank. Thermodynamic expressions of random coil state are derived.

As stated in Chapter 1, understanding the random configurations of proteins is important due to several reasons. Thus, a better statistical understanding of denatured proteins is required for answering questions referring to functional properties of proteins. The number of states available to the denatured chain may vary from an enormous set to only a few in numbers as observed in switches. The general statistical mechanical model that we adopt is not restricted with this variation. The size of available states is determined by the probabilities of the latter, and several sources for such probabilities are either available and may be extracted from various databases, or may be generated by suitable training techniques of bioinformatics, depending on the constraints and requirements of the problem at hand. In the present study, we extract the probabilities from the Ramachandran plots obtained from the coil library [66] which is accepted to be representative of the random coiled state of proteins [7, 51, 67]. Having characterized the probabilities from the knowledge data base, we apply the matrix multiplication technique to obtain the partition function, and the thermodynamic functions such as energy, entropy and heat capacity for the denatured state. Finally we present random coil results for thermodynamic functions for several proteins whose primary sequences are chosen from the Protein Data Bank.

3.1 STATISTICAL EVALUATION

A denatured protein assumes a multitude of conformations, each subject to a certain probability determined by the configurational features of the residues which are either of local or nonlocal nature. Local effects result from interactions among neighboring amino acids along the chain. We refer to this state the random coiled state of the protein. Determination of the conformation of a chain using near neighbor interactions only reduces the problem to a Markov process. Nonlocal effects are those among residues separated by more than two residues along the chain. Having adopted the probabilities from the coil library, where the sequence-distant long-range interaction are absent because secondary or tertiary structures are lacking, is a good approximation to the Markov nature of the coiled state.

Markov statistics of denatured proteins have an important place in protein statistics in general, because: (i) This is the first approximation to the difficult problem of non-Markov behavior, (ii) Markov behavior is responsible for a large body of observed phenomena, (iii) There is already a powerful and successful Markov model of characterizing the conformations of polymers, i.e., the Rotational Isomeric States (RIS) model that has been studied in some detail. The specific aim of the present paper is to extend the RIS model to calculate the thermodynamic properties of denatured chains using data generated from the denatured components of chains from the PDB.

A common and most straightforward practice for determining torsion angle propensities is the use of knowledge based potentials. The Protein Data Bank (PDB) is the most common source. The frequency of occurrence of a given amino acid at a given torsion state leads to the probabilities. For calculations of the random denatured conformations of proteins, a coil library serves as the source of information where torsion angle data is taken from the set of amino acids those are not in helical or beta structures. Although the coil library serves as a plausible source for the denatured states of proteins, it is not the only one. One may make modifications to this data set depending on the constraints set on the conformations of the denatured proteins. In this paper, we use the Rose Protein Coil Library which can be downloaded from http://www.roselab.jhu.edu/coil/ [66].

3.1.1 STATES

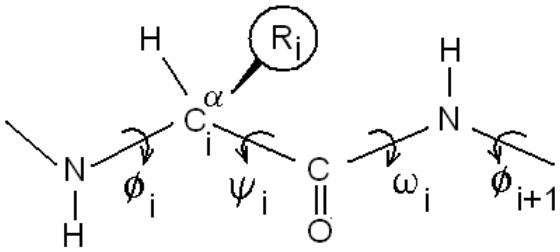


Figure 2. Torsion angles of the i[th] amino acid.

The backbone torsion angles for the i[th] amino acid are shown in Figure 2. Each bond can assume different angles, with different preferences. Each residue has three torsion angles, ϕ,ψ, and ω. The occurrence of a residue in a given ϕ and ψ state, irrespective of its type is presented in Figure 3. An examination of this figure shows that the choice of isomeric states for the ϕ and ψ angles is more complicated than the choice in synthetic polymer applications. In the latter, usually there are a few states like trans, gauche+ and gauche-, and their combinations for two successive bonds along the chain. In the protein case, one sees several discrete states as can be observed from the distribution of the points in Figure 3. Furthermore, the states are centered on different regions for the successive ϕ and ψ angles, and for different amino acids.

We perform the construction of states separately for ϕ,ψ, and ω angles. The states for the ϕ angles are chosen as follows: The points in Figure 3 are projected onto the ϕ axis, and 13 states are identified visually as the following intervals:

Table 1. The states for the angle Φ

State	Interval
ϕ_1	(-180,-150)
ϕ_2	(-150,-120)
ϕ_3	(-120,-105)
ϕ_4	(-105,-75)
ϕ_5	(-75,-40)

ϕ_6	(-40,-20)
ϕ_7	(-20,-10)
ϕ_8	(-10,30)
ϕ_9	(30,70)
ϕ_{10}	(70,105)
ϕ_{11}	(105,130)
ϕ_{12}	(130,155)
ϕ_{13}	(155,180)

The choice is made such that the clusters of points seen in Figure 3 are demarcated, as accurately as possible, by the intervals chosen. This is not a unique procedure, however, because the clusters are centered on different angular positions for different residues and the intervals chosen along the ϕ axis, which are not all equal to each other in magnitude, are only best approximations. Similarly, the points are projected onto the ψ axis, and the thirteen intervals are chosen as:

Table 2. The states for the angle ψ

State	Interval
ψ_1	(-180,-160)
ψ_2	(-160,-135)
ψ_3	(-135,-105)
ψ_4	(-105,-75)
ψ_5	(-75,-40)
ψ_6	(-40,-15)
ψ_7	(-15, 20)
ψ_8	(20, 60)
ψ_9	(60, 90)

ψ_{10}	(90,110)
ψ_{11}	(110,130)
ψ_{12}	(130,160)
ψ_{13}	(160,180)

For the ω angle, there are two states, one is either (-180,-160) or (160,180), and the other is (-20, 20). The states chosen in this manner are representative of the regions given by Karplus [23] and also in Reference [68]. Thus, we identified 13 states for the angle ϕ, 13 states for ψ, and 2 states for ω as rotational isomeric states.

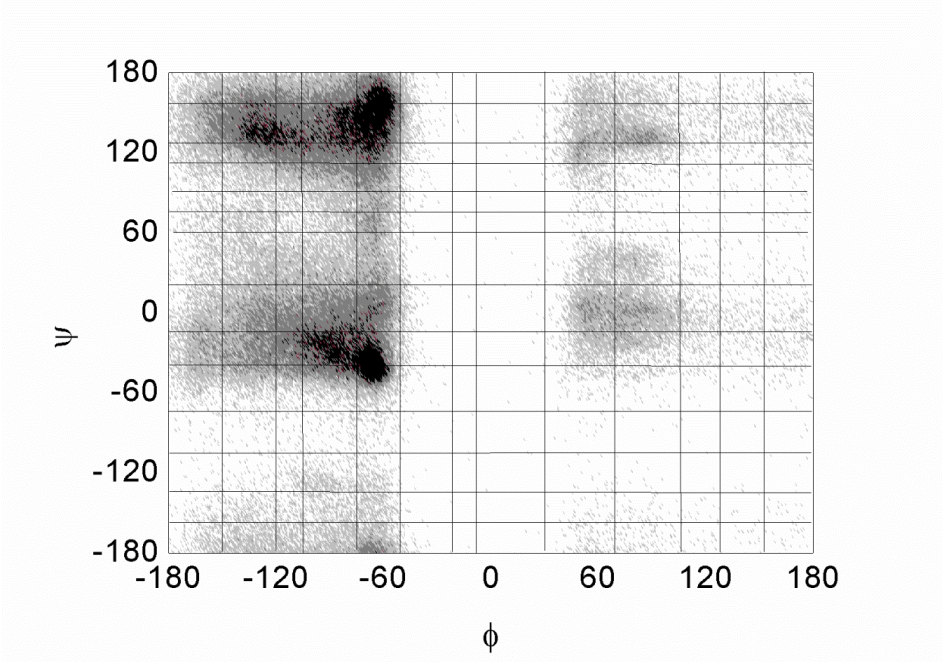


Figure 3. Regions of the Φ-Ψ plane.

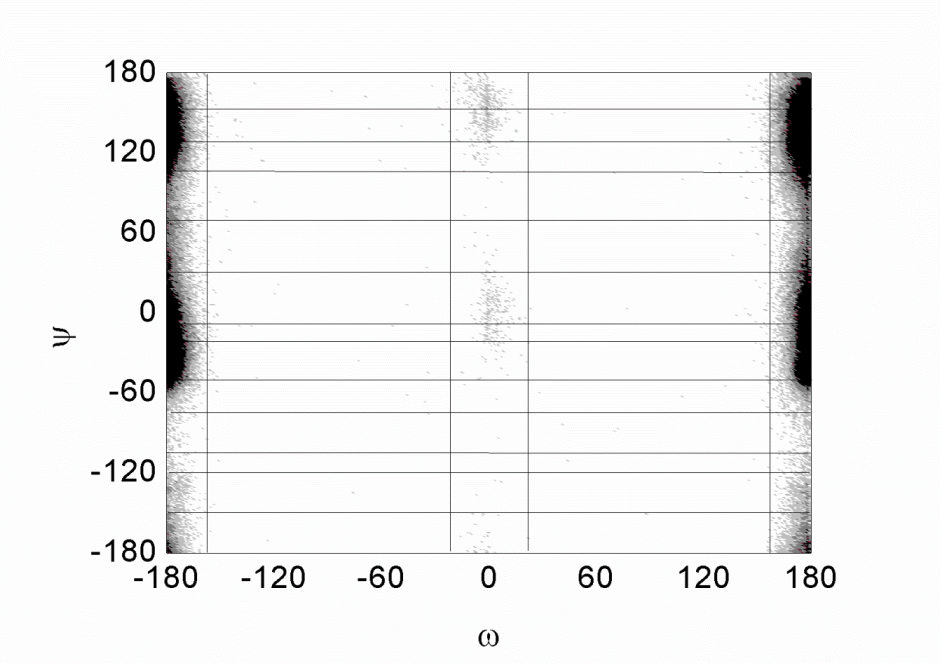


Figure 4. Regions of the ω-ψ plane.

The distributions of the ϕ, ψ, and ω angles are presented in Figure 5, Figure 6, and Figure 7 respectively. The differences in distribution for different amino acid types may be seen in Figure 8 and Figure 9.

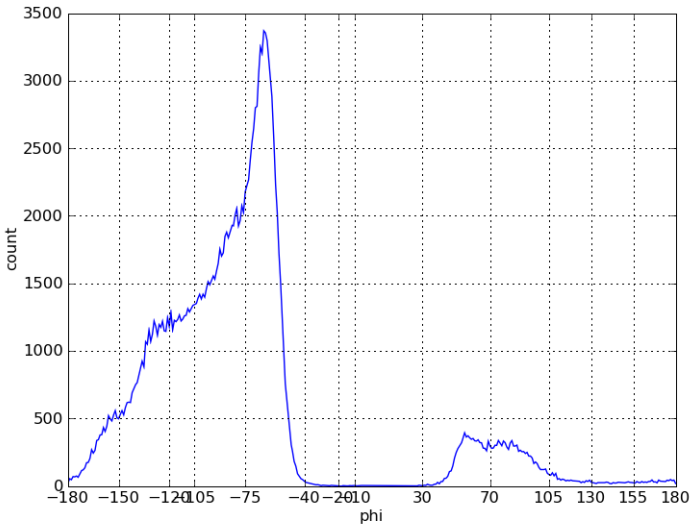


Figure 5. The distribution of the Φ angles in coil database.

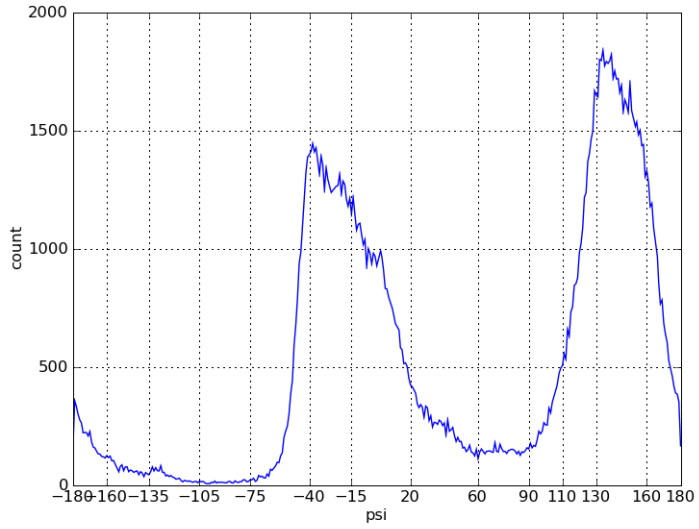


Figure 6. The distribution of the ψ angles in coil database

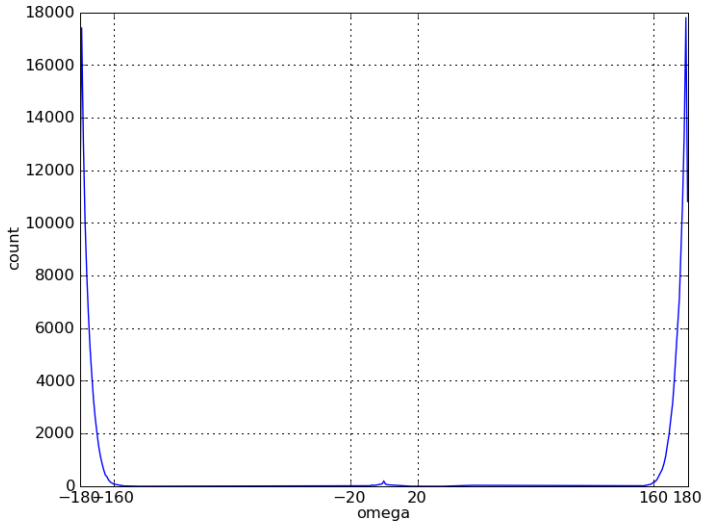


Figure 7. The distribution of the ω angles in coil database.

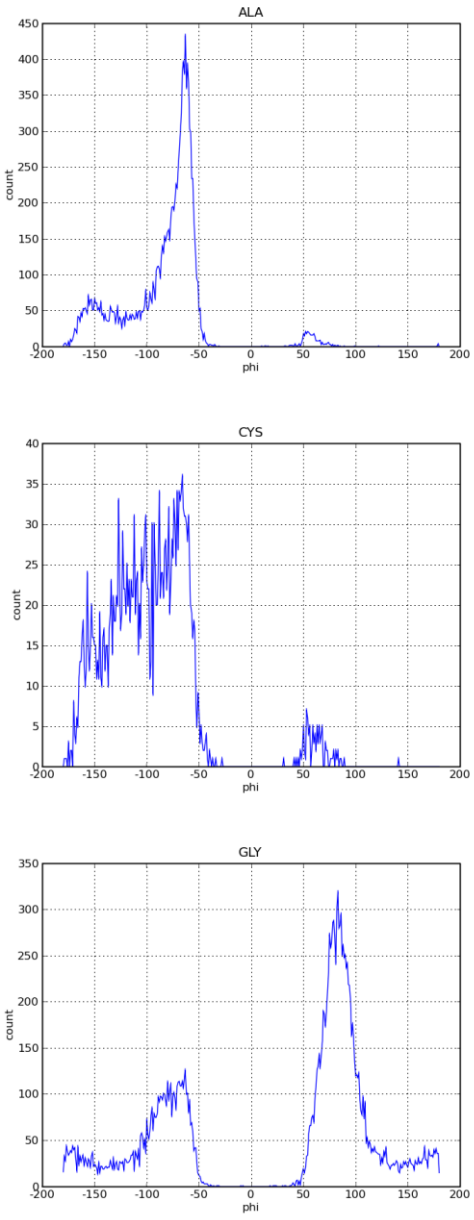


Figure 8. Distributions of Φ angle for ALA, CYS, and GLY

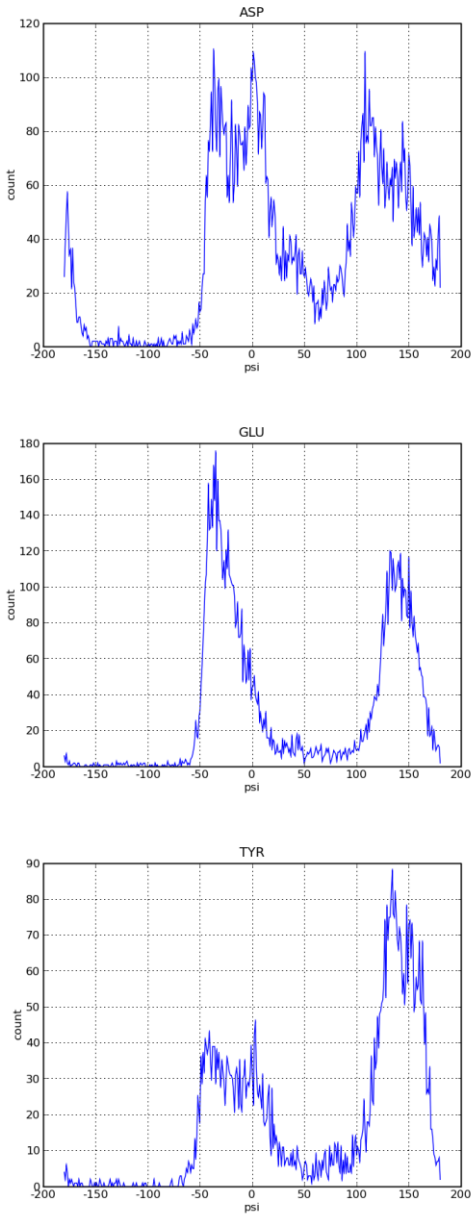


Figure 9. Distributions of ψ angle for ASP, GLU, and TYR

3.1.2 STATE PROBABILITIES

The pair wise dependent probabilities of observed states of angles $P_X(\phi_i,\psi_i)$, $P_X(\psi_i,\omega_i)$, and $P_{XY}(\omega_i,\phi_{i+1})$ are defined as

$$P_X(\phi_i,\psi_i)=\frac{N_X(\phi_i,\psi_i)}{\sum N_X}$$
$$P_X(\psi_i,\omega_i)=\frac{N_X(\psi_i,\omega_i)}{\sum N_X} \qquad (3.1)$$
$$P_{XY}(\omega_i,\phi_{i+1})=\frac{N_{XY}(\omega_i,\phi_{i+1})}{\sum N_{XY}}$$

where $N_X(\phi_i,\psi_i)$ is the number of residue type X observed in the indicated states, and $\sum N_X$ is the total number of conformations [19, 68]. Similarly, $N_{XY}(\omega_i,\phi_{i+1})$ is the number of dipeptides of XY in the given conformations. Here, $P_X(\phi_i,\psi_i)$ is the probability of observing residue X to be in state (ϕ_i,ψ_i), $P_X(\psi_i,\omega_i)$ is the probability of observing residue X to be in state (ψ_i,ω_i), and $P_{XY}(\omega_i,\phi_{i+1})$ is the joint probability of observing residue X in state (ω_i) and Y in state (ϕ_{i+1}). The neighbor-dependence introduced in the third of (3.1) is a dependence that originates from the residue type differences. Otherwise, (3.1) acknowledge the Flory isolated pair hypothesis. The dependence introduced by neighbor types is clearly seen in Figure 10, where the probabilities obtained by summing $P_{XY}(\omega_i,\phi_{i+1})$ over ω_i for $Y=\text{ALA}$ and $X=\text{TRP, GLY or PRO}$ are presented by the three curves which marked differences from each other.

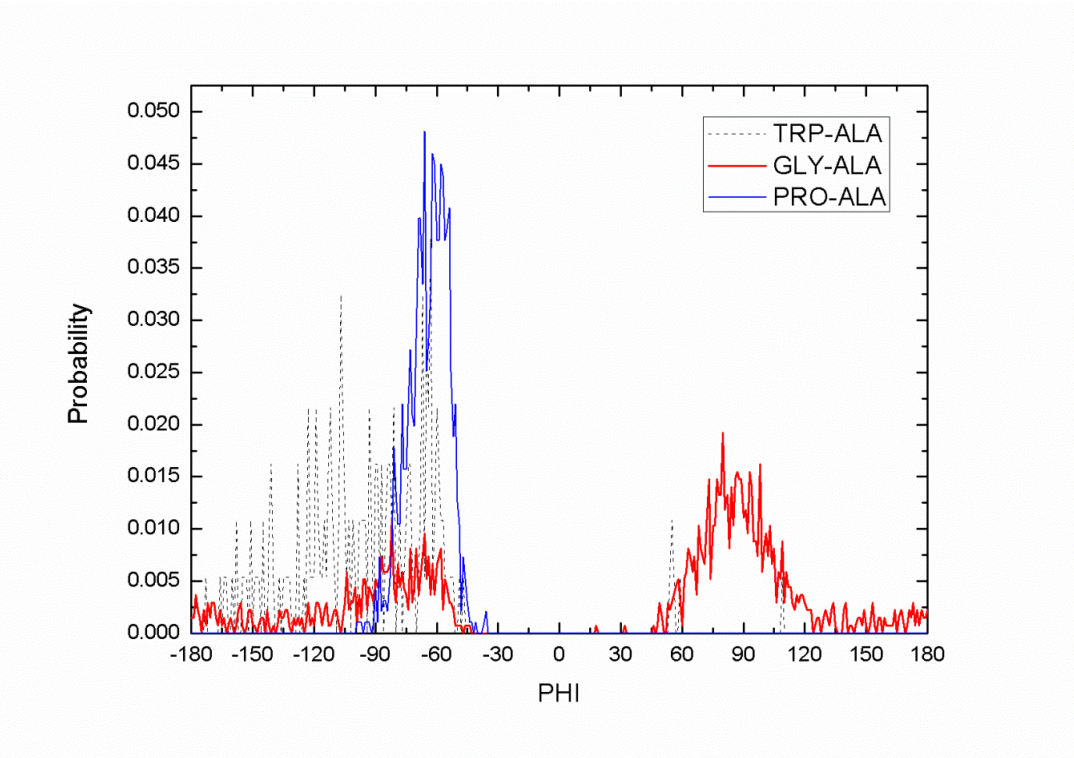


Figure 10. Probabilities obtained by summing $P_{XY}(\omega_i,\Phi_{i+1})$ over ω_i for Y=ALA and X=TRP,GLY, or PRO.

The conformational energies are defined as

$$E_X(\phi_i,\psi_i) = -RT\ln\left(\frac{P_X(\phi_i,\psi_i)}{P_X^0(\phi_i)P_X^0(\psi_i)}\right)$$
$$E_X(\psi_i,\omega_i) = -RT\ln\left(\frac{P_X(\psi_i,\omega_i)}{P_X^0(\psi_i)P_X^0(\omega_i)}\right) \qquad (3.2)$$
$$E_{XY}(\omega_i,\phi_{i+1}) = -RT\ln\left(\frac{P_{XY}(\omega_i,\phi_{i+1})}{P_{XY}^0(\omega_i)P_{XY}^0(\phi_{i+1})}\right)$$

where the superscript 0 indicates the uniform distribution probabilities. Hence, they are directly proportional to the size of the angular intervals of the states; $P_X^0(\phi_i) = P_X^0(\psi_i) = 1/13$ and

$P_X^0(\omega_i) = 1/2$. Statistical weights $u_{\phi_i\psi_i}$, $u_{\psi_i\omega_i}$, and $u_{\omega_i\phi_{i+1}}$ corresponding to the energies may be defined by

$$\begin{aligned} u_{\phi_i\psi_i;X} &= \exp\left(-E_X(\phi_i,\psi_i)/RT\right) \\ u_{\psi_i\omega_i;X} &= \exp\left(-E_X(\psi_i,\omega_i)/RT\right) \\ u_{\omega_i\phi_{i+1};XY} &= \exp\left(-E_{XY}(\omega_i,\phi_{i+1})/RT\right) \end{aligned} \tag{3.3}$$

where R is the gas constant, T is the temperature.

The statistical weight matrix for a configuration can be written as a product of statistical weights of each bond pair (ϕ,ψ), (ψ,ω), and (ω,ϕ). For this purpose, the statistical weight matrix for a given residue X is defined as $U_X^{\phi\psi} = \left[u_{\phi_i\psi_i}\right]_X$, $U_X^{\psi\omega} = \left[u_{\psi_i\omega_i}\right]_X$, and $U_{XY}^{\omega\phi} = \left[u_{\omega_i\phi_{i+1}}\right]_{XY}$. Depending on the number of states of each angle, dimensions of the statistical weight matrices $U_X^{\phi\psi}$, $U_X^{\psi\omega}$, and $U_{XY}^{\omega\phi}$ are 13×13, 13×2, and 2×13 respectively. The superscripts (ϕ,ψ), (ψ,ω), and (ω,ϕ) identify the bond pairs over which statistical weights are calculated.

3.1.3 Calculation Of The Thermodynamic Quantities

The statistical weight for a configuration specified by a set of rotational states of the several bonds of the chain may be written as a product of statistical weights of each bond pair. The complete set of all such products can be generated by matrix multiplication. The computational background that is developed in this section is based on and is an extension of Flory's macromolecules paper [12].

The partition sum of statistical weights for all configurations of the chain is given by

$$Z = J^* U_1^{\phi\psi} U_1^{\psi\omega} U_1^{\omega\phi} U_2^{\phi\psi} U^{\psi\omega}{}_2 \ldots U_n^{\phi\psi} U_n^{\psi\omega} J \tag{3.4}$$

where $J^* = \begin{bmatrix}1 & 0 & \ldots & 0\end{bmatrix}$, and $J = column\begin{bmatrix}1 & 1\ldots1\end{bmatrix}$.

The thermodynamic properties such as the entropy and the energy of a peptide, and the coefficients derived from them depend not only on a single conformation of the peptide, but on

all possible configurations. In the remaining equations, we give the relevant expressions for calculating these averages.

3.1.4 HELMHOLTZ FREE ENERGY

Since the Helmholtz free energy in canonical formalism is additive over the energies, it can be calculated using the partition function of the chain [13].

$$-\beta F = \ln Z \tag{3.5}$$

where $\beta = 1/kT$.

3.1.5 MEAN ENERGY

The average energy is given by

$$E = -\frac{d}{d\beta}(\ln Z) = -\frac{1}{Z}\frac{dZ}{d\beta} \tag{3.6}$$

The matrix multiplication formalism of the partition function leads to matrix multiplication scheme of its derivatives in the following way

$$\frac{dZ}{d\beta} = L^*\left(\prod \hat{U}_i\right)L \tag{3.7}$$

where $L^* = \begin{bmatrix} J^* & 0\ldots 0 \end{bmatrix}$, $L = column\begin{bmatrix} 0 & \ldots & 0 & J \end{bmatrix}$ and $\hat{U}$ is the super matrix whose elements are matrices

$$\hat{U} = \begin{bmatrix} U & U'_\beta \\ 0 & U \end{bmatrix} \tag{3.8}$$

$$U'_\beta = \frac{dU}{d\beta} \tag{3.9}$$

Therefore, the mean energy can be obtained using the following multiplication scheme

$$E = -\frac{1}{Z}L^*\left(\prod G_i\right)L \tag{3.10}$$

where

$$G_i = \begin{bmatrix} U & U'_\beta \\ 0 & U \end{bmatrix}_i \tag{3.11}$$

3.1.6 ENTROPY

The entropy of the chain can be expressed in terms of Z and its derivatives with respect to β. Following the equality $S/k = \beta^2 \, dF/d\beta$,

$$\frac{S}{k} = \beta^2 \left(\frac{1}{\beta^2} \ln Z - \frac{1}{\beta} \frac{\partial \ln Z}{\partial \beta} \right) = \ln Z + \beta E \tag{3.12}$$

is obtained.

Using the matrix multiplication formalism of Z and its first derivative with respect to β, the entropy can be calculated as

$$\frac{S}{k} = \ln\left(J^* \left(\prod U_i \right) J \right) - \beta \frac{L^* \left(\prod G_i \right) L}{J^* \left(\left[\prod U_i \right] \right) J} \tag{3.13}$$

3.1.7 HEAT CAPACITY

The heat capacity is one of the most important properties of the proteins, both native and denatured. When force acting on the chain is taken as zero, denoted below by the subscript $f = 0$, the heat capacity can be calculated as

$$C_{f=0} = \left(\frac{\partial E}{\partial T} \right)_{f=0} = k\beta^2 \frac{\partial^2 \ln Z}{\partial \beta^2} \tag{3.14}$$

Similar to (3.7), second derivative can be obtained as

$$\frac{\partial^2 Z}{\partial \beta^2} = M^* \left(\prod \hat{\hat{U}}_i \right) M \tag{3.15}$$

where $M^* = \begin{bmatrix} J^* & 0 \ldots 0 & 0 \ldots 0 & 0 \ldots 0 \end{bmatrix}$ and $M = column\begin{bmatrix} 0 \ldots 0 & 0 \ldots 0 & 0 \ldots 0 & J \end{bmatrix}$, and

$$\hat{\hat{U}} = \begin{bmatrix} U & U'_\beta & U'_\beta & U''_\beta \\ 0 & U & 0 & U'_\beta \\ 0 & 0 & U & U'_\beta \\ 0 & 0 & 0 & U \end{bmatrix} \tag{3.16}$$

$$U'_\beta = \frac{dU}{d\beta},\ U''_\beta = \frac{\partial^2 U}{\partial \beta^2} \tag{3.17}$$

It is possible to write the second derivative of $\ln Z$ on the right hand side of the equation in terms of the first and second derivatives of the partition function

$$\begin{aligned} \frac{\partial^2 \ln Z}{\partial \beta^2} &= \frac{\partial}{\partial \beta}\left(\frac{\partial \ln Z}{\partial Z}\frac{\partial Z}{\partial \beta} \right) \\ &= \frac{\partial}{\partial \beta}\left(\frac{1}{Z}\frac{\partial Z}{\partial \beta} \right) \\ &= -\frac{1}{Z^2}\left(\frac{\partial Z}{\partial \beta} \right)^2 + \frac{1}{Z}\left(\frac{\partial^2 Z}{\partial \beta^2} \right) \end{aligned} \tag{3.18}$$

The matrix multiplication form of Z and its derivatives allow the heat capacity to be calculated by the matrix notation

$$\begin{aligned} C_{f=0} &= k\beta^2 \left\{ -\frac{1}{Z^2}\left[L^* \left(\prod \hat{U}_i \right) L \right]^2 + \frac{1}{Z}\left[M^* \left(\prod \hat{\hat{U}}_i \right) M \right] \right\} \\ &= k\beta^2 \left\{ -\left[\frac{L^* \left(\prod \hat{U}_i \right) L}{J^* \left(\prod U_i \right) J} \right]^2 + \left[\frac{M^* \left(\prod \hat{\hat{U}}_i \right) M}{J^* \left(\prod U_i \right) J} \right] \right\} \end{aligned} \tag{3.19}$$

3.2 RESULTS

In this section, the free energy, energy, entropy, and heat capacity of peptides of different sizes ranging from 10 to 800 amino acids are calculated using the RIS model, over a temperature range of 200-700 K. Table 3 lists the representative protein set taken from the non-redundant PDB. The selection of the proteins is made according to their numbers of residues which varied evenly between 10 and 802.

Table 3. The protein dataset used in calculations

Number of Residues	Protein, Chain ID	Number of Residues	Protein, Chain ID
10	1FYN_B	349	3KM8_A
40	3E7R_L	408	2WKN_A
120	1C2A_A	456	3NPL_A
160	1CZT_A	545	3OTQ_A
200	1YKN_A	802	3IQM_A
226	3K6P_A		

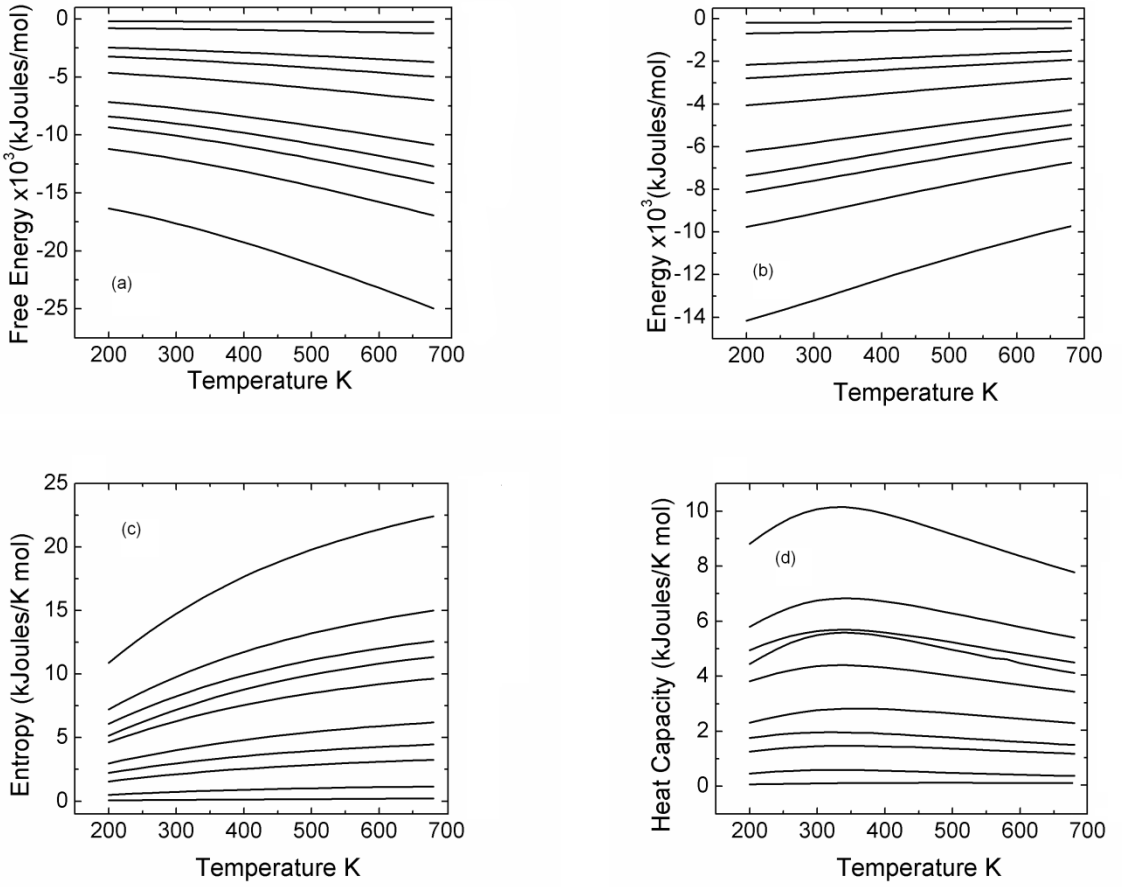


Figure 11. (a) The free energy as a function of temperature, T, for different length proteins. (b) energy as a function of T. (c) entropy as a function of T. (d) heat capacity as a function of T. The curves in parts (a),(b), and (c) are ordered from top to bottom represent proteins with the following numbers of residues: 10, 40, 120, 160, 226, 349, 408, 456, 545, and 802, respectively. In part (d) they are in reverse order.

The variation of the free energy, energy, entropy and heat capacity is evaluated by repeating the calculations over a temperature range of 200-700 K. Results are presented in Figure 11.

The curves shown in the four panels of Figure 11 are not independent from each other, and are related by the thermodynamic relations by equations (3.5), (3.6), (3.12), and (3.14). It is seen that the curves in the figures all scale with the number of residues N. In order to find analytical functions that will give the curves shown in Figure 11, we first chose an analytical form for the heat capacity as

$$C_{f=0}(T,N) = NT^3\left(Ae^{BT} + Ce^{DT}\right) \tag{3.20}$$

keeping in mind the thermodynamic postulates. We inspired the Debye model of heat capacity in a solid that shows the dependence of T^3. Then, by integration subject to the conditions imposed by (3.5), (3.6), (3.12), and (3.14), we obtain the remaining thermodynamic functions as given in equations (3.21)-(3.23) by curve fitting as

A :	1.5×10^{-6} kJoules/K^4 mol
B :	-7.2×10^{-3} 1/K
C :	2.6×10^{-5} kJoules/K^4 mol
D :	-2.3×10^{-2} 1/K
E :	–4083 kJoules/mol

$$S(T,N) = \frac{AD^3Ne^{BT}(B^2T^2 - 2BT + 2) + CB^3Ne^{DT}(D^2T^2 - 2DT + 2)}{B^3D^3} - 2N\left(\frac{A}{B^3} + \frac{C}{D^3}\right) \tag{3.21}$$

$$F(T,N) = -\frac{AD^3Ne^{BT}(B^2T^2 - 4T + \frac{6}{B}) + CB^3Ne^{DT}(D^2T^2 - 4T + \frac{6}{D})}{B^3D^3} + 2NT\left(\frac{A}{B^3} + \frac{C}{D^3}\right) + EN \tag{3.22}$$

$$U = F + TS = \frac{AD^3Ne^{BT}(B^2T^3 - 3BT^2 + 6T - \frac{6}{B}) + CB^3Ne^{DT}(D^2T^3 - 3DT^2 + 6T - \frac{6}{D})}{B^3D^3} + EN \tag{3.23}$$

In Figure 12, we compare the results obtained by (3.21)-(3.23) with the results of calculations for four peptides of sizes 20, 200, 408, and 802.

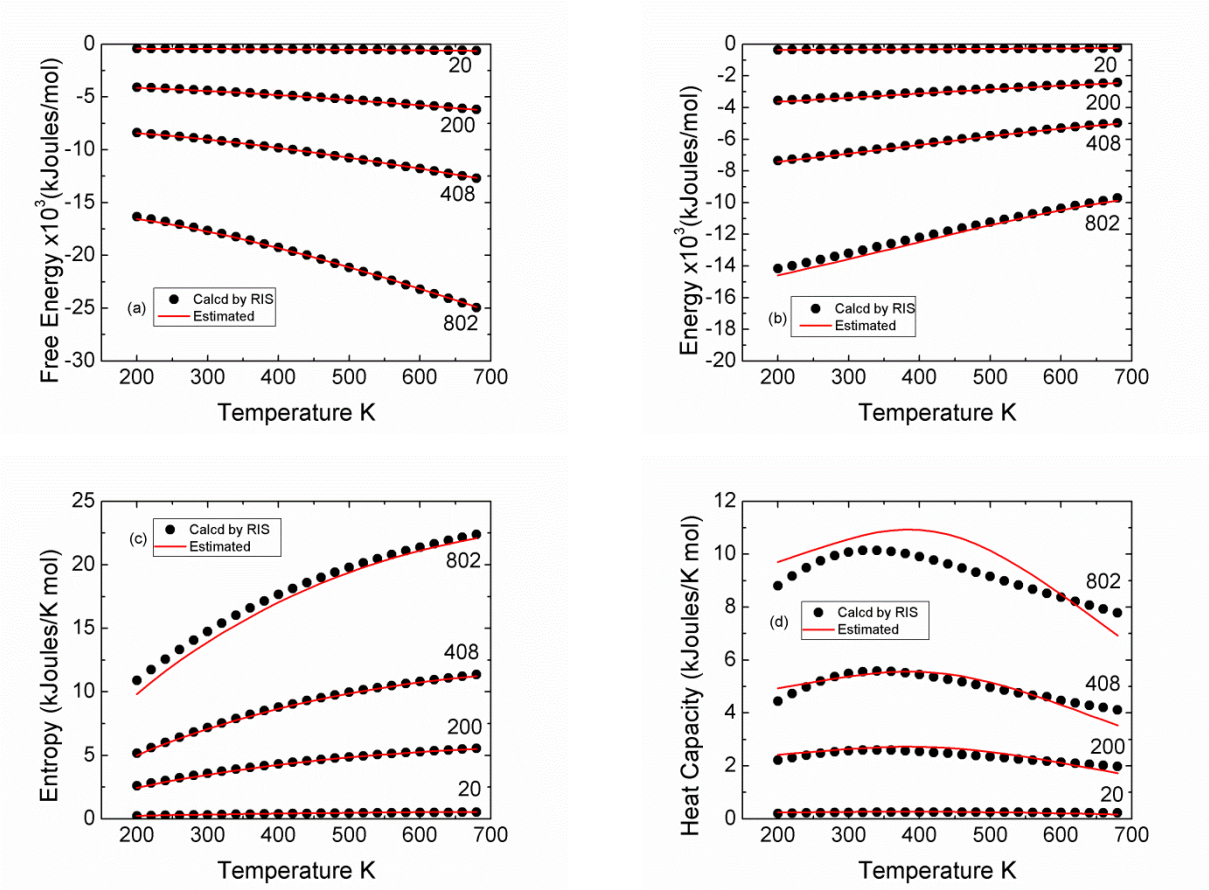


Figure 12. Comparison of (a) free energy, (b) mean energy, (c) entropy, and (d) heat capacity estimates. Exact values are calculated by matrix multiplication scheme, estimated values are calculated by fundamental relation. The lengths of chains are shown on each curve.

3.3 CONCLUDING REMARKS

We adopted the RIS model for calculating the thermodynamic functions of proteins in the random coil state. The use of the RIS model depends critically on two items: (i) the choice of the states, and (ii) the choice of the database with which the probabilities of these states are evaluated. The states are described in terms of the populated regions on the Ramachandran map, and the possible states for the ϕ and ψ angles of different amino acids are determined following the work of Karplus [23]. In order to apply the RIS model, however, the states available to the

torsion angles ϕ, ψ, and ω are required separately. The state space is obtained in our formulation as 13 states for ϕ and 13 states for ψ, and two states for ω. Evaluation of the probabilities follows the choice of the state space. For proof of principle, we used a coil library for the determination of the probabilities. One could alternatively construct a databank of known denatured proteins, or a subset of them depending on the nature of the investigation, and extract the probabilities from that state. Once the states are determined, the RIS model is independent of the databases used. We observed that the per residue thermodynamic properties of proteins in the random coil state scales only with the temperature. While entropy and energy increases with the temperature, free energy decreases. Heat capacity represents a decrease around 340 Kelvin that implies an energy barrier for a possible transition state. The explicit expressions that we determined for the thermodynamic functions form a thermodynamically consistent set which may be used to obtain other thermodynamic potentials by applying the known Legendre transformation techniques [13].

Chapter 4

PREDICTING MOST PROBABLE CONFORMATIONS OF A GIVEN PEPTIDE SEQUENCE IN THE RANDOM COIL STATE

In this chapter, we present a computational scheme for finding high probability conformations of peptides. The scheme calculates the probability of a given conformation of the given peptide sequence using the probability distribution of torsion states.

Backbone conformations of a protein can be described by $\phi - \psi$ torsion angles of the amino acids that specify the rotational freedom. It is already known that some values of torsion angles are more populated than the others on the Ramachandran map and these preferences are different for different types of amino acids [69]. Furthermore, preferences for a given residue are dependent on the states of the neighboring residues [7, 14, 15, 19, 67, 70-72]. In this work, the problem of predicting the high probability conformations of a protein using the rotational preferences obtained by knowledge-based approaches is discussed. Neighbor dependence is taken into consideration both in constructing the probabilities and in the generation of conformations.

Most of the computational work in predicting peptide conformations has been confined to secondary structure prediction for the native state. In the present work, we do not restrict the problem to the determination of secondary structure. We specifically focus on determining the high probability conformations of peptides in the random coil state. For this reason, we use a Coil Library for determining the probabilities [66]. Use of a secondary structure library would be more appropriate to use for the secondary structure prediction of a given sequence, which we do not pursue here.

The torsion state of a residue is defined by a region on the Ramachandran map and these regions are separated by energy barriers. Existing studies on torsion angle predictions use different number of states with different definitions. For example, Bystroff et al., took eleven states covering the full Ramachandran map in their hidden Markov model for local sequence-patterns in proteins (HMMSTR) [73]. Kuang et al. assumed a smaller number of states each of

which correspond to a large number of states [74]. They used four major torsion states occupying 80% of the Ramachandran map and achieved 77.3% accuracy in predicting torsion angles using SVM method. Zimmermann and Hansmann emphasized that the torsion state prediction would provide more structural information than secondary structure predictions since the secondary structure has a complex definition [75]. Estimated from the work of Lovell et al. [76], they introduced ten regions. However, due to few number of samples, they based their results on three regions (right handed alpha helix, beta strand, and outside of these regions). The achieved 3-state accuracy is around 82% using SVMs. Xin et al. presented two probabilistic methods (MEMM and CRF) to predict torsion angles using sequence profiles of residues and used fifteen regions proposed by Shortle that covers 43% of the Ramachandran map to define different coarse-grained classes [77, 78]. The achieved prediction accuracies (that is below 70%) of each class are given.

Previous studies all provide the prediction of the most probable torsion state for each residue and do not consider a probability distribution. However, instead of predicting only the most probable torsion state or the secondary structure, predicting the probability distribution of the states or secondary structures is an approach that may be useful for sampling the conformational space and may provide a detailed analysis of the space [79-81]. Along this line, Helles and Fonseca proposed an analysis of probability distribution of torsion angles of coil residues using neural networks [79]. They calculated the distribution of dihedral angles for $30^\circ \times 30^\circ$ bins and showed that the prediction accuracy increases as the number of considered bins increases. Helles and Fonseca showed that the neural network model they adopted outperforms the basic statistical method that simply predicts the most populated bin in the database.

In this work, our aim is to find a probability distribution of the torsion states of a given peptide sequence in the random coil state. We adopt the Rotational isomeric States (RIS) formalism from polymer theory [1] for calculating the probability distribution. Here, we identify eleven torsion states following the work of Karplus [23] that indicates the distinct regions not only for helices or strands but also for loops, beta bridges, etc., (See the notation in References [23, 51, 68]). This knowledge provides an efficient conformational sampling. We derive the probability distribution based on a Hidden Markov Model (HMM) and the Viterbi algorithm that has been described in Section 2.4. Dependence of the conformational probability of a given

residue on the type and conformation of its neighbors is taken into consideration in constructing the probabilities, in the RIS calculations and in the Viterbi algorithm. Our aim is similar to that of Helles and Fonseca. We depart in the method of calculations, where their approach is based on a neural network and ours on the generation of the partition function of a chain by the matrix multiplication scheme. Our ultimate aim, however, is to determine the most probable conformation and an ensemble of conformations that are of high probability. Recently, ensembles of intrinsically disordered proteins (IDP) have been constructed by Daughdrill et. al [82], which is similar to our approach in spirit but different in methodology, where our high probability conformations are generated using an appropriate library, the RIS scheme, a Hidden Markov Model and the Viterbi algorithm with multistep backtracking [50]. As we explain in detail in the following section, we base the present calculations on a coil library. However, choice of an IDP library would clearly lead to the generation of high probability conformations of IDP's, which are known to have statistical differences than random coils as pointed out by Schweitzer-Stenner [83].

The number of occurrences of the residues in the databank in the defined states leads to the corresponding probabilities. The determination of high probability conformations of a given peptide sequence of length N is a computationally difficult problem. Considering a state space of eleven torsion states leads to 11^N possible conformations from which one has to choose the optimum one. The Markov assumption and the use of the Viterbi algorithm simplify the problem significantly.

4.1 METHODS AND MATERIALS

4.1.1 KNOWLEDGE BASED DATABASE

Since we consider the conformations as random denatured forms and since the Coil Library stores the fragments from the disordered conformations that are neither alpha-helix nor beta-strand, the Coil Library is a good source for calculation of the probabilities [66]. We used Coil Library given by the Rose group http://www.roselab.jhu.edu/coil/, September 2011 version. The library contains fragments that have less than 20% sequence identity, better than 1.6 Ångstrom resolutions and a refinement factor of 0.25 or better and contains 24112 fragments extracted from

the Protein Data Bank. The library contains segments obtained after removal of secondary structures, the segments that are classified as alpha helix or beta strand and the one-residue coils. We remove the chains including UNK or ASX types of residues or the chains with less than 3 residues. The remaining set contains 261548 residues. The fragment size distribution is given Figure 13. The probabilities depend on the choice of the state space. Since the Coil Library is a good representative set of proteins in the coiled state and our interest is the coiled conformations of proteins, occurrence probabilities of residues in torsion states are calculated over the coil library.

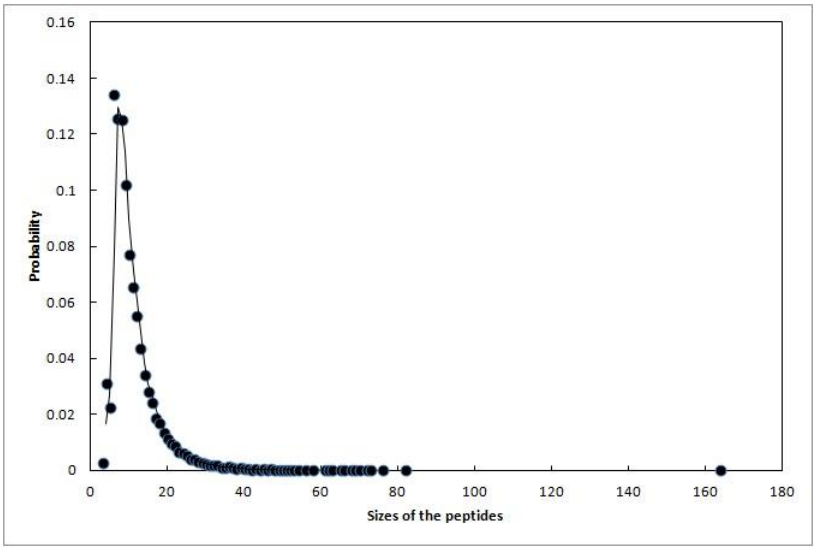


Figure 13. Distribution of the sizes of the peptides in the Coil library. Total number of peptides is 24057.

4.1.2 TORSION STATES

We use the eleven torsion states introduced by Karplus [23]. The descriptions of the states are given in Table 4.

Table 4. Description of the twelve torsion states

State 1	ε'	Mirror image of the extended region ε
State 2	ε	The extensive regions, $\phi > 0,\ \psi \approx \pm 180^{\circ}$
State 3	α_R	Right-handed alpha helix

State 4	γ	Tight turn region
State 5	δ_R	The right handed bridge region between two β-strands
State 6	δ_L	Mirror image of the δ_R region
State 7	ζ	Region observed mostly in residues preceding Pro
State 8	γ'	Inverse tight region
State 9	α_L	Mirror image of α_R
State 10	β_S	Extended beta sheet forming region
State 11	β_P	Region with extended polyproline-like helices

4.1.3 METHODS

It has been previously shown that the torsion preferences of a residue depend on the type and state of neighboring residues [14, 19, 67]. In Figure 14, the left neighbor dependence of Arginine is shown as an example. The frequencies are plotted in the 30×30 degree regions of Ramachandran map. Each 30 degree ψ region is shown in a different color in order to increase the clarity. For certain pairs, neighbor dependence of state probabilities are significant and are included into our calculations. We treat the problem of the prediction of torsion state as a Markov process and consider the correlations between the near neighbor residues.

In the following sections, we first define the Hidden Markov model of the problem and then we provide the calculation of the parameters of the HMM in detail. We then introduce the Viterbi algorithm for searching the state sequence that has the maximum probability of occurrence. The probability distribution of the torsion states of the residues is determined using the Viterbi algorithm. Lastly, we define multistep backtracking algorithm that is used to guess the high probability conformations.

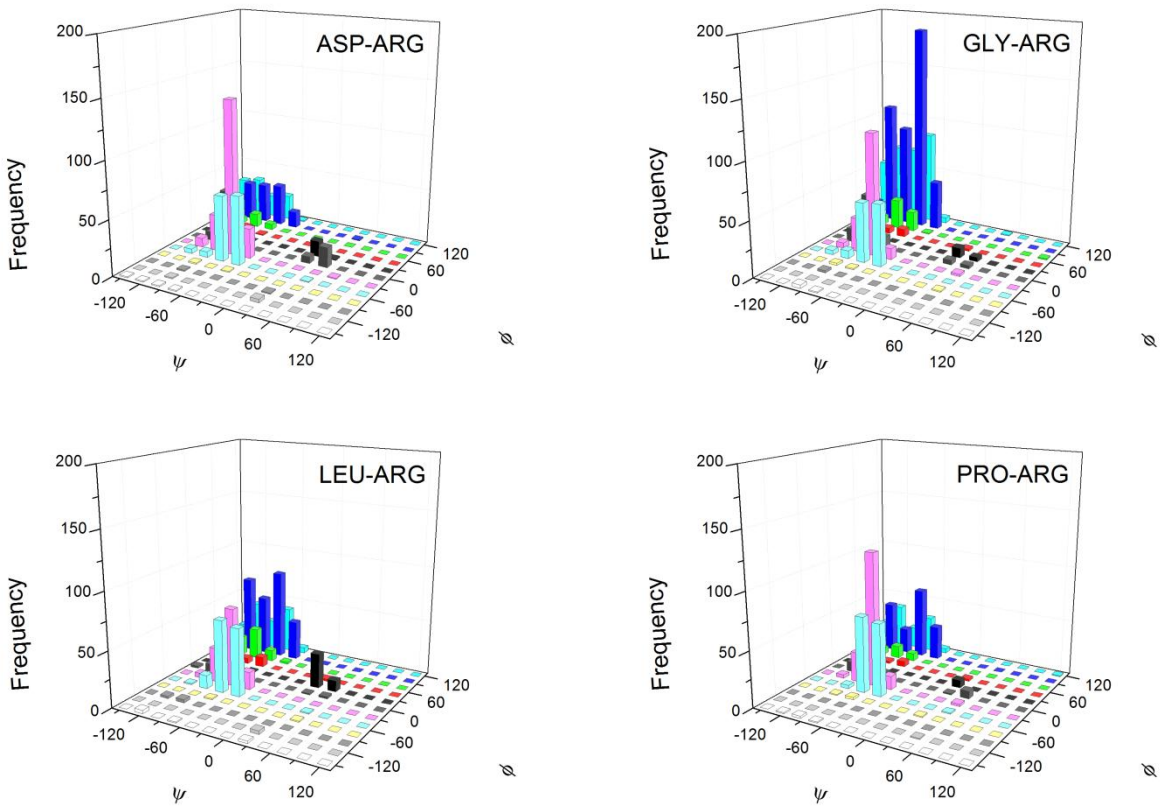


Figure 14. Arginine has different $\phi-\psi$ distributions depending on the previous near neighbor.

4.1.4 HIDDEN MARKOV MODEL

An introduction to hidden Markov models is given in 2.4. Here we describe the model and its parameters in terms of our problem. In a hidden Markov model there is an observed sequence of emitted symbols which is defined by $O=O_1,O_2,O_3,\ldots$, and a sequence of states visited which is defined by $Q=q_1,q_2,q_3,\ldots.$ Further information on Hidden Markov models can be found in [50, 51]. The observed sequence of our model is the given peptide sequence and the torsion state sequence is the hidden state sequence to be determined. The definitions of components of the model are given below:

1) $A=\{a_1,a_2,\ldots,a_{20}\}$ The set of 20 amino acids
2) $S=S_1,S_2,\ldots,S_{11}$ The set of 11 torsion states

3) $P = (p_{ij})$ The transition probability matrix where $p_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$ that is the probability of t+1st amino acid is in state S_j given that tth amino acid is in the state S_i.

4) $b_i(a) = P(S_i \text{ emits symbol a})$ for each S_i and a, The emission probabilities that is the probability of state S_i emits amino acid a..

5) $\Pi = (\pi_i)$ An initial distribution vector where $\pi_i = P(q_1 = S_i)$

4.1.5 EVALUATION OF A PRIORI PROBABILITIES

Here we adopt the transition probabilities for each dipeptide of the given sequence and the emission probabilities of each residue from the Coil Library.

The pair wise dependent probabilities $P_{XY}(S_i, S_j)$ where S_i is a torsion state defined by the two torsion angles (ϕ, ψ) and $i, j = 1, 2, \ldots, 11$ are determined according to

$$\mathrm{P}_{XY}(S_i, S_j) = \frac{N_{XY}(S_i, S_j)}{\sum N_{XY}} \tag{3.24}$$

Here, $N_{XY}(S_i, S_j)$ indicates the number of residue pairs observed having the indicated values of the argument. The term $\sum N_{XY}$ in the denominator is the total number of observed dipeptides of XY in all possible states. The probabilities calculated in this manner are the *a priori* probabilities; because the conditionality of the dipeptide under consideration that being embedded into the given specific amino acid sequence of the protein has not been used.

We define a conformational energy for a given residue Y in the dipeptide XY along the primary sequence of the protein as

$$E_{XY}(S_i, S_j) = -RT \ln\left(\frac{\mathrm{P}_{XY}(S_i, S_j)}{\mathrm{P}_X^0(S_i)\,\mathrm{P}_Y^0(S_j)} \right) \tag{3.25}$$

where the superscript zero denotes the uniform distribution probabilities, i.e., those valid when all angles are equally probable that is $1/11$.

In this study, we use the rotational isomeric state formalism in which each residue is treated as occurring in one or another of several discrete torsion states to obtain the statistics of the chain [1].

4.1.6 CALCULATION OF A POSTERIORI PROBABILITIES

The statistical weight matrix $U_i = \left[u_{\eta\zeta} \right]$ for a given residue pair $i-1$ and i is determined by statistical weights $u_{\eta\zeta}$ corresponding to the energies $E_{\eta\zeta}$ following the work of Flory [1]. The details these definitions are given in 2.2.

$$u_{\eta\zeta;i} = \exp(-E_{\eta\zeta;i} / RT) \tag{3.26}$$

So, *a priori* probabilities led us to statistical weights that would be the primary quantities for characterizing the partition function. The partition function, Z, of the chain of N repeat units is given by

$$Z = J^* \left[\prod_{i=2}^{N} U_i \right] J \tag{3.27}$$

where $J^* = [1 \quad 1 \quad \cdots \quad 1]$, and $J = column[1 \quad 1 \quad \cdots \quad 1]$ of order 1×11 and 11×1 respectively. We choose all elements of J^* as ones to allow all possible states in the first residue as opposed to Flory's notion, $J^* = [1 \quad 0 \quad \cdots \quad 0]$, that fixes the first residue's state as the first state.

The probability $p_{\eta;i}$ that residue i will be in state η is estimated as the fraction of the sum of the statistical weights for all configurations for which residue i is in state η over partition function Z

$$p_{\eta;i} = Z^{-1} J^* \left[\prod_{m=2}^{i-1} U_m \right] U'_{\eta;i} \left[\prod_{m=i+1}^{N} U_m \right] J \tag{3.28}$$

Here, $U'_{\eta;i}$ is the matrix obtained by equating the entries of all of its columns to zero except those of column η.

Similarly, the joint probability $p_{\eta\zeta;i-1,i}$ that residue $i-1$ is in state η and residue i is in state ζ simultaneously is given by

$$p_{\eta\zeta;i-1,i} = Z^{-1} J^{*} \left[\prod_{m=2}^{i-1} U_m \right] U'_{\eta\zeta;i} \left[\prod_{m=i+1}^{N} U_m \right] J \tag{3.29}$$

where $U'_{\eta\zeta;i}$ is the matrix obtained by vanishing all elements of U_i with the exception of $u_{\eta\zeta}$. The transition probability $p_{\eta|\zeta}$ is simply the conditional probability that residue i will be in state ζ, given that residue $i-1$ is in state η is determined as the quotient of the joint probability, divided by $p_{\eta;i-1}$.

$$p_{\eta|\zeta} = \mathrm{P}(q_i = \zeta \mid q_{i-1} = \eta) = \frac{p_{\eta\zeta;i-1,i}}{p_{\eta;i-1}} \tag{3.30}$$

4.1.7 THE VITERBI ALGORITHM

Using the Viterbi algorithm is an efficient way of calculating the most probable state sequence. The problem is to find a state sequence $Q = q_1, q_2, \ldots, q_N$ given an observed sequence $O = O_1, O_2, \ldots O_N$ that maximizes $P(Q \mid O)$. That is $\arg\max_Q P(Q \mid O)$ meaning the given amino acid sequence takes its most probable conformation. Here, argmax denotes the maximum over the full state sequence set Q of the arguments of the probability function. In the first part of the algorithm, we obtain $\max_Q P(Q \mid O)$. For arbitrary t and i,

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} \mathrm{P}(q_1, q_2, \ldots, q_{t-1}, q_t = S_i \text{ and } O_1, O_2, \ldots, O_t) \tag{3.31}$$

is defined. That is the maximum probability of all ways to end in state S_i at t^{th} amino acid and given the amino acids $O_1, O_2, \ldots O_t$.

$$\delta_1(i) = \mathrm{P}(q_1 = S_i \text{ and } O_1) \tag{3.32}$$

Then, the joint probability of Q and O is maximized over all Q as

$$\max_{Q} \mathrm{P}(Q \text{ and } O) = \max_{i} \delta_N(i) \tag{3.33}$$

Since the maximum conditional probability is given as

$$\max_{Q} \mathrm{P}(Q \mid O) = \max_{Q} \frac{\mathrm{P}(Q \text{ and } O)}{\mathrm{P}(O)} \tag{3.34}$$

and $P(O)$ is fixed for all Q,

$$\max_{Q} \mathrm{P}(Q \mid O) = \max_{Q} \mathrm{P}(Q \text{ and } O) \tag{3.35}$$

Hence,

$$\arg\max_{Q} \mathrm{P}(Q \mid O) = \arg\max_{Q} \mathrm{P}(Q \text{ and } O) \tag{3.36}$$

a) First part of the algorithm

$$\delta_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le 11 \tag{3.37}$$

In the initialization step, δ_1 is calculated as an array of size eleven holding the probabilities of eleven torsion states for the first amino acid O_1 of the given peptide sequence. The probability δ_i is calculated by multiplying the initial probability of first residue is in state *i* with the emission probability that state *i* emits the first residue. The elements of initial probability vector are chosen as 1 so that each state has equal probability. Then, we calculate all δ_t's inductively.

$$\delta_t(j) = \max_{1 \le i \le 11} \delta_{t-1}(i)\, p_{ij}\, b_j(O_t), \quad 2 \le t \le N,\ 1 \le j \le 11 \tag{3.38}$$

Here N denotes the number of residues of the given peptide sequence.

b) Second part of the algorithm

The states q_i's are obtained by backtracking technique that searches for the path with the highest probability. We define

$$\psi_N = \arg\max_{1 \le i \le 11} \delta_N(i) \tag{3.39}$$

and determine the most probable state of last residue as $q_N = S_{\psi_N}$. The remaining states are then obtained by successive backtracking

$$\psi_t = \arg\max_{1 \le i \le 11} \delta_t(i)\, p_{i\psi_{t+1}} \tag{3.40}$$

and then equating $q_t = S_{\psi_t}$.

Example

We may consider a simple example to visualize the steps of the Viterbi algorithm. Assume that we have an observed sequence of amino acids ALA-GLY-VAL of length $N = 3$. Let $S = \{\alpha, \beta, \gamma\}$ be the set of torsion states. The scheme may be seen in Figure 15.

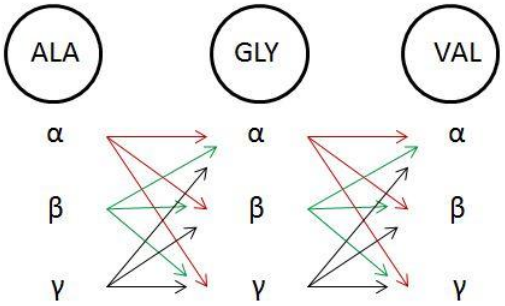


Figure 15. The hidden and observed states for a toy example. The arrows show the probability of transition from one hidden state to another.

In the first part of the algorithm we calculate the partial probabilities, δ's, and partial best paths. For each intermediate and terminating state, there is a specific most probable path to that state. For each of the three states of VAL the most probable paths may be found as given in Figure 16.

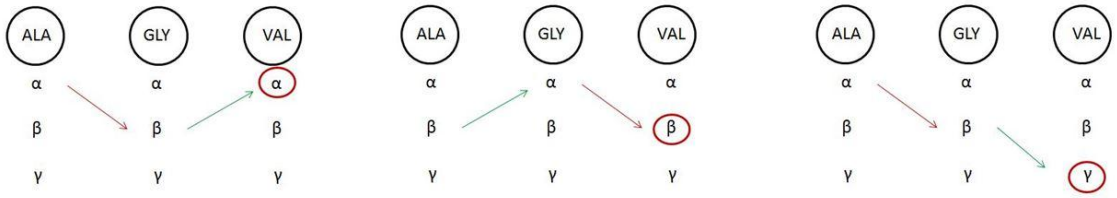


Figure 16. $\delta_3(i)$ are shown for i=1,2,3. That is the maximum probability of all sequences ending at state i at time t=3 and the partial best path is the sequence which achieves the maximum probability.

In the second part, we search for the most probable state sequence given the observed sequence after calculating the partial probabilities $\delta_t(i)$. For each state we hold a back pointer ψ_t via finding the states that maximize the probability.

4.1.8 Conformations Of Lower Probabilities

Changing preferences during backtracking allows for the generation of conformations with lower probabilities. Thus, for the t'th residue, we define

$$\psi_t^{(k)} = \arg\max_{1\le i\le 11}^{(k)} \delta_t(i)\, p_{i\psi_{t+1}} \tag{3.41}$$

where the superscript (k) denotes the kth maximum value. Here, $k=1$ refers to the maximum value, $k=2$ refers second maximum, etc. For each residue, besides the most probable state, we also determine the $2^{nd}, 3^{rd}, \ldots, 11^{th}$ maximum probable state. We use this definition to determine conformations of lower probabilities during the backtracking stage. We refer to this method of generating conformations of lower probabilities as multistep backtracking.

4.1.9 Multistep Backtracking

In this algorithm, we keep the n conformations with highest probability. The multistep backtracking algorithm proceeds as follows:

1. We define

$$\psi_N(j) = \arg\max^{(j)}_{1\le i\le 11} \delta_N(i), \quad 1 \le j \le 11$$
$$\lambda_N(j) = \max^{(j)}_{1\le i\le 11} \delta_N(i), \quad 1 \le j \le 11 \qquad (3.42)$$

where the superscript (j) denotes the jth maximum value. In λ_N, we keep the probabilities of states of the last residue in decreasing order; in ψ_N we keep the corresponding states. Then,

2. For the remaining amino acids we define

$$\psi_t(j) = \arg\max^{(j)}_{1\le i\le 11} \lambda_{t+1}(i)\delta_t(j)\, p_{j\psi_{t+1}(i)}, \quad 1 \le t<N$$
$$\lambda_t(j) = \max^{(j)}_{1\le i\le 11} \lambda_{t+1}(i)\delta_t(j)\, p_{j\psi_{t+1}(i)}, \quad 1 \le t<N \qquad (3.43)$$

Here, we calculate the probability of the state sequence of the region going backward from N^{th} to the t^{th} residue by multiplying the probability of the state sequence of the region from N^{th} to the $t+1^{st}$ residue by the transition probability from t^{th} residue to $t+1^{st}$ residue. The probabilities are stored in decreasing order.

There are a total of 11^N possible conformations for the sequence of N residues. In order to keep the number of conformations to be considered at a manageable level, we apply some pruning techniques. (i) We remove the state sequences with zero probability of occurring to decrease the complexity. Those are the impossible paths of the HMM. (ii) At each t'th step, if the number of stored sequences becomes larger than a specified threshold value m, we remove the state sequences those have lowest probabilities and continue with the m sequences with highest probability. The prediction accuracy increases if one keeps the number of stored paths, m, large. On the other hand, it is not feasible when considering long peptides (number of residues > 20).

For an amino acid sequence $O = a_1 - a_2 - a_3 - a_4 - a_5 - a_6$ of length 6, assume that the size of the state set is 3, $S = \{s_1, s_2, s_3\}$. Then, to find $n = 2$ most probable state sequences we may select $m = 4,\ m > n$. Then, we keep a list of maximum probable state sequences and a list of occurrence

probabilities of the corresponding state sequences. A schematic representation of the example may be given as follows

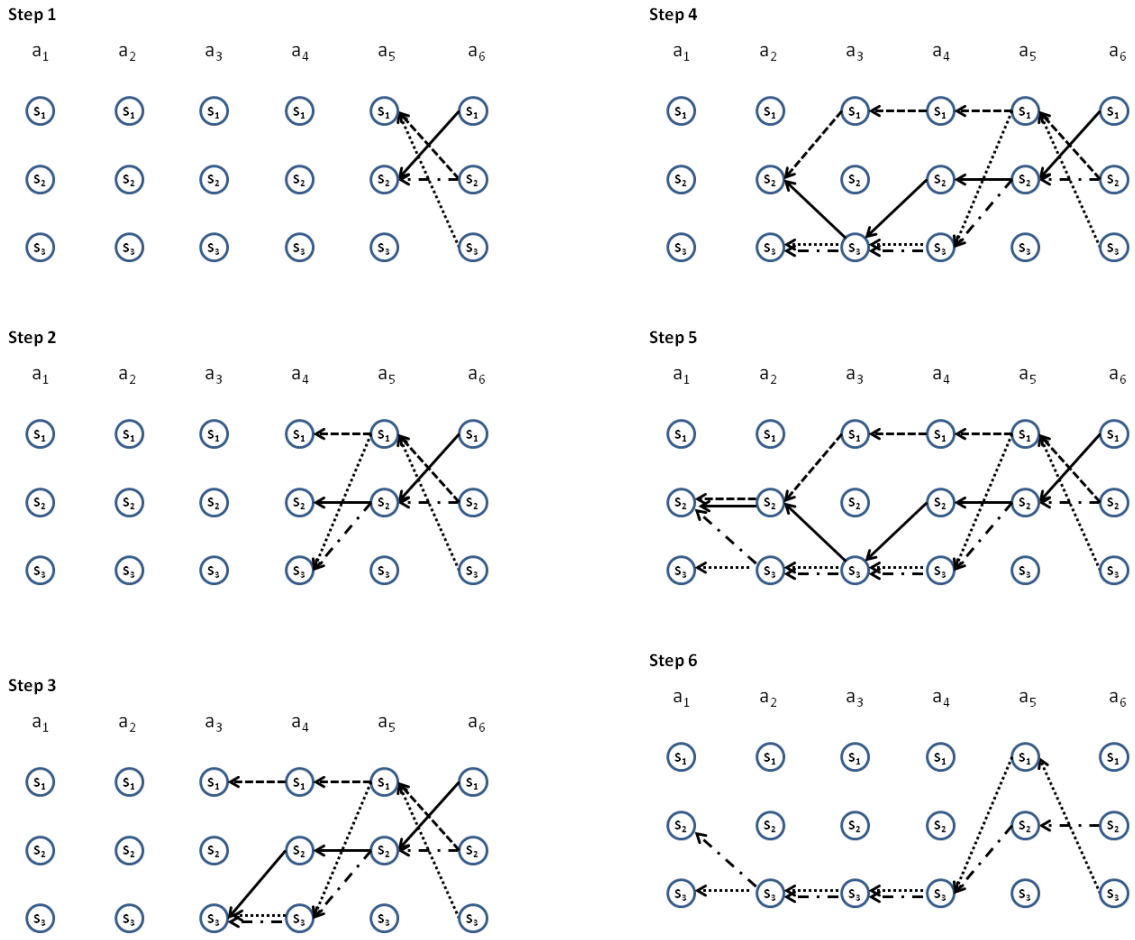


Figure 17. The steps of multistep backtracking for finding the 2 conformations with highest probability of an observed sequence of length 6.

4.2 RESULTS AND DISCUSSION

Using equation (3.41), we calculate the probability distribution of the eleven states for each amino acid in a given peptide sequence. Thus, given a peptide sequence of length N, and an amino acid at the k^{th} position, we can estimate the probability of its observed state, with $1 \le k \le N$. Prediction accuracy is defined as the fraction of amino acids that are predicted to be in

the most probable state. More specifically, it is calculated as the number of correctly predicted residues of type X in the predicted most probable state sequences using the Viterbi algorithm, over total number of residues of type X .

$$\text{prediction accuracy} = n_{\text{correct}} / n_{\text{total}} \tag{3.44}$$

Most populated states are determined as the number of residues of type X in the most populated state in the Coil library over total number of residues of type X .

We ran 5-fold cross validation for training and testing. The dataset is randomly divided into five subsets, then $4/5$ of the fragments are used to obtain training parameters, and $1/5$ are used for testing. The procedure is repeated five times, and the prediction accuracy is calculated by averaging the five accuracies. The training parameters, the transition and the emission probabilities, are derived each time by equations (3.24)-(3.30) using the corresponding training set. Then, the torsion states are predicted using (3.37)-(3.41) for each peptide in the corresponding test set. Lastly, the prediction accuracies determined for each test set and the average is calculated.

Table 5. Prediction accuracies for each residue type

Amino Acid Type	Most Populated	Viterbi	Improvement
ALA	24.32	31.96	31.41
GLU	28.67	34.01	18.61
LYS	26.22	33.67	28.40
MET	33.72	35.21	4.40
GLN	26.77	32.39	20.99
ARG	29.72	33.2	11.71
LEU	31.66	35.51	12.18
PHE	41.14	41.09	0.00
TYR	40.65	41.68	2.54
TRP	35.58	34.69	0.00
CYS	38.71	39.41	1.81
HIS	32.36	34.74	7.35

ASN	28.24	27.9	0.00
ASP	28.96	35.66	23.15
SER	28.23	35.21	24.74
THR	35.04	38.66	10.33
ILE	55.43	55.8	0.67
VAL	55.59	55.83	0.44
GLY	34.58	38.6	11.62
PRO	55.36	59.93	8.25
Average	35.55	38.76	9.03

In Table 5, the prediction accuracies of each amino acid type using the Viterbi algorithm are provided. Simply, we consider the improvement gained by using the Viterbi algorithm comparing to guessing the most populated state in the library. Since the near neighbor effects are taken into account using the Viterbi algorithm, prediction accuracies of some amino acid types are considerably higher. This comparison approach is first applied by Helles and Fonseca to determine the accuracy of neural networks method that is used to predict dihedral angle probability distribution for protein coil residues [79]. They compared the accuracies with the probabilities of guessing the most populated state in the data set for each amino acid type. Here, we show the performance of the Viterbi algorithm comparing the accuracies with the most populated state percentages for each amino acid type in a similar manner.

4.2.1 Consideration Of Less Probable States Of Residues

It is possible to generalize the prediction accuracy definition as the number of correctly predicted residues of type X in the k^{th} most probable state sequences using the Viterbi algorithm, over total number of residues of type X.

$$\text{prediction accuracy } (k) = n_{\text{correct}}^{(k)} / n_{\text{total}} \tag{3.45}$$

Then, we may add up the prediction accuracies and define the success rate as

$$\text{success rate } (i) = \sum_{k=0}^{i} \text{prediction accuracy } (k) \tag{3.46}$$

Since we know the probability of each state for each residue in a given peptide sequence, we can tell the probability of seeing the residue in either the most or the next most probable states, and so on. In Figure 18, we present the improvements when considering not only the most probable states but also the second, third, etc., most probable states. Increasing values of the abscissa values correspond to considering less and less probable states in the prediction.

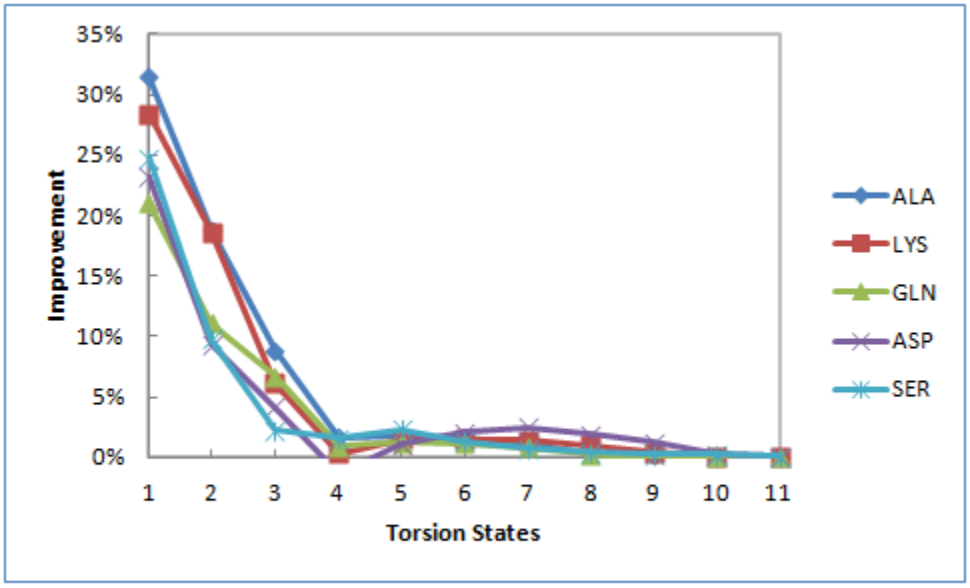


Figure 18. Percentage of improvements of Ala, Lys, Gln, Asp, and Ser are shown.

In Figure 19, we present the success rates of each amino acid type. The well predicted amino acids are proline, valine, and isoleucine. On the other hand, the small amino acids glycine, and asparagine shows lower rates of success. It is observed that the prediction method works better on hydrophobic residues. For example the three hydrophobic residues, Pro, Val, and Ile have $\phi-\psi$ distributions that are sharply peaked. This decreases the total number of accessible regions in the Ramachandran plot. Majority of the states have zero probability leaving behind only a small region over which the predictions are to be made. The better prediction of hydrophobic residues may be related to this characteristic distribution on the Ramachandran map.

In Figure 20, we present the success rates of each torsion states. Clearly, prediction of the torsion states that represents the tight turn regions (γ and γ') is not successful. This is not surprising because these states are the rarest torsion states in the database. The well predicted states are $\beta_S, \delta_L, \varepsilon, \delta_R,$ and β_P.

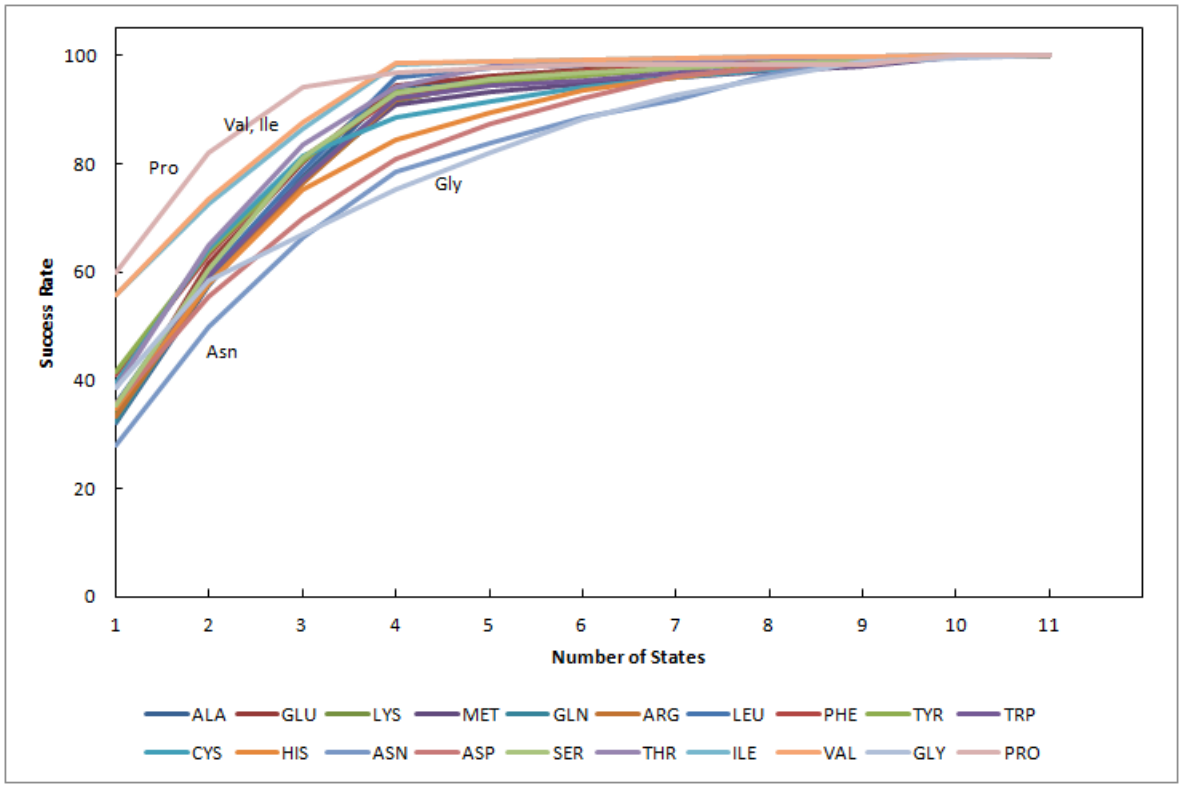


Figure 19. Success rates of each amino acid types. x axis shows the number of included states.

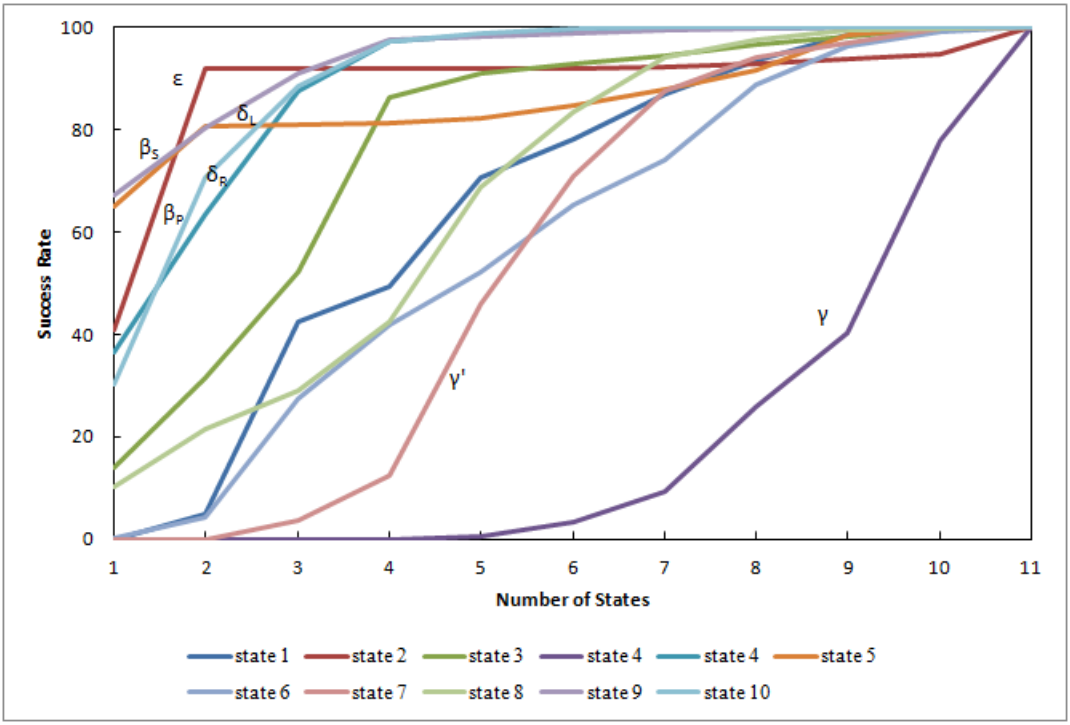


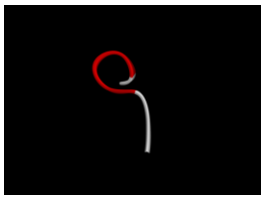
Figure 20. Success rates of each torsion states. x axis shows the number of included states.

4.2.2 Determination Of High Probability Conformations

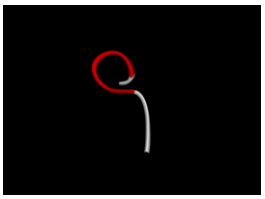
The most probable torsion state sequence is predicted using the Viterbi algorithm. The transition and emission probabilities are calculated from the Coil database. In the first step of the Viterbi algorithm, the state sequences (Q's) that maximize $P(Q|O)$ given the peptide sequence O are determined. Then, in the second step, the sequence $Q = q_1, q_2, \ldots, q_N$ that has the maximum probability is backtracked where each q_i represents the torsion state of the i^{th} residue. Therefore, after the first step, we may also designate a list of q_i's for each residue ordered by the probability of occurring. However, our main goal is to find most probable "n" conformations of a given peptide sequence. Hence, using the Viterbi algorithm with multistep backtracking described in Methods section we derive most probable "n" conformations of the peptide.

As an example, consider the sequence "GLN-VAL-CYS-ALA-ASN-PRO-GLU-LYS-LYS-TRP". The most probable twelve conformations are predicted as shown in Figure 21. It is clear that some conformations repeat although the torsion states of some residues in the state sequences

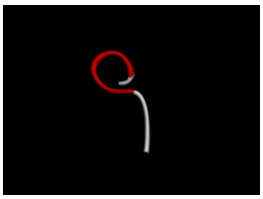
are different. Examination of the states listed below each figure shows that the differences are all in the first residue, which results in the similarity of the conformations. The repeating conformations indicate high entropy for that conformation.



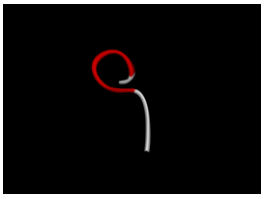
1) β_S-β_S-β_S-β_P-α_L-β_P-α_R-α_R-α_R-β_S



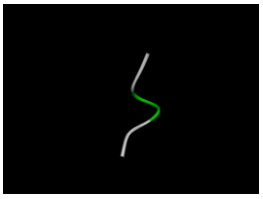
2) δ_R-β_S-β_S-β_P-α_L-β_P-α_R-α_R-α_R-β_S



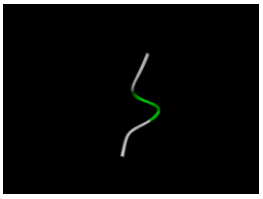
3) β_P-β_S-β_S-β_P-α_L-β_P-α_R-α_R-α_R-β_S



4) α_R-β_S-β_S-β_P-α_L-β_P-α_R-α_R-α_R-β_S



5) β_S-β_S-β_P-β_P-β_P-α_R-α_R-α_R-β_S-β_S



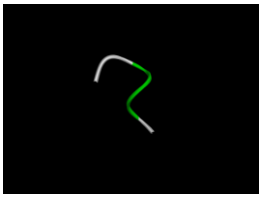
6) δ_R-β_S-β_P-β_P-β_P-α_R-α_R-α_R-β_S-β_S

7) β_S-β_S-β_P-δ_R-α_L-β_P-α_R-α_R-α_R-β_S

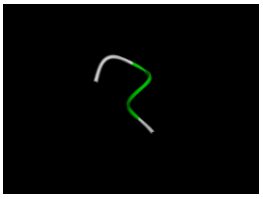
8) δ_R-β_S-β_P-δ_R-α_L-β_P-α_R-α_R-α_R-β_S

9) β_S-β_S-β_S-δ_R-β_P-δ_R-α_R-α_R-α_R-β_S

10) δ_R-β_S-β_S-δ_R-β_P-δ_R-α_R-α_R-α_R-β_S



11) β_S-β_S-β_P-α_R-β_P-δ_R-α_R-α_R-α_R-β_S



12) δ_R-β_S-β_P-α_R-β_P-δ_R-α_R-α_R-α_R-β_S

Figure 21. Most probable 12 conformations of the sequence "Gln-Val-Cys-Ala-Asn-Pro-Glu-Lys-Lys-Trp". From top left to bottom right the state sequences are ordered from the most probable to the 12th most probable conformation. The symbols indicate the torsion states of the residues for the respective conformations. The conformations are determined using the Viterbi algorithm with multistep backtracking with n=12, and m=1000.

In previous studies, the prediction of most probable torsion state of a residue has been considered [74, 75]. Helles et al. considered the problem of predicting the probability distribution of the states of the coil residues in a given peptide instead of predicting only the most probable state [79]. As a different approach, we predict the high probable conformations rather than the probabilities of torsion states of the individual residues. Therefore, this is the first study that handles the prediction of high probability conformations.

The predictive capability of a given method depends on the nature of the library used, and the number of states that are used to define the probability space. The non-redundant PDB data set contains information on secondary structures, information on structural motifs and their relation to conformations which may be incorporated into the probability space to be used for predictions. A coil database contains less information, only on residue types and neighbor dependences. An IDP database contains information on the specific structural features of the disordered chains. Comparing the predictions from different data sets is not meaningful. However, just in order to see the relative performances of other methods and ours, we use results of Bystroff et al. [73] and Kuang et al. [74]. Both of these studies use the non-redundant PDB data base. Bystroff et al. [73] used eleven states using a different definition from our torsion state definition and proposed an HMM model based on a library of sequence-structure motifs rather than the coil regions. Kuang et al. [74] defined four states (A, B, G, E) and used an SVM method based on protein sequence profiles. Since these methods use different torsion state definitions, Kuang et al. [74] suggested to group the states given by Bystroff et al. [73] into four states (A', B', G', E') for comparison. The grouped states are only approximately equivalent. Then, the prediction accuracies are determined as 74% for HMMSTR and 77.3% for SVM methods. We follow the same idea and group our eleven torsion states into four states as ($A'' = \alpha_R + \delta_R$, $B'' = \beta_P + \beta_S + \varepsilon' + \gamma' + \zeta$, $G'' = \alpha_L + \delta_L + \gamma$, $E'' = \varepsilon$). Then, the prediction accuracy of the Viterbi algorithm is 59.2% for predicting the most probable state of a residue in a given peptide sequence of the coil library. The additional 15% improvement may approximately be ascribed to the additional structural information content in the non-redundant PDB data set over that of the coil dataset.

Predicting highly probable conformations of a given peptide sequence is an important issue in targeted drug design. To be effective, a short peptide must match a given region of the surface of

a protein, producing a change in the activities of proteins that can affect certain diseases [68]. The conformations of the peptide that fulfil this constraint have to be among the most probable ones for maximum stability.

Another important concept is the intrinsically disordered proteins those don't naturally adopt a unique three dimensional structure. Therefore, their structure ranges from the totally random coil to molten globular [83, 84]. Use of a torsion state library for IDP's and the present computational model will lead to the prediction of high probability conformations of disordered proteins.

4.3 CONCLUDING REMARKS

Predicting "n" high probability conformations of a given peptide problem is discussed. The proposed method is based on RIS model and Viterbi algorithm with multistep backtracking. The necessary parameters of the HMM are derived from a coil library depending on the residue types and the near neighbor effects.

While using the Viterbi algorithm the long-range interactions are not included and may be essential for the state prediction. One may introduce second, or third neighbor effects that also would improve the prediction accuracies, since the coil regions are not exactly random [85]. Also, the scarcity of data for some torsion states in the Coil library lead to difficulty in prediction.

The present work discusses the prediction of conformations of random coiled proteins from primary sequence. We showed that for certain amino acid types, introduction of residue type and neighbor dependence improved the predictions of conformations of proteins in the coil state. Thus, we showed that the Flory Isolated Pair Hypothesis does not hold in general.

Chapter 5

CONFORMATIONAL TRANSITIONS IN THE RAMACHANDRAN SPACE OF AMINO ACIDS BY THE DYNAMIC ROTATIONAL ISOMERIC STATE (DRIS) MODEL

The dynamic rotational isomeric state (DRIS) approach is utilized to predict local dynamics of 20 amino acids. An exhaustive sampling of amino acid conformations is given to understand the intrinsic properties of the amino acids. The transition rates between rotational isomeric states are calculated from molecular dynamics (MD) simulations of Gly-Gly-X-Gly-Gly pentapeptides where X represents one of the 20 amino acids. A computational approach is given for measuring relaxation times of the amino acids. The results are in good agreement with fluorescence quenching rate measurements [86].

Proteins are not rigid structures, they are dynamic molecules and characterized by conformational ensembles at equilibrium. Proteins perform most of their function through rearrangements of their conformations, such as allosteric rearrangements, large scale fluctuations, tail motions etc. Therefore, the knowledge of only the average structure of a protein does not provide a complete description. The knowledge of the popularity of the states, the transition rates and the pathways are all necessary for a full understanding of a protein [87]. Proteins exhibit both internal motions including bond vibrations, bond stretching, side-chain motions, rotational transitions and global motions as rotational and translational motions [63, 88]. In this study, we consider dynamics at the single residue level and focus on the internal motions resulting from conformational transitions exhibited by the twenty amino acids. We adopt the dynamic rotational isomeric states (DRIS) formalism in explaining and interpreting the residue dynamics. The DRIS approach has been developed for predicting local internal dynamics of polymer chains by Bahar and Erman [56]. The model requires the knowledge of the exact locations and probabilities of all torsion states of each amino acid and the transition rates between torsion states. These parameters are determined in this work by an exhaustive sampling of amino acid Ramachandran plots by

Molecular dynamics (MD) simulations. In a sense, the DRIS formalism is used to organize the MD simulations results in a compact and efficient way as explained in detail in Chapter 2.

The DRIS model is an extension of the equilibrium theory of chain statistics by Flory [1] to the dynamic domain by Jernigan [64] and is conveniently used to calculate the internal time correlation functions of a chain and different dynamic properties associated with the transitions between the isomeric states [56-63].

A computational approach is given for measuring relaxation times of the amino acids. Here, the aim is to investigate the conformations of the twenty amino acids whose dynamics evolves only from transitions from one torsion minimum to another as it would obtain in the random coil state. The preferences of the amino acids extracted from the Protein Data Bank are biased by the presence of secondary and tertiary structure, and long range interactions such as contacts and hydrogen bonds of a given residue with the rest of the protein [89]. Hence, instead of analyzing the protein structures from the protein data bank (PDB) or running simulations of proteins we perform MD simulations of Gly-Gly-X-Gly-Gly pentapeptides. Here X represents the amino acid to be examined. The GGXGG peptides have been commonly used as models for the random coil state [89-94].

We use MD simulations to identify isomeric states of the amino acids and to calculate the equilibrium probabilities of the corresponding states. Then, the rates of the transitions between the isomeric states are calculated using the DRIS approach based on the parameters obtained via MD simulations. The relaxation rates determined via DRIS model are compared with the fluorescence quenching rate constants obtained experimentally by Huang and Nau [86]. Results of our calculations agree with their experimental findings.

5.1 METHODS AND MATERIALS

The details of the generalized DRIS model are already given in Section 2.5. Here we apply the DRIS model to analyze the dynamics of the transitions of (ϕ,ψ) angles of the amino acids. These transitions are determined by the transitions between the specified regions of the Ramachandran map. Those regions are separated by energy barriers. We have defined $\nu = 8$

discrete regions for this purpose based on the molecular dynamics simulations we have performed.

Let the torsion state of an amino acid be defined by

$$\{\Phi\}_i = \{\varphi_i, \psi_i\} \tag{4.1}$$

Based on the interdependence between the neighbor amino acids, the transitions may be defined as (i) first order (independent), According to this, the statistics of $\{\Phi\}_i$ is independent of the statistics of its neighbors, (ii) second order (pairwise dependent), where the statistics of $\{\Phi\}_i$ depends on the statistics of its preceding neighbor, and (iii) third order (triplewise dependent) where the statistics of $\{\Phi\}_i$ depends on the statistics of its preceding two neighbors. In this work, we adopt the Flory independence hypothesis and assume first order statistics [1]. For a first order transition of an amino acid, we define the master equation as

$$\frac{dP(t)}{dt} = AP(t) \tag{4.2}$$

where the vector $P(t)$ is the 8-dimensional vector of the probabilities of 8 states, and the transition rate matrix A is the 8×8 dimensional matrix. The state of a given amino acid is defined by one of the torsion states $\{\Phi\}_k$, $k = 1,\ldots,8$. Then, A_{ij}, the ijth element of the rate matrix, denotes the rate of the transition from state $\{\Phi\}_j$ to the state $\{\Phi\}_i$.

The formal solution of the master equation gives

$$P(t) = \exp\{At\}P(t=0) \tag{4.3}$$

$P(t=0)$ represents the vector of the equilibrium probabilities of the 8 states. The eigendecomposition of the matrix A gives $A = B\Lambda B^{-1}$. Here B is the matrix whose ith column is the ith eigenvector of A, Λ is the diagonal matrix whose diagonal elements are the eigenvalues λ_i, and B^{-1} is the inverse of B. Then,

$$P(t) = B\exp\{\Lambda t\}B^{-1}P(t=0) \tag{4.4}$$

The term $Be^{\Lambda t}B^{-1}$ defines the time-delayed conditional probability matrix $C(t)$. The element C_{ij} denotes the probability of being at state $\{\Phi\}_i$ at time t, given the initial state $\{\Phi\}_j$ at t=0. The elements of each column of C_{ij} comprise all possible transitions from a given initial state to one of the 8 possible states. Hence, the summation of each column is unity.

$$C(t) = B\exp\{\Lambda t\}B^{-1} \tag{4.5}$$

The total time-dependent joint probability matrix $P(t)$ is defined as

$$P(t) = C(t)\ \mathrm{diag}P(t=0) \tag{4.6}$$

The element P_{ij} represents the joint probability of occurrence of state $\{\Phi\}_i$ at time t and $\{\Phi\}_j$ at t=0. P_{ij} may be defined as

$$P_{ij} = \sum_k B_{ik}\exp\{\lambda_k t\}B_{kj}^{-1}P_j(0) \tag{4.7}$$

Knowledge of the joint probability matrix $P(t)$, or alternatively knowledge of conditional probability matrix $C(t)$ with the equilibrium probabilities $P(t=0)$ gives a complete description of the dynamics of a given amino acid.

For any dynamic property f_{ij} which is a function of the conformations $\{\Phi\}_i$ and $\{\Phi\}_j$, the average of the property $\langle f_{ij}\rangle$ over all possible transitions taking place from conformation j to i can be calculated by assigning a stochastic weight to each transition

$$\langle f_{ij}\rangle = \sum_i\sum_j P_{ij}(t)\,f_{ij} \tag{4.8}$$

5.1.1 Molecular Dynamics Simulations For Determination Of The States And Equilibrium Probabilities

A capped Gly-Gly-X-Gly-Gly peptide model is used for the analysis of the amino acid X. The simulations are performed using Gromacs 4.5.5 package [95]. The –N and the –C termini of the pentapeptide models Gly-Gly-X-Gly-Gly are capped with ACE and NME respectively. The initial conformation is selected as extended.

OPLS/AA all atom force field is used and the system is solvated using four-atom water model TIP4P [96, 97]. The electrostatic interactions are calculated using PME-Switch method with Coulomb cut-off 1.0 nm. The Van der Waals interactions are calculated using Switch method with VDW cut-off 1.0 nm.

An energy minimization using the conjugate gradient method with 5000 maximum number of steps and 0.001 ps time step is performed. Then to assure that the solvent configuration matches the peptide, the solvent and the hydrogen atoms are relaxed for 2500 steps while the peptide was restrained.

Before the production MD step, first a 20 ps temperature coupling MD simulation and then a 40 ps pressure coupling MD simulation is performed. During the 20 ps NVT ensemble simulation the Berendsen temperature coupling method is used [98]. The coupling constant is assigned as 0.1 ps. Then, during a 40 ps NPT ensemble simulation Berendsen temperature and pressure coupling methods are used. The temperature coupling constant is again assigned as 0.1 ps and the pressure coupling constant is taken as 0.5 ps.

A simulated annealing procedure is necessary to avoid a local minimum and to sample the energy surface properly [99]. We used 80 ps simulated annealing cycles during the production MD simulation. In each cycle, the temperature of the system is increased from 310 K to 1010 K in the first 20 ps. Then in the second 20 ps part the temperature of the system is decreased from 1010 K to 310 K. Finally in the last 40 ps part the temperature of the system is kept constant at 310 K. This equilibration part is necessary since the collected data should be sampled from a Boltzmann distribution. Using the simulated annealing method, a different initial conformation is

selected in each cycle. Since the system is equilibrated after heating and cooling procedures, we assure that the selected conformations are sampled from Boltzmann distribution.

5.1.2 DETERMINATION OF THE ROTATIONAL ISOMERIC STATES

The Ramachandran plots showing the distributions of the populated regions of 20 amino acids in the Gly-Gly-X-Gly-Gly peptide are constructed. The Ramachandran plot of each amino acid is obtained directly from the simulation trajectories. The common populated regions are selected as the torsion states.

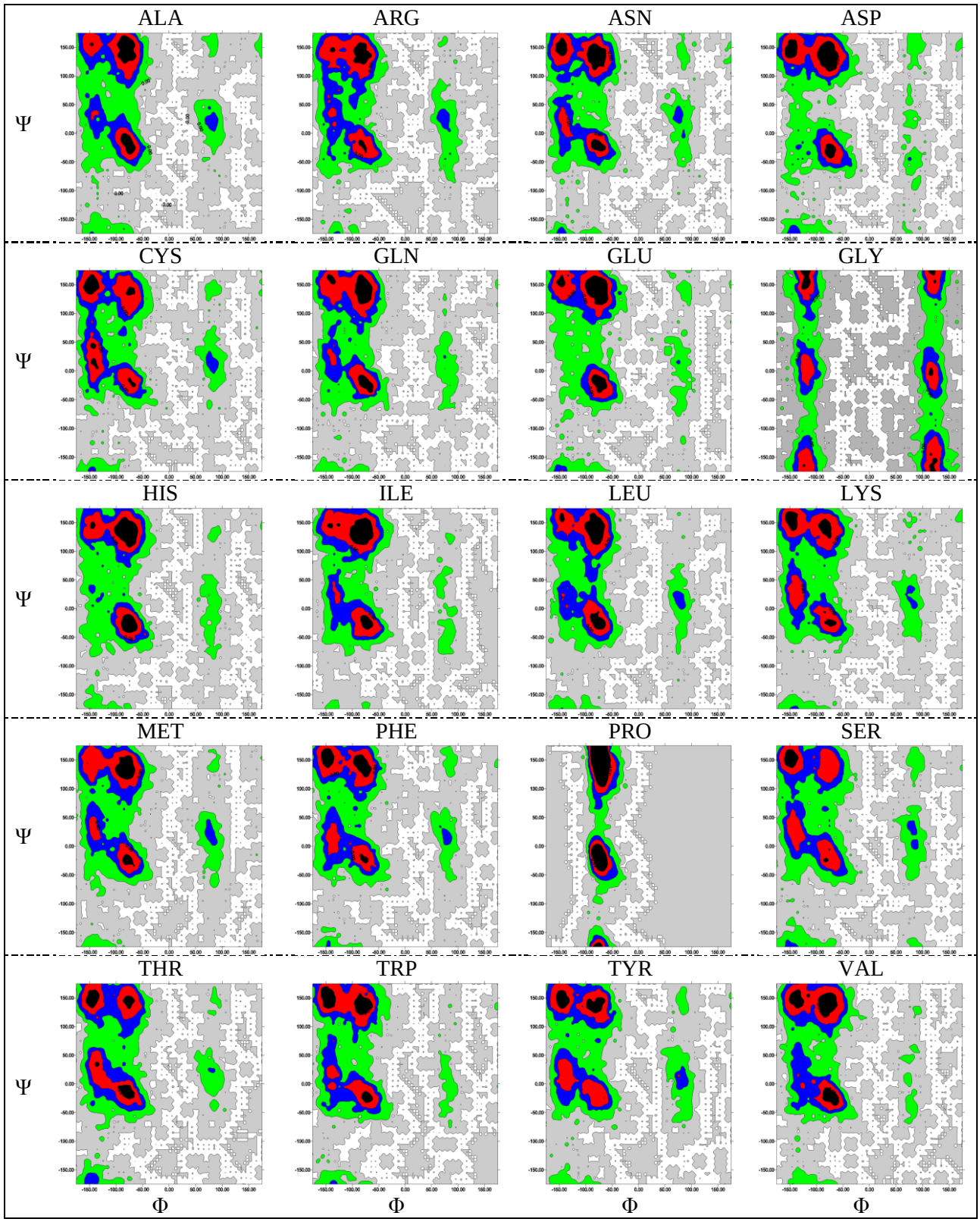


Figure 22. Ramachandran plots for the twenty amino acids. The populations are obtained from MD simulations of GGXGG peptides.

In Figure 22, a contour plot for each amino acid plotted using $5^{\circ} \times 5^{\circ}$ grids where the x-, y-axes represent φ and ψ angle distributions, respectively. The percentage of points in each grid are colored as: $[0, 0.05)$: gray; $[0.05, 0.2)$: green; $[0.2, 0.4)$: blue; $[0.4, 0.8)$: red; $[0.8, 1]$: black that has been used by Beck et al. [89]. Based on the Ramachandran plots, 8 torsion states are determined. The selected regions are shown in Figure 23. We have defined 7 states based on the popularity of the regions and identified 8th state as the all other regions. Since Proline has a constrained phi angle, we have determined 4 torsion states for Proline. The identified states of the Proline are shown in Figure 24. Again, the 4th state is defined as the all other allowed regions.

The regions are not so different from the regions introduced by Karplus [23]. Unal et al. also defined eleven states based on a knowledge based database which are similar to the regions we have defined by MD simulations [51, 68, 100]. In a knowledge based database, the preferences of the amino acids are biased by the presence of the protein. Furthermore, a coil library that is constructed by removing helical or beta structures may still include influences of conformations of proteins [89]. Therefore we perform MD simulations of pentapeptides which would present better representatives for the random coil structures. The boundaries of the states are selected in such a way that the transitions between these states would be analyzed explicitly. We have defined rectangular shaped states centered on the most populated regions.

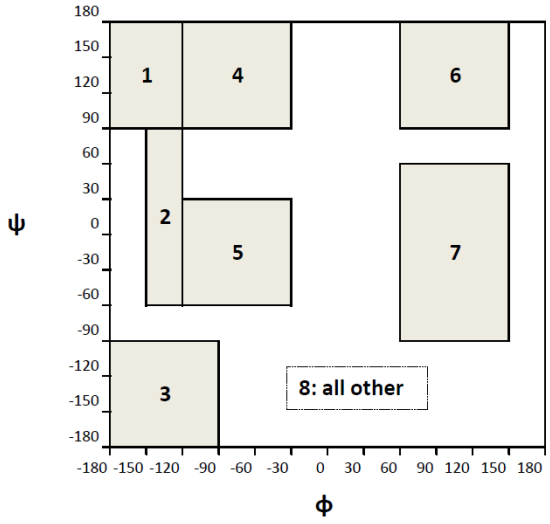


Figure 23. The representation of 8 states on Ramachandran map.

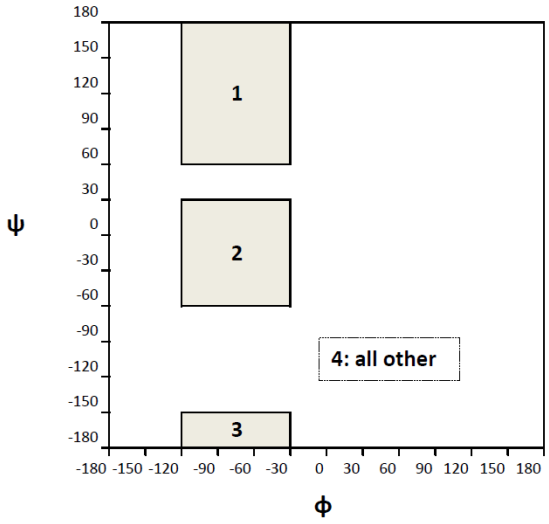


Figure 24. The representation of 4 states of Proline.

5.1.3 Calculating The Equilibrium Probabilities

The equilibrium probabilities are calculated from the equilibrium conformations data collected via MD simulations. For each amino acid, we simply count the number of states visited. The probability of observing an amino acid X in a state i is calculated as

$$P^X(s_i) = \frac{N^X(s_i)}{\sum_i N^X(s_i)} \tag{4.9}$$

where $N^X(s_i)$ represents the number of occurrences of amino acid X in a state i during the simulation ant the summation term indicates the total number of occurrences in all possible states by amino acid X .

The equilibrium probabilities of each amino acid obtained by MD simulations are presented in Table 6.

Table 6. The equilibrium probabilities of the amino acids based on 8 states

States:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
ALA	0.14	0.08	0.03	0.33	0.21	0.01	0.05	0.13
ARG	0.17	0.11	0.02	0.30	0.19	0.01	0.07	0.13

ASN	0.20	0.11	0.03	0.31	0.18	0.01	0.05	0.11
ASP	0.21	0.05	0.04	0.37	0.19	0.02	0.03	0.08
CYS	0.22	0.16	0.03	0.23	0.16	0.01	0.05	0.14
GLN	0.17	0.10	0.02	0.33	0.21	0.01	0.04	0.12
GLU	0.17	0.04	0.04	0.40	0.21	0.01	0.04	0.09
GLY	0.10	0.07	0.14	0.09	0.07	0.13	0.16	0.25
HIS	0.16	0.06	0.02	0.37	0.23	0.01	0.04	0.11
ILE	0.16	0.10	0.01	0.39	0.22	0.01	0.03	0.08
LEU	0.14	0.08	0.03	0.34	0.23	0.01	0.05	0.11
LYS	0.20	0.13	0.03	0.27	0.18	0.01	0.05	0.12
MET	0.17	0.10	0.02	0.32	0.19	0.01	0.06	0.12
PHE	0.21	0.13	0.02	0.28	0.18	0.01	0.05	0.12
PRO	0.00	0.00	0.01	0.65	0.25	0.00	0.00	0.10
SER	0.21	0.13	0.04	0.22	0.19	0.02	0.06	0.15
THR	0.20	0.14	0.04	0.23	0.22	0.01	0.04	0.11
TRP	0.22	0.11	0.03	0.30	0.20	0.01	0.03	0.10
TYR	0.20	0.13	0.03	0.26	0.18	0.02	0.07	0.12
VAL	0.19	0.10	0.02	0.35	0.23	0.01	0.02	0.08

5.2 CALCULATION OF THE RATES

In the previous applications of the DRIS, the rate constants related to transitions were calculated by using a kinetic scheme [56, 57]. The rates were determined by the activation energies E_a and a front factor A_0.

$$r = A_0 \exp\{-E_a / RT\} \tag{4.10}$$

The activation energies are calculated from the heights of the saddle points in energy maps. The front factor A_0 is represented by Kramers' expression for high-friction Brownian motions [56]. It is the frequency of passing the energy barrier at a given temperature. Then the similarity transformation of the rate matrix A whose elements are the rates leads to the conditional probability matrix $C(t)$.

Alternatively, MD simulation trajectories of the chains may be used to determine the time-delayed conditional probability curves $C(t)$ [63]. We follow this approach and obtain the conditional probability curves for each amino acid. Then, the initial slopes of the curves give the transition rates following the equation (4.5). Eeach curve converges to the equilibrium probabilities of the final state at infinite time.

5.2.1 Calculation Of The Time-Delayed Conditional Probabilities

Given that we observe a state s_i at time $t = 0$, the probability of observing a state s_j at time t is given as

$$C(s_j, t; s_i, 0) = \frac{P(s_i, 0; s_j, t)}{P(s_i, 0)} \tag{4.11}$$

where $P(s_i, 0; s_j, t)$ is the joint probability that the system is in state i at $t = 0$ and in state j at time t.

For a chosen time step $t = \tau$, we record the joint observations of s_i and s_j in time interval τ throughout the simulation. Then, the joint probability that the system is in state s_i at $t = 0$ and in state s_j at $t = \tau$ is

$$P(s_i, 0; s_j, \tau) = \frac{N_{ij}}{N_{total}} \tag{4.12}$$

Here N_{ij} is the number of times when the system is in s_i at a time t and in s_j at time $t + \tau$, and N_{total} is the total number of observations.

After repeating the counting scheme for different chosen $t = \tau, 2\tau, 3\tau, \ldots$ we obtain the joint probabilities. Then the joint probabilities lead determination of the time-delayed conditional probabilities.

The time-delayed conditional probability curves C_τ are obtained for each amino acid using this approach. The transition probabilities for all possible transition between eight states are

represented. All of the curves are expected to approach to the equilibrium probability at infinite time. However, this condition is not satisfied in reasonable time for MD simulations at lower temperatures. Hence, the simulations are performed for higher temperatures as 400 K, 600 K, 800 K, and 1000 K. Then, the aim is to estimate the rates for 310 K based on the values obtained for higher temperatures.

In Figure 25, the conditional probability curves are given for Alanine evaluated by MD simulations at 400 K, 600 K, 800 K, and 1000 K. It is evident that the curves converge to the equilibrium probability faster for the higher temperatures. Based on this fact, the calculations are performed for higher temperatures. Then, the dynamical behavior at 310 K is estimated for each amino acid.

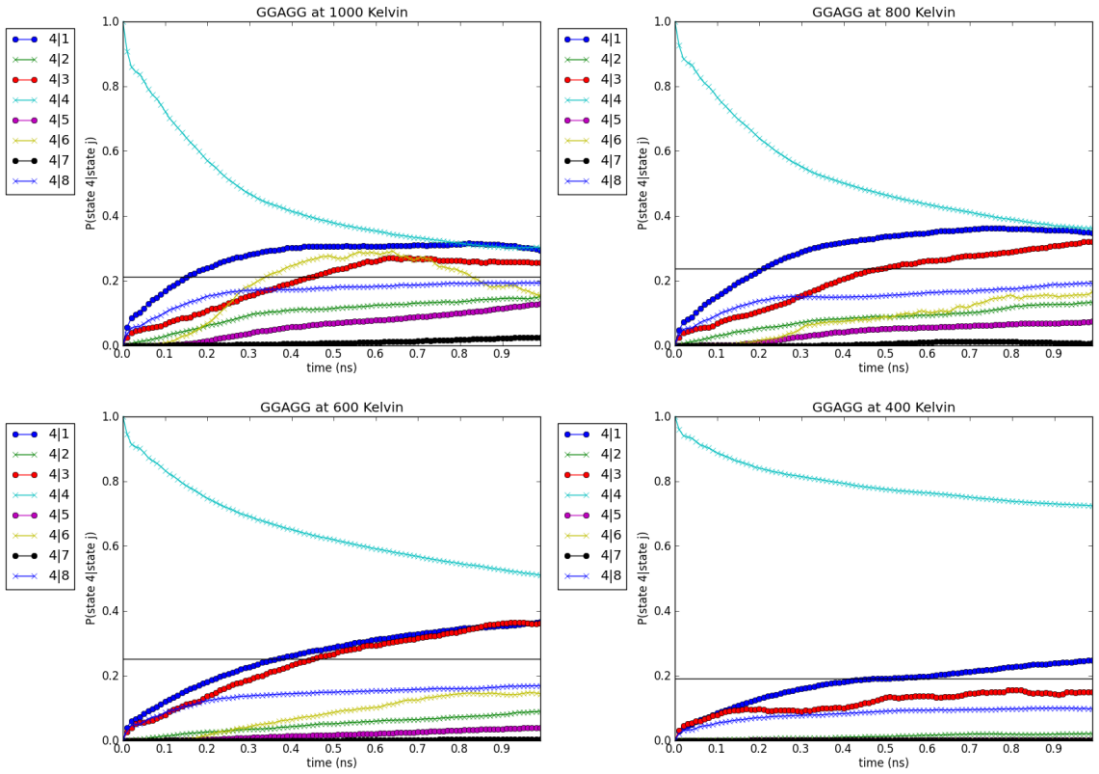


Figure 25. Time-delayed conditional (transition) probabilities for Alanine at 1000 K, 800 K, 600 K, and 400 K. The probabilities are shown for all possible transitions from the eight states to the state 4. The black solid lines represent the equilibrium probability of the state 4 for Alanine.

5.2.2 *Determination Of The Rate Matrices*

The initial slopes of the time-delayed conditional probability curves give the transition rate matrix elements. Since $C_\tau = \exp\{A\tau\}$,

$$\left.\frac{dC_\tau}{d\tau}\right|_{\tau=0} = A\exp\{A\tau\}\big|_{\tau=0} = A \tag{4.13}$$

One may derive the derivative at $\tau = 0$ using the finite differences. The forward difference is defined by

$$\Delta f(x) = f(x+h) - f(x) \tag{4.14}$$

Furthermore, the higher order differences are derived by

$$\Delta^k f(x) = \Delta^{k-1} f(x+h) - \Delta^{k-1} f(x) \tag{4.15}$$

Therefore, $f'(x=0)$ may be estimated by the sum of forward differences as follows

$$\left.\frac{df}{dx}\right|_{x=0} \cong \frac{1}{h}\left(\Delta f(0) - \frac{1}{2}\Delta^2 f(0) + \frac{1}{3}\Delta^3 f(0) + \cdots\right) \tag{4.16}$$

The rates are estimated using finite differences up to 5 order, using only the first six data points of the conditional probability curves. After construction of the rate matrices, the matrices of eigenvectors and eigenvalues are determined using the eigenvalue decomposition (4.2)-(4.4). Then, any dynamic property may be obtained by going through equations (4.7) and (4.8) .

5.2.3 *Molecular Dynamics Simulations For Determination Of The Rates*

The same peptide model Gly-Gly-X-Gly-Gly and the same parameters that have been used in MD simulations for determinations of the states are used. However, this time we don't use a simulated annealing procedure since the aim is to collect data for the transitions happened during the simulations. 1 ns simulations are performed for each type of amino acids and the data collected in 1 ps periods.

5.3 INTERNAL FLEXIBILITY

Any dynamic property associated with conformational transitions, $\langle f(\tau)\rangle$, may be computed by using the rate matrices. Similarity transformation of the rate matrix ($A = B\Lambda B^{-1}$) leads to the determination of a mean transient property as follows

$$\langle f(\tau)\rangle = \sum_{\gamma} k_{\gamma} \exp(\lambda_{\gamma}\tau) \tag{4.17}$$

where λ_{γ} is the γ^{th} eigenvalue of Λ and k_{γ} denotes the amplitude factor given by

$$k_{\gamma} = \sum_{\alpha}\sum_{\beta} B_{\alpha\gamma}\left[B^{-1}\right]_{\gamma\beta} P_0(\Phi_{\beta}) f(\Phi_{\alpha};\Phi_{\beta}) \tag{4.18}$$

We are interested in calculation of the local relaxation properties of the amino acids. The internal dynamics of a given residue X may be analyzed by a function that is associated with conformational transitions taking place between the torsion states. Hence, we consider two vectors; (i) l_1 between the atoms C_{i-1} and N_i, and (ii) l_2 between the atoms N_{i+1} and C^{α}_{i+1} of the backbone of $Gly_{i-2} - Gly_{i-1} - X_i - Gly_{i+1} - Gly_{i+2}$ for this purpose. The conformations of the vectors will change with time related to the values of (ϕ,ψ).

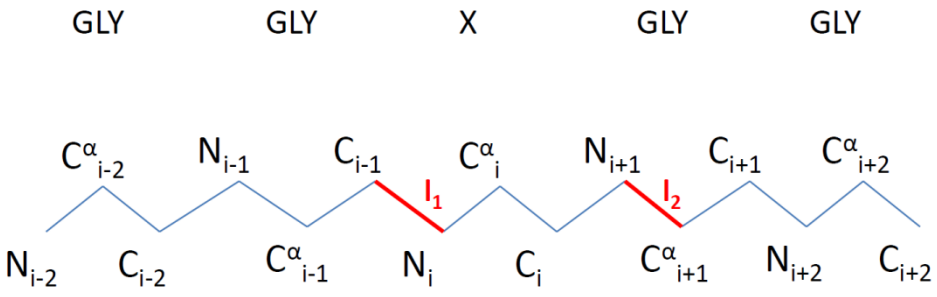


Figure 26. l_1 and l_2 vectors of the GGXGG peptide

For this purpose we define two functions;

i. $\langle f_1(\tau)\rangle$ that gives the average cosine of the angle between l_1 vectors at $t=0$ and $t=\tau$ when the vector l_2 is kept fixed,

ii. $\langle f_2(\tau)\rangle$ that gives the average cosine of the angle between the cross product of the two vectors, $l_3 = l_1 \times l_2$, that gives information on conformational change in the normal.

Here, $\langle\ \rangle$ means averaging over all time steps. Then, $\langle f_1(\tau)\rangle$ is evaluated by

$$\langle f_1(\tau)\rangle = \left\langle \frac{l_1(0)\bullet l_1(\tau)}{\|l_1\|\|l_1\|} \right\rangle = \sum_{\gamma} k_\gamma \exp\{\lambda_\gamma \tau\} \tag{4.19}$$

$$k_\gamma = \sum_i \sum_j B_{i\gamma} \left[B^{-1}\right]_{\gamma j} P_0(s_j) \frac{l_1 \bullet l_1(s_i; s_j)}{\|l_1\|\|l_1\|} \tag{4.20}$$

Here $l_1 \bullet l_1(s_i; s_j)$ is the dot product of the l_1 vector with itself when the transition takes place from state s_j to state s_i.

Similarly, the cross product of the two vectors, $l_3 = l_1 \times l_2$, is analyzed by calculating $\langle f_2(\tau)\rangle$.

$$\langle f_2(\tau)\rangle = \left\langle \frac{l_3(0)\bullet l_3(\tau)}{\|l_3\|\|l_3\|} \right\rangle = \sum_{\gamma} k_\gamma \exp\{\lambda_\gamma \tau\} \tag{4.21}$$

$$k_\gamma = \sum_i \sum_j B_{i\gamma} \left[B^{-1}\right]_{\gamma j} P_0(s_j) \frac{l_3 \bullet l_3(s_i; s_j)}{\|l_3\|\|l_3\|} \tag{4.22}$$

5.4 RESULTS AND DISCUSSION

Figure 27 and Figure 28 show the two correlation functions $\langle f_1(t) \rangle$ and $\langle f_2(t) \rangle$ over time for some amino acids (Asn, Cys, Gly, Trp) at T=400, 600, 800, 1000 Kelvin respectively. The decay curves presented in the figure show that these functions exhibit exponential decay behavior. Hence, it is possible to obtain the relaxation times by fitting the curves with formula $A + B\exp(-t/C)$. Then, the parameter C that gives the relaxation time τ_R is determined for all amino acids at different temperatures.

Arrhenius' equation is a mathematical expression that gives the relation between rate constants and temperature. The equation was proposed by Svante Arrhenius in 1889 [101] by means of experiments. That is

$$k = A\exp(-E_a / k_B T) \tag{4.23}$$

where k is the rate constant of a chemical reaction. A is the pre-exponential factor or the frequency factor of the reaction. E_a is the activation energy, k_B is the Boltzmann constant and T is the absolute temperature. Taking the natural logarithm of the equation yields

$$\ln(k) = \ln(A) - \frac{E_a}{k_B T} \tag{4.24}$$

When $\ln(k)$ is plotted as a function of $1/T$, this relation, called the Arrhenius' equation, gives a straight line since it is in the form of $y = mx + b$.

Hence, our aim is to estimate the straight lines by linear least squares fitting technique after obtaining the relaxation times. These lines determined for each amino acid at different temperatures lead to relaxation times at 310 K by extending the lines until $1/T = 1/310 \approx 0.003$.

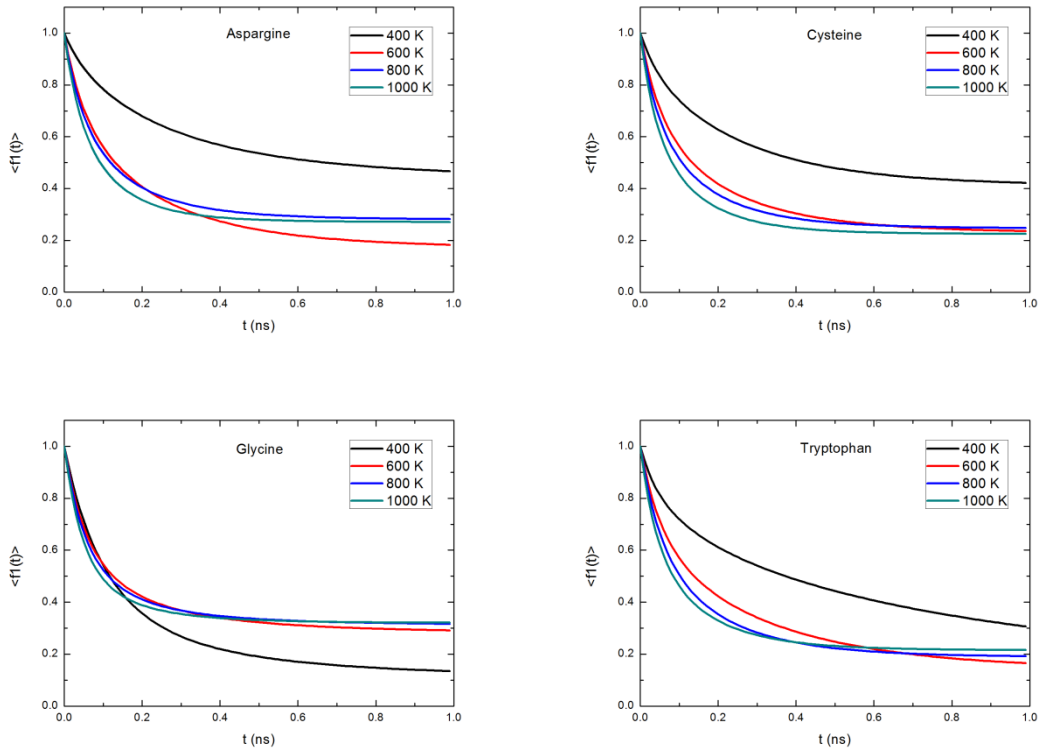


Figure 27. The mean dynamic properties $<f_1(t)>$ for Asn, Cys, Gly, Trp at 400, 600, 800 and 1000 Kelvin.

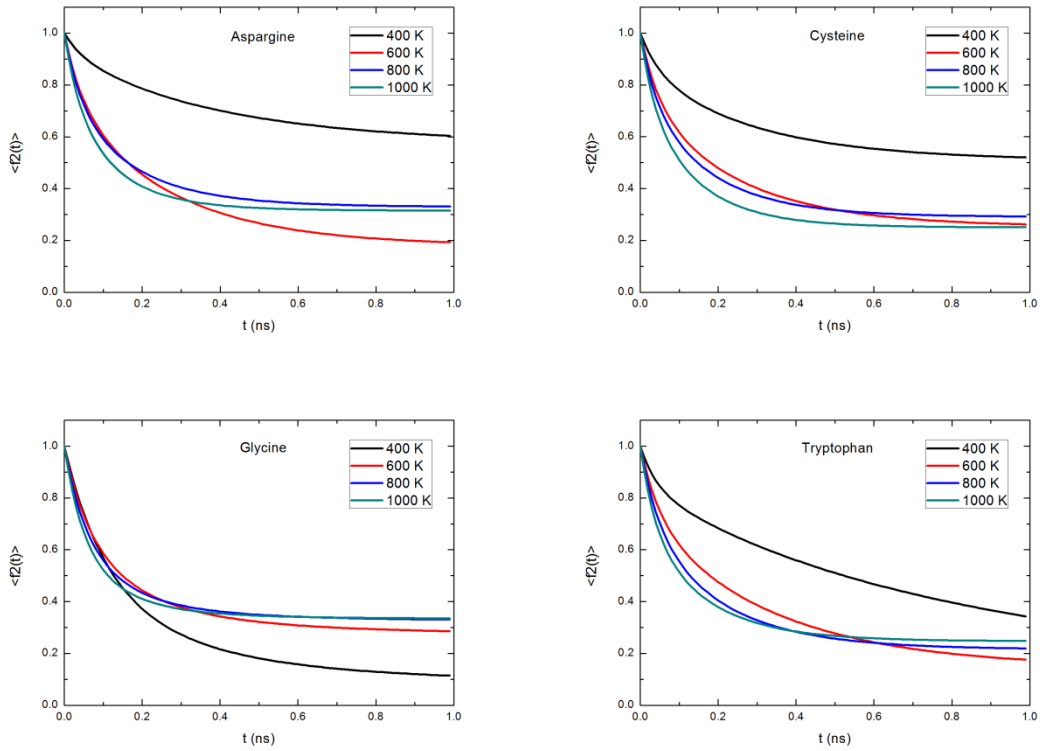


Figure 28. The mean dynamic properties $<f_2(t)>$ for Asn, Cys, Gly, Trp at 400, 600, 800 and 1000 Kelvin.

The relaxation times determined via fitting $\langle f_1(t) \rangle$ to exponential decays are shown in Figure 29 for some amino acids. Curve fitting the natural logarithm of relaxation times to linear lines yield the red lines shown in the graphs. The relaxation times at T=310 K are estimated by this lines since the dynamic functions obey Arrhenius' equation.

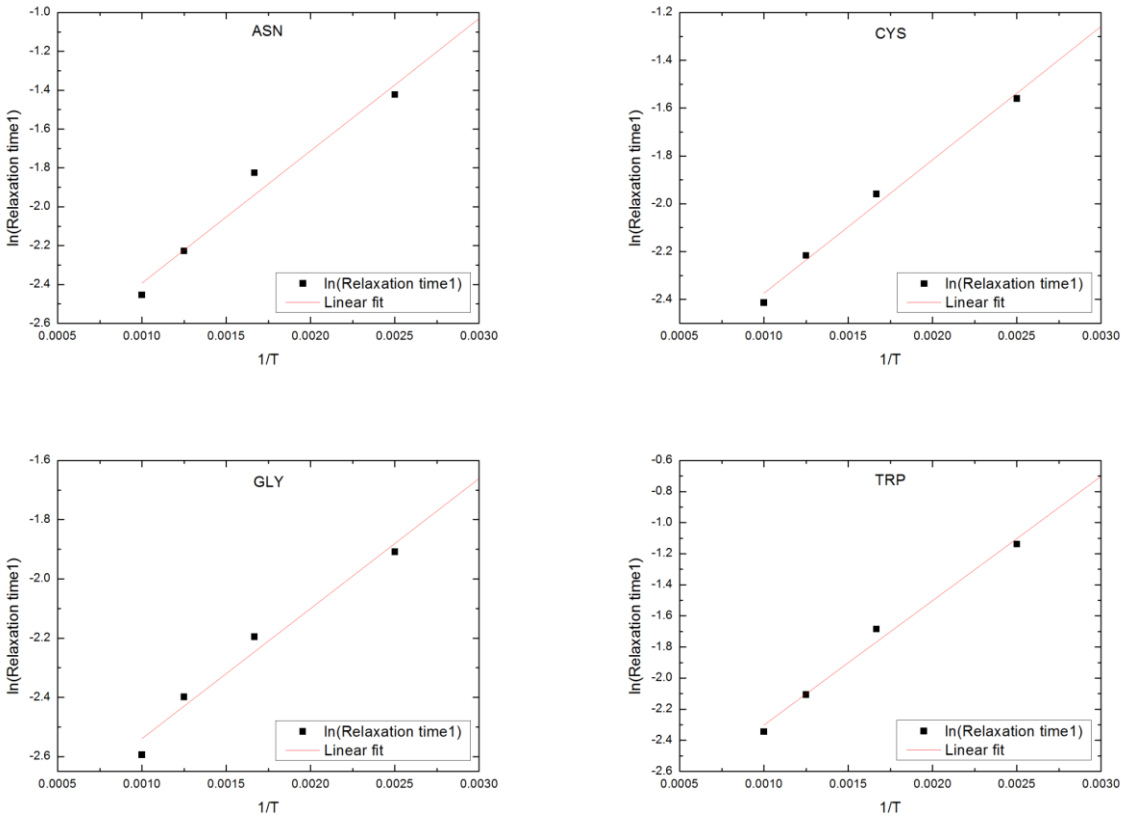


Figure 29. Relaxation times for amino acids Asn, Cys, Gly, and Trp calculated via <f1(t)>. Red line represents the linear curve fitting.

Similarly, the relaxation times obtained via fitting $\langle f_2(t) \rangle$ are presented in Figure 30.

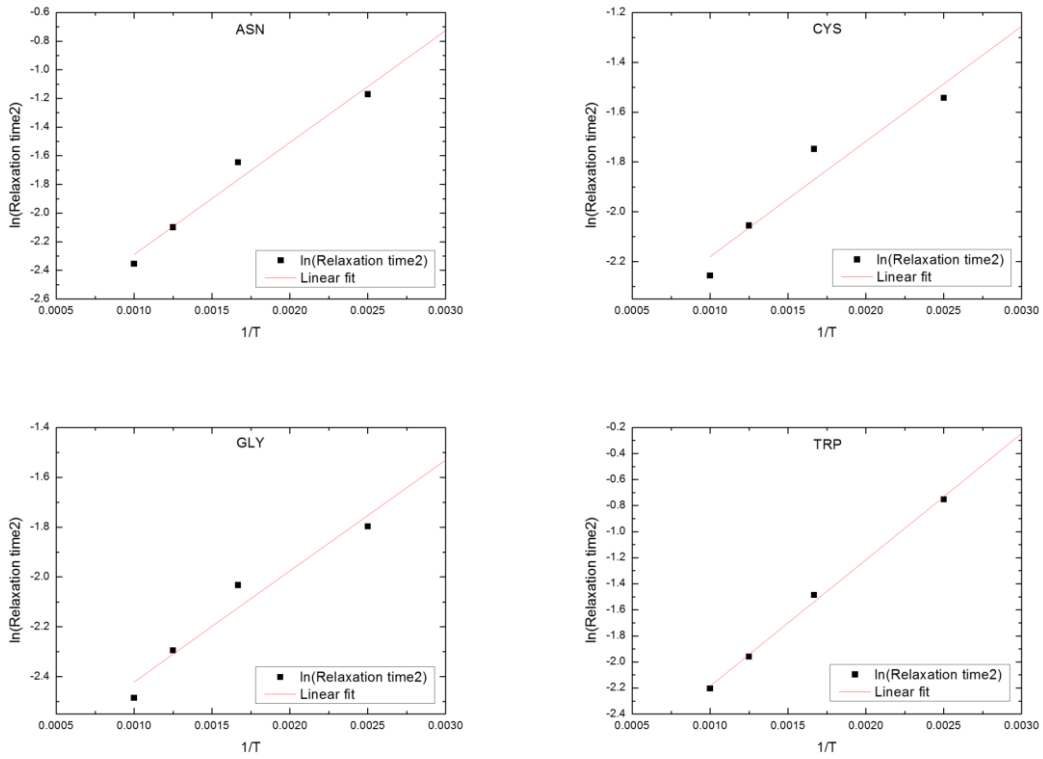


Figure 30. Relaxation times for amino acids Asn, Cys, Gly, and Trp calculated via <f2(t)>. Red line represents the linear curve fitting.

In Table 7, the estimated relaxation times τ_1 and τ_2 for T=310 K are presented. Glycine and Serine that are known as small amino acids relax faster than the other amino acids. Glycine is the least restricted amino acid while Proline has restrictions in Ramachandran map and has the largest relaxation time.

Isoleucine and Threonine relax more slowly than the other amino acids. These amino acids are restricted in the conformations since they contain two non-hydrogen substituent attached to their C-beta carbon. They are known as C-beta branched amino acids. Hence, it is not surprising that these amino acids relax slower than the other amino acids.

Table 7. Estimated relaxation times for twenty amino acids at 310 Kelvin.

Amino Acid	Relaxation time1 (τ_1)	Relaxation time2 (τ_2)
ALA	0.33	0.40
ARG	0.32	0.38
ASN	0.36	0.48
ASP	0.29	0.33
CYS	0.28	0.28
GLN	0.54	0.81
GLU	0.39	0.56
GLY	0.19	0.22
HIS	0.45	0.61
ILE	0.55	1.12
LEU	0.27	0.33
LYS	0.49	0.70
MET	0.40	0.58
PHE	0.41	0.54
PRO	1.51	1.51
SER	0.23	0.21
THR	0.72	1.05
TRP	0.50	0.78
TYR	0.39	0.53
VAL	0.37	0.52

Huang and Nau determined the fluorescence lifetimes of unstructured peptides as a function of the amino acid type [86]. It has been reported that the quenching rate constant is a scale for the flexibility of amino acids. The rate constants of sixteen amino acids are given except for Trp, Tyr, Cys, and Met since these four amino acids are themselves are known as efficient quenchers [102].

In Figure 31, we compare our results with the fluorescence results of Huang and Nau [86]. In general, the results are in good agreement. The quenching rates are determined as a measure for flexibility and the results are related with the residue size.

In Figure 32, we provide a comparison of quenching time with molecular weight. We removed Proline since it is an outlier with a relaxation time value of more than 10 microseconds. The quenching times of Ile and Val are higher than the others. These amino acids are β-branched amino acids that have restricted main chain conformations. On the other hand, β-branched Thr seems to be more flexible than Ile and Val. Huang and Nau proposed that since Thr has a secondary hydroxyl group, it limits the flexibility of Thr but much less than Val and Ile. Furthermore, if we compare Leu and Ile; Leucine seems to be more flexible than Isoleucine. Huang and Nau state that this is because Ile is a β-branched amino acid.

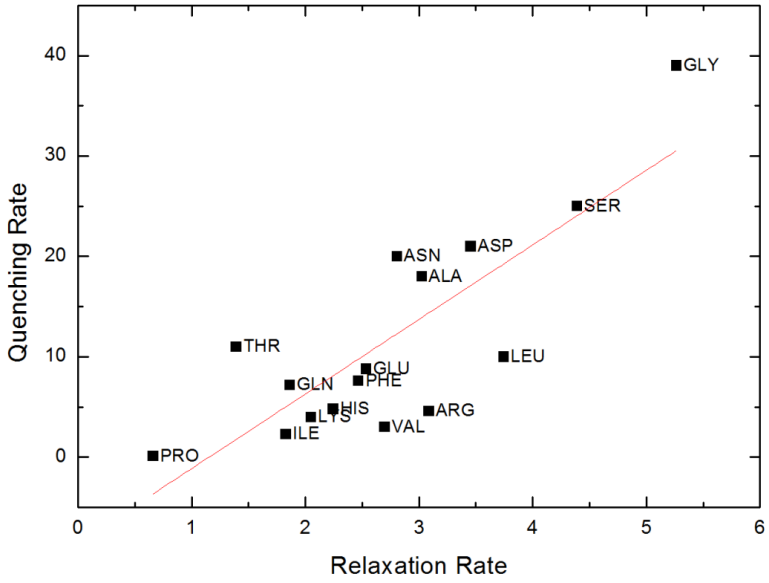


Figure 31. Comparison of the relaxation rates determined via DRIS with the quenching rates obtained experimentally by [86].

The scale between relaxation time determined by DRIS and molecular weight is given in Figure 33. The relation is determined by excluding Thr since it is an outlier. Here two C-beta branched amino acids, Thr and Ile, relax more slowly than the other amino acids while the other C-beta branched amino acid Val relaxes faster than these and our explanation parallels that of Huang and Nau.

The order of relaxation time τ_1 is Gly< Ser< Leu< Cys< Asp< Arg< Ala< Asn< Val< Tyr,Glu< Met< Phe< His< Lys< Trp< Gln< Ile< Thr< Pro. Similarly, the order of relaxation time τ_2 is Ser< Gly< Cys< Asp, Leu< Arg< Ala< Asn< Val< Tyr< Phe< Glu< Met< His< Lys< Trp< Gln< Thr< Ile< Pro.

Although the results fit together with the data given by Huang and Nau [86], there are some differences. This may be a consequence of different macrocyclization probabilities of different amino acid sequences, which plays a role in the experiments of Huang and Nau but is not consequential in our single amino acid calculations. The proportion of macrocyclic constituents at equilibrium is related to the macrocyclizatiıon probability that is the probability of coincidence of the two ends of a sequence [103-105]. Hence, it effects the statistical configurations of the chain. Therefore, it is related to chain flexibility and the choice of chain length may affect measurements of the quenching.

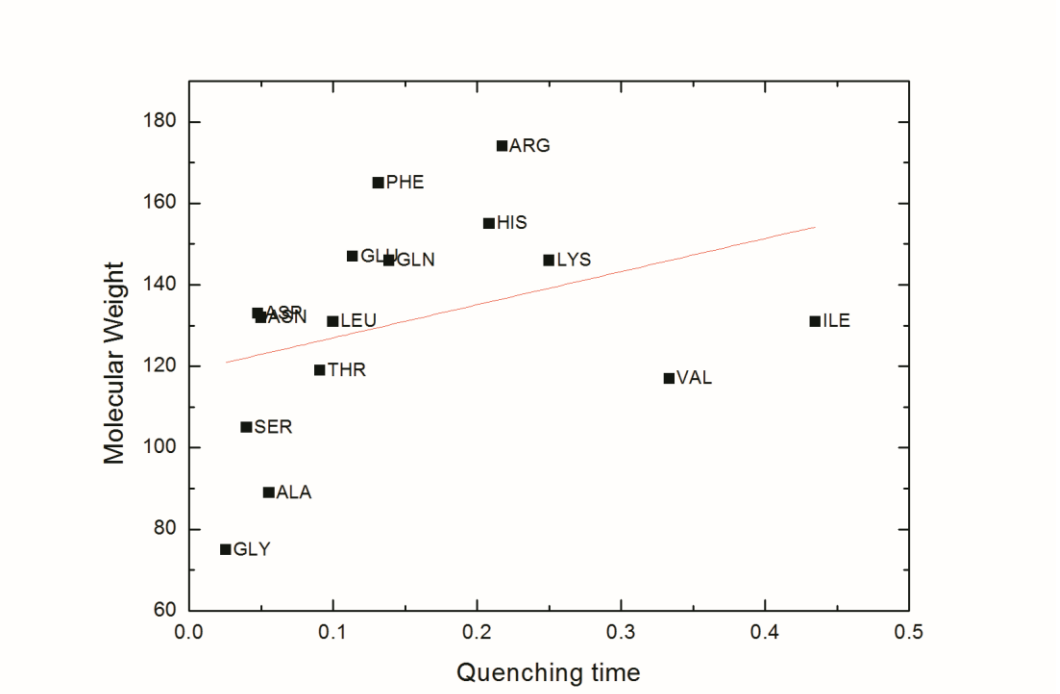


Figure 32. The relation between quenching time [86] and molecular weight of the fifteen amino acids. Pro is excluded.

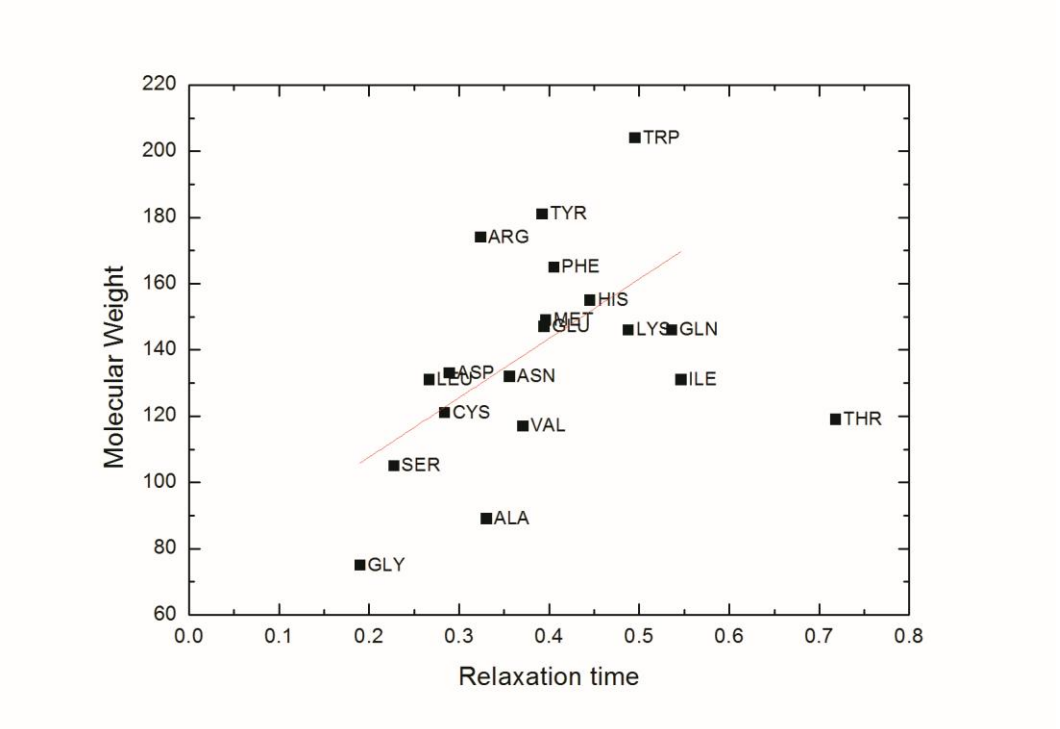


Figure 33. The relation between relaxation time determined by DRIS and molecular weight of all amino acids except Pro and Thr. The relation is determined by excluding Thr since it is an outlier.

5.5 CONCLUDING REMARKS

The rotational transitions between the torsion states of the bonds have an important role in local dynamics. DRIS is an efficient computational method that yields determination of a dynamic property of the chain as a function of local transitions based on the frequencies of each state.

In this study, the rates of first order transitions between the specified states are determined by the conditional probability curves that obtained from MD simulation trajectories of amino acids. Since the probability curves converges to the equilibrium probabilities faster at higher temperatures, the calculations are performed for 400, 600, 800, and 1000 K. The initial slopes of the time-delayed conditional probability curves give the transition rates that to be implemented in the rate matrices. The identified dynamic properties $\langle f_1(t) \rangle$ and $\langle f_2(t) \rangle$ are determined based on

the DRIS method. Since these functions decays exponentially, respective relaxation times at different temperatures are obtained.

Our calculations showed that $\langle f_1(t) \rangle$ and $\langle f_2(t) \rangle$ have Arrhenius behavior. Based on this observation, we extrapolated the rates for higher temperatures and obtain the relaxation times at 310 K for all amino acids.

Relaxation times of the amino acids are related to the flexibility of the molecules, smaller relaxation times indicating higher flexibility while larger relaxation times indicating higher rigidity.

We applied the DRIS method to analyze the internal dynamics of the amino acids. The DRIS method is a general method that can be applied to determine the internal dynamics of the peptides. One may adopt the same method to analyze dynamics of different length peptides in the random coil state.

Chapter 6

CONCLUSION

In this dissertation, we analyzed the properties of the protein in the random coil state. Having a complete description of the random coil state is motivated by (i) understanding primary sequence of the proteins, (ii) characterization of the denatured proteins, (iii) analyzing the intrinsically disordered proteins. The functional, structural and dynamical properties for the random coil state are investigated.

The statistical thermodynamics properties of the proteins are obtained using rotational isomeric states (RIS) model of polymer theory and based on the knowledge based data. The method is utilized to determine conformational energy, entropy, mean energy and heat capacity for the random coil state. The explicit thermodynamic expressions are derived for these functions using a curve fitting technique. The results show that per residue thermodynamic properties scale only with the temperature for the proteins in the random coil state. We provided a computational scheme to analyze the thermodynamic properties for the random coil state. This scheme can be extended to determine other thermodynamic potentials by using existing Legendre transformation techniques of the thermodynamic theory.

In the second study, we proposed a computational method for predicting "n" high probability conformations of a given peptide sequence in the random coil state. The presented scheme is based on a hidden Markov model and Viterbi algorithm. A priori probabilities are determined from a knowledge based database, and a posteriori probabilities are obtained using rotational isomeric states (RIS) model. Existing studies are mainly focused on predicting the most probable state of the residues that does not provide a conformational sampling. However, prediction of a probability distribution is necessary for a proper analysis of the conformational space. The proposed method is the first study that predicts the high probability conformations of a given peptide sequence in the random coil state. This approach would be useful in targeted drug design since finding the conformation that matches a given surface area is necessary to be effective.

Furthermore, using a torsion state library for the intrinsically disordered proteins one may predict the high probability conformations using the proposed scheme.

In the last study, we have analyzed conformational transitions of amino acids by DRIS model. The torsion states of the amino acids and the probabilities of the states are determined by MD simulations of Gly-Gly-X-Gly-Gly pentapeptides where X represents a given amino acid. Dynamic properties are important for a full description of the amino acids. The DRIS model based on the MD simulations is an efficient computational scheme for providing local dynamic properties. The relaxation times of the amino acids for given dynamic properties are calculated. The results are in good agreement with experimental results and indicate that the relaxation time of the amino acids scale with the flexibility of the molecule.

BIBLIOGRAPHY

1. Flory, P.J., *Statistical mechanics of chain molecules*. 1969, New York,: Interscience Publishers. xix, 432 p.
2. Brant, D.A. and P.J. Flory, *The Configuration of Random Polypeptide Chains. II. Theory.* Journal of the American Chemical Society, 1965. **87**(13): p. 2791-2800.
3. Flory, P.J. and R.L. Jernigan, *Second and Fourth Moments of Chain Molecules*. Vol. 42. 1965: AIP. 3509-3519.
4. Brant, D.A., A.E. Tonelli, and P.J. Flory, *The Configurational Statistics of Random Poly(lactic acid) Chains. II. Theory.* Macromolecules, 1969. **2**(3): p. 228-235.
5. Conrad, J.C. and P.J. Flory, *Moments and Distribution Functions for Polypeptide Chains. Poly-L-alanine.* Macromolecules, 1976. **9**(1): p. 41-47.
6. Rehahn, M., W.L. Mattice, and U. Suter, *Rotational isomeric state models in macromolecular systems*. 1997: Springer.
7. Engin, O., M. Sayar, and B. Erman, *The introduction of hydrogen bond and hydrophobicity effects into the rotational isomeric states model for conformational analysis of unfolded peptides.* Phys Biol, 2009. **6**(1): p. 016001.
8. Tanford, C., *Protein Denaturation*, in *Advances in Protein Chemistry*, M.L.A.J.T.E. C.B. Anfinsen and M.R. Frederic, Editors. 1968, Academic Press. p. 121-282.
9. Dill, K.A. and D. Shortle, *Denatured states of proteins.* Annu Rev Biochem, 1991. **60**: p. 795-825.
10. Tompa, P., *Unstructural biology coming of age.* Curr Opin Struct Biol, 2011. **21**(3): p. 419-25.
11. Orosz, F. and J. Ovádi, *Proteins without 3D structure: definition, detection and beyond.* Bioinformatics, 2011. **27**(11): p. 1449-1454.
12. Flory, P.J., *Foundations of Rotational Isomeric State Theory and General Methods for Generating Configurational Averages.* Macromolecules, 1974. **7**(3): p. 381-392.
13. Callen, H.B., *Thermodynamics and an introduction to thermostatistics*. 2nd ed ed. 1985, New York: Wiley. xvi, 493 p.
14. Pappu, R.V., R. Srinivasan, and G.D. Rose, *The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding.* Proc Natl Acad Sci U S A, 2000. **97**(23): p. 12565-70.
15. Zaman, M.H., et al., *Investigations into Sequence and Conformational Dependence of Backbone Entropy, Inter-basin Dynamics and the Flory Isolated-pair Hypothesis for Peptides.* Journal of Molecular Biology, 2003. **331**(3): p. 693-711.
16. Ohkubo, Y.Z. and C.L. Brooks, 3rd, *Exploring Flory's isolated-pair hypothesis: statistical mechanics of helix-coil transitions in polyalanine and the C-peptide from RNase A.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 13916-21.
17. Jha, A.K., et al., *Statistical coil model of the unfolded state: resolving the reconciliation problem.* Proc Natl Acad Sci U S A, 2005. **102**(37): p. 13099-104.
18. Jha, A.K., et al., *Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library.* Biochemistry, 2005. **44**(28): p. 9691-702.
19. Keskin, O., et al., *Relationships between amino acid sequence and backbone torsion angle preferences.* Proteins: Structure, Function, and Bioinformatics, 2004. **55**(4): p. 992-998.

20. Esposito, L., et al., *Correlation between [omega] and [psi] Dihedral Angles in Protein Structures.* Journal of Molecular Biology, 2005. **347**(3): p. 483-487.
21. Colubri, A., et al., *Minimalist representations and the importance of nearest neighbor effects in protein folding simulations.* J Mol Biol, 2006. **363**(4): p. 835-57.
22. Lennox, K.P., et al., *Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics.* J Am Stat Assoc, 2009. **104**(486): p. 586-596.
23. Karplus, P.A., *Experimentally observed conformation-dependent geometry and hidden strain in proteins.* Protein Sci, 1996. **5**(7): p. 1406-20.
24. Ovacik, A.M., *Statistical mechanics and local dynamics of denaturated proteins*, in *Computational Science and Engineering*. 2005, Koc University: Istanbul. p. 75.
25. Szasz, C.S., et al., *Protein disorder prevails under crowded conditions.* Biochemistry, 2011. **50**(26): p. 5834-44.
26. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins, 2000. **41**(3): p. 415-27.
27. Sickmeier, M., et al., *DisProt: the Database of Disordered Proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D786-93.
28. Dunker, A.K., et al., *Protein disorder and the evolution of molecular recognition: theory, predictions and observations.* Pac Symp Biocomput, 1998: p. 473-84.
29. Romero, P., et al., *Thousands of proteins likely to have long disordered regions.* Pac Symp Biocomput, 1998: p. 437-48.
30. Garner, E., et al., *Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization.* Genome Inform Ser Workshop Genome Inform, 1998. **9**: p. 201-213.
31. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.* J Mol Biol, 1999. **293**(2): p. 321-31.
32. Dunker, A.K., et al., *Intrinsic disorder and protein function.* Biochemistry, 2002. **41**(21): p. 6573-82.
33. Dunker, A.K., et al., *Intrinsic protein disorder in complete genomes.* Genome Inform Ser Workshop Genome Inform, 2000. **11**: p. 161-71.
34. Tompa, P., *Intrinsically unstructured proteins evolve by repeat expansion.* Bioessays, 2003. **25**(9): p. 847-55.
35. Gunasekaran, K., et al., *Extended disordered proteins: targeting function with less scaffold.* Trends Biochem Sci, 2003. **28**(2): p. 81-5.
36. Fuxreiter, M., et al., *Preformed structural elements feature in partner recognition by intrinsically unstructured proteins.* J Mol Biol, 2004. **338**(5): p. 1015-26.
37. Tompa, P. and P. Csermely, *The role of structural disorder in the function of RNA and protein chaperones.* FASEB J, 2004. **18**(11): p. 1169-75.
38. Tompa, P., *The interplay between structure and function in intrinsically unstructured proteins.* FEBS Lett, 2005. **579**(15): p. 3346-54.
39. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nature Reviews Molecular Cell Biology, 2005. **6**(3): p. 197-208.
40. Dunker, A.K., et al., *Flexible nets. The roles of intrinsic disorder in protein interaction networks.* FEBS J, 2005. **272**(20): p. 5129-48.
41. Dunker, A.K., et al., *Intrinsically disordered protein.* J Mol Graph Model, 2001. **19**(1): p. 26-59.

42. Romero, P., et al., *Sequence complexity of disordered protein.* Proteins, 2001. **42**(1): p. 38-48.
43. Brown, C.J., A.K. Johnson, and G.W. Daughdrill, *Comparing models of evolution for ordered and disordered proteins.* Mol Biol Evol, 2010. **27**(3): p. 609-21.
44. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Intrinsically disordered proteins in human diseases: introducing the D2 concept.* Annu Rev Biophys, 2008. **37**: p. 215-46.
45. Dosztanyi, Z., B. Meszaros, and I. Simon, *Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins.* Brief Bioinform, 2010. **11**(2): p. 225-43.
46. He, B., et al., *Predicting intrinsic disorder in proteins: an overview.* Cell Res, 2009. **19**(8): p. 929-49.
47. Drews, J., *Drug discovery: a historical perspective.* Science, 2000. **287**(5460): p. 1960-4.
48. Tompa, P., *Structure and function of intrinscially disordered proteins*. 2010, Boca Raton: Chapman & Hall/CRC Press. xxvii, 331 p.
49. Cheng, Y., et al., *Rational drug design via intrinsically disordered protein.* Trends Biotechnol, 2006. **24**(10): p. 435-42.
50. Ewens, W.J. and G.R. Grant, *Statistical methods in bioinformatics : an introduction*. 2001, New York: Springer. xix, 476 p.
51. Unal, B., A. Gursoy, and B. Erman, *VitAL: Viterbi algorithm for de novo peptide design.* PloS one, 2010. **5**(6): p. e10926.
52. Rabiner, L., *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE, 1989. **77**(2): p. 257-286.
53. Bishop, M.J. and E.A. Thompson, *Maximum likelihood alignment of DNA sequences.* Journal of Molecular Biology, 1986. **190**(2): p. 159-165.
54. Viterbi, A.J., *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* Information Theory, IEEE Transactions on, 1967. **13**(2): p. 260-269.
55. Forney, G.D., Jr., *The viterbi algorithm.* Proceedings of the IEEE, 1973. **61**(3): p. 268-278.
56. Bahar, I. and B. Erman, *Investigation of local motions in polymers by the dynamic rotational isomeric state model.* Macromolecules, 1987. **20**(6): p. 1368-1376.
57. Bahar, I., B. Erman, and L. Monnerie, *Application of the dynamic rotational isomeric states model to poly(ethylene oxide) and comparison with nuclear magnetic relaxation data.* Macromolecules, 1989. **22**(5): p. 2396-2403.
58. Bahar, I. and B. Erman, *Activation energies of local conformational transitions in polymer chains.* Macromolecules, 1987. **20**(9): p. 2310-2311.
59. Bahar, I., B. Erman, and L. Monnerie, *Comparison of dynamic rotational isomeric state results with previous expressions for local chain motion.* Macromolecules, 1989. **22**(1): p. 431-437.
60. Bahar, I., *Stochastics of rotational isomeric transitions in polymer chains.* The Journal of Chemical Physics, 1989. **91**(10): p. 6525-6531.
61. Bahar, I. and W.L. Mattice, *Efficient calculation of the intramolecular contribution to orientational autocorrelation functions using dynamic rotational isomeric state theory.* Macromolecules, 1990. **23**(10): p. 2719-2723.
62. Bahar, I., N. Neuburger, and W.L. Mattice, *Mechanism of local conformational transitions in poly(dialkylsiloxanes). Molecular dynamics simulations and dynamic rotational isomeric state approach.* Macromolecules, 1992. **25**(18): p. 4619-4625.
63. Haliloglu, T., et al., *A dynamic rotational isomeric state approach for extension of the time scale of the local dynamics observed in fully atomistic molecular dynamics*

simulations: Application to polybutadiene. The Journal of Chemical Physics, 1996. **104**(12): p. 4828-4834.

64. Jernigan, R.L., *Internal Relaxation in Short Chains Bearing Terminal Polar Groups*, in *Dielectric Properties of Polymers*, F. Karasz, Editor. 1972, Springer US. p. 99-128.
65. BAHAR, #160, and I., *Stochastics of rotational isomeric transitions in polymer chains*. Vol. 91. 1989, Melville, NY, ETATS-UNIS: American Institute of Physics.
66. Fitzkee, N.C., P.J. Fleming, and G.D. Rose, *The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB.* Proteins, 2005. **58**(4): p. 852-4.
67. Ormeci, L., et al., *Computational basis of knowledge-based conformational probabilities derived from local- and long-range interactions in proteins.* Proteins: Structure, Function, and Bioinformatics, 2007. **66**(1): p. 29-40.
68. Unal, E.B., A. Gursoy, and B. Erman, *Conformational energies and entropies of peptides, and the peptide-protein binding problem.* Phys Biol, 2009. **6**(3): p. 036014.
69. Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan, *Stereochemistry of polypeptide chain configurations.* Journal of Molecular Biology, 1963. **7**(1): p. 95-99.
70. Avbelj, F. and R.L. Baldwin, *Origin of the neighboring residue effect on peptide backbone conformation.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(30): p. 10967-10972.
71. Avbelj, F., et al., *Intrinsic backbone preferences are fully present in blocked amino acids.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(5): p. 1272-1277.
72. Gibrat, J.F., B. Robson, and J. Garnier, *Influence of the local amino acid sequence upon the zones of the torsional angles phi and psi adopted by residues in proteins.* Biochemistry, 1991. **30**(6): p. 1578-86.
73. Bystroff, C., V. Thorsson, and D. Baker, *HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins.* Journal of Molecular Biology, 2000. **301**(1): p. 173-190.
74. Kuang, R., C.S. Leslie, and A.-S. Yang, *Protein backbone angle prediction with machine learning approaches.* Bioinformatics, 2004. **20**(10): p. 1612-1621.
75. Zimmermann, O. and U.H.E. Hansmann, *Support vector machines for prediction of dihedral angle regions.* Bioinformatics, 2006. **22**(24): p. 3009-3015.
76. Lovell, S.C., et al., *Structure validation by Cα geometry: ϕ,ψ and Cβ deviation.* Proteins: Structure, Function, and Bioinformatics, 2003. **50**(3): p. 437-450.
77. Shortle, D., *Composites of local structure propensities: Evidence for local encoding of long-range structure.* Protein Science, 2002. **11**(1): p. 18-26.
78. Xin, G., et al. *Protein Backbone Dihedral Angle Prediction Based on Probabilistic Models*. in *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*. 2010.
79. Helles, G. and R. Fonseca, *Predicting dihedral angle probability distributions for protein coil residues from primary sequence using neural networks.* BMC Bioinformatics, 2009. **10**: p. 338.
80. Boomsma, W., et al., *A generative, probabilistic model of local protein structure.* Proceedings of the National Academy of Sciences, 2008. **105**(26): p. 8932-8937.
81. Wang, Z., et al., *Protein 8-class secondary structure prediction using conditional neural fields.* Proteomics, 2011. **11**(19): p. 3786-92.

82. Daughdrill, G.W., et al., *Understanding the structural ensembles of a highly extended disordered protein.* Molecular BioSystems, 2012. **8**(1): p. 308-319.
83. Schweitzer-Stenner, R., *Conformational propensities and residual structures in unfolded peptides and proteins.* Molecular BioSystems, 2012. **8**(1): p. 122-133.
84. Uversky, V.N., *Natively unfolded proteins: a point where biology waits for physics.* Protein Sci, 2002. **11**(4): p. 739-56.
85. Fitzkee, N.C. and G.D. Rose, *Reassessing random-coil statistics in unfolded proteins.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(34): p. 12497-12502.
86. Huang, F. and W.M. Nau, *A Conformational Flexibility Scale for Amino Acids in Peptides.* Angewandte Chemie International Edition, 2003. **42**(20): p. 2269-2272.
87. Kleckner, I.R. and M.P. Foster, *An introduction to NMR-based approaches for measuring protein dynamics.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2011. **1814**(8): p. 942-968.
88. Hamaneh, M.B., L. Zhang, and M. Buck, *A Direct Coupling between Global and Internal Motions in a Single Domain Protein? MD Investigation of Extreme Scenarios.* Biophysical journal, 2011. **101**(1): p. 196-204.
89. Beck, D.A.C., et al., *The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins.* Proceedings of the National Academy of Sciences, 2008. **105**(34): p. 12259-12264.
90. Plaxco, K., et al., *The effects of guanidine hydrochloride on the 'random coil' conformations and NMR chemical shifts of the peptide series GGXGG.* Journal of Biomolecular NMR, 1997. **10**(3): p. 221-230.
91. Schwarzinger, S., et al., *Random coil chemical shifts in acidic 8 M urea: Implementation of random coil shift data in NMRView.* Journal of Biomolecular NMR, 2000. **18**(1): p. 43-48.
92. Ding, L., et al., *The Pentapeptide GGAGG Has PII Conformation.* Journal of the American Chemical Society, 2003. **125**(27): p. 8092-8093.
93. Shi, Z., et al., *Conformation of the backbone in unfolded proteins.* Chem Rev, 2006. **106**(5): p. 1877-97.
94. Wang, L. and J. Markley, *Empirical correlation between protein backbone 15N and 13C secondary chemical shifts and its application to nitrogen chemical shift re-referencing.* Journal of Biomolecular NMR, 2009. **44**(2): p. 95-99.
95. Van Der Spoel, D., et al., *GROMACS: Fast, flexible, and free.* Journal of Computational Chemistry, 2005. **26**(16): p. 1701-1718.
96. Jorgensen, W., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
97. Jorgensen, W.L., D.S. Maxwell, and J. Tirado-Rives, *Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids.* Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.
98. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath.* The Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
99. Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi, *Optimization by Simulated Annealing.* Science, 1983. **220**(4598): p. 671-680.
100. Bayrak, C.S. and B. Erman, *Predicting most probable conformations of a given peptide sequence in the random coil state.* Molecular BioSystems, 2012. **8**(11): p. 3010-3016.

101. Laidler, K.J., *The development of the Arrhenius equation.* Journal of Chemical Education, 1984. **61**(6): p. 494.
102. Hudgins, R.R., et al., *A Fluorescence-Based Method for Direct Measurement of Submicrosecond Intramolecular Contact Formation in Biopolymers: An Exploratory Study with Polypeptides.* Journal of the American Chemical Society, 2002. **124**(4): p. 556-564.
103. Flory, P.J. and J.A. Semlyen, *Macrocyclization Equilibrium Constants and the Statistical Configuration of Poly(dimethylsiloxane) Chains.* Journal of the American Chemical Society, 1966. **88**(14): p. 3209-3212.
104. Suter, U.W., M. Mutter, and P.J. Flory, *Macrocyclization equilibriums. 2. Poly(dimethylsiloxane).* Journal of the American Chemical Society, 1976. **98**(19): p. 5740-5745.
105. Mutter, M., U.W. Suter, and P.J. Flory, *Macrocyclization equilibriums. 3. Poly(6-aminocaproamide).* Journal of the American Chemical Society, 1976. **98**(19): p. 5745-5748.